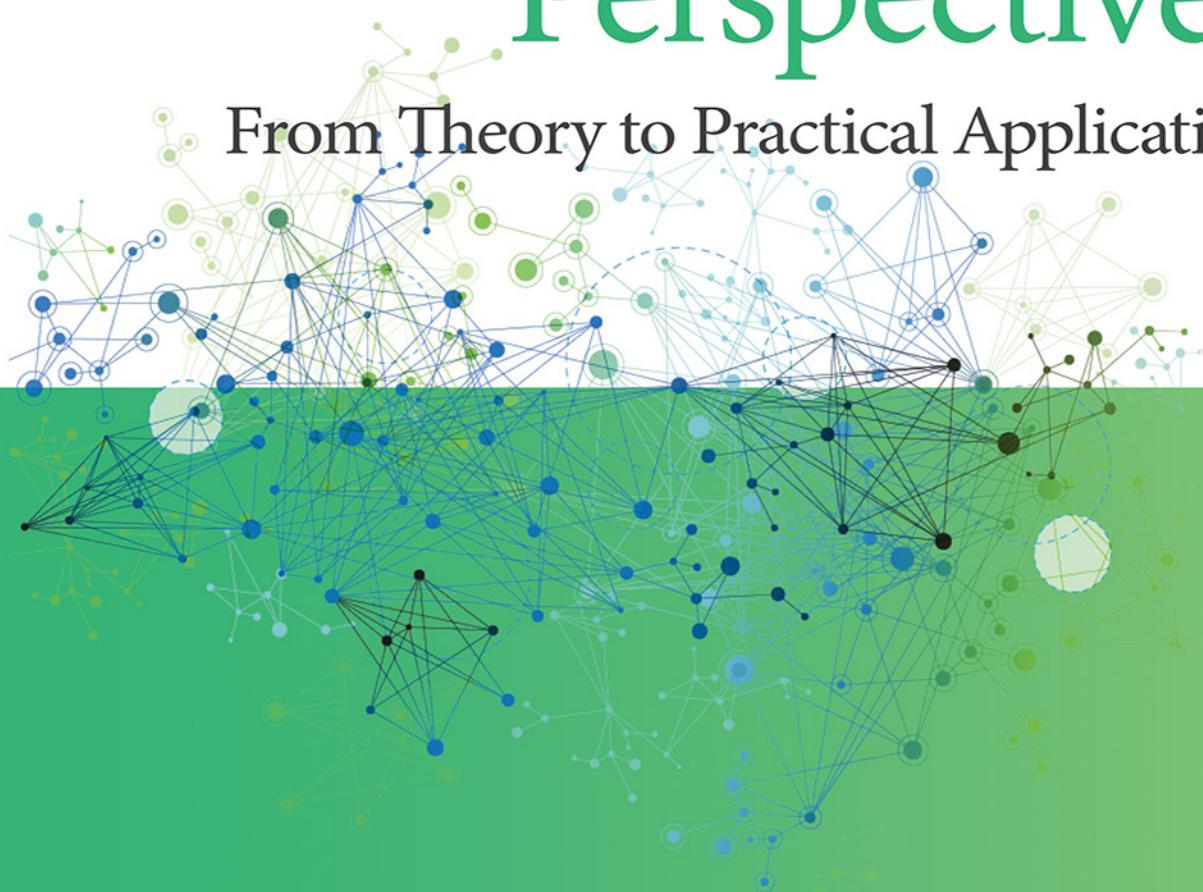


Metabolomics Perspectives

From Theory to Practical Application



Edited by
Jacopo Troisi



Metabolomics Perspectives

From Theory to Practical Application

This page intentionally left blank

Metabolomics Perspectives

From Theory to Practical Application

Edited by

Jacopo Troisi

*Department of Medicine, Surgery and Dentistry
“Scuola Medica Salernitana”, University of Salerno, Baronissi,
Salerno, Italy;
Theoreo Srl—Spin-off Company of the University of Salerno,
Montecorvino Pugliano, Salerno, Italy;
Department of Chemistry and Biology “A. Zambelli”,
University of Salerno, Fisciano, Salerno, Italy*



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2022 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-323-85062-9

For Information on all Academic Press publications
visit our website at <https://www.elsevier.com/books-and-journals>

Publisher: Stacy Masucci
Acquisitions Editor: Peter B. Linsley
Editorial Project Manager: Susan Ikeda
Production Project Manager: Punithavathy Govindaradjane
Cover Designer: Christian Bilbow
Typeset by MPS Limited, Chennai, India



Working together
to grow libraries in
developing countries
www.elsevier.com • www.bookaid.org

Contents

List of contributors	xvii
Forewordxxi
Introduction.....	xxiii

Section 1 Fundamentals

CHAPTER 1 System biology	3
<i>Elizabeth C. Plunk, Weston S. Chambers, and Sean M. Richards</i>	
Introduction.....	3
Genomics	4
Introduction.....	4
Genomic tools.....	4
Epigenetics.....	7
Introduction.....	7
Epigenetic tools	9
Transcriptomics	11
Introduction.....	11
Transcriptomic tools	12
Proteomics	13
Introduction.....	13
Proteomic tools	14
Metabolomics.....	16
Introduction.....	16
Metabolomic tools	16
References.....	19
CHAPTER 2 Experimental design in metabolomics	27
<i>Allycia Y. Lee, Jacopo Troisi, and Steven J.K. Symes</i>	
Introduction.....	27
Applications of metabolomic experiments	28
Biomarker discovery.....	28
Detection of altered biochemical pathways	29
Monitoring of response to stimuli.....	30
Untargeted and targeted approaches	30
Untargeted metabolomics	30
Targeted metabolomics	31
Sample types.....	31

Metabolically active versus metabolically inactive	31
Tissue and cells	32
Whole blood, plasma, and serum.....	33
Urine	36
Other biofluids	37
Analytical methodologies.....	42
Nuclear magnetic resonance.....	42
Mass spectrometry	42
Techniques without sampling.....	44
Sample preparation	45
Quenching	45
Extraction.....	46
Sample clean up.....	47
Nuclear magnetic resonance.....	49
Gas chromatography-mass spectrometry	49
Liquid chromatography-mass spectrometry.....	52
Identification and quantification of metabolites	52
Quantification	52
Identification.....	54
Quality control.....	55
Conclusion	56
References.....	57
Further reading	61
 CHAPTER 3 Separation techniques	63
<i>Martina Catani, Simona Felletti, and Flavio Antonio Franchina</i>	
The role of the separation processes in metabolomics research	63
Sample preparation	65
Sample extraction techniques.....	66
Derivatization.....	69
Fundamentals of chromatography	69
Definitions and classifications.....	70
Retention.....	72
Selectivity	74
Efficiency of separation.....	74
Resolution	75
Peak capacity	76
Qualitative and quantitative analysis in chromatography.....	76

Liquid chromatography	78
Instrumentation	78
Principal separation modes.....	79
Detectors	81
Gas chromatography	84
Mobile phase and flow control.....	85
Temperature zones.....	87
Sample introduction and inlets.....	87
Column, stationary phases, and separation	88
Detectors	90
Multidimensional chromatography	92
Concept of multidimensionality	92
Practical and instrumental aspects	96
Other separation techniques	99
Capillary electrophoresis	99
Supercritical fluid chromatography.....	100
Chiral chromatography	101
References.....	103

CHAPTER 4 Mass spectrometry in metabolomics 109

Angela Amoresano and Piero Pucci

Mass spectrometry	109
Mass spectrum	109
Isotopes	110
Resolution and accuracy.....	111
Mass spectrometer	113
System for sample introduction	114
Ion sources	114
Mass analyzer	114
Ion detector	114
Ion sources	115
Mass analyzers.....	119
Tandem mass spectrometry	125
Instruments for tandem mass spectrometry analysis	125
Tandem mass spectrometry scan modes	127
Untargeted metabolomics in complex samples	129
Analytical techniques in mass spectrometry -based metabolomics	132
Gas chromatography-mass spectrometry	133
Liquid chromatography-tandem mass spectrometry	133

Imaging mass spectrometry	134
Data analysis.....	136
Applications	136
Metabolomic analysis for clinical biomarker discovery.....	138
Metabolomics in drug development.....	139
Metabolomics in nutrition science	140
Metabolomics in toxicology	141
Metabolomics in forensic science	142
References.....	144
Further reading	147

CHAPTER 5 Nuclear magnetic resonance in metabolomics 149

*Abdul-Hamid Emwas, Kacper Szczepski,
Benjamin Gabriel Poulson, Ryan McKay, Leonardo Tenori,
Edoardo Saccenti, Joanna Lachowicz, and
Mariusz Jaremko*

Introduction.....	149
Nuclear magnetic resonance spectroscopy	151
1D nuclear magnetic resonance	151
2D nuclear magnetic resonance spectroscopy	170
High-resolution magic-angle spinning nuclear magnetic resonance spectroscopy	174
Pure shift nuclear magnetic resonance.....	176
Recent advances	177
Improvements in nuclear magnetic resonance hardware and techniques and additional tools to aid in metabolomics studies.....	177
Nuclear magnetic resonance magnets.....	178
Nuclear magnetic resonance probes.....	179
Flow probes	179
Metabolomics databases and nuclear magnetic resonance software programs	181
Databases for nuclear magnetic resonance-based metabolomics	181
Use of software to analyze metabolite nuclear magnetic resonance data.....	183
Advantages of nuclear magnetic resonance spectroscopy.....	185
Reproducibility	186
Challenges and limitations	187
Sample preparation	188

Summary and future perspectives	191
References.....	192
CHAPTER 6 Targeted metabolomics	219
<i>Michele Costanzo, Marianna Caterino, and Margherita Ruoppolo</i>	
Targeted metabolomics	219
Inborn errors of metabolism.....	224
Application of targeted metabolomics to the newborn screening of inborn errors of metabolism	225
Examples of inborn error of metabolism diagnosed by the newborn screening	228
Methylmalonic acidemias.....	228
Propionic acidemia	229
Glutaric acidemia.....	229
Isovaleric acidemia.....	232
Phenylketonuria	232
Hereditary tyrosinemas.....	233
Maple syrup urine disease	233
Conclusion	234
References.....	234
CHAPTER 7 Approaches in untargeted metabolomics	237
<i>Jacopo Troisi, Sean M. Richards, Giovanni Scala, and Annamaria Landolfi</i>	
Introduction.....	237
Local and nonlocal metabolomics effects.....	238
Untargeted metabolomics application	240
Metabolomics profiling	242
Cardiovascular disease	244
Neurodegenerative disease	247
Limitations	249
Sources of metabolome variability	250
Key trends in untargeted metabolomics.....	252
Metabolome coverage.....	252
Moving metabolomics from laboratories to clinics	253
Metabolomics pipeline standardization.....	254
Sample size	254
Independent cohort to validate the results	254
Cause/effects disambiguation	256
Conclusion	258
References.....	258

Section 2 Data analysis

CHAPTER 8 Techniques for converting metabolomic data for analysis	265
<i>Jacopo Troisi, Sean M. Richards, Giovanni Troisi, and Giovanni Scala</i>	
Introduction.....	265
Data preprocessing	266
Mass spectrometry-based experiments.....	266
Nuclear magnetic resonance.....	271
Normalization	275
Internal standard normalization.....	275
Data pretreatment	278
Centering.....	280
Scaling	281
Conclusion	284
References.....	284
CHAPTER 9 Data analysis in metabolomics: from information to knowledge.....	287
<i>Jacopo Troisi, Giovanni Troisi, Giovanni Scala, and Sean M. Richards</i>	
Introduction.....	287
Exploratory analysis	287
Univariate approach.....	290
Multivariate approach.....	299
Unsupervised machine learning analysis	305
Introduction.....	305
Cluster analysis.....	306
Conclusion	314
Supervised machine learning.....	314
Introduction.....	314
Decision trees	317
Naïve Bayesian	322
Discriminant analysis	324
Artificial neural network	324
Support vector machine.....	329
Regressive models	332
Partial least square discriminant analysis	338
Classification model validation	344

Leave-one-out cross-validation	346
Leave-k-out cross-validation	346
<i>k</i> -fold cross validation	347
Permutation test	347
Class imbalance	348
Metrics to estimate the classification performances.....	349
Sampling strategies.....	351
Machine learning algorithms modification.....	354
Ensemble machine learning	354
Bagging.....	355
Boosting.....	356
Features selection	359
Features filtering.....	360
Boruta's algorithm.....	361
Genetic algorithm	362
Features generation.....	365
Embedded methods.....	367
Conclusions.....	367
Hyperparameters optimization	368
Parameters and hyperparameters in machine learning	368
Hyperparameters tuning	369
Appendix.....	372
References.....	374

CHAPTER 10 Relevant metabolites' selection strategies..... 381

Jos Hageman

Introduction.....	381
Low-level variable selection	382
Unsupervised low-level variable selection	383
Supervised low-level variable selection.....	384
Medium-level variable selection	386
Variable selection or wrapper methods.....	386
Stepwise regression	387
Global optimization algorithms.....	387
High-level variable selection.....	388
Embedded methods for the selection of variables.....	388
Decision trees	390
Random forests	391
Support vector machine.....	392
Heuristic approach.....	392

Bootstrap and stability selection	393
Cross validation	394
Concluding remarks.....	395
References.....	396
CHAPTER 11 Pathway analysis	399
<i>Rachel Cavill and Jildau Bouwman</i>	
Metabolites ontology	399
Introduction to ontologies	399
Ontologies for metabolites	399
Common metabolite databases.....	401
Common pathway databases	402
Metabolic pathway analysis	403
Overrepresentation.....	404
Enrichment.....	406
Metabolite set enrichment analysis	407
Kolmogorov–Smirnov test	407
Wilcoxon signed rank test.....	408
Topological methods	408
Tools for metabolomic pathway analysis	408
Conclusions.....	410
References.....	410

Section 3 Application

CHAPTER 12 Cell culture metabolomics and lipidomics	415
<i>Irina Alecu, Carmen Daniela Sosa-Miranda, Jagdeep K. Sandhu, Steffany A.L. Bennett, and Miroslava Cuperlovic-Culf</i>	
Introduction.....	415
Sample processing and experimentation for cell culture lipidomics and metabolomics.....	417
Methods for optimized metabolite and lipid extractions for cell culture analysis	417
Analysis of metabolic processes including metabolic flux	423
Methods and protocols for isolation and metabolomics of small extracellular vesicles from cell culture supernatants	427
Cell culture for isolation of small extracellular vesicles	429
Isolation of small extracellular vesicles using ultracentrifugation.....	430

Differential ultracentrifugation.....	430
Density gradient ultracentrifugation.....	430
Isolation of small extracellular vesicles using tangential flow filtration.....	432
Characterization of small extracellular vesicles	433
Metabolite extractions from cells and small extracellular vesicles.....	433
Sample preparation and analysis with nuclear magnetic resonance spectroscopy	435
Cell culture metabolomics and lipidomics data analysis.....	435
Cell culture metabolomics and cell modeling for the design and optimization of cell culture applications	436
Determination of major metabolic pathways or network from metabolomics or fluxomics data	437
Network analysis in cell culture metabolomics	437
Mechanistic modeling for cell culture optimization, design, and information gathering.....	441
Machine learning and hybrid models and artificial intelligence for cell design	444
References.....	447
 CHAPTER 13 Single cell metabolomics	457
<i>Minakshi Prasad, Mayukh Ghosh, and Rajesh Kumar</i>	
Introduction.....	457
Single-cell metabolomics in microbial technology	460
Single-cell metabolomics in plant science and agriculture	480
Diversified animal applications.....	481
Single-cell metabolomics in developmental biology	484
Single-cell metabolomics in aging and senescence study	485
Single-cell metabolomics in stem cell biology	486
Single-cell metabolomics in functional genomics	488
Single-cell metabolomics in nutrition research	489
Single-cell metabolomics in environmental biology	490
Single-cell metabolomics in system biology	491
Single-cell metabolomics in immunology	492
Single-cell metabolomics in detection of metabolite dynamicity and pathway modulation	494
Single-cell metabolomics in clinical metabolism and disease perspective	496
Conclusion and future prospect.....	498
References.....	499

CHAPTER 14 Gut microbiota-derived metabolites in host physiology	515
<i>Francesco Strati and Federica Facciotti</i>	
Introduction.....	515
Metabolomics methods in host-microbiome studies.....	516
Fermentable metabolites and short chain fatty acids.....	518
Secondary bile acids	520
Amino acids- and tryptophan-derived metabolites	521
Additional microbially derived metabolites.....	523
Perspectives and future directions.....	524
References.....	525
Further reading	533
CHAPTER 15 MALDI—mass spectrometry imaging: the metabolomic visualization	535
<i>Emanuela Salviati, Eduardo Sommella, and Pietro Campiglia</i>	
Introduction.....	535
Basics of MALDI mass spectrometry imaging.....	536
Matrix choice and application.....	537
Tissue preparation for MALDI mass spectrometry imaging analysis.....	539
MALDI mass spectrometry imaging instrumentation	541
MALDI mass spectrometry imaging of endogenous metabolites.....	542
Metabolite annotation and quantitation in MALDI mass spectrometry imaging	547
Conclusion and future perspectives	547
References.....	548
CHAPTER 16 Metabolomics for oncology	553
<i>Susan Costantini and Alfredo Budillon</i>	
Introduction.....	553
Reprogramming of cancer cell metabolism	555
Glucose and Warburg effect.....	555
Lactate shuttle due to tumor hypoxia and Warburg effect.....	556
Glutamine metabolism.....	556
Serine metabolism	557
Methionine metabolism	558
Metabolism of arginine and ornithine involved in linking tricarboxylic acid and urea cycles.....	559

Proline metabolism	559
Lipid synthesis pathway	560
Nucleotide biosynthesis pathway	561
Applications and examples of human cancer metabolomics....	561
Serum/plasma metabolomics studies	562
Urine metabolomics studies	571
Tissue metabolomics studies	576
Fecal metabolomics studies.....	581
Saliva metabolomics studies	584
Metabolomics studies on other biological matrices	587
Conclusion	588
References.....	588
CHAPTER 17 Metabolomics as a tool for precision medicine	605
<i>Edoardo Saccenti and Leonardo Tenori</i>	
Systems approaches and systems medicine	605
Individual phenotyping using nuclear magnetic resonance.....	608
Applications	613
References.....	616
CHAPTER 18 Metabolomics in public health	625
<i>Pierpaolo Cavallo</i>	
Introduction.....	625
Data integration	627
System biology and metabolomics in public health	628
Longitudinal and life-long studies in metabolomics	630
Quantitative methods are necessary	631
Big data and metabolomics in public health	634
Policies, training, and resources.....	637
Final remarks	638
References.....	641
Index	643

This page intentionally left blank

List of contributors

Irina Alecu

Neural Regeneration Laboratory, Ottawa Institute of Systems Biology, Brain and Mind Research Institute, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada

Angela Amoresano

Department of Chemical Sciences, University of Naples Federico II, Naples, Italy

Steffany A.L. Bennett

Neural Regeneration Laboratory, Ottawa Institute of Systems Biology, Brain and Mind Research Institute, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada

Jildau Bouwman

Department of Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research (TNO), Zeist, The Netherlands

Alfredo Budillon

Experimental Pharmacology Unit—Istituto Nazionale Tumori-IRCCS Fondazione G. Pascale, Naples, Italy

Pietro Campiglia

Department of Pharmacy, University of Salerno, Fisciano, Salerno, Italy

Martina Catani

Department of Chemical, Pharmaceutical, and Agricultural Sciences, University of Ferrara, Ferrara, Italy

Marianna Caterino

Department of Molecular Medicine and Medical Biotechnology, University of Naples “Federico II”, Naples, Italy; CEINGE—Biotecnologie Avanzate s.c.ar.l., Naples, Italy

Pierpaolo Cavallo

Department of Physics, University of Salerno, Fisciano, Salerno, Italy; Complex Systems Institute-National Research Council (ISC-CNR), Rome, Italy

Rachel Cavill

Data Science and Knowledge Engineering, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands

Weston S. Chambers

Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States

Susan Costantini

Experimental Pharmacology Unit—Istituto Nazionale Tumori-IRCCS Fondazione G. Pascale, Naples, Italy

Michele Costanzo

Department of Molecular Medicine and Medical Biotechnology, University of Naples “Federico II”, Naples, Italy; CEINGE—Biotecnologie Avanzate s.c.ar.l., Naples, Italy

Miroslava Cuperlovic-Culf

Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada; Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada

Abdul-Hamid Emwas

Core Labs, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Federica Facciotti

Mucosal Immunology Lab, Department of Experimental Oncology, IEO—European Institute of Oncology IRCCS, Milan, Italy

Simona Felletti

Department of Chemical, Pharmaceutical, and Agricultural Sciences, University of Ferrara, Ferrara, Italy

Flavio Antonio Franchina

Department of Chemical, Pharmaceutical, and Agricultural Sciences, University of Ferrara, Ferrara, Italy

Mayukh Ghosh

Department of Veterinary Physiology and Biochemistry, RGSC, Banaras Hindu University, Mirzapur, India

Jos Hageman

Biometris, Applied Statistics, Wageningen University & Research, Wageningen, The Netherlands

Mariusz Jaremkó

Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Rajesh Kumar

Department of Veterinary Physiology and Biochemistry, Lala Lajpat Rai University of Veterinary and Animal Sciences, Hisar, India

Joanna Lachowicz

Department of Medical Sciences and Public Health, Università di Cagliari, Cittadella Universitaria, Monserrato, Italy

Annamaria Landolfi

Department of Medicine, Surgery and Dentistry, “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy

Allycia Y. Lee

Department of Chemistry and Physics, University of Tennessee-Chattanooga, Chattanooga, TN, United States

Ryan McKay

Department of Chemistry, University of Alberta, Edmonton, AB, Canada

Elizabeth C. Plunk

Department of Environmental Medicine, University of Rochester Medical School, Rochester, NY, United States

Benjamin Gabriel Poulsen

Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Minakshi Prasad

Department of Animal Biotechnology, Lala Lajpat Rai University of Veterinary and Animal Sciences, Hisar, India

Piero Pucci

CEINGE Advanced Biotechnology, Naples, Italy

Sean M. Richards

Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States; Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States

Margherita Ruoppolo

Department of Molecular Medicine and Medical Biotechnology, University of Naples “Federico II”, Naples, Italy; CEINGE—Biotecnologie Avanzate s.c.ar.l., Naples, Italy

Edoardo Saccenti

Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands

Emanuela Salviati

Department of Pharmacy, University of Salerno, Fisciano, Salerno, Italy

Jagdeep K. Sandhu

Human Health Therapeutics Research Centre, National Research Council of Canada, Ottawa, ON, Canada; Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada

Giovanni Scala

Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy

Eduardo Sommella

Department of Pharmacy, University of Salerno, Fisciano, Salerno, Italy

Carmen Daniela Sosa-Miranda

Human Health Therapeutics Research Centre, National Research Council of Canada, Ottawa, ON, Canada; Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada

Francesco Strati

Mucosal Immunology Lab, Department of Experimental Oncology, IEO—European Institute of Oncology IRCCS, Milan, Italy

Steven J.K. Symes

Department of Chemistry and Physics, University of Tennessee-Chattanooga, Chattanooga, TN, United States; Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States

Kacper Szczepski

Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Leonardo Tenori

Department of Chemistry and Magnetic Resonance Center (CERM), University of Florence, Florence, Italy

Giovanni Troisi

Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy

Jacopo Troisi

Department of Medicine, Surgery and Dentistry, “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy; Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy; Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy

Foreword

Before the completion of the human genome project, the paradigm of human diseases was based on the assumption that being born with genes associated to any disease was a signed destiny to develop it. With the discovery that we are made by only 23,000 genes, we now appreciate that the environment plays a major role in determining if and when we develop chronic inflammatory diseases. Indeed, over the past few decades, we have witnessed a dramatic rise of chronic inflammatory diseases affecting humankind with different type of outcomes depending on the socioeconomic environment in which children are born and raised. Children living in impoverished areas around the world often develop stunted growth from 4 to 24 months of age, a concern that is heightened by potential lasting consequences of impaired cognitive development throughout their lifespan. While in the past it was believed that malnutrition was the driving factor leading to these clinical outcomes, we now know that chronic, subclinical inflammation most likely is the key element leading to poor physical and intellectual growth. Similarly, chronic inflammatory processes starting in childhood seem to be responsible of chronic inflammatory disease “epidemics,” including allergic, autoimmune, metabolic, neurodegenerative, and tumoral diseases, detected in industrialized countries during the past few decades that can develop at any age. The combination of pre-, peri-, and postnatal factors may influence if and when the immune system unleash inflammation. Under ideal conditions, including healthy pregnancy, normal delivery, appropriate feeding (breast-feeding, natural food), limited use of antibiotics, and few infections during the first 1000 days of life, the human microbiome stays in balance and train the immune system to generate inflammation only when there are extreme conditions to be protected against “enemies,” so maintaining normal health and prevent aberrant pro-inflammatory or allergic responses. Conversely, increase in C-section practice even when not medically indicated, decrease in breast-feeding practice, and excessive use of antibiotics cause an imbalanced microbiome (dysbiosis). Therefore training the immune system to continuously unleash inflammation also when not needed leads to chronic inflammatory diseases in genetically predisposed individuals. While research around the role of the microbiome is growing exponentially, clinical applicability is lacking. Current studies evaluate the microbiota at different taxonomic levels, at different time points, from different sites, by different platforms, and with different computational strategies. This is due to rapid growth in the field, challenged by narrow focus of individual studies, small sample sizes, cross-sectional design, and lack of standardization. The focus has also mainly been on the bacterial microbiota, while viruses, parasites, and fungi are also likely to be important members of this large coevolving ecosystem that lives on and within us. Further mechanistic studies to understand these human–microbe interactions will be important. We are much more likely to discover clinically meaningful and successful interventions if they are designed based on established mechanistic

understanding. Therefore we need to transition from descriptive to mechanistic studies of the microbiome, for promising translational medicine to be possible. To implement this transition, we need to appreciate that studies on human genetics, microbiome, immune functions, and environmental factors need to be highly integrated following the overall theme that human phenotypes are mainly dictated by the activation of specific metabolic pathways that may change key functions influencing the balance between health and disease. Specifically, it is now clear that the human microbiome may epigenetically influence the expression of a variety of genes controlling key metabolic functions that control the shift from genetic predisposition to clinical outcome. Therefore this book provides the state-of-the-art of current knowledge on the human metabolome and how the use of novel technologies coupled with robust metadata and artificial intelligence analysis may lead to a radical paradigm shift of the future of human health by allowing personalized interventions (precision medicine) and, even more impactful and exciting, the possibility of disease interception and primary prevention. If properly implemented, these studies have the potential to dramatically impact our understanding of and approach to a variety of complex chronic inflammatory diseases. Only then breakthrough treatment and prevention strategies will likely emerge.

Alessio Fasano

*Mucosal Immunology and Biology Research Center, Harvard Medical School,
Massachusetts General Hospital for Children, Boston, MA, United States;
European Biomedical Research Institute of Salerno, Salerno, Italy*

Introduction

Jacopo Troisi^{1,2,3}

¹*Department of Medicine, Surgery and Dentistry “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy*

²*Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy*

³*Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy*

Over the past centuries, scientists have been animated by the belief that complex questions can be addressed by dividing them into smaller, simpler problems. This approach, known as reductionism, is mainly based on the idea that information about single components is sufficient to explain the whole. In other words, the fundamental assumption is that the sum of all the answers to the small questions into which a problem has been broken down leads to the answer of the original problem.

This, although not totally correct, has been effective in addressing almost all the scientific and technical issues that humankind has faced over time. However, several topics, such as cancer biology and behavior, psychiatric disease onset and progression, chronic disease, and neurodegeneration (just to cite a few) cannot be reduced to smaller parts because these are driven by complex biological networks that involve the cross-talk of different parts.

Currently, we are approaching a new era in which the increased availability of data, coupled with the growing computational power to manage them, are providing novel and promising tools to address even the most challenging questions.

To that end, the metabolomic perspective offers new points of view by rejecting the tendency to reduce problems into small parts, but rather addressing the complexity as a whole in order to both appreciate the stateliness of life mechanisms and build new strategies to understand and improve the quality of life. In such scenario, it is reasonable to think that the next-generation medicine will be largely based on such omics-type approaches.

Unfortunately, the study of complexity has not been the prerogative of life science scholars so far. However, the bioanalytical techniques are progressing, allowing the collection of information from thousands of molecules within a cell, tissue, or biological fluid. This amount of information presents its own challenges, requiring even more specialized knowledge for its appropriate use.

Starting from these premises, the aim of this book is to navigate readers along a path to explain all the needed steps to obtain data ([Section 1](#)) and to extract knowledge from the data ([Section 2](#)) in an *omics* perspective. This book also

contains a third section in which the principal applications of metabolomics in the life science research are explained. The ultimate goal of the book is to provide a pathway for a paradigm shift for the way complex biological problems are addressed and understood. For access to the R package and other datasets, please visit <http://www.metabolomicsperspectives.com>.

SECTION

Fundamentals 1

This page intentionally left blank

System biology

1

Elizabeth C. Plunk¹, Weston S. Chambers², and Sean M. Richards^{2,3}

¹*Department of Environmental Medicine, University of Rochester Medical School, Rochester, NY, United States*

²*Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States*

³*Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States*

Introduction

Systems biology integrates multiple levels of information to develop a comprehensive understanding of organismal biochemical and physiological functions. The “omics” technologies represent analytical approaches of a holistic nature to molecules (genes, proteins, metabolites, etc.) that make up a cell, a tissue or an organism; they are capable of performing hundreds/thousands of measurements simultaneously on a biological sample to generate a unique status of the state of an organism. Genomics, transcriptomics, proteomics, and metabolomics are today the main omics technologies. Besides these better-known applications, there are other technologies such as interactomics, energomics, and fluxomics that use specific acquisitions to arrive at an understanding of the overall biological system. The integration of these technologies in a holistic perspective is the specific aim of systems biology. The several “omics” disciplines produce different, usually complementary, information. Genomics indicates an organism is capable of doing, transcriptomics describes what it is about to do, and proteomics indicates what it is doing. Metabolomics describes the interaction between genes and environment. These disciplines can be used individually, but they can also be used in tandem (e.g., proteogenomics). In this chapter, the omics are introduced and tools that are used to collect information are described. Examples of several studies are provided that were conducted to better understand disease as well as to expand the field of interest. Over the past few decades, the science around the omics has progressed significantly, and researchers are continuing to work on how to improve these sciences for better clinical practices such as diagnoses and treatments of disease.

Genomics

Introduction

The human genome, including exons, introns, and intergenic regions, consists of 3.2 billion base pairs, or nucleotides, and 20,000 protein-coding genes (Turnbull, 2018), and for the most part, the nucleotides' location on the chromosome is known (Christensen & Murray, 2007). Most of the human genome is identical in the world's population; in fact, two randomly selected individuals have approximately 99.9% of DNA in common (Guttmacher & Collins, 2005). The remaining 0.1% is composed of DNA sequence variants, which contribute to differences in physical appearance and can be used to identify and define racial and ethnic groups (Guttmacher & Collins, 2005). Owing to geographical and historical reasons, DNA variant frequencies can differ between groups of people; however, it is uncommon for DNA sequence variants to only be found within a single group (Guttmacher & Collins, 2005).

The science of genetics and its applications are mostly restricted to singular genes, while genomics utilizes the genome in its entirety (Molster et al., 2018). An example of this is shown by the International HapMap Project, which has identified the location of single-nucleotide polymorphisms (SNPs), the most common type of genetic variation. This holistic approach allows researchers in the field of genomics to study multiple genes and their interactions with each other while also taking environmental factors into account (Molster et al., 2018). Genomics has contributed to the realization that there are very few diseases lacking genetic influence and that common complex diseases are multifaceted, being affected by genetics as well as lifestyle and environmental factors (Cleeren et al., 2011). There have been rapid developments in the field of genomics in the past two decades (Molster et al., 2018), and these developments have blurred the distinction between genetic and environmental diseases (Cleeren et al., 2011).

Genomic tools

In 2001, the first draft of the human genome was published, and it was completed in 2004, bringing an end to the Human Genome Project (HGP) (Naidoo et al., 2011). This monumental development led to studies aimed at finding the number of genes as well as gene density, nonprotein-coding RNA genes, pervasive transcription, high copy number repeat sequences, and evolutionary conservation (Naidoo et al., 2011). Before the HGP, genetics was simply used to analyze chromosomes and uncover the genes responsible for Mendelian diseases (Urban, 2015). Microarray techniques, allowing for a single test to analyze several million predetermined items of genetic information, and next-generation sequencing (NGS) technologies led to the improvement of genomics (Urban, 2015).

Microarray techniques have been an important tool in learning about the genome systematically and comprehensively; the advantages of these techniques

include simplicity, affordability, flexibility, and quickness (Brown & Botstein, 1999). Microarray-based comparative genomic hybridization is used to analyze the whole genome in a single experiment. On a genome wide scale, microarrays allow for the measurement of transcripts of every gene at once as well as the expression pattern that explains the function of that gene (Brown & Botstein, 1999). Microarrays can also easily obtain information on what the promoter of each gene is transducing along with what biochemical processes are present in the cell (Brown & Botstein, 1999).

Originally, scientists could only study one human gene fragment at a time; however, with the current development of NGS, scientists are now able to sequence far greater blocks of DNA in concert (Turnbull, 2018). NGS, also known as massively parallel sequencing, represents an effective way to capture a large amount of genomic information about cancer (Gagan & Van Allen, 2015). NGS works by binding each DNA fragment to an array. DNA polymerase goes through and incorporates labeled nucleotides in order of the DNA (Gagan & Van Allen, 2015). Then a high-resolution camera takes a snapshot of the signals emitted from each nucleotide, from this the computer can determine the location of the nucleotide and create a single DNA sequence, called a read (Gagan & Van Allen, 2015).

NGS allows clinicians to test patients with suspected genetic disorders in a more efficient method (Turnbull, 2018). NGS includes two models: short-read sequencing and long-read sequencing (Goodwin et al., 2016). Short-read sequencing is more affordable and provides higher-accuracy data and is used in population research and clinical variant discovery, while long-read sequencing is better for de novo investigations of the genome as well as full-length isoform sequencing (Goodwin et al., 2016). Recently, NGS has advanced to the point of being used for clinical diagnosis of pathogens and identifying cancers, for example.

Whole genome sequencing (WGS) only relies on a minimal amount of DNA (Roos et al., 2018) and can quickly identify the majority of genetic variants in a single genome (Sobreira et al., 2010). WGS can be used to identify genetic information in an individual patient, and this information can be used for diagnosis and treatment plans (Ashley et al., 2010). While many genetic variants had previously been identified to diagnose Mendelian diseases (diseases caused by a single mutated gene), the use of WGS has been responsible for the identification of more genetic variants causing disease (Sobreira et al., 2010). By comparing the genome of individuals with a dominant Mendelian disease to the genome of family members without the disease, scientists can determine which DNA variant is responsible for a disease (Sobreira et al., 2010).

Genome-wide association studies (GWASs) have come to the forefront for studying complex diseases and traits, which were originally studied by candidate-gene association techniques (Naidoo et al., 2011). GWASs tend to pinpoint genetic risk factors that are moderate risks (i.e., environmental risk factors) and not pinpoint genetic risk factors associated with high risk genetic factors associated with single-gene disorders (Christensen & Murray, 2007). GWASs have been useful in determining genetic variants responsible for various drug outcomes

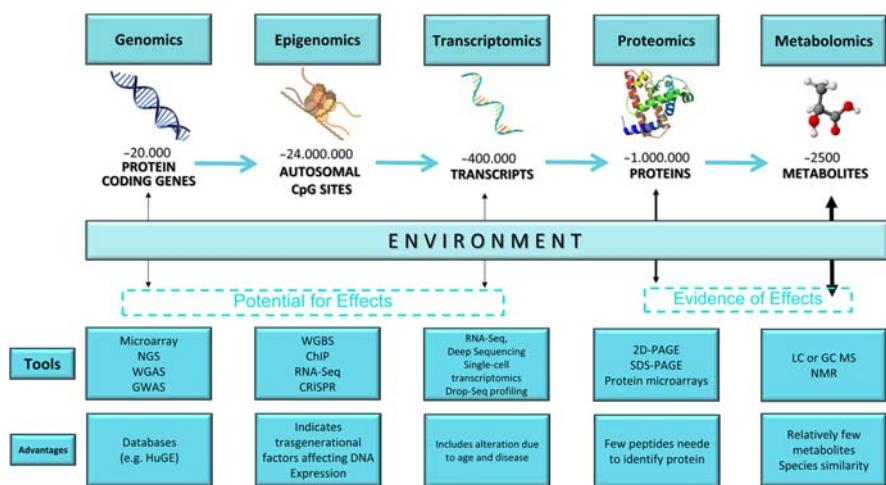
including adverse reactions. However, GWASs are unable to identify many of the presumed disease-associated variants (Naidoo et al., 2011).

Pharmacogenomics is the study of genetic factors influencing drug response (Weinshilboum & Wang, 2006), which aids in the efforts to improve drug success, safety, and dosage (Urban, 2015; Wang et al., 2011). The use of pharmacogenomics in the clinical setting, although limited, is being used in the governmental regulation of drug development (Weinshilboum & Wang, 2006), and it is internationally investigated (Turnbull, 2018). Monogenic traits involving drug metabolism, were the first to be analyzed, and currently, pharmacokinetics [factors that influence the concentration of a drug reaching its target(s)], pharmacodynamics (the drug's interaction with the receptor), and genome-wide analysis are types of pharmacogenomics being utilized (Weinshilboum & Wang, 2006).

A variety of illnesses can be identified based on genetics and biochemistry, increasing the ability to accurately diagnose and better treat patients (Urban, 2015). For example, genomics serves as a useful tool in studying cancer. Pre-NGS, only specific genomic loci known for mutating and resulting in cancer (i.e., hotspots) were examined during tumor genotyping (Gagan & Van Allen, 2015). Now, known cancer genes, whole-exome, whole-genome, and whole-transcriptome assays can be performed (Gagan & Van Allen, 2015). Precision oncology examines otherwise unexplained drug resistance, genetic links to tumorigenesis, responses and tumor recurrence, and investigates effects of drug combinations (Collins & Varmus, 2015). With current techniques, traditional pathologic diagnosis is unlikely to be replaced, however, precision cancer medicine and NGS allow for a more complete understanding of cancer etiology (Gagan & Van Allen, 2015). Tumor genomics is used to identify the tumor mutations which are reliable enough to predict drug response, tumor behavior, cancer prognosis, and surveil for early recurrence (Turnbull, 2018). Tumor genomic techniques and applications will surely improve as groups such as the International Cancer Genome Consortium, the Cancer Genome Atlas, and the 100,000 Genomes Project continue their work (Turnbull, 2018).

Genomic technologies are promising when looking at the prospects of other diseases as well. Franks & Poveda (2017) suggest that by using genomics as a tool for precision diabetes medicine, the susceptibility of an individual to developing diabetes can be predicted. Furthermore, prognostic biomarkers will be available to know better when lifestyle changes should be made, and if diabetes is developed, the proper treatment route that can be taken (Franks & Poveda, 2017).

Stewart et al. (2010) crossbred mice polygenic for type 2 diabetes (characteristics like obesity, hyperinsulinemia, impaired glucose uptake and tolerance, hyperlipidemia, and hyperglycemia) with mice lacking these characteristics. The F1 and F2 mice resulting from this crossing were phenotyped, and F2 mice were genotyped. After genome wide genotyping, quantitative trait locus (QTL) mapping found that QTLs exist on chromosomes 1,4, 8, and 11 for hypertriglyceridemia. Chromosome 1 and 3 had QTLs for hypercholesterolemia while chromosome 4 had QTLs for hyperglycemia. Body weight appeared to be linked to QTLs on chromosome 1 and

**FIGURE 1.1**

Required Implications, tools and advantages of the omics.

11. Both hypertriglyceridemia and hypercholesterolemia had links to chromosome 1 which contains *Apoa2* gene, a candidate for hyperlipidemia.

Because human genomes can have more than genetic variants (Urban, 2015), public access databases, such as the Human Genome Epidemiology (HuGE) Network (Lin et al., 2006), and computational algorithms have been pivotal in the identification of normal and pathogenic variants in genomic variation. HuGE is a database including published data on population-based epidemiological studies of human genes. Although these databases exist, clinically, WGS and other genomic techniques are expensive (Goldenberg et al., 2013). Furthermore, the genome is quite extensive in comparison to other system biologies (Dunn & Ellis, 2005). However, with tools such as the Precision Medicine Initiative, doctors and scientists will be able to understand the mechanisms of diseases, be proactive against diseases, and predict the most optimal treatment for the individual using information provided by the genome (Collins & Varmus, 2015). Indeed, precision medicine is expected to become a model for the future of medicine and medical research and is suspected to result in earlier diagnosis and even prevention of diseases (Ashley, 2016; Urban, 2015) (Fig. 1.1).

Epigenetics

Introduction

Epigenetics translates to the above genetics. This field of research includes studying modifications of gene expression through DNA methylation, posttranslational

histone modifications, and noncoding RNAs, including microRNAs (Klose & Bird, 2006; Li, 2002). DNA methylation and histone modifications are heritable (Li, 2002). These modifications control the expression of genes in multiple ways, including alteration of the organization of the structure of chromatin, either packing it tightly into heterochromatin or unpacking it into euchromatin (Bender, 2004). The tightly compact regions of heterochromatin restrict access for transcription, while the loosely-packed regions of euchromatin are most accessible for transcription. These two configurations result in downregulation and upregulation of gene expression, respectively. Inhibition of gene expression and global gene silencing are results of epigenetic modifications (Bird & Wolffe, 1999; Li, 2002). Aberrant chromatin modifications can lead to X-chromosome inactivation and genomic imprinting (Klose & Bird, 2006; Li, 2002).

DNA methylation provides a heritable mark to direct the formation of heterochromatin (Bender, 2004). It contributes to silencing the transcription of regions of DNA, and this methylation-mediated silencing has a role in the etiology of human disease and potential for the treatment of disease (Bird & Wolffe, 1999). CpG dinucleotides are cytosine bases located 5' to a guanosine base (Jones & Baylin, 2002), and they are the primary sites of methylation, resulting in the formation of methyl-CpGs (Bird & Wolffe, 1999). The presence of methylated-CpGs in a specific region of DNA can have direct and indirect effects on transcription. Regulation of DNA methylation plays a role in growth and development; an example being the genome-wide demethylation that occurs after fertilization, followed by remethylation after the implantation of the blastocyst. The highest density of nonmethylated CpGs are contained in CpG islands, which contain regulatory DNA that are required for gene transcription, such as promoter regions. Aberrant methylation of CpG sites has been found to contribute to tumorigenesis and the development of cancer in humans (Sandoval et al., 2011). The hypermethylation of CpG islands in tumor-suppressor genes result in their inactivation, and the hypomethylation of the genomes of cancer cells results in more uncontrolled gene expression.

Histone modifications are another mechanism of epigenetic regulation of gene transcription. Histones are basic proteins found in the chromatin of all eukaryotic cells that serve a role in packing the DNA, and the association of histones with DNA forms a nucleosome (Li, 2020). There are four core histones: H2A, H2B, H3, and H4. DNA base-pairs wrap around the histone octamer, which is comprised of two copies of the four core histones to create the nucleosome. Histone proteins are modified posttranslationally, and there are a large number of different posttranslational modifications (PTMs) to histones that interact with one another to regulate gene transcription (Bannister & Kouzarides, 2011). Histone proteins contain highly basic (N)-terminal tails that can protrude from their own nucleosome and interact with adjacent nucleosomes. These tails are where histone modifications occur, which affect internucleosomal interactions and result in a change in chromatin structure.

One of the many different kinds of RNA that are involved in biological processes is microRNA (miRNA), and while some types of RNA are involved in

gene expression, like mRNA and tRNA, miRNA plays a role in silencing the expression of genes (Eulalio et al., 2008). There are more than 2000 different miRNAs that have been detected in humans, and each miRNA targets between one to several dozen specific mRNAs (Duchaine & Fabian, 2019). MicroRNAs are approximately 22 nucleotides long, and they work to silence gene expression post transcriptionally by forming complexes that bind to target mRNAs at their 5' untranslated regions (Eulalio et al., 2008). This mechanism of gene silencing differs from DNA methylation and histone modifications by inhibiting the initial transcription of DNA. In order for miRNAs to perform this function, they assemble together with Argonaute proteins, which are small RNA-binding proteins (Zamore & Haley, 2005), to form miRNA-induced silencing complexes (miRISCs). The miRNAs guide the Argonaute proteins to the target mRNAs to become silenced (Eulalio et al., 2008).

Carcinogenesis is often attributed to genetic mutations; however, epigenetic mechanisms are also important in the initiation and progression of human cancer (Jones & Baylin, 2002). Epigenetic changes, especially aberrant hypermethylation of promoter regions, affect virtually every step in tumor progression (Jones & Baylin, 2002). The hypermethylation of promoter regions has been found in virtually every type of human neoplasm and has been associated with inappropriate silencing of transcription. Importantly, epigenetic changes associated with cancer, unlike genetic mutations, are potentially reversible (Bannister & Kouzarides, 2011).

Epigenetic tools

Bisulfite modification is fundamental for the majority of assays used in measuring DNA methylation and consists of converting cytosine bases to uracil. Whole genome bisulfite sequencing (WGBS) is considered the most comprehensive method for evaluating the methylation state of almost every CpG site in the genome. Legendre et al. (2015) utilized WGBS in an attempt to identify changes in DNA methylation that may be used to predict metastatic breast cancer (MBC). The researchers obtained plasma samples from patients who comprised three cohorts: patients diagnosed with MBC, disease-free survivors of primary breast cancer, and patients who were healthy (H). DNA was extracted from each sample and subjected to bisulfite conversion and later amplified using PCR. The bisulfite-modified DNA was sequenced and the researchers were able to identify 21 methylation hotspots exhibited in patients with MBC.

Chromatin immunoprecipitation (ChIP) is the most direct way to identify locations of histone modifications (Furey, 2012), and the most comprehensive technique for the evaluation of the state of global histone modification is ChIP-sequencing (ChIP-seq) (Li, 2020). ChIP involves using antibodies towards specific histone modifications, with public access databases that can be used to determine the optimal antibodies necessary to measure these histone modifications (Kagohara et al., 2018). Using the ChIP-seq method, Zhao et al. (2016) were able to identify a decrease in the levels of multiple histones involved in transcriptional repression in

breast cancer cells. The researchers were also able to identify an increase in the levels of a histone deacetylase enzyme in the cancer cells.

RNA-sequencing (RNA-seq) is a method that can be used for whole genome analysis of noncoding RNAs (ncRNAs) (Li, 2020). For RNA-seq, RNA is extracted from an isolated cell or tissue population in totality, and then a specific RNA species, such as miRNA or long noncoding RNA (lncRNA), can be isolated using different protocols (Kukurba & Montgomery, 2015). The subsets of RNA are then converted to complementary DNA (cDNA) via reverse transcription, and then the fragments of cDNA are ligated by sequencing adapters. Finally, PCR amplification and sequencing can be performed. Fan and Liu (2016) conducted a study using RNA-seq data on lncRNA expression in esophageal cancer tissue. They identified 265 differentially expressed lncRNAs between cancer tissue and normal esophageal tissue, and they identified eight specific lncRNAs that were used to establish a predictive model that can be used to classify patients into high-risk and low-risk categories with significantly different survival rates.

There have also been developments in epigenome editing in recent years, which may be useful to address the relationship between epigenetic modulation of specific regions of DNA and subsequent changes in gene expression (Huang et al., 2017). Utilizing clustered, regularly interspaced, short palindromic repeats (CRISPR), researchers have designed a CRISPR-Cas9-acetyltransferase fusion protein to target genomic sites and achieve histone acetylation to activate nearby gene expression (Hilton et al., 2015). CRISPR has also been utilized for DNA methylation via the conjugation of the Cas9 protein with repetitive peptide epitopes, which recruit multiple copies of DNA methyltransferase 3A to amplify methylation (Huang et al., 2017). Thus CRISPR-Cas9 technology is helping to determine the effects of epigenetic modulation on specific areas of the human genome (Hilton et al., 2015).

Using CRISPR-Cas9-based acetyltransferase, Hilton et al. (2015) were able to activate transcription in target genes. Transcriptional activation was achieved by fusing nuclease-null dCas9 protein to the catalytic core of the human acetyltransferase p300. This fusion event caused the acetylation of histone H3 lysine 27 at the target site, and this in turn led to the transcriptional activation of the target genes. This is an example of how targeted acetylation can cause transcriptional activation therefore affecting gene regulation via epigenetic processes.

Melamed et al. (2015) took cord blood samples from 10 pregnancies from assisted reproductive technologies (ART) and eight from controls. Using Illumina Infinium Human Methylation27 array, Melamed et al. (2015) found that 733 of the 27,578 CpG sites analyzed were significantly differentially methylated between the ART group and control. Within the ART group, a higher variation of DNA methylation was observed suggesting random genome-wide changes in DNA methylation, or epigenetic instability associated with ART pregnancies or the subfertility resulting in ART pregnancies.

This section is not an exhaustive listing of epigenetic technologies, and a more extensive list can be found (Li, 2020).

Transcriptomics

Introduction

Transcriptomics is used to identify the transcriptional activity that is present in cells and tissues (Hegde et al., 2003). The transcriptome consists of the coding and ncRNAs that are transcribed in cells, tissues, or organs during normal physiological or pathological conditions (Assis et al., 2014) and can be defined by mRNA expression. It displays the genes that are being actively expressed at a particular moment (Horgan & Kenny, 2011).

The transcriptome, unlike the genome, can vary depending on the phase of the cell cycle as well as the cell type, the organ analyzed, exposure to drugs, aging, and the presence of diseases (Assis et al., 2014). Key goals of transcriptomics include: cataloging all components of the transcriptome, which include mRNAs, ncRNA's, and small RNAs, but exclude rRNAs (Urban, 2015; Wang et al., 2011); establishing the transcriptional structure of genes in terms of start sites, 5' and 3' ends, splicing patterns, and other posttranscriptional modifications; and quantifying the changes in expression of each transcript during stages of development and under different conditions.

The understanding of life and disease can be increased by studying the transcriptome of developing adults and pathological tissues with spatial single-cell transcriptomics (Sandberg, 2014). By doing so, discoveries could be made regarding the role of the transcriptome in intercellular communication, polarity formation, and gradients. Another benefit of single-cell transcriptomics is the ability to discover high-resolution transcriptional maps which will allow the vision of both stable and transient cellular states during differentiation or reprogramming (Sandberg, 2014).

Even though only approximately 3% of the human genome encodes proteins, 80% of the human genome is transcribed into RNA (Hasin et al., 2017). ncRNAs are RNA molecules that do not encode for a protein (Mattick & Makunin, 2006), and they play essential roles in many physiological processes, such as endocrine regulation and neuron development (Hasin et al., 2017). A subset of ncRNAs called long noncoding RNAs (lncRNAs) are involved in many cellular processes, including cell differentiation, organogenesis, and tissue homeostasis (Schmitz et al., 2016).

Defects in ncRNAs have been implicated in the etiology of many different pathological conditions (Wang et al., 2013), and the genetic cause for some diseases may be caused by mutations within ncRNAs (Mattick & Makunin, 2006). lncRNAs have been shown to be involved in cancer and cardiovascular disease (Schmitz et al., 2016). Transcriptomics may prove to be useful for discovering ncRNAs that may be used as novel therapeutic targets or disease biomarkers.

Transcriptomics has promise in the field of oncology as it has been used to determine drug response in patients with breast carcinomas (Chang et al., 2003). Alterations in expression, sequence, or target sites of miRNAs may be a

significant source of carcinogenesis, and miRNA profiling may be used as an accurate diagnostic tool for the classification of different cancers ([Mattick & Makunin, 2006](#)). In the future, therapeutic approaches and biomarkers of diseases may be studied by analyzing the total set of miRNAs, called the miRNome; miRNAs are approximately 22 nucleotides long, making them the shortest known functional eukaryotic RNAs ([Assis et al., 2014](#)).

Transcriptomic tools

Original transcriptome analyses used large nylon arrays and high-density filters that contained colony cDNA (or PCR products) ([Assis et al., 2014](#)). This technique is still useful while testing sets of up to a few thousand genes. For researchers who are interested in studying organisms that lack reference genomes, the development of de novo methods for transcriptome characterization is of particular interest ([McGettigan, 2013](#)).

Two types of transcriptomic approaches are hybridization-based and sequence-based ([Wang et al., 2009](#)). RNA sequencing (RNA Seq) has high resolution and sensitivity maps and quantifies transcriptomes of known and unknown genomic sequences, while hybridization-based approaches are limited to just identifying known genomic sequences ([Wang et al., 2009; Wang et al., 2013](#)); however, hybridization-based approaches have been successful in determining global gene expression ([Raiol et al., 2014](#)). Due to this advantage of RNA Seq, unknown transcribed regions and splicing isoforms, even in low abundance ([Raiol et al., 2014; Roos et al., 2018](#)) of known genes have been discovered, and RNA Seq has been used to map 5' and 3' boundaries for numerous genes ([Wang et al., 2009](#)).

Deep sequencing, a NGS approach, is used to sequence a DNA fragment many times while delivering great accuracy ([Malone & Oliver, 2011](#)). However, deep sequencing is prone to losses and/or biases in data, so single-cell transcriptomics has been used in order to avoid this ([Sandberg, 2014](#)). Single-cell transcriptomics measures the reverse transcription of RNA to complementary DNA (cDNA). Then PCR is performed to amplify the DNA or in vitro transcription before deep sequencing occurs.

Single-cell transcriptomics can be used in the future to develop a comprehensive and qualitative reference “atlas” of every human cell type in both adult and fetal tissues ([Camp & Treutlein, 2017](#)). This information can be used to identify transcription factors specific to certain cell types and cell-to-cell communication through assumed receptor-ligand pairs, and ultimately human tissues can be reverse engineered.

Of recent, droplet-based microfluidics (Drop-seq profiling) has been used to profile the transcriptome of individual cells. [Alles et al. \(2017\)](#) used the methanol-based fixation protocol ([Stoeckius et al., 2009](#)) and analyzed both live and fixed mixtures of cultured human (HEK) and mouse (3T3) cells and found that methanol fixation does not alter the number of genes and transcripts in a cell. Once this was established, Alles et al. also used Drop-seq profiling on

methanol-fixed cells from dissociated *Drosophila* embryos as well as sorted mouse hindbrain cells. This experiment proved that Drop-seq profiling of cultured and primary cells works well with methanol-fixed cells.

[Patino et al. \(2005\)](#) used a serial analysis of gene expression technique to examine the transcriptomes of monocytes from patients undergoing carotid endarterectomy. Finkel-Biskis-Jinkins osteosarcoma (FOS) was increased in subjects from the carotid endarterectomy group more than in the control group, and subjects with the highest levels of FOS were patients who had also undergone coronary revascularization. FOS is responsible for inflammation and calcification, and further studies showed that in vitro inhibition of FOS decreased monocyte activation which gives insight into its pathogenesis. The findings of this study illustrate that transcriptomics can be used to better understand atherosclerosis (and other diseases) biologically and clinically.

[Kim et al. \(2014\)](#) studied the effects of statins, which are prescribed to patients to reduce low density lipoprotein cholesterol (LDLC) levels. Statins act by inhibiting 3-hydroxy-3-methylglutaryl coenzyme A reductase (*HMGCR*) which is the enzyme responsible for catalyzing the rate-limiting step of cholesterol biosynthesis. However, the result of statin treatment varies on an individual basis based on phenotypic and genetic factors. [Kim et al. \(2014\)](#) identified 100 signature genes between high- and low-responders to statin treatment and developed a model to describe the role of the signature genes. Results indicated that 12.3% of the variance in statin-mediated LDLC changes was attributable to the signature genes. When SNPs either associated with expression levels of the signature genes (eQTLs) or previously reported to be associated with statin response were included, the model predicted 15.0% of the variance. Results such as these illustrate how the use of transcriptomics can predict the effectiveness of pharmaceuticals and indicate pathways that affect drug efficacy.

Transcriptomics can be used to better understand disease as well as aid clinically in treatment plans. Limitations of identifying the transcriptome without identifying the proteome include: the transcriptome does not account for RNA splicing, differential RNA and protein turnover, PTMs, allosteric protein interactions, and proteolytic processing events that result in an alteration in protein synthesis that would not be predicted in the mRNA ([Hegde et al., 2003](#)).

Proteomics

Introduction

Proteins are manufactured to carry out important cellular functions leading to the growth, differentiation, proliferation, and death of cells, and aberration of structure and expression of proteins may be an indicator for the presence of disease ([Vlahou & Fountoulakis, 2005](#)). Proteomics is the large-scale study of the proteins and their structure and function within a cell, organ system, or organism ([Horgan](#)

& Kenny, 2011). It can be utilized to investigate proteins and determine their relative abundance, variability due to PTMs, and protein-protein interactions, which can be used to recognize and distinguish the flow of information through protein networks (Dupree et al., 2020; Petricoin et al., 2002).

During normal physiological and pathological conditions, the proteome defines what proteins are present in a cell type, tissue or organ (Assis et al., 2014). Proteomics is useful in the study of health and disease as well as drug discovery because it allows in a single experiment a snap-shot of thousands of proteins (Frantzi et al., 2019; Roos et al., 2018). During the evaluation of each protein, scientists can determine the protein signature of cells or tissues (Roos et al., 2018). The protein signature can then be used to best determine a treatment for a disease. Clinical proteomics is a subdiscipline of proteomics that involves the utilization of proteomic technologies at the bedside of patients in order to diagnose and treat disease (Petricoin et al., 2002).

Proteomics research began in the mid-1990s as a result of the need for disease research where other approaches were limited (Dupree et al., 2020; Hanash, 2003). The emergence of proteomics began with goals for understanding delineation of altered protein expression, discovering biomarkers for earlier diagnosis of a disease, and aiding in drug development by determining novel targets for therapeutics. The field uses a wide range of methodology today due to modern developments in research technology.

Proteomic tools

One of the first techniques used to profile protein expression in disease used two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) to separate proteins from the nonprotein components of cells and then identify the proteins using mass spectrometry (MS) (Hanash, 2003). 2D-PAGE, as well as other gel-based techniques like sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), is still used to this day to separate complex protein samples (Aslam et al., 2017). Along with gel-based techniques, liquid based techniques, which also separate the proteins and use MS, were created because the gel-technique needed greater sample quantity, which is a limitation in clinical work.

A conventional method for protein purification is chromatography, of which there are different types that can be used to sort proteins based on their physical and chemical properties (Aslam et al., 2017). Size exclusion chromatography is used to separate molecules based on size in solution, and it utilizes a porous gel mixture with distinct pore sizes. Ion-exchange chromatography separates proteins based on their surface charges by using specific ion-exchange adsorbents (Fekete et al., 2015). Affinity chromatography utilizes a binding ligand in order to retain desired proteins for analysis or observe biological interactions (Hage et al., 2012).

Enzyme-linked immunosorbent assays (ELISAs) and western blotting are technologies that may be used for analysis of selective proteins (Aslam et al., 2017). The ELISA method is used to quantitatively analyze molecules via antigen-

antibody reactions that produce a change in color when exposed to an enzyme-linked conjugate and an enzyme substrate (Aydin, 2015). The change in color is proportional to the concentration of the antigen, for which the antibodies used are specific for. Western blotting involves the transfer of proteins that have been separated by SDS-PAGE or other gel techniques and is particularly useful for detection of proteins in low abundance (Kurien & Scofield, 2006). Both of these techniques are useful for analysis of individual proteins but are unable to determine expression levels of proteins.

Grishina et al. (2017) utilized SDS-PAGE and western blotting to identify novel allergenic proteins in shrimp, sesame, hazelnuts, and pistachios (Grishina et al., 2017). The researchers first obtained protein extracts from the foods and separated them on immobilized pH-gradient (IPG) strips and then proceeded to use SDS-PAGE to resolve the proteins based on their molecular weights. The proteins were then transferred to membranes for blotting, and the allergens were identified using blood serum of allergic subjects. The immunolabeled proteins were identified using Edman sequencing, a method consisting of multiple chemical reactions which cleave N-terminal amino acids from proteins and identify them by using MS (Aslam et al., 2017).

MS and protein microarrays are techniques used on tissue without denaturing the proteins (Hanash, 2003). MS requires the use of a matrix-assisted laser desorption/ionization (MALDI) plate. With this method, protein expression can be compared between healthy and diseased tissue by identifying peptide sequences, protein abundance, PTMs, and protein interactions (Frantzi et al., 2019). A limitation, however, is that proteins in low concentrations may not be detected during the MS analysis (Roos et al., 2018). Protein microarrays are important in determining how proteins differ between healthy and diseased tissue. The term “microarray” refers to thousands of miniature assays that occur on one plate, which is accomplished by immobilizing proteins, cell lysates, or antibodies, similar to ELISA assays, on glass slides (Sutandy et al., 2013). The concentration of proteins and their interactions between other biomolecules can be analyzed with protein microarrays.

The underlying cause of disease must be understood in order to treat symptoms and cure (Frantzi et al., 2019). As mentioned above, the underlying cause of the disease can be elucidated by measuring protein expression in relevant cells. To further emphasize the role of proteomics in disease, many pharmaceuticals target proteins in a cell; therefore proteomics can be used to study drug targets, protein interactions, and protein-based biomarkers, all of which aid in drug development.

Proteomics and genomics can be used in concert, called proteogenomics, for increasing information for individual disease treatment (Ang et al., 2019). Genetic components such as copy-number variations, SNPs, missense variants (nonsynonymous SNPs), nonsense variants, 5' untranslated region, 3' untranslated region variants, splice-site mutations, and programmed frameshifts on protein abundances can be better understood with the use of proteogenomics (Roos et al.,

2018). By studying protein to gene, proteomics can be used to discover protein effects on a gene. Then genomic techniques can be used to study the candidate gene.

Irregularities in the human proteome can result in pathogenesis, and proteomics research gives insight into the cause-and-effect relationship between protein aberrations and disease. The knowledge that has been gathered by proteomics research greatly compliments the data that has been gained from genomics research, and the combination of these two omics has the potential to largely impact the development of future diagnostic and therapeutic approaches (Dupree et al., 2020).

Metabolomics

Introduction

The field of metabolomics studies the array of metabolites (the metabolome) of biochemical processes within cells, tissues, and organs. The metabolome reveals the phenotype of a cell, tissue, or organism (Fanos et al., 2012) and consists of all associated low molecular weight compounds (<1500 Da) including organic, inorganic, and elemental metabolites (Dunn & Ellis, 2005). These metabolites are products of biochemical processes; identifying these metabolites can provide information about how the system responds to genetic and environmental changes (Fiehn, 2002). Metabolic networks can be studied by identifying and quantifying the metabolites present in a multitude of samples (Weckwerth, 2003). There are many applications of metabolomics including the discovery of biomarkers in healthy and diseased tissues, assessing efficacy of pharmaceuticals, and determining biochemical pathways associated with diseases, among others (Le Gall et al., 2003).

Metabolomic tools

The general process of determining the metabolome is as follows: sample preparation, data collection, data processing, and interpretation (Yan & Xu, 2018). Because of the diverse size, hydrophobicity, and volatility of metabolites, techniques used to identify and quantify the metabolome must be very sensitive to ensure that metabolites are not excluded during the preparation of the sample (Dunn & Ellis, 2005). The most commonly used analytical methods in metabolomics are NMR spectroscopy and gas or liquid chromatography (GC or LC) paired with MS (Emwas et al., 2019), which allow for the rapid identification and quantitation of metabolites in multiple samples (Dunn & Ellis, 2005; Weckwerth, 2003). However, to successfully employ GC-MS, each metabolite's volatility must be enhanced through chemical derivatization first (Fiehn, 2002) and the metabolites must be thermally stable (Dunn & Ellis, 2005). The GC functions to

separate the now volatile and thermally stable compounds from the rest of the sample; MS works after GC to define the separated compounds. While GC-MS is effective for analysis of volatile molecules, LC-MS can be used to analyze polar, involatile molecules without the need for derivatization (Gika et al., 2014). Liquid-based separations provide the most versatile tools for analysis of multiple molecules belonging to different groups and having different chemical properties in the same sample. NMR spectroscopy has multiple advantages over GC-MS and LC-MS, including its ease of sample preparation, and its nondestructive quality. However, these advantages are often outweighed by the fact that GC-MS and LC-MS are more sensitive analytical techniques than NMR.

Metabolomes can be determined from all ages of organisms at all taxonomic levels. Indeed, metabolomes ranging from bacteria to plants to humans have been characterized (Sumner et al., 2003). The implications of metabolomics in newborns is promising as the metabolomic analysis can be used to predict diseases that could occur in adulthood, allowing for earlier interventions in those diseases (Fanos et al., 2012, 2013) (see also Chapter 18: Metabolomics in Public Health). A broad range of biofluids and tissues can be utilized in metabolomic studies. These include blood, urine, and saliva. On a metabolic level, serum is useful in determining the physiological and pathological status of a disease, diagnosis of disease, and allows for the identification of early metabolic markers of disease (Zhang et al., 2012). Urine samples provide information of metabolic activity over a period of time, and in neonatology studies, urine is the easiest biofluid to obtain and utilize (Atzori et al., 2009).

Metabolomics has been successfully used in characterizing a wide range of metabolic signatures including but not limited to gut biomes (Le Gall et al., 2011; Sitkin & Pokrotnieks, 2018), fetal anomalies (Troisi et al., 2017, 2021; Troisi, Sarno, et al., 2018), cancer biology (Troisi et al., 2017, 2021; Troisi, Landolfi, et al., 2018), Parkinson's Disease (PD) (Han et al., 2017; Troisi et al., 2019), Dementia (Santos et al., 2020), and Alzheimer's Disease (Fiandaca et al., 2015; Mapstone et al., 2014). By characterizing and comparing the metabolome between groups, biochemical pathways associated with unique metabolites may be targeted to determine disease etiology or pharmaceutical efficacy, for example. Indeed, metabolomes are often analyzed to differentiate between healthy and diseased cells, tissues, organs, or patients and ultimately determine which metabolites are most associated with a disease.

Even though metabolomes may consist of hundreds of metabolites, oftentimes, just a few metabolites are responsible for the differences in the metabolomes. For example, Han et al. (2017), using LC and MS found that PD metabolite profiles differed significantly from healthy controls while using a 5-metabolite panel. Similarly, they employed an 8-metabolite panel to distinguish between PD patients without dementia and PD patients with early signs of dementia.

Ulcerative colitis (UC) and Crohn's disease (CD) are chronic inflammatory bowel diseases (IBDs) that usually begin in young adulthood and continue throughout life. It is understood that genetics, environment, and intestinal

microbial factors are responsible for the disease (Poggioli & Renzi, 2019). It is known that changes in gut microbiota are associated with changes in serum microbial metabolites levels. Sitkin and Pokrotnieks (2018) compared metabolomes of healthy volunteers to patients with UC and found reduced levels of butyrate and deduced that a specific bacterium ratio was responsible (*Bacteroides fragilis/Faecalibacterium prausnitzii*). Dawiskiba et al. (2014) compared metabolites found in serum and urine samples of patients with UC, CD, and healthy controls. They found in serum samples of patients with IBD that the metabolites N-acetylated compounds and phenylalanine were increased, low-density and very low-density lipoproteins were decreased compared to healthy controls. In the patients with IBD, their urine had increased glycine and decreased acetoacetate.

Intrauterine growth restriction (IUGR) is the reduced growth of a fetus during the fetal development period; it is a condition that can lead to health problems later in life (Forsdahl, 1978). Nissen et al. (2011) studied IUGR in piglets in order to better understand how fetal metabolic programming correlated with fetal metabolism and postnatal development of the metabolic disorders. Using plasma samples from both low birth weight and high birth weight piglets to compare the metabolomes, glucose concentrations were found to positively correlate with birth weight (Nissen et al., 2011). In contrast, myo-inositol and D-chiro-inositol showed a negative correlation with birth weight. Bhandari et al. (2008) found that D-chiro-inositol inhibits glucose-stimulated insulin release which can leave the insulin-responsive tissues sensitive to insulin which could explain the birth weight. This example illustrates the utility of metabolomics: case metabolomes are compared to control metabolomes, differences in individual metabolites are discovered and subsequently linked to specific metabolic pathways, effectively linking the disease to metabolic pathways enhanced or inhibited in correlation with the disease.

Metabolomics has great utility for determining disease etiology which may lead to a cure. Prostate cancer (PCa) is a leading type of cancer in men worldwide. Current clinical biomarkers are not successful in early diagnosis or patient prognosis (Gómez-Cebrián et al., 2019). Metabolite biomarkers could provide early diagnosis, patient prognosis, and monitoring of the disease. In PCa metabolomic studies using both NMR and MS techniques, alterations in biochemical cycles involved in formation of polyamines, tricarboxylic acid cycle, glycolysis, one carbon metabolism, nucleotide synthesis, amino acid, fatty acid, and lipid metabolism influenced the development and progression of PCa. By targeting these biochemical pathways for future research, the cause and a potential cure for PCa may be determined.

According to the estimates of the World Alzheimer Report 2015, there are 46.8 million people living with dementia worldwide, and this number is expected to almost double every 20 years (Ali et al., 2015). Currently, no cures exists, and testing an individual for the current known biomarkers of the early disease is invasive, time-consuming, and expensive (Mapstone et al., 2014). Blood-based biomarkers would be more useful and could provide a better outcome for prevention. Mapstone et al. (2014) found 10 phosphatidylcholines that predicted

phenoconversion in AD within 2–3 years. This finding can be used for the screening of individuals at risk for AD. Likewise, for late-onset Alzheimer's disease (LOAD), researchers created a 24 plasma based metabolite panel after a 5 year observational study to be used to predict the likelihood of phenoconversion to the clinical stages of LOAD (Fiandaca et al., 2015). A panel like this could eventually be used in the clinical setting to identify patients who are at higher risk of phenoconversion to the clinical stages which could result in the patient receiving treatment before the phenoconversion.

There is great promise in the use of metabolomics to diagnose diseases as well as to better understand the cause of the disease. An advantage of metabolomics compared to genomics and proteomics is that there are fewer metabolites than genes or proteins in a cell. Furthermore, due to metabolite commonality between species, identifications of metabolites can be shared between fields (Schrimpe-Rutledge et al., 2016). Metabolomics can be used to evaluate the interactions between genes and the environment by identifying epigenetic differences, but because metabolites are fragile, sample preparation can result in the loss of molecular structures (Fanos et al., 2013).

References

- Ali, G., Guerchet, M., Wu, Y., Prince, M., & Prina, M. (2015). The global prevalence of dementia. In *World Alzheimer report 2015. The global impact of dementia an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International.
- Alles, J., Karaiskos, N., Praktiknjo, S. D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S., Schreyer, L., Boltengagen, A., & Birchmeier, C. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biology*, 15(1), 1–14.
- Ang, M. Y., Low, T. Y., Lee, P. Y., Nazarie, W. F. W. M., Guryev, V., & Jamal, R. (2019). Proteogenomics: From next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. *Clinica Chimica Acta*, 498, 38–46.
- Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, 17(9), 507.
- Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A., Pushkarev, D., Neff, N. F., Hudgins, L., Gong, L., Hodges, L. M., Berlin, D. S., Thorn, C. F., Sangkuhl, K., Hebert, J. M., ... Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *Lancet (London, England)*, 375(9725), 1525–1535. Available from [https://doi.org/10.1016/S0140-6736\(10\)60452-7](https://doi.org/10.1016/S0140-6736(10)60452-7).
- Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and their applications. *Journal of Chromatographic Science*, 55(2), 182–196.
- Assis, A. F., Oliveira, E. H., Donate, P. B., Giuliatti, S., Nguyen, C., & Passos, G. A. (2014). *What is the transcriptome and how it is evaluated? Transcriptomics in health and disease* (pp. 3–48). Springer.

- Atzori, L., Antonucci, R., Barberini, L., Griffin, J. L., & Fanos, V. (2009). Metabolomics: A new tool for the neonatologist. *The Journal of Maternal-Fetal & Neonatal Medicine*, 22(Suppl. 3), 50–53.
- Aydin, S. (2015). A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA. *Peptides*, 72, 4–15.
- Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3), 381–395.
- Bender, J. (2004). DNA methylation and epigenetics. *Annual Review of Plant Biology*, 55, 41–68.
- Bhandari, R., Juluri, K. R., Resnick, A. C., & Snyder, S. H. (2008). Gene deletion of inositol hexakisphosphate kinase 1 reveals inositol pyrophosphate regulation of insulin secretion, growth, and spermiogenesis. *Proceedings of the National Academy of Sciences*, 105(7), 2349–2353.
- Bird, A. P., & Wolffe, A. P. (1999). Methylation-induced repression—Belts, braces, and chromatin. *Cell*, 99(5), 451–454.
- Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(1), 33–37.
- Camp, J. G., & Treutlein, B. (2017). Human organomics: A fresh approach to understanding human development using single-cell transcriptomics. *Development (Cambridge, England)*, 144(9), 1584–1587.
- Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., & Allred, D. C. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, 362(9381), 362–369.
- Christensen, K., & Murray, J. C. (2007). What genome-wide association studies can do for medicine. *The New England Journal of Medicine*, 356(11), 1094–1097.
- Cleeren, E., Van der Heyden, J., Brand, A., & Van Oyen, H. (2011). Public health in the genomic era: Will Public Health Genomics contribute to major changes in the prevention of common diseases? *Archives of Public Health*, 69(1), 1–12.
- Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793–795.
- Dawiskiba, T., Deja, S., Mulak, A., Ząbek, A., Jawień, E., Pawełka, D., Banasik, M., Mastalerz-Migas, A., Balcerzak, W., & Kaliszewski, K. (2014). Serum and urine metabolomic fingerprinting in diagnostics of inflammatory bowel diseases. *World Journal of Gastroenterology: WJG*, 20(1), 163.
- Duchaine, T. F., & Fabian, M. R. (2019). Mechanistic insights into microRNA-mediated gene silencing. *Cold Spring Harbor Perspectives in Biology*, 11(3), a032771.
- Dunn, W. B., & Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. *Trends in Analytical Chemistry*, 24(4), 285–294.
- Dupree, E. J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B. A., & Darie, C. C. (2020). A critical review of bottom-up proteomics: The good, the bad, and the future of this field. *Proteomes*, 8(3), 14.
- Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G., Raftery, D., Alahmari, F., Jaremko, L., & Jaremko, M. (2019). NMR spectroscopy for metabolomics research. *Metabolites*, 9(7), 123.
- Eulalio, A., Huntzinger, E., & Izaurralde, E. (2008). Getting to the root of miRNA-mediated gene silencing. *Cell*, 132(1), 9–14.

- Fan, Q., & Liu, B. (2016). Identification of a RNA-Seq based 8-long non-coding RNA signature predicting survival in esophageal cancer. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 22, 5163.
- Fanos, V., Antonucci, R., Barberini, L., Noto, A., & Atzori, L. (2012). Clinical application of metabolomics in neonatology. *The Journal of Maternal-Fetal & Neonatal Medicine*, 25(Suppl. 1), 104–109.
- Fanos, V., Van den Anker, J., Noto, A., Mussap, M., & Atzori, L. (2013). *Metabolomics in neonatology: Fact or fiction? Seminars in fetal and neonatal medicine* (Vol. 18, pp. 3–12). Elsevier, Issue 1.
- Fekete, S., Beck, A., Veuthey, J.-L., & Guillarme, D. (2015). Ion-exchange chromatography for the characterization of biopharmaceuticals. *Journal of Pharmaceutical and Biomedical Analysis*, 113, 43–55.
- Fiandaca, M. S., Zhong, X., Cheema, A. K., Orquiza, M. H., Chidambaram, S., Tan, M. T., Gresenz, C. R., FitzGerald, K. T., Nalls, M. A., & Singleton, A. B. (2015). Plasma 24-metabolite panel predicts preclinical transition to clinical stages of Alzheimer's disease. *Frontiers in Neurology*, 6, 237.
- Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Functional Genomics*, 48, 155–171.
- Forsdahl, A. (1978). Living conditions in childhood and subsequent development of risk factors for arteriosclerotic heart disease. The cardiovascular survey in Finnmark 1974–75. *Journal of Epidemiology & Community Health*, 32(1), 34–37.
- Franks, P. W., & Poveda, A. (2017). Lifestyle and precision diabetes medicine: Will genomics help optimise the prediction, prevention and treatment of type 2 diabetes through lifestyle therapy? *Diabetologia*, 60(5), 784–792.
- Frantzi, M., Latosinska, A., & Mischak, H. (2019). Proteomics in drug development: The dawn of a new era? *PROTEOMICS—Clinical Applications*, 13(2), 1800087.
- Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12), 840–852.
- Gagan, J., & Van Allen, E. M. (2015). Next-generation sequencing to guide cancer therapy. *Genome Medicine*, 7(1), 1–10.
- Gika, H. G., Theodoridis, G. A., Plumb, R. S., & Wilson, I. D. (2014). Current practice of liquid chromatography–Mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis*, 87, 12–25.
- Goldenberg, A. J., Marshall, P. A., & Sharp, R. R. (2013). Next-generation disadvantages: Identifying potential barriers to integrating genomics into underserved medical settings. *Personalized Medicine*, 10(7), 623–625.
- Gómez-Cebrián, N., Rojas-Benedicto, A., Albors-Vaquer, A., López-Guerrero, J. A., Pineda-Lucena, A., & Puchades-Carrasco, L. (2019). Metabolomics contributions to the discovery of prostate cancer biomarkers. *Metabolites*, 9(3), 48.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333.
- Grishina, G., Bardina, L., & Grishin, A. (2017). *2D-electrophoresis and immunoblotting in food allergy. Food allergens* (pp. 59–69). Springer.
- Guttmacher, A. E., & Collins, F. S. (2005). Realizing the promise of genomics in biomedical research. *JAMA: The Journal of the American Medical Association*, 294(11), 1399–1402.

- Hage, D. S., Anguizola, J. A., Bi, C., Li, R., Matsuda, R., Papastavros, E., Pfaunmiller, E., Vargas, J., & Zheng, X. (2012). Pharmaceutical and biomedical applications of affinity chromatography: Recent trends and developments. *Journal of Pharmaceutical and Biomedical Analysis*, 69, 93–105.
- Han, W., Sapkota, S., Camicioli, R., Dixon, R. A., & Li, L. (2017). Profiling novel metabolic biomarkers for Parkinson's disease using in-depth metabolomic analysis. *Movement Disorders*, 32(12), 1720–1728.
- Hanash, S. (2003). Disease proteomics. *Nature*, 422(6928), 226–232.
- Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 1–15.
- Hegde, P. S., White, I. R., & Debouck, C. (2003). Interplay of transcriptomics and proteomics. *Current Opinion in Biotechnology*, 14(6), 647–651.
- Hilton, I. B., D'ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33 (5), 510–517.
- Horgan, R. P., & Kenny, L. C. (2011). 'Omic' technologies: Genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3), 189–195.
- Huang, Y.-H., Su, J., Lei, Y., Brunetti, L., Gundry, M. C., Zhang, X., Jeong, M., Li, W., & Goodell, M. A. (2017). DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biology*, 18(1), 1–11.
- Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6), 415–428.
- Kagohara, L. T., Stein-O'Brien, G. L., Kelley, D., Flam, E., Wick, H. C., Danilova, L. V., Easwaran, H., Favorov, A. V., Qian, J., & Gaykalova, D. A. (2018). Epigenetic regulation of gene expression in cancer: Techniques, resources and analysis. *Briefings in Functional Genomics*, 17(1), 49–63.
- Kim, K., Bolotin, E., Theusch, E., Huang, H., Medina, M. W., & Krauss, R. M. (2014). Prediction of LDL cholesterol response to statin using transcriptomic and genetic variation. *Genome Biology*, 15(9), 1–12.
- Klose, R. J., & Bird, A. P. (2006). Genomic DNA methylation: The mark and its mediators. *Trends in Biochemical Sciences*, 31(2), 89–97.
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11), pdb-top084970.
- Kurien, B. T., & Scofield, R. H. (2006). Western blotting. *Methods (San Diego, Calif.)*, 38 (4), 283–293.
- Le Gall, G., Colquhoun, I. J., Davis, A. L., Collins, G. J., & Verhoeven, M. E. (2003). Metabolite profiling of tomato (*Lycopersicon esculentum*) using ¹H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *Journal of Agricultural and Food Chemistry*, 51(9), 2447–2456.
- Le Gall, G., Noor, S. O., Ridgway, K., Scovell, L., Jamieson, C., Johnson, I. T., Colquhoun, I. J., Kemsley, E. K., & Narbad, A. (2011). Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *Journal of Proteome Research*, 10(9), 4208–4218.
- Legendre, C., Gooden, G. C., Johnson, K., Martinez, R. A., Liang, W. S., & Salgia, B. (2015). Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clinical Epigenetics*, 7(1), 1–10.

- Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews. Genetics*, 3(9), 662–673.
- Li, Y. (2020). Modern epigenetics methods in biological research. *Methods*.
- Lin, B. K., Clyne, M., Walsh, M., Gomez, O., Yu, W., Gwinn, M., & Khoury, M. J. (2006). Tracking the epidemiology of human genes in the literature: The HuGE Published Literature database. *American Journal of Epidemiology*, 164(1), 1–4.
- Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1), 1–9.
- Mapstone, M., Cheema, A. K., Fiandaca, M. S., Zhong, X., Mhyre, T. R., MacArthur, L. H., Hall, W. J., Fisher, S. G., Peterson, D. R., & Haley, J. M. (2014). Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine*, 20(4), 415–418.
- Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, 15(Suppl. 1), R17–R29.
- McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, 17(1), 4–11.
- Melamed, N., Choufani, S., Wilkins-Haug, L. E., Koren, G., & Weksberg, R. (2015). Comparison of genome-wide and gene-specific DNA methylation between ART and naturally conceived pregnancies. *Epigenetics: Official Journal of the DNA Methylation Society*, 10(6), 474–483.
- Molster, C. M., Bowman, F. L., Bilkey, G. A., Cho, A. S., Burns, B. L., Nowak, K. J., & Dawkins, H. J. (2018). The evolution of public health genomics: Exploring its past, present, and future. *Frontiers in Public Health*, 6, 247.
- Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N., & Ku, C.-S. (2011). Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics*, 5(6), 1–46.
- Nissen, P. M., Nebel, C., Oksbjerg, N., & Bertram, H. C. (2011). Metabolomics reveals relationship between plasma inositol and birth weight: Possible markers for fetal programming of type 2 diabetes. *Journal of Biomedicine and Biotechnology*, 2011.
- Patino, W. D., Mian, O. Y., Kang, J.-G., Matoba, S., Bartlett, L. D., Holbrook, B., Trout, H. H., Kozloff, L., & Hwang, P. M. (2005). Circulating transcriptome reveals markers of atherosclerosis. *Proceedings of the National Academy of Sciences*, 102(9), 3423–3428.
- Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C., & Liotta, L. A. (2002). Clinical proteomics: Translating benchside promise into bedside reality. *Nature Reviews Drug Discovery*, 1(9), 683–695.
- Poggiali, G., & Renzi, N. (2019). *Presentation and natural course of ulcerative colitis. Ulcerative colitis* (pp. 17–28). Springer.
- Raiol, T., Agustinho, D., Cristina, K., Simi, K., De Souza Silva, C., De, S., Silva, M., Walter, I., Silva-Pereira, I., Brígido, M., Agustinho, D., & Walter, M. (2014). Transcriptome analysis throughout RNA-seq. <https://doi.org/10.1007/978-3-319-11985-4_2>.
- Roos, A., Thompson, R., Horvath, R., Lochmüller, H., & Sickmann, A. (2018). Intersection of proteomics and genomics to “solve the unsolved” in rare disorders such as neurodegenerative and neuromuscular diseases. *PROTEOMICS—Clinical Applications*, 12(2), 1700073.
- Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1), 22–24.

- Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., & Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics: Official Journal of the DNA Methylation Society*, 6(6), 692–702.
- Santos, A. L. M., Vitório, J. G., de Paiva, M. J. N., Porto, B. L. S., Guimarães, H. C., Canuto, G. A. B., das Graças Carvalho, M., de Souza, L. C., de Toledo, J. S., & Caramelli, P. (2020). Frontotemporal dementia: Plasma metabolomic signature using gas chromatography–mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 189, 113424.
- Schmitz, S. U., Grote, P., & Herrmann, B. G. (2016). Mechanisms of long noncoding RNA function in development and disease. *Cellular and Molecular Life Sciences*, 73(13), 2491–2509.
- Schrimepe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies—Challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12), 1897–1905.
- Sitkin, S., & Pokrotnieks, J. (2018). Alterations in polyunsaturated fatty acid metabolism and reduced serum eicosadienoic acid level in ulcerative colitis: Is there a place for metabolomic fatty acid biomarkers in IBD? *Digestive Diseases and Sciences*, 63(9), 2480–2481.
- Sobreira, N. L., Cirulli, E. T., Avramopoulos, D., Wohler, E., Oswald, G. L., Stevens, E. L., Ge, D., Shianna, K. V., Smith, J. P., & Maia, J. M. (2010). Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genetics*, 6(6), e1000991.
- Stewart, T. P., Kim, H. Y., Saxton, A. M., & Kim, J. H. (2010). Genetic and genomic analysis of hyperlipidemia, obesity and diabetes using (C57BL/6J × TALLYHO/JngJ) F2 mice. *BMC Genomics*, 11(1), 1–17.
- Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.-P., Friedländer, M. R., Li, N., Chen, W., Piano, F., & Rajewsky, N. (2009). Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. *Nature Methods*, 6(10), 745–751.
- Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: Large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6), 817–836.
- Sutandy, F. R., Qian, J., Chen, C., & Zhu, H. (2013). Overview of protein microarrays. *Current Protocols in Protein Science*, 72(1), 27, 1.
- Troisi, J., Cavallo, P., Richards, S., Symes, S., Colucci, A., Sarno, L., Landolfi, A., Scala, G., Adair, D., & Ciccone, C. (2021). Non-invasive screening for congenital heart defects using a serum metabolomics approach. *Prenatal Diagnosis*.
- Troisi, J., Landolfi, A., Sarno, L., Richards, S., Symes, S., Adair, D., Ciccone, C., Scala, G., Martinelli, P., & Guida, M. (2018). A metabolomics-based approach for non-invasive screening of fetal central nervous system anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 14(6), 1–10.
- Troisi, J., Landolfi, A., Vitale, C., Longo, K., Cozzolino, A., Squillante, M., Savanelli, M. C., Barone, P., & Amboni, M. (2019). A metabolomic signature of treated and drug-naïve patients with Parkinson's disease: A pilot study. *Metabolomics: Official Journal of the Metabolomic Society*, 15(6), 1–11.
- Troisi, J., Sarno, L., Landolfi, A., Scala, G., Martinelli, P., Venturella, R., Di Cello, A., Zullo, F., & Guida, M. (2018). Metabolomic signature of endometrial cancer. *Journal of Proteome Research*, 17(2), 804–812.

- Troisi, J., Sarno, L., Martinelli, P., Di Carlo, C., Landolfi, A., Scala, G., Rinaldi, M., D'alessandro, P., Ciccone, C., & Guida, M. (2017). A metabolomics-based approach for non-invasive diagnosis of chromosomal anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 13(11), 1–12.
- Turnbull, C. (2018). Genomics in medicine. *Medicine*, 46(12), 774–779. Available from <https://doi.org/10.1016/j.mpmed.2018.09.013>.
- Urban, M. F. (2015). Genomics in medicine: From promise to practice. *SAMJ: South African Medical Journal*, 105(7), 545–547.
- Vlahou, A., & Fountoulakis, M. (2005). Proteomic approaches in the search for disease biomarkers. *Journal of Chromatography B*, 814(1), 11–19.
- Wang, L., McLeod, H. L., & Weinshilboum, R. M. (2011). Genomics and drug response. *New England Journal of Medicine*, 364(12), 1144–1153.
- Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., Hu, Y., Xu, J., Yi, L., & Yang, J. (2013). Mammalian ncRNA-disease repository: A global view of ncRNA-mediated disease network. *Cell Death & Disease*, 4(8), e765.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology*, 54(1), 669–689.
- Weinshilboum, R. M., & Wang, L. (2006). Pharmacogenetics and pharmacogenomics: Development, science, and translation. *Annual Review of Genomics and Human Genetics*, 7, 223–245.
- Yan, M., & Xu, G. (2018). Current and future perspectives of functional metabolomics in disease studies—A review. *Analytica Chimica Acta*, 1037, 41–54.
- Zamore, P. D., & Haley, B. (2005). Ribo-gnome: The big world of small RNAs. *Science (New York, N.Y.)*, 309(5740), 1519–1524.
- Zhang, A., Sun, H., & Wang, X. (2012). Serum metabolomics as a novel diagnostic approach for disease: A systematic review. *Analytical and Bioanalytical Chemistry*, 404(4), 1239–1245.
- Zhao, Q.-Y., Lei, P.-J., Zhang, X., Zheng, J.-Y., Wang, H.-Y., Zhao, J., Li, Y.-M., Ye, M., Li, L., & Wei, G. (2016). Global histone modification profiling reveals the epigenomic dynamics during malignant transformation in a four-stage breast cancer model. *Clinical Epigenetics*, 8(1), 1–15.

This page intentionally left blank

Experimental design in metabolomics

2

Allycia Y. Lee¹, Jacopo Troisi^{2,3,4}, and Steven J.K. Symes^{1,5}

¹*Department of Chemistry and Physics, University of Tennessee-Chattanooga, Chattanooga, TN, United States*

²*Department of Medicine, Surgery and Dentistry “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy*

³*Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy*

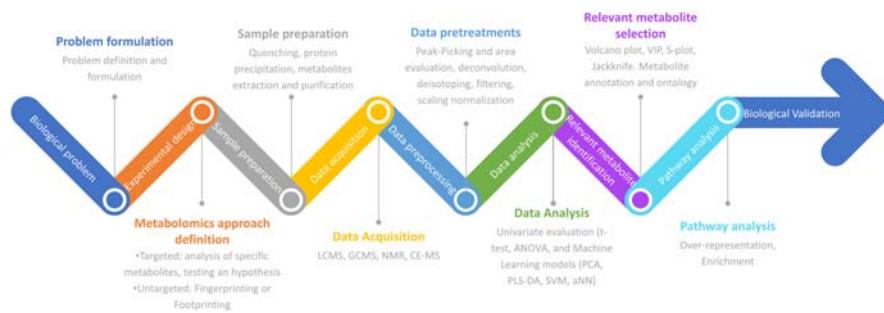
⁴*Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy*

⁵*Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States*

Introduction

Metabolomics is the study of metabolites, which are low-molecular weight (<1.5 kDa) organic and inorganic molecules that are the substrates and products of the biochemical processes within the body. Metabolites are involved in essential cellular functions and can be produced by the host organism or be derived from microorganisms, diet, and other external sources. Together, the pattern of abundances and molecular identities of all metabolites constitute the “metabolome” of a given system under study. Thus one can refer to a “cellular metabolome,” derived from either laboratory cell cultures or specific cells harvested from a lifeform, a “tissue metabolome,” perhaps derived from a biopsy specimen, or a “biofluid metabolome” derived from a chosen fluid of interest. The biological significance of metabolites lies in the fact that they represent the end of the line defined by the processes of gene expression, transcription, and translation to proteins. Whereas genes and proteins may or may not be expressed or translated, respectively, under a given set of conditions, the presence and quantity of a particular metabolite necessarily implies whether or not an upstream process occurred and their pattern is thus a direct reflection of the observed phenotype.

When designing a metabolomic experiment, a general process must be evaluated with different considerations at each step of the overall process. The process starts by deciding on the focus and purpose of the experiment. Based on the focus of the study, the design of the metabolomics experiment can begin to take shape. First, an appropriate sample type must be chosen, followed by a suitable method of analysis. Next, the sample preparation and data acquisition methods should be developed. Ideally, the sample preparation should improve the quality of the data acquisition without

**FIGURE 2.1**

Workflow in the design of a metabolomic experiment. The general pipeline of metabolomics experiment.

introducing systematic errors (Lu et al., 2017). Finally, there are many data processing and analysis steps that need to take place to identify and quantify the metabolites and interpret the data (Worley & Powers, 2013). The general pipeline of metabolomic experiments is illustrated in Fig. 2.1. This chapter highlights the important considerations that should be made when designing a metabolomic experiment as well as some of the more popular options concerning metabolomic experiments.

Applications of metabolomic experiments

Metabolomics can be used to study microbial, mammalian, and plant metabolisms. The metabolomic analysis of biological samples yields metabolomic signatures that can be used to determine details of normal and abnormal metabolic pathways. Metabolomic experiments could simply aim to characterize the metabolome of a certain sample type or disease state, but there are several applications that the experiment could focus on beyond just measuring the metabolome. Depending on the focus of the experiment, there are different considerations for the overall design of the experiment. The applications presented here are some of the more popular ones, for more details see Section 3, Application, of this book.

Biomarker discovery

Biomarker discovery has long been a focus of metabolomic studies in which the goal is to identify specific metabolites, or groups of metabolites, that are impacted by disease states or environmental factors (Johnson et al., 2016). Once identified, the affected metabolites can then be mapped to specific biochemical (i.e., metabolic) pathways in an effort to determine which pathways are dysregulated and consequently provide insight into possible therapeutic options. Knowledge of affected metabolites resulting from a given pathology can also be related to the

genes and proteins responsible for their regulation. Many cancers and other diseases have been studied using metabolomics to better understand the disease, aid in the detection of the disease, and/or aid in the development of treatment options (Beger, 2013; Holmes et al., 2008; Madsen et al., 2010; Raffone et al., 2020; Smith et al., 2019; Troisi et al., 2021). Biomarker detection is primarily studied using biofluids (Chetwynd et al., 2017). The two most challenging aspects of biomarker discovery are metabolite identification and biomarker validation (Johnson et al., 2016). Hundreds of metabolites could potentially be detected in a single sample, but not all of the identities of these metabolites will be able to be determined. In addition, it is likely that there will be variations in which metabolites are detected between samples. Biomarker validation is often challenging because it often requires an entirely separate metabolomic experiment. As an example, biomarkers might be potentially identified using untargeted metabolomics and then followed up with a targeted metabolomic experiment to confirm and validate the results. The difficulty in validating a biomarker often lies in the general lack of targeted follow up experiments (Johnson et al., 2016).

Case-control studies are often the basis of biomarker validation experiments (Bahado-Singh et al., 2013; Troisi, Landolfi, et al., 2018; Troisi, Sarno, et al., 2018; Troisi et al., 2021). After determining appropriate inclusion/exclusion criteria for a specific study, the chosen sampling strategy is employed and then those samples are processed for a particular analytical technique. For successful biomarker identification, the most important concept is consistency; all sampling, every step of sample preparation, and every step of analysis (both analytical and statistical) must be done in the exact same way. Only in this way is there any hope of truly determining if there is a single metabolite, or group of metabolites, that are quantifiably different between cases and controls. Such metabolites can then be considered a “biomarker” for the pathology under consideration. Challenges can arise due to the effect of individual factors on the variation of the metabolome. Age, genetics, diet, and other lifestyle factors all create variation in the metabolome (Bouchard-Mercier et al., 2013; Lankadurai et al., 2013). Well-designed case-control studies attempt to minimize the effects of these individual factors from impacting the metabolome. When considering case-control studies, there are ways to overcome the challenges originating from biological variation, such as incorporation of appropriate quality control (QC) materials throughout the analytical campaign, statistical power calculations to determine appropriate sample size (Blaise et al., 2016; Trutschel et al., 2015), patient questionnaires to help mitigate confounding variables, metabolite normalization, identification of statistically important metabolites (Want & Masson, 2011), and the use of databases to aid in metabolite identifications (Wishart et al., 2018).

Detection of altered biochemical pathways

Metabolomics can be taken a step further than biomarker identification. Once biomarkers for a given disease are determined, they can be utilized to detect how that disease impacted the biochemical pathways of the body. The key metabolites

can be traced to specific metabolic pathways to understand which processes in the metabolic pathways are being affected in the body to produce the disease phenotype (Johnson et al., 2016).

The detection of altered biochemical pathways requires more advanced metabolomic techniques than general metabolomic characterization and biomarker discovery. Some of the techniques that can be utilized are stable isotope tracing and integration of the data with other orthogonal data sets such as those resulting from other -omic studies. Tracing of the biochemical pathways also requires a firm understanding of biochemical pathways and interactions to be able to determine which parts would need to be affected to understand the observed changes in the metabolome. This application of metabolomics is one that may need to be combined with data from other -omics research (Johnson et al., 2016).

Monitoring of response to stimuli

This application of metabolomic experiments is similar to the detection of altered biochemical pathways but focuses more on short-term, more external stimuli. Metabolomics allows for the opportunity to monitor how stimuli like drugs, toxins, or general environmental stimuli affect the metabolome and their related biochemical pathways. Another example would be studies monitoring the responses of bacteria and other microbes to environmental conditions.

Untargeted and targeted approaches

Untargeted metabolomics

Untargeted metabolomics, also called global metabolomics, is the broad analysis of metabolites in a sample. This approach is used when there is no prior knowledge or hypothesis about the metabolome of interest. Untargeted metabolomics is beneficial because it allows for a relatively unbiased analysis of metabolites that derive from multiple metabolic pathways. As such, the untargeted approach is considered hypothesis-generating and is an ideal starting point for new metabolomics studies that lack previous data on what metabolites to expect. This method is also beneficial for the identification of novel metabolites.

In the strictest sense of the word, the “metabolome” consists of all possible metabolites within a given system at a particular instant in time. It is important to consider that the metabolites that are detected, even when using a so-called “untargeted” approach, are necessarily impacted by the sample preparation step(s) and the chosen analytical method. While multiple metabolite classes can be detected in an untargeted approach, it is currently not possible to obtain data for all metabolite classes with a single method given the extreme range of metabolite physico-chemical properties: polarity, charge states, acid-base properties, molecular weights, functional groups present, among others. Factors, such as pH, matrix

composition, and column chemistry used during sample preparation and analysis, can alter which metabolite classes are detected using a particular analytical methodology (Johnson et al., 2016). Because complex datasets are produced, multifaceted computational tools are required to aid in the identification and correlation of the metabolites.

Targeted metabolomics

Targeted metabolomics is the analysis of a subset of metabolites that have been identified based on prior knowledge of the system to be studied. The lower number of metabolites allows for higher sensitivity and selectivity compared to untargeted metabolomics since the methods of analysis are optimized for the specific set of chosen metabolites. A targeted analysis often follows from the data generated from an untargeted analysis and serves to validate and expand upon the previous data.

Sample types

This section will cover the main sample types that have been used in metabolic research. For each sample type, consideration must be given to the timing and overall methods of collection, sample preparation, and analytical measurement, as these decisions directly impact the quality of the data generated and ultimately influence the robustness of subsequent biological interpretation (Álvarez-Sánchez, Priego-Capote, & de Castro, et al., 2010b; Chetwynd et al., 2017; Duportet et al., 2012; Vuckovic, 2012). The types of samples that can be collected and studied can be categorized as tissues, primary or immortalized cells, and biofluids. The sample type to be utilized is dependent upon the system to be studied and may also depend on budget and ethics. In some cases, the collection of an ideal sample may be too expensive or ethically unacceptable, so a surrogate sample type must be used instead (Chetwynd et al., 2017). Plans for the transport and storage of collected samples should also be made prior to collection as the metabolome of many sample types can be affected over time if not handled properly. Discussion of specific sample preparation considerations unique to each sample type will be included.

Metabolically active versus metabolically inactive

The sample type that is chosen should be qualitatively and quantitatively representative of the biological system to be studied. All sample types can be classed as either metabolically active or metabolically inactive samples. If processed properly, metabolically active samples can provide a snapshot of the metabolism at the time of collection. These types of samples include tissue, blood, and saliva. Metabolically inactive samples, on the other hand, provide information about the

metabolism over some period of time. Some biofluids, such as urine, are considered metabolically inactive (Chetwynd et al., 2017).

In general, metabolically active samples represent the intracellular metabolome, or endometabolome. While these types of samples are very valuable for studying specific aspects of metabolism, they typically require more work. Due to the intracellular nature of these types of samples, the metabolism must be quenched and the metabolites extracted in a timely manner (Álvarez-Sánchez, Priego-Capote, & De Castro, 2010a; Chetwynd et al., 2017; Duportet et al., 2012; Vuckovic, 2012). Furthermore, these samples tend to require more extensive sample preparation, which could become an issue if the samples are not stored and handled properly as metabolically active samples have metabolomes that are more likely to change during sample preparation. It is recommended that the samples, specifically biofluids, are kept on ice while being processed and stored long-term at -80°C (Chetwynd et al., 2017).

Metabolically inactive samples, on the other hand, represent the extracellular metabolome, also called the exometabolome or metabolic footprint. These types of samples may not typically require quenching, allowing for minimal sample preparation and high-throughput of metabolite analysis.

Tissue and cells

Tissue and cell analysis is best for the study of specific, localized systems within an organism. Information obtained through the study of cells and tissues can complement information obtained from the study of biofluids, which would provide whole organism metabolic profiling. These types of samples are better for studying response to stimuli and pathogenesis, especially when investigating the details of the biochemical mechanisms (Chetwynd et al., 2017; Johnson et al., 2016). As a result, cells and tissues are very good candidates for the metabolomic study of cancers (Johnson et al., 2016).

Tissues

There are many types of tissues in the body, and a wide variety of them have been used in metabolomic studies. These tissues include muscle, cardiac tissue, placental tissue, skin, and blood vessels. Most of the time, collection of tissue samples involves invasive procedures that require an operating room. Some exceptions to this would be the placenta, which is delivered after the delivery of the fetus, skin biopsies, and feces. Collection of tissues is more difficult than other sample types due to the invasiveness of the collection and inhomogeneity of the samples. Careful consideration must be taken during sample collection to ensure that the samples are from identical anatomical areas on different subjects. This means a specific area, such as a lobe or region (inner vs outer), of an organ is being sampled for each subject.

Tissues are a metabolically active sample, so they must be quenched as soon as possible after collection and be stored appropriately until analysis. Storage of

the collected samples should be at -80°C with no more than 48 hours spent at 4°C .

Primary and immortalized cells

The metabolomic study of cells is beneficial in the sense that it is relatively unaffected by the confounding factors that other sample types are prone to. The data collected from primary and immortalized cells can complement whole organism metabolomic information obtained from the usage of other sample types and with other “omic” sciences. Primary cells refer to the normal, healthy cells in the body that undergo normal cell cycles. Immortalized cells is a term that is often used to describe tumor and cancer cells as they do not follow regular cell cycles and often do not undergo apoptosis. Cell metabolomics has been used for drug discovery, foodomics, and metabolic tracing.

To study primary and immortalized cells, the cell sample of interest is grown in cell cultures. Special considerations must be made towards the growth of cells to allow for consistency and reproducibility because there are many challenges in sample preparation that can be time-consuming and may lead to sample degradation. Some of these issues are growth medium variability, metabolic quenching, and extraction. There are several experimental design recommendations to help reduce the impact of these issues. For one, it is recommended that sample extraction and analysis are randomized. It is also recommended that studies should be conducted using six biological replicates, which are cultures that are grown in separate containers but are treated the same way. For the culture media, it is recommended that the same media is used for all cell cultures. The growth medium should be chosen very carefully as a suboptimal medium may cause changes to the metabolome. The growth medium should also be completely washed from the obtained cell pellet to avoid contamination of the cells. These recommendations should reduce the variability in the cell cultures. Another important consideration in growing cell cultures is to ensure that the method prevents leakage of the intracellular metabolites into the growth medium.

Metabolic quenching and extraction are important steps of sample preparation for cell cultures (Canelas et al., 2009; Duportet et al., 2012). These steps should take place as soon as possible since active enzymes can still change the metabolite concentration over time. Cells in suspension and adherent cells are two different types of cell cultures that may occur. There are some differences in the preparation of cells in suspension and adherent cells with adherent cells typically requiring more steps as the cells need to be separated.

Whole blood, plasma, and serum

Human blood is the most widely studied biospecimen in metabolomics (Nagana Gowda & Raftery, 2017). More than 3000 metabolites have been detected and quantified in human blood, whereas many thousands of other small molecules have either been detected and/or are expected to be present. The serum metabolome has

been described in detail (Psychogios et al., 2011) and a free, searchable database (<https://serummetabolome.ca/>) is available online. Blood is a metabolically active sample that is representative of many tissues and systems within the body. As such, blood is a good option for untargeted metabolomics. There are various components of the blood that can be analyzed in metabolomics. The three types of blood samples that can be processed include whole blood, plasma, and serum. Whole blood is a normal blood sample that has not been processed in any way and still contains all of its cells and platelets. Plasma and serum are the liquid components of whole blood that only differ in the presence (plasma) or absence (serum) of fibrinogen and other clotting factors (Yu et al., 2011).

Most metabolomic studies will utilize plasma or serum samples, but whole blood may be better in cases of nuclear magnetic resonance (NMR) analysis, specifically when studying coenzymes and antioxidants (Nagana Gowda & Raftery, 2017; Stringer et al., 2015). Whole blood is also preferred for studies that would benefit from analysis of red blood cells (Nagana Gowda & Raftery, 2017). For studies using other methods of analysis, serum or plasma are more likely to be used. Generally, serum and plasma samples have been shown to yield similar results as long as the preparation of the samples is the same. There are some cases where serum or plasma may be preferred over the other. For example, serum is preferred for studying cardiac troponin, and serum is preferred for oral glucose tolerance tests (Yu et al., 2011). Plasma produces more reproducible results than serum, but serum provides greater sensitivity as the metabolites are more concentrated, which is beneficial for biomarker detection (Yu et al., 2011).

The collection of blood samples requires an individual who has been trained in the collection of blood, such as phlebotomists and nurses, and is typically performed in a clinical setting. An exception would be the collection of dried blood spots (DBS) which can be done by the subject and delivered by mail to the clinic. Whole blood samples are the easiest to obtain as it is simply the unaltered blood drawn from an individual. Plasma, which still retains its clotting factors, is isolated from whole blood by adding an anticoagulant, such as EDTA or heparin, and centrifuging to separate the plasma from the red blood cells, white blood cells, and platelets. Lithium heparin (MW >10 kDa) would be the preferred anticoagulant for metabolomic studies as it has a higher molecular mass than most metabolites (MW < 5 kDa) and is not an endogenous compound (Chetwynd et al., 2017). There are specialized tubes used in clinical settings that are coated with the anticoagulant for the collection of plasma samples. To obtain a serum sample, the blood is allowed to clot before being centrifuged. The clotting will separate the fibrinogen and other clotting factors from the remaining liquid component. Blood samples should be stored at -80°C in 0.5 or 1.0 mL aliquots. When being processed, the sample should be thawed on ice and processed at 4°C . The low temperatures quench the metabolic activity that may occur due to the presence of proteins and enzymes.

The least invasive method to collect blood samples utilizes the DBS technique. Here, 1 drop of blood from a heel, toe, or finger-prick is deposited on a

preprinted paper card followed by sample desiccation (see Fig. 2.2). This collection system is especially useful for fragile patients or for children. Indeed, it is also used for neonatal screening (see Chapter 6: Targeted Metabolomics). DBS samples have been used for untargeted metabolomics (Troisi et al., 2019; Ward et al., 2021).

Blood samples tend to require more sample preparation than other biofluids because of their composition. Blood is made up of cells, platelets, water, metabolites, proteins, RNA, and DNA. There is also a wide range of metabolites from small ionic species to the larger lipids. Sample preparation, specifically extraction, will depend on the metabolite classes to be studied as not all metabolite classes can be studied simultaneously. Various extraction methods will be discussed later in this chapter. For gas chromatography-mass spectrometry (GC-MS) analysis, the sample should be dried and derivatized, which will also be discussed later. Liquid chromatography-mass spectrometry (LC-MS) analysis does not require drying of the sample but may need to be diluted prior to analysis. Solid-phase extraction (SPE) may also be applied for LC-MS applications.

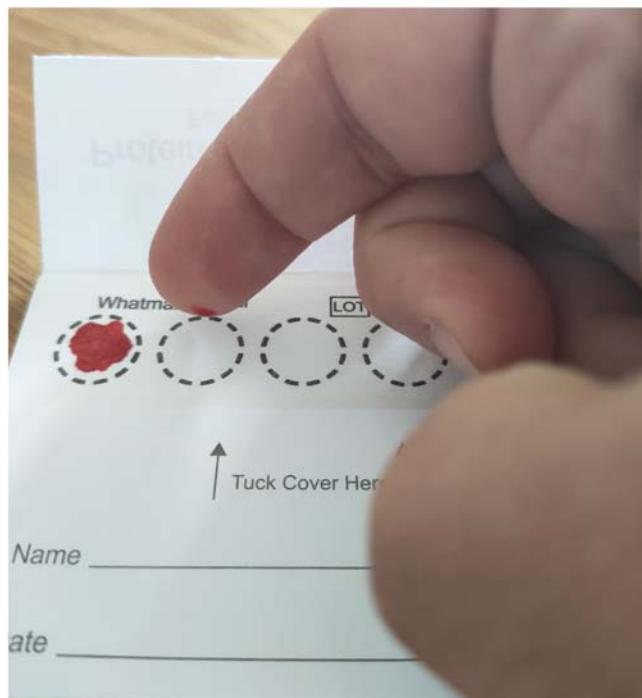


FIGURE 2.2

Dried Blood Spot samples collection. One drop of blood is deposited on a preprinted paper card followed by samples desiccation.

Urine

Urine is also one of the most studied biofluids in metabolomic research because of its ease of collection. Urine is composed of endogenous and exogenous metabolites that have been filtered from the bloodstream by the kidneys. In contrast to other biofluids, the collection of urine requires no training and can be collected by the subjects themselves in home or in clinic. Urine is typically used for studying whole body metabolic response to disease and other stimuli. In other words, urine would be a good choice for untargeted metabolomics. It has been used in biomarker discovery, drug investigations, determination of nutritional status and detection of environmental toxins. So far, over three thousand metabolites have been successfully characterized in human urine. Like the serum metabolome, the human urine metabolome has also been characterized in great detail ([Bouatra et al., 2013](https://urinemetabolome.ca/)) along with an online database (<https://urinemetabolome.ca/>).

The urinary metabolome is heavily affected by the time of collection. Urine samples are generally taken using three methods: first morning void, spot urine, and 24 hour urine. First morning void samples are collected from the first urination of the day, ideally in the morning when the body has had an overnight fast, minimizing the effects of food and medication on the metabolome. This type of sample is the preferred method for metabolomic studies. Spot urine samples are collected during some other urination during the day. The key to spot samples is to have all subjects collect spot samples at time points that would be uniform between all subjects; first morning void is a type of spot sampling. Spot urine samples are useful for analyzing the effects of diet or medication on the metabolome. The metabolites in spot urine samples are prone to fluctuation in excretion by diurnal or cosine rhythm. The last urine sample type is a method of reducing the impact of these rhythms. Twenty-four hour urine samples are pooled samples that contain samples from all urinations in a 24 hour period. This sampling reduces circadian variation but will likely be affected by food and medications. When collecting urine samples, it is recommended that the urine is collected mid-stream in order to avoid bacterial contamination. Urine is normally sterile while in the bladder, but bacteria can be introduced to the urine from the urethra during urination.

Urine collection from small children is challenging and several methods have been described. One employs collection bags that can be attached to the child genitalia and the other utilizes cotton balls inserted into the diaper. The former is burdened by frequent redness in the peri-genital area where the sachet is made to adhere. The second system, on the other hand, does not present this inconvenience but the collected urine must be separated from the cotton wool and this can introduce a disturbance to the profiling especially for untargeted studies.

Urine samples degrade quickly at room temperature, so samples should be frozen as soon as possible after collection. For sample preparation and analysis, the urinary metabolome can be kept at 4°C for up to 48 hours without significant alteration. For long term storage, a minimum of -20°C is required for

stable storage for at least 6 months; -80°C is the recommended storage temperature. Metabolite stability in urine has not been tested for more than 6 months. Based on LC-MS analysis, the urinary metabolome can withstand up to nine freeze thaw cycles without significant changes (Chetwynd et al., 2017).

Urine is a metabolically inactive sample type. These types of samples generally do not require as much sample preparation as metabolically active samples. The preparation of urine will vary depending on the method of analysis. For LC-MS and NMR, centrifuged neat or diluted urine may be injected; however, both preparation methods have their issues when analyzed via LC-MS. Neat urine means that the urine metabolites are unmodified by sample processing. While neat urine would allow for the acquisition of the pure metabolome, the sensitivity of low abundance metabolites suffers during LC-MS analysis. This lack of sensitivity is caused by the coelution of low concentration metabolites with high abundance peaks and ion suppression. Ion suppression occurs due to the high salt concentration of urine, which forms adducts within the ESI source and fouls the LC column and ESI source. Dilution of the urine sample will help reduce ion suppression, but it may also lower the signal of low abundance peaks below the limit of detection (LOD). Similarly, in GC-MS analysis, the high concentration of urea that is present in urine could coelute with other lower abundance metabolites, reducing the number of detectable metabolites. Pretreatment with the enzyme urease is an option for removing urea (Michell et al., 2008), but this pretreatment has been reported to have negative effects (Kind et al., 2007). SPE and headspace (HS) solid-phase microextraction (SPME) are good choices for sample preparation as well.

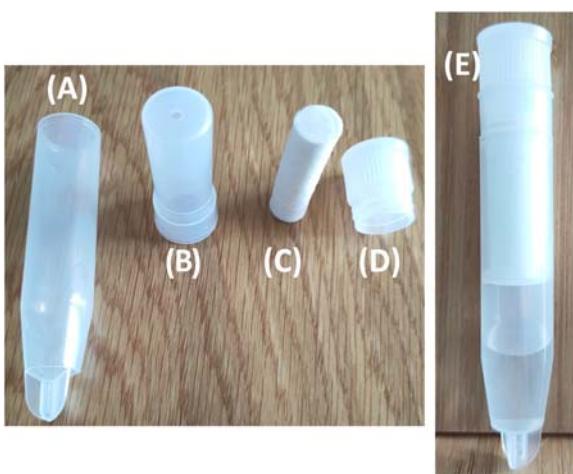
Other biofluids

It seems that most, if not all, biofluids have the potential to be studied using metabolomics. Based on the production of a particular biofluid, specific body systems or diseases can be studied. This section covers some of the less used biofluids and the specific instances in which they might be used.

Saliva

Saliva is perhaps an underutilized biofluid for metabolomic analysis, but it has the potential to serve as a surrogate sample for plasma. It is thought that the metabolome of saliva will be reflective of the plasma metabolome and would allow for noninvasive sample collection. One drawback of saliva samples in comparison to plasma would be the lower concentration of salivary metabolites, but this decreased concentration could be improved by including a preconcentration step in the sample preparation. The saliva metabolome has also been characterized (Dame et al., 2015) (see <https://salivametabolome.ca/>).

There are two subtypes of saliva samples: stimulated and unstimulated. Stimulated saliva samples are collected by using citric acid to induce the production of saliva. Unstimulated or resting saliva samples allow for the natural production of saliva. An investigation of both sample types revealed that stimulated

**FIGURE 2.3**

Saliva sample collection system. Saliva collection kit (A). Collection tube (B). Holed swab tube (C). Swab (D). Tube Top (E). Collection tube after centrifugation.

samples have lower concentration of citric acid cycle and amino acid metabolites due to dilution. As a result, it is recommended to use unstimulated saliva collection for metabolomic studies. To avoid contamination, collection typically occurs after a period of fasting and delayed oral hygiene.

Several saliva collection systems can be used. The gold standard is the Salivette collection tube consisting of a tube with an internal mini-tube with a holed bottom where a cotton swab is deposited after 1 minute of chewing to make it soaked in saliva. This system (tube, holed mini-tube and swab), after centrifugation, allows the separation of purified saliva (free from food debris, mucous, etc.) from the swab (see Fig. 2.3).

Saliva is considered to be a metabolically active sample, so similar storage and sample preparation considerations as those made for blood samples may be utilized. In other words, the samples should be frozen at -80°C as soon as possible after collection. For slightly shorter periods of storage, saliva is stable for up to 3 weeks at -20°C . For the preparation of these samples, the sample should first be centrifuged to remove unwanted debris from the oral cavity, like food and cellular constituents. Most saliva analyses have been conducted using NMR, but some LC-MS and GC-MS studies have also occurred (Chetwynd et al., 2017; Mueller et al., 2014). For NMR, 0.2 M phosphate buffer and D_2O have been used to buffer and dilute the samples.

Cerebrospinal fluid

Cerebrospinal fluid (CSF) is the fluid that resides in and aids in fluid regulation and nutrient transport in the central nervous system (CNS) of the body, namely

the spinal column and brain. CSF would be the recommended sample type for studying neurological diseases, especially those that originate in the CNS because metabolite abundances in CSF are dependent on metabolism in the brain (de Paiva et al., 2014). Alzheimer's disease and amyotrophic lateral sclerosis are examples of the diseases that could benefit from the metabolomic analysis of CSF (Chetwynd et al., 2017).

CSF samples must be collected in a clinical setting by a healthcare professional via lumbar puncture (see Fig. 2.4). The sample should be clear with no blood. Due to the invasiveness of the sample collection, CSF would be a difficult sample to obtain routinely. The sample should be centrifuged and then frozen at -20°C or lower for long term storage. For short-term storage, the samples would be stable at 4°C for up to 48 hours.

CSF samples are considered to be metabolically active. As such, similar sample preparation considerations made for other metabolically active biofluids have been applied to CSF (Wishart et al., 2008). To date, a significant portion of the



FIGURE 2.4

Cerebrospinal fluid (CSF) collection during rachicentesis. CSF samples collection in a clinical setting by a healthcare professional via lumbar puncture.

CSF metabolome has been studied with ^1H -NMR and GC-MS approaches (Holmes et al., 2006; Pears et al., 2007; Sweatman et al., 1993; Wishart et al., 2008). As for serum, urine and saliva, a database has also been generated for CSF (Wishart et al., 2008) (<https://csfmetabolome.ca/>).

Amniotic fluid and breast milk

Amniotic fluid is the ideal biofluid for the study of feto-maternal health because it derives from both the maternal and fetal tissues (Bardanzellu & Fanos, 2019). The composition of amniotic fluid affects the growth of the fetus and represents the health of both the mother and the fetus. Metabolomic analysis of amniotic fluid provides dynamic information about the fetus development and other pregnancy conditions, such as the study of the effects of Chinese medicine on pregnancy (Shan et al., 2019) or the understanding of spontaneous preterm births (Menon et al., 2014). Gestational diabetes and congenital diseases have also been studied using metabolomic analysis of amniotic fluid (Bardanzellu & Fanos, 2019). NMR, LC-MS, and GC-MS have all been used to analyze the metabolome of amniotic fluid with each method yielding different subsets of metabolites (Shan et al., 2019). NMR is beneficial for low volume samples like amniotic fluid because it is nondestructive and has high reproducibility, but it has low sensitivity. GC-MS and LC-MS are becoming an increasingly popular option due to their increased sensitivity.

Amniotic fluid collection requires a medical intervention since it requires an ultrasound guided collection using a specialized syringe designed for this purpose (see Fig. 2.5). Such collection is also a challenge from an ethical point of view because the procedure does present an inherent risk of spontaneous termination of pregnancy.

Breast milk is the preferred sample type for analyzing the nutritional intake of infants. Breast milk is similar to urine in that it is also affected by diurnal cycles and diet, so the same types of sample collection could be considered (Ten-Doménech et al., 2020). In the case of a spot collection, the time of the mother's last meal should be noted. A 24-hour pool would aid in reducing the effects of diurnal variation, but this collection method raises ethical concerns as it may conflict with the feeding of an infant (Ten-Doménech et al., 2020). Collected samples should be pasteurized and stored at -80°C ; -20°C is also an acceptable temperature for short-term storage. Pasteurization is meant to reduce the chance of contamination by infectious agents; however, the process can cause some changes to the properties of the milk. The effects of storage time and freeze-thaw cycles on metabolome composition have not been investigated. LC-MS, GC-MS, CE-MS, and NMR are all analytical methods that can be utilized for the study of breast milk, and there are several extraction methods and analytical techniques that can be utilized to extract certain subsets of metabolites (Chetwynd et al., 2017; Ten-Doménech et al., 2020).

**FIGURE 2.5**

Amniotic fluid collection. Ultrasound guided amniotic fluid collection during amniocentesis.

Sweat and tears

Sweat can be modified during some disease states, allowing it to become a more popular option in studying diseases. Sweat collection must be done under standardized conditions as the anatomical site, skin, environmental temperature and humidity can affect the metabolome of a sweat sample (Hussain et al., 2017). Sweat can be induced by a device called a sweat inducer that heats an area of the skin to produce sweat. There is a small number of studies that have used sweat as the sample type. Of those, LC-MS or NMR were used with minimal sample preparation (Chetwynd et al., 2017).

Tears are produced by the tear film, which is a thin, fluid layer that covers the anterior surface of the eyeball and has roles in the protection and health of the eyes (Chen et al., 2011). Tears would be the ideal sample type for any studies involving ocular health. The metabolomic study of tears is limited by the small volume that can be collected in a single instance. The collection of tear samples is noninvasive, but the samples are normally only 5–10 µL in volume (Chen et al., 2011). Tears can be collected using Schirmer paper, capillary tubes, and microglass pipettes. So far, tears have primarily been used to study dry eye disorders using ¹H-NMR or LC-MS (Yazdani et al., 2019).

Analytical methodologies

Even though data acquisition occurs after sampling in the pipeline of a metabolic experiment, it is necessary to decide on the method of analysis first. Some methodologies may require greater sample preparation, and different analytical methods will allow for the detection of different metabolite classes (Lu et al., 2017). As such, it is important to identify the method of analysis prior to sample preparation. Only a few of the most prominent methodologies will be discussed in this chapter. Other methodologies will be discussed in other chapters.

Nuclear magnetic resonance

NMR was the first analytical method used in metabolomics and aided in the development of the field itself (Chetwynd et al., 2017). NMR was a good option for metabolomics because it is able to detect all organic metabolites, and its signal intensities are proportional to the concentration. Of the different modes that can be utilized, one dimensional ^1H NMR is the most sensitive and the most used NMR experiment for metabolomics.

Though NMR was the first analytical method utilized, it has a significantly lower sensitivity when compared to the mass spectrometry (MS)-based techniques that have been gaining popularity. Additionally, NMR has a limited resolution, which causes difficulties in analyzing complex metabolite samples. At most an NMR can detect up to the 50–70 most abundant metabolites in a sample (Snytnikova et al., 2019). While a much smaller number of metabolites is detected in NMR analyses, there are still issues that can arise with overlap of metabolite signals. Most metabolite signals will fall within the chemical shift range of 1–10 ppm, causing overlap with almost all aliphatic signals with those of metabolites. This overlap makes identification and quantification more difficult. To compensate for these difficulties, the software side on NMR-based metabolomic experiments has improved greatly. Over the years, many techniques have been developed to help in the analysis of the spectra, such as database searches, statistical deconvolution, and multidimensional analysis.

Mass spectrometry

Mass spectrometry is the basis for many metabolomic analyses. It is becoming the preferred analytical method because it offers the most sensitive detection and broadest metabolome coverage, being able to detect hundreds of metabolites with a single method (Lu et al., 2017; Snytnikova et al., 2019). Though MS does provide the best metabolite coverage, it is important to note that these methods will only be able to detect metabolites that can be ionized. MS is usually preceded by a column-based separation technique, with GC and LC being the most frequently

used. Direct injection mass spectrometry can be used, but is best performed with a high-resolution, high mass accuracy instrument in order to minimize the isobaric interferences resulting from a multitude of metabolites within a given sample. Incorporation of a chromatographic step also helps minimize ion suppression effects that may compromise detection of lower abundance metabolites. Direct injection can be useful for the analysis of enzyme assay and cellular extracts, but it serves more as a discovery tool and should be followed by an analysis involving chromatographic separation (Lu et al., 2017).

There are a variety of mass spectrometers that can be used for metabolomics. The best mass spectrometer for a given experiment will depend on whether an experiment will have a targeted or untargeted approach. For a targeted approach, a triple quadrupole is preferred because you can monitor for the appearance of parent/product fragmentation reactions, also called selected reaction monitoring. For untargeted approaches, the high sensitivity of a single quadrupole mass spectrometer makes for a highly capable instrument, despite the lower resolution compared to other mass spectrometers. For example, time-of-flight (TOF) or Orbitrap mass analyzers have substantially higher mass resolution and thus the ability to determine “exact” masses. Both TOF and Orbitrap instruments, including hybrid instruments that also incorporate a quadrupole, are powerful platforms that are used for targeted and untargeted experiments. Tandem MS/MS systems offer the highest possible selectivity and are best suited for the identification of unknowns. Generally, any of these mass spectrometers would allow for specific, sensitive, and quantitative metabolite analysis. With this in mind, the choice of mass spectrometer should be decided upon carefully based upon the analytical requirements, budget, and accessibility of a given instrument.

Gas chromatography

GC-MS is a technique that is capable of measuring a wide range of water-soluble metabolites (Danielsson et al., 2012; Troisi et al., 2021). In fact, GC-MS is the best analytical method for low molecular weight and volatile metabolites. It can detect small and uncharged species, like very short fatty acids, sugars, acids, and alcohols, that may not be detected using LC-MS. GC-MS is also the only universally applicable method for the analysis of essential oils and other volatile compounds (Lu et al., 2017). Due to the very long column lengths, GC has the highest separation efficiency and cleanest chromatograms, which takes some pressure off of the performance of the MS (Danielsson et al., 2012). For GC, a single quadrupole is very effective, but a TOF mass analyzer could also be used if higher mass resolution is needed.

Based on the principle of how a GC operates, it is only effective for compounds that will volatilize (without decomposing) within the operating temperature. For this reason, GC-MS analysis is almost always preceded by the derivatization of metabolites to maximize metabolite coverage. It is crucial that as many of the compounds within the sample are volatile because nonvolatile compounds can degrade inside the GC inlet and cause cross-contamination. While hot

injection is a technique that can aid in the volatilization of compounds from the injection syringe, it can impact the quantification of thermolabile compounds even after derivatization. These considerations are important because decomposition of any metabolites will be recorded in the chromatogram and could increase the difficulty of interpretation (Lu et al., 2017). To summarize, both the samples and inlet of the GC should be clean and maintained throughout the experiment (Fiehn et al., 2000).

Hard ionization techniques, such as electron impact (EI), are often used with GC-MS to create many reproducible fragments that aid in identification of metabolites. Though hard ionization techniques are useful for identification, these techniques will produce few, or no, molecular ion peaks. If the molecular ion is of interest for a particular study, a soft ionization technique such as chemical ionization would be useful.

Liquid chromatography

LC-MS is a versatile analytical method that allows for the coverage of many metabolites and is especially suited for the analysis of fatty acids and lipids using reversed-phase chromatography. Unfortunately, there is not an individual LC method that allows for the analysis of all metabolites simultaneously, so multiple LC methods are typically needed to maximize the metabolites detected. LC-MS is limited to the analysis of ionizable metabolites, therefore it is important that the ionization source that will be used can ionize most of the metabolites of interest.

LC-MS is a very powerful analytical tool for metabolomics, but it does have a few drawbacks. For one, LC-MS can be less precise than other methods. There can be significant variation of integrated peak areas, between runs and between days, with an average relative standard deviation of about 10% (Lu et al., 2017). LC-MS methods also tend to have issues with ion suppression and differential adduct formation. Ion suppression is where high concentration ions mask lower abundance ions due to coelution. This can be caused by matrix components or analytes. Differential adduct formation occurs due to the coelution of metabolites and salts and causes the formation of adduct peaks, reducing the molecular ion peak of the metabolite. These issues are not insurmountable, but it is very important to keep them in mind while developing a method. These issues can usually be managed through sample preparation and improved chromatographic separation.

Techniques without sampling

Techniques that do not require the use of sampling would be situations where the metabolome needs to be studied *in vivo*. These types of techniques provide real-time data and allow for decisions to be made during a procedure. An example would be the iKnife, which stands for intelligent knife (Chetwynd et al., 2017). This is a tool used during surgery where a surgeon uses an electrosurgical knife to apply an electric current to the tissue being cut. The electric current vaporizes

the tissue into smoke that can be sucked into a MS in the operating room (Chetwynd et al., 2017). The MS data would help the surgeon differentiate between healthy tissue and a tumor. More applications like this are being developed for other types of procedures.

Sample preparation

Sample preparation for any study will be dependent upon the chosen sample type, class of metabolites to be analyzed, and analytical method. In general, sample preparation should be as simple, fast, and reproducible as possible. Many key sample preparation techniques will be discussed, but few of them may actually be necessary depending on the experimental design. Ultimately, the sample preparation phase could be as short as a few minutes or as long as a day, so it is important to work quickly and efficiently to avoid any time-dependent alterations to the metabolome. Recall that for most samples, a holding time of 48 hours at 4°C should not be exceeded to minimize metabolite degradation.

Quenching

The ultimate goal of quenching is to terminate any enzymatic activity within the sample in an effort to produce a stable extract that quantitatively reflects the metabolome. Quenching is imperative for highly metabolically active samples, such as cells and tissues. Quenching can be attained through a combination of heat, cold, acid, base, and organic solvents. It is important to quench the sample as soon as possible after collection because the turnover rate for some metabolites can be as fast as 1 second (Martano et al., 2015; Mashego et al., 2007; Van Gulik et al., 2012; Vuckovic, 2012). This step must be performed carefully as there are often issues with the alteration of the metabolome during sample collection or incomplete termination of enzyme activity. Incomplete quenching can be monitored experimentally through the observation of transformation of isotope-labeled standards that have been spiked into the quenching solvent.

For studies involving cells, the first step is to separate the cells from the growth medium. Pelleting of the cells via centrifugation takes too long and can affect the metabolome, so this is not a recommended method of isolation (Lu et al., 2017). For cells in suspension, the cells should be isolated using fast filtration, like MxP FastQuench, and then placed immediately into the quenching solvent (Munger et al., 2008). For adherent cells, the quenching solvent can be added directly via aspiration without removal of the media (Lorenz et al., 2011). After separating the cells from the media, washing the cells with warm phosphate-buffered saline (PBS) may be helpful. Cold PBS is not preferred as it increases the chance of metabolite leakage from the cells (Mashego et al., 2007; Wittmann et al., 2004). A situation in which washing with PBS would be

necessary is the analysis of intracellular amino acids as growth media often contain a high concentration of amino acids (Lu et al., 2017). Once the cells have been separated, the direct addition of a hot or cold organic solvent will quench the cells. Boiling ethanol is a typical choice despite the chance for analyte thermal degradation, although the high temperature will reliably denature all enzymes in the sample. While cold organic solvents are also an option, they have a higher risk of slow or incomplete quenching. If a cold solvent is to be used, 4°C isotonic saline would be a possible option as it is less likely to cause metabolite leakage compared to cold 100% methanol (Chetwynd et al., 2017).

For tissues, metabolic quenching of the samples normally involves the washing of the tissue in saline or a phosphate-buffered solution before being placed in liquid nitrogen. Unfortunately, the heat transfer between liquid nitrogen and warm tissues is not very fast, so smashing the tissue against precooled metal plates is a faster alternative that has been developed (Maharjan & Ferenci, 2003). Ideally, quenching would occur immediately after collection. In cases where the sample must be collected in an operating theater and liquid nitrogen or other quenching apparatus is not allowed, the tissue sample should be quickly transported on ice to a location where the sample can be quenched. The inability to immediately quench a sample in the operating room should be carefully considered if studying metabolic pathways with high metabolic flux (Chetwynd et al., 2017). Once the sample has been quenched, tissue samples must be homogenized before moving on to extraction. Homogenization usually involves a physical technique, like a mortar and pestle, a cryomill machine, or a stainless steel or silica particle covered ball. Extraction solutions tend to be added during the homogenization, allowing for cell lysis and metabolite extraction to take place simultaneously.

Extraction

Extraction is a process that aims to release metabolites from within a sample's cells and increase the quantitative yield of metabolites. The extraction method is important to optimize for a given study since different approaches can lead to contradictory interpretation of the resulting metabolomes (Duporet et al., 2012). For most studies, it would likely be preferred to only extract free metabolites as opposed to protein bound metabolites. The extraction of protein bound metabolites would only be necessary when studying metabolites like NADP⁺ that are often bound to protein.

The primary method of extraction is liquid-liquid extraction (LLE) which utilizes organic solvents to precipitate higher molecular mass biochemicals, such as proteins, RNA, and DNA. Most, if not all, metabolites will remain in solution. Generally, many metabolites can be extracted using a variety of organic solvents. Commonly used organic solvents are methanol, acetonitrile, isopropyl alcohol (IPA) and acetone. IPA is frequently used for the extraction of lipids (Jiye et al., 2005).

A specific type of LLE that may be utilized is monophasic extraction, which involves one miscible solvent system. These extraction solutions are well suited to

extract a wide range of metabolites. This type of extraction is often used for cell cultures, tissues, and even some biofluids. Typical combinations of organic solvents used are methanol/water, water/acetonitrile, and methanol/water/chloroform (Troisi, Landolfi, et al., 2018; Vorkas et al., 2015). For cells and tissues, 2:2:1 acetonitrile/methanol/water with 0.1 M formic acid is an effective solution to quench the sample and extract the metabolites (Rabinowitz & Kimball, 2007). In this case, ammonium bicarbonate should be added afterwards to counteract the acidity of the formic acid. For serum or plasma samples, just methanol may be enough.

Another type of LLE is biphasic extraction which utilizes two immiscible solvent layers (Patterson et al., 2015). The two layers allow for water-soluble metabolites to be separated from lipids. Lipids may need to be separated from water-soluble metabolites for a couple of reasons. First, lipids would be better analyzed under different conditions than that of water-soluble metabolites. The separated analyses would also provide more information overall when put together. Second, glycerophospholipids as well as some other lipids can interfere with the data quality and sensitivity in untargeted metabolomics, so researchers may choose to exclude lipids from the extracted sample (Chetwynd et al., 2017). The common solvent system used for biphasic extraction is water/methanol with a nonpolar solvent and chloroform, dichloromethane, or methyl tert-butyl ether (MTBE) (Moldoveanu & David, 2019). Typically, the aqueous layer (water-soluble metabolites) is on top when using chloroform and dichloromethane and on bottom when using MTBE. When doing a biphasic extraction, it is easy to access the top layer, but the bottom layer requires crossing through the layer of cellular debris between the two layers. If possible, it would be best to use the solvent system that would place the metabolites of interest in the top layer to avoid contamination by cellular debris.

It is important to keep in mind that the quantity of metabolites does not necessarily indicate that a successful extraction occurred. Sometimes the concentration can be due to artificial production of another metabolite during sample processing. It should also be noted that two extractions can help to increase the metabolite yield, but more than two extractions is likely to be more detrimental (Lu et al., 2017).

Sample clean up

Sample clean up is necessary for most metabolite samples with and without quenching and extraction. Sample clean up could be as simple as a dilution or much more complicated depending on the sample. These processes make alterations to the matrix and/or metabolite concentrations in order to reduce the chance of damaging the analytical instrument used or to increase the quality of the data.

Solvent removal

After extraction, samples often need to be cleaned up as the organic solvents used in quenching and extraction could cause problems in later steps or during

analysis. These potential issues are prevented by removing the organic solvents via drying (Vuckovic, 2012). The different approaches for drying a sample include the use of a lyophilizer, a N₂ evaporator, or a Speedvac at room temperature. Most metabolites will not have any issues being dried, but redox-active species (NADPH and GSH) could potentially oxidize, changing the metabolite concentrations (Lu et al., 2017). The removal of the solvent can also serve as a sample concentration step. By drying the sample, the resulting concentration could be adjusted to an extent depending on the amount of dilution that follows. Solvent removal is frequently performed for GC-MS analysis because as much water as possible needs to be removed from the samples so as to not interfere with the derivatization of the metabolites (see below). This process could also be used to adjust the solvent in a sample prior to analysis by LC-MS.

Solid-phase extraction

SPE is a process that includes both sample cleanup and sample concentration. These methods can remove unwanted urinary salts, matrix components, and proteins, allowing for lower abundance metabolites to be detected. SPE improves the signals of most metabolites in blood-based samples (de Paiva et al., 2014). A specific method of SPE known as SPME is an alternative technique that is useful in untargeted metabolomics (de Paiva et al., 2014). This technique is solvent-free and combines sample cleanup and concentration. The two modes of SPME are HS and direct immersion. HS-SPME is a useful alternative to derivatization for GC-MS analysis. In HS-SPME, the sample is heated, and volatile metabolites vaporize and are trapped on SPME fibers. The metabolites are then thermally released onto the GC column when the SPME fibers are introduced into the heated GC inlet. HS-SPME allows for quick processing and concentration of samples without solvents, reducing cost and environmental impact (Chetwynd et al., 2017). These methods are popular for cleaning up urine and other high salt concentration samples.

Ultrafiltration

Ultrafiltration is a technique that is useful in the study of polar metabolites, but it may hinder the extraction of hydrophobic metabolites (de Paiva et al., 2014). Ultrafiltration utilizes special filters that can physically separate small molecules (metabolites) from larger ones (proteins). It is a simple procedure that more efficiently precipitates proteins from a sample than the use of solvents (de Paiva et al., 2014). Ultrafiltration may also improve the stability of the filtered metabolite. This technique is recommended for studies utilizing NMR or LC-MS.

Controlling metabolite concentrations

Depending on analyte abundances and the chosen method of analysis, samples may need to be preprocessed in order to improve the detection of metabolites. Preconcentration of the metabolites will help increase the detection of low abundance metabolites or of dilute samples in general. Preconcentration can be

achieved by decreasing the volume of the solvent or completely removing the solvent and redissolving the metabolites in a smaller volume of new solvent. SPE is also an option for preconcentration.

Dilution is the opposite of preconcentration and is often used for samples that have a very high abundance of metabolites. While in many cases, it is best to have a high abundance of metabolites, there are some cases where too high of a concentration can cause issues during analysis. For example, dilution is very useful for urine samples because they have a high abundance of salts and ions that have ion suppression effects in LC-MS analysis (Chetwynd et al., 2017). Dilution will help reduce the ion suppression, but it should be noted that it may also lower the signal of low abundance peaks to below the LOD. Another case where dilution is necessary is where some of the metabolites in a sample are in such high abundance that the corresponding signals overlap with the signals of lower abundance metabolites. Dilution has the potential to reveal some less abundant metabolites that previously coeluted with high abundance metabolites as long as the concentration remains above the LOD (Chetwynd et al., 2017).

In some cases, only the concentration of one metabolite needs to be altered because of its incredibly high abundance. This would require a specific compound that would only target the metabolite that needs to be reduced. One possible option is the use of enzymes. As mentioned in the discussion on urine samples, high concentrations of urea can be eliminated through the use of the enzyme urease (Michell et al., 2008). It is not very likely that the removal of an individual metabolite may be necessary for samples other than urine, but it is something that should be considered.

Nuclear magnetic resonance

Metabolomic experiments utilizing NMR do not tend to require as much sample preparation as the MS-based techniques do. Generally, NMR sample preparation should maximize metabolite extraction and minimize the presence of proteins and lipids. As such, extraction of water-soluble metabolites and ultrafiltration are often used in the preparation of NMR samples (Snytnikova et al., 2019). The samples may also require pre-concentration steps, and it is unlikely that dilution would be necessary due to the low sensitivity of NMR. In comparison to MS-based methodologies, there are no unique sample preparation steps that are required for NMR.

Gas chromatography-mass spectrometry

Derivatization

Derivatization is perhaps the most important sample preparation step for GC-MS analysis since most metabolites are polar and heat-labile. The process of derivatization replaces polar (active) hydrogens on metabolites with a nonpolar

substituent to prevent H-bonding and therefore increase the volatility of the metabolites (Moldoveanu & David, 2019). The functional groups that are most commonly affected by derivatization are $-\text{OH}$, $-\text{COOH}$, $-\text{SH}$, $-\text{NH}_2$, and $-\text{CONH}$. Derivatization of metabolites should be optimized in each laboratory as the derivatization conditions can impact the LOD, sensitivity, and selectivity of the GC-MS (Danielsson et al., 2012). This process can also improve chromatographic separation and stability of the metabolites (Danielsson et al., 2012). Derivatization often occurs as a two-step process.

The first step of derivatization is methoximation which involves reaction with the compound methoxyamine hydrochloride, often dissolved in the aprotic solvent pyridine (Troisi et al., 2018). Methoxyamine reacts with ketones and α -ketoacids to serve as a protecting agent and prevent decarboxylation (Fig. 2.6) (Danielsson et al., 2012). The methoxyamine also reacts with sugar tautomers to cause ring opening and protects aldehydes and ketones (Fig. 2.7A). This reduces the number of sugar tautomers that can form during derivatization, lowering the number of chromatographic peaks which facilitates better separation and quantification of metabolites (Fiehn et al., 2000). This first derivatization step is typically shorter than the silylation step.

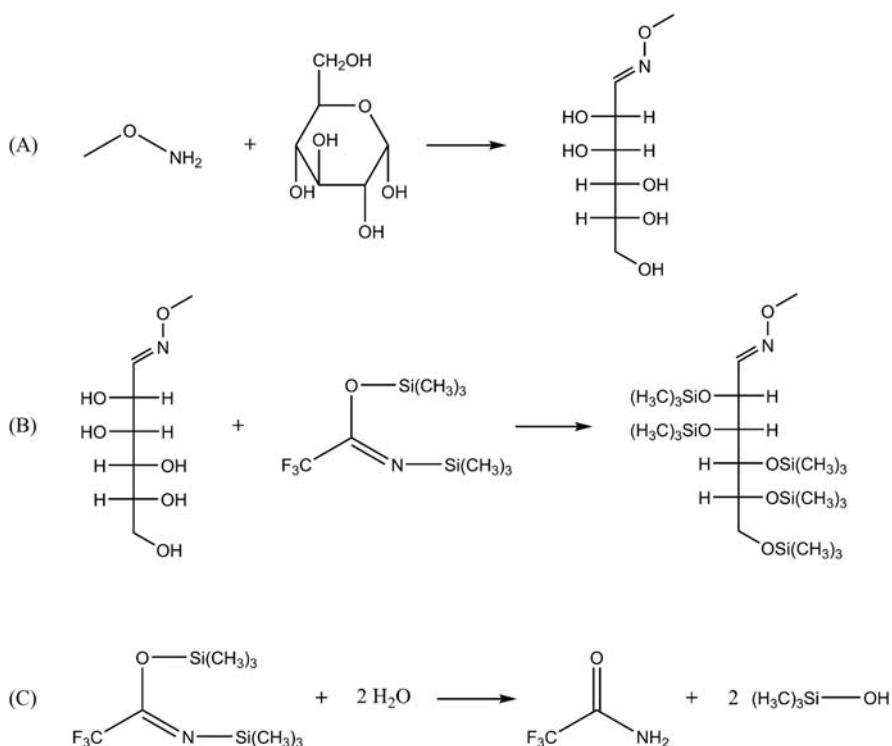
The second part of derivatization is trimethylsilylation derivatization (Fig. 2.7B). In this process, active hydrogens are replaced by trimethylsilyl (TMS) groups to produce TMS derivatives (Moldoveanu & David, 2019). The two most commonly used TMS reagents are *N*, *O*-bis(trimethylsilyl)-trifluoroacetamide (BSTFA) and *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA). Generally, BSTFA is a better silyl donor than MSTFA. Dryness, as well as solvent and sample purity, have the largest impact on the success of derivatization, whereas derivatization time and temperature can vary widely and should be optimized in each laboratory (Lu et al., 2017). Derivatization occurs best in a solvent with no active hydrogens, such as pyridine, dimethylformamide, and acetonitrile (Moldoveanu & David, 2019).

On the topic of dryness, the amount of water in the sample must be as minimal as possible prior to derivatization. Water interferes with the silylation process because water will react more quickly with the derivatization reagent than the metabolites. The derivatization of water forms trimethylsilanol (Fig. 2.7C), which separates into a separate solvent layer, impeding the derivatization of metabolites (Moldoveanu & David, 2019). This becomes important when considering that metabolite extraction and purification steps often use aqueous solutions, and water



FIGURE 2.6

Methoximation reaction on ketonic group of propanone.

**FIGURE 2.7**

Derivatization reactions. (A) Methoximation and (B) BSTFA trimethylsilylation of glucose. (C) Trimethylsilylation of water.

constitutes the majority of all biofluid sample types. Derivatization is severely hampered in samples with a water content of more than 10% (Moldoveanu & David, 2019). As a result, silylation-based derivatization methods must include a sample dry-down step, such as lyophilization or N_2 evaporation, prior to addition of derivatization reagents. In addition, great care must be taken to ensure that all derivatization reagents are as anhydrous as possible. For example, most suppliers offer anhydrous reagents packaged in special bottles that maintain reagent integrity. Alternatively, proper use of molecular sieves can be used in order to dry the reagents prior to use.

An alternative option to TMS derivatization is *N*-*tert*-butyldimethylsilyl-*N*-methyl-trifluoroacetamide (MTBSTFA). Metabolite derivatization with MTBSTFA yields *tert*-butyldimethylsilyl (TBDMS) derivatives, some of which may be more stable than their corresponding TMS derivatives (Danielsson et al., 2012). MTBSTFA is larger than BSTFA and MSTFA, which allows it to reduce the yield and volatility of glucose and large carbohydrates to the point that it

cannot be detected (Danielsson et al., 2012). MTBSTFA would, therefore not be suitable for studies that need to look at carbohydrates but may be useful if the sample is likely to have large carbohydrate signals that may mask other metabolites. Another advantage is that TMS derivatization often yields both mono- and di-silylated amines, while TBDMS derivatization is less likely to form multiple derivatives due to the bulkiness of the reagent (Danielsson et al., 2012).

Liquid chromatography-mass spectrometry

Unlike GC-MS, LC-MS does not require derivatization of metabolites as successful analysis does not depend on analyte volatility. In addition to quenching and extraction of metabolites from cells or biofluids, many of the sample cleanup techniques above (e.g., ultrafiltration or SPE) would be suitable for LC-MS analyses. Nonpolar metabolite extraction is most often used with LC-MS since LC-MS often utilizes a nonpolar stationary phase. Polar metabolites can also be analyzed with LC-MS using hydrophilic interaction LC. Solvent removal can be an important step in LC-MS depending on the other solvents used in sample preparation. The final sample diluent is important because it can impact the chromatography. For example, a sample diluent that is a stronger eluent than the LC initial conditions can lead to band broadening or distorted peak shapes. For further details regarding LC-MS analysis of metabolites, see Chapter 3, Separation Techniques, and Chapter 4, Mass Spectrometry in Metabolomics.

Identification and quantification of metabolites

Quantification

When quantifying metabolites, there are two methods: absolute quantification and relative quantification. Absolute quantification involves determining the exact concentration of a given metabolite in a sample. The two ways to determine absolute concentration are (1) comparison of signal intensity of the native metabolite of interest to that of a known amount of spiked-in, isotopically-labeled internal standard—assuming one is available or can be generated (Bennett et al., 2008), and assuming the response factor of the instrument has been determined, or (2) by comparison to a calibration curve using external standards. Calibration curves and/or instrument response factors are mandatory for quantification purposes because a given analytical instrument produces different signal intensities for a given concentration of different analytes. As a result, a ratio comparison of raw signal intensity between different metabolites in the same sample is not necessarily reflective of the ratio of metabolite abundances. This is ultimately rooted in how strongly a particular analyte generates a signal. In the case of MS-based detection, different analytes can have dramatically different ionization efficiencies so that an equimolar solution of two metabolites could produce signals that vary

by $>10^6$. A further compounding factor is that metabolites will be extracted and/or derivatized with different efficiencies, complicating efforts to determine the true concentration of a given analyte in a sample of interest. Appropriate method development steps (e.g., spike-recovery experiments, analysis of certified reference materials, etc.) can mitigate these issues. Absolute quantification is usually more suitable for targeted metabolomics experiments because it is easier to more precisely analyze a smaller number of metabolites ([Kapoore & Vaidyanathan, 2016](#)).

Relative quantification involves comparison of the instrument response of a metabolite relative to that of an internal standard. Such a ratio can then be compared to the analogous ratio determined in a separate sample so that a relative change between samples can be calculated. Assuming the signal responses are within the linear range of the instrument (which must be determined separately), and assuming that all samples have been processed and analyzed identically, then such relative signal changes can be equated with a fold-change in concentration between the samples. Relative quantification is more widely used for untargeted metabolomics ([Kapoore & Vaidyanathan, 2016](#)) and is especially helpful when trying to determine if there is a difference in metabolite abundance between two conditions, such as comparing a case versus a control ([Troisi et al., 2021](#)).

Calibration curve technique

The development of calibration curves (also called standard curves) for individual metabolites is a technique that allows absolute quantification of metabolite concentrations for any analyte for which a pure compound is available. To create the calibration curves, metabolite standards covering a range of known concentrations are analyzed. The gold standard is to use the standard addition method in which standards are spiked directly into the samples in order to compensate for matrix effects. Alternatively, standards can be spiked into a sample diluent that closely mimics the matrix in an effort to generate matrix-matched standards. The corresponding signals (often normalized to an internal standard) are plotted against the known concentration to derive the calibration curve for each metabolite standard. The signals from the unknown samples are then compared to the calibration curve in order to infer the unknown concentration. This technique is useful for quantification of NMR or MS-based signals and is applicable for any sample type.

Internal standard and isotope dilution

The use of internal standards to aid in the quantification of metabolites is common in metabolomics. For this type of quantification, the samples are usually spiked with a standard reference material that is an internal standard mix ([Kapoore & Vaidyanathan, 2016](#)). Another way of using internal standards is to utilize a single compound as an internal standard. The use of an individual internal standard that should ideally have the same concentration in each sample can be used to normalize the metabolite signals. By normalizing the signals to the

internal standard peak area, variation in the metabolome concentrations due to sample preparation or analysis can be accounted for.

Isotope dilution is a specific type of internal standard quantification. In isotope dilution, the sample is spiked with isotopically labeled compounds, preferably quantitative isotopic tracers. The isotopes that are commonly used are ^2H , ^{13}C , ^{15}N , and ^{18}O because metabolites labeled with these isotopes are chemically similar to the unlabeled metabolites (Kapoore & Vaidyanathan, 2016). These isotopes also compensate for ion suppression as they also match the ionization properties of the unlabeled metabolites. As a result, isotope dilution can reduce sample preparation and instrumental bias. The key factor to using isotope dilution is that labeled and unlabeled versions of a metabolite must have different masses to avoid isotopic interference. This technique is not usually a practical approach for the analysis of individual metabolites, but it can be useful for biochemical pathway analysis.

Identification

The identification of metabolites is a crucial yet challenging part of understanding the data gathered from metabolomic experiments. There are many methods and resources that can aid in the identification of metabolites though some may be specific to certain methods of analysis. Generally, a combination of methods will be utilized to identify metabolites from their corresponding signals. It should be noted that not every signal present in a spectrum or chromatogram will correspond to a unique metabolite (e.g., the same sugar may give several derivatization products that chromatograph at separate times) nor will every signal necessarily be identifiable. It is often the case that a chromatogram may show reproducible peaks that are demonstrably above the signal to noise ratio and yet the spectral quality of the corresponding mass spectrum may be too poor to be reliably identified. Importantly, such features can still be incorporated into multivariate statistical models that aim to determine if a group of metabolomes (e.g., cases) are statistically distinguishable from a separate group (e.g., controls). Such features often provide direction for future studies aimed at deciphering their true identities.

Public libraries and databases

Over the years of metabolomic experiments and the analysis of metabolites, many databases have been developed to aid in the identification of metabolites. Most of these databases can help to identify metabolites based on spectral data, such as NMR and MS spectra. Some of the global metabolite databases are Human Metabolome Database (HMDB), METLIN, MassBank, and the Golm Metabolome Database, all of which can be accessed via web search. These databases are tremendous resources for metabolomic researchers as they contain abundant information on the metabolites that have been detected from a wide variety of sample types analyzed with the myriad analytical technologies available in

laboratories worldwide. As an example, the HMDB (<http://www.hmdb.ca>) is a freely available database that contains >114,000 metabolite entries for metabolites spanning >6 orders of magnitude in concentration—from mM to <nM. Another database that can be used to help identify MS-analyzed metabolites is the NIST Mass Spectral Library (<http://www.nist.gov>). This database contains reference mass spectra for many hundreds of thousands of “small” molecules, a large fraction of which are known human or plant metabolites. The Library is purchased from a distributor and incorporates directly into existing LC-MS or GC-MS software allowing experimentally obtained spectra to be searched and compared against Library reference spectra. Potential matches are then ranked based on the statistical similarity between the compared spectra. Libraries are available for both EI generated mass spectra as well as tandem mass spectra.

Databases and spectral libraries are extremely useful for identifying compounds in general, but there are some things to consider. To compare two spectra, the spectra must be run under the same conditions. For example, there can be considerable uncertainty in the results of metabolite identification for NMR spectra as NMR shifts are sensitive to pH, osmolality, and ion concentrations (Lu et al., 2017). Different conditions can produce different results, so it is difficult to rely simply on databases. Similarly, mass spectra can be impacted by the type of ionization and the conditions of the MS. Overall, databases and libraries are certainly valuable tools in the identification of metabolites, but they should not be the only thing that should be relied on.

Metabolomics Standards Initiative

There are four levels of rigor with which metabolite identifications are recognized among the literature. A community effort called the Metabolomics Standards Initiative (MSI) has provided guidelines that define the various degrees of confidence with which chromatographic or spectral features are labeled (Sumner et al., 2007). In summary, MSI-Level 1 are fully identified compounds that have at least two independent measurements (e.g., retention time and mass spectrum) that match authentic standards; Level 2 are putative identifications that rely on spectral matching to a database, but lack comparison to an authentic standard; Level 3 are putative compound class identifications; and Level 4 are completely unknown compounds but their spectral or chromatographic features can still be statistically differentiated from the background.

Quality control

There are many safeguards that should be incorporated into a successful metabolomics campaign. Of prime importance is the requirement that all samples that are to be directly compared to one another must be collected, processed, measured, and statistically analyzed in exactly the same way. Resulting raw data

(e.g., NMR spectrum or chromatogram) should be statistically aligned before comparison. In high-throughput campaigns, where hundreds or thousands of samples are analyzed, on-going and proper instrument maintenance is mandatory (Fiehn, 2016). The concept of randomization should be practiced throughout every aspect of data collection. Samples, which include cases, controls, unknowns, standards, QC materials, pooled samples, certified reference materials, etc., should be quenched, extracted, processed, and measured in a random order. Don't load the autosampler and proceed to measure all of the controls and then all of the cases, for example. Plan the analytical campaign such that a restricted number of (random) samples are analyzed in batches. For example, a sequence might consist of 25 samples, a reagent blank, a method blank, QC sample(s), and a pooled sample. The next batch would then be run in a different order. Within the batch of 25 samples should be included at least one repeat injection of a randomly selected sample within that batch.

Using GC-MS as an example, an analytical batch might be considered validated if the following conditions are met: the reagent blank does not generate any chromatographic peaks; the peak areas of all analytes in the analytical standard (normalized by the internal standard area) are within 10% of the expected value; the standard deviation of peak area (normalized to the internal standard) for the 100 highest intensity peaks of the repeated injection are $\leq 15\%$ of the respective signals in the original injection; and, the pooled sample must cluster with all other pooled samples within 5% of the total area using a PLS-DA model built using all the samples analyzed.

Conclusion

Once the experiment has been designed and performed, the metabolomic data still need to be interpreted and validated. Interpretation of the results will involve various statistical analyses and could move further into the analysis of the metabolites in the context of biochemical pathways. Validation often requires a followup metabolomic experiment, or it could be paired with another -omic science to confirm the biological process. The design of the metabolomic experiment is just one part of the overall study because the data analysis can require significant time and energy to complete. There are many ways that one could interpret data depending on the focus of the study just as there are many ways to analyze the same metabolome. Care must be taken throughout the entire study to ensure the most accurate and detailed results.

Metabolomics is a field that continues to grow, and there is so much potential for what metabolomic experiments could accomplish. There is no one correct way to design a metabolomic experiment, and each experiment is going to be unique because of the focus, subject population, and available resources (i.e., funds and analytical instrumentation) of each experiment. The considerations for the design of metabolomic experiments presented here simply serve as a guide to help in the

development of new experiments. The hope is that the field will continue to improve and develop more techniques to aid in the discovery and analysis of metabolites.

References

- Álvarez-Sánchez, B., Priego-Capote, F., & de Castro, M. L. (2010a). Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. *TrAC Trends in Analytical Chemistry*, 29(2), 111–119.
- Álvarez-Sánchez, B., Priego-Capote, F., & de Castro, M. L. (2010b). Metabolomics analysis II. Preparation of biological samples prior to detection. *TrAC Trends in Analytical Chemistry*, 29(2), 120–127.
- Bahado-Singh, R. O., Akolekar, R., Chelliah, A., Mandal, R., Dong, E., Kruger, M., Wishart, D. S., & Nicolaides, K. (2013). Metabolomic analysis for first-trimester trisomy 18 detection. *American Journal of Obstetrics and Gynecology*, 209(1), 65, e1.
- Bardanzellu, F., & Fanos, V. (2019). The choice of amniotic fluid in metabolomics for the monitoring of fetus health-update. *Expert Review of Proteomics*, 16(6), 487–499.
- Beger, R. D. (2013). A review of applications of metabolomics in cancer. *Metabolites*, 3 (3), 552–574.
- Bennett, B. D., Yuan, J., Kimball, E. H., & Rabinowitz, J. D. (2008). Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. *Nature Protocols*, 3(8), 1299–1311.
- Blaise, B. J., Correia, G., Tin, A., Young, J. H., Vergnaud, A.-C., Lewis, M., Pearce, J. T., Elliott, P., Nicholson, J. K., & Holmes, E. (2016). Power analysis and sample size determination in metabolic phenotyping. *Analytical Chemistry*, 88(10), 5179–5188.
- Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., & Liu, P. (2013). The human urine metabolome. *PLoS One*, 8(9), e73076.
- Bouchard-Mercier, A., Rudkowska, I., Lemieux, S., Couture, P., & Vohl, M.-C. (2013). The metabolic signature associated with the Western dietary pattern: A cross-sectional study. *Nutrition Journal*, 12(1), 1–9.
- Canelas, A. B., ten Pierick, A., Ras, C., Seifar, R. M., van Dam, J. C., van Gulik, W. M., & Heijnen, J. J. (2009). Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics. *Analytical Chemistry*, 81(17), 7379–7389.
- Chen, L., Zhou, L., Chan, E. C., Neo, J., & Beuerman, R. W. (2011). Characterization of the human tear metabolome by LC–MS/MS. *Journal of Proteome Research*, 10(10), 4876–4882.
- Chetwynd, A. J., Dunn, W. B., & Rodriguez-Blanco, G. (2017). Collection and preparation of clinical samples for metabolomics. *Metabolomics: From Fundamentals to Clinical Applications*, 965, 19–44.
- Dame, Z. T., Aziat, F., Mandal, R., Krishnamurthy, R., Bouatra, S., Borzouie, S., Guo, A. C., Sajed, T., Deng, L., & Lin, H. (2015). The human saliva metabolome. *Metabolomics: Official Journal of the Metabolomic Society*, 11(6), 1864–1883.
- Danielsson, A. P., Moritz, T., Mulder, H., & Spégel, P. (2012). Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical

- experimental design. *Metabolomics: Official Journal of the Metabolomic Society*, 8(1), 50–63.
- de Paiva, M. J. N., Menezes, H. C., & de Lourdes Cardeal, Z. (2014). Sampling and analysis of metabolomes in biological fluids. *Analyst*, 139(15), 3683–3694.
- Duporet, X., Aggio, R. B. M., Carneiro, S., & Villas-Bôas, S. G. (2012). The biological interpretation of metabolomic data can be misled by the extraction method used. *Metabolomics: Official Journal of the Metabolomic Society*, 8(3), 410–421.
- Fiehn, O. (2016). Metabolomics by gas chromatography–mass spectrometry: Combined targeted and untargeted profiling. *Current Protocols in Molecular Biology*, 114(1), 30–34.
- Fiehn, O., Kopka, J., Trethewey, R. N., & Willmitzer, L. (2000). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry*, 72(15), 3573–3580.
- Holmes, E., Tsang, T. M., Huang, J. T.-J., Leweke, F. M., Koethe, D., Gerth, C. W., Nolden, B. M., Gross, S., Schreiber, D., & Nicholson, J. K. (2006). Metabolic profiling of CSF: Evidence that early intervention may impact on disease progression and outcome in schizophrenia. *PLoS Medicine*, 3(8), e327.
- Holmes, E., Wilson, I. D., & Nicholson, J. K. (2008). Metabolic phenotyping in health and disease. *Cell*, 134(5), 714–717.
- Hussain, J. N., Mantri, N., & Cohen, M. M. (2017). Working up a good sweat—The challenges of standardising sweat collection for metabolomics analysis. *The Clinical Biochemist Reviews*, 38(1), 13.
- Jiye, A., Trygg, J., Gullberg, J., Johansson, A. I., Jonsson, P., Antti, H., Marklund, S. L., & Moritz, T. (2005). Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry*, 77(24), 8086–8094.
- Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews. Molecular Cell Biology*, 17(7), 451–459.
- Kapoore, R. V., & Vaidyanathan, S. (2016). Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2079), 20150363.
- Kind, T., Tolstikov, V., Fiehn, O., & Weiss, R. H. (2007). A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical Biochemistry*, 363(2), 185–195.
- Lankadurai, B. P., Nagato, E. G., & Simpson, M. J. (2013). Environmental metabolomics: An emerging approach to study organism responses to environmental stressors. *Environmental Reviews*, 21(3), 180–205.
- Lorenz, M. A., Burant, C. F., & Kennedy, R. T. (2011). Reducing time and increasing sensitivity in sample preparation for adherent mammalian cell metabolomics. *Analytical Chemistry*, 83(9), 3406–3414.
- Lu, W., Su, X., Klein, M. S., Lewis, I. A., Fiehn, O., & Rabinowitz, J. D. (2017). Metabolite measurement: Pitfalls to avoid and practices to follow. *Annual Review of Biochemistry*, 86, 277–304.
- Madsen, R., Lundstedt, T., & Trygg, J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta*, 659(1–2), 23–33.
- Maharjan, R. P., & Ferenci, T. (2003). Global metabolite analysis: The influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Analytical Biochemistry*, 313(1), 145–154.

- Martano, G., Delmotte, N., Kiefer, P., Christen, P., Kentner, D., Bumann, D., & Vorholt, J. A. (2015). Fast sampling method for mammalian cell metabolic analyses using liquid chromatography–mass spectrometry. *Nature Protocols*, 10(1), 1–11.
- Mashego, M. R., Rumbold, K., De Mey, M., Vandamme, E., Soetaert, W., & Heijnen, J. J. (2007). Microbial metabolomics: Past, present and future methodologies. *Biotechnology Letters*, 29(1), 1–16.
- Menon, R., Jones, J., Gunst, P. R., Kacerovsky, M., Fortunato, S. J., Saade, G. R., & Basraon, S. (2014). Amniotic fluid metabolomic analysis in spontaneous preterm birth. *Reproductive Sciences*, 21(6), 791–803.
- Michell, A. W., Mosedale, D., Grainger, D. J., & Barker, R. A. (2008). Metabolomic analysis of urine and serum in Parkinson's disease. *Metabolomics: Official Journal of the Metabolomic Society*, 4(3), 191–201.
- Moldoveanu, S. C. & David, V. (2019). Derivatization methods in GC and GC/MS. In *Gas chromatography derivatization, sample preparation, application*.
- Mueller, D. C., Piller, M., Niessner, R., Scherer, M., & Scherer, G. (2014). Untargeted metabolomic profiling in saliva of smokers and nonsmokers by a validated GC-TOF-MS method. *Journal of Proteome Research*, 13(3), 1602–1613.
- Munger, J., Bennett, B. D., Parikh, A., Feng, X.-J., McArdle, J., Rabitz, H. A., Shenk, T., & Rabinowitz, J. D. (2008). Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. *Nature Biotechnology*, 26(10), 1179–1186.
- Nagana Gowda, G., & Raftery, D. (2017). Whole blood metabolomics by ¹H NMR spectroscopy provides a new opportunity to evaluate coenzymes and antioxidants. *Analytical Chemistry*, 89(8), 4620–4627.
- Patterson, R. E., Ducrocq, A. J., McDougall, D. J., Garrett, T. J., & Yost, R. A. (2015). Comparison of blood plasma sample preparation methods for combined LC–MS lipidomics and metabolomics. *Journal of Chromatography B*, 1002, 260–266.
- Pears, M. R., Salek, R. M., Palmer, D. N., Kay, G. W., Mortishire-Smith, R. J., & Griffin, J. L. (2007). Metabolomic investigation of CLN6 neuronal ceroid lipofuscinosis in affected South Hampshire sheep. *Journal of Neuroscience Research*, 85(15), 3494–3504.
- Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., & Gautam, B. (2011). The human serum metabolome. *PLoS One*, 6(2), e16957.
- Rabinowitz, J. D., & Kimball, E. (2007). Acidic acetonitrile for cellular metabolome extraction from *Escherichia coli*. *Analytical Chemistry*, 79(16), 6167–6173.
- Raffone, A., Troisi, J., Boccia, D., Travagliino, A., Capuano, G., Insabato, L., Mollo, A., Guida, M., & Zullo, F. (2020). Metabolomics in endometrial cancer diagnosis: A systematic review. *Acta Obstetricia et Gynecologica Scandinavica*, 99(9), 1135–1146.
- Shan, J., Xie, T., Xu, J., Zhou, H., & Zhao, X. (2019). Metabolomics of the amniotic fluid: Is it a feasible approach to evaluate the safety of Chinese medicine during pregnancy? *Journal of Applied Toxicology*, 39(1), 163–171.
- Smith, A. M., King, J. J., West, P. R., Ludwig, M. A., Donley, E. L., Burrier, R. E., & Amaral, D. G. (2019). Amino acid dysregulation metabotypes: Potential biomarkers for diagnosis and individualized treatment for subtypes of autism spectrum disorder. *Biological Psychiatry*, 85(4), 345–354.
- Snytnikova, O. A., Khlichkina, A. A., Sagdeev, R. Z., & Tsentalovich, Y. P. (2019). Evaluation of sample preparation protocols for quantitative NMR-based metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*, 15(6), 1–9.

- Stringer, K. A., Younger, J. G., McHugh, C., Yeomans, L., Finkel, M. A., Puskarich, M. A., Jones, A. E., Trexel, J., & Karnovsky, A. (2015). Whole blood reveals more metabolic detail of the human metabolome than serum as measured by ^1H -NMR spectroscopy: Implications for sepsis metabolomics. *Shock (Augusta, Ga.)*, 44(3), 200.
- Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., & Griffin, J. L. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics: Official Journal of the Metabolomic Society*, 3(3), 211–221.
- Sweatman, B. C., Farrant, R. D., Holmes, E., Ghauri, F. Y., Nicholson, J. K., & Lindon, J. C. (1993). 600 MHz ^1H -NMR spectroscopy of human cerebrospinal fluid: Effects of sample manipulation and assignment of resonances. *Journal of Pharmaceutical and Biomedical Analysis*, 11(8), 651–664.
- Ten-Doménech, I., Ramos-Garcia, V., Piñeiro-Ramos, J. D., Gormaz, M., Parra-Llorca, A., Vento, M., Kuligowski, J., & Quintás, G. (2020). Current practice in untargeted human milk metabolomics. *Metabolites*, 10(2), 43.
- Troisi, J., Cavallo, P., Richards, S., Symes, S., Colucci, A., Sarno, L., Landolfi, A., Scala, G., Adair, D., & Ciccone, C. (2021). Noninvasive screening for congenital heart defects using a serum metabolomics approach. *Prenatal Diagnosis*, 41, 743–753.
- Troisi, J., Cinque, C., Giugliano, L., Symes, S., Richards, S., Adair, D., Cavallo, P., Sarno, L., Scala, G., & Caiazza, M. (2019). Metabolomic change due to combined treatment with myo-inositol, D-chiro-inositol and glucomannan in polycystic ovarian syndrome patients: A pilot study. *Journal of Ovarian Research*, 12(1), 1–11.
- Troisi, J., Landolfi, A., Sarno, L., Richards, S., Symes, S., Adair, D., Ciccone, C., Scala, G., Martinelli, P., & Guida, M. (2018). A metabolomics-based approach for non-invasive screening of fetal central nervous system anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 14(6), 1–10.
- Troisi, J., Sarno, L., Landolfi, A., Scala, G., Martinelli, P., Venturella, R., Di Cello, A., Zullo, F., & Guida, M. (2018). Metabolomic signature of endometrial cancer. *Journal of Proteome Research*, 17(2), 804–812.
- Trutschel, D., Schmidt, S., Grosse, I., & Neumann, S. (2015). Experiment design beyond gut feeling: Statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics: Official Journal of the Metabolomic Society*, 11(4), 851–860.
- Van Gulik, W. M., Canelas, A. B., Taymaz-Nikerel, H., Douma, R. D., de Jonge, L. P., & Heijnen, J. J. (2012). *Fast sampling of the cellular metabolome*. *Microbial systems biology* (pp. 279–306). Springer.
- Vorkas, P. A., Isaac, G., Anwar, M. A., Davies, A. H., Want, E. J., Nicholson, J. K., & Holmes, E. (2015). Untargeted UPLC-MS profiling pipeline to expand tissue metabolome coverage: Application to cardiovascular disease. *Analytical Chemistry*, 87(8), 4184–4193.
- Vuckovic, D. (2012). Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Analytical and Bioanalytical Chemistry*, 403(6), 1523–1548.
- Want, E., & Masson, P. (2011). *Processing and analysis of GC/LC-MS-based metabolomics data. Metabolic profiling* (pp. 277–298). Springer.
- Ward, C., Nallamshetty, S., Wartrous, J.D., Acres, E., Long, T., Mathews, I.T., Sharma, S., Cheng, S., Imam, F., & Jain, M. (2021). *Nontargeted mass spectrometry of dried blood spots for interrogation of the human circulating metabolome*.

- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., & Karu, N. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617.
- Wishart, D. S., Lewis, M. J., Morrissey, J. A., Flegel, M. D., Jeroncic, K., Xiong, Y., Cheng, D., Eisner, R., Gautam, B., & Tzur, D. (2008). The human cerebrospinal fluid metabolome. *Journal of Chromatography B*, 871(2), 164–173.
- Wittmann, C., Krömer, J. O., Kiefer, P., Binz, T., & Heinzle, E. (2004). Impact of the cold shock phenomenon on quantification of intracellular metabolites in bacteria. *Analytical Biochemistry*, 327(1), 135–139.
- Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1), 92–107.
- Yazdani, M., Elgstøen, K. B. P., Rootwelt, H., Shahdadfar, A., Utheim, Ø. A., & Utheim, T. P. (2019). Tear metabolomics in dry eye disease: A review. *International Journal of Molecular Sciences*, 20(15), 3755.
- Yu, Z., Kastenmüller, G., He, Y., Belcredi, P., Möller, G., Prehn, C., Mendes, J., Wahl, S., Roemisch-Margl, W., & Ceglarek, U. (2011). Differences between human plasma and serum metabolite profiles. *PLoS One*, 6(7), e21230.

Further reading

- Troisi, J., Raffone, A., Travaglino, A., Belli, G., Belli, C., Anand, S., Giugliano, L., Cavallo, P., Scala, G., & Symes, S. (2020). Development and validation of a serum metabolomic signature for endometrial cancer screening in postmenopausal women. *JAMA Network Open*, 3(9), e2018327.

This page intentionally left blank

Separation techniques

3

Martina Catani, Simona Felletti, and Flavio Antonio Franchina

*Department of Chemical, Pharmaceutical, and Agricultural Sciences, University of Ferrara,
Ferrara, Italy*

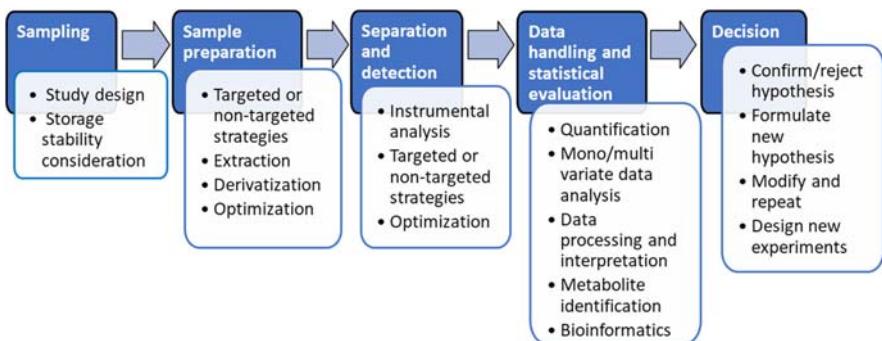
The role of the separation processes in metabolomics research

The analytical workflow for complex samples consists of several and subsequent steps, which typically include the sampling, the sample preparation, the separation/detection, the data handling, and statistical treatment, as shown in Fig. 3.1 (Pawlizyn & Lord, 2011).

The sampling step generally involves a study design to get the appropriate samples and size that properly define the object or problem being characterized. Also, careful consideration is needed for the storage condition, to assure the sample integrity at the time of the extraction/analysis. The goal of the sample preparation step is to modify sample properties to facilitate a successful separation and detection of the analytes of interest with the appropriate instrumentation. During the separation and detection step, the processed sample containing the analytes of interest is separated into its constituents (often employing a chromatographic or an electrodriven technique) which are subsequently identified and quantified.

The data handling and statistical evaluation can consist of the data integration from multiple techniques, followed by univariate or multivariate analysis to find differences in metabolites among the sample and can provide the absolute concentration of the target compounds. Univariate and multivariate represent two approaches for the statistical analysis, which are described in Chapter 9, Data Analysis in Metabolomics: From Information to Knowledge. In brief, univariate considers the analysis of a single variable at the time to describe the sample, for example, the concentration of a single analyte in it. On the other hand, the multivariate approach involves the simultaneous observation and analysis of more than one variable trying to find a relationship, for example, the correlation between the concentration of multiple analytes and a specific metabolic status.

From the statistical analysis, whether univariate or multivariate, initial hypotheses are tested or novel ones are generated, and appropriate decisions are taken, for example, for further investigation. In metabolomics, an important part of the

**FIGURE 3.1**

Typical steps for sample analysis in a metabolomics study.

data evaluation is also devoted to analyte identification (e.g., *via* mass spectrometry (MS) and retention time) and its correlation with the system biology ([Godzien et al., 2018](#)).

As emphasized in [Fig. 3.1](#), the analytical steps take place in a succeeding order, and errors introduced in any of the preceding steps will be propagating and cumulated to the next one, resulting in overall poor procedure performance.

It is important to optimize and tailor each of the steps with the application strategy and specific goals. Often in metabolomics, two general strategies can be identified, namely, targeted and nontargeted. Targeted analyses aim at the quantitative determination of a limited number of metabolites. On the other hand, nontargeted analyses provide a semiquantitative determination of a large number of metabolites (ideally all) and are often used in discovery studies. Nontargeted analyses can be intended as profiling, in which a large group of metabolites related to a specific metabolic pathway or a class of compounds is considered, or as fingerprinting, in which patterns of metabolites that change in a given biological system is sought for a rapid classification of the samples and extensive metabolite identification is generally not used.

A well-planned analytical protocol should include a proper design of experiment, optimized and validated sample preparation, separation/detection, and data processing methods. Moreover, careful quality controls during the whole analytical procedure are essential to monitor/correct potential bias (e.g., in the preparation of the sample) and limit the instrumental variability (see [Chapter 2: Experimental Design in Metabolomics](#)).

An ideal analytical system would permit sufficient coverage of the metabolome, allowing simultaneous analysis of a wide concentration range in a high-throughput and robust manner. Moreover, it should allow the reliable identification of unknown metabolites. However, in practice, no single methodology is sufficient for the analysis of the global metabolites profile and each methodology has its limitations. Currently, the methodologies used in metabolomics for final

separation and identification are based mainly on chromatographic methods, such as liquid chromatography (LC) and gas chromatography (GC) combined with MS detection, and direct shot-gun methodologies based on MS. The selection of a suitable technique depends on several factors, such as the strategy (i.e., targeted vs nontargeted), the type of analytes and matrix, the amount of sample available, and the concentration of the analytes.

Sample preparation

Sample preparation consists of one or a series of operations needed to modify the sample to deliver the analytes of interest in a suitable form to the chromatographic process or to improve the analysis results and/or performance. It is the most crucial step in the analytical procedure for implementation in any application. Moreover, it represents the limiting factor in chemical analysis since, on one hand, it is usually the most time consuming and, on the other hand, it can introduce errors at the first step of the analytical process which will be retained through the following steps (Moldoveanu, 2014; Pawliszyn & Lord, 2011).

The ideal sample preparation approach would involve the direct introduction of the intact sample into the analytical instrument; however, this is rarely possible and some sample pre-treatments are inevitably required. Hence, sample pre-treatment protocols must be developed and applied, avoiding sample composition alteration and impurities contamination during the handling, assuring the removal of interferences, and concentrating the analytes at detectable levels to be quantified precisely and accurately.

The common objectives of a generic sample preparation can be summarized as the reduction of sample size, pre-concentration, the simplification of the matrix or the enhancement in the release of the analytes of interests from the matrix, the cleanup and the removal of interferences, making the sample compatible with the following analytical technique or making it stable for longer-term storage.

Several procedures can be used to achieve the aforementioned objectives, such as mechanical (e.g., grinding, filtration, centrifugation, etc.) and phase-change separations (e.g., distillation, vaporization, precipitation from solutions), dilutions, headspace sampling, solvent and sorbent extractions, derivatizations, and preliminary chromatographic separation before the core chromatographic analysis. All the procedures exist in a multitude of versions, thanks to the continuous advancements in miniaturization, automation, and configurations. They are considered as steps that introduce a degree of selectivity towards the analyte(s) of interest.

Depending on the application, the analytes of interest can be one, a set of compounds, or an entire chemical class(es) of compounds, defining the targeted or nontargeted strategy. Different approaches are needed for target-compound analysis and global profiling: in targeted analyses, the defined number of analytes of interest makes simpler a highly selective sample pre-treatment and analysis, leading to sensitive and precise quantification of metabolite concentrations. As

introduced in Section The role of the separation processes in metabolomics research, nontargeted strategies (e.g., fingerprinting or profiling) are used in discovery research and tend to use wide-selectivity analytical steps to ensure adequate analytes coverage, characterize in-depth the samples, track and/or compare a large number of analytes between groups of samples to address specific research questions. In metabolite profiling or fingerprinting, it is difficult to find the optimal conditions for all types of analytes, thus some compromises must be made in the selection of the final experimental conditions. Additionally, to exhaustively cover and analyze such a broad range of analytes, analytical multiplatform strategies are ideally used (Haggarty & Burgess, 2017). Then, once the analytes of interest are identified and characterized (e.g., biomarkers of a specific disease), highly-selective targeted methods can be developed and applied, allowing faster and simpler workflows and instrumentation.

Nevertheless, the sample preparation procedure must ensure the preservation of the information from the original sample. Besides the targeted or nontargeted approach, the careful selection and optimization of these procedures depend mainly on the physicochemical characteristics of the analytes of interest, sample matrix, and the available instrumentation.

Since most of the metabolomics studies considering biological samples (e.g., biological fluids, tissues, cells, etc.) are designed to take a reliable snapshot of the metabolome, the metabolic flux should be stopped or inhibited. To do so, an additional step, called “quenching”, is introduced before the preparation of the sample. The quenching is important for the deactivation of the remaining metabolic activity (i.e., enzymatic processes, protein degradation) that otherwise would alter the levels of some metabolites during storage and preparation, giving a distorted snapshot of the sample metabolic status. To inhibit the metabolic processes, several procedures exist and can involve freeze-drying, heating, acid treatment, or the addition of organic solvent, enzyme inhibitors, or antioxidants (Hyötyläinen, 2013; Winder & Dunn, 2011).

Sample extraction techniques

Different techniques have been developed based on the type of analytes and sample matrices (Mushtaq et al., 2014). Depending on the condition used, each extraction procedure differs in speed, selectivity, ease of use, and automation possibility. The most suitable method and conditions should be carefully selected and optimized, keeping in mind the objective and the sample matrix (e.g., high/low-fat content, liquid/solid sample). In all techniques, the extraction phase is in contact with the sample matrix and analytes are transported between the phases. Based on the type of configuration between the extraction phase and the sample, the most common extraction techniques used in metabolomics can be classified as static or dynamic. In static extractions, fixed volumes of extraction phase are used, whereas in dynamic extractions the extraction phase flows continuously through the sample.

Examples of extraction techniques that use static (or in batch) and dynamic (or flow through) principles are discussed below and showed in Fig. 3.2.

Usually, the extraction process can be either solvent- or sorbent-based, depending on whether the analytes of interest are partitioned into an extraction solvent or a sorptive material. Extractions can also be exhaustive or non-exhaustive, depending on whether the depletion of the analyte of interest from the original sample matrix is complete or not. Other ways to identify/characterize the type of extraction technique are based on the nature of the extractant phase, liquid or solid, that defines liquid-phase and solid-phase extractions, respectively.

In liquid-phase extractions (LPEs), the compounds of interest are transferred from one phase, the sample or sample-containing phase, to another liquid phase where further processing and/or analysis can occur (Poole, 2020a). The classical example of LPE is the conventional solvent liquid-liquid extraction (LLE) in which the analytes are partitioned between two immiscible liquids of fixed volumes (static extraction) in a separation funnel (Fig. 3.2A).

Moreover, these solvent-based extraction methodologies can be assisted by additional physical mechanisms to ensure analytes stability and high recovery rates: microwaves assisted extraction, ultrasounds assisted extraction, and

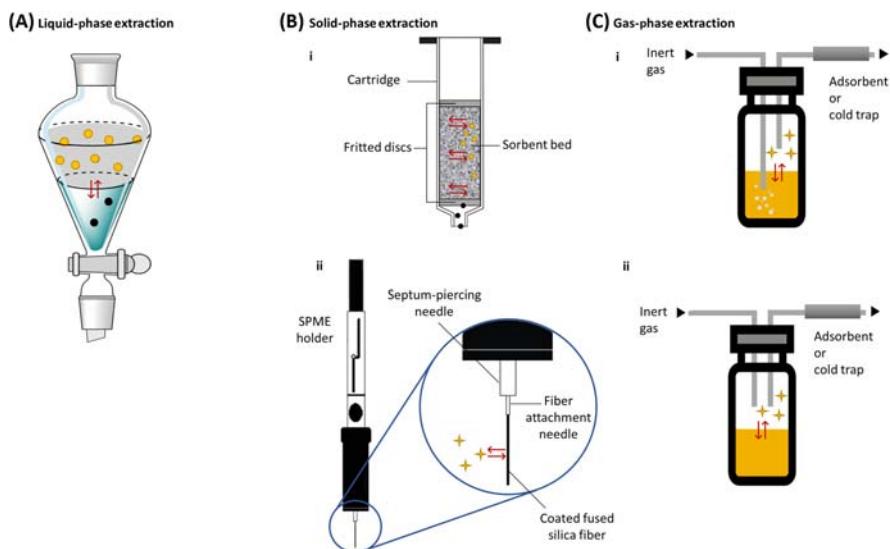


FIGURE 3.2

(A) Illustration of the classical liquid-liquid extraction (a static extraction), as an example of liquid-phase extractions; (B) Illustration of the SPE cartridge (a dynamic extraction) (i), and SPME (a static extraction) (ii), as examples of solid-phase extractions; (C) Illustration of the dynamic headspace sampling (i), and purge and trap sampling (ii), as examples of gas-phase extractions.

high-pressure liquid extraction (Lopez-Avila & Luque de Castro, 2014; Luque de Castro & Delgado-Povedano, 2014; Richter et al., 1996). If the extractant fluid is maintained at its supercritical phase, a supercritical fluid extraction can be obtained (Abbas et al., 2008). Indeed, gases such as CO₂ and N₂O become supercritical fluids when kept at their critical point and they can be used for extraction and purification purposes. The supercritical fluid state holds intermediate characteristics between gas and liquid, in particular the density and viscosity, which can give more convenient and faster extractions.

In solid-phase extractions (SPEs), the analyte extraction from the liquid sample occurs through its transfer to a solid sorbent (sorbent-based extractions) (Poole, 2020b). Typically, the sample containing the compounds of interest flows over the solid sorbent which retains the compounds based on their favorable interactions. Two interaction types can be distinguished: adsorption mechanisms (in most cases) when the analytes interact on its external surface, and absorption mechanisms when a noncrystalline polymer sorbent is used and the analytes partition into the phase. Subsequently, the target compounds are recovered from the sorbent by solvent displacement or thermal desorption. Different sorbent types with different physicochemical characteristics are available, offering a various degree of selectivity which can be tuned based on the analytes of interest. The typical solid adsorbents used in SPE can be classified based on the nature of the material type (i.e., inorganics, carbons, porous polymers), and based on the type of interaction (i.e., polar, nonpolar, ionic or mixed, immunosorbents, molecularly imprinted polymers, restricted-access media) (Diehl, 2007). Besides the continuous development in novel materials and selectivity, a main advantage of the adsorbent materials is the extended surface area for more efficient extractions.

Among SPEs, some techniques can be used both in the headspace, in direct contact with the sample (e.g., immersion in a liquid), or through a membrane protection approach. The SPE cartridge and solid-phase microextraction (SPME) are surely among the most promising and consolidated SPEs techniques. The SPE cartridge represents a modern and dynamic alternative to LLE and it is based on the principle that the components of interest are retained on a specific sorbent, packed into a disposable minicolumn (cartridge) (Fig. 3.2B_i). The further miniaturization and evolution towards solventless/green techniques brought to the development of SPME (Vuckovic et al., 2012). Its most used form consists of a fused silica fiber coated with the extraction phase(s) (Fig. 3.2B_{ii}). It performs an extraction in static mode, with the fiber exposed to the headspace of the sample or in direct contact with it. SPME is a non-exhaustive technique since only a negligible portion of the analytes is extracted. Thus it holds a lot of potential for direct *in vivo* sampling to capture a reliable metabolic snapshot, avoiding the risk of metabolic alteration with the sample handling (Risticevic et al., 2020).

Also gas-phase extractions (GPEs) exist and specifically involve the use of a gaseous phase to obtain the analytes from the sample matrix: in this case, the extraction is directed to the more volatile analytes and is often followed by a trapping step into a sorbent tube or trap (Franchina, Zanella, et al., 2020). Indeed, most of the GPE techniques rely on subsequent sorbent-based

interactions with the analytes to be isolated, as in SPEs. Among the GPEs, the purge and trap (P&T) and the dynamic headspace sampling (DHS) are the most representative. In these dynamic headspace techniques, an inert gas is used to continuously strip the volatile analytes off the sample, which are directed to and enriched into a sorbent tube (Liberto et al., 2020). In the most classical P&T technique, the gas bubbles into the liquid sample or liquid media containing the sample (Fig. 3.2C_i); in DHS instead, the headspace of the sample is purged with the gas and collected (Fig. 3.2C_{ii}). As discussed, GPEs are naturally highly-suitable for volatile compounds, which are thermally desorbed and transferred into a GC or an ionization source for direct MS analysis. On the other hand, SPEs and LPEs are suitable for the nonvolatile compounds, and either LC and GC separation, or direct MS can be exploited, depending on the analytes of interest.

Derivatization

The derivatization is an important and common procedure used for the preparation of the samples. It involves a chemical modification of the analytes, the matrix, or the whole sample and it can be performed at different stages (i.e., pre/post-extraction, pre/post-separation), depending on the targeted analyte (s) and the objective. It can be used for different purposes, such as for solubility improvement, for clean-up or fractionation purposes, and for separation or detection enhancement. The derivatization strongly affects the nonpolar/polar property of the analyte, which modifies the interactions during the subsequent analytical separation. For liquid chromatographic or capillary electrophoretic analysis (see Sections Liquid chromatography and Other separation techniques), modification of analyte polarity is usually aimed at improving the separation or, in the case of UV detection (see Section Detectors in Liquid chromatography), the addition of chromophores is intended to improve the overall detection. In this last case, the detector response for the derivatized form is enhanced thanks to the higher UV absorbance provided by the chromophore. For GC analysis (see Section Gas chromatography), the derivatization process usually aims at increasing the volatility and reducing the polarity of some chemical classes, or enhancing the thermal stability of the analytes to avoid their decomposition in the injection port. A common derivatization method in biological samples is the transformation of fatty acids into their silyl or ester derivatives, which make them suitable for GC analysis. For a complete description of the derivatization processes, the reader can refer to dedicated literature (Moldoveanu & David, 2015).

Fundamentals of chromatography

Chromatography represents the most widely used analytical tool for the separation of samples into their constituents. Although MS allows detecting more

components than can actually be separated by any pre-fractionation step, the precision and robustness of the identification, and especially the quantification of the analytes, remain problematic for complex samples. Compounds reaching the detector at the same time (or coeluted) influence each other's intensity signals, altering the ionization process and promoting either suppression or enhancement of the signal in a way not predictable. Therefore a proper chromatographic separation enables reliable and precise measurements. In metabolomics, chromatography is undoubtedly the most used technique for the physical separation of the analytes, and this section is devoted to its general principles and definitions.

Definitions and classifications

The term “chromatography” has been coined by the Russian botanic Mikhail S. Tswett at the beginning of the 20th century. The term is composed of two Greek words, *khroma* (“color”) and *graphein* (“to write”). Therefore chromatography means “color writing.” Tswett developed this separation method with the goal of isolating pigments in plants such as chlorophylls, carotenes, and xanthophylls. He employed a glass column filled with calcium carbonate as adsorbent through which the plant pigments mixture was percolated by using a petrol ether/ethanol mixture as eluent. During the elution, the initial mixture was separated into many different colored bands migrating with different velocities throughout the column (Ettré, 1975). At the outlet, he placed different flasks where the different separated components were collected (Fig. 3.3).

The separation mechanism of all chromatographic techniques is based on the distribution of the sample components between two phases, one of which usually in

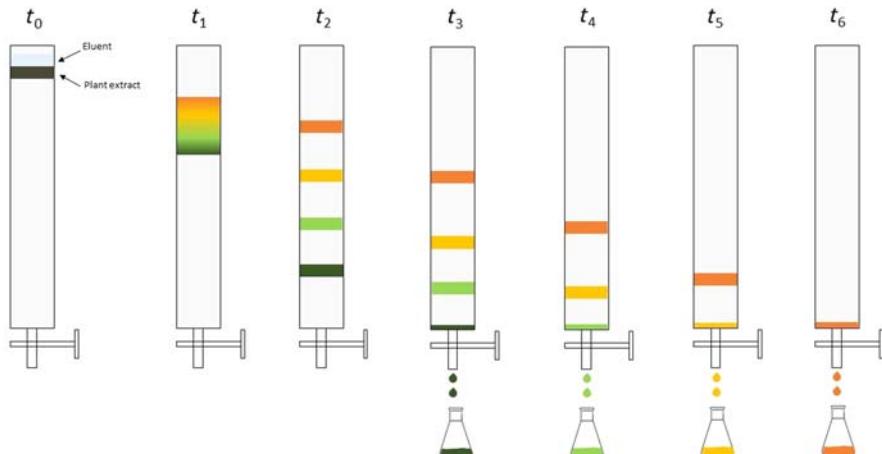


FIGURE 3.3

Schematic representation of Tswett's experiment.

the form of a porous, bulk liquid or thick film fixed in the separation bed, and it is called the stationary phase, and the other one being a fluid that percolates through and around the stationary phase, defined as mobile phase. Chromatographic methods can be classified in different ways. The most important distinction is between planar chromatography and column chromatography. In the first case, the separation process takes place on an inert plane, such as glass (thin layer chromatography, TLC) or a strip of paper on which the stationary phase is supported. In the second case, the stationary phase is placed inside a narrow tube and the mobile phase is introduced inside the system under pressure or by gravity. In this chapter, only column chromatography will be considered. The other most common classification is made on the phase of the mobile phase used, which can be either a liquid or a gas, characterizing LC and GC, respectively. As mobile phase, also supercritical fluids (such as pressurized CO₂) can be used, defining supercritical fluid chromatography (SFC), which holds intermediate properties between LC and GC. Another distinction regards the type of support for the stationary phase. If the adsorbent is a packed porous bed inside the whole internal volume of the column, the technique is called packed-column chromatography; on the other hand, if the stationary phase is located on the internal wall of the column, the technique is defined as open-tubular (OT) or capillary chromatography. Table 3.1 reports the most common classifications of chromatographic methods.

Table 3.1 Classification of column chromatographic methods.

General classification	Type of method	Type of stationary phase	Principle of separation
Gas chromatography	Gas-solid (GSC)	Underderivatized solid support	Adsorption
	Gas-liquid (GLC)	Liquid layer on a solid support or column wall	Partition between gas and liquid
Liquid chromatography	Adsorption (liquid-solid)	Solid support	Adsorption
	Partition (liquid-liquid)	Liquid layer adsorbed on a solid support or derivatized solid support	Partition between two immiscible liquid
	Ion-exchange	Charged solid support	Electrostatic interactions
	Size exclusion	Inert porous support	Sieving (separation according to size)
	Affinity	Solid support with immobilized biologic ligand	Highly specific binding interactions (mainly biological)
		Solid support (derivatized or underderivatized)	Partition between supercritical fluid and a liquid

Retention

Chromatographic separations take place according to the different migration of solutes from the inlet to the outlet of the column. The process by which solutes are washed through the stationary phase through the movement of the mobile phase (eluent) is called elution. The mobile phase-solutes mixture exiting the column (eluate) will then enter a detector, usually placed at the outlet of the column (as it will be described in Section Detectors in Liquid chromatography and Gas chromatography). The detector selectively responds to specific chemo-physical properties of the solutes; therefore the outcome of separation can be visualized by constructing a graph showing the detector signal versus time. A schematic representation of a chromatogram is reported in Fig. 3.4.

Each chromatographic separation is characterized by a specific time (that is different for each column) called dead or void time (indicated by t_M) corresponding to the time necessary for the mobile phase to pass through the column, from the inlet to the outlet. It can be practically measured by injecting an unretained molecule. All components in the chromatographic system spend time t_M in the

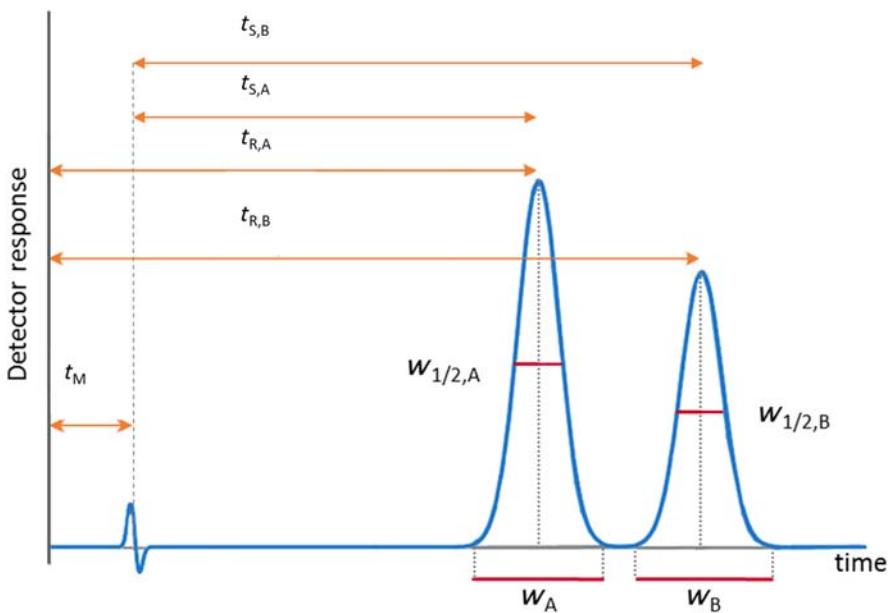


FIGURE 3.4

Schematic representation of a chromatogram. The peaks can be described by a Gaussian function with standard deviation σ . The symbol w ($=4\sigma$) represents the width at the base of the peak while $w_{1/2}$ ($=2.35\sigma$) is the peak width at half-height.

mobile phase, but retained ones also spend time in the stationary phase, indicated with t_S . Therefore the total retention time (t_R) represents the time the analyte is retarded by the column, that is the time effectively spent in the column and can be defined as:

$$t_R = t_S + t_M \quad (3.1)$$

Since in chromatography different volumetric flow rates (F_v) can be used, it is usually convenient using the retention volume V_R , instead of t_R :

$$V_R = t_R \times F_v \quad (3.2)$$

However, every column is characterized by its own void time. Therefore in order to compare retention on different columns, the most important parameter that can be used is the adimensional retention factor, k_A :

$$k_A = \frac{t_R - t_M}{t_M} = \frac{t_S}{t_M} \quad (3.3)$$

Retention is strongly affected by thermodynamic equilibria. Indeed, each solute is characterized by its migration velocity, which depends on its distribution extent between mobile and stationary phases. Let us consider a solute A migrating along the column. The local equilibrium describing its distribution between the mobile and the stationary phase can be written as:



where M and S represent the mobile and stationary phase, respectively. This equilibrium is regulated by a distribution constant, K_A , defined as:

$$K_A = \frac{[A]_S}{[A]_M} \quad (3.5)$$

where $[A]$ is the concentration of the solute. According to Eq. (3.5), the higher the concentration of the analyte in the stationary phase, the higher the distribution constant. In other words, the stronger the interactions of the analyte with the stationary phase, the higher the distribution constant.

Eq. (3.3) can be also expressed as:

$$k_A = \frac{(mol \text{ } A)_S}{(mol \text{ } A)_M} \quad (3.6)$$

By expressing now moles as the product between concentration and volume, Eq. (3.6) can be rearranged into:

$$k_A = \frac{[A]_S V_S}{[A]_M V_M} = K_A \frac{V_S}{V_M} \quad (3.7)$$

being (V_S/V_M) the phase ratio (often indicated with β), that is, the ratio between the volumes of stationary phase and mobile phase, respectively.

Selectivity

[Eq. \(3.7\)](#) indicates that two species can be separated as long as they are characterized by different distribution constants. The higher the difference, the better the separation. The parameter that allows defining the extent of separation is called selectivity (α) and it can be calculated from the ratio between the retention factors of two species. By considering the two peaks represented in [Fig. 3.4](#) relative to the separation of two species *A* and *B*, where *B* is more retained, the selectivity is defined as:

$$\alpha = \frac{k_B}{k_A} \quad (3.8)$$

Its minimum value is 1, indicating two overlapped peaks. Therefore a separation occurs when the selectivity factor is higher than 1.

Efficiency of separation

The efficiency of a separation is basically the ability of the system of eluting the chromatographic bands as Gaussian narrower peaks. The narrower and more symmetrical the peak, the higher the efficiency. It is typically quantified in terms of plate height, H , or number of theoretical plates, N . These concepts have been elaborated by A.J.P. Martin and R.L.M. Synge (awarded with Nobel Prize in 1952) who adapted a model usually employed to describe fractional distillation columns to chromatography. In practice, they considered the chromatographic column as a sort of fractionating column composed of a series of bubble-cap like plates where equilibrium conditions prevail. However, it must be taken in mind that this is just a model, and it does not describe what is happening inside the column from a physical point of view. N and H are inversely proportional and they are related as in the following equation:

$$N = \frac{L}{H} \quad (3.9)$$

being L the column length. Efficiency increases when N becomes greater and H becomes smaller. Typical N values can vary between a few hundred to many hundreds of thousands. H and N can be practically calculated from the chromatogram. H can be calculated as follows:

$$H = \frac{\sigma^2}{L} \quad (3.10)$$

where σ^2 is the peak variance, and σ is the standard deviation of the Gaussian curve. On the other hand, two equivalent formulas can be used for N :

$$N = 16 \left(\frac{t_R}{w} \right)^2 = 5.54 \left(\frac{t_R}{w_{1/2}} \right)^2 \quad (3.11)$$

being $w (= 4\sigma)$ and $w_{1/2} (= 2.35\sigma)$ the width at the base of the peak and its width at the half of the height, respectively (see [Fig. 3.4](#)).

An efficient separation is characterized by narrow Gaussian peaks. However, peaks get unavoidably broader during their migration along the column and the extent of band broadening is higher for those highly retained compounds that spend a considerable amount of time inside the column. The efficiency of a column is usually evaluated through the van Deemter equation (Giddings, 1964):

$$H = A + \frac{B}{u} + Cu \quad (3.12)$$

where u is the linear velocity of the mobile phase, expressed as:

$$u = \frac{L}{t_M} \quad (3.13)$$

The van Deemter equation contains three terms that account for the main sources of band broadening during the migration of analytes inside the column. The first term A is called eddy-diffusion and it is related to the multiple flow paths of unequal length that should exist inside a packed bed (Knox, 1999). This term is usually negligible when open tubular columns are used, such as for most GC separations. The second term, B/u , describes the natural diffusion of molecules from a more concentrated region to a more dilute one. The term B is called longitudinal diffusion coefficient and strictly depends on the diffusion coefficients of the analytes in mobile and stationary phases. This term is maximum in absence of flux while it diminishes by increasing the mobile phase velocity. The last term Cu , called solid-liquid mass transfer resistance, is then related to the finite time required for an analyte to reach the equilibrium between the stationary and the mobile phase. It is directly proportional to the mobile phase velocity. The sum of these three terms generates the well-known van Deemter curve, schematically represented in Fig. 3.5.

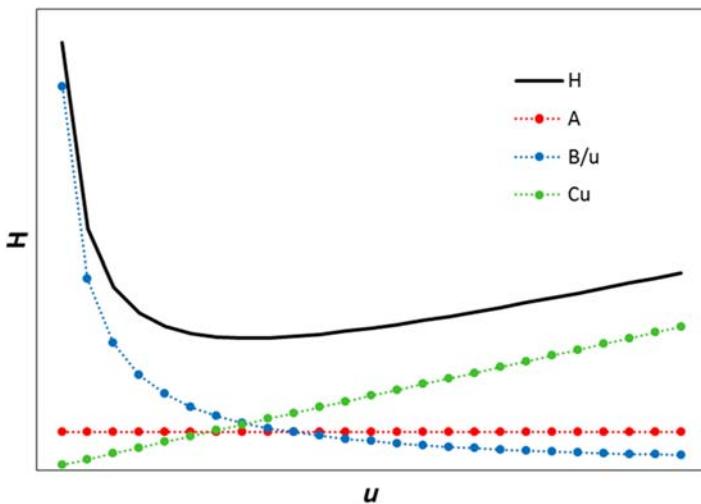
As it can be observed, this curve has a minimum that represents the optimal conditions at which the column can give the best kinetic performance (minimum value of H). A flat curve at high flow velocities indicates that the flow rate can be increased over the optimal one without considerable loss of efficiency (very important characteristics if fast or ultrafast separations need to be performed) (Ismail et al., 2016).

Resolution

Resolution is a very useful parameter to describe the ability of a column to separate two components. It can be directly calculated by a chromatogram according to the following relation:

$$R_S = \frac{2[t_{R,B} - t_{R,A}]}{w_A + w_B} \quad (3.14)$$

where A and B refer to the peaks reported in Fig. 3.4. It is generally acknowledged that a resolution of 1.5 is indicative of a complete baseline separation.

**FIGURE 3.5**

Example of van Deemter curve (black solid line) and each contribution to band broadening. Blue: longitudinal diffusion (B/u); red: eddy diffusion (A); green: solid-liquid mass transfer resistance (Cu).

Resolution is affected by both thermodynamic and kinetic factors. This is accounted for in the Purnell equation, which can be alternatively used to calculate resolution:

$$R_S = \frac{\sqrt{N}}{4} \left(\frac{\alpha - 1}{\alpha} \right) \left(\frac{k_B}{1 + k_B} \right) \quad (3.15)$$

Peak capacity

The peak capacity of a column is a useful parameter that defines the maximum number of peaks that can be separated at a given resolution value (usually 1) in a given time interval (Calvin Giddings, 1967). It can be calculated as follows:

$$n_C = \frac{t_{R,n} - t_{R,1}}{w} \quad (3.16)$$

where $t_{R,n}$ and $t_{R,1}$ are the retention times of the last eluting compound and the first one, respectively, while w is taken as the average peak width of a series of peaks of the chromatogram.

Qualitative and quantitative analysis in chromatography

Chromatography enables both qualitative (detection and identification) and quantitative analyses. The chromatographic parameter which carries qualitative

information is the retention time because it is correlated to the properties, thus the identity, of the solute. However, the sole use of the retention parameters cannot univocally confirm peak identity, and selective detectors or the use of standards are often used to help with the identification of classes of compounds. However, the use of a MS is nowadays the most informative and powerful solution for qualitative analysis. Relative retention times are more reproducible than total retention times, and some methods exist (for example, the retention index system) that exploit the relative retention concept to increase confidence in analyte identification (Section Column, stationary phases, and separation). The detector signals are proportional to the quantity of each solute (analyte) and the area under the peak or the peak height can be used to obtain quantitative information. However, peak area integration is generally the preferred and most used. The common approaches to carry out a quantitative analysis are the area normalization, and the external or internal standard addition. As the name suggests, area normalization is the calculation of the area percent that is assumed to be equal to weight percent (Fig. 3.6A). This method is simple and is often useful if a semiquantitative (relative quantification) analysis is sufficient. However,

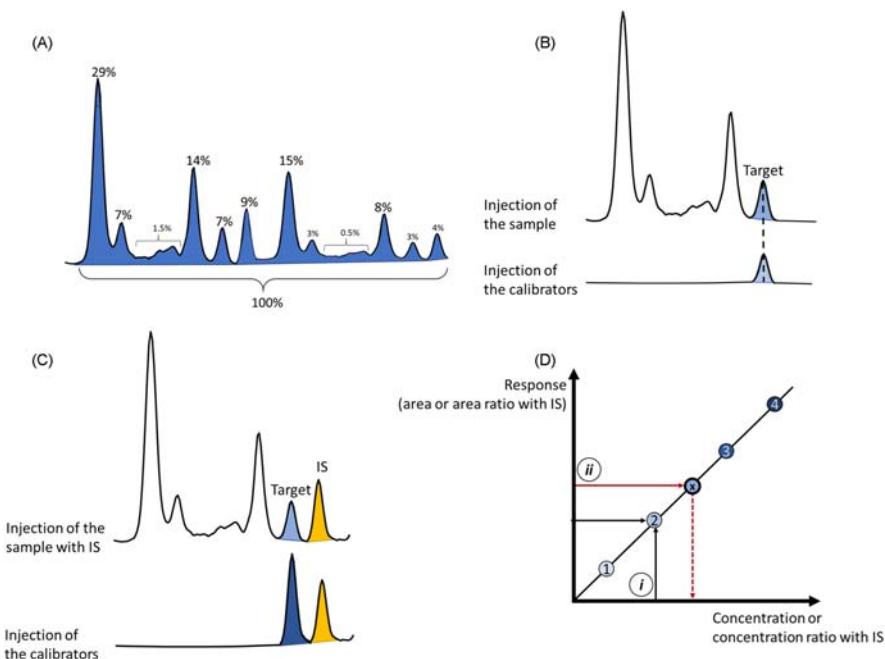


FIGURE 3.6

Graphical illustrations of the area normalization (A), the external and internal standard (B)–(C) for quantitative analysis, and a representative calibration curve (D).

this method to be accurate requires that all the analytes are detected and their detector response is uniform.

In the external standard method (Fig. 3.6B), a pure reference standard (ideally the same target compound to be quantified) is employed. Solutions at different concentrations of the target analyte (called calibrators, and indicated with numbered dots in Fig. 3.6D) are injected to obtain a calibration curve (Fig. 3.6D). The known concentration is plotted on the x -axis with the corresponding area on the y -axis, as illustrated by the black arrows in step *i* of Fig. 3.6D. Then, the sample containing the analyte to be quantified is analyzed under the same conditions and its response (the area) is measured, which allows the extrapolation of the concentration from the calibration curve (illustrated by the red arrows in step *ii* of Fig. 3.6D). In the internal standard methodology (Fig. 3.6C), a fixed amount of a given compound (different from the analyte to be quantified), called internal standard (IS), is added to the sample and to each calibrator. The calibration curve is constructed by plotting the ratio between the concentration of the calibrators and the IS, in which the concentration ratio is plotted on the x -axis and the area ratio is plotted on the y -axis (step *i* of Fig. 3.6D). When the sample is analyzed, the area of the target analyte and internal standard is measured and the ratio is determined; this allows to determine the concentration ratio with the IS (step *ii* of Fig. 3.6D) and finally the concentration of the target analyte. Qualitative and especially quantitative analysis are mostly influenced by the characteristics of the detector used which can provide an additional degree of selectivity and sensitivity to the method (Morris, 1974).

Liquid chromatography

Instrumentation

The last 50 years have seen the birth of the modern LC, under the name of high-performance LC (HPLC). Conversely to Tswett experiment (see Section Definitions and classifications), this separation technique makes use of pumps (up to 400 bar) able to mechanically pump the liquid mobile phase through the column packed with porous particles with a diameter of 3–5 μm . The continuous need for higher efficiency of separation and faster analysis time has led to the reduction of the particle diameter. The use of smaller and smaller particles turned out to be a huge limitation for conventional HPLC instruments, being unable to withstand the too high back pressure generated by the column. Only in the early 2000s, ultrahigh-pressure LC (UHPLC) instrumentations were introduced in the market. UHPLC can reach a maximum operating pressure up to 1200–1500 bar using the last generation stationary phases prepared either on sub-2 μm totally porous particles or sub-3 μm superficially porous particles (Fig. 3.7). The main advantages of UHPLC over conventional HPLC consist of faster analysis, lower solvent consumption, and higher efficiency of separation.

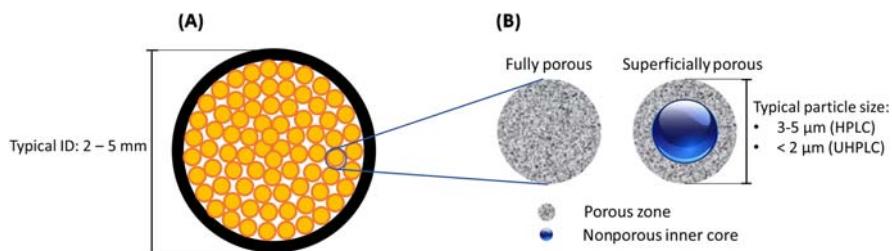


FIGURE 3.7

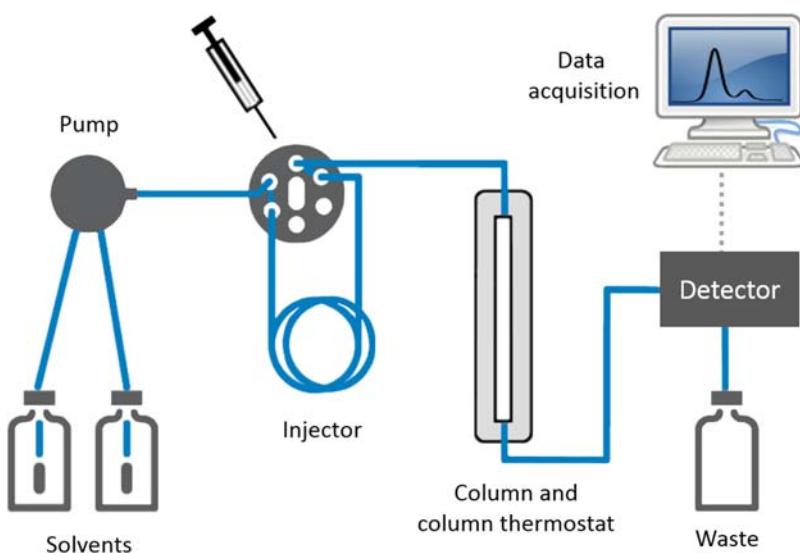
Graphical representation of the longitudinal section of a packed column (A), and the section of a fully porous and a superficially porous particle (B) designed for liquid chromatography.

HPLC or UHPLC systems consist of four main components: pump, injector, column compartment (often thermostated), and detector (Fig. 3.8). The mobile phase is pushed into the systems through a high-pressure pump. The sample is introduced into the flow of the mobile phase through an injector (manual or automated injection) and carried into the column where the separation takes place. A detector is placed at the outlet of the column in order to directly detect the eluted samples. The information from the detector is sent to a computer that generates a graph, called chromatogram, by plotting the detector response *versus* the time of analysis (see also Section Retention and Fig. 3.4).

The basis of the success of (U)HPLC lies in its versatility, indeed it can be applied to different fields, including biological and chemical samples (drugs, pesticides, additives, proteins, etc.) thanks to the availability of a broad variety of stationary phases, mobile phases, and detection methods. The most used elution modes and detectors will be discussed in the following sections.

Principal separation modes

LC was originally performed in normal phase mode (NPLC), in which the stationary phase is polar (i.e., silica, amino, cyano, etc.) and the mobile phase nonpolar (usually, a mixture of alkanes and alcohols), for the separation of hydrophilic compounds. Retention occurs due to the competition for adsorption on the stationary phase between the molecules of the analyte and the mobile phase (Jandera, 2010; Yin & Xu, 2014). The most used elution mode for the separation of a wide variety of samples, ranging from strongly hydrophobic to ionic compounds, is reversed-phase LC (RPLC). It makes use of a polar mobile phase (i.e., water, alcohol, acetonitrile, etc.) and an apolar stationary phase (i.e., C8, C18, phenyl-hexyl, etc.). Nevertheless, highly polar nonionizable analytes cannot be retained under RP chromatographic conditions. To overcome this problem, the so-called hydrophilic interaction chromatography (HILIC) has been developed. HILIC employs stationary phases typical of NP (e.g., amino, cyano, etc.) and mobile

**FIGURE 3.8**

Schematic representation of an HPLC (or UHPLC) system equipped with solvent reservoir, pump, injector and sample loop, column and column thermostat, detector, and data acquisition system. *HPLC*, high-performance liquid chromatography; *UHPLC*, ultrahigh-pressure liquid chromatography.

phases similar to RP mode (a mixture of organic solvent and water), with both adsorption and partitioning retention mechanisms (Jandera, 2010). Fig. 3.9 shows the complementarity of HILIC with respect to NPLC and RPLC. In particular, HILIC allows achieving higher sensitivity when MS is employed as a detection method (see Section Detectors in Liquid chromatography) for highly polar analytes, thanks to the high content of organic solvent employed.

As already pointed out, a separation method to be effective requires the correct selection of process parameters and the right combination of mobile and stationary phases depending on the nature of the analyte. If complex mixtures of solutes with a wide range of capacity factors (or polarity) are considered, gradient elution chromatography is the method of choice to achieve more efficient and fast separation. Conversely to isocratic elution mode, in gradient elution the composition of the mobile phase is altered during a chromatographic run (Fig. 3.10). In addition to the mobile phase composition, column temperature, injection volume, and the mobile phase flow rate can also be changed in order to optimize the separation.

In the last years, the continuous improvement in both instrument and column technologies has led UHPLC instruments equipped with sub- $2\mu\text{m}$ columns to be routinely used in the majority of laboratories, with gradient elution RP as the preferred chromatographic method for metabolic profiling (especially for nonpolar and moderately polar analytes). This technique can indeed provide high resolution and

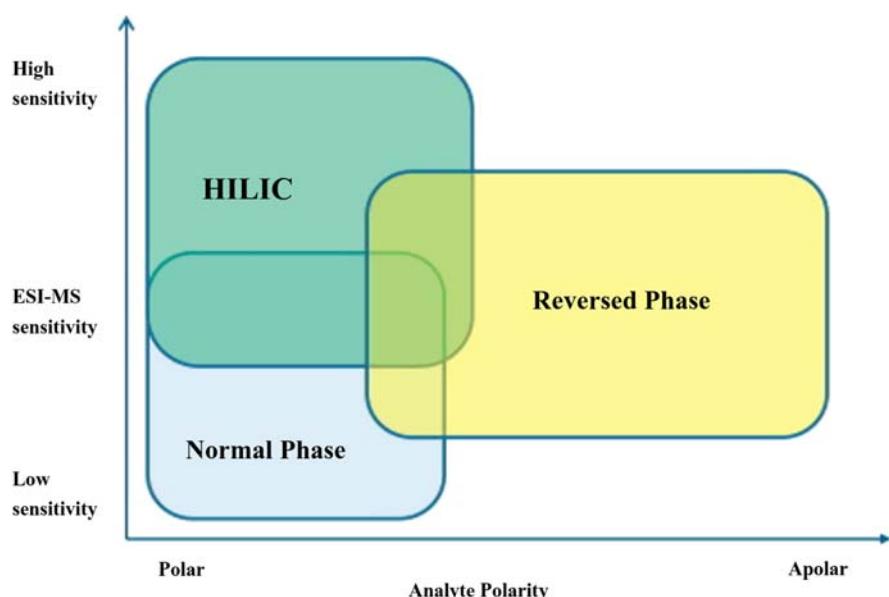


FIGURE 3.9

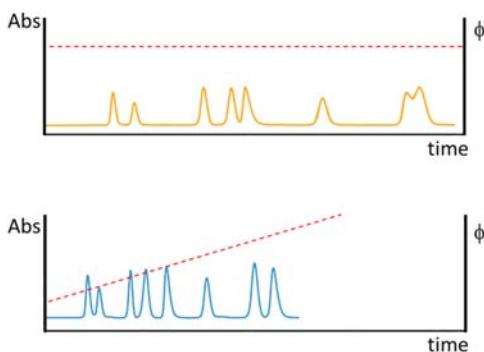
Comparison between HILIC, NP, and RP chromatographic modes.

From Kahsay, G., Song H., Van Schepdael, A., Cabooter, D., Adams, E. (2014). Hydrophilic interaction chromatography (HILIC) in the analysis of antibiotics. *Journal of Pharmaceutical and Biomedical Analysis* 87, 142–154.

sensitivity, high throughput analysis, good repeatability, and wide metabolite coverage at the same time (Gaikwad, 2013; Yin & Xu, 2014). On the other hand, for the separation and analysis of polar metabolites, HILIC mode can be effectively used as a complementary method to RP (Spagou et al., 2010; Yin & Xu, 2014). As an example, in the case of urine, tissue, or cell extracts samples, a combination of these two techniques can be advantageous due to the high percentage of both hydrophilic and polar metabolites (Chen et al., 2009; Fei et al., 2014; Yin & Xu, 2014).

Detectors

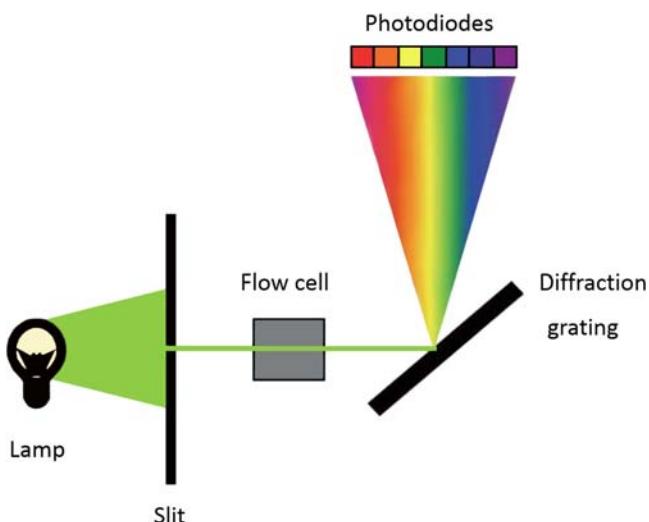
The identification and quantification of analytes can be achieved through the use of a variety of methods of detection available for LC, including diode array detectors (DAD), also commonly called the photodiode-array (PDA), refractive index detectors, fluorescence detector, evaporative light scattering detection (ELSD) and the MS. Among those, the two most common detectors for metabolite analysis coupled with (U)HPLC are the UV-Vis absorbance spectrophotometry and the MS.

**FIGURE 3.10**

Comparison between isocratic (top) and gradient elution (bottom) LC separation.
 Φ = mobile phase composition.

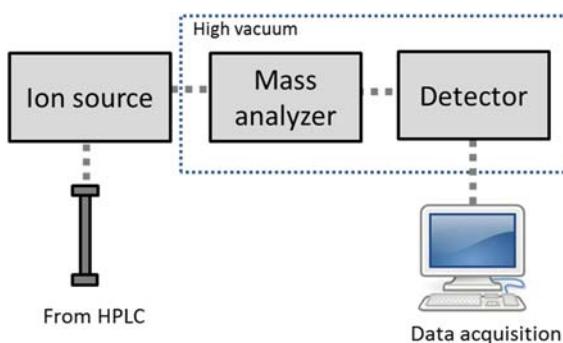
Absorbance detectors are based on the capacity of many compounds to absorb UV-Vis light at a defined wavelength due to the presence of chromophores in their structure. The detection of the sample occurs by measuring the difference in absorbance between the pure mobile phase and the mobile phase containing the sample (Dolan, 2016). The success of absorbance-based detectors for LC analysis lies, on the one hand, in their stability and ease of use and, on the other hand, in the ability to collect the entire UV–Vis spectrum (190–700 nm) of the eluting analyte through the so-called photodiode array detector, PDA (Fig. 3.11). The recorded spectra are a useful tool for the evaluation of the purity, the identification of compounds through the comparison with library spectra, and for the determination of the most suitable wavelength.

An LC instrument can be interfaced to a MS (LC-MS) as schematically shown in Fig. 3.12. The combination of LC and MS permits to separate and identify analytes based on the mass-to-charge (m/z) ratios of ions derived by fragmentation (Fig. 3.13) through the comparison of these patterns with spectral reference libraries (Stein, 2012). It must be pointed out that LC and MS would be incompatible without a proper interface placed in between the two parts of the instruments. Indeed, while the mobile phase used in LC is a pressurized liquid, MS operates under high vacuum. Therefore the eluate from the LC column cannot be directly introduced into the MS. Furthermore, LC usually operates at ambient temperature while MS requires elevated temperature. Among the various interfaces already developed, Atmospheric Pressure Ionization (API) interfaces are commonly used. For metabolomics application especially two types of API interfaces are particularly useful, namely, electrospray ionization and atmospheric pressure chemical ionization (APCI). The former one is most suitable for polar and semipolar analytes, while the latter is mostly used for neutral or nonpolar compounds. A more detailed overview of MS and ionization methods is given in Chapter 4, Mass Spectrometry in Metabolomics.

**FIGURE 3.11**

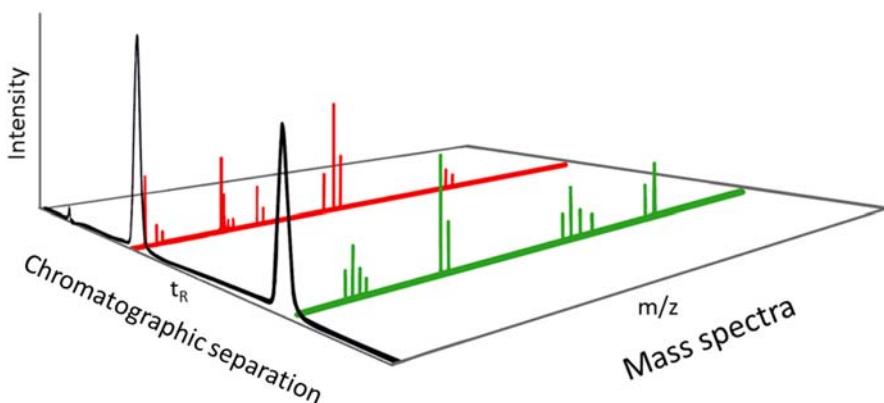
Schematic diagram for a diode-array detector.

From Dolan. (2016). How does it work? Part IV: Ultraviolet detectors. *LC-GC North Am.*, 34, 534–539.

**FIGURE 3.12**

Schematic representation of LC-MS.

LC-MS is the dominant analytical method for metabolomics thanks to its wide metabolite coverage and applicability in both nontargeted and targeted metabolomics analyses, high selectivity, efficiency, and sensitivity. Moreover, it can provide several metabolome information with less extensive sample preparation (Spagou et al., 2010).

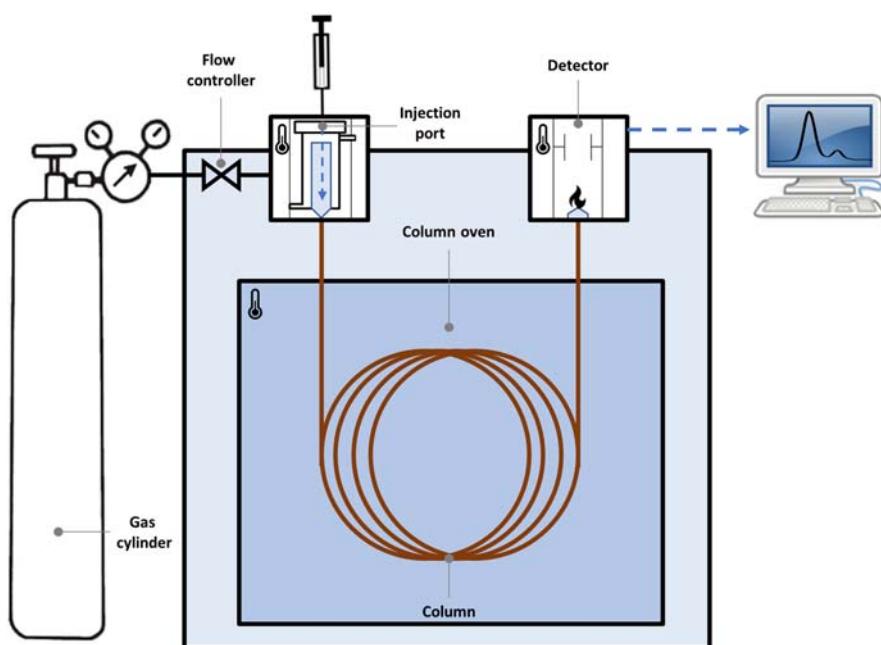
**FIGURE 3.13**

Schematic 3D representation of a chromatogram generated by chromatography coupled to mass spectrometry. Left axes: chromatogram of eluted peaks from the chromatographic column, right axes: mass spectra obtained for each peak.

Gas chromatography

GC is nowadays a very common tool available for chemical analysis in the laboratory, and it represents the technique of choice for the analysis of volatile compounds that are thermally stable at their vaporization temperature. However, as discussed in the previous paragraph (Section Derivatization), for those thermally-labile compounds, derivatization can be used to increase thermal stability and allow GC analysis. The scheme of a typical GC instrument is illustrated in Fig. 3.14 and the key components are highlighted. A measured amount of sample is introduced into the gaseous mobile phase stream through a hot injector. Sample injection can be achieved manually or automatically by using an autosampler. From the injector, the sample enters the chromatographic column, located inside the GC oven, in which it is physically separated in its components. The carrier gas, accurately regulated with flow- and pressure-controllers, transports the separated components of the sample into a detector that generates an electrical signal, related to the nature and concentration of each analyte, which is recorded and can be further processed into data systems. Depending on the detector, the chromatogram contains various qualitative and quantitative information regarding the components of the sample injected (Section Qualitative and quantitative analysis in chromatography).

The thermal stability of the column stationary phase usually sets the upper temperature limit of the GC separation, with the majority carried out at temperatures below 350°C, even if some special columns can reach up to 430°C, which generally makes possible the elution of analytes with molecular weight < 1000 Da. The GC separation, especially in temperature-programmed (see Sections Mobile phase

**FIGURE 3.14**

Schematic representation of a gas chromatography system and its components.

and flow control and Temperature zones), can be considered as a continuous online distillation process, in which the elution follows the boiling point of the analytes. However, even though the boiling point of the analyte plays the most significant role in the elution process, also the interactions analyte-stationary phase determines the effectiveness of the final separation. The technological advancement of GC hardware and columns has brought to the reduction of instrumental footprint, operating costs, maintenance, and development of portable devices. Ultrafast and fast GC separations (from seconds to few minutes) can also be achieved by maintaining high column efficiency, thanks to fast heating technologies and microbore or column geometries (Poole, 2012).

Mobile phase and flow control

The mobile phase in GC is an inert and ultrapure gas, or carrier gas, which is applied under pressure to the column inlet. The inertness of the gas means that not only it does not react with the analytes, but also that, in contrast to LC and SFC, it has no role in the analytes retention (no chemical interaction, such as adsorption/desorption or partition effects), with its primary purpose being to carry the sample through the column. The intrinsic characteristics of the carrier gas

(e.g., density, viscosity, compressibility) will affect column efficiency and separation time. Indeed, as a compressible gas, it expands with the increase of temperature. This results in a change in its viscosity, which impacts the resulting flow, and then its velocity inside the column. The optimal carrier gas linear velocity is characteristic for each gas and is derived from the van Deemter plot (Section Fundamentals of chromatography and Fig. 3.5). Usually, the linear gas velocity is set by changing the flow rate until the maximum plate number (or minimum plate height) is achieved. With conventional columns ($30\text{ m} \times 0.25\text{ mm}$, using helium as carrier), a column flow or a linear velocity of around $0.9\text{--}1.2\text{ mL min}^{-1}$ or $30\text{--}35\text{ cm s}^{-1}$ can be considered optimum values. Among the suitable types of carrier gas, helium is the most commonly used thanks to its safety, detectors compatibility, and intermediate properties among the other gases that allow to achieve efficient separation at acceptable separation times, despite its cost. Other carrier gases available are hydrogen and nitrogen. The accurate measurement and control of carrier gas flow are essential for both column efficiency and reproducibility for qualitative and quantitative analysis. The temperature control allows isothermal and temperature-program analysis (Fig. 3.15), a concept similar to the isocratic and gradient elution in LC (Fig. 3.10). However, most GC separations are carried out at temperature-program and constant flow or linear velocity conditions (McNair, 2019). In the latter case, in order to maintain a constant flow within the column over the GC analysis, the pressure varies

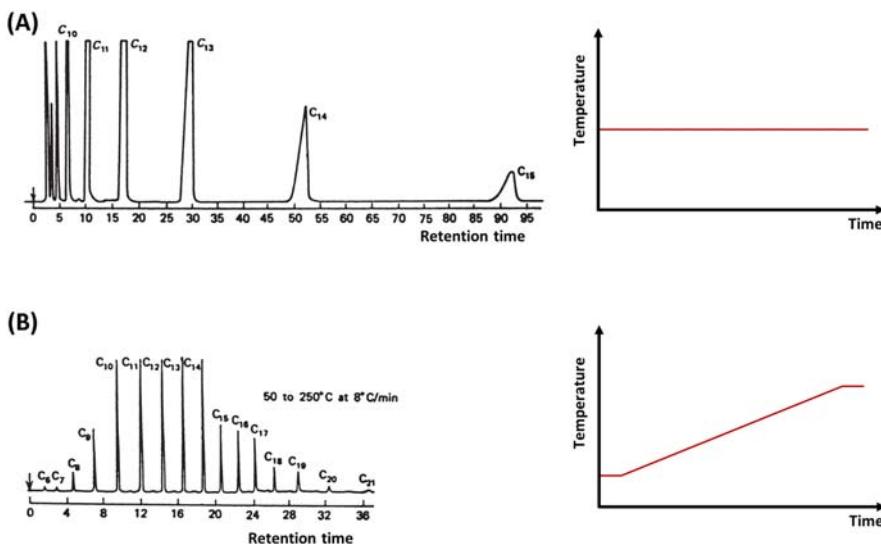


FIGURE 3.15

Graphical representation of isothermal (A) and temperature-program (B) separations.

Modified from McNair, H. M., Miller, J. M., & Snow, N. H. (c.2019). Basic gas chromatography. In *Basic gas chromatography*. Wiley. <https://doi.org/10.1002/9781119450795>.

(pressure-program). This situation allows for the elution of a wide boiling sample mixture and it maintains the ideal separation efficiency through the entire analysis.

Temperature zones

During the entire GC separation process, the whole system is maintained at a temperature sufficient to keep the sample components in the vapor phase. Since temperature is the most effective way to influence the separation, its control and regulation must be accurate and precise. Generally, three temperature-controlled zones exist in a GC system: the injector port, the column, and the detector. The injection block should be hot enough to vaporize the sample rapidly so that no loss in efficiency results from the injection technique, and usually works under isothermal conditions, or temperature-programmed. The column is located in the GC oven and its temperature should be high enough so that the sample components pass through it at a reasonable speed and may be under isothermal temperature or temperature-programmed conditions during the analysis (Fig. 3.15). The detector and its connections (e.g., transfer line) from the column exit are maintained at isothermal conditions and must be hot enough to prevent the recondensation of the sample. As it will be discussed, the measure and proper regulation of the temperatures, pressures, and flows is fundamental to ensure precise and accurate qualitative and quantitative analysis.

Sample introduction and inlets

In GC, the temperature of the injector allows the volatilization of the analytes, a series of valves control the amount of sample that enters the chromatographic column and determine the injection mode: split or splitless (S/SP). The split injection is preferred for concentrated samples since only a preset portion of the sample vapor is transferred into the column to avoid chromatographic overload and thus poor separation. Splitless is a variation of the split mode, very useful for trace-level analytes, and in which essentially the whole injected amount enters the column. In addition to the split/splitless inlet, the programmed temperature vaporizer (PTV) allows more flexible control of the inlet temperature. In the latter, the sample can be injected into a cold injector, which is then rapidly heated using a defined temperature program. Analytes evaporate according to their vapor pressure, reducing temperature stress for thermally labile compounds and preventing their degradation. It is also possible the injection of large volumes of a liquid sample (e.g., $>5\text{ }\mu\text{L}$ till $\sim 100\text{ }\mu\text{L}$), thanks to the proper elimination of the solvent in which the sample is diluted. The S/SL and PTV inlets are also compatible with SPME (see Section Sample preparation) injection, in which the analytes are thermally desorbed from the fiber into the inlet and transferred into the column. An additional thermal desorption unit is instead necessary prior to the aforementioned inlets for special applications, for example when using thermal desorption

tubes for the collection of volatile analytes (e.g., exhaled breath, cell culture metabolites). In this case, the injection process involves 2 steps: a first thermal release of the analytes from the sampling device into the inlet block, followed by the injection into the column.

Column, stationary phases, and separation

Even if packed columns exist, GC separations are mostly performed using capillary columns (OT) (Fig. 3.16). These are fused-silica columns of variable length (1–60 m) and inner diameter (0.05–0.5 mm), generally surrounded by an external polyamide protective layer. The stationary phase is coated on the inner wall of the column as a thin film, defined as wall-coated OT (WCOT); it can be dispersed on inert particles adhering to the column wall, so-called support-coated OT (SCOT); or it can consist of a solid porous film on the column wall, or porous-layer OT columns (PLOT). A liquid or viscous liquid stationary phase usually characterizes WCOT and SCOT columns, and the separation here is attained by analyte partitioning into the liquid (GLC, Table 3.1), while with the solid porous layer of PLOT columns the separation is based on adsorption mechanisms (GSC, Table 3.1). The WCOT columns are by far the most used nowadays, including for metabolomics applications, and different stationary phase chemistries exist.

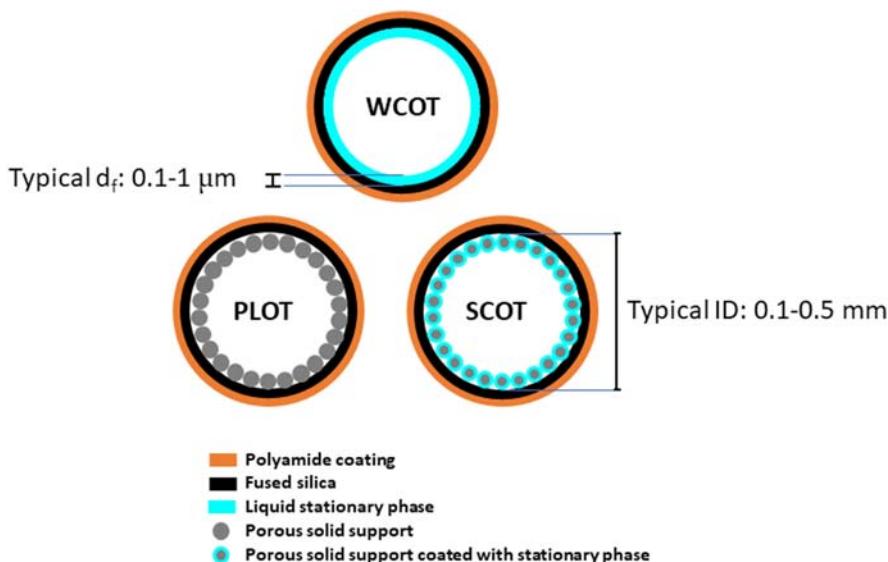


FIGURE 3.16

Types of open-tubular columns in GC.

As discussed before, since no interaction with the mobile phase exists, the retention times of analytes are determined by the difference in their vapor pressures and their different interaction with the stationary phase. The vapor pressure indeed influences the volatility, and it is inversely related to the molecular weight of the analyte (high vapor pressures characterize low molecular weight analytes, which are more volatile). As a general rule, high molecular weight analytes elute at higher elution times compared to low molecular weight analytes. On the other hand, physico-chemical interactions between the analyte and the stationary phase (i.e., hydrogen bonding, dipolar interaction, steric affinity) shift the analytes to later relative retention times, compared to less strongly-interacting analytes. The stationary phases can be classified in a scale of polarity, with the polysiloxane (silicone) backbone generally functionalized with noninteracting functional groups, like methyl or octyl for the nonpolar phases, while polar phases are functionalized with groups such as cyanopropyl, hydroxyl, and phenyl. The nonpolar phases tend to separate analytes based on their vapor pressures since there is no or little specific chemical interaction. Another nonsilicone phase is the common and polar phase based on polyethylene glycol. The recently introduced ionic liquid represents an attractive new alternative to silicone and glycol phases (Mazzucotelli et al., 2019). Ionic liquids are a class of organic solvents with a very low melting point that includes organic cations and either inorganic or organic counterions, and as stationary phases they offer a unique selectivity for polar compounds.

However, owing to the huge chemical diversity of metabolites, there is no single ideal column chemistry capable of separating the entire metabolome. Instead, multiple stationary phase chemistries tailored to different classes of metabolites or their derivatives and high-efficiency/resolution columns might need to be used in parallel to achieve complete separation. Another powerful strategy to achieve high-resolution separations is the use of comprehensive two-dimensional GC, or GC × GC (see Section Multidimensional chromatography).

In any case, the selection of an appropriate stationary phase is critical for a successful separation. Indeed, analytes and polar phases with similar chemical properties will have a stronger affinity for each other, translating into higher retention. For example, nonpolar and polar compounds are better separated using nonpolar column and strongly polar column, respectively; if the analytes of interest have a wide range of boiling points, they are better separated on nonpolar columns; on the other hand, isomers or compounds with little difference in their boiling point are better separated on strongly polar columns. For the analysis of low boiling compounds, the use of longer columns with thicker coatings and shorter columns with thinner coatings will enhance the separation, respectively.

A unique qualitative feature that is commonly used in GC is the exploitation of the retention index system (Fig. 3.17), that is the standardization of the elution time of the analytes of interest in relation to chemically homologous series of standard compounds (e.g., linear saturated hydrocarbons, aromatic hydrocarbons, etc.).

The retention index system adds a level of identification confidence by comparing the measured indices with the available database, it can be used with all

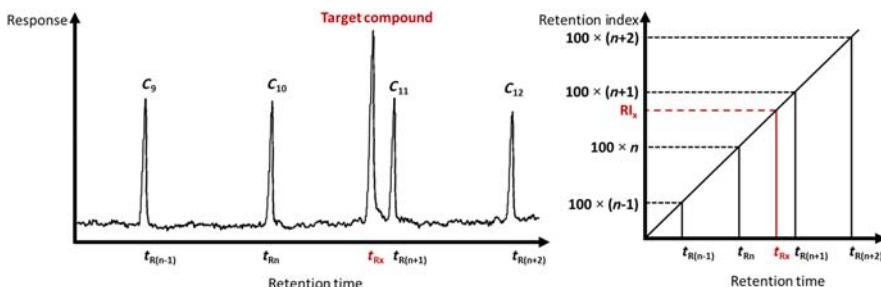
**FIGURE 3.17**

Illustration of the retention index method in GC using the homologous series of linear alkanes.

classes of analytes, and it is very reliable. Especially when combined with mass spectral information (e.g., MS library comparison), the retention index is particularly helpful to reduce the risk of misidentification. This system was initially developed for isothermal condition by Kovats, though it is mostly used nowadays in programmed temperature using the formula introduced by van den Dool and D. J. Kratz ([d'Acampora Zellner et al., 2008](#)):

$$RI_x = 100 \left(\frac{t_{RX} - t_{Rn}}{t_{R(n+1)} - t_{Rn}} + n \right) \quad (3.17)$$

where t_{Rn} , $t_{R(n+1)}$ and t_{RX} , are the retention times (in minutes) of the two n -alkanes containing n and $n + 1$ carbons and of the compound of interest, respectively. This distinctive index is primarily dependent on the type of stationary phase interaction with the analyte, being nearly independent of the temperature ramp of the method, film thickness, column length, column diameter, and carrier gas velocity.

Detectors

A detector senses the effluents from the column and provides a record of the separation in the form of a chromatogram. The detector signals are proportionate to the quantity of each solute making possible a quantitative analysis. In the context of metabolomics, surely MS detection is the most important. In some instances, especially in targeted methodology and for quantitative analysis, the flame ionization detector (FID) is used. This detector is schematically illustrated in [Fig. 3.18](#).

Column effluent enters the FID and mixes with hydrogen combustion gas and a make-up gas (if required) in the lower part (as shown in [Fig. 3.18](#)). This gas mixture burns in the upper part in an excess of air and all organic compounds are decomposed and ionized in the flame. These ions produced by the combustion are collected by an electrode and converted to a voltage that will represent the detector signal. The FID has the desirable characteristics of high sensitivity, linearity, and yet it is relatively simple and inexpensive. The FID is considered a universal

detector, responding to most carbon-containing compounds and having pretty consistent response factors. All these aspects make the FID the most used in GC considering all application fields, however, it does not provide any chemical information on the analytes.

On the other side, MS detection holds the leading position for qualitative analysis, giving a snapshot (mass spectra) of the chemical structure of the analytes (Fig. 3.13). MS, its principles, operations, and components will be thoroughly discussed in the following Chapter 4, Mass Spectrometry in Metabolomics.

The MS coupling to GC represents a much easier task compared to LC, in which special interfaces are needed to volatilize and ionize the analytes in the liquid effluent. Indeed, the physical state of eluent going out from the GC columns is ideal and immediately compatible with the vacuum conditions of the MS. The most used and standardized ionization method is the electronic impact (EI), a hard ionization method that can produce highly-informative and compound-specific fragmentograms. Thanks to its reproducibility, mass spectral libraries are available which facilitate the identification of compounds. Softer EI is also possible and it produces less fragmented molecules, with the possibility to obtain information on the molecular weight (Beccaria et al., 2018). Other possible soft ionization techniques exist, such as the chemical ionization, the field ionization, the photo ionization and the APCI.

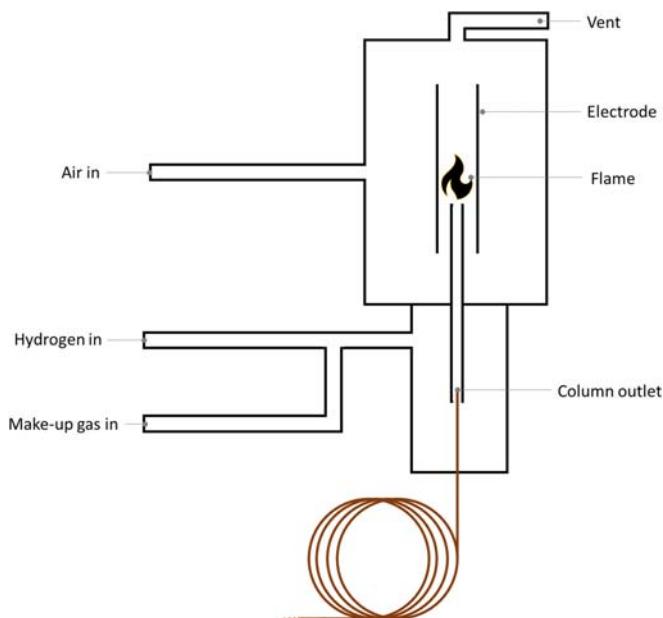


FIGURE 3.18

Schematic illustration of an FID.

Multidimensional chromatography

At present, conventional monodimensional (1D) chromatography is the most commonly applied method for the separation of real-world samples. However, it has become increasingly clear that the baseline separation of all the constituents of a sample or specific analytes of interest from the matrix is often difficult with a single chromatography column. The separation efficiency, as discussed in Section Efficiency of separation, can be increased, for example, with longer (and/or narrower) columns, often at the cost of analysis time, until reaching instrumental limits (e.g., back pressure). However, the most effective way of enhancing the separation efficiency (and the selectivity) of a chromatographic system, with equivalent detection conditions and analysis time, is by using a multidimensional chromatographic system, in which an extra separation dimension is present.

Concept of multidimensionality

A series of combinations of different separation mechanisms can be used to create multidimensional separation systems. A significant number of combinations involving different types of chromatographic processes (LC, GC, SFC) have already been implemented successfully, and the experimental results nicely illustrated the potential of high separation power typically associated with these advanced techniques (Blumberg, 2011; François et al., 2011; Jandera, 2011).

Practically, a basic multidimensional separation can be obtained using TLC by exploiting the first separation with one solvent; after this, by rotating the chromatographic plane by 90 degrees, a second and independent/orthogonal separation will develop, if a different solvent is used: in this case, all the components will undergo a multidimensional separation. Another multidimensional chromatography experiment can be achieved by collecting off-line a fraction of LC eluate after the column and reinjecting it into a different LC column for further separation: in this case, this fraction undergoes a two-dimensional separation. In modern multidimensional instruments, the on-line transfer between the first and second dimensions is generally enabled and it can be controlled carefully by the user. The basic requirements for a multiple separation to be considered multidimensional were discussed by Giddings in 1990 (Giddings, 1964), with two conditions to be satisfied:

1. the components of the mixture should be subjected to two (or more) separation steps in which their retention is governed by different factors; and
2. the analytes that have been resolved in the first step should remain separated until the following separation process is completed.

When two (or more) independent separation mechanisms are used, an equal number of parameters will define the identity of an analyte (Marriott and Wu, 2012). Considering two-dimensional (2D) chromatography, each analyte is characterized by two different retention times rather than by a single one (as in conventional 1D LC or GC).

The second condition requires the distinct analysis of relatively small fractions of eluent from the first dimension (^1D) column to the second one (^2D), to maintain the separation already achieved in ^1D . If the dimensions are based on different interaction mechanisms, the separation is said to be “orthogonal” (Venkatramani et al., 1996). The more the correlation between the dimensions, the more is the redundancy of the information generated.

To illustrate the concept of orthogonality, Fig. 3.19 represents three examples of various degrees of correlation between two separation dimensions. In the case of total orthogonal separation, the peaks (round dots) are distributed over the entire plane (Fig. 3.19A). The more the dimensions are correlated, the more the analytes will be aligned along the diagonal (Fig. 3.19B). In the extreme case of total correlation (Fig. 3.19C), the analytes will have the same retention in the two dimensions, resulting in the equivalent 1D separation along the diagonal, with no additional information given by the second dimension, because identical.

The separation dimensions must be correctly selected to form efficient multidimensional systems and to obtain ordered distributions of the compounds for a true increase of useful information. Giddings introduced the notion of sample dimensionality (S), representing the number of independent variables describing the properties of the sample compounds (Giddings, 1995). The dimensions can be expressed at several levels: π -aromaticity interactions, chirality, hydrogen bonds, ion mobility, size or shape of molecules, chemical functions, volatility/number of carbon atoms, degree of branching, etc.

There are two types of multidimensional chromatography: heart-cut or classical multidimensional (indicated as LC-LC or GC-GC) and comprehensive two-dimensional chromatography (indicated as $\text{LC} \times \text{LC}$ or $\text{GC} \times \text{GC}$) (Marriott and Wu, 2012). In heart-cut, two different columns are used, but only a small portion of the material eluting from the ^1D is introduced for further separation

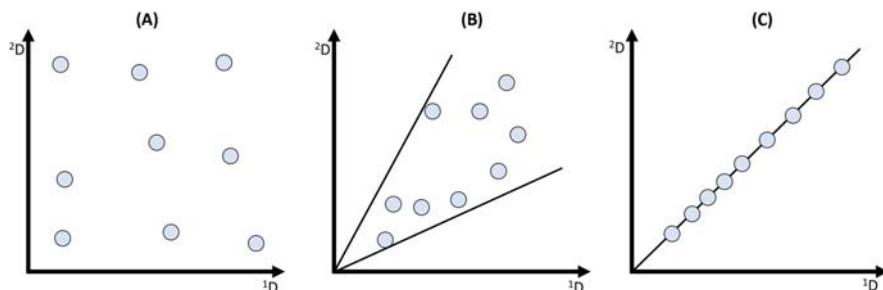


FIGURE 3.19

Illustration of the various degrees of correlation and the separation space utilization between two separation dimensions: orthogonal separation (A), separation with correlation (B), and separation with an absolute correlation.

From Venter, A. (2004) PhD Thesis: Comprehensive Two-Dimensional Supercritical Fluid and Gas Chromatography (SFCxGC), University of Pretoria.

into the ^2D . The number of peaks that a chromatographic analysis can resolve can be expressed by the system peak capacity (n_c) (Section Peak capacity) (Calvin Giddings, 1967). In heart-cutting, the peak capacity equals the sum of that of the first and second dimensions, the latter multiplied by the number (i) of heart-cuts:

$$n_{tot} = [n_{c1} + (n_{c2} \times i)] \quad (3.18)$$

In an ideal comprehensive chromatographic system, the total peak capacity becomes that of the first dimension multiplied by that of the second dimension:

$$n_{tot} = n_{c1} \times n_{c2} \quad (3.19)$$

The increased separation power in multidimensional chromatography is graphically illustrated in figure Fig. 3.20.

The usefulness of the comprehensive approach can be schematically illustrated for the separation of a complex mixture of analytes (Fig. 3.21): a hypothetical sample that contains a large number of analytes that differ in shape, color, and size, is considered with a dimensionality of three, following Giddings's guidelines. In these conditions, there is virtually no chance to separate all the analytes using a conventional single dimensional system. Exploiting a 1D system, the separation can either be performed according to the size, but the color and the shape

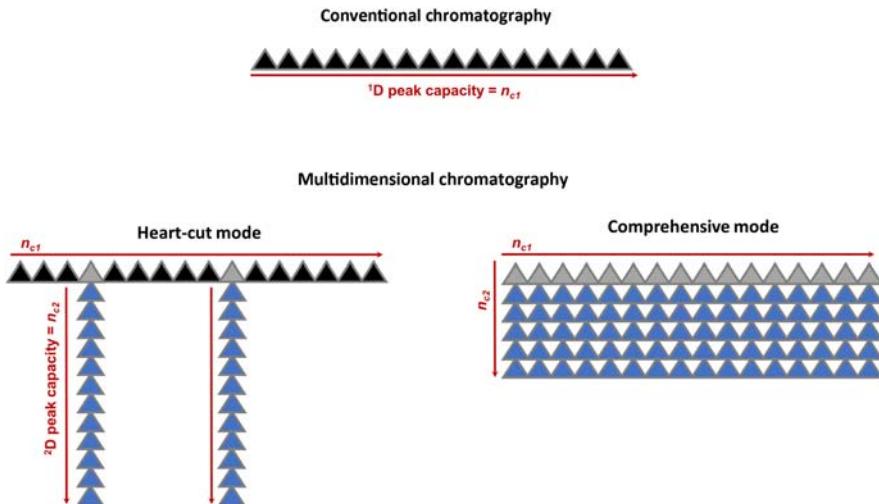
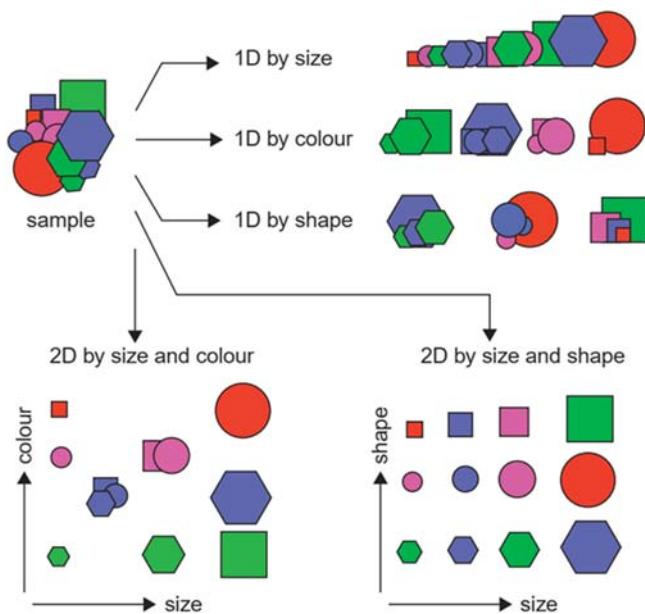


FIGURE 3.20

Illustration of the theoretical separation power accessible by the 1D and the 2D chromatography.

From Ramos, L., & Brinkman, U. A. (c.2009). Chapter 1 Multidimensionality in gas chromatography: General concepts. In *Comprehensive analytical chemistry* (vol. 55, pp. 3–14). [https://doi.org/10.1016/S0166-526X\(09\)05501-9](https://doi.org/10.1016/S0166-526X(09)05501-9).

**FIGURE 3.21**

Graphical illustration of the separation for a virtual sample ($S = 3$) with 1D and comprehensive 2D chromatographic methods.

From Semard, G., Adahchour, M., & Focant, J. F. (c.2009). Chapter 2 Basic instrumentation for GC \times GC. In Comprehensive analytical chemistry (vol. 55, pp. 15–48). [https://doi.org/10.1016/S0166-526X\(09\)05502-0](https://doi.org/10.1016/S0166-526X(09)05502-0).

would remain unseparated; another option is to obtain the separation according to the color, this time with the size and the shape remaining unseparated; finally, the third option is the separation according to the shape, and again the size and the color will remain unseparated. A logical approach to achieve the separation of all the constituents of this hypothetical sample is to use an orthogonal two-dimensional separation system with a dimensionality that can match the dimensionality of the sample (Giddings, 1987). In this case, most of the available separation space is used more efficiently to accommodate the analytes and create a highly structured elution pattern (Fig. 3.21).

Resuming, the main general advantages of comprehensive multidimensional chromatography over conventional 1D chromatographic methods, are essentially five (Mostafa et al., 2011; François, Sandra, Sciarrone, et al., 2011; Jandera, 2011):

1. speed (in terms of resolved number of peaks per time unit);
2. selectivity;
3. separation (higher peak capacity);

4. sensitivity through the isolation of chemical noise from the peak(s) of interest. In the case of GC \times GC using thermal modulators, the analyte band compression generated by modulation improves also the signal-to-noise ratio; and
5. spatial order (formation structured 2D chromatograms for chemically-similar compounds).

Practical and instrumental aspects

A LC \times LC or GC \times GC system can be constructed using the same equipment and accessory employed for conventional 1D LC or GC (Fig. 3.22). The additional piece of hardware is an interface or a device, called modulator, to connect the 2 dimensions.

In the case of LC \times LC and GC \times GC, samples are introduced by an injector, allowing all conventional injection techniques to be used onto the first column (^1D) where the first separation occurs; the eluate is then fractionated and re-injected through the modulator onto the second column (^2D) with a different stationary phase for further separation. The two columns can be situated in a single oven, or in two different ones, the latter option providing a higher degree of flexibility during method optimization. Generally, the column used in the ^1D can be a conventional LC or GC column, and the ^2D column is suitable for fast separations (in LC, smaller particles; in GC, narrow and shorter columns). The unique and key component of a comprehensive 2D chromatography system is the modulator, which is placed between the two columns and it accumulates or samples narrow bands of the eluate of the first column (as a rule of thumb 3–4 times) for a fast re-injection into the second column. This process is showed in the inserts of

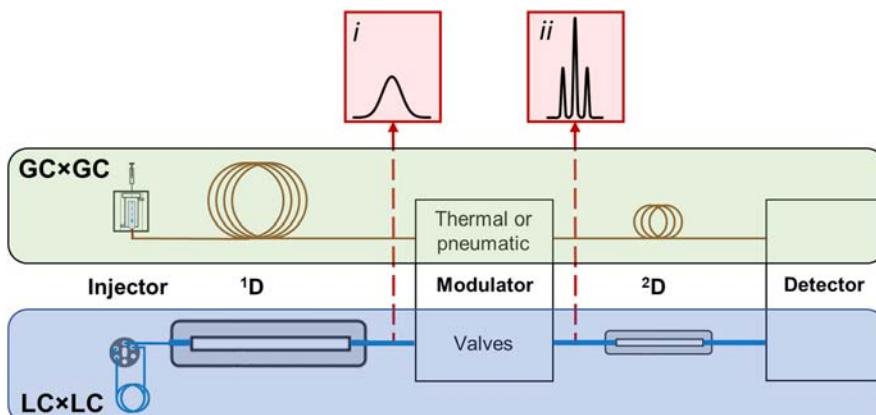


FIGURE 3.22

Side-by-side representation of the main hardware parts of a GC \times GC (top) and a LC \times LC (bottom) system. In the upper inserts, “i” and “ii” illustrate the analyte peak before and after the process of modulation, respectively.

Fig. 3.23 and the left part of **Fig. 3.23**: the transfer process is performed at regular and consecutive time intervals, which define the modulation time. Narrow and higher multiple peaks emerge from the ²D column, whose area under the curve corresponds to the area of the 1D peak.

Modulators are typical of the chromatographic processes involved: in LC × LC, the transfer relies on the use of switching valves (Česla & Křenková, 2017); in GC × GC, the most common modulators exploit the use of temperature (thermal modulators), or pressures (pneumatic modulators) (Tranchida et al., 2011).

The raw detected signal from a comprehensive chromatographic process appears extremely complex, containing an incredibly high number of peaks, and difficult to interpret. Indeed, the comprehensive chromatographic signal requires further data transformation and visualization to make it more readable and interpretable **Fig. 3.23**. Dedicated software packages stack second-dimension chromatograms side by side and, considering the modulation time, derive the first- and second-dimension retention times for each peak. Peak areas are obtained by summing the areas relative to each modulated peak while the signal intensity is considered as the height of the tallest modulated peak of the same compound.

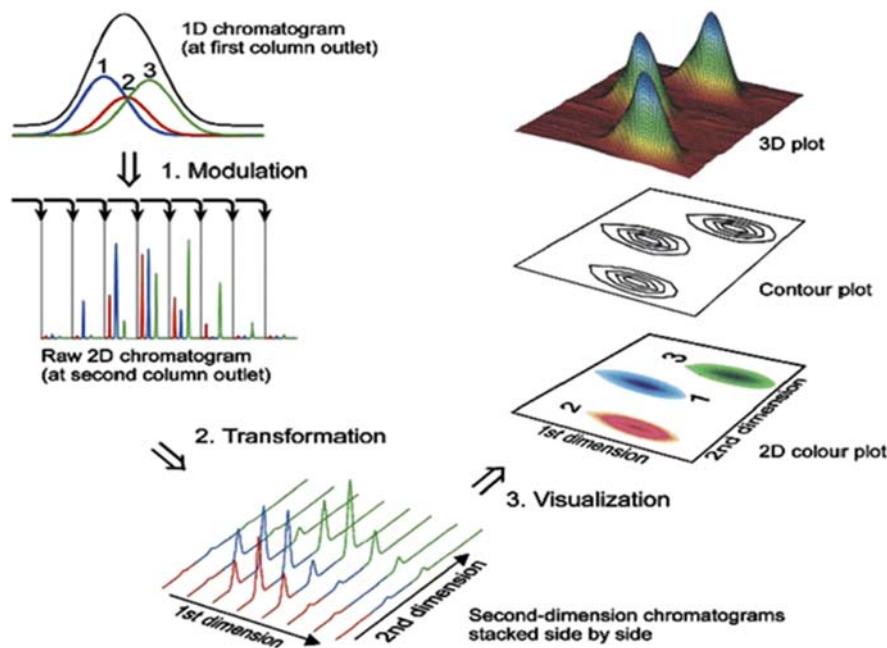


FIGURE 3.23

Data generation and visualization in comprehensive two-dimensional chromatography.

From Adahchour, M., Beens, J., Vreuls, R. J., & Brinkman, U. A. (c.2006). Recent developments in comprehensive two-dimensional gas chromatography (GC × GC). I. Introduction and instrumental set-up. *TrAC - Trends in Analytical Chemistry*, 25(5), 438–454. <https://doi.org/10.1016/j.trac.2006.03.002>.

The 2D chromatograms can be visualized in different ways, and specifically, by the means of colors (2D color plot), by contour line (2D contour plot), and by a 3D visualization (3D plot). Each analyte peak appears as a blob in the 2D plot, in which the color gives the sense of the intensity of the response, which develops towards the z-axis in a 3D space.

Regarding the detection step, generally, the detectors used for conventional 1D chromatography can be used also in 2D chromatography. However, because of the narrower peaks obtained with the modulator, fast detection acquisition rate and low internal volumes are the detector requirements for a successful comprehensive two-dimensional chromatographic separation.

To this end, the modern MS instrumentation is suitable for fast separations and can be used in combination with multidimensional chromatographic systems, making it one of the most powerful analytical tools for the detailed characterization of complex samples (Mondello, 2011). Nowadays, various research lines in metabolomics are present using LC \times LC and GC \times GC-MS, in which their superior separation power is exploited for metabolite profiling and fingerprinting (Beckstrom et al., 2011; Di Giovanni et al., 2020; Franchina et al., 2018, 2019; Franchina, Dubois et al., 2020; Franchina et al., 2021; François et al., 2009; Higgins Keppler et al., 2018; Jeong et al., 2010; Navarro-Reig et al., 2017; Stoll et al., 2017; Toro-Uribe et al., 2018; Wang et al., 2010; Welthagen et al., 2005; Wilson et al., 2007; Yang et al., 2020). As an example, a 2D GC chromatogram obtained from a biological sample is showed in Fig. 3.24. Here, a urine sample was derivatized and analyzed via

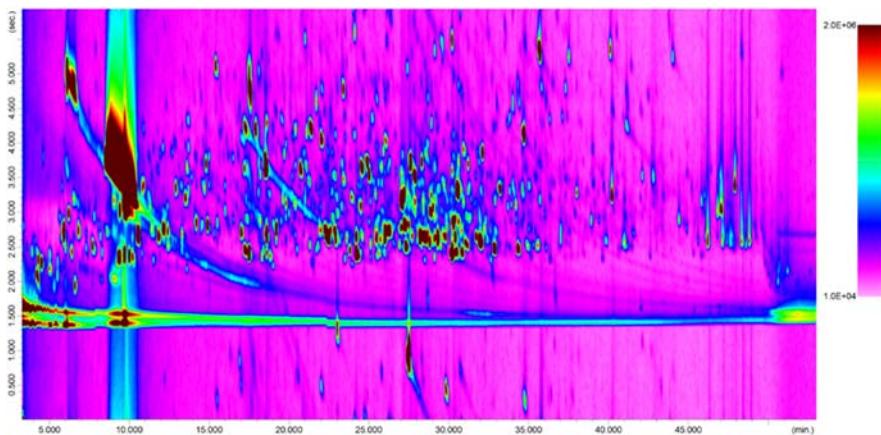


FIGURE 3.24

GC \times GC chromatogram of a urine sample after derivatization.

From Weinert, C.H., Egert, B., & Kulling, S.E. (2015). On the applicability of comprehensive two-dimensional gas chromatography combined with a fast-scanning quadrupole mass spectrometer for untargeted large-scale metabolomics. *Journal of Chromatography A*, 1405, 156–167. <https://doi.org/10.1016/j.chroma.2015.04.011>

GC × GC-MS ([Weinert et al., 2015](#)). Several classes of metabolites were detected and identified, such as amino acids, fatty acids, carbohydrates, and small polar components of glycolysis and the Krebs cycle. The 2D approach made possible the separation of the hundreds of metabolites that would appear unresolved in a single column.

Other separation techniques

Capillary electrophoresis

Capillary electrophoresis (CE) is increasingly used for the separation of metabolites ([Ramautar et al., 2006](#)). This separation method combines both chromatographic and electrophoretic separation mechanisms to separate species according to their mass-to-charge (m/z) ratio. Charged analytes are separated based on their different electrophoretic mobility under the application of an electric field. The electrophoretic mobility depends on different properties of the analyte, such as charge, size, ratio. In any case, the greater the electric field, the greater the mobility of ions. Neutral species are not affected by the electric field and therefore they do not move. The migration rate (ν) of an ion under the electric field is defined by the following equation:

$$\nu = \mu_e \frac{V}{L} \quad (3.20)$$

where μ_e is the electrophoretic mobility, V the applied voltage, and L the length of the capillary. The instrumentation used for CE is very simple ([Fig. 3.25](#)). Two buffers

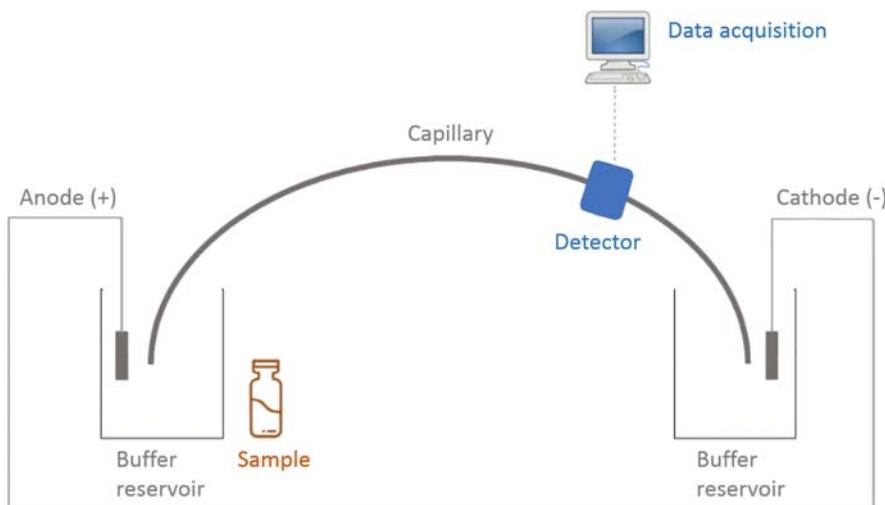


FIGURE 3.25

Schematic representation of a CE system.

reservoirs are connected by a fused-silica capillary with an internal diameter of typically 10–100 µm. Inside the buffer reservoirs, there are also two platinum electrodes. A potential of 5–50 kV is applied at the two electrodes. The sample is introduced by pressure injection when the end of one capillary is inserted into the sample vial.

The peculiarity of CE is the electro osmotic flow (EOF). It is generated by the electrical double layer at the silica/solution interface. Thanks to EOF, the buffer solution flows from one buffer reservoir to the other, just as if it was being pumped. The rate of EOF is generally higher than the migration rate of single ions and it is sufficient to direct all the species (positive, negative, and neutrals) towards one direction, going through the detector. What is generated is an electropherogram, which is very similar to a chromatogram but with narrower peaks. The most used popular detector for CE is UV but for metabolites determination CE coupled to MS is particularly useful. The advantage of CE over LC or GC for metabolomics applications is that it is an orthogonal separation method that allows to separate analytes on mechanisms not based on their interaction with the stationary phase. For this reason, it can be applied for the determination of a large library of compounds (Soga et al., 2003). In addition, only a few nanoliters are necessary for injection, therefore it is particularly suitable for those cases in which the sample volume is limited.

Supercritical fluid chromatography

SFC has been recently started to be applied to metabolomics, in particular for separating a wide range of polar metabolites (Shulaev & Isaac, 2018). SFC is a unique separation method with intermediate properties between LC and GC. Indeed, the mobile phase used is composed of a supercritical fluid (usually CO₂) and an organic modifier (usually an alcohol, such as methanol or ethanol). Density is one of the most important parameters in SFC, playing a fundamental role not only on molecular interactions (hence retention), but also on viscosity, diffusivity, and mobile phase velocity. A significant variation in density could lead to the formation of radial temperature gradients that have a detrimental effect on column efficiency. Therefore modern SFC instruments are designed to ensure adiabatic conditions of the column and to have a strict control of pressure.

The most common detector is UV, but also other detection methods are widely employed such as ELSD. For metabolomics applications, the MS is the most suitable detector and a schematic representation of a SFC-MS equipment is showed in Fig. 3.26.

SFC has been recently applied for the determination of metabolites belonging to different families, including alkaloids, cannabinoids, carotenoids, lipids, steroids, and tocopherols, to name but a few (Shulaev & Isaac, 2018). Owing to its orthogonal properties with respect to LC, SFC seems to be particularly suitable to be used as one dimension in multidimensional separations.

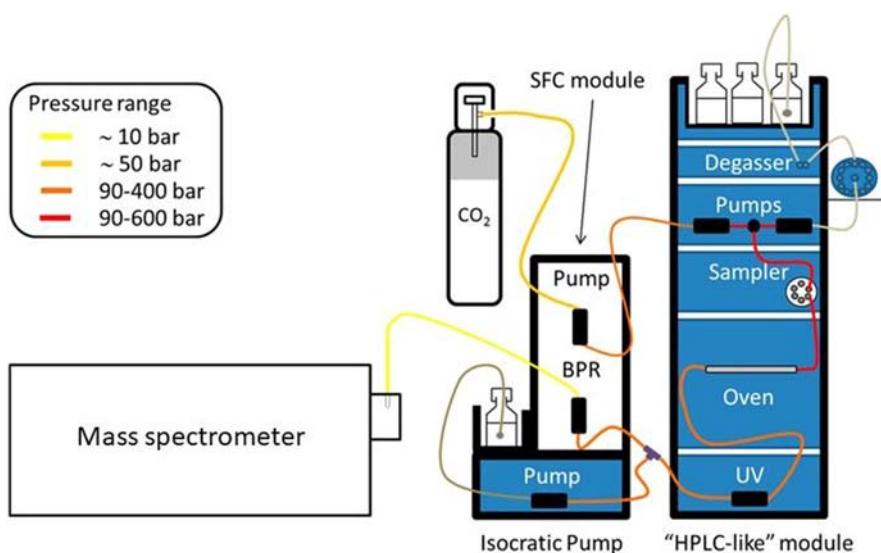


FIGURE 3.26

Schematic representation of an SFC-MS equipment. The key legend refers to the pressure range of the different tubing sections.

Modified from Laboureur, L., Ollero, M., & Touboul, D. (c.2015). Lipidomics by supercritical fluid chromatography. *International Journal of Molecular Sciences*, 16(6), 13868–13884. <https://doi.org/10.3390/ijms160613868>.

Chiral chromatography

Many metabolites that can be found in living organisms are chiral (e.g., nonsuperimposable mirror images of each other), and some of them are considered as specific biomarkers of diseases (e.g., some D-amino acids) (Kimura et al., 2016). Therefore the possibility to specifically recognize enantiomers is acquiring increasing importance in metabolomics. Chromatography is one of the most powerful tools for chiral recognition. The fundamental basis for the separation of enantiomers is their transformation into diastereoisomers. This can be achieved in mainly three different ways (Cavazzini et al., 2011; Lämmerhofer, 2010). The first *indirect approach* involves the use of a pre-column derivatization. The two enantiomers are derivatized with a single enantiomer of chiral molecules to achieve two distinct diastereoisomers that can be separated on an achiral column. The second approach makes use of a chiral auxiliary agent to be added to the mobile phase (this is possible especially in LC) to generate transient diastereoisomers that can be separated on an achiral stationary phase. This method is not recommended since it is very expensive and the presence of the chiral auxiliary agent in the mobile phase may cause severe interference in detection. The third method is the most widely used and it is called *direct approach*. In this case, the

enantiomers directly form transient diastereoisomers with a chiral stationary phase (CSP), where a single enantiomer of chiral molecules (called chiral selector) is covalently immobilized or adsorbed on an appropriate support. CSPs can be used in LC, GC, SFC, and also CE. The physical basis for retention relies on the different energies of the two diastereoisomeric complexes transiently formed. The stability of the complex is driven by the extent of noncovalent interactions between enantiomers and CSP, such as H-bonds, π - π interactions, ionic interactions, dipole stackings, and van Der Waals forces. In many cases, these interactions may also induce specific steric fittings of enantiomers on and within the CSP. CSPs can be classified based on the chiral selector in: (1) macromolecular chiral selectors, including biopolymers (e.g., proteins, polysaccharides derivatives such as cellulose or amylose), synthetic polymers (e.g., polymethacrylamide); (2) macrocyclic chiral selectors, including macrocyclic antibiotics (e.g., teicoplanin, vancomycin), cyclodextrins, chiral crown ethers; (3) low-molecular mass chiral selectors, including pirkle-type donor-acceptor (e.g., Whelk-O1, DACH-DNB), chiral ion-exchangers or zwitterions, and chiral chelating agents. Discrimination of chiral metabolites in biological samples is not an easy task due to the complexity of the sample and due to the presence of other many possible chiral compounds. However, much effort is currently put into the development of selective and sensitive chiral chromatographic methods. As an example, the Fig. 3.27 reports the chromatograms relative to the separations of D,L-proline and D,L-pipercolic acid enantiomers (Nakano et al., 2019). D-proline is considered a

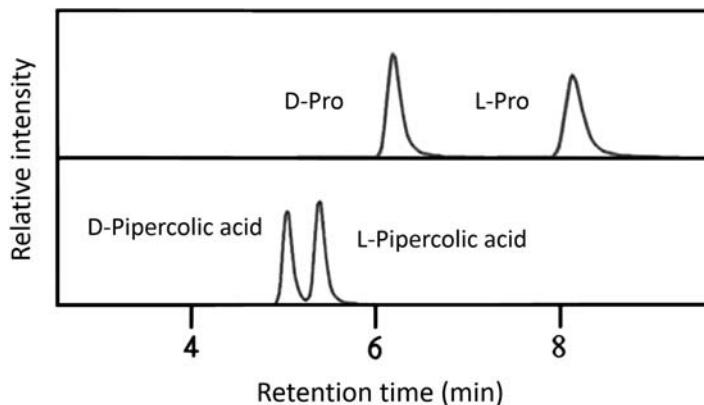


FIGURE 3.27

LC-MS chromatogram of D,L-proline (top) and D,L-pipercolic acid (bottom) enantiomers on a zwitterionic CSP; mobile phase: 25 mM formic acid 25 mM ammonium formate in MeOH and water (98/2, v/v); flow rate: 0.4 mL/min; isocratic elution.

Modified from Nakano, Y., Taniguchi, M., & Fukusaki, E. (2019). High-sensitive liquid chromatography-tandem mass spectrometry-based chiral metabolic profiling focusing on amino acids and related metabolites. *Journal of Bioscience and Bioengineering*, 127(4), 520–527. <https://doi.org/10.1016/j.jbiosc.2018.10.003>

biomarker and its levels can be associated with age, diabetes mellitus, and kidney functions, to name but a few (Kimura et al., 2016). Pipecolic acid is a biomarker for peroxisomal disorders (Steinberg et al., 2008). The two enantiomers have different origins, indeed D-pipecolic acid is derived from diet and intestinal bacteria, while its corresponding L-enantiomer is an endogenous intermediate of the L-lysine pathway. The chromatogram shown in Fig. 3.27 refers to a LC-MS separation obtained with a zwitterionic CSP in reversed-phase conditions.

Chromatography allows calculating the enantiomeric excess directly from the chromatogram. Let us imagine having two peaks corresponding to two enantiomers (1 and 2). The enantiomeric excess (ee %) can be calculated from the areas of the two peaks where enantiomer-1 is more abundant than enantiomer-2:

$$ee\% = \frac{A_1 - A_2}{A_1 + A_2} \times 100 \quad (3.21)$$

For a racemic mixture, the enantiomeric excess is 0% and that for a pure enantiomer is 100%.

References

- Abbas, K. A., Mohamed, A., Abdulamir, A. S., & Abas, H. A. (2008). A review on supercritical fluid extraction as new analytical method. *American Journal of Biochemistry and Biotechnology*, 4(4), 345–353. Available from <https://doi.org/10.3844/ajbbsp.2008.345.353>.
- Beccaria, M., Franchina, F. A., Nasir, M., Mellors, T., Hill, J. E., & Purcaro, G. (2018). Investigation of mycobacteria fatty acid profile using different ionization energies in GC-MS. *Analytical and Bioanalytical Chemistry*, 410(30), 7987–7996. Available from <https://doi.org/10.1007/s00216-018-1421-z>.
- Beckstrom, A. C., Humston, E. M., Snyder, L. R., Synovec, R. E., & Juul, S. E. (2011). Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model. *Journal of Chromatography A*, 1218(14), 1899–1906. Available from <https://doi.org/10.1016/j.chroma.2011.01.086>.
- Blumberg, L. M. (2011). *Multidimensional gas chromatography: Theoretical considerations. Comprehensive chromatography in combination with mass spectrometry* (pp. 13–63). John Wiley & Sons, Inc. Available from <https://doi.org/10.1002/9781118003466.ch2>.
- Calvin Giddings, J. (1967). Maximum number of components resolvable by gel filtration and other elution chromatographic methods. *Analytical Chemistry*, 39(8), 1027–1028. Available from <https://doi.org/10.1021/ac60252a025>.
- Cavazzini, A., Pasti, L., Massi, A., Marchetti, N., & Dondi, F. (2011). Recent applications in chiral high performance liquid chromatography: A review. *In Analytica Chimica Acta*, 706(Issue 2), 205–222. Available from <https://doi.org/10.1016/j.aca.2011.08.038>.
- Česla, P., & Křenková, J. (2017). Fraction transfer process in on-line comprehensive two-dimensional liquid-phase separations. *Journal of Separation Science*, 40(1), 109–123. Available from <https://doi.org/10.1002/jssc.201600921>.

- Chen, J., Wang, W., Lv, S., Yin, P., Zhao, X., Lu, X., Zhang, F., & Xu, G. (2009). Metabonomics study of liver cancer based on ultra performance liquid chromatography coupled to mass spectrometry with HILIC and RPLC separations. *Analytica Chimica Acta*, 650(1), 3–9. Available from <https://doi.org/10.1016/j.aca.2009.03.039>.
- d'Acampora Zellner, B., Bicchi, C., Dugo, P., Rubiolo, P., Dugo, G., & Mondello, L. (2008). Linear retention indices in gas chromatographic analysis: A review. *Flavour and Fragrance Journal*, 23(5), 297–314. Available from <https://doi.org/10.1002/ffj.1887>.
- Diehl, D. M. (2007). *EXTRACTION | sorptive extraction methods. Encyclopedia of separation science* (pp. 1–7). Elsevier, <https://doi.org/10.1016/b978-012226770-3/10678-8>.
- Di Giovanni, N., Meuwis, M. A., Louis, E., & Focant, J. F. (2020). Untargeted serum metabolic profiling by comprehensive two-dimensional gas chromatography-high-resolution time-of-flight mass spectrometry. *Journal of Proteome Research*, 19(3), 1013–1028. Available from <https://doi.org/10.1021/acs.jproteome.9b00535>.
- Dolan. (2016). How does it work? Part IV: Ultraviolet detectors. *LC-GC North America*, 34, 534–539.
- Ettre, L. S. (1975). The M. S. Tswett chromatography medal. *Chromatographia*, 8(11), 603–604. Available from <https://doi.org/10.1007/bf02286256>.
- Fei, F., Bowdish, D. M. E., & McCarry, B. E. (2014). Comprehensive and simultaneous coverage of lipid and polar metabolites for endogenous cellular metabolomics using HILIC-TOF-MS. *Analytical and Bioanalytical Chemistry*, 406(15), 3723–3733. Available from <https://doi.org/10.1007/s00216-014-7797-5>.
- Franchina, F. A., Dubois, L. M., & Focant, J. F. (2020). In-depth cannabis multiclass metabolite profiling using sorptive extraction and multidimensional gas chromatography with low- and high-resolution mass spectrometry. *Analytical Chemistry*, 92(15), 10512–10520. Available from <https://doi.org/10.1021/acs.analchem.0c01301>.
- Franchina, F. A., Zanella, D., Dubois, L. M., & Focant, J. F. (2020). The role of sample preparation in multidimensional gas chromatographic separations for non-targeted analysis with the focus on recent biomedical, food, and plant applications. *Journal of Separation Science*, 44, jssc.202000855-jssc.202000855. Available from <https://doi.org/10.1002/jssc.202000855>.
- Franchina, F. A., Mellors, T. R., Aliyeva, M., Wagner, J., Daphtry, N., Lundblad, L. K. A., Fortune, S. M., Rubin, E. J., & Hill, J. E. (2018). Towards the use of breath for detecting mycobacterial infection: A case study in a murine model. *Journal of Breath Research*, 12(2). Available from <https://doi.org/10.1088/1752-7163/aaa016>.
- Franchina, F. A., Purcaro, G., Burklund, A., Beccaria, M., & Hill, J. E. (2019). Evaluation of different adsorbent materials for the untargeted and targeted bacterial VOC analysis using GC × GC-MS. *Analytica Chimica Acta*, 1066, 146–153. Available from <https://doi.org/10.1016/j.aca.2019.03.027>.
- Franchina, F. A., Zanella, D., Dejong, T., & Focant, J. F. (2021). Impact of the adsorbent material on volatile metabolites during in vitro and in vivo bio-sampling. *Talanta*, 222, 121569. Available from <https://doi.org/10.1016/j.talanta.2020.121569>.
- François, I., Cabooter, D., Sandra, K., Lynen, F., Desmet, G., & Sandra, P. (2009). Tryptic digest analysis by comprehensive reversed phase x two reversed phase liquid chromatography (RP-LC x RP-LC) at different pH's. *Journal of Separation Science*, 32(8), 1137–1144. Available from <https://doi.org/10.1002/jssc.200800578>.
- François, I., Sandra, K., & Sandra, P. (2011). *History, evolution, and optimization aspects of comprehensive two-dimensional liquid chromatography. Comprehensive chromatography in combination with mass spectrometry* (pp. 281–330). John Wiley & Sons, Inc. Available from <https://doi.org/10.1002/9781118003466.ch8>.

- François, I., Sandra, P., Sciarrone, D., & Mondello, L. (2011). *Other comprehensive chromatography methods. Comprehensive chromatography in combination with mass spectrometry* (pp. 429–448). John Wiley & Sons, Inc. Available from <https://doi.org/10.1002/9781118003466.ch11>.
- Gaikwad, N. W. (2013). Ultra performance liquid chromatography-tandem mass spectrometry method for profiling of steroid metabolome in human tissue. *Analytical Chemistry*, 85(10), 4951–4960. Available from <https://doi.org/10.1021/ac400016e>.
- Giddings, J. C. (1964). Reduced plate height equation: A common link between chromatographic methods. *Journal of Chromatography A*, 13(C), 301–304. Available from [https://doi.org/10.1016/s0021-9673\(01\)95123-4](https://doi.org/10.1016/s0021-9673(01)95123-4).
- Giddings, J. C. (1987). Concepts and comparisons in multidimensional separation. *Journal of High Resolution Chromatography*, 10(5), 319–323. Available from <https://doi.org/10.1002/jhrc.1240100517>.
- Giddings, J. C. (1995). Sample dimensionality: A predictor of order-disorder in component peak distribution in multidimensional separation. *Journal of Chromatography A*, 703 (1–2), 3–15. Available from [https://doi.org/10.1016/0021-9673\(95\)00249-M](https://doi.org/10.1016/0021-9673(95)00249-M).
- Godzien, J., Gil de la Fuente, A., Otero, A., & Barbas, C. (2018). *Metabolite annotation and identification*. *Comprehensive analytical chemistry* (82, pp. 415–445). , <https://doi.org/10.1016/bs.coac.2018.07.004>.
- Haggarty, J., & Burgess, K. E. (2017). Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology*, 43, 77–85. Available from <https://doi.org/10.1016/j.copbio.2016.09.006>.
- Higgins Keppler, E. A., Jenkins, C. L., Davis, T. J., & Bean, H. D. (2018). Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics. *TrAC—Trends in Analytical Chemistry*, 109, 275–286. Available from <https://doi.org/10.1016/j.trac.2018.10.015>.
- Hyötyläinen, T. (2013). Sample collection, storage and preparation. In *RSC chromatography monographs*, Issue 18, pp. 11–42. <<https://doi.org/10.1039/9781849737272-00011>>.
- Ismail, O. H., Catani, M., Pasti, L., Cavazzini, A., Ciogli, A., Villani, C., Kotoni, D., Gasparini, F., & Bell, D. S. (2016). Experimental evidence of the kinetic performance achievable with columns packed with new 1.9 µm fully porous particles of narrow particle size distribution. *Journal of Chromatography A*, 1454, 86–92. Available from <https://doi.org/10.1016/j.chroma.2016.05.038>.
- Jandera, P. (2010). *Gradient elution mode. Handbook of HPLC* (2nd ed., pp. 119–154). CRC Press, <https://doi.org/10.1201/ebk1574445541-c5>.
- Jandera, P. (2011). *Multidimensional liquid chromatography: Theoretical considerations. Comprehensive chromatography in combination with mass spectrometry* (pp. 65–92). John Wiley & Sons, Inc, <https://doi.org/10.1002/9781118003466.ch3>.
- Jeong, E. K., Cha, H. J., Ha, Y. W., Kim, Y. S., Ha, I. J., & Na, Y. C. (2010). Development and optimization of a method for the separation of platycosides in Platycodi Radix by comprehensive two-dimensional liquid chromatography with mass spectrometric detection. *Journal of Chromatography A*, 1217(26), 4375–4382. Available from <https://doi.org/10.1016/j.chroma.2010.04.053>.
- Kimura, T., Hamase, K., Miyoshi, Y., Yamamoto, R., Yasuda, K., Mita, M., Rakugi, H., Hayashi, T., & Isaka, Y. (2016). Chiral amino acid metabolomics for novel biomarker screening in the prognosis of chronic kidney disease. *Scientific Reports*, 6. Available from <https://doi.org/10.1038/srep26137>.

- Knox, J. H. (1999). Band dispersion in chromatography—A new view of A-term dispersion. *Journal of Chromatography A*, 831(1), 3–15. Available from [https://doi.org/10.1016/S0021-9673\(98\)00497-X](https://doi.org/10.1016/S0021-9673(98)00497-X).
- Lämmerhofer, M. (2010). Chiral recognition by enantioselective liquid chromatography: Mechanisms and modern chiral stationary phases. *Journal of Chromatography A*, 1217(Issue 6), 814–856. Available from <https://doi.org/10.1016/j.chroma.2009.10.022>.
- Liberto, E., Bicchi, C., Cagliero, C., Cordero, C., Rubiolo, P., & Sgorbini, B. (2020). *Chapter 1: Headspace sampling: An “evergreen” method in constant evolution to characterize food flavors through their volatile fraction*, . *Food chemistry, function and analysis* (Vols. 2020, pp. 3–37). Royal Society of Chemistry, January, Issue 17; <https://doi.org/10.1039/9781788015752-00001>.
- Lopez-Avila, V., & Luque de Castro, M. D. (2014). *Microwave-assisted extraction. Reference module in chemistry, molecular sciences and chemical engineering*. Elsevier, <https://doi.org/10.1016/b978-0-12-409547-2.11172-2>.
- Luque de Castro, M. D., & Delgado-Povedano, M. M. (2014). Ultrasound: A subexploited tool for sample preparation in metabolomics. *Analytica Chimica Acta*, 806, 74–84. Available from <https://doi.org/10.1016/j.aca.2013.10.053>.
- Marriott, S. & Wu (2012). *Nomenclature and conventions in comprehensive multidimensional chromatography—an update*. LC GC Eur.
- Mazzucotelli, M., Bicchi, C., Marengo, A., Rubiolo, P., Galli, S., Anderson, J. L., Sgorbini, B., & Cagliero, C. (2019). Ionic liquids as stationary phases for gas chromatography—Unusual selectivity of ionic liquids with a phosphonium cation and different anions in the flavor, fragrance and essential oil analyses. *Journal of Chromatography A*, 1583, 124–135. Available from <https://doi.org/10.1016/j.chroma.2018.11.032>.
- McNair (2019). Temperature programming. In *Basic gas chromatography*. <<https://doi.org/10.1002/9781119450795.ch6>>.
- Moldoveanu (2014). *Modern sample preparation for chromatography*. <<https://doi.org/10.1016/C2011-0-00093-5>>.
- Moldoveanu, S., & David, V. (2015). *The role of derivatization in chromatography. Modern sample preparation for chromatography* (pp. 307–331). Elsevier, <https://doi.org/10.1016/b978-0-444-54319-6.00009-8>.
- Mondello, L. (2011). *Comprehensive chromatography in combination with mass spectrometry. Comprehensive chromatography in combination with mass spectrometry*. John Wiley & Sons, Inc., <https://doi.org/10.1002/9781118003466>.
- Morris (1974). *Chromatographic detectors*. <<https://doi.org/10.1201/9781482273564>>.
- Mostafa, A., Górecki, T., Tranchida, P. Q., & Mondello, L. (2011). *History, evolution, and optimization aspects of comprehensive two-dimensional gas chromatography. Comprehensive chromatography in combination with mass spectrometry* (pp. 93–144). John Wiley & Sons, Inc., <https://doi.org/10.1002/9781118003466.ch4>.
- Mushtaq, M. Y., Choi, Y. H., Verpoorte, R., & Wilson, E. G. (2014). Extraction for metabolomics: Access to the metabolome. *Phytochemical Analysis*, 25(4), 291–306. Available from <https://doi.org/10.1002/pca.2505>.
- Navarro-Reig, M., Jaumot, J., Baglai, A., Vivó-Truyols, G., Schoenmakers, P. J., & Tauler, R. (2017). Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution. *Analytical Chemistry*, 89(14), 7675–7683. Available from <https://doi.org/10.1021/acs.analchem.7b01648>.

- Nakano, Y., Taniguchi, M., & Fukusaki, E. (2019). High-sensitive liquid chromatography-tandem mass spectrometry-based chiral metabolic profiling focusing on amino acids and related metabolites. *Journal of Bioscience and Bioengineering*, 127(4), 520–527. Available from <https://doi.org/10.1016/j.jbiosc.2018.10.003>.
- Pawlizyn, J., & Lord, H. L. (2011). *Handbook of sample preparation. Handbook of sample preparation*. John Wiley & Sons, Inc., <https://doi.org/10.1002/9780813823621>.
- Poole, C. F. (2012). *Gas-solid chromatography (PLOT columns). Gas chromatography* (pp. 123–136). Elsevier. Available from <https://doi.org/10.1016/B978-0-12-385540-4.00005-5>.
- Poole, C. F. (2020a). *Core concepts and milestones in the development of solid-phase extraction. Solid-phase extraction* (pp. 1–36). Elsevier, <https://doi.org/10.1016/B978-0-12-816906-3.00001-7>.
- Poole, C. F. (2020b). *Milestones in the development of liquid-phase extraction techniques. Liquid-phase extraction* (pp. 1–44). Elsevier, <https://doi.org/10.1016/B978-0-12-816911-7.00001-3>.
- Ramautar, R., Demirci, A., & Jong, G. Jd (2006). Capillary electrophoresis in metabolomics. *TrAC—Trends in Analytical Chemistry*, 25(5), 455–466. Available from <https://doi.org/10.1016/j.trac.2006.02.004>.
- Richter, B. E., Jones, B. A., Ezzell, J. L., Porter, N. L., Avdalovic, N., & Pohl, C. (1996). Accelerated solvent extraction: A technique for sample preparation. *Analytical Chemistry*, 68(6), 1033–1039. Available from <https://doi.org/10.1021/ac9508199>.
- Risticevic, S., Souza-Silva, E. A., Gionfriddo, E., DeEll, J. R., Cochran, J., Hopkins, W. S., & Pawliszyn, J. (2020). Application of in vivo solid phase microextraction (SPME) in capturing metabolome of apple (*Malus × domestica* Borkh.) fruit. *Scientific Reports*, 10(1), 1–11. Available from <https://doi.org/10.1038/s41598-020-63817-8>.
- Shulaev, V., & Isaac, G. (2018). Supercritical fluid chromatography coupled to mass spectrometry— A metabolomics perspective. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 1092, 499–505. Available from <https://doi.org/10.1016/j.jchromb.2018.06.021>.
- Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., & Nishioka, T. (2003). Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *Journal of Proteome Research*, 2(5), 488–494. Available from <https://doi.org/10.1021/pr034020m>.
- Spagou, K., Tsoukali, H., Raikos, N., Gika, H., Wilson, I. D., & Theodoridis, G. (2010). Hydrophilic interaction chromatography coupled to MS for metabolomic/metabolomic studies. *Journal of Separation Science*, 33(6–7), 716–727. Available from <https://doi.org/10.1002/jssc.200900803>.
- Stoll, D. R., & Carr, P. W. (2017). Two-dimensional liquid chromatography: A state of the art tutorial. *Analytical Chemistry*, 89(1), 519–531. Available from <https://doi.org/10.1021/acs.analchem.6b03506>.
- Stein, S. (2012). Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical Chemistry*, 84(17), 7274–7282. Available from <https://doi.org/10.1021/ac301205z>.
- Steinberg, S., Jones, R., Tiffany, C., & Moser, A. (2008). Investigational methods for peroxisomal disorders. *Current Protocols in Human Genetics*, 58(Suppl. 58). Available from <https://doi.org/10.1002/0471142905.hg1706s58>.
- Tranchida, P. Q., Purcaro, G., Dugo, P., & Mondello, L. (2011). Modulators for comprehensive two-dimensional gas chromatography. *TrAC—Trends in Analytical Chemistry*, 30(9), 1437–1461. Available from <https://doi.org/10.1016/j.trac.2011.06.010>.

- Toro-Uribe, S., Montero, L., López-Giraldo, L., Ibáñez, E., & Herrero, M. (2018). Characterization of secondary metabolites from green cocoa beans using focusing-modulated comprehensive two-dimensional liquid chromatography coupled to tandem mass spectrometry. *Analytica Chimica Acta*, 1036, 204–213. Available from <https://doi.org/10.1016/j.aca.2018.06.068>.
- Venkatramani, C. J., Xu, J., & Phillips, J. B. (1996). Separation orthogonality in temperature-programmed comprehensive two-dimensional gas chromatography. *Analytical Chemistry*, 68(9), 1486–1492. Available from <https://doi.org/10.1021/ac951048b>.
- Vuckovic, D., Risticevic, S., & Pawliszyn, J. (2012). *Solid-phase microextraction protocols. Handbook of solid phase microextraction* (pp. 455–478). Elsevier, <https://doi.org/10.1016/B978-0-12-416017-0.00013-9>.
- Wang, C., Wang, S., Fan, G., & Zou, H. (2010). Screening of antinociceptive components in Corydalis yanhusuo W.T. Wang by comprehensive two-dimensional liquid chromatography/tandem mass spectrometry. *Analytical and Bioanalytical Chemistry*, 396(5), 1731–1740. Available from <https://doi.org/10.1007/s00216-009-3409-1>.
- Welthagen, W., Shellie, R. A., Spranger, J., Ristow, M., Zimmermann, R., & Fiehn, O. (2005). Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC × GC-TOF) for high resolution metabolomics: Biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice. *Metabolomics: Official Journal of the Metabolomic Society*, 1(1), 65–73. Available from <https://doi.org/10.1007/s11306-005-1108-2>.
- Wilson, S. R., Jankowski, M., Pepaj, M., Mihailova, A., Boix, F., Vivo Truyols, G., Lundanes, E., & Greibrokk, T. (2007). 2D LC separation and determination of bradykinin in rat muscle tissue dialysate with on-line SPE-HILIC-SPE-RP-MS. *Chromatographia*, 66(7–8), 469–474. Available from <https://doi.org/10.1365/s10337-007-0341-4>.
- Weinert, C. H., Egert, B., & Kulling, S. E. (2015). On the applicability of comprehensive two-dimensional gas chromatography combined with a fast-scanning quadrupole mass spectrometer for untargeted large-scale metabolomics. *Journal of Chromatography A*, 1405, 156–167. Available from <https://doi.org/10.1016/j.chroma.2015.04.011>.
- Winder, C. L., & Dunn, W. B. (2011). Fit-for-purpose quenching and extraction protocols for metabolic profiling of yeast using chromatography-mass spectrometry platforms. *Methods in Molecular Biology*, 759, 225–238. Available from https://doi.org/10.1007/978-1-61779-173-4_14.
- Yang, L., Nie, H., Zhao, F., Song, S., Meng, Y., Bai, Y., & Liu, H. (2020). A novel online two-dimensional supercritical fluid chromatography/reversed phase liquid chromatography–mass spectrometry method for lipid profiling. *Analytical and Bioanalytical Chemistry*, 412(10), 2225–2235. Available from <https://doi.org/10.1007/s00216-019-02242-x>.
- Yin, P., & Xu, G. (2014). Current state-of-the-art of nontargeted metabolomics based on liquid chromatography-mass spectrometry with special emphasis in clinical applications. *Journal of Chromatography A*, 1374, 1–13. Available from <https://doi.org/10.1016/j.chroma.2014.11.050>.

Mass spectrometry in metabolomics

4

Angela Amoresano¹ and Piero Pucci²

¹Department of Chemical Sciences, University of Naples Federico II, Naples, Italy

²CEINGE Advanced Biotechnology, Naples, Italy

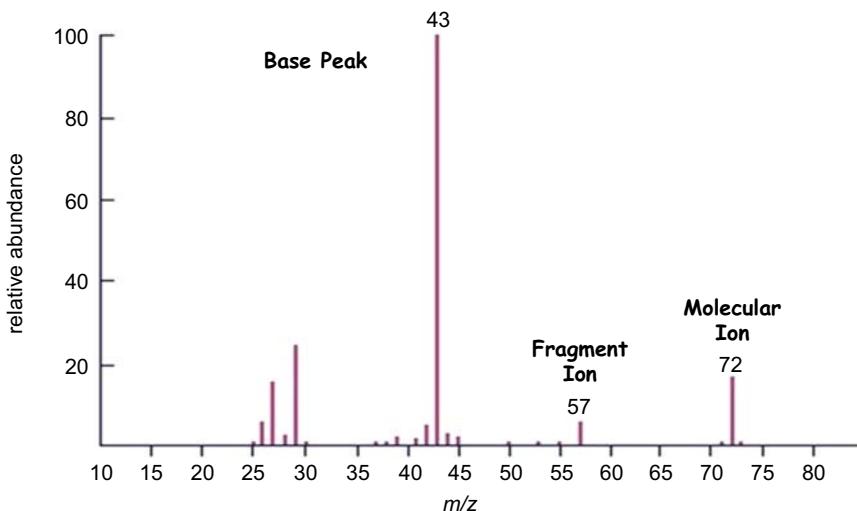
Mass spectrometry

Mass spectrometry (MS) is an analytic technique capable of identifying a wide set of molecules by measuring their molecular weight as the mass-to-charge ratio (m/z) of gaseous ions in electric and magnetic fields. In a mass spectrometer, the ionized charged particles are detected electrically, while in the mass spectrograph, the ions are evaluated by photographic or nonelectrical tools. The instrumentation, now available, generally use electrical detectors; thus, this technique is generally referred to as MS. What MS can do? Unknown compounds can be identified, structure and chemical properties of different classes of analytes can be investigated and target compounds can be identified and quantified even in complex mixtures. Such different goals can be pursued by using different types of mass spectrometers designed by combining sources and analyzers with specialized characteristics (Awad et al., 2015). The high resolution, accuracy, sensitivity, and the possibility to carry out multiple analyses led to an enormous spreading of MS, now extensively used in an ever-increasing number of applications in different fields of science and technology from clinics, health, forensics to geology, environment, industry and foods (Baidoo & Benites, 2019).

Mass spectrum

The result of a mass spectral analysis is a graph, called mass spectrum, reporting the relative abundance of ions versus their m/z . The vertical signals occurring in the spectrum, represent the analytes having a specific m/z value and the length of each bar represents the relative abundance of the ion (Fig. 4.1).

The base peak is the most intense ion whose abundance is arbitrary fixed at 100%. With the exception of few cases, the ions showed a single charge; in this case the mass of the analytes corresponds to the detected m/z value. In the mass spectrum of a pure analyte, the highest-mass ion represents the molecular ion,

**FIGURE 4.1**

Mass spectrum. An example of mass spectrum. The characteristic ions are indicated.

corresponding to the intact molecule ionized in the gas phase, while the other ions occurring at lower mass value are fragments generated from the molecular ion.

Isotopes

Most elements consist of two or more types of atoms that have the same atomic number but differ for the numbers of neutrons, named isotopes. All isotopes of a given element have the same chemical properties but different atomic masses. This suggests that rather than one single peak, in the mass spectrum any analyte originates a cluster of signals defined as the isotope pattern of the molecule. The mass number is denoted within the upper left corner of an atom, like ^{12}C for the carbon 12 isotope containing 6 protons and 6 neutrons. Several isotopes of a component are often found in nature and are defined as natural isotopes. The natural isotope with the lowest mass, like ^1H , ^{12}C , ^{14}N , ^{16}O , ^{31}P and ^{32}S is referred to as monoisotopic. The mass spectrometer is able to separate and detect ions of slightly different masses and then to distinguish the different isotopes of a given element (Table 4.1).

The mass spectrometer is able to measure either the monoisotopic mass or the average mass of the analytes. The monoisotopic mass of a molecule corresponds to the sum of the masses of the constituting atoms taking into account only the most abundant isotope. The molecular ion corresponding to the monoisotopic mass of a compound is then an homogeneous species formed by all the molecules that are only composed by the most abundant isotopes of the constituting element.

Table 4.1 Isotopic abundance: relative abundance isotopes and their mass in Dalton for the six elements most abundant in biomolecules.

Element	Isotope	Abundance (%)	Mass (Da)	Average mass (Da)
Hydrogen	1H	99.988	1,007825	1,00794
	2H	0.012	2,014102	
Carbon	12C	98.93	12	12,0107
	13C	1.07	13,003355	
Nitrogen	14N	99.636	14,003074	14,0067
	15N	0.364	15,00109	
Oxygen	16O	99.757	15,994915	15,9994
	17O	0.038	16,999131	
	18O	0.205	17,99916	
Phosphor	31P	100	30,9737762	30,973762
Sulfur	32S	94.99	31,972071	32,065
	33S	0.75	32,971459	
	34S	4.25	33,967867	
	36S	0.01	35,967081	

When isotopes are clearly resolved, the monoisotopic mass is used as it is the most accurate measurement. However, when the molecular weight of the compound is increasing, the probability to find molecules only composed of the most abundant isotopes decreases up to zero. When this limit is achieved, the mass spectrometer will measure the average mass of the analytes. The average mass (or atomic weight) of a component is a weighted average of all of the isotopes occurring in that compound, in which the mass of each isotope is multiplied by the natural abundance of that particular isotope.

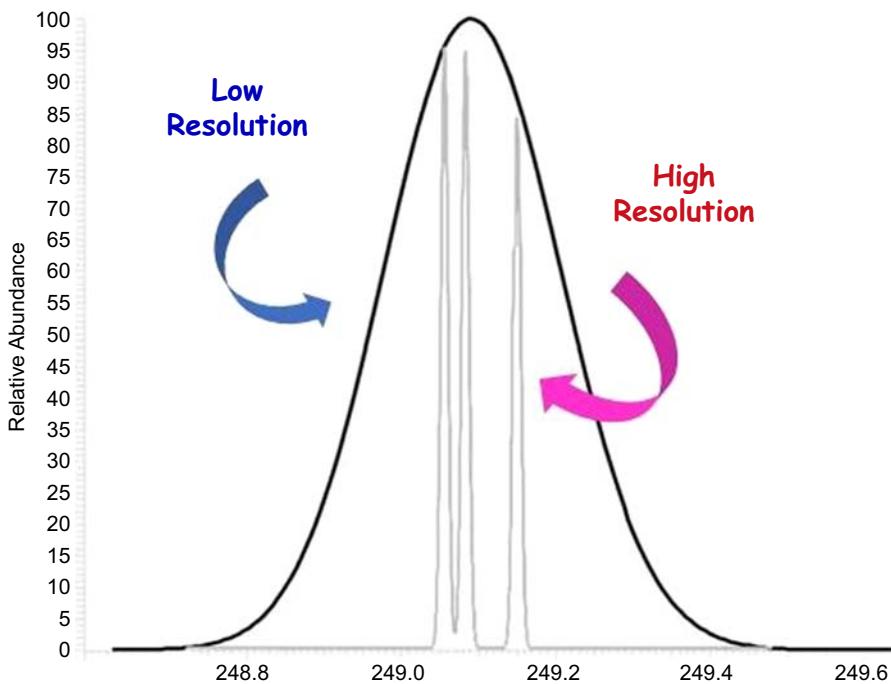
Resolution and accuracy

The overall performances of a mass spectrometer can be measured by several parameters including resolution, precision, accuracy, and sensitivity. Resolution is defined as the ability of the instrument to separate two adjacent peaks. Resolution can be calculated by the following formula:

$$R = m/\Delta m \quad (4.1)$$

where Δm is the mass difference between two adjacent resolved peaks and m is the nominal mass of the first peak (Fig. 4.2). This equation shows that resolution in mass cannot be considered a fixed value, but it depends on the mass range under investigation. As an example, to separate two peaks with $\Delta m = 0.1$, R must be 4000 when $m = 400$, but resolution should increase to 20,000 if $m = 2000$.

Modern mass spectrometers are able to distinguish (resolve) ions differing by few decimals of Dalton without difficulty. However, resolution is highly

**FIGURE 4.2**

Mass resolution. Mass spectrum peak resolution.

dependent upon the specific mass spectrometer. Quadrupole mass spectrometers are normally considered low resolution instruments while high resolution mass spectrometers are usually equipped with orbitrap analyzers (see below).

Precision refers to the reproducibility of mass measurements of a specific ion reflecting random errors. Random errors cause measurements to fall on either side of the average experimental measurement and affect the precision of the set of measurements. When a set of mass measurements of one ion species lie close together, we say the measurements are precise. Precision is commonly expressed in terms of the coefficient of variation (relative standard deviation) for a series of measurements on a single sample (internal precision) or as the minimum detectable difference in isotope ratio between a pair of samples introduced sequentially from a dual-inlet system. Mass accuracy is defined as the proximity of the experimental measurements to the true value, the exact or theoretical mass, of the analyte. Accuracy is very often expressed as the error associated with mass measurements. The error is defined by the following formula:

$$\frac{\text{experimental mass} - \text{true mass}}{\text{true mass}} \times 10^6 \quad (4.2)$$

and is expressed in parts per million (ppm). The accuracy of mass measurement directly determines the usefulness of mass spectrometric experiments, and much effort in instrumentation development is directed at improving this key parameter. Today, Time of Flight (TOF) instruments equipped with energy correcting reflectrons can reach low ppm values. Moreover, orbitrap mass spectrometers reduce the mass measurement to a frequency measurement and are therefore potentially capable of exceedingly high mass accuracy. Mass accuracy and mass resolution are connected, and instruments introduced during the last decades radically improved in these two attributes.

The sensitivity of a mass spectrometer may be described as the intensity of the signal-to-noise ratio (S/N) recorded for a fixed concentration of the analyte. The limit of detection is determined from the analyte S/N and is the lowest concentration of a substance where its signal can be distinguished from system noise. Several factors can affect sensitivity including the effectiveness of producing gas-phase ions from analytes in solution (ionization efficiency), the ability to transfer them to the analyzer (transmission efficiency) and the specific type of analyzer used (Iwasaki et al., 2011).

Mass spectrometer

The design of a mass spectrometer includes four essential associated components: the system of sample introduction, the ion source, the mass analyzer and the detector (Fig. 4.3). Compounds to be analyzed are usually ionized within the source, then sorted and separated on the basis of their mass/charge ratio within the analyzer and finally measured and recorded within the mass spectrum by the detectors. Nowadays, a number of differently designed mass spectrometers are available, all sharing the potential to measure mass-to-charge values of ions, although the principles of operation and therefore the sort of experiments that can be performed on these instruments differ greatly.



FIGURE 4.3

Schematic representation of a mass spectrometer. Working process of a mass spectrometer, from sample introduction system to the mass spectrum display.

System for sample introduction

The analytes to be investigated can be introduced into the mass spectrometer by essentially three systems depending on the chemical nature of the sample. Direct introduction constitutes the simplest sample introduction method. The analyte is introduced directly into the source region of the mass spectrometer through a needle valve either in the gas phase or using a short capillary tube for liquid samples. When the sample is volatile or can be made volatile, Gas Chromatography (GC) can be used for introducing the analytes into a mass spectrometer. Complex mixtures are routinely separated by GC and MS is used to identify and quantitate the individual components. Analogously, Liquid Chromatography (LC) methods are used to introduce soluble, nonvolatile or thermally labile compounds into the mass spectrometer. In this case the sample is ionized directly from the condensed phase.

Ion sources

The ion source produces gas phase ions from the sample allowing them to fly within the mass spectrometer and to be manipulated by external electric and/or magnetic fields before reaching the detector ([Bhardwaj & Hanley, 2014](#)).

Mass analyzer

A mass analyzer is the component of the mass spectrometer that separates ions based on their mass to charge ratios driving them to the detector. A variety of mass analyzers are currently available, each of which has its specific mode of operation and is characterized by specific features including speed of operation, resolution, mass accuracy and sensitivity. Moreover mass spectrometers can be equipped with a combination of more than one mass analyzer to allow specific experiments to be carried out.

Ion detector

Detectors play an important role in mass spectrometer for measuring the separated charged ions and should be endowed with desirable properties including high amplification, fast time response, low noise, high collection efficiency. Electron multiplier is typically used in any type of MS as it is able to provide the considerable amplification of the signals needed to detect the quite small number of ions leaving the mass analyzer. Faraday cups and ion-to-photon detectors can also be used for specific applications.

The different types of mass spectrometers are often distinguished by their specific combination of ion source and mass analyzer ([Xia & McLuckey, 2008](#)). The most widespread ionization methods in biochemical analyses are Electron Impact (EI), Electrospray Ionization (ESI) and Matrix Assisted Laser Desorption

Ionization (MALDI). Mass analyzers as quadrupoles (Q), ion traps (IT), time-of-flight (TOF), Orbitrap, or combination of these in “hybrid instruments,” are commonly used in the biochemical field because of their good resolution and sensitivity. Coupling of GC or LC and tandem Mass Spectrometry (MS/MS) methods are also widely used for the analysis of complex mixtures or for quantitative measurements (Byliński et al., 2017).

Ion sources

EI ion source

Electron ionization (EI, formerly referred to as Electron Impact) is an ionization method in which energetic electrons interact with gas phase molecules to produce ions (Fig. 4.4). EI was one of the first ionization techniques developed for MS and is still popular today. This technique is considered a hard ionization method, since it uses highly energetic electrons to produce ions. This leads to extensive fragmentation, which contributes to the structural determination of unknown compounds, but originates complex spectra when a mixture of compounds is analyzed. EI is mainly addressed to the analysis of small, thermally stable and volatile organic compounds and can be very efficiently coupled to GC as this analysis is in perfect match with the electron ionization conditions.

The EI source usually consists of a cathode (filament), an ion chamber, an electron receiver (anode), the ion repeller and the focus lenses. Under high vacuum conditions, a current is applied to the filament to produce electrons that are

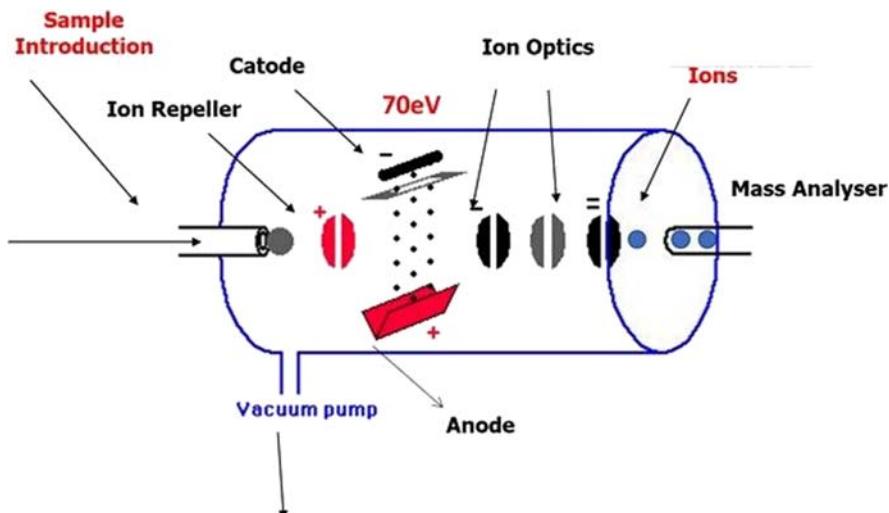


FIGURE 4.4

Electron ionization source. Schematic of electron ionization source.

accelerated from the filament to the anode. During this process, the sample introduced in the gas phase is converted to ions by collision with the high energy electron beam. The collision of electrons with a molecule can generate ions according to the following reactions:



where the molecule (M) loses one electron producing a radical cation and



where the molecule (M) adsorbs one electron producing a radical anion. Both ions are normally produced during the ionization process and the analysis can be carried out either in positive or negative mode by appropriately selecting the ion repeller potential. Since the charge of the molecular ion (parent ion) formed during these processes is $+1$, the measured m/z ratio exactly corresponds to the molecular mass of the compound. However, the parent ion has very high internal energy and usually tends to fragments within the ion source by breaking a single covalent bond. Fragmentation statistically occurs on many covalent bonds originating a mixture of molecular ion and fragments that are accelerated outside the ion source and analyzed by the mass spectrometer. Therefore the mass spectrum usually shows several peaks corresponding to the molecular ion and the fragments. It should be underlined that during fragmentation one fragment retains the charge whether the other is neutral and the mass spectrometer can only detect the charged species.

The fragmentation process depends on the stability of the compound and its molecular structure leading to a specific fragmentation pattern that constitutes a sort of fingerprint of the molecule. Unknown molecules can then be identified by comparing their fragmentation spectra to specific databases containing the fragmentation pattern of a multitude of compounds (Wang et al., 2018).

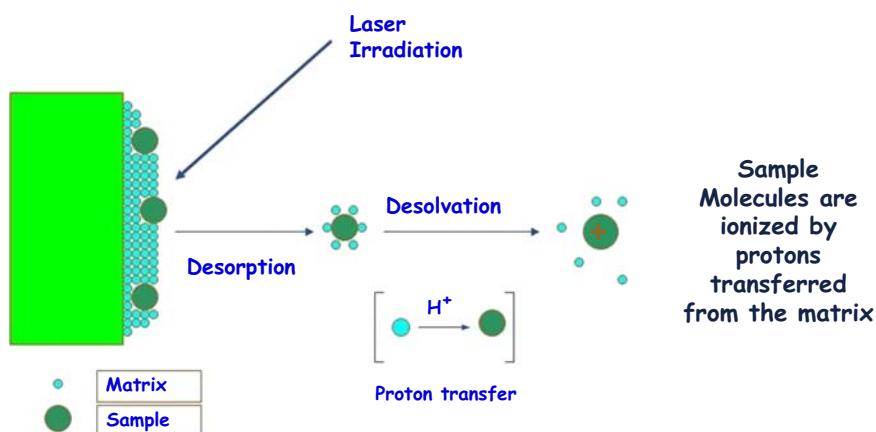
Matrix-assisted laser desorption ionization ion source

MALDI source generally uses a laser light irradiation (pulsed nitrogen laser beam at 337 nm) to induce sample ionization (Fig. 4.5).

The sample is pre-mixed on a metal target with a highly absorbing matrix, consisting of small aromatic molecules, that upon drying, co-crystallizes with the sample. The most frequently used matrices are aromatic organic acids that absorb in the region of the laser wavelength (Table 4.2).

Several functions are accomplished by the matrix:

1. Isolate the analyte molecules from one another such that the incident laser radiation primarily hits the matrix, rather than the analyte
2. Provide a source of protons (H^+ ions) to ionize the analyte molecules
3. Absorb the ultraviolet light, thus converting the incident laser energy into molecular electronic energy, which may be used both for desorption and ionization.

**FIGURE 4.5**

MALDI ionization source. Schematic representation of MALDI ionization source.

Table 4.2 MALDI-MS matrices commonly used for MALDI-MS analyses.

Compound	Acronym	Application to
Picolinic acid	PA	Oligonucleotides, DNA
3-Hydroxypicolinic acid	HPA, 3-HPA	Oligonucleotides, DNA
2,5-Dihydroxybenzoic acid	DHB	Proteins, oligosaccharides
α -Cyano-4-hydroxycinnamic acid	α -CHCA, 4-HCCA, CHCA	Peptides, smaller proteins, triacylglycerols, numerous other compounds
4-Chloro- α -cyano-cinnamic acid	CICCA	Peptides
3,5-Dimethoxy-4-hydroxycinnamic acid	SA	Proteins
2-(4-Hydroxyphenylazo)benzoic acid	HABA	Peptides, proteins, glycoproteins, polystyrene
2,6-Dihydroxyacetophenone	DHAP	Glycopeptides, phosphopeptides, proteins
2,4,6-Trihydroxyacetophenone	THAP	Solid-supported oligonucleotides

The ionization process takes place within the ion source under high vacuum and a large electrical field between the target and the extraction plates. Energy deposition from the laser beam into the matrix molecules induces sublimation of the matrix crystals and subsequent expansion into the gas phase. Following desorption, the sample is either protonated or deprotonated by the matrix according to its acid-base properties. Protonated molecular ions ($M + H$)⁺ are detected in positive ion mode, while deprotonated molecular ions ($M - H$)⁻ can be measured in negative ionization mode.

Since the ionization process is strongly dependent by the sample-matrix pair, specific matrices have been developed for the analysis of different compounds. Table 4.2 shows the most common MALDI matrices used in biomolecular application (Brown & Lennon, 1995; Wolk & Clark, 2018).

The MALDI source is quite tolerant to contaminants and the ionization energy is mild originating stable molecular ions without any tendency to fragment. Interpretation of the MALDI spectrum is then very simple as it essentially consists of molecular ions. MALDI-MS is then very useful in the analysis of very complex mixtures as any component will only give raise to its corresponding molecular ion (Vestal & Campbell, 2005).

Electrospray ion source

ESI ionization is obtained when a high voltage is applied to a sample in solution to create an aerosol (Fig. 4.6). The solution containing the sample is introduced into the ion source by a needle from which the sample emerges forming a dispersed spray of highly charged droplets. Solvent evaporation is facilitate by the assistance of a drying gas causing charged droplets to diminish in size to a limit when the surface tension of the droplets is not able to match the charge repulsion leading to Coulomb explosion and to ions release. This limit is called the Rayleigh Limit and is defined as the minimum drop size at which the Coulomb explosion occurs. Ions are then directed towards the mass spectrometer with the aid of a co-axially nebulizing gas.

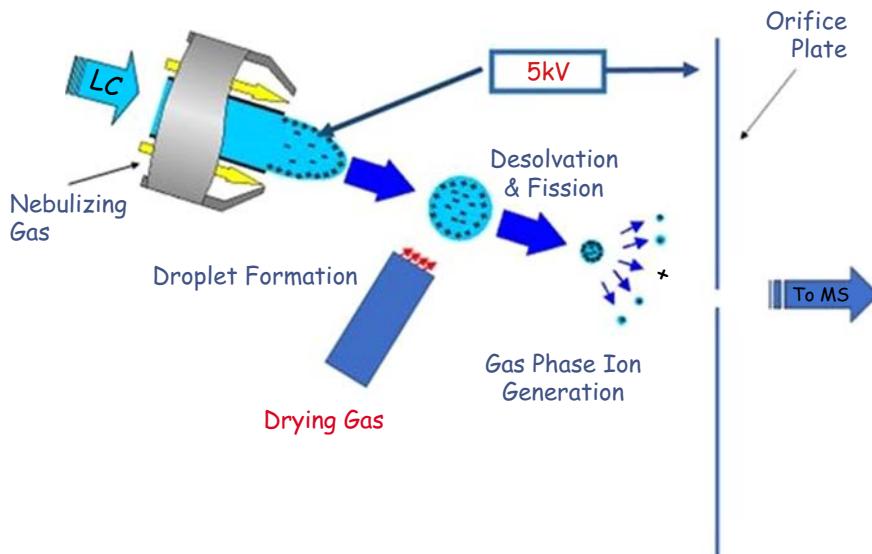


FIGURE 4.6

Electrospray ionization source. Schematic of electrospray ionization source.

ESI is a soft ionization technique, since the sample is ionized at low energy conditions preventing in source fragmentation. A specific feature of the ES ionization consists in the possible production of multiple-charged ions, that is a single molecule might generate several mass signals all of which share the same mass but different charge state. This makes ES ionization not amenable to the analysis of complex mixture unless a chromatographic pre-fractionation of the sample is introduced leading to the procedure called LC-MS. However, the occurrence of multiply charged ions greatly extends the mass range of the analysis as the mass spectrometer will measure the mass to charge ratio, making ES very useful for the analysis of biological macromolecules (Whitehouse et al., 1985).

An excellent enhancement of ESI ion sources has come from the reduction of the flow of the liquid needed to create the spray to micro- or nano-scale level. This device results in a higher ionization efficiency because the charge density at the Rayleigh limit increases significantly with decreasing droplet size. Moreover, the utilization of micro or nano flow results in a higher sensitivity of the analyses as the concentration of the analyte within the solvent droplets increases (Van Berkel, 2003).

Mass analyzers

Following ionization, the analytes fly into the second region of the mass spectrometer, the mass analyzer, whose main function is devoted to the separation of the ions on the basis of their m/z values (Haag, 2016). The specific features of mass analyzers include:

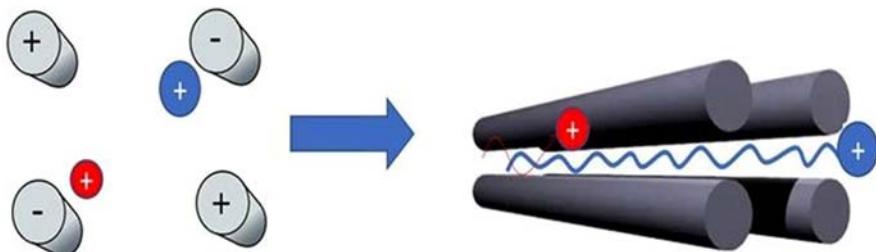
- mass range: interval of m/z values that the analyzer is able to monitor
- resolution power: the ability to distinguish between two ions with similar m/z ratios
- transmission efficiency: the percentage of selected ions that travel through the analyzer and reach the detector without being deflected.

There are several types of mass analyzers currently used for the analysis of biomolecules including Quadrupole (Q), TOF, IT (either geometrical or linear) and Orbitrap or combination of these.

Quadrupole mass analyzer

Quadrupole mass analyzer consists of 4 cylindrical or elliptical rods, set parallel to each other (Dawson, 2013). Each opposing rod pair is electrically connected, and a radio frequency (RF) voltage with a DC offset voltage is applied between one pair of rods and the other with opposite rods having potentials of the same sign (Fig. 4.7).

Ions travel down the quadrupole between the rods. Voltages and RF applied to the rods affect trajectory of ions. Only ions of a certain m/z will have stable trajectories reaching the detector for a given ratio of voltages, other ions have unstable trajectories and will collide with the rods. This permits the complete

**FIGURE 4.7**

Quadrupole. Schematic representation of a quadrupole.

scan of a range of m/z values by continuously varying the applied voltage. Alternatively, selection of a single ion with a particular m/z can also be possible by simply fixing the voltages to a specific value (Select Ion Monitoring, SIM).

Time of flight mass analyzer

The TOF mass analyzer consists of a flying tube under high vacuum in which ions are accelerated by an electric field of known strength (Brown & Lennon, 1995). This acceleration results in all ions with the same charge having the same kinetic energy. Their velocity then depends on the mass and separation is achieved on the principle that higher the mass of the ion, lower its velocity. The time that an ion takes to reach the detector flying along the tube is measured. This time will depend on the velocity of the ion, and therefore on its m/z ratio that can then be determined.

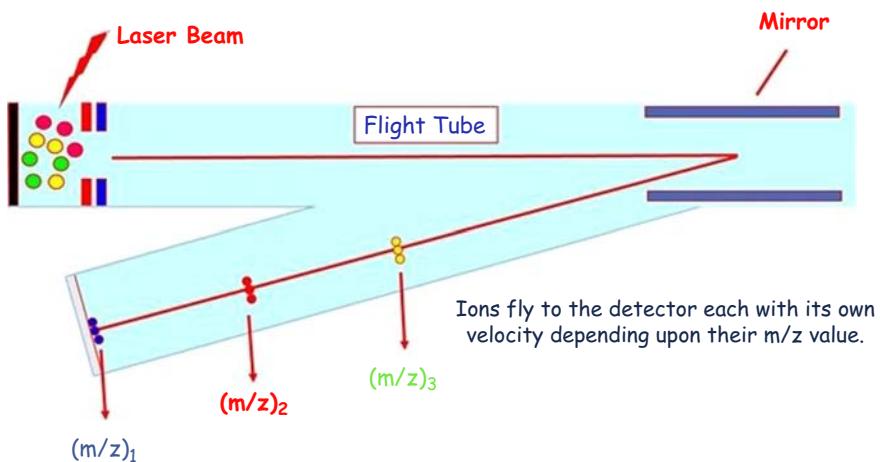
Mass resolution of TOF analyzers has been improved by the “Delay Extraction” procedure to compensate for the spread of initial kinetic energies among ions and by the introduction of an electrostatic mirror (reflectron). The delay extraction voltage is applied to the ions produced during the desorption/ionization process for approximately 100 ns or less allowing the initial burst of ions and neutrals produced by the laser pulse to equilibrate, lining up all together before they are accelerated into the flight tube.

The electrostatic mirror was introduced to increase the length of the ion path and to correct small discrepancies in distribution of kinetic energy among ions with identical m/z values thus increasing the resolving power of the TOF. The reflectron uses a constant electrostatic field to reflect the ion beam toward the detector (Fig. 4.8).

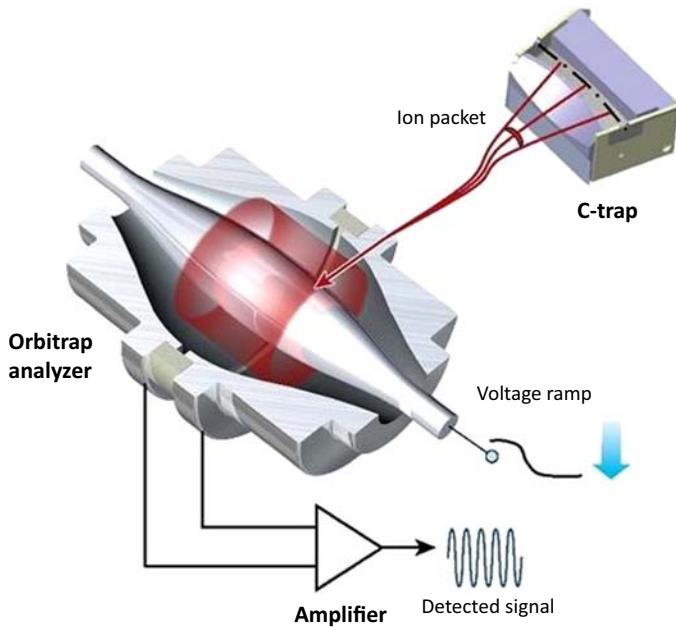
The more energetic ions penetrate deeper into the reflectron taking a slightly longer path to the detector. Less energetic ions of the same m/z penetrate a shorter distance into the reflectron and, correspondingly, take a shorter path to the detector (Debois et al., 2012).

Orbitrap mass analyzer

The Orbitrap is a high-performance mass analyzer that traps ions in the electrostatic field produced by two electrodes: a coaxial inner spindle-like electrode and an outer barrel-like electrode (Fig. 4.9).

**FIGURE 4.8**

Time of flight analyzer. Schematic representation of TOF analyzer.

**FIGURE 4.9**

Orbitrap mass analyzer. Cross section of the C-trap ion accumulation device and the Orbitrap mass analyzer with an example of an ion trajectory. During the voltage ramp, the ion packets enter the Orbitrap mass analyzer forming rings that induce current, which is detected by the amplifier (<https://mass-spec.chem.ufl.edu>).

From <http://mass-spec.chem.ufl.edu>.

Ions are trapped in a harmonic orbital motion around the spindle shuttling back and forth over its long axis in periodic motion with frequencies dependent only on their m/z values (Makarov, 2000). The image current induced from the axial oscillating ions on the outer electrode is detected and after amplification is converted into a frequency spectrum using a Fourier transform algorithm. The frequency of oscillation is related to the mass/charge ratio as illustrated by the following equation:

$$\omega_z = \sqrt{\frac{k}{m/z}} \quad (4.5)$$

All the ions are detected simultaneously over some given period of time and resolution can be improved by increasing the strength of the field or the detection period.

Nowadays, several instruments equipped with a orbitrap mass analyzed have been developed including Linear trap quadrupole Orbitrap (LTQ), Q Exactive, and Orbitrap Fusion mass spectrometers. The LTQ Orbitrap was the first instrument including an Orbitrap mass analyzer to be developed. The design of this mass spectrometer consists of an IT followed by the Orbitrap mass analyzer (Fig. 4.10). Within this particular device, the ions are first stored in the Linear IT then axially ejected and accumulated in the C-Trap. Finally, a small cloud of squeezed ions are axially injected into the Orbitrap performing axial oscillation and producing the frequency spectrum. Both MS and MS_n spectra might be recorded with this instrument by using either the Orbitrap analyzer for highest resolution and mass accuracy or the IT analyzer for highest speed and sensitivity. The most common operation mode consists in the acquisition of full scan spectra within the Orbitrap analyzer and data-dependent MS/MS scans within the IT analyzer. This scan mode takes full advantage of resolution and mass accuracy of the

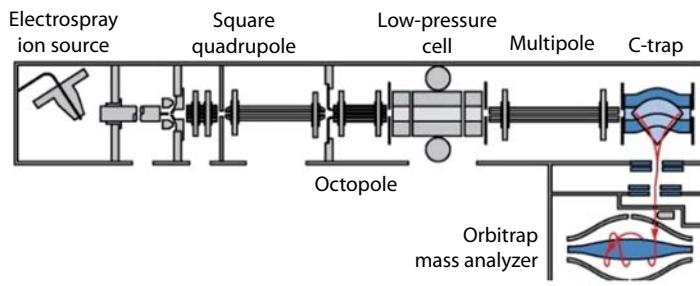


FIGURE 4.10

Linear trap quadrupole Orbitrap. Schematic of the LTQ Orbitrap mass spectrometer with traditional ion trap followed by Orbitrap mass spectrometer architecture (<http://mass-spec.chem.ufl.edu>)

From <http://mass-spec.chem.ufl.edu>.

Orbitrap for the detection of precursor ions even in complex mixture and the speed and sensitivity for MS/MS spectra of the IT device.

Despite the good analytical performances, LTQ Orbitrap has some limitations as the sole fragmentation method available is the Collision Induced Dissociation (CID) within the IT. The instrument was then implemented with additional fragmentation techniques including Higher Energy Collision-Induced Dissociation (HCD) and the newly developed Electron Transfer Dissociation (ETD).

The subsequently developed Q Exactive instrument was designed with a quadrupole mass filter located in front of the Orbitrap analyzer (Fig. 4.11). Isolation of precursor ions is normally carried out in the quadrupole whereas the Orbitrap analyzer is used for acquiring both full scan and MS/MS spectra, using an HCD cell for fragmentation.

The Orbitrap Fusion instrument, incorporating three mass analyzers namely a quadrupole, an Orbitrap and a linear IT, combines advanced IT Orbitrap hybrid technology with the quadrupole Orbitrap hybrid systems (Fig. 4.12). This architecture enabled significant performance improvements. The quadrupole is used to select ions for subsequent fragmentation in MS/MS. The ion path is set up in such a way that selected ions can be directed to either the Orbitrap or the linear IT. Precursor fragmentation can take place with a variety of fragmentation mechanisms in the ion-routing multipole (HCD), in the IT (CID, ETD) or both (ETCD) with fragment detection in the linear IT or Orbitrap mass analyzers at any stage of MSn analysis increasing the achievable amount of structural information (Nikolaev et al., 2016; Perry et al., 2008).

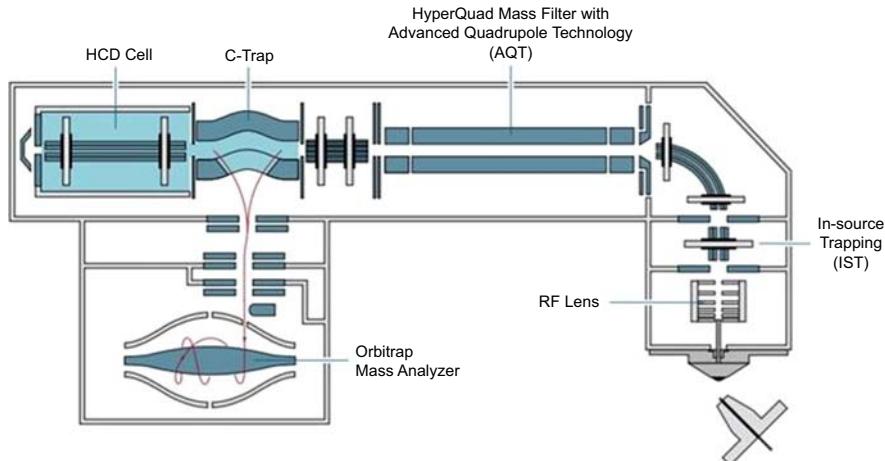
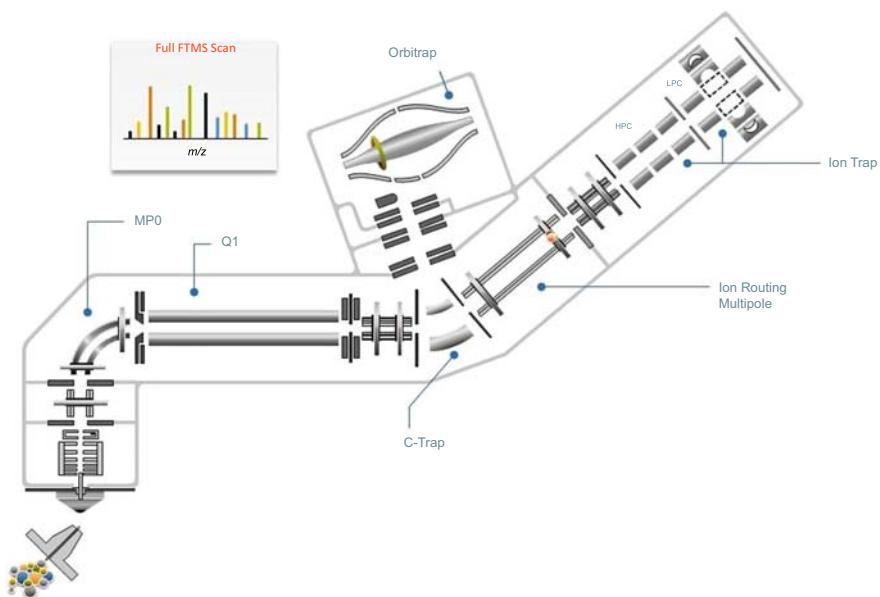


FIGURE 4.11

Quadrupole/Orbitrap hybrid mass spectrometer. Quadrupole/Orbitrap hybrid mass spectrometer Q Exactive architecture. Abbreviation: HCD, higher energy collision-induced dissociation (<http://www.thermofisher.com>).

From <http://www.thermofisher.com>.

**FIGURE 4.12**

Orbitrap fusion. Schematic of the Orbitrap Fusion a Quadrupole/Orbitrap/ion trap hybrid mass spectrometer (<http://www.thermofisher.com>).

From <http://www.thermofisher.com>.

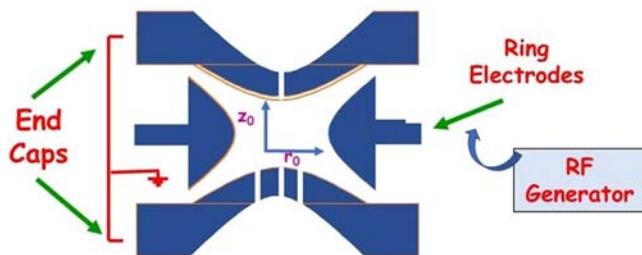
The instrument architecture facilitates the rapid execution of complex experiments and is ideal for the analysis of complex analytes mixtures for its characteristic ability to concurrently isolate ions with one analyzer and separately detect ions within the two remaining analyzers (Zhang et al., 2005; Zubarev & Makarov, 2013).

Quadrupole ion trap

The evolution of the quadrupole resulted in the development of the quadrupole IT also called “geometrical” or 3D IT. The IT consists of three electrodes, a hoop electrode and two hemispherical end caps electrodes generating dynamic electric fields which are employed to trap ions within a small volume (Fig. 4.13).

Ions are dynamically stored within the IT traveling on stable paths. The path of a specific ion can be altered by changing the voltage or the RF. When the path becomes unstable the ion is ejected from the trap and can be detected. The full scan spectra can be obtained by scanning the voltage and RF potentials to eject ions of successive m/z ratios from the trap into the detector.

The ability to trap and accumulate ions constitutes the main advantages of the ion-trap mass spectrometer that shows very high efficiency, increased sensitivity and extended S/N ratio of the measurements. Moreover, IT can also be utilized in

**FIGURE 4.13**

Ion trap mass analyzer. Schematic representation of a tridimensional ion trap.

MS/MS analyses as they can isolate precursor ions and provide fragmentation experiments (see below).

An evolution of 3D IT consists in the Linear IT that uses a set of quadrupole rods to confine ions radially and a static electrical potential on-end electrodes to confine the ions axially. The linear form of the trap can be used as a selective mass filter like a quadrupole, or as an actual trap by creating a potential wall for the ions along the axis of the electrodes. Advantages of the linear trap design are increased ion storage capacity, faster scan times, and simplicity of construction (McLuckey et al., 1994).

Tandem mass spectrometry

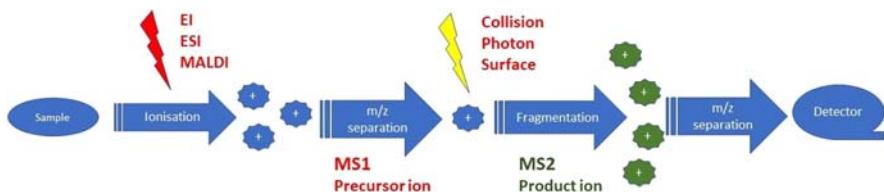
MS/MS is employed to provide structural information on the molecules under investigation by controlled fragmentation of their specific molecular ions within the mass spectrometer and identification of the fragment ions (Fig. 4.14).

MS/MS also enables specific compounds to be detected in complex mixtures due to their characteristic fragmentation patterns. MS/MS is performed by a two stages analysis; first the molecules of a given sample are ionized and the specific ion of interest is isolated on the basis of its m/z -ratio. The selected ion is then introduced into the collision cell and fragmented by collision-induced dissociation, ion-molecule reaction, or photodissociation originating a number of fragment ions which in turn are separated by their m/z -ratio and detected.

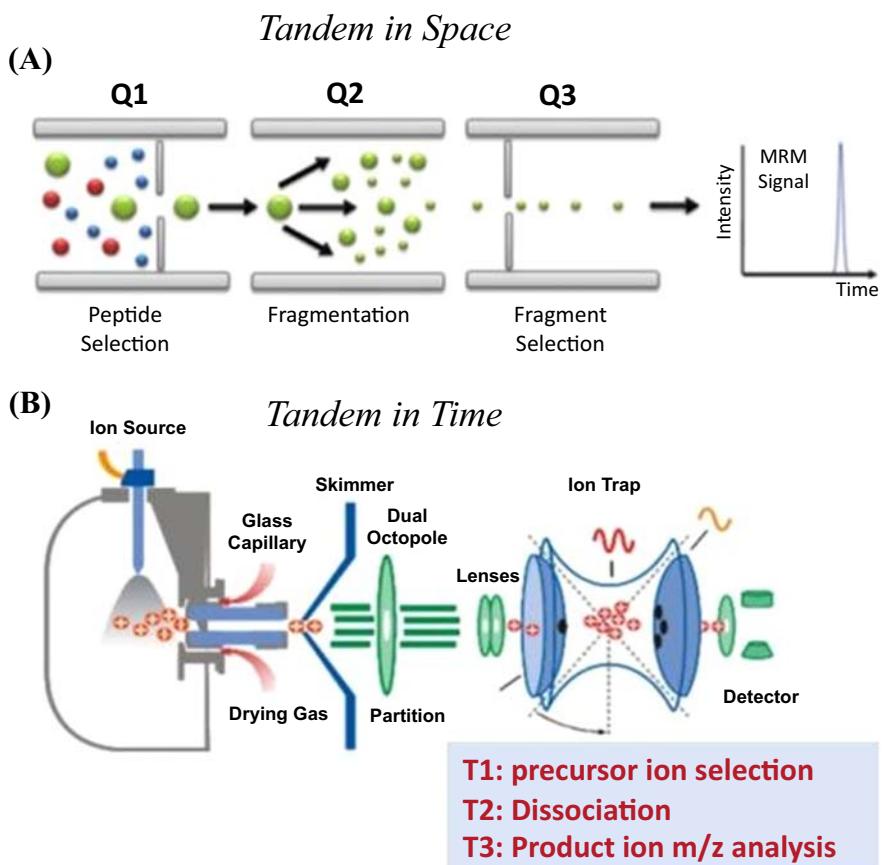
The fragmentation step makes it possible to identify and separate ions that have very similar m/z -ratios in regular mass spectrometers.

Instruments for tandem mass spectrometry analysis

MS/MS experiments can be design according to two different approaches, tandem in space or tandem in time, essentially depending on the instruments available (Fig. 4.15).

**FIGURE 4.14**

Tandem mass spectrometry. Schematic of tandem mass spectrometry.

**FIGURE 4.15**

Tandem mass spectrometer. Scheme of tandem in space (A) and in time (B) mass spectrometer (<http://www.mrmatlas.org>, <http://www.slideserve.com>).

From <http://www.mrmatlas.org>.

In the “Tandem in space” procedure, ion selection, ion fragmentation and fragments analysis, occur in three different regions of the mass spectrometer meanwhile the ion beam is flying from the ion source to the detector. Instruments to perform “Tandem in space” experiments are equipped with two or more mass analyzers coupled together; the parent ion of interest is isolated by the first analyzer (MS 1), fragmented in the collision cell and the fragments [product ions (PIs)] are analyzed by the second analyzer (MS 2). Typical instruments used for the “Tandem in space” approach are Triple Quadrupole or TOF-TOF mass spectrometers where the two analyzers in series are identical and hybrid instruments like Q-TOF equipped with two different analyzers.

In the “Tandem in time” approach, the three steps that is isolation, fragmentation and analysis, occur within the same region of the mass spectrometer but at different times. Trapping mass spectrometers either quadrupole IT or Linear IT instruments are normally used for the “tandem in time” procedure.

The ion of interest is first isolated in the IT by ejecting all the others outside the trap. The ion is then fragmented and the PIs are analyzed by scanning the voltage and RF potentials.

Tandem mass spectrometry scan modes

When MS/MS in space is performed with a triple quadrupole instruments, a number of different scan modes can be employed each of which has its own applications and provides different information. The four main scan possible experiments using MS/MS with an in-space design are PI scan, precursor ion scan (PIS), neutral loss scan (NL) and multiple reaction monitoring (MRM).

Product ion scan

The precursor ion with a specific m/z ratio is selected by the first quadrupole (Q1), introduced into the collision cell (Q2) and fragmented. The resulting fragment ions (or PIs) are then separated by scanning the second quadrupole (Q3) and detected providing the MS/MS spectrum containing several structural information on the analyte under examination.

Precursor ion scan

The first quadrupole (Q1) operates in a full scan mode to detect all precursor ions that upon fragmentation in the collision cell (Q2) generate a selected fragment with a specific m/z ratio. The second quadrupole (Q3) is set at a fixed value corresponding to the m/z ratio of the selected fragment ion. This scan mode is usually employed to monitor a specific set of molecules within a complex mixture that contain a common functional group that can be released upon fragmentation.

Neutral loss scan

In the constant NL scan both the Q1 and Q3 analyzers are scanned simultaneously and collect data across the entire m/z range. However, the second analyzer (Q3)

is scanned with a mass offset from the first mass analyzer. This offset correlates with the specific mass value of a neutral fragment that is released from the ions transmitted through Q1 upon fragmentation in the collision cell (Q2). This scan mode is commonly used in the selective identification of closely related class of compounds in a mixture as it is able to detect all precursors that undergo the loss of a specified common neutral molecule.

Multiple reaction monitoring

MRM is a highly specific and sensitive MS technique that can selectively quantify known compounds within complex mixtures. Using an MRM data acquisition method targeted experiments addressed to the identification and quantification of one or more specific molecules can be designed. In this approach the triple quadrupole MS firstly targets the ion(s) corresponding to the compound(s) of interest with subsequent fragmentation of the target ion(s) to produce a range of daughter ions. One (or more) of these fragment daughter ions can be selected for quantitation purposes.

Before starting MRM experiments, it is critical to design mass transitions from the precursor ion and its corresponding PIs for the target compound(s). Based on the predesigned transition lists, the precursor ion is preselected in the first quadrupole (Q1) and transmitted to the collision cell (Q2) for fragmentation. The resultant fragment (product) ions will be driven to the second quadrupole (Q3), which is set to detect only the predefined PIs (Fig. 4.16).

Only compounds that meet both these criteria, that is, specific parent ion and specific daughter ions corresponding to the mass of the molecule of interest are isolated within the mass spectrometer. Because of the two selection steps, the specificity of MRM is ensured. By ignoring all other ions that flow into the mass spectrometer the experiment gains sensitivity, whilst maintaining exquisite accuracy.

The MRM mode constitutes a sort of a double mass filter that drastically reduces the noise and increases sensitivity and selectivity. In contrast to standard full scan studies, MRM measurements are quantitative analyses strictly targeting a predetermined set of molecules and depend upon specific MRM transitions for the targeted species. Moreover, triple quadrupole systems allow the detection of

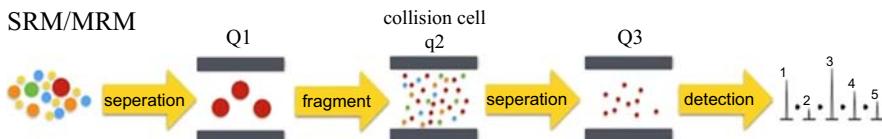


FIGURE 4.16

Selected reaction monitoring experiment. Diagram showing the selected reaction monitoring experiment. In this triple quad, Q1 and Q3 act like a mass filter whereas Q2 acts as a collision cell for selected peptide ion (<http://www.thermofisher.com>).

From <http://www.thermofisher.com>.

many MRM transitions enabling the quantitation of many targeted analytes in a single experiment. When necessary, additional MRM transitions can be included to increase selectivity in the identification and quantification of very similar compounds in complex mixtures. Finally, MRM methods offer both absolute structural specificity for the analyte and relative or absolute quantification of analyte when stable, isotopically labeled standards are added to a sample in known amount.

Since nowadays most chemical analysis needs high sensitivity, each MRM transition is maximized by optimal tuning of the acquisition parameters of the mass spectrometer. The maximum sensitivity in MRM mode depends on the ionization efficiency of the compounds, the transfer into the analyzer and the dissociation into intense fragments. The setting of parameters for the ionization process and ion optics are critical. Optimal conditions are determined by employing a set of reference compounds spanning the m/z range of the instrument. Fragmentation conditions can be further tuned to extend the signal response for each target. The fragmentation patterns are affected by different collision conditions, the nature of the collision gas and its pressure. These instrumental parameters and the optimal should then be meticulously determined (Kitteringham et al., 2009; Koal & Deigner, 2010).

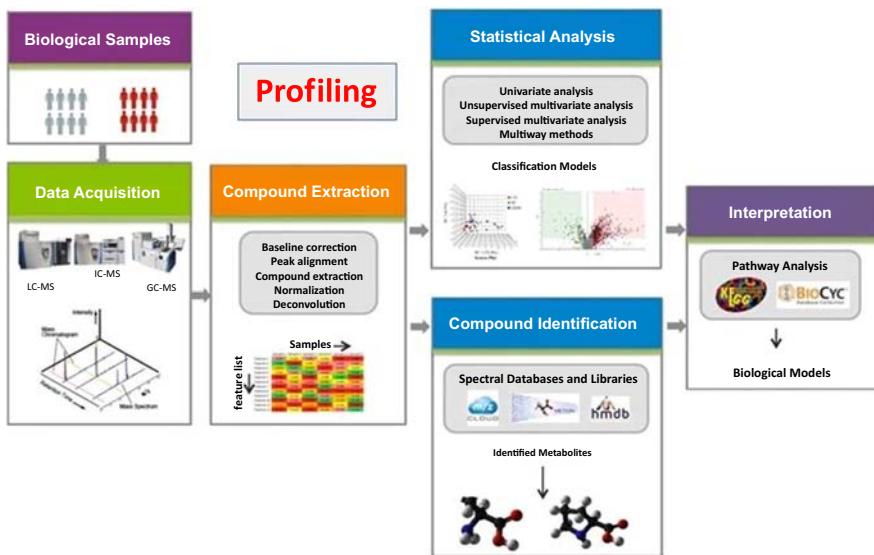
Once developed and validated, the MRM methods can be utilized in any application that involves identification and quantification of targeted molecules in any field of science.

Untargeted metabolomics in complex samples

Metabolomics, as all post-genomics disciplines, is addressed to high-speed, high-throughput, and comprehensive analysis of metabolites in biological samples by integrating high-throughput analysis techniques and bioinformatics (Fig. 4.17). Metabolomics plays a key role in the investigation of the complex metabolic networks in different conditions, including health and disease.

The strategies and methodologies used in metabolomic analyses were developed during the past 20 years in the field of drugs analysis, in the study of disease-related biomarkers, toxicology, molecular mechanisms and human biochemistry. In general, metabolome is less complex than the proteome even if metabolites display different physico-chemical properties with a wide range of concentration, solubility, polarity, and volatility compared to proteins. However, the wide range of different properties among thousands of analytes and their dynamic change reflecting endogenous and exogenous species interacting with each other (e.g., drugs, toxins, microorganisms, and nutrients) make metabolomics analyses very challenging.

The ability of metabolomics to highlight even subtle changes of the cell status is one of the main advantage of this discipline and functional information

**FIGURE 4.17**

Untargeted mass spectrometry based metabolomics workflow (<http://www.thermofisher.com>).

From <http://www.thermofisher.com>.

originated by metabolomics analyses increase by orders of magnitude when compared with traditional (bio)chemical or genetic screening. The state of a given biological system is described in detail by analyzing the composition, interaction or changes of metabolites. Nuclear Magnetic Resonance (NMR) and MS are the two main analytical techniques extensively applied within the metabolomics field to detect a huge number of metabolites in a single analysis. Both these techniques can achieve metabolites identification and quantification in complex samples.

Innovations and developments in MS have made this technique particularly suited for metabolic analyses so that metabolomics is now considered a specific field of study for MS. This allowed academic and industrial research labs to enlarge the list of multiomics disciplines adding metabolomics to genomics and proteomics as main field of investigation (Alves et al., 2020; Dettmer et al., 2007).

MS can typically identify several species at very low level in the femtomolar to attomolar range. However, analytes in a metabolomic sample comprise a highly complex dynamic mixture that is changing from second to second. This ensemble has to be simplified prior to detection by separation methodologies. Hyphenated approaches in which either GC, ion chromatography or LC are coupled with MS or MS/MS are able to routinely analyze thousands of metabolites in a single run. Nowadays, LC-MS/MS and GC-MS are considered powerful and high-throughput

techniques for metabolic profiling (Cui et al., 2018; Drexler et al., 2011; Fiehn, 2016).

In metabolomics analyses, single MS or MS/MS data can be accumulated and employed for relative quantification or profiling of complex mixture of metabolites, provided that their identity can be revealed. Identification of metabolites is generally achieved by searching through different databases using mass spectral data. The first metabolite database (called METLIN) was developed in 2005 using fragmentation data from MS/MS experiments. The Human Metabolome Database is perhaps the most extensive public metabolomic spectral database to date.

However, due to the lack of a genetic template for metabolites in contrast to proteins, metabolomics databases are generally considered incomplete. Incomplete metabolite databases can sometime be implemented by in-silico prediction of the fragmentation pattern of molecules not included in the databank (Psychogios et al., 2011; Wishart et al., 2018).

Metabolomics studies can be performed along two different but complementary lines of investigations, targeted metabolomics and untargeted metabolomics. Targeted Metabolomics is recognized as a quantitative method for the identification and measurement of well-defined groups of specific metabolic compounds in cells or organisms and requires a priori knowledge of the molecules to be analyzed. This approach will be discussed in a different chapter (see Chapter 6: Targeted Metabolomics).

Untargeted, or discovery-based, metabolomics focuses on global detection and relative quantitation of metabolites in a biological sample often involving the comparison of the metabolome between the control and test groups, to identify differences between their metabolite profiles which may be relevant to specific biological conditions.

How does an untargeted method work? Many untargeted approaches are based on the acquisition of as many species as possible, then the metabolites are properly annotated and all metabolic changes are reviewed. Data obtained by untargeted approaches can further be used for relative quantification of compounds in different conditions and they often represent the starting point for further studies with targeted approaches.

Untargeted metabolomics studies use two general approaches: the first is based on the accurate measurement of the molecular mass for each detected metabolite followed by statistical elaboration for relative quantification. However, the fragmentation patterns of a specific subset of molecules is often generated by data dependent acquisition for their unambiguous identification.

Alternatively, the data independent acquisition (DIA) method can be used in which measurement of molecular mass is integrated with the MS/MS fragmentation pattern for all precursor ions either simultaneously (MSE) or in finite mass ranges (SWATH). However, when analyzing complex mixtures of metabolites, the DIA method generates very complicated fragmentation spectra making the assignment of fragment ions to the corresponding precursor ion very challenging. This identification is pursued on the basis of different parameters such as

retention time, mass, and drift time without taking in consideration metabolite signal intensity. Finally, these data are used for compound identification by screening metabolite databases ([Gertsman & Barshop, 2018](#); [Gika et al., 2019](#); [Koal & Deigner, 2010](#); [Schrime-Rutledge et al., 2016](#)).

Analytical techniques in mass spectrometry -based metabolomics

The first step in metabolomics experiments consists in the extraction of metabolites from the biological matrix. Metabolites are extracted from a wide range of samples including cells, body fluids, microbes, plants and fungi using a variety of methods and solvent systems. These compounds consist of polar metabolites, small, hydrophilic metabolites such as amino acids, nucleotides, sugars and small organic acids often involved in primary metabolism, and non-polar metabolites, generally not directly involved in primary metabolism, such as antibiotics, polyketides, phenolics. Sample preparation methods must then be optimized for each type of sample and/or for specific metabolites of interest.

Extracted metabolites are often analyzed by using hyphenated techniques such as GC-MS, LC-MS, or LC-MS/MS. The separation step is instrumental to reduce the high complexity of the biological sample allowing different sets of molecules to be eluted at different times before being analyzed by a wide range of instrumental and technical MS variants.

LC and GC separation techniques are based on the different interaction of the metabolites with the adsorbent column phase resulting in different times of traveling through the column (Retention Times). Following mass spectral analysis, the resulting chromatogram showing the m/z values and the retention times of each detected metabolite constitute the result of the LC-MS and/or GC-MS-based metabolomics analysis.

Nowadays, the high sensitivity, high-throughput and the ability to detect thousands of molecules in a single analysis of complex biological samples make MS widely used in metabolomics. In this respect, the continuous development of new MS technologies and the improvement of existing instrumentation is a crucial point in metabolomics. Notwithstanding, no universal MS analytical approach exists for measuring all the metabolite universe, composed by thousands of small molecules with different chemical characteristics. A proper selection of the most suitable instrument and method in terms of resolution, sensitivity, throughput and, last but not the least, price should then be performed.

Finally, it should be underlined that critical to all MS-based approaches is the efficient desorption and ionization of metabolites, where the resulting gas phase ions can be separated by mass analyzers such as quadrupole, TOF, and IT ([Cui et al., 2018](#); [Fiehn, 2016](#); [Swenson & Northen, 2019](#)).

Gas chromatography-mass spectrometry

GC-MS is the most standardized method in metabolomics, with almost 50 years of established protocols for metabolite analyses. Compared with other analytical tools, GC-MS is one of the most efficient, sensitive, and reliable tools for the analysis of volatile metabolites and thermally stable compounds. However, relatively few metabolites are truly volatile and so, many metabolites can only be analyzed by GC-MS following chemical derivatization. Through this process, a large portion of small molecule metabolites, especially those found centrally in metabolism, enter the range of feasible GC-MS analysis. Thus, the derivatization reaction contributes to the isolation and detection of not only volatile and non-polar compounds, but also polar metabolites like fatty acids, amino acids, amines, sterols, and sugars.

The main advantages of using GC-MS for metabolomics are its high chromatographic separation power, high peak capacity, reproducible retention times, robust quantitation, high selectivity and sensitivity. Moreover, GC-MS produces reproducible molecular fragmentation patterns making it an integral tool for metabolite identification. For over 40 years, mass spectra and chromatographic retention times of metabolites have been accumulated in publicly available libraries, most notably in the Mass Spectral Library collection of the United States National Institute of Standards and Technology (NIST). Standardized fragmentation patterns are then available in the library leading to a fingerprint of specific molecules.

In summary, GC-MS has the benefits of being relatively inexpensive and straightforward to work and shows a good stability and repeatability and fast compound identification using existing commercial spectral libraries.

Liquid chromatography-tandem mass spectrometry

High-performance LC (HPLC) coupled on-line to MS (LC-MS) combines the advantages of both techniques, that is, the high selectivity and separation efficiency of chromatography and the structural information and further increase in selectivity of MS. In addition, coupling LC to MS is relatively straightforward to accomplish, due to the development of ESI. Due to different ion sources and dealing modes LC-MS and LC-MS/MS approaches generate different information compared to GC-MS methods. Different ion scanning ranges, a higher tolerance to volatility, a wider metabolite coverage, and simpler approaches to sample preparation, are the main characteristics of LC-MS and LC-MS/MS. Moreover, with the diffusion of UPLC, the separation degree, peak capacity and sensitivity resulted to be greatly improved, thus making UPLC-MS/MS the leading technology in metabolomic research. Different chromatographic columns are often selected consistent with the polarity of the analytes. As an example, lipid-soluble metabolites are analyzed by using a reversed-phase chromatographic column with

C18 or C8 stationary phases. On the contrary, some highly polar and charged metabolites are separated by a more hydrophilic chromatographic column. In LC-MS/MS analyses, identification of the metabolites can be carried out in the linear scan mode (MS) making use of accurate measurement of their molecular mass or daughter scan mode (MS-MS) generating fragmentation patterns for each compound.

Even if LC-MS/MS has become a leading technology for the analyses of both polar and nonpolar small molecules and shows an even increasing selectivity and data content, LC methods are time-consuming (minutes to hours) when compared to automated immunoassays. These parameters are particularly important as LC-MS has started to penetrate into large and medium sized hospitals and clinical laboratories spanning hundreds of different tests, ranging from rare and cryptic analytes to high volume tests in drug/toxicology, newborn screening and endocrinology. In clinical applications sample throughput is an essential need as hundreds and sometimes thousands of daily clinical tests are not unusual. The most obvious factor responsible for the limited throughput of LC-MS/MS is the time necessary for chromatography. This pressing needs to improve LC-MS/MS systems led to development of direct infusion or flow injection analyses in which the sample is simply diluted and injected directly into the MS with or without in-line sample clean-up or guard column.

An emerging technology recently introduced in metabolomics research is represented by ion mobility MS in which the ion mobility separation device (IM) is coupled to LC-MS based analyses (LC-IM-MS). Ion mobility is able to separate gas phase ions on the basis of their size-to-charge ratio or gas phase packing efficiency, thus adding a third dimension to polarity and mass separations. An increased peak capacity, a reduced chromatography time without decreasing resolution and the separation of isomeric co-eluting species are among the advantages offered by including ion mobility in the process. Furthermore, bidimensional chromatography using multi column approaches in combination with mass spectral devices can also be employed to increase separation of complex mixtures of metabolites before MS analysis.

Finally, with more advanced MS systems, multiple sequential rounds of MS (e.g., MS3) can often be performed to attain fine structural elucidation. In general, MS/MS with MS_n experiments were shown to be effective to achieve more confident identification during metabolite analyses when querying MS/MS mass spectral libraries.

Imaging mass spectrometry

Exploring the metabolic differences directly on tissues is essential for the comprehensive understanding of how multicellular organisms function. MS imaging (IMS) is an attractive technique to visualize the spatial distribution of molecules, as biomarkers, metabolites, peptides or proteins by their molecular masses.

Recently, IMS approaches were proposed for *in vivo* or *in vitro* detection and visualization of metabolites in tissues or cells.

After collecting a mass spectrum at one spot, the sample is scanned to cover the entire surface of the sample. MS data can then be used to map the distribution of selected peaks in the resulting spectra, corresponding to the specific compounds of interest, across the sample. This results in pictures of the spatially resolved distribution of the selected compounds pixel by pixel, originating a gallery of pictures depicting the local distribution of each compound within the sample.

Although IMS is generally considered a qualitative method, the signal generated by this technique is proportional to the relative abundance of the analyte, making quantification possible. Other widely used traditional methodologies that is radiochemistry and immunohistochemistry can achieve the same goal as IMS; however they are limited in their abilities to analyze multiple samples at once, and very often they need *a priori* knowledge of the samples being studied.

Most common ionization technologies in the field of IMS are DESI imaging, MALDI imaging and secondary ion MS imaging. Matrix-assisted laser desorption/ionization MS imaging (MALDI IMS) is by far the most frequent MS techniques used in IMS studies. This label-free technique is able to identify multiple metabolites and determine both their distribution and relative abundance *in situ* in a *m/z* range of 500 kDa and a spatial resolution of 20 mm. A typical MALDI IMS experiment is performed by coating a thin tissue section mounted onto a target with an appropriate matrix solution. This solution serves to extract analytes of interest from the underlying tissue and upon solvent evaporation the extracted molecules are cocrystallized with the matrix. Mass spectra are then acquired across the tissue at defined geometrical coordinates. The resulting dataset contains hundreds to thousands of individual spectra consisting of all ions detected at each location of acquisition. Specific software are then used to organize the mass spectra into a format where each spectrum represents a discrete pixel and the distribution and intensity of any of the detected species can be viewed across the tissue as an ion density map or image. In a single MALDI IMS experiment it is possible to detect hundreds or even thousands of discrete signals across a tissue from a diverse set of analytes.

The impact of MALDI imaging in metabolomics stems from the unique ability of this method to correlate chemistry and biology, to carefully link tissue histology and metabolite ion maps at meaningful levels of spatial resolution providing a greater understanding of metabolite tissue distribution. New developments in MALDI-based imaging, like matrix-free approaches, faster scan rates and smaller spot size, together with new improved software, will result in the acquisition of more precise and detailed images with wider metabolite coverage.

The IMS techniques have become high-throughput molecular tools extensively used in histological research field, providing an effective method to monitor the distribution of endogenous metabolites over time and space. These approaches can be successfully applied in analyses for diagnosis of carcinoma and may even be extended to whole body imaging in different diseases. As they can obtain a

comprehensive and high-throughput characterization of metabolic changes in micro regions, IMS technologies provide methods and perspectives for the study of metabolic mechanisms in different diseases. Key challenges in MS imaging range from sample preparation (including optimization of cell fixation) to statistical analysis of huge datasets, and therefore to the balance between spatial resolution of tissue slice imaging and sensitivity of analyte detection ([Cole & Clench, 2015](#); [Debois et al., 2012](#); [Wu et al., 2013](#)).

Data analysis

In an untargeted metabolomics study, crucial steps are the identification but also the quantitation of all class of detected species. As with other “omics” techniques, metabolomic analysis generates large-scale and complex datasets containing many different information, that is peak areas, peak retention times and spectral information. This data may also contain many experimental artifacts, and sophisticated software is required for high-throughput and efficient analysis, to provide statistical power to eliminate systematic bias, confidently identify compounds and explore significant findings. Various data analysis tools have been proposed including multiple univariate and multivariate statistical methods, commonly known as chemometric methods, to extract biologically relevant information.

Metabolomics data analysis workflows usually consist of several steps, including feature extraction, compound identification, statistical analysis and interpretation. Data analysis is a significant part of the metabolomics workflow, with compound identification being the major bottleneck. The typical untargeted metabolomic workflow starts with the processing of the spectral data which are highly dependent on the analytical technique used (e.g., NMR, LC-MS, or GC-MS). The general structure of the metabolomics data in the dataset is then analyzed by chemometric methods. Univariate and multivariate data analyses are instrumental to define how the different metabolic information are related with the phenotype associated with the samples. Finally, models attempting to describe the biological pathways and/or networks associated with the observed data are constructed.

Bioinformatics plays a crucial role in the integration of metabolomic data with other “omics” procedures (e.g., genomics, transcriptomics and/or proteomics). Some tools most commonly used in metabolomic analysis providing different methodological options for spectral processing, data analysis, or pathway analysis are freely available on the net ([Chong et al., 2018](#); [Swenson & Northen, 2019](#); [Wishart et al., 2018](#); [Zampieri et al., 2017](#)).

Applications

The rapid growth of metabolomics was essentially due to the possibility to simultaneously study tens to hundreds of metabolites in complex biological samples

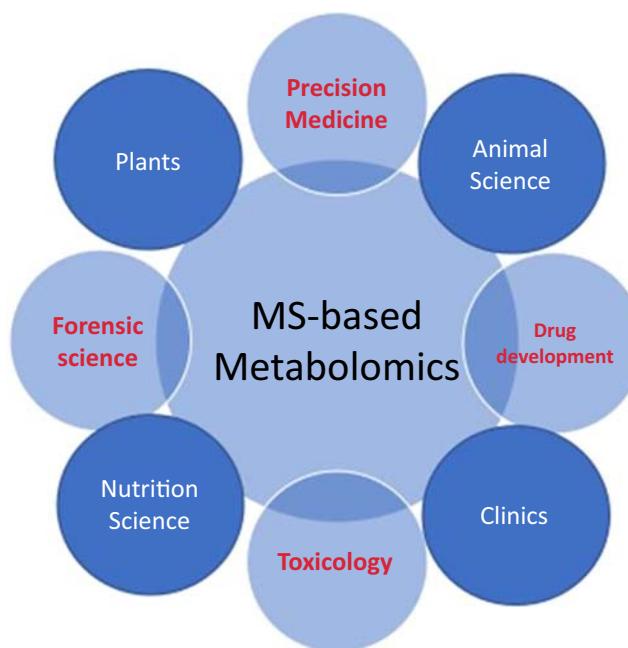


FIGURE 4.18

Untargeted metabolomics applications.

thus widening its range of applications (Fig. 4.18). Numerous examples of untargeted metabolomics experiments are present in literature, considering its ability to detect subtle changes in large datasets through comprehensive metabolic measurements. In the recent years, metabolomics has been applied in multiple fields, from plant biology, clinical biomarker discovery, drug development and discovery, toxicology, forensics and nutrition science.

Untargeted metabolite profiling is considered a powerful discovery tool for diagnostic purposes in the identification of metabolites associated with a disease. Personalized medicine, the ultimate customization of healthcare, requires metabolomics for quick medical diagnosis to identify disease. In healthcare, classical biochemical tests are currently used to measure individual metabolite concentrations to define disease states (e.g., the blood-glucose level in the case of diabetes). Metabolomics offers the potential for the rapid identification of hundreds of metabolites, making it possible to study the commonalities, properties, and laws of each of the metabolic components. Metabolomics is more closely linked to physiology than genomics and proteomics. The disease causes changes in the pathophysiological process of the body, which eventually causes corresponding changes in metabolites. Metabolomics is able to detect even subtle differences between normal and disease states and to identify the onset of disease much earlier (S. J. Kim et al., 2021; Y.-M. Kim & Heyman, 2018; Schrimpe-Rutledge et al., 2016).

Biomarker discovery is another area where metabolomics can help decision making. Biomarkers constitute objective indications of medical state observed from outside the patient which can be measured accurately and reproducibly. By analyzing certain metabolites and comparing them with normal human metabolites, specific biomarkers of the disease can be defined. In metabolomics, biomarkers can be used to distinguish two groups of samples, typically a disease and control group. For example, a metabolite reliably present in disease samples, but not in healthy individuals would be classified as a biomarker. Samples of urine, saliva, bile, or seminal fluid contain highly informative metabolites, and can be readily analyzed through metabolomics fingerprinting or profiling, for the purpose of biomarker discovery. Urine metabolomics has recently emerged as an interesting field for the discovery of biomarkers to detect subtle metabolic changes in response to a specific disease or therapeutic intervention. Urine, compared to other biological fluids, is characterized by its ease of collection, richness in metabolites and its ability to reflect imbalances of all biochemical pathways within the body (Khamis et al., 2017). MS was shown to provide the best sensitivity, selectivity and identification capabilities to analyze the majority of the metabolite composition in the urine. Furthermore, differential isotope tagging techniques have provided a solution to ion suppression from urine matrix thus allowing for quantitative analysis.

Plant metabolomics is particularly interesting because of the range and functions of primary and secondary metabolites in plants. About 300 distinct metabolites could be routinely identified per sample a decade ago, and the number is gradually increasing over time. Moreover, as the development of new pesticides is critical to meet the growing demands on farming, metabolomics can help to estimate associated risks by informative snapshots acquired at different time points during plant development.

Application of metabolomics has also been proposed for comprehensive and quantitative analysis of the amine- and phenol-containing metabolites in fecal samples by using a chemical isotope labeling LC-MS method. Thousands of metabolites could be identified providing the most comprehensive profile of the amine/phenol sub-metabolome ever detected in human fecal samples (Alvarez & Naldrett, 2021; Zampieri et al., 2017).

Metabolomic profiling of tissues is still poorly pursued, however, a characterization of the metabolite patterns in mouse brain, liver, kidney and skeletal muscle was proposed by integrating solid phase microextraction of analytes and LC-MS. All the targeted organs showed differences in alpha-amino acids, purine nucleotides and fatty acid esters thus contributing to the definition of the baseline metabolome of organs (Y. Wang et al., 2018).

Metabolomic analysis for clinical biomarker discovery

Early diagnosis of disease is especially important for the success in treatment and to limit the extent of disease. Therefore, identification of specific diagnostic

biomarkers within the initial stage of a disease is the main goal of metabolomics. One of the first examples of using biomarkers dates back to 1846 when H. Bence-Jones identified a primary tumor marker, the Bence Jones protein, a biomarker with a diagnostic value for myeloma. However, early diagnosis is not always feasible as several diseases lack a specific marker or they are caused by several factors thus needing the simultaneous quantification of several biomarkers. More than a century had to pass before metabolomics could provide an excellent possibility of efficient screening of biomarkers in disease diagnostic.

Diagnosis of disease often requires invasive examination, like biopsies, or computerized tomography scans. Metabolomics measurements of biomarkers open different scenarios for broad clinical applications, including screening studies, diagnosis and detection, prognosis and prediction, and monitoring of disease progress. Rapid and effective quantitative analysis of specific biomarkers can provide prognostic information about disease outcomes, suggest early treatment for patients and select and evaluate targeted therapies.

Biomarkers can play a crucial role for the diagnosis and prognosis of disease in personalized treatment and precision medicine. However, their discovery and development are challenging, requiring a huge amount of economic investment, rigorous protocols and technology but also multidisciplinary teams of clinicians, biologists, chemists and other experts. Large-scale cohort of patients, multicenter studies and comprehensive clinical information together with large scale, multidisciplinary efforts are needed for biomarkers discovery and their validation. Among numerous potential biomarkers under examination, only about 1% entered clinical practice ([French, 2017; Luan et al., 2019](#)).

Metabolomics in drug development

Biology, chemistry and pharmacy create the complex network leading to drug development. Traditionally, drugs are screened in molecular libraries, trying to optimize their characteristics, and their molecular affinity, selectivity, metabolic stability, oral bioavailability are deeply investigated. Only after satisfying the above conditions, a little number of molecules enter the step of clinical trials but most of them fail to reach satisfactory levels.

The application of metabolomics technology has gradually penetrated into various fields of drug development such as the study of drugs, mechanisms of action and compatibility. Metabolomics can investigate the specific pharmacological effects, the toxic and side effects and the metabolic mechanisms of the drugs, the compatibility rules and the affected metabolic pathways, significantly accelerating the process of drug discovery, and contributing to the definition of suitable clinical plans.

Metabolomics can also provide information on drug efficacy and safety by monitoring the changes in metabolic disturbances *in vivo* and *in vitro*. This function is typically applied in animal model validation and in preclinical studies of drug development. Moreover, the process of different treatments of the drug could

also be distinguished by metabolic methods. Finally, metabolomics investigations can contribute to the identification of potential new therapeutic targets thus opening the way to new drugs development (Drexler et al., 2011; Papac & Shahrokh, 2001; Wen & Zhu, 2015).

Metabolomics in nutrition science

Since the 1970s several epidemiological and clinical studies have provided clear evidence that diet plays a key role in the development of several diseases, including diabetes, cardiovascular diseases and cancer. It is now well established that selected foods, nutrients, and dietary patterns interact with various metabolic processes affecting several diseases in positive or negative mode. As an example, high salt intake influences blood pressure, or high meat intake increases the incidence of cancers. On the other side, the positive effects of classical Mediterranean diet, and its antioxidant power, are well established. The prevention or development of many diseases are tightly affected by diet, and in this respect, food intake is a crucial environmental factor, because food ingredients may alter the composition of proteins and metabolites.

Because of the complexity of the components present in food and how they may interact with the biochemical networks of living organisms, holistic approaches, capable of gathering comprehensive, high throughput amounts of data, are extremely well suited to enhance our understanding of the role of food in health and disease. Vegetables and fruits are full of trace elements required by humans, and many other foods provide essential proteins, lipids and sugars to the body.

In this context, global metabolite analysis is becoming an appealing research tool for nutrigenomics and nutrigenetics scientists. The application of metabolomics in the field of nutrition is called nutritional metabolomics, and refers to the systematic study of the interaction between diet and organism metabolism using metabolomics in different health and disease states of organisms. At present, the world is facing the double problem of nutritional deficiencies and overnutrition. The prevalence of nutrition-related chronic diseases is increasing year by year. Nutritional metabolic diseases such as diabetes, high blood pressure and obesity are serious threats to national health. Studies have shown that personalized diets can change the metabolism of living organisms, which in turn affects their health status.

Nutritional metabolomics is revealing as a useful method to monitor dietary uptake and related systemic biological changes. Metabolomics is a high throughput and sensitive procedure for revealing the complex relationship between dietary exposure and chronic diseases with metabolic molecular changes. The main aims of nutritional metabolomics rely on the definition of the effects of diet compounds on metabolism, the detection of dietary biomarkers, the identification of dietary-related diseases and disease biomarkers and the use of nutrition as a tool to monitor specific molecular mechanisms. Compared with traditional dietary

evaluation methods, metabolomics provides more objective and comprehensive dietary uptake measurements, thus improving our capability to assess the real effects of diet on metabolic networks. In this context, monitoring untargeted metabolite profiles following specific dietary protocols, by LC and GC-MS/MS, might provide very accurate measurements of diet compounds thus effectively defining the consequences of diet/food on metabolic pathways ([Rafiq et al., 2021](#)).

Metabolomics in toxicology

In toxicology, metabolomics is the -omics discipline that is most closely related to classical knowledge of disturbed biochemical pathways and it is now one the most relevant disciplines. The first toxicological applications of metabolomics were for mechanistic research, but different ways to use the technology in a regulatory context are being explored.

Metabolomics allows rapid identification of potential targets of dangerous compounds, can give information on target organs, can evaluate and predict toxicity changes in organisms and often can help to improve our understanding regarding the mode-of-action of a given compound. Metabolomic procedures are commonly applied to provide a detailed analysis of altered metabolic pathways that are targeted by harmful chemicals helping researchers and key players in the medical field to understand the mechanism of harmful chemicals. Moreover, such insights aid the discovery of biomarkers that either indicate pathophysiological conditions or help the monitoring of the efficacy of drug therapies.

The main advantage of using metabolomics in toxicology is that alterations in metabolism are “downstream” events to those that occur at genetic, transcriptomic, and proteomic levels. Thus, metabolomics facilitates the understanding of direct cellular phenotypes that are induced by toxic insults. More important, metabolomics can be used with non-invasive or low-invasive sampling of biological fluids (blood, urine, etc.) to gain information on specific organ toxicities since changes in extracellular metabolites reproduce the intracellular scenario. Another important benefit of metabolomics is that information on a large number of metabolites can be obtained through a single analysis, leading to a comprehensive and rapid understanding of the level of drug toxicity.

As an example, metabolomics is the prevalent technical method to monitor toxicity problems within traditional Chinese medicines (TCM) due to the complex composition of TCM, its multi-component and multitarget mode of action. Ren and coworkers ([Ren et al., 2020](#)) studied the toxic effect of Aconiti kusnezoffii radix in rats in order to define the correct dose to be used. The toxicity of Aconiti kusnezoffii radix, a substance extensively used in TCM, had already been observed before by histopathology. By comparing the histopathology data with the changes in metabolites detected under the action of Aconiti kusnezoffii radix, the key metabolic pathways involved in the administration of the radix were determined and the compatibility and detoxification effects were clarified.

A LC-MS/MS-based metabolomics approach was proposed by Xia et al. (Y. [Xia & McLuckey, 2008](#)) to demonstrate the protective role of some herbs against the hepatorenal toxicity associated with overdose or long-term use of cinnabar and realgar.

Finally, in vitro metabolomics analyses can also be useful to predict toxicological effects of unknown chemicals in biological systems using known reference chemicals. The metabolic profile of the known reference chemical is compared with its toxicological profile *in vivo* originating a toxicity pattern which defines the metabolic profile related to the specific toxic compound. Afterward, the metabolic profile of the unknown chemical is aligned with the toxicity pattern and the degree of overlap is calculated, suggesting the possible toxicologic effects of the unknown chemicals.

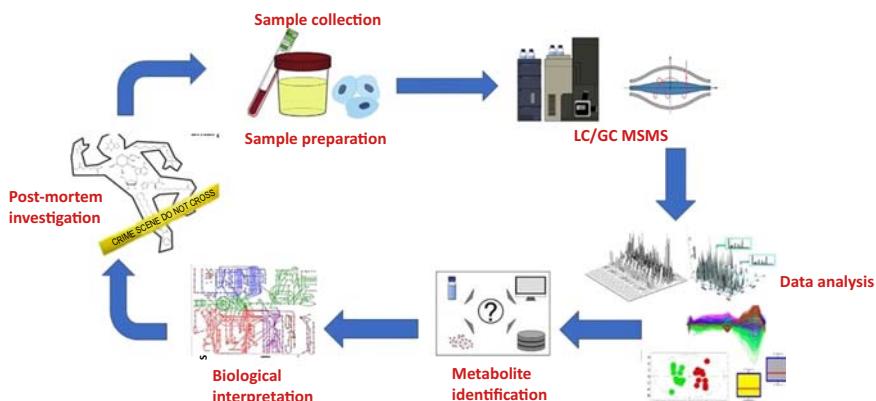
Since metabolomic approaches lead to the determination of hundreds of metabolites in a single analysis, the definition of quality criteria to avoid artifacts and to simplify the comparison of data for toxicological studies is of the utmost importance. Apart from the biological variability, analytics and sample processing also are potential sources of variability, thus the accordance of metabolomics protocols both at biological and analytical levels are mandatory. Reference standards and reference systems, negative and positive test controls are essential to demonstrate the solidity of the analysis and to integrate metabolomics results with other -omics approaches, that is proteomics and transcriptomics data ([Olesti et al., 2021](#); [Szeremeta et al., 2021](#)).

Metabolomics in forensic science

Several scientific disciplines, such as medicine, chemistry, physics, and biology, contribute to generate the forensic science for the supplying of tools for all actors involved in the sphere of criminal inquiry. Nowadays, metabolomics is increasingly used in forensics science for the identification of peculiar metabolic fingerprints and specific ante-mortem and post-mortem profiles ([Fig. 4.19](#)).

Determining the time since death (PMI) is essential in crime investigation as the knowledge of a time frame can help to investigate the possible causes of death, to clarify the death circumstances and to assess any potential information made by suspects. Several specific biomarkers related to postmortem changes of metabolites including amino acids sugars and carboxylic acids have been proposed and many of them were seen to have statistically strong correlation with PMI and the potential to estimate the time since death.

The use of metabolomics opens the possibility of identifying some novel markers of forensic interest in the field of drug abuse, especially new psychoactive substances (NPS). Untargeted metabolomic-related procedures might be highly beneficial to provide fast response to suspected NPS consumption and aid in the overall diagnostics of drug abuse or overdose. The holistic profiling approach can be used to profile the range of chemical xenobiotics and their metabolites in humans as often the parent compound itself remains undetectable in specimens,

**FIGURE 4.19**

Forensics untargeted metabolomics pipeline.

and drug intake detection will only be possible over one or a few unique metabolites. Moreover, in the case of common primary metabolites for several structurally related compounds, other minor metabolites might be necessary to prove the intake of a particular illegal drug.

Novel metabolomics approach connected to the drug of abuse did not only focus on biomarkers indicating acute drug consumption, but also on the identification of guide for drug addiction, the intensity of drug addiction or the interpretation of the level of intoxication. In this respect, new degradation pathways have been defined by metabolomics workflow to better depict addiction of known drugs of abuse and to counteract new emergent drugs within the clandestine market.

Several deaths but also several crimes are due to abuse of opium alkaloids, such as morphine, codeine and heroin. A number of well-established biomarkers of these opium derived drugs including 6-acetylmorphine, morphine, codeine, codeine-6-glucuronide, 6-acetylcodeine, noscapine, papaverine and thebaine are commonly used to detect drug addiction. However, recent metabolomics investigation pointed out to new altered metabolic mechanisms to monitor heroin abuse. Alterations in L-tryptophan, 5-hydroxytryptamine and 5-hydroxyindoleacetate metabolism were suggested to be related to long-term drug addiction with relevant clinical and forensic implications.

A major advantage of metabolomics in forensics is represented by the possibility of integrating data originating from different body compartments even considering the low amount and complexity of most forensic samples. Blood, urine, saliva, feces, tissues, etc., are the typical samples detected on a crime scene which can be submitted to metabolomic analyses.

It is well known that the identity of criminals can be identified by the examinations of fingerprints occurring on a crime scene. Recently, MALDI imaging was proposed as an alternative method to detect fingerprints on a crime scene by

analyzing the spatial distribution of specific fatty acids occurring on human fingertip. The resulting MS images were demonstrated to be of higher quality in comparison with classical procedures and independent from the surface where the fingerprint was deposited. Moreover, the metabolomic investigation was a non-invasive methodology leaving the fingerprint available for afterwards classical analyses.

Besides the fatty acids originating fingerprint images, MALDI IMS was successfully employed in the detection of other specific metabolites occurring within the fingerprints, thus providing further information on exogenous chemicals crystallized in the samples. Specific drugs including amphetamines, alkaloids, opioids, cannabinoids and specialty drugs and their metabolites could be detected in MALDI IMS within fingermarks suggesting illegal activities and/or drug abuse of the suspect. Finally, plastic trace elements of a specific type of condom detected within a fingerprint led to a rape charge of the suspect (Akçan et al., 2020; Benson et al., 2006; Du et al., 2020; Szeremeta et al., 2021).

References

- Akçan, R., Taştekin, B., Yıldırım, M. Ş., Aydogan, H. C., & Sağlam, N. (2020). Omics era in forensic medicine: Towards a new age. *Turkish Journal of Medical Sciences*, 50(5), 1480–1490.
- Alvarez, S., & Naldrett, M. J. (2021). *Mass spectrometry based untargeted metabolomics for plant systems biology*. Emerging topics in life sciences. Portland Press.
- Alves, S., Paris, A., & Rathahao-Paris, E. (2020). Mass spectrometry-based metabolomics for an in-depth questioning of human health. *Advances in Clinical Chemistry*, 99, 147–191.
- Awad, H., Khamis, M. M., & El-Aneed, A. (2015).). Mass spectrometry, review of the basics: Ionization. *Applied Spectroscopy Reviews*, 50(2), 158–175.
- Baidoo, E. E., & Benites, V. T. (2019). *Mass spectrometry-based microbial metabolomics: Techniques, analysis, and applications*. *Microbial Metabolomics* (pp. 11–69). Springer.
- Benson, S., Lennard, C., Maynard, P., & Roux, C. (2006). Forensic applications of isotope ratio mass spectrometry—A review. *Forensic Science International*, 157(1), 1–22.
- Bhardwaj, C., & Hanley, L. (2014). Ion sources for mass spectrometric identification and imaging of molecular species. *Natural Product Reports*, 31(6), 756–767.
- Brown, R. S., & Lennon, J. J. (1995). Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer. *Analytical Chemistry*, 67(13), 1998–2003.
- Byliński, H., Gębicki, J., Dymerski, T., & Namieśnik, J. (2017). Direct analysis of samples of various origin and composition using specific types of mass spectrometry. *Critical Reviews in Analytical Chemistry*, 47(4), 340–358.
- Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., & Xia, J. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1), W486–W494.

- Cole, L. M., & Clench, M. R. (2015). Mass spectrometry imaging tools in oncology. *Biomarkers in Medicine*, 9(9), 863–868.
- Cui, L., Lu, H., & Lee, Y. H. (2018). Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases. *Mass Spectrometry Reviews*, 37(6), 772–792.
- Dawson, P. H. (2013). *Quadrupole mass spectrometry and its applications*. Elsevier.
- Debois, D., Smargiasso, N., Demeure, K., Asakawa, D., Zimmerman, T. A., Quinton, L., & De Pauw, E. (2012). MALDI in-source decay, from sequencing to imaging. *Applications of MALDI-TOF Spectroscopy*, 117–141.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78.
- Drexler, D. M., Reily, M. D., & Shipkova, P. A. (2011). Advances in mass spectrometry applied to pharmaceutical metabolomics. *Analytical and Bioanalytical Chemistry*, 399 (8), 2645–2653.
- Du, T., Mengxi, M., Ye, X., Tu, C., Jin, K., Chen, S., Liu, N., Xie, J., & Shen, Y. (2020). Research progress of metabolomics in forensic pathology. *Fa Yi Xue Za Zhi*, 36(3), 347–353.
- Fiehn, O. (2016). Metabolomics by gas chromatography–mass spectrometry: Combined targeted and untargeted profiling. *Current Protocols in Molecular Biology*, 114(1), 30–34.
- French, D. (2017). Advances in clinical mass spectrometry. *Advances in Clinical Chemistry*, 79, 153–198.
- Gertsman, I., & Barshop, B. A. (2018). Promises and pitfalls of untargeted metabolomics. *Journal of Inherited Metabolic Disease*, 41(3), 355–366.
- Gika, H., Virgiliou, C., Theodoridis, G., Plumb, R. S., & Wilson, I. D. (2019). Untargeted LC/MS-based metabolic phenotyping (metabonomics/metabolomics): The state of the art. *Journal of Chromatography B*, 1117, 136–147.
- Haag, A. M. (2016). Mass analyzers and mass spectrometers. modern proteomics—sample preparation. *Analysis and Practical Applications*, 157–169.
- Iwasaki, Y., Nakano, Y., Mochizuki, K., Nomoto, M., Takahashi, Y., Ito, R., Saito, K., & Nakazawa, H. (2011). A new strategy for ionization enhancement by derivatization for mass spectrometry. *Journal of Chromatography B*, 879(17–18), 1159–1165.
- Khamis, M. M., Adamko, D. J., & El-Aneed, A. (2017). Mass spectrometric based approaches in urine metabolomics and biomarker discovery. *Mass Spectrometry Reviews*, 36(2), 115–134.
- Kim, S. J., Song, H. E., Lee, H. Y., & Yoo, H. J. (2021). Mass spectrometry-based metabolomics in translational research. *Advanced imaging and bio techniques for convergence science*, 509.
- Kim, Y.-M., & Heyman, H. M. (2018). *Mass spectrometry-based metabolomics. Fungal genomics* (pp. 107–118). Springer.
- Kitteringham, N. R., Jenkins, R. E., Lane, C. S., Elliott, V. L., & Park, B. K. (2009). Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. *Journal of Chromatography B*, 877(13), 1229–1239.
- Koal, T., & Deigner, H.-P. (2010). Challenges in mass spectrometry based targeted metabolomics. *Current Molecular Medicine*, 10(2), 216–226.
- Luan, H., Wang, X., & Cai, Z. (2019). Mass spectrometry-based metabolomics: Targeting the crosstalk between gut microbiota and brain in neurodegenerative disorders. *Mass Spectrometry Reviews*, 38(1), 22–33.

- Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72(6), 1156–1162.
- McLuckey, S. A., Van Berkel, G. J., Goeringer, D. E., & Glish, G. L. (1994). Ion trap mass spectrometry using high-pressure ionization. *Analytical Chemistry*, 66(14), 737A–743A.
- Nikolaev, E. N., Kostyukevich, Y. I., & Vladimirov, G. N. (2016). Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations. *Mass Spectrometry Reviews*, 35(2), 219–258.
- Olesti, E., González-Ruiz, V., Wilks, M. F., Boccard, J., & Rudaz, S. (2021). Approaches in metabolomics for regulatory toxicology applications. *Analyst*, 146(6), 1820–1834.
- Papac, D. I., & Shahrokh, Z. (2001). Mass spectrometry innovations in drug discovery and development. *Pharmaceutical Research*, 18(2), 131–145.
- Perry, R. H., Cooks, R. G., & Noll, R. J. (2008). Orbitrap mass spectrometry: Instrumentation, ion motion and applications. *Mass Spectrometry Reviews*, 27(6), 661–699.
- Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., & Gautam, B. (2011). The human serum metabolome. *PLoS One*, 6(2), e16957.
- Rafiq, T., Azab, S. M., Teo, K. K., Thabane, L., Anand, S. S., Morrison, K. M., de Souza, R. J., & Britz-McKibbin, P. (2021). Nutritional metabolomics and the classification of dietary biomarker candidates: A critical review. *Advances in Nutrition*.
- Ren, J., Dong, H., Han, Y., Yang, L., Zhang, A., Sun, H., Li, Y., Yan, G., & Wang, X.-J. (2020). Network pharmacology combined with metabolomics approach to investigate the protective role and detoxification mechanism of Yunnan Baiyao formulation. *Phytomedicine*, 77, 153266. Available from <https://doi.org/10.1016/j.phymed.2020.153266>.
- Schrinne-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12), 1897–1905.
- Swenson, T. L., & Northen, T. R. (2019). *Untargeted soil metabolomics using liquid chromatography–mass spectrometry and gas chromatography–mass spectrometry*. *Microbial metabolomics* (pp. 97–109). Springer.
- Szeremeta, M., Pietrowska, K., Niemcunowicz-Janica, A., Kretowski, A., & Ciborowski, M. (2021). Applications of metabolomics in forensic toxicology and forensic medicine. *International Journal of Molecular Sciences*, 22(6), 3010.
- Van Berkel, G. J. (2003). An overview of some recent developments in ionization methods for mass spectrometry. *European Journal of Mass Spectrometry*, 9(6), 539–562.
- Vestal, M. L., & Campbell, J. M. (2005). Tandem time-of-flight mass spectrometry. *Methods in Enzymology*, 402, 79–108.
- Wang, Y., Sun, J., Qiao, J., Ouyang, J., & Na, N. (2018). A “soft” and “hard” ionization method for comprehensive studies of molecules. *Analytical Chemistry*, 90(24), 14095–14099.
- Wen, B., & Zhu, M. (2015). Applications of mass spectrometry in drug metabolism: 50 years of progress. *Drug Metabolism Reviews*, 47(1), 71–87.
- Whitehouse, C. M., Dreyer, R. N., Yamashita, M., & Fenn, J. B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, 57(3), 675–679.
- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., & Karu, N. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617.

- Wolk, D. M., & Clark, A. E. (2018). Matrix-assisted laser desorption time of flight mass spectrometry. *Clinics in Laboratory Medicine*, 38(3), 471–486.
- Wu, C., Dill, A. L., Eberlin, L. S., Cooks, R. G., & Ifa, D. R. (2013). Mass spectrometry imaging under ambient conditions. *Mass Spectrometry Reviews*, 32(3), 218–243.
- Xia, Y., & McLuckey, S. A. (2008). Evolution of instrumentation for the study of gas-phase ion/ion chemistry via mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 19(2), 173–189.
- Zampieri, M., Sekar, K., Zamboni, N., & Sauer, U. (2017). Frontiers of high-throughput metabolomics. *Current Opinion in Chemical Biology*, 36, 15–23.
- Zhang, J., McCombie, G., Guenat, C., & Knochenmuss, R. (2005). FT-ICR mass spectrometry in the drug discovery process. *Drug Discovery Today*, 10(9), 635–642.
- Zubarev, R. A., & Makarov, A. (2013). *Orbitrap mass spectrometry*. ACS Publications.

Further reading

- Wang, S., Blair, I. A., & Mesaros, C. (2019). Analytical methods for mass spectrometry-based metabolomics studies. *Advancements of Mass Spectrometry in Biomedical Research*, 635–647.
- Xia, F., Li, A., Chai, Y., Xiao, X., Wan, J., Li, P., & Wang, Y. (2018). UPLC/Q-TOFMS-based metabolomics approach to reveal the protective role of other herbs in An-Gong-Niu-Huang Wan against the hepatorenal toxicity of cinnabar and realgar. *Frontiers in Pharmacology*, 9, 618.

This page intentionally left blank

Nuclear magnetic resonance in metabolomics

5

Abdul-Hamid Emwas¹, Kacper Szczepski², Benjamin Gabriel Poulsen²,
Ryan McKay³, Leonardo Tenori⁴, Edoardo Saccenti⁵, Joanna Lachowicz⁶, and
Mariusz Jaremko²

¹Core Labs, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

²Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

³Department of Chemistry, University of Alberta, Edmonton, AB, Canada

⁴Department of Chemistry and Magnetic Resonance Center (CERM), University of Florence, Florence, Italy

⁵Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands

⁶Department of Medical Sciences and Public Health, Università di Cagliari, Cittadella Universitaria, Monserrato, Italy

Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a versatile analytical tool that has been used for decades to identify, quantitate, and structurally elucidate molecules. Traditionally, as a requirement of any chemistry department, NMR branched out into biophysical problems in the late 80s and early 90s with peptide/protein structure/function studies (Alsiary et al., 2020; Cavanagh et al., 1995; Chu et al., 2010; Marion, 2013; Wüthrich, 1986). NMR has several unique advantages over related methods, including nondestructive and cumulative scanning, high reproducibility, and being a nonbiased method. NMR spectroscopy can examine a molecule at its atomic level, providing a potent tool to distinguish composition a priori, kinetics, energetics, and otherwise difficult to elucidate structural isomers (Dhahri et al., 2020). In contrast to other analytical tools commonly used in metabolomics studies such as GC–MS (gas chromatography–mass spectrometry) (Emwas, Al-Talla et al., 2015; Zhang et al., 2017) and liquid chromatography (LC)–MS (Raji et al., 2013; Wu et al., 2021; Zhang et al., 2017), NMR does not require structural and/or chemical manipulation, nor extra steps for sample preparation or metabolite isolation prior to measurement such as ionization and chemical derivatization (Emwas, 2015; Emwas, Salek et al., 2013). NMR is inherently quantitative as each signal intensity is directly proportional to the atomic concentration of the originating resonance in the mixture or sample (Emwas et al.,

2016). Beside its major advantages (i.e., nondestructive and minimal sample preparation requirements), 1D-NMR, especially ^1H NMR, is a relatively fast method with a single sample acquisition taking typically seconds to minutes, therefore hundreds of samples can be analyzed in a single day. It has been our experience that careful, consistent sample thawing and preparation is the practical bottle neck, not the instrument access time. The inclusion of cryogenically cooled NMR probes and the related practical signal to noise increases of 3–4-fold (corresponding to a 9–16-fold decrease in experiment time) have made NMR data acquisition even faster. Thus NMR is particularly useful in comparative studies that could involve a high number of samples, for example, hepatitis studies involving hundreds of thousands of volunteers donating multiple biofluid samples on a daily basis (Duarte et al., 2014; Sarfaraz et al., 2016; Wang et al., 2014; Zheng, Chen et al., 2017).

NMR applications are not limited to samples in the liquid state; solid-state NMR spectroscopy is also a well-established field (Alahmari et al., 2018; Apperley et al., 2012; Ashbrook et al., 2018; Chisca et al., 2015; Mroue et al., 2010; Renault et al., 2010; Separovic & Sani, 2020). Recently, tissue samples including human, animal, plant and marine tissues have been studied using HR-MAS (magic angle spinning) NMR approaches (Bunescu et al., 2010; Heude et al., 2015; Kaebisch et al., 2017; Rocha et al., 2010; Taglienti et al., 2020).

One of the most important advantages of NMR-based metabolomics is that liquid samples can be detected in mild/neutral conditions without the need for sample alterations such as chemical modification or cleavage, pressure/vacuum, or high-temperature conditions, and with only minimal preparation steps (e.g., typically an addition of a small amount of deuterated “lock” solvent and an internal reference standard with or without buffering capabilities) (Abdul Jameel et al., 2021; Harris et al., 2007; Kijewska et al., 2021; Markley et al., 1998; Mercier et al., 2011; Sheedy et al., 2010).

The inclusion of multidimensional NMR experiments, typically utilized to provide atomic 3D coordination/neighbor/distance information on each atom, can provide additional powerful insights (albeit at the cost of extra instrument time). Taking advantage of the fact that NMR measurements are nondestructive, one can record (and re-record) many NMR experiments on the same sample over different periods of time, thus providing an influential platform for enhanced signal to noise (i.e., adding experimental time together) and/or kinetic studies where the spectra are recorded during the course of a reaction period. This provides “real time” measurement and monitoring of sample status (e.g., stability), chemical reactions, and molecule interactions (e.g., protein/ligand binding), to name a few.

As stated, one of the most important advantages of NMR-based metabolomics is that samples can be detected in neutral or mild conditions usually without the need for mechanical manipulation (e.g., sonication, heating/cooling, chemical reaction/modification). This particular advantage allows NMR-based metabolomics studies to monitor metabolic flux for some micro-organisms, such as

bacteria and cell lines, and the corresponding metabolite, leading to a new metabolomics field called fluxomics (Winter & Krömer, 2013).

Nuclear magnetic resonance spectroscopy

NMR spectra can selectively observe isotopes of spin active (Keeler, 2011) single atoms, that is, 1D NMR experiments with the most common being ^1H , ^{13}C , ^{31}P and ^{15}N NMR. Popularity is due to natural abundance and magnetic susceptibility increasing signal intensity. Moreover, inductive magnetic correlation between two types of atoms can be routinely investigated through multidimensional experiments, for example, 2D NMR. While requiring much more instrument time [despite recent time-saving advances (Aljuhani et al., 2019; Cui, Zhu et al., 2019; Guennec et al., 2014; Qiu et al., 2019)], multidimensional experiments can resolve spectral overlap and/or atomic ambiguity (Cui, Zhu et al., 2019; Féraud et al., 2020; Le Guennec et al., 2015; Mattar et al., 2004). In this section, the most common metabolomics-relevant 1D NMR as well as 2D NMR experiments will be reviewed.

1D nuclear magnetic resonance

1D NMR is concerned with determining the resonant frequencies of a single type of NMR-active nucleus (i.e., ^1H , ^{13}C , ^{15}N , etc.) that depend on the chemical environment around the nuclei. In short, 1D NMR has properties that enable researchers to determine:

1. the type of molecule involved (i.e., aromatics, aliphatics, amino acids, etc.) in an experiment;
2. how the atoms are connected to each other.

These two properties of 1D NMR form the backbone for metabolomics research, and we proceed to describe the use and application of 1D NMR experiments to metabolomics according to the most common types of nuclei involved in metabolomics studies. The nuclei discussed (in order) are ^1H , ^{13}C , ^{15}N , ^{31}P , and ^{19}F .

1D ^1H nuclear magnetic resonance spectroscopy

Simple one-dimensional ^1H NMR spectroscopy is the most common approach in metabolomics studies and consists of an excitation “pulse” (e.g., micro-seconds) with an acquisition period (seconds). If the experiment is to be repeated (e.g., to improve signal to noise), then an interscan delay (seconds) is required to re-establish equilibrium prior to the subsequent scan. The short experimental time (e.g., a few minutes) with minimal sample preparation offers a high-throughput method appropriate for metabolomics studies seeking statistical confirmation.

Studies in metabolomics generally fall into one of three categories (Sahoo et al., 2020):

1. metabolic profiling
2. metabolic fingerprinting
3. metabonomics, that is, a quantitative study over time of the metabolic response to stimuli (Holmes et al., 2008)

It is important to clarify our definition of metabonomics as the literature has been somewhat confusing, with “metabolomics” and “metabonomics” often used interchangeably and/or for slightly different interpretations. Regardless, NMR spectroscopy can easily be applied to metabolomic studies falling into any one of these three categories due to the ease of application (Fan & Lane, 2016; Kanwal et al., 2020), versatility (Agrawal, 2020), reproducibility, and important nondestructive nature (Viola et al., 2006) allowing a sample to be studied and analyzed numerous times. Sample stability becomes the limiting factor (Sykes, 2007).

NMR spectroscopy (often advantageously coupled with MS) remains a method of choice, even when considering the benefits that the sole use of MS may provide. Indeed, NMR and MS are often combined in metabolomics studies (Abd Ghafar et al., 2020; Laserna et al., 2020; Nizioł et al., 2020; Tayyari et al., 2013; Vassilev et al., 2020) for their mutually complementary benefits in both identifying and analyzing metabolites (Bhinderwala, Wase et al., 2018). However, in this book chapter we will focus on the role of NMR in metabolomics and we will discuss the roles of NMR-detectable nuclei (1D and 2D experiments) in the field of Metabolomics. Strengths and weaknesses of some nuclei for 1D NMR spectroscopy are listed in Table 5.1. Examples of each of the nuclei and their respective metabolomics studies are listed in Table 5.2.

¹H 1D nuclear magnetic resonance in metabolomic studies

A basic 1D ¹H NMR experiment (i.e., delay → solvent suppression → excitation pulse → acquire) is by far the most commonly used NMR experiment in metabolomics (Chandra et al., 2021). This type of experiment can be applied to any NMR active nuclei (Gallo et al., 2019; Zhang et al., 2020). The ¹H nucleus has the highest relative receptivity (Sanders & Hunter, 1993) (aside from the ³H nucleus, which has an extremely low natural abundance therefore making absolute receptivity impractical) and the highest natural abundance at 99.99%. These properties make obtaining a ¹H NMR spectrum of any liquid sample (e.g., urine, blood plasma) relatively rapid, and simple to acquire, process, and analyze. A typical 1D ¹H NMR experiment usually takes only a few minutes with less frequent cases taking hours. Hydrogen atoms in the spectrometer will induce resonances (i.e., peaks) at distinct frequencies based on their different magnetic/chemical environments (e.g., aliphatic, aromatic, hydrophilic, distance to functional groups, exchangeable/exposed neighbors, etc.). Furthermore, the signal intensity (integrated area) is directly proportional to the quantity of nuclei in a molecule, and

Table 5.1 Strengths and weaknesses of chosen nuclei for 1D NMR spectroscopy metabolomics.

Nuclei	Strengths	Weaknesses
¹ H	<p>Rapid acquisition</p> <p>Ability to identify many (approximately 50–200) metabolites (with or without the aid of software/databases)</p> <p>Signal intensity is directly proportional to metabolite concentration, and the number of nuclei in the sample</p> <p>Ease of analysis (if there is sufficient resolution and resolving power)</p> <p>Well-defined peaks (narrow line widths)</p> <p>Ideal for untargeted analysis</p>	<p>Narrow chemical shift window (0–10 ppm)</p> <p>Low resolution may result in spectral overlap</p> <p>Solvent suppression may be needed, and may obfuscate metabolite signal</p> <p>May not provide sufficient information to determine complete atom connectivity</p>
¹³ C	<p>Broad chemical shift dispersion (~ 200 ppm)</p> <p>High resolution, less spectral overlap</p> <p>More stable to pH and other sample conditions (Edison et al., 2019)</p> <p>Minimal amount of homonuclear coupling (i.e., ¹³C-¹³C) at natural abundance (~ 1.1%)</p> <p>Ideal metabolic tracer (Fan et al., 2016; Kovtunov et al., 2014; Markley et al., 2017)</p> <p>Directly measures the backbone structures of metabolites (Clendinen et al., 2014)</p>	<p>Low natural abundance (~ 1.1%)</p> <p>Low sensitivity</p> <p>Experiment may take several hours</p> <p>Long relaxation delays (on the NMR timescale)</p>
¹⁵ N	<p>Broad chemical shift dispersion (~ 100 ppm)</p> <p>Present in a number of important metabolites</p> <p>Expands the coverage of the metabolome (Bhinderwala, Lonergan et al., 2018)</p>	<p>Low natural abundance (0.37%)</p> <p>Low sensitivity</p> <p>Long experimental time</p>
³¹ P	<p>Highly abundant (nearly 100%)</p> <p>Relatively high sensitivity</p> <p>Powerful metabolic imaging tool</p>	<p>More sample preparation required for biological samples (i.e., proteins)</p> <p>Small chemical shift window</p> <p>Peak overlap</p> <p>Very sensitive to environmental conditions (pH, T, solvent, etc.)</p> <p>Not many ³¹P reference spectra available</p> <p>³¹P containing compounds are not very stable</p> <p>Structurally similar to other metabolites, making compound identification difficult</p> <p>Not many reference spectra currently available</p>
¹⁹ F	<p>High natural abundance (100%)</p> <p>High sensitivity</p>	<p>Not many metabolites contain ¹⁹F</p>

Table 5.2 Applications of 1D NMR techniques in metabolomics.

Nucleus	Purpose(s) of Study	NMR Methods	Results	Paper
¹ H	To determine the metabolic differences between three Quinoa Ecotypes found in Ecuador before and after treatment by either washing, cooking, and/or germination.	¹ H NMR	Regardless of the Quinoa ecotype, germination (as opposed to washing or cooking treatments) caused the greatest increase of metabolites.	Lalaleo et al. (2020)
	1. To identify potential biomarkers for aflatoxin B1 ingestion in dairy cows. 2. To evaluate the effect of adding clay and/or yeast fermentation products on identified biomarkers.	¹ H NMR	Of 15 total metabolites (acetic acid, 12 amino acids, mannose, and ethanol) identified through biomarker analysis, ethanol was most influenced by the study's conditions (control, toxin, toxin with clay, and control with yeast fermentation product), and was therefore chosen as the candidate biomarker of aflatoxin B1 ingestion in dairy cows that had not ingested a sequestering agent.	Ogunade and Jiang (2019)
	To identify and monitor metabolites specific to breast cancer patients.	¹ H NMR HSQC	36 metabolites in both groups were identified, with creatine, glycine, serine, dimethylamine, trimethylamine N-oxide, α -hydroxyisobutyrate,mannitol, glutamine, cis-aconitate, and trigonelline showing the highest levels of sensitivity and specificity for differentiating breast cancer patients from healthy individuals.	Silva et al. (2019)
	To monitor the effects of treatment with VSL #3 on children with non-alcoholic fatty liver disease (NAFLD) in order to identify non-invasive biomarkers.	¹ H NMR COSY TOCSY ¹ H- ¹³ C HSQC ¹ H- ¹³ C HMBC	VSL#3 treatment-dependent urinary metabolites involved in amino-acid metabolism, nucleic acid degradation, and creatine metabolism of children with NAFLD may be considered non-invasive and effective biomarkers to evaluate their response to treatment.	Miccheli et al. (2015)
	To evaluate the effects of Algerian date (Deglet) seeds on the metabolome of LPS-IFN- γ -induced RAW 264.7 cells.	¹ H NMR	Treating RAW 264.7 cells with Deglet seed interfere with the energy and amino acid metabolism; Deglet seeds could serve as food with anti-inflammatory properties.	Abdul-Hamid et al. (2019)
	To determine a (more) complete biochemical signature of autism spectrum disorders (ASD).	¹ H-NMR ¹ H- ¹³ C HSQC	The differences between the metabolic profiles of the urine of ASD patients and normal, healthy individuals may serve as strong indicators for the diagnosis of ASD.	Nadal-Desbarats et al. (2014)

To find biomarkers that distinguish the diagnosis of schizophrenia from bipolar disorder.	^1H NMR $^1\text{H}-^{13}\text{C}$ HSQC $^1\text{H}-^{13}\text{C}$ HMBC	Blood serum metabolomics can be used to differentiate patients with schizophrenia and bipolar disorder (although there are many similarities between the two), and from healthy control individuals. Several metabolites were identified that distinguish serum, plasma, and plasma subtypes from each other. Correction for inter-individual variation was necessary to identify the distinguishing metabolites. Changes in concentration of postprandial metabolites (in serum) can be linked to food intake.	Tasic et al. (2019) Kaluarachchi et al. (2018)
To discriminate between the metabolic profiles of serum, plasma, and plasma subtypes.	^1H NMR COSY TOSCY $^1\text{H}-^{13}\text{C}$ HSQC UPLC-MS		Radjursoga et al. (2018)
To analyze the metabolic responses (in serum) to food intake according to three different diets: <ol style="list-style-type: none">1. Vegan2. Lacto ovo-vegetarian3. Omnivore	^1H NMR $^1\text{H}-^{13}\text{C}$ HSQC $^1\text{H}-^1\text{H}$ TOCSY		
To assess the use of ^1H NMR metabolomics in distinguishing the metabolites (of saliva) of healthy controls, mild cognitive impairment sufferers, and Alzheimer's disease patients.	^1H NMR	^1H NMR metabolomics has the potential to diagnose Alzheimer's disease in its early stages of development.	Yilmaz et al. (2017)
To elucidate the relationship between inositol 1,4,5 triphosphate receptor and metabolic processes.	^1H NMR $^1\text{H}-^1\text{H}$ COSY $^1\text{H}-^1\text{H}$ TOCSY $^1\text{H}-^{13}\text{C}$ HSQC	1. Differences in the metabolic profiles among the subjects clearly distinguished healthy controls from breast cancer patients. 2. Relationship identified between key metabolites (glucose, lactate, glutamate, lysine, alanine, pyruvate, NAG, and some lipids) and the high expression of 1,4,5 triphosphate receptor in breast cancer patients.	Singh et al. (2017)
To examine the reaction (growth, nutrient uptake) of <i>Zea mays</i> plants upon inoculation with <i>Trichoderma</i> and treatment (or lack therefore) of different types of fertilizers.	^1H NMR COSY TOCSY NOESY $^1\text{H}-^{13}\text{C}$ HSQC $^1\text{H}-^{13}\text{C}$ HMBC	Combining <i>Trichoderma</i> with compost fertilizer may increase phosphate uptake in plants with the necessary phosphate nutrients.	Vinci et al. (2018)
To see if ^1H -NMR based metabolomics can distinguish animal urine samples (dog, cat, horse, monkey, etc.) from human urine samples.	^1H NMR	Several characteristic metabolites were identified in animal urine samples that were not present in human urine samples, some of which could be used	Lee et al. (2019)

(Continued)

Table 5.2 Applications of 1D NMR techniques in metabolomics. *Continued*

Nucleus	Purpose(s) of Study	NMR Methods	Results	Paper
¹³ C	To see if uranium inhibits renal gluconeogenesis in humans and mice.	¹³ C NMR	as biomarkers to distinguish animal urine samples from human urine samples.	Renault et al. (2010)
	To test the ability of metabolomics in discriminating between five drugs found in citrus fruits.	¹ H NMR ¹³ C NMR	Naturally occurring uranium inhibits lactate metabolism in humans and mice.	
	To assess if ¹³ C-MRS of hyperpolarized [1- ¹³ C] pyruvate can differentiate between responding and resistant BRAFV600E melanoma cells and xenografts.	HP- ¹³ C-MRS (1D ¹³ C NMR) EPR	¹³ C based metabolomics may be a useful method to classify and distinguish drugs found in citrus fruits.	Tsujimoto et al. (2018)
	To investigate aspartate metabolism in hepatocellular carcinoma.	¹ H NMR ¹³ C NMR	The hyperpolarized lactate/pyruvate ratio may be an early indicator of response to vemurafenib (or other BRAF inhibitors) in melanoma.	
	To monitor the effect of daily administration of low amounts of 3-iodothyronamine (a metabolite) to obese mice.	¹ H NMR ¹³ C NMR	Changes in aspartate metabolism are characteristic of hepatocellular carcinoma.	Acciardo et al. (2020)
	To isolate and identify secondary metabolites from the oil-derived fungus <i>Aspergillus</i> isolated from the rhizospheric soil of <i>Phoenix dactylifera</i> (Date palm tree).	¹ H NMR ¹³ C, DEPT-135 NMR COSY ¹ H, ¹³ C-HMBC ¹ H, ¹³ C-HSQC	Subchronic effects of 3-iodothyronamine administration may include lipolysis and protein breakdown, 3-iodothyronamine may have a lasting effect on weight maintenance in mice.	
	To develop and test analysis methods to improve quantification of pyruvate to lactate, a key indicator of cancer cell metabolism.	¹³ C - NMR	One novel compound (1-(4-hydroxy-2,6-dimethoxy-3,5-dimethylphenyl)-2-methyl-1-butanone), and four secondary metabolites (citricin, dihydrocitrinone, 2, 3, 4-trimethyl-5, 7-dihydroxy-2, 3-dihydrobenzofuran, and orcinol) were identified, with the novel compound showing strong antimicrobial activity against <i>Staphylococcus aureus</i> , and significant growth inhibition against <i>Candida albicans</i> and <i>Candida parapsilosis</i> .	Darpolar et al. (2014)
	To see if biological metabolites hyperpolarized via dissolution dynamic nuclear polarization (d-DNP) and cross polarization (CP) yield readable NMR spectra.	¹³ C NMR ¹ H- ¹³ C HSQC	NMR analysis of hyperpolarized carbon-13 pyruvate, coupled with dynamic imaging and kinetic modeling, provides quantitative assessments of prostate cancer metabolism.	
			d-DNP combined with CP can enhance the ¹³ C signal of biological metabolites in less time than a standard 1D ¹³ C NMR or ¹ H- ¹³ C HSQC spectrum with comparable (or even enhanced) quality of spectra obtained from standard experiments.	Larson et al. (2018)

Renault et al.
(2010)

Tsujimoto
et al. (2018)

Acciardo
et al. (2020)

Darpolar
et al. (2014)

Haviland
et al. (2013)

Orfali and
Perveen
(2019)

Larson et al.
(2018)

Dumez et al.
(2015)

	To find acceptable candidates for in vivo MRI (NMR) imaging to determine glucose cellular uptake using para-hydrogen-induced polarization (PHIP). To investigate the toxicity of using para-hydrogen-induced polarization Side Arm Hydrogenation approach (PHIP-SAH) to hyperpolarize metabolites in order to enable metabolic imaging of prostate cancer cell lines. To develop an economic approach to synthesize hyperpolarized pyruvate.	¹ H NMR ¹³ C NMR ¹³ C NMR ¹³ C NMR	PHIP hyperpolarized glucose molecules may serve as biomarkers for tumor progression. Aqueous solutions of hyperpolarized metabolites have moderate levels of toxicity. PHIP-SAH can be used to hyperpolarize acetate in a quick and cost-effective manner. This hyperpolarized acetate can then be converted into pyruvate, a common metabolite used for in vivo metabolic profiling of cancer cells. The NMR signal of ¹³ C nuclear polarization of 1- ¹³ C-phospholactate-d ₂ was enhanced by greater than 3×10^7 fold.	Reineri et al. (2010) Cavallari et al. (2020) Reineri et al. (2015)
¹⁵ N	To synthesize 1- ¹³ C-phosphoenolpyruvate-d ₂ , a precursor for parahydrogen-induced polarization (PHIP) of 1- ¹³ C- phospholactate-d ₂ . To explore methods (chemical derivatization + ¹³ C NMR techniques) to improve characterization of bodily fluids such as urine and serum.	¹ H NMR ¹³ C NMR ¹ H- ¹³ C HSQC	Chemical derivatization of human bodily fluids (urine and serum) coupled with ¹³ C labeled compounds and ¹³ C based NMR techniques are sufficient to derive high quality NMR spectra of said human bodily fluids.	Shchepin et al. (2014) Shanaiah et al. (2007)
	To investigate the primary nitrogen metabolism of the N ₂ -fixing root nodule symbiosis <i>Alnus incana</i> (L.)—Frankia.	¹⁵ N NMR ³¹ P NMR	The assimilation of NH ⁴⁺ in <i>A. incana</i> root nodules primarily occurs through the GS-GOGAT (glutamine synthetase (GS) and glutamate synthase (GOGAT)) pathway.	Lundberg and Lundquist (2004) Theis et al. (2015)
	To demonstrate the use of SABRE-SHEATH (signal amplification by reversible exchange in shield enables alignment transfer to heteronuclei) to enhance the signal of ¹⁵ N molecules. To observe the effects of using SABRE-SHEATH (reversible exchange in shield enables alignment transfer to heteronuclei) to hyperpolarize imidazole- ¹⁵ N ₂ .	¹⁵ N NMR ¹⁵ N NMR	SABRE-SHEATH effectively hyperpolarizes pyridine and nicotinamide (vitamin B ₃ amide), ¹⁵ N containing metabolites. ¹⁵ N NMR signal of imidazole- ¹⁵ N ₂ was enhanced ~2000-fold, and imidazole- ¹⁵ N ₂ could be used for in vivo pH sensing.	Shchepin et al. (2016)

(Continued)

Table 5.2 Applications of 1D NMR techniques in metabolomics. *Continued*

Nucleus	Purpose(s) of Study	NMR Methods	Results	Paper
³¹ P	To explore how using ¹⁵ N labeled choline affects metabolomics experiments (with MS and NMR) and metabolomics analysis.	¹ H NMR ¹ H- ¹⁵ N 2D HSQC	Adding ¹⁵ N labeled choline significantly improves the detection of metabolites containing the carboxyl group (-COOH), using both NMR and MS techniques.	Tayyari et al. (2013)
	To determine the metabolic effects of adding methylseleninic acid (an anti-cancer agent) on A549 lung cancer cells.	1D- ¹ H NMR TOCSY ¹ H- ¹³ C-HSQC ¹ H- ¹⁵ N HSQC	Methylseleninic acid inhibits nucleotide turnover, and the incorporation of nucleotides into RNA.	Fan et al. (2012)
	To improve detection of low concentration metabolites containing carboxyl groups via ¹⁵ N labeling with ethanolamine.	1D ¹ H NMR COSY ¹ H- ¹⁵ N HSQC	Improved identification and quantification of metabolites containing carboxyl groups.	Ye et al. (2009)
	To understand how the metabolism of prostate cancer subtypes vary.	¹ H NMR ³¹ P NMR	Significant differences in metabolite patterns were detected between prostate cancer samples and benign tissue samples.	Dudka et al. (2020)
	To investigate the foliar phosphorous metabolism of trees of a French Guiana rainforest.	³¹ P NMR	Large (59%) differences in phosphorus metabolism between sympatric tree species were observed, indicating that phosphorus metabolism of sympatric trees is species specific.	Gargallo-Garriga et al. (2020)
	To elucidate the metabolic profiles of <i>Aphanizomenon flos-aquae</i> (AFA) cyanobacteria from Klamath Lake (Oregon state, USA).	COSY TOCSY ROESY ³¹ P ¹ H, ³¹ P – HMBC ¹ H, ³¹ P-HSQC-TOCSY ¹ H, ¹³ C – HSQC ¹ H, ¹³ C - HMBC ³¹ P-MRSI (NMR)	Nucleoside 2',3'-cyclic monophosphates contributing to a 20 ppm ³¹ P NMR signal were identified, and fully characterized.	Zambon et al. (2019)
	To determine possibility of non-invasive early detection of breast cancer after the first cycle of neoadjuvant therapy.	¹ H NMR ³¹ P NMR ¹ H, ³¹ P Fast-HMQC	It is possible to observe a slight change in ³¹ P metabolites in breast cancer patients after the first cycle of neoadjuvant therapy.	Krikken et al. (2019)
	To present an alternative to determining pH-sensitive metabolites via NMR.		Following chemical shift changes in only ¹ H or ³¹ P does not yield enough information to determine pH sensitivity. However, if ³¹ P chemical shift changes are combined (with a 2D experiment), identification of pH sensitive metabolites is easier.	Koskela et al. (2018)

	To study the effects of H ₂ O ₂ induced stress upon metabolites in C2C112 myotubules.	¹ H NMR ¹³ C NMR ³¹ P NMR ¹ H, ¹ H COSY ¹ H, ¹³ C HSQC	H ₂ O ₂ induced stress decreased the amount of amino acids (with the exception of alanine), and increased the amount of lactate, indicating that alanine synthesis de novo is likely connected to lactate release from myotubules.	Straadt et al. (2010)
¹⁹ F	To trace the metabolism of leniolisib, a PI3K inhibitor, in healthy controls.	¹⁹ F NMR	Elimination of leniolisib occurred mostly through internal metabolic processes.	Pearson et al. (2019)
	To understand how much the lens accumulates possibly toxic products of vitamin C degradation.	¹⁹ F NMR	Fluoro-dehydroascorbate and ascorbic acid, a fluorine containing metabolite, are taken up into HLE-B3 cells.	Satake et al. (2003)
	To probe ascorbic acid homeostasis and degradation in diabetes.	¹⁹ F NMR	Diabetes pushes ascorbic acid homeostasis towards a higher oxidative state in liver, kidney, spleen, and plasma, but tends to a lower oxidative state in brain, adrenal glands, and heart.	Nishikawa et al. (2003)
	To monitor the plasma and metabolic stability of BLT-F2 and BLT-S-F6, two tumor targeting drug conjugates.	¹ H NMR ¹³ C NMR ¹⁹ F NMR	Both ¹⁹ F NMR probes could be used as metabolic tracing agents in cancer studies.	Seitz et al. (2015)

Select applications of 1D NMR techniques in metabolomics.

therefore can be used to determine the molecular concentration (Cui, Liew et al., 2019; Emwas, Roy et al., 2019).

Owing to these properties of qualification and quantitation, a single ^1H NMR spectrum can often provide enough information to readily identify and elucidate the proportions of anywhere between 30 and 200 metabolites (Bouatra et al., 2013; Emwas, Roy et al., 2019) depending on the nature of the studied sample (biofluids, species, etc.) (Holmes et al., 2008; Li et al., 2017; Tarachiwin et al., 2007; Wei et al., 2009; Yilmaz et al., 2017).

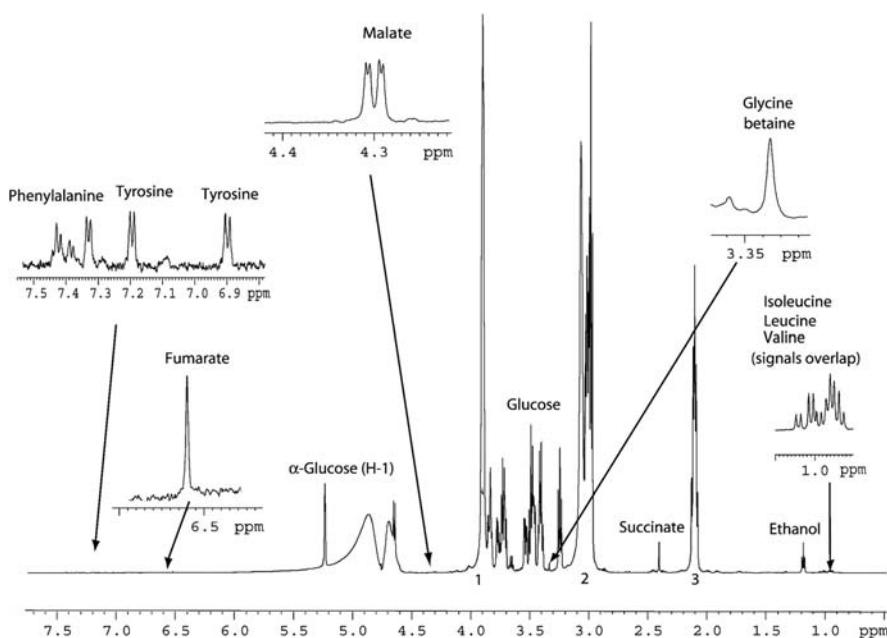
Metabolite identification is usually augmented by tools such as freely available software programs, for example, CAMERA and MetaboMiner (Spicer et al., 2017), commercial software, for example, Chenomx Inc. (Weljie et al., 2006), Mestrelab's Mnova (Mestrelab Research S.L.—Analytical Chemistry Software, 2021), and CRAFT (KCMurthy, 2013), and/or a database such as the Human Metabolome Database (HMDB) (Wishart et al., 2007, 2009, 2018). These tools facilitate metabolite identification from the raw NMR data and assist in organizing/presenting results.

1D ^1H NMR experiments are typically used in untargeted metabolomics studies (Blasco et al., 2014; Emwas, Roy et al., 2019; Flores et al., 2020; Hasanpour et al., 2020; Karaman et al., 2016; Lee et al., 2019; Luke et al., 2020; Madrid-Gambin et al., 2018; Silva et al., 2019; Stringer et al., 2014) as opposed to targeted metabolomic studies. “Untargeted” simply means that all the measurable analytes, including chemical unknowns, undergo a full comprehensive analysis including advanced chemometric techniques (e.g., multivariate analysis). This is necessary to eliminate possible outliers and to make the dataset(s) easier to manage (Roberts et al., 2012). “Targeted” metabolomics refers to the measurement of only defined groups of characterized and annotated metabolites (Roberts et al., 2012).

1D ^1H NMR has a marvelous ability to discriminate (i.e., reveal the differences between) metabolic profiles and quantify their respective metabolites. Fig. 5.1, for example, shows a 1D ^1H NMR spectrum that distinguishes between major nutrients and metabolites of a culture of *Aspergillus fumigatus*, according to their individual peak assignments. The abilities of 1D ^1H NMR to identify and quantify metabolites are perhaps the single greatest advantages 1D ^1H NMR can offer to the field of metabolomics, and arise from the relatively high sensitivity and highly quantifiable nature (Emwas, Roy et al., 2019) of the ^1H nucleus. Representative cases that demonstrate the sensitivity and quantifiability of 1D ^1H NMR are discussed (see below). Some representative additional studies that use 1D ^1H NMR are discussed (see below). Moreover, other studies in which 1D ^1H NMR has been involved (with or without additional techniques) are listed in Table 5.2.

1D ^1H nuclear magnetic resonance examples

Silva et al. (2019) used an untargeted ^1H -NMR based metabolomics approach in order to identify and monitor metabolites specific to breast cancer patients. To do

**FIGURE 5.1**

Standard example of a labeled 1D ^1H NMR spectrum for metabolomics research.

Copied with permission from Plummer, R., Bodkin, J., Power, D., Pantarat, N., Bubb, W. A., Kuchel, P. W., & Sorrell, T. C. (2007). Effect of caspofungin on metabolite profiles of *Aspergillus* species determined by nuclear magnetic resonance spectroscopy. *Antimicrobial Agents and Chemotherapy*, 51 (11), 4077–4084.

<https://doi.org/10.1128/AACO.00602-7>.

so, they took urine samples from 38 healthy controls (HCs), and 40 breast cancer patients, and analyzed the 1D ^1H NMR spectra of the samples. These spectra enabled the rapid identification of 33 metabolites in both study groups. Further multivariate statistical analysis allowed the identification of ten key metabolites (creatinine, glycine, serine, dimethylamine, trimethylamine N-oxide, α -hydroxyisobutyrate, mannitol, glutamine, cis-aconitate, and trigonelline), which showed the highest sensitivity levels and specificity between both study groups (HCs and breast cancer patients); these metabolites could therefore be useful in the diagnosis of breast cancer as they successfully discriminate the urinary profiles of breast cancer patients from those of HCs (Silva et al., 2019). This study is an excellent example of the potential of untargeted 1D ^1H NMR-based metabolomics.

A similar approach was taken by Rocha et al. (2011). In their subsequent study of lung cancer involving blood plasma samples from 85 lung cancer patients (55 males, 30 females) and 78 HCs (38 male, 40 female), the authors successfully identified 36 metabolites. Increased levels of pyruvate, lactate,

VLDL + LDL were observed, along with decreased levels of glucose, citrate, formate, acetate, several amino acids (alanine, glutamine, histidine, tyrosine, valine), methanol, and HDL. These changes could be linked to known characteristics of cancer such as increased glycolysis, glutaminolysis, and gluconeogenesis, a suppressed Krebs cycle, and a decreased catabolism of lipids. Blood plasma samples were taken from a diverse population of lung cancer patients {i.e., male and female, varying histopathology, different TNM stages [tumor (T), nodes (N), and metastases (M)]} and HCs (i.e., smokers and nonsmokers, male and female). Despite the variation and diversity among the samples from the lung cancer patients, 1D ^1H NMR metabolomics consistently discriminated the metabolic profiles of blood plasma samples of lung cancer patients from those of HCs (Rocha et al., 2011).

An earlier study by Jordan et al. (2010) also involved 1D ^1H NMR to distinguish lung cancer patients from HCs. Though they took fewer serum samples (total = 21) from a smaller, less diverse population of subjects [lung cancer patients with either squamous cell carcinoma (SCC) or adenocarcinoma (AC), and HCs] than Rocha et al. (2011), they nonetheless demonstrated the potential of serum NMR metabolomics to discriminate between the two lung cancer types (SCC or AC), as well as between cancer patients and HCs. Although the Jordan study was more limited and less robust than the Rocha study (Rocha et al., 2011), both illustrate the ability of 1D ^1H NMR to differentiate the metabolic profiles of diseased patients from those of HCs, and provided preliminary evidence for 1D ^1H NMR-based metabolomics becoming a key method for disease diagnosis. This method may be especially critical in the earlier stages of the disease.

1D ^1H NMR has also been used to identify and quantify metabolites specific to patients with psychiatric disorders (Sethi et al., 2017; Tasic et al., 2017; Yilmaz et al., 2017). Samples were taken from over 100 individuals [60 HCs, 50 schizophrenia (SCZ) patients, and 45 patients with bipolar disorder (BD)], with the intent to find metabolites that distinguish the three study groups (HC, SCZ, BD) from each other. 1D ^1H NMR spectra allowed Tasic et al. (2019) to identify and quantify metabolites specific to each study group; for example, SCZ patients had a unique presence of isovaleryl carnitine, pantothenate, mannitol, glycine, and GABA, whereas BD patients had 2,3-diphospho-D-glyceric acid, N-acetyl aspartyl-glutamic acid, and monoethyl malonate. SCZ and BD patients both possessed 6-hydroxydopamine, while HC individuals did not. Higher lipid levels were also observed in SCZ and BD as compared to HC individuals (Tasic et al., 2019). Taken together, this approach may support the diagnosis of patients with psychiatric diseases such as SCZ and BD, and the studies lay the foundation for more robust and reliable diagnosis of psychiatric diseases.

Despite several advantages and successes of 1D ^1H NMR in metabolomics, this technique does have drawbacks. The high sensitivity of the ^1H nucleus may cause an “information overload.” The presence of several hundred metabolites in the analyte, each with several NMR signals, may cause severe NMR spectral overlap. Thus NMR is a “double-edged sword” for metabolomics: it provides a

wealth of information that can enable the identification of many metabolites but, at the same time, cripples the identification of individual molecule types.

Fortunately, the effect of spectral overlap can be mitigated and overcome with the use of 2D NMR techniques such as Heteronuclear Single Quantum Coherence (HSQC), Heteronuclear Multiple Bond Correlations (HMBC), and COSY. 2D NMR techniques usually provide enough resolution and resolving power to overcome the problem of peak overlap, and to reveal the connectivity of atoms in greater detail. 2D NMR techniques are also used to confirm the chemical shifts assignments of 1D ^1H NMR spectra. 2D NMR techniques may be used in addition to 1D ^1H NMR experiments (Beckonert et al., 2007; Cao et al., 2020; Feraud et al., 2019) (also see Table 5.2) even if the latter provides enough information to identify the metabolites. Adding information from additional nuclei and/or multidimensional correlation experiments (e.g., ^1H - ^{13}C) comes at the price of extended instrument time (Van et al., 2008) because another variable (i.e., and additional nucleus) is involved (Cavanagh et al., 1995; Claridge, 2016).

For example, Jiang et al. (2013) studied the metabolic changes occurring in plasma, urine and liver extracts from hamsters fed a high-fat/high-cholesterol diet using a 1D ^1H NMR-based approach. Even though over 100 metabolites were identified from just the ^1H spectra of the samples, additional 2D NMR experiments (^1H , ^1H -TOCSY and ^1H , ^{13}C -HSQC) were performed to confirm the assignments (Jiang et al., 2013).

In 2010, Jung et al. also used a 2D NMR technique in addition to 1D ^1H NMR to determine the geographical origin of beef samples (Jung et al., 2010). From ^1H NMR spectra, they identified 25 metabolites. Overlapping signals from a few metabolites were resolved with 2D NMR techniques (^1H , ^1H -TOCSY, ^1H , ^{13}C -HMBC, and ^1H , ^{13}C -HSQC), and the 2D NMR techniques were also used to validate the metabolites identified from the 1D spectra (Jung et al., 2010).

Clearly, using 2D NMR techniques to aid 1D ^1H NMR chemical assignment and to resolve peak overlap could be a standard in metabolomics studies (Martineau & Giraudeau, 2019), however 2D NMR techniques are not common because they take longer to perform (hours to days) (Emwas, Alghrably et al., 2019; Emwas, Roy et al., 2019), and take more experience to process and interpret.

1D ^1H NMR also suffers from the effects of solvent and/or attempts at suppression. Signals from the solvent may completely cover signals from metabolites. The concentration of ^1H in pure water (H_2O) is $\sim 110 \text{ mol L}^{-1}$, which is significantly higher than the concentration of metabolites (nM–mM) in the water solution (for example). As most metabolomics studies are carried out in water (Emwas, Roy et al., 2019), solvent suppression becomes mandatory, either by attempting to deuterate the solvent (i.e., replace H_2O with highly pure D_2O), and/or by applying an NMR pulse sequence. Much work has gone into overcoming the effects of solvent suppression and the results can be seen in Giraudeau et al. (2015), McKay (2009), and McKay (2011).

1D ^{13}C nuclear magnetic resonance in metabolomic studies

In terms of inherent resolution and reduced spectral overlap, 1D ^{13}C NMR is far superior to 1D ^1H NMR. ^{13}C NMR has a much broader observed chemical shift range (~ 200 ppm) than that of ^1H (~ 10 ppm) (Edison et al., 2019), and therefore offers better resolution than 1D ^1H NMR. This leads to higher quality spectra with narrow line width peaks (Emwas, Roy et al., 2019), and facilitates the identification of organic functional groups (i.e., aromatics, aliphatics, and carbonyl-containing compounds). This is especially true for metabolites and most organic molecules containing a carbon backbone. Hence, 1D ^{13}C NMR provides a direct way to measure the primary structures of many metabolites (Clendinen et al., 2014). Furthermore, 1D ^{13}C NMR signals are less sensitive to environmental changes such as pH (Edison et al., 2019), unlike the 1D ^1H NMR peak frequencies, which are more sensitive to changes in pH (Dona et al., 2016; Tredwell et al., 2016; Tynkkynen et al., 2009). The lack of ^{13}C to ^{13}C homonuclear coupling at natural abundance also makes ^{13}C spectra easier to interpret (Edison et al., 2019).

While these advantages may seem to make ^{13}C spectra an obvious first choice, 1D ^{13}C NMR suffers from inherently low sensitivity. The ^{13}C isotope is 62 times (cube of the gyromagnetic ratio difference) less sensitive than the ^1H nucleus, and the natural abundance of ^{13}C is far lower (99.99% for ^1H , $\sim 1.1\%$ for ^{13}C). The absolute receptivity difference is more than 5717-fold. This inherently low sensitivity of the ^{13}C nucleus, combined with its low natural abundance, significantly impedes its use in metabolomics studies. There are several avenues to overcome this weakness. Metabolites can be isotopically enriched with ^{13}C nuclei; the most straightforward approach to overcoming the low natural abundance (Clendinen et al., 2015). This approach is most beneficial to 2D correlation-based NMR techniques (e.g., INADEQUATE, HSQC, and HMBC) that involve the ^{13}C nucleus (Clendinen et al., 2015; Geier et al., 2019; Lewis et al., 2010; Otto et al., 2015; Pan et al., 2016). Isotopic enrichment (often termed labeling) is usually accomplished by feeding the organism of interest (e.g., bacteria) with ^{13}C -enriched nutrients (e.g., U- $^{13}\text{C}_6$ glucose, or fully labeled maximal media) (Malloy et al., 2010).

A new indirect and useful method to detect ^{13}C at low and/or natural abundance concentrations involves hyperpolarization or a variant of such. Hyperpolarization implies transferring the nuclear spins of a sensitive nucleus (such as ^1H) to a less sensitive nucleus (such as ^{13}C) (Emwas et al., 2008; Hill et al., 2018; Kovtunov et al., 2018; Ludwig et al., 2010), thus increasing the overall sensitivity of the less sensitive nuclei, often by a theoretical factor of 10^4 – 10^5 (Altes & Salerno, 2004), although practical applications rarely achieve this ideal factor.

This in turn can reduce a standard 1D ^{13}C NMR experiment from several hours to several minutes, an acquisition time comparable to that of the standard 1D ^1H NMR experiment (Emwas, Roy et al., 2019). As mentioned, this

theoretical enhancement is rarely fully achieved. Hyperpolarization is, however, still used to enhance the readability of ^{13}C at natural abundance ($\sim 1.1\%$), especially for *in vivo* metabolic imaging in which the hyperpolarized metabolite and its concentration can be tracked in real-time (Walker & Happer, 1997; Wang et al., 2019). Methods to hyperpolarize ^{13}C metabolites include PHIP (Parahydrogen Induced Polarization), SABRE (Signal Amplification by Reversible Exchange), and DNP (Dynamic Nuclear Polarization). It must be noted that hyperpolarization methods are invasive, and this must be taken into account when considering subsequent sample testing. Applications of each technique are discussed below, and more applications are listed in Table 5.2.

Zacharias et al. (2016) employed PHIP to hyperpolarize ^{13}C labeled succinate (SUC) and its derivative diethyl succinate (DES). This allowed the authors to monitor the uptake and cellular conversion (or lack thereof) of SUC and DES in five cancer allograft animal models: breast (4T1), Renal Cell Carcinoma (RENCA), colon (CT26), lymphoma NSO, and lymphoma A20 via ^{13}C MRS and MRI (both essentially inductive magnetic resonance techniques similar to NMR). They found that RENCA metabolized SUC and DES, while the other cancers did not (Zacharias et al., 2016). In a related study, Zacharias et al. (2012) created hyperpolarized DES- $1-^{13}\text{C}-2,3-d_2$ (via the PHIP method) to monitor the Krebs cycle in real-time. Downstream metabolites (malate, succinate, fumarate, and aspartate) of hyperpolarized DES were identified *in vivo* via high resolution 1D ^{13}C NMR spectra. Both studies serve as powerful indicators of the large capacity of ^{13}C NMR (with ^{13}C molecules metabolites hyperpolarized via PHIP) to track metabolites related to diseases such as cancer in real-time. This could have enormous applications in the health and pharmaceutical industries.

Similar to PHIP, SABRE is another hyperpolarization technique to increase the sensitivity of ^{13}C nuclei. In a novel study involving a low concentration 4-methylpyridine [4MP, a metabolite involved in bacteria metabolism (Khasaeva et al., 2016)], Richardson et al. (2018) used the SABRE technique to enhance their ^{13}C NMR signal of 4-methylpyridine (4MP). Without SABRE, the total time required to collect the ^{13}C signal of 4MP was 52 hours (4096 cumulative scans). With SABRE, the same experiment took only 15 seconds (i.e., 1 scan). The concentration of 4-MP for both experiments was in the millimolar range, and both experiments were performed on a benchtop NMR spectrometer (BNMR) with a low field strength (Richardson et al., 2018). Clearly, SABRE is a powerful tool to increase ^{13}C sensitivity, and to significantly reduce total experimental time for metabolite measurements.

An interesting case study of SABRE comes from the work of Lloyd et al. (2012). Using para-hydrogen as the polarization source, they hyperpolarized quinoline, a metabolite mostly found in plants (Diaz et al., 2015). Of note, the ^{13}C spectrum of the quinoline molecule typically shows good resolution and a large peak distribution. 1D and 2D NMR techniques were performed with the SABRE technique and resulted in high resolution NMR spectra. The time to record these experiments was reduced significantly from several hours to several minutes, and

it is important to note that all quinolone samples had low concentrations (μM — mM) (Lloyd et al., 2012); this study therefore establishes the mighty potential of hyperpolarization in metabolomics studies.

Although PHIP and SABRE are prime hyperpolarization techniques with enormous potential for metabolomics applications, DNP is much more common in the scientific literature than either PHIP or SABRE. Indeed, a number of key metabolites ($1-\text{}^{13}\text{C}$ -pyruvic acid, ^{13}C -bicarbonate, $1-\text{}^{13}\text{C}$ -fumarate, and $5-\text{}^{13}\text{C}$ -glutamine) have been successfully hyperpolarized via DNP. These metabolites, however, were already isotopically enriched with ^{13}C (Nikolaou et al., 2015). DNP however still works effectively for 1D ^{13}C NMR experiments at natural ^{13}C abundance.

An excellent example of using DNP at natural ^{13}C abundance comes from the work of Dey et al. (2020). They utilized DNP to hyperpolarize plant metabolites of red and green tomato plant extract. This made it possible to collect clean and well-resolved 1D ^{13}C NMR spectra (1 scan each) from both sets of plant extracts (red tomato vs. green tomato). Additional statistical analysis of ^{13}C NMR data was able to distinguish between both sets of plant extracts (Dey et al., 2020). This study effectively showed that DNP worked well to hyperpolarize and enhance the sensitivity of ^{13}C NMR at natural abundance. It also demonstrates that DNP is suitable for distinguishing metabolic profiles of similar species.

DEPT (distortions enhancement by polarization transfer) NMR is also a useful technique in NMR metabolomics that shows how other nuclei are coupled to the carbon nucleus. DEPT is used to distinguish between a CH_3 group (methyl), a CH_2 group (methylene), and a CH group (methine). DEPT has a pulse set at 45, 90, or 135 degrees in three separate experiments. DEPT can increase the sensitivity of ^{13}C by a factor of 4, and is therefore useful in metabolomics studies.

For example, Kamal et al. (2012), found four bioactive metabolites from a culture medium of a bacterial strain of a new *Pseudomonas* sp. that had antimicrobial and biosurfactant activities, and used DEPT-135 (along with other NMR methods) to identify the metabolites as 1-hydroxyphenazine, phenazine-1-carboxylic acid, rhamnolipid-1, and rhamnolipid-2. Li et al. (2019) used DEPT as part of their proposal to statistically correlate NMR spectra with LC–MS data, facilitating metabolite structure identification.

Hyperpolarization techniques do suffer from some limitations, one being that sample or NMR probe deterioration can occur as a result of the irradiation pulses that are required for hyperpolarization techniques to work (Richardson et al., 2018). A brief summary of the advantages and disadvantages of the hyperpolarization techniques presented here (PHIP, SABRE, and DNP) is provided in Table 5.2. More studies involving hyperpolarized ^{13}C metabolites are listed in Table 5.2. For those desiring a deeper level of theory regarding hyperpolarization methods and how they work, they are referred to the following published manuscripts (Barskiy et al., 2017; Halse, 2016; Kovtunov et al., 2018; Meier et al., 2014; Nikolaou et al., 2015).

Aside from isotopic enrichment and the application of hyperpolarization techniques, another practical approach to improving ^{13}C sensitivity for 1D ^{13}C NMR

experiments involves the use of enhanced NMR equipment. For example, NMR cryoprobe technology can be used in the probe and electronics. This involves reducing the temperature of the equipment down to $\sim 20\text{K}$ as a way to reduce electronic noise, and results in a two to fourfold improvement in signal to noise (Emwas, Roy et al., 2019). One can also use higher magnetic field strengths, but ^{13}C (and other lower gyromagnetic ratio nuclei) is relatively insensitive to the magnetic field strength (Halse, 2016).

1D ^{15}N nuclear magnetic resonance in metabolomics

As with ^{13}C NMR, ^{15}N also has a large chemical shift dispersion ($\sim 100\text{ ppm}$) (Emwas, Roy et al., 2019) allowing for improved resolution and identification of nitrogen-containing compounds. ^{15}N however has a low natural abundance ($\sim 0.37\%$, even lower than that of ^{13}C), and an even lower sensitivity than that of ^{13}C . Direct observation of ^{15}N nuclei using a standard 1D ^{15}N NMR experiment is difficult and time consuming unless steps are taken to mitigate the physical features. A number of important metabolites, for example, amino acids, alkaloids, purines, pyrimidines, and terpenoids (Ramirez et al., 2019; Song et al., 2020) contain nitrogen, and the ability to study these atoms can provide a more complete picture of the metabolome (Bhinderwala, Lonergan et al., 2018; Kanamori, 2017).

Chemically tagging metabolites with an ^{15}N labeled compound is one way to increase the sensitivity, and this has been demonstrated successfully by Tayyari et al. (2013) and Ye et al. (2009). Chemically tagging metabolites with ^{15}N labeled atoms is mostly beneficial to 2D ^{15}N -NMR techniques and is rarely, if ever, applied to 1D ^{15}N NMR due to cost and synthesis/expression considerations.

Direct observation of ^{15}N containing metabolites is mostly done with hyperpolarization techniques (similar for ^{13}C), as hyperpolarization can significantly improve the sensitivity and observability of ^{15}N containing metabolites. In fact, hyperpolarization seems to be the method of choice for 1D ^{15}N NMR when it comes to metabolomics.

Barskiy et al. (2016), for example, used SABRE-SHEATH (SABRE in SHield Enables Alignment Transfer to Heteronuclei) to directly hyperpolarize the ^{15}N sites of metronidazole, a precursor of two major oxidative products, an acid metabolite and a hydroxy metabolite (Pendland et al., 1994). By using SABRE-SHEATH, the authors were able to hyperpolarize the ^{15}N nuclei in metronidazole to over 20% in less than one minute, thus increasing the sensitivity of ^{15}N in metronidazole and its observability via 1D ^{15}N NMR at low concentration (50 mM) in a short amount of time (Kanamori, 2017). An increase in metronidazole nitrogen sensitivity could have useful applications such as direct *in vivo* imaging of mechanisms of action or hypoxia sensing, as metronidazole is an antimicrobial drug against species such as *Entamoeba histolytica*, *Giardia lamblia*, and *Trichomonas vaginalis* (Freeman et al., 1997). A similar study by Shchepin et al. (2019) also examined the effects of hyperpolarizing metronidazole via SABRE-SHEATH. During the process, they were able to transfer the hyperpolarization of iridium hydrides to a distance of up to six chemical bonds. The polarization level

of metronidazole achieved by Shchepin et al. ($\sim 15\%$) (Shchepin et al., 2019) was smaller than that achieved by Barskiy et al. ($\sim 20\%$) (Barskiy et al., 2016), and yet both groups of researchers achieved their respective levels of metronidazole hyperpolarization in under one minute. These two studies demonstrate the potential of hyperpolarization-based techniques to increase the observability of 1D ^{15}N NMR spectra of ^{15}N containing metabolites, which could also eventually transform 1D ^{15}N NMR into a live, real-time metabolic imaging tool.

Hyperpolarization could also enable the molecular imaging of viruses, as demonstrated in the study by Shchepin et al. (Shchepin & Chekmenev, 2014). They measured the T_1 relaxation times of hyperpolarized ^{15}N labeled Azidothymidine (AZT), an antiretroviral medication used to prevent and treat HIV/AIDS (Eckhardt et al., 2017). In 2014, they also determined that NMR T_1 relaxation times were sufficiently long to allow *in vivo* imaging of ^{15}N -AZT and its relevant kinetics after injection into an HIV-infected patient. Strictly speaking, Shchepin et al. did not generate 1D ^{15}N NMR spectra of ^{15}N -AZT; nevertheless, the long T_1 times measured indicated that the 1D ^{15}N NMR spectra of hyperpolarized ^{15}N -AZT would be of higher quality than those of nonpolarized ^{15}N AZT. This study thus points to the potential of hyperpolarized ^{15}N -containing molecules and metabolites in analyzing and monitoring the “metabolism” of viruses or even virus-like particles.

Additional examples of hyperpolarized ^{15}N NMR experiments are listed in Table 5.2. We hope the readers agree that hyperpolarization of ^{15}N containing metabolites is valuable, and perhaps crucial for future metabolomics applications.

^{31}P nuclear magnetic resonance in metabolomic studies

1D ^{31}P NMR has a unique niche for structural elucidation and metabolomics studies (Babgi et al., 2021; Tomah Al-Masri et al., 2012). Phosphate metabolism is quite diverse across different living systems and organs (e.g., liver, muscle tissues, and kidney) (Felsenfeld & Levine, 2015; Gattineni & Friedman, 2015; Moe & Daoud, 2014; Quinn, 2012; Silver et al., 2002; Tebben et al., 2013; Uday et al., 2019) and several phosphorus-containing metabolites [especially those that transport the inorganic phosphate group (Bhinderwala et al., 2020)] are vital intermediaries and regulators of essential biochemical pathways, including glycolysis (Berg et al., 2002), the Krebs cycle (Enderle, 2012), and fatty acid β -oxidation (Bosc et al., 2020). As such, 1D ^{31}P NMR is a powerful method to study metabolic profiles, and to expand the coverage of the metabolome (Bhinderwala et al., 2020). ^{31}P has a high natural abundance (100%), and a relatively good sensitivity (lower than that of ^1H , but higher than those of ^{13}C or ^{15}N), and ^{31}P NMR does not usually require an external supplemental source (Bhinderwala et al., 2020).

Despite these advantages, ^{31}P NMR suffers from severe drawbacks (listed in Table 5.1) which have limited its applicability in metabolomics (Emwas, Roy et al., 2019). The ^{31}P nucleus is extremely sensitive to pH changes (Blaive et al., 2000; Zheng, Liu et al., 2017), solvent selection, and the experimental temperature. ^{31}P resonances also have a limited chemical shift which dispersion tends to

be broad and causes obscured peak splitting (Bhinderwala et al., 2020). 1D ^{31}P NMR also experiences signal overlap from phosphorylated compounds (Markley et al., 2017).

1D ^{31}P reference NMR spectra are generally lacking in databases such as the HMDB, which further limits the use of ^{31}P NMR in metabolomics applications (Bhinderwala et al., 2020). However, the scientific literature contains a plethora of ^{31}P NMR studies applied to metabolomics (Buchli et al., 1994; Cady et al., 1983; Carlbom et al., 2017; Chorao et al., 2010; Gout et al., 2011; Komoroski et al., 2008; Levine et al., 2003; Park & Park, 2001; Qiao et al., 2006; Shah et al., 2014; Sterin et al., 2001; Thebault et al., 2009; Tiret et al., 2016; Tokumaru et al., 2009; Vauclare et al., 2013; Wijnen et al., 2012). Many studies have employed 1D ^{31}P NMR as an *in vivo* metabolic imaging technique. A couple of examples demonstrating the use of 1D ^{31}P NMR in metabolomics are discussed below. Additional examples of 1D ^{31}P NMR in metabolomics are listed in Table 5.2.

Sterin et al. (2001) used ^{31}P spectra to test the effects of antimitotic drugs on breast cancer cell metabolism. Analysis of observations revealed a correlation between the mechanism of action of selected anticancer drugs (i.e., paclitaxel, vincristine, colchicine, nocodazole, methotrexate, and doxorubicin), and observed differences in breast cancer cell metabolism [i.e., hormonal response, estrogen receptors (positive/negative), and metastatic potential]. Specifically, Sterin et al. discovered that the antimicrotubule drugs (paclitaxel, vincristine, colchicine, and nocodazole) increased the amount of intracellular glycerophosphorylcholine (GPC), an intracellular metabolite, whereas the nonantimicrotubule drugs (methotrexate and Adriamycin) did not. These results suggest that the level of intracellular GPC is indicative of cellular microtubule functionality (Sterin et al., 2001), and demonstrate that ^{31}P NMR has the ability and potential to determine the effects of drugs on the target of interest. This could also extend beyond cancer cells to targets such as proteins.

Levine et al. (2003) showed that ^{31}P NMR can be extended to other living systems such as fish. They used 1D ^{31}P NMR to assess how adding acetyl-L-carnitine (ALCAR) and myo-inositol influenced the levels of phosphate-containing metabolites in zebrafish. Zebrafish present a unique model for studying high-energy phosphate and membrane phospholipid metabolism in living systems. Addition of ALCAR and myo-inositol decreased the levels of phosphodiesters and inorganic orthophosphate, and increased levels of phosphocreatine in the zebrafish (Levine et al., 2003). This is important because ALCAR and myo-inositol are both antidepressant drugs (Chiechio et al., 2018; Nasca et al., 2013; Nemets et al., 2001), and monitoring the metabolic profiles and effects of antidepressants or any other type of drug via 1D ^{31}P NMR may allow scientists to obtain a more complete picture of how these drugs affect patients. This is especially true of ^{31}P NMR, which has been, and most likely will continue to be, used as an *in vivo* metabolic imaging tool of living systems.

Past scientific work has proven the utility of 1D ^{31}P NMR in *in vivo* metabolic imaging, and other meaningful applications. However, more work, such as

the creation of more 1D ^{31}P NMR reference spectra (Bhinderwala et al., 2020), is required to make 1D ^{31}P NMR a standardized tool for metabolomics research. The results of such work will be worthwhile and meaningful for a number of applications, which will most likely have patient benefits (e.g., the health industry will be better able to monitor the effect of a drug on patients with the use of 1D ^{31}P NMR).

^{19}F in metabolomic studies

^{19}F NMR studies are rarer than any of the previously mentioned 1D NMR experiments, but still deserve a mention. ^{19}F has a sensitivity akin to that of ^1H , and like ^{31}P , has a natural abundance of 100%. ^{19}F is not common in naturally occurring metabolites, but a number of drugs containing a fluorine atom and their respective metabolisms have been studied in detail (Park et al., 2001).

For example, Pawłowski et al. (2019) combined ^{19}F NMR and in silico techniques to study 5-fluorouracil (5-FU) (an anticancer drug containing fluorine) metabolism in yeast under low ATP conditions. Lutz and Hull (1999) used 1D ^{19}F NMR to study how some of 5-FU's anabolites [5-fluorouracil (FUra), 5-fluorouridine (FUr), 5-fluoro-2'-deoxyuridine (FdUr), 5-fluorouridine-5'-monophosphate (FUMP), FdUMP, 5-fluorouridine-5'-diphosphate (FUDP), FUTP and 5-fluorouridine-5'-diphospho(1)- α -D-glucose (FUDPG)] responded to different pH values. These anabolites (anabolic metabolites) are key to fluoropyrimidine chemotherapy in cancer treatment (Lutz & Hull, 1999), and thus an understanding of their behavior under different environmental conditions may prove useful for future treatment options.

1D ^{19}F NMR may find additional applications in environmental studies, as ^{19}F is a convenient way to study the biodegradation of environmental pollutants, and to obtain a quick initial scan of the metabolic profiles of newly isolated organisms (Boersma et al., 2001). Nevertheless, future studies are needed to fully appreciate and expand the use of 1D ^{19}F NMR in metabolomics studies.

2D nuclear magnetic resonance spectroscopy

Mono-dimensional (1D) proton (^1H) NMR spectra from complex biological samples like blood and urine or plant extracts are crowded with overlapping peaks of hundreds to thousands of molecules, making it difficult, if not impossible, to obtain an accurate peak assignment and disentangle the information content attributable to each molecule.

Multidimensional NMR experiments, and particularly two-dimensional (2D NMR), can be used for peak assignment and structural determination of compounds (Dona et al., 2016; Emwas, Roy et al., 2019). The use of 2D experiments is not widespread in metabolomics mostly because the acquisition of a 2D spectrum typically need the repetition of several hundreds of 1D experiments, leading to acquisition times between a few tens of minutes and several hours, depending on the type of experiment (Emwas, Roy et al., 2019; Giraudeau, 2020). This

makes it impossible to use 2D experiments in a high-throughput setting. However, over the years several approaches have been proposed to accelerate the acquisition of 2D spectra (Giraudieu, 2020; Rouger et al., 2017) including fast repetition techniques (Farjon, 2017), spectral aliasing (Vitorge et al., 2009), non-uniform sampling (Mobli & Hoch, 2014), Hadamard (Kupče & Freeman, 2003), and UF (Frydman et al., 2002) spectroscopy.

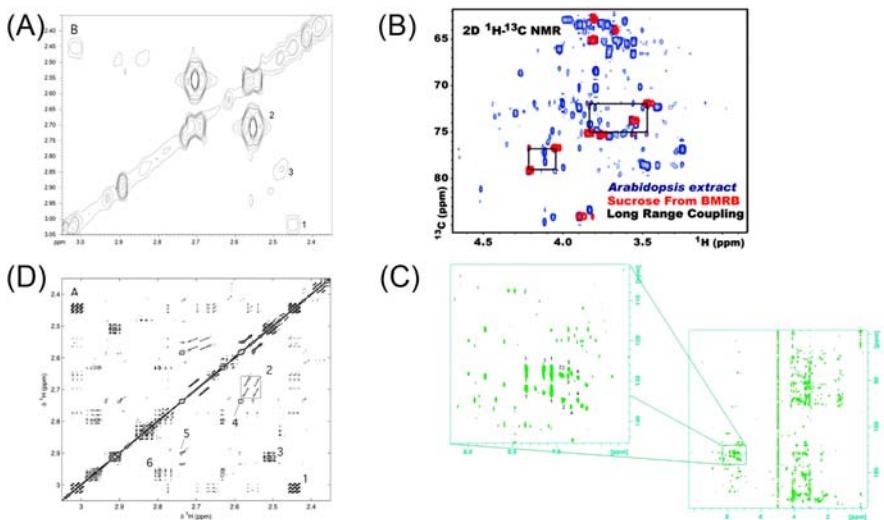
Several 2D NMR experiments are available and have been used in metabolomics for several applications including molecular identification, structural elucidation, and kinetic or energetic analysis (Chu et al., 2010; Emwas, Al-Talla et al., 2013; Sahoo et al., 2020). 2D NMR experiments can be used to overcome the problem of overlapping resonances by spreading a different type of information about peaks onto a second dimension. Correlation methods exploit chemical shifts from covalently attached neighbors, while other methods exploit other chemo physical properties, like the coupling constant J (which contains information about relative bond distances and angles) or the diffusion time (which is the time need by a molecule to diffuse along the magnetic) (Emwas, Roy et al., 2019).

The correlation spectroscopy (COSY) (Alonso et al., 1989) experiment is the simplest of all 2D NMR experiments and provides information on homonuclear correlations between coupled nuclei (^1H - ^1H). It has been widely used for molecular identification and structural elucidation (Hunt et al., 1984; Kono, 2013; Lown & Hanstock, 1985; Macura et al., 1983) and has proven to be particularly useful for metabolomics research (Blasco et al., 2010; Flores-Sanchez et al., 2012; Kim et al., 2010; Le Guennec et al., 2012; Sekiyama et al., 2011).

TOCSY (total COSY), also known as the homonuclear Hartmann–Hahn experiment (Braunschweiler & Ernst, 1983) is an extension of the COSY experiment wherein the chemical shift of a given nucleus is correlated with the chemical shift of other nuclei within the total (or near total) spin system of a given compound. Typical applications of 2D-TOCSY are the structural elucidation of carbohydrates and peptides since all protons belonging to the same sugar residue or to a single amino acid will appear correlated (Johnson et al., 1995), or metabolite identification. An example of TOCSY is given in Fig. 5.2A.

COSY (like COSY and TOCSY-like spectroscopy) is not limited to homonuclear correlations; therefore, it can also be used for measuring heteronuclear correlations, that is, plotting the chemical shift from the ^1H against the chemical shift of other atoms (like ^{13}C or ^{15}N) as in the HSQC and HMBC.

In an ^1H - ^{13}C -HSQC spectrum (Bodenhausen & Ruben, 1980), the chemical shifts of proton and carbon atoms that are directly bonded are mapped, providing only one cross-peak for each H–C coupled pair. Similarly, an ^1H - ^{13}C -HSQC maps proton and carbon atoms that are directly bonded. HSQC experiments are very useful for resolving and assigning overlapping proton signals, particularly for metabolite signals arising from complex biofluid mixtures (Emwas, Roy et al., 2019). An example of an H - ^{13}C -HSQC spectrum of sucrose overlaid onto an *Arabidopsis* extract is given in Fig. 5.2B.

**FIGURE 5.2**

(A) Example of TOCSY NMR spectrum. Example of TOCSY NMR spectrum obtained from a single mouse urine sample from the 3–2.4 ppm chemical shift region: keys: (1) 2-oxoglutarate; (2) citrate; (3) 3-hydroxyphenylpropionate; (4) methylamine and dimethylamine correlation; (5) dimethylamine and trimethylamine correlation. (B) ^1H – ^{13}C HSQC NMR spectrum of sucrose from the Biological Magnetic Resonance Data Bank (BMRDB) (Ulrich et al., 2007) (red) overlaid onto an aqueous whole-plant extract from *Arabidopsis thaliana* (blue) (Lewis et al., 2007). (C) 2D HMBC spectrum of a urine sample acquired at 900 MHz. The zoomed zone represents the aromatic region. Aromatic peaks of Hippurate (1), PAG (2), Histidine (3) and mHPPA (4) can be distinguished. (D) Two-dimensional representation of STOCSY of ^1H NMR spectra obtained from 599 urine samples; numeric key as for panel (A).

(A and D) Adapted from Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., & Holmes, E. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Analytical Chemistry*, 77(5), 1282–1289; (B) Adapted from Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G., Raftery, D., Alahmari, F., Jaremko, L., & Jaremko, M. (2019). NMR spectroscopy for metabolomics research. *Metabolites*, 9(7), 123; (C) From Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schäfer, H., Schuțz, B., Spraul, M., & Tenori, L. (2009). Individual human phenotypes in metabolic space and time. *Journal of Proteome Research*, 8(9), 4264–4271.

Heteronuclear multiple-quantum correlation spectroscopy (HMQC) is similar to HSBQ. It provides correlation, as in HSQC, and the two methods give similar quality results for small to medium-sized molecules, but HSQC is more appropriate for larger molecules (Keeler, 2011).

HMBC (Bax & Summers, 1986) is similar to HSQC but the HMBC experiment reveals correlations between nuclei that are separated by two or more

chemical bonds. The utility of ^1H - ^{13}C HMBC in metabolomics has been shown by the possibility of easily distinguishing some of the aromatic peaks of hippurate, phenylacetylglycine, and histidine in urine samples (Bernini et al., 2009), as shown in Fig. 5.2C.

The combined use of HSQC and HMBC has proven to be particularly useful in metabolomics for the identification of new compounds from plant extracts (Liang et al., 2006), spider venom (Taggi et al., 2004), and insects (Dossey et al., 2007).

J-resolved (JRES) experiments (Aue et al., 1976) are by far the most used in metabolomics because of their simplicity and short acquisition time (Giraudeau, 2020; Ludwig & Viant, 2010; Mahrous & Farag, 2015). Through JRES experiment, simplified projection of the proton spectrum, in which all peaks from a multiplet appear as a singlet, is obtained by plotting the proton spectrum along one dimension and the coupling constant (*J* value) of each signal along the second dimension. JRES experiments have been applied to resolve overlapping resonances of metabolites identification in human biofluids such as urine, blood plasma, and cerebral spinal fluid (Foxall et al., 1993; Lutz et al., 1998; Yang et al., 2008). Moreover, JRES spectra connection between neighboring protons can be established, and information about the *J* value of each signal can be used to distinguish between some isomers such as α and β anomers of sugars and glycosides or *cis* and *trans* isomers of olefinic compounds (Mahrous & Farag, 2015).

In 2D diffusion-ordered (DOSY) experiments (Stilbs, 1987), a similar approach is used by plotting the proton spectrum along one dimension and the diffusion coefficient related to each NMR signal along the second dimension. Since molecules with different molecular weights have different diffusion coefficients, it is in principle possible to identify them. It is possible to obtain a good degree of separation between compounds that differ substantially in their molecular weights (Mahrous & Farag, 2015). 2D-DOSY have been used in the assignment of the anomeric protons of mono-, di-, or oligosaccharides in apple and grape juices (Gil et al., 2004).

Building on the concepts of the TOCSY experiment, an idea has been proposed to exploit the statistical correlation among peaks that can be calculated using multiple mono-dimensional. The Statistical TOCSY (STOCSY) (Cloarec et al., 2005) uses multicollinearity of the intensity variables in a set of 1-D spectra to generate a pseudo-two-dimensional NMR spectrum that displays the correlation among the intensities of the various peaks across the whole sample. This has been used to assign and identify relevant metabolites in a metabolomic study of a model of insulin resistance (Cloarec et al., 2005). An example of STOCSY obtained from a stack of X samples is given in Fig. 5.2D and can be compared with a TOCSY spectrum obtained from one sample, Fig. 5.2C. The method can be expanded to statistical total correlation spectroscopy editing (STOCSY-E) (Sands et al., 2009) and to the hetero-nuclear HET-STOCSY by combining mono-dimensional ^{31}P and ^1H NMR signals and used for biomarker detection.

In the case of 2D spectra, there is no linear correlation between the molar concentration of any compound and the area under the curve of the corresponding

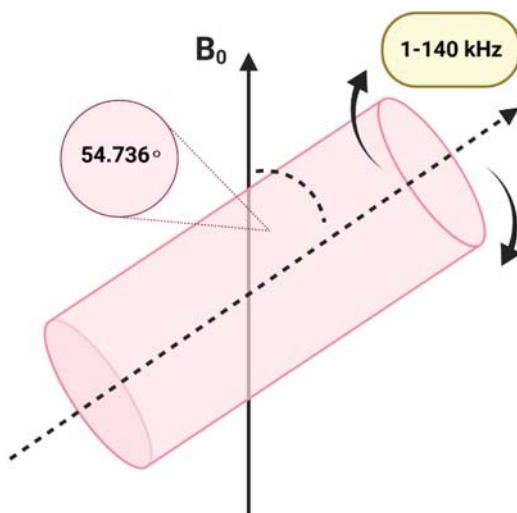
signals because other factors like the resonance specific signal attenuation contribute to the intensity of the cross-peaks and their corresponding integration volumes (Mahrous & Farag, 2015). Several approaches have been proposed to extract quantitative information from 2D spectra (Giraudeau et al., 2007; Lewis et al., 2007; Michel & Akoka, 2004; Parsons et al., 2007; Pathan et al., 2011). However, none of the methods is routinely applied in metabolomics for quantification purposes.

Contextually, 2D spectra have seldom been used directly for (multivariate) data analysis since this kind of analysis needs either to transform 2D spectra to 1D spectra or the use of a multivariate approach that can handle three-dimensional data since there is a data matrix (the 2D spectra) for each sample. However, some applications have been proposed for the use of NMR 2D spectra showing the value of using 2D (Bertelli et al., 2010; Brinson et al., 2020; Farag et al., 2014; Sharma et al., 2017) or providing some guidelines (Brinson et al., 2020).

High-resolution magic-angle spinning nuclear magnetic resonance spectroscopy

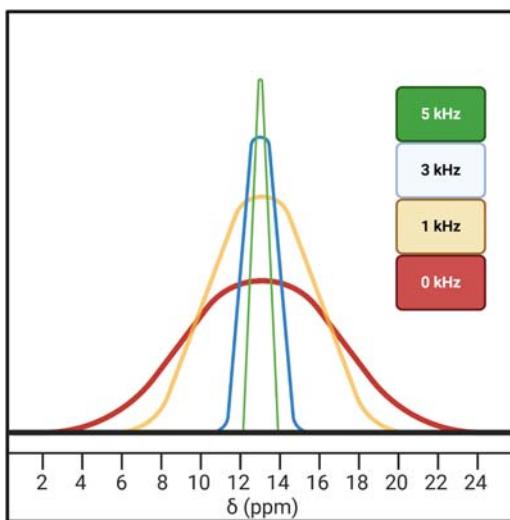
High-Resolution Magic-Angle Spinning NMR Spectroscopy (HR-MAS) is a versatile tool that utilizes the MAS to measure suspended solid- or gel-like samples that contain some internal molecular motions (Händel et al., 2003). In MAS NMR [similar to solid-state NMR (ssNMR)], in order to obtain high resolution information from a spectra, the sample has to be positioned in a rotor that rapidly rotates at the fixed angle of 54.736 degrees (the so-called “magic angle”) with respect to the Z-axis of the static magnetic field (Fig. 5.3). This rotation is required to average the anisotropic spin interactions and remove the dipolar coupling, which normally cause line-broadening effects on the spectra, as shown in Fig. 5.4 (Ashbrook et al., 2018; Beckonert et al., 2010; Watts, 2005). For example, tissue samples are usually spun at approximately 4–6 kHz. Under this kind of condition, the acquisition of NMR spectra of the tissue is comparable with solution state NMR (Beckonert et al., 2010). The problem however lies in the preservation of the tissue: the speed of rotation affects the morphology and metabolite profile of the sample (Gogiashvili et al., 2019). For example, two hours of rotation at 3 kHz destroys 15%–19% of adipocytes (Weybright et al., 2005). The amount of sample used can vary from a standard 65 µL that gives an improved signal to noise ratio, however the required size could be reduced down to 12 µL by using spacers. The use of spacers improves the field homogeneity and the peak line shape due to the compact spherical sample volume (Beckonert et al., 2010). Currently, MAS probes enable metabolomics researchers to measure different types of nuclei such as ^1H , ^{13}C , ^{15}N and ^{31}P for human and animal tissue samples (Esmaeili et al., 2014; Levin et al., 2009; McNally et al., 2006).

The application of HR-MAS was found to be useful in an “integrative metabolomics” approach when investigating metabolic profiles of different tissue types

**FIGURE 5.3**

MAS NMR. MAS NMR rotor with an angle of 54.736 degrees with respect to the z axis of the static magnetic field. The rotor is spinning with a frequency of 1 for up to 140 kHz (Lin et al., 2018), which diminishes the line broadening effects.

Created with BioRender.com.

**FIGURE 5.4**

Rotating speed effect on NMR spectral line widths. The effects of different rotating speeds on resulting NMR spectral line widths. The removal of dipolar coupling combined with averaging anisotropic spin interactions results in narrower resonances and therefore higher resolution spectra.

Created with BioRender.com.

and bio fluids (Beckonert et al., 2010). One of the examples is the work done by Chan et al. (2009) in which they analyzed colon mucosae in order to define metabolites that could help to discriminate malignant (colorectal cancer) forms of mucosae from normal mucosae. They utilized HR-MAS NMR (1D NOESY and 1D Carr-Purcell-Meiboom-Gill) and GC-MS to profile tumor specimens and compare them with normal mucosae. The results obtained from these experiments revealed a higher level of both saturated and unsaturated lipids and/or fatty acids in normal mucosae compared to malignant samples, and an increase in levels of lactate and glycine with the decrease of glucose and arachidonic acid levels in tumor specimens (Chan et al., 2009).

Another prominent example of HR-MAS in practice is the use of ^1H HR-MAS NMR to investigate human bone cancer biology. The work done by Tavel et al. (2016) focused on multiple myeloma—a type of bone marrow cancer. The goal of the study was to gain insight and understand the rules behind the genetic heterogeneity within the same tumor lesion. For that, the authors combined ^1H HR-MAS, and multivariate analysis, and complemented this with information obtained through histo-morphological observations. Although the differences between lipids in the two morphologically distinct types of the tumor (termed as “oily” and “calcified”) were not significant, the peak at 2.75 ppm that was significantly lower in the calcified tumor samples as compared to the oily ones suggested the existence of a differential lipid composition (Tavel et al., 2016).

HR-MAS NMR could also be used in more industry types of applications, for example, to control the process of production. An example of such an approach was taken by García-García et al. (2018) where the authors investigated the manufacturing process of dry fermented *salchichón* type sausages and predict their early days of ripening. They collected the samples after 0, 2, 4, 7, 11, and 14 days of drying and then measured them using ^1H HR-MAS NMR. The results obtained showed that *salchichón* production is a three-stage process that includes formulation, fermentation and drying-ripening. Each of the processes could be distinguished by metabolomic profiles related to microbial activity. For example, the beginnings of the fermentation stage were characterized by the increase of signals related to ethanol (signal 9), acetic acid (signal 24) and 2,3-butanediol (signals 7, 8, 89). However, on the 14th day of ripening (the profile of final product), those signals were reduced and an increase in signals of Ile, Leu, Val, α -Ala, Glu, Gln, Met, Thr, Phe, Tyr, and Trp was observed (García-García et al., 2018). Those signals were signs of proteolysis and were related to the future flavor development of *salchichón* (Díaz et al., 1997).

Pure shift nuclear magnetic resonance

Additionally, the proton-proton homonuclear scalar coupling results in splitting of signals into complex multiples, creating in some cases more severely overlapped

spectra. When compared to ^{13}C (at the state natural isotopic abundance), ^{13}C NMR spectra tend to be better separated with singlets only present (assuming perfect ^1H decoupling during observation). Although the homonuclear ^{13}C – ^{13}C couplings can be seen in the spectra, they are extremely weak and usually hidden in the baseline noise (Zanger, 2015).

While the methods for suppression of heteronuclear J couplings have existed for many decades, development still continues (Kogler et al., 1983; Kövér & Batta, 1987; Rutar, 1984; Schilling et al., 2014). The removal of homonuclear J couplings (i.e., short range through bond ^1H – ^1H) has been more challenging. The goal of pure shift NMR (homonuclear broadband decoupling) is to convert all of the signals into singlets by manipulating the acquired time-domain observations, and exhibit the average evolution only under the effects of chemical shift evolution while excluding J-coupling (Zanger, 2015). To achieve this, a combination of novel NMR pulse sequences with innovative data acquisition and data processing techniques has been developed. Most pure shift NMR spectra are obtained using one of two different classes of experiments: J-refocusing experiments in which evolution under scalar coupling is refocused, or constant-time experiments with a constant amount of evolution (Adams, 2007). Both are based on a somewhat similar approach. First, a portion of free induction decay (FID) signal is collected for each increment of an evolution period. Second, the increments are combined to create an interferogram that can be later transformed as a conventional FID to obtain pure shift spectra (Adams, 2007; Emwas, Roy et al., 2019; Zanger & Sterk, 1997). More technical details about the pulse sequences and different types of pure shift experiments can be found in (Adams, 2007; Castañar, 2017; Zanger, 2015).

The pure shift NMR offers a great opportunity for applications in metabolomics. Although the main disadvantage of this method is the general loss of sensitivity, significant advances have been made to overcome this challenge. For example, utilization of pure shift (real-time BIRD) ^1H – ^{13}C HSQC-SI experiments on the metabolomics studies have been shown to increase the sensitivity of a spectra by about 40%–50% over a traditional 2D HSQC-SI experiment, and improve the chemical shift matching against NMR metabolomics databases (Timári et al., 2019). Another example shows that the use of SAPPHIRE-PSYCHE experiments combined with Statistical TOTal Correlation SpectroscopY (STOCSY) allows for easier identification of metabolites present in the mixture of *Physalis peruviana* fruits, and even enabled the identification of a new metabolite that was omitted by standard proton NMR (Lopez et al., 2019).

Recent advances

Improvements in nuclear magnetic resonance hardware and techniques and additional tools to aid in metabolomics studies

Efforts are underway to improve the inherently low sensitivity and limited resolution of NMR magnets. Many metabolomic studies use NMR to identify and

quantify metabolites present at very low concentrations (nM— μ M) ([Psychogios et al., 2011](#)) (also see <https://serummetabolome.ca/concentrations>). Correctly identifying and quantifying metabolites requires good quality NMR spectra, which may take several hours to a few days to obtain, even with high field NMR magnets (600–950 MHz). High field NMR magnets require routine liquid helium and nitrogen filling, which adds to the cost of running NMR experiments.

To acquire the large number of samples expected, for example, in population screening programs or in epidemiological studies, it is necessary to improve the technique to reach an even higher throughput, possibly in the range of hundreds per day or more.

A smart approach to deal with this issue was proposed by [Mulder et al. \(2019\)](#). The authors suggest the addition of a small amount of a commercial gadolinium contrast agent solution to urine samples. The resulting effect is a drastic shortening of the relaxation times (T_1) of the molecules in solution, thus permitting the reduction of the relaxation delay parameter. An increase in speed of up to fourfold with respect to standard protocols was reported, without appreciable effects on spectral quality.

Nuclear magnetic resonance magnets

Currently, 950 MHz and 1 GHz NMR magnets are quite commonly available for commercial applications. Intriguingly, 1.2 GHz magnets have been also introduced on the market, opening a new era of ultra-high field applications. Noteworthy, the first ^1H NMR spectrum at 1.2 GHz of a urine sample has been recently acquired at CERM ([Banci et al., 2019](#)). 1.2 GHz magnets increase the sensitivity and resolution of the experiments, thus potentially allowing the detection of a higher number of metabolites in biofluids ([Wishart, 2019](#)). Larger magnets, however, typically cost more to use and maintain, and also take up a significant amount of lab space. As a result, BNMR spectrometers with field strengths of ~ 100 MHz that take up much less space are gaining interest from researchers.

BNMR spectrometers have permanent magnets that are maintenance free and do not require dedicated lab space for their use ([Blümich, 2019](#)). Because of their permanent magnets, BNMR have lower magnetic fields (60, 90, 100 MHz), and do not offer as high resolution as the larger high magnetic field NMR magnets. However, their low cost and portability are attractive reasons for using them over larger magnets with higher fields, especially for academic laboratories with low budgets ([Wishart, 2019](#)). BNMR are also easy for novice NMR spectroscopists to use ([Lawson et al., 2020; Romero et al., 2020; van Beek, 2021](#)) and are even used for academic teaching purposes ([Riegel & Leskowitz, 2016](#)). Studies with BNMR show that BNMR (low-field magnets) can produce NMR spectra that are comparable to high magnetic fields for biofluid samples of patients with tuberculosis ([Izquierdo-Garcia et al., 2019](#)) and type II diabetes ([Percival et al., 2019](#)), demonstrating the value of BNMR for disease diagnosis.

The use of BNMR in metabolomics has emerged in recent years. Percival et al. (2019) developed a protocol for the analysis of urine, saliva, and blood serum, all of which are readily available biofluids on a BNMR (60 MHz). The authors successfully detected common diabetic markers, particularly α -glucose (≤ 2.8 mM) and acetone (25 μ M) in urine samples (Percival et al., 2019). In addition, Edgar et al. (2021) used a 60 MHz BNMR to perform a multicomponent metabolomics analysis for urine samples of patients with type II diabetes. The authors were able to successfully detect and quantify ~ 15 metabolite biomarkers for type II diabetes, which could be monitored to achieve rapid diagnosis and prognosis of patients (Edgar et al., 2021).

It is possible that BNMRs may become more popular for future metabolomics studies than high field magnets, but they may never achieve the full level of resolution and sensitivity of more costly, high-field magnetic fields. Nevertheless, whether or not time and cost are of concern, BNMRs and high-field NMR are of great value and have tremendous potential for metabolomic studies involving the diagnosis and treatment of diseases.

Nuclear magnetic resonance probes

The probe is essential for NMR experiments as it contains the necessary electronics and hardware to transmit and receive radiofrequency energy into the NMR sample of interest (Emwas, Roy et al., 2019). The probe can be designed in such a way as to accommodate various sizes of NMR tubes, and to have channels for the reading of multiple nuclei (Hong et al., 2018; Webb, 2006). Several changes have been made to probes over the last 20 years that increase the sensitivity of NMR, making it possible in some cases to enhance the signal of NMR-active nuclei with low natural abundance (such as ^{13}C) (Emwas, Roy et al., 2019; Keun, Beckonert et al., 2002).

Cryoprobes, for example, can increase the sensitivity of NMR by reducing the level of thermal noise generated by electronic circuits. As a result, the signal to noise ratio can increase by a factor of four, which is useful for NMR-nuclei with low sensitivity such as ^{13}C (Webb, 2006). For example, Keun, Beckonert et al. (2002) demonstrated the use of cryoprobes in obtaining information-rich and high-quality 1D ^{13}C NMR spectra of rat urine. Cryoprobes are considered a major advancement in NMR technology (Kovacs et al., 2005), and will likely play a larger role in future metabolomic studies involving NMR-active nuclei with low sensitivity.

Flow probes

Flow NMR is a term used to describe various techniques in which the sample is flowing through a tube into the NMR probe and out to a reservoir. Samples are not individually contained and measured in an NMR tube. The field of flow NMR is based on the combination with LC (i.e., LC-NMR), in which the first

experimental outlet of an HPLC system is connected directly to the NMR probe, which dates back to the end of the 1970s (Keifer, 2007; Watanabe & Niki, 1978). Currently, the field has evolved with various techniques such as Direct Injection NMR (DI-NMR) (Keifer et al., 2000), Flow Injection Analysis NMR (FIA-NMR) (Keifer, 2003) and Solid-Phase-Extraction NMR (SPE-NMR) (Griffiths & Horton, 1998). The hardware for LC-NMR consists of an HPLC fluid pump, some form of injector loop and valve control for the introduction of the sample into the system, a separating column, and one or more HPLC detectors (e.g., UV detector, conductivity detector), and this material is output to the NMR flow probe (Keifer, 2003). The sample is loaded into the injector loop and manually or automatically injected at the start of the LC-NMR run (Keifer, 2007; 2003). In DI-NMR, the hardware can consist of only a Gilson 215 Liquids Handler and an NMR flow probe (Keifer et al., 2000; Keifer, 2003). The samples (in solution state) are automatically injected into an NMR flow cell and then withdrawn the same way it came in, with rinsing of the flow cell in between the different samples (Keifer et al., 2000; Keifer, 2003, 2007). On the other hand, FIA-NMR is a simplified version of LC-NMR where the chromatographic column and the detectors are removed, leaving only a pump, an injector loop, a valve and a connector between the injector loop and the NMR flow probe (Keifer, 2003; Keifer, 2007).

Each of the methods have an NMR flow probe in common. The flow probe has to be adapted for both continuous sample injection and removal, for example, the connection to the HPLC column/system, and be designed to minimalize cross-contamination of samples and the formation of air bubbles in the flow cell (a glass or quartz sample tube with openings at the top and the bottom) (Haner & Keifer, 2007; Keifer, 2007). Two classes of probes exist: probes with the sample tube positioned vertically (like a standard NMR tube) and with saddle-shaped RF coils, and probes with the capillary sample tube positioned horizontally, with a solenoidal RF coil (Haner & Keifer, 2007). Most of those probes are optimized for ^1H detection and ^2H lock with the option of extra RF channels either single tuned (^{13}C) or double ($^{13}\text{C}/^{15}\text{N}$) (Haner & Keifer, 2007). Additionally, since the flow probe is directly connected to the HPLC system, it has to be adapted to the limitations of HPLC. For example, the flow cells usually have $\sim 60 \mu\text{L}$ volumes, which is a typical analytical LC peak (Haner & Keifer, 2007; Keifer, 2007). It is also worth mentioning that cryogenic flow probes became available some time ago, which increased sensitivity and resolution, and enabled faster acquisition of NMR data (Haner & Keifer, 2007; Keifer, 2007). Flow NMR has several complications such as cross-contamination, which can be alleviated by multiple rinses in-between. However, this requires time and increased solvent volumes, leading to the dilution of samples. Solvent suppression also becomes an issue, as the desire is often to scan samples as they flow by negating the ability to actively and slowly reduce solvent signals. Different solvent suppression techniques are therefore required (Altieri et al., 1996; Hoult, 1976; McKay, 2009; Price, 1999). Otherwise sample flow has to be halted while NMR acquisition occurs, which limits the advantages of flow systems.

Metabolomics databases and nuclear magnetic resonance software programs

A plethora of metabolomics databases and NMR software programs (Metabolomics Software and Servers) (Bingol, 2018; Ellinger et al., 2013; Giraudeau, 2020; Halabalaki et al., 2014; Izquierdo-Garcia et al., 2021; Johnson & Lange, 2015; Lipfert et al., 2019; Nagana Gowda & Raftery, 2017; Okazaki & Saito, 2012; Wishart et al., 2009, 2019), both free and licensed, is available to assist researchers with the identification and quantification of metabolites and natural products (Emwas, Roy et al., 2019; Johnson & Lange, 2015). Databases provide reference spectra that assist researchers in identifying the metabolites in their samples (Johnson & Lange, 2015), and/or separating known compounds from unknown compounds (Lai et al., 2018), an especially useful tool for untargeted metabolomics studies (Bingol, 2018). Software programs can be tuned to accelerate (Ellinger et al., 2013; Puchades-Carrasco et al., 2016; Spicer et al., 2017) and automate (Beirnaert et al., 2018; Bingol, 2018; de Brouwer & Stegeman, 2011; Howarth et al., 2020; Johnson & Lange, 2015; Kern et al., 2019) the identification and quantification of metabolites.

Since MS is also widely used in metabolomics studies, many databases include NMR and MS reference spectra, while others only contain NMR or MS spectra (Johnson & Lange, 2015). Some databases contain no spectral data at all (Johnson & Lange, 2015). Since the focus of this chapter is the role of NMR in metabolomics, we will limit our discussion to databases that contain NMR spectral data and to software tools that are capable of analyzing NMR spectra of metabolites. We introduce some of these tools below, and briefly describe some improvements that could be made for future applications.

Databases for nuclear magnetic resonance-based metabolomics

The HMDB, found at <https://hmdb.ca/>, is probably the most widely used database for metabolomics studies, and is considered the standard metabolic database for human metabolic studies (Emwas, Roy et al., 2019; Wishart et al., 2009). Created in 2007 by David Wishart (Wishart et al., 2007, 2009, 2012), it contains over 100,000 metabolites, with many data fields hyperlinked to additional open-source databases. Since its creation, the number of experimental and predicted NMR spectra in HMDB has increased fourfold. HMDB also contains data about many drugs and drug metabolites, toxins and environmental pollutants, pathway diagrams for human metabolic and disease pathways, food components and food additives in its subdatabases DrugBank, T3DB, SMPDB and FooDB, respectively, all of which extend its utility beyond metabolomics studies into clinical and environmental studies. The number of metabolites for each of these subdatabases extends into the thousands. It is likely that HMDB will become the de facto database for NMR-based metabolomics studies (Emwas, Roy et al., 2019) if it is not already, and Wishart and his colleagues will continue to increase the number of

metabolites in the database, and their relevant spectral information. Though not nearly as extensive as HMDB, the Biological Magnetic Resonance Bank (BMRB, found at <https://bmrbl.io/>) (Ulrich et al., 2007) contains information regarding where the cataloged compounds (~1000 total) were collected, the solution conditions, how the data (i.e., NMR spectra) were obtained, and the NMR pulse sequences used to obtain the NMR spectra (Markley et al., 2007). It has several types of NMR spectra, including ^1H , ^{13}C , DEPT90, DEPT 135, ^1H J-resolved, ^1H - ^{13}C HSQC, ^1H - ^{13}C HMBC, ^1H - ^1H TOCSY, and ^1H - ^1H COSY spectra (Johnson & Lange, 2015). The BMRB database is searchable by compound name, structure, 1D and 2D HSQC peak lists, and solvent and field strength. BMRB has several output files that can be easily read by third party NMR software. It also contains a bulk download option, which is extremely convenient for researchers who need to do their work offline.

The Madison-Qingdao Metabolomics Consortium Database (MMCD) (found at <http://mmcd.nmrfa.m.wisc.edu/>) contains over 20,000 metabolites, with standard ^1H , ^{13}C , ^1H , ^1H -TOCSY, ^1H , ^{13}C -HSQC NMR spectra for 794 of their cataloged compounds. As of 2015, MMCD contained a total of 5,256 NMR spectra. Far from simply holding useful NMR (and MS) data for researchers, MMCD is an efficient and flexible query system that accepts input in the forms of text, molecular structure search, NMR parameters, MS parameters, and/or miscellaneous (i.e., reference, data source, organism, etc.). These search parameters may include information about isotopomer molecular weight, nomenclature, physical properties, empirical and calculated chemical shifts, and/or NMR sample conditions. It is possible to submit queries as a single file or as a batch (Cui et al., 2008). All of these properties make MMCD a flexible and useful tool for metabolomics research, though improvements such as more data entries and more synchronization with other databases would help to increase its already large versatility.

The last NMR-metabolomics database we will discuss in detail is the Birmingham Metabolite Library BML-NMR (found at <http://www.bml-nmr.org>). BML-NMR contains comparatively few metabolites (208 total) compared to the databases discussed above. However, the metabolites it contains are widely detected in metabolomics studies, and its standard solutions were prepared at pH values (6.6, 7.0, 7.4), which are close to physiological pH (7.35–7.45) (Duarte, Jaremko et al., 2020) enabling the effect of pH on the chemical shifts of the NMR signals to be quantified. 1D and 2D J-resolved NMR spectra, quantified peak lists, and metadata that comply with the Metabolomics Standards Initiative (Rubtsov et al., 2007) are included in the database (Ludwig et al., 2012). Though not as comprehensive as other metabolomic databases such as HMDB, the BML-NMR database provides detailed and high-quality data of metabolite standards that serve as a useful starting point for the identification of metabolites, and for filtering the “known” metabolites from the “unknown” metabolites.

Despite a plethora of metabolomic databases to aid in the detection, identification, and quantification of metabolites, much more work must be done to “unify” them to enhance their capacity in NMR-based metabolomics research. A crucial

element would be to have a downloadable, uniform/common file output system (e.g., every file is.xml or.txt) across the multiple databases. Virtually all metabolomic databases have their own unique file output, with some being vendor-specific, and some having no form of downloadable data whatsoever (Johnson & Lange, 2015). The creation of a common output file format (regardless of the databases from which NMR data is taken) would make it convenient for scientists to analyze standard and known metabolites. A second option would be to create a software program that converts the various output files into one, searchable file format (such as a text file).

The more pressing need, however, is to increase the amount of available data on metabolites. Currently, only a tiny fraction of the known metabolome has been analyzed, while the rest remaining uncharacterized and named as the “dark matter” of the metabolome (da Silva et al., 2015; Wishart et al., 2018). Most of the databases are limited to NMR spectra involving ^1H and ^{13}C nuclei, with comparatively few spectra involving other important nuclei such as ^{19}F , ^{15}N (Emwas, Roy et al., 2019) and ^{31}P (Bhinderwala et al., 2020). Though increasing the amount of data in the NMR metabolomics databases would take a considerable amount of time, the effort would pay off since many important diseases such as cancer are metabolic in nature (Čuperlović-Culf, 2012; Palmnas & Vogel, 2013; Wishart et al., 2016), and the increase in data amount and availability would most likely help in the identification of novel biomarkers (Gebregiworgis & Powers, 2012; Serkova & Niemann, 2006; Smolinska et al., 2012).

For those interested in a more in-depth review of metabolomics databases containing NMR spectra, we recommend the following (Ellinger et al., 2013; Halabalaki et al., 2014; Izquierdo-Garcia et al., 2021; Johnson & Lange, 2015; Misra & van der Hooft, 2016; Puchades-Carrasco et al., 2016; Wishart, 2019).

Use of software to analyze metabolite nuclear magnetic resonance data

In order to correctly identify and quantify metabolites, NMR data are subject to additional analyses, which include pre and post data processing (Ellinger et al., 2013; Emwas et al., 2018; Euceda et al., 2015; Karaman, 2017), normalization (Kohl et al., 2012; Roberts et al., 2014; Zacharias et al., 2018) and statistical methods such as multivariate analysis (Bartel et al., 2013; De Livera et al., 2013; Misra & van der Hooft, 2016; Puchades-Carrasco et al., 2016; Ren et al., 2015; Saccenti et al., 2014). Several NMR software programs are available to aid in the analysis and interpretation of NMR metabolite data. There are both open source and commercially available NMR software for metabolomics studies (Ellinger et al., 2013; Lewis et al., 2009; Puchades-Carrasco et al., 2016), though the availability of and interest in open-source NMR software have increased in recent years (O’Sullivan et al., 2007). Below, we discuss a few of the available NMR software tools, and their uses in metabolomics research.

The BATMAN (Bayesian AuTomed Metabolite Analyzer for NMR data) program is a freely available, R-based package used to deconvolute, analyze, and quantify the peaks of metabolites in ^1H NMR spectra (Hao et al., 2012). BATMAN implements a Bayesian model, followed by a Markov chain Monte Carlo algorithm to automatically quantify metabolites, and then assign the metabolites based on user-defined parameters (e.g., list of metabolites, chemical shift region) (Hao et al., 2012). BATMAN can also account for shifts in the position of peaks commonly observed in NMR spectra (Izquierdo-Garcia et al., 2021). In short, BATMAN automatically assigns the peaks and gives concentration estimates of metabolites in the sample (Hao et al., 2012), thereby saving time that would otherwise be spent assigning and picking peaks manually, which may be especially applicable in discriminating diseased patients from HCs. For example, Padayachee et al. (2019) compared BATMAN with the common binning approach to discriminate between serum samples of lung cancer patients and those of HCs. Though their use of BATMAN was not completely automatic (i.e., they had to spend significant amounts of time using the software to analyze the samples), they found that BATMAN had a fair predicting power for metabolite concentrations at high magnet field strengths (900 MHz) (Padayachee et al., 2019). This example illustrates the clinical relevance of BATMAN for clinical studies and could be used in metabolic studies to discover new biomarkers for diseases.

The BAYESIL program (Web-based) is another software that implements the Bayesian model to automatically identify and quantify metabolites from 1D ^1H NMR spectra of ultra-filtered plasma, serum, or cerebrospinal fluid (CSF) (Ravanbakhsh et al., 2015). For BAYESIL to work properly and optimally, a strict protocol (see http://bayesil.ca/spectra_collection) must be followed. BAYESIL automatically performs processing steps, such as Fourier transformation, phasing, solvent filtering, chemical shift referencing, baseline correction and reference line shape convolution. It is the first fully-automatic publicly-accessible system to quantify NMR spectra, and has an identification accuracy similar to and even higher than that of highly trained spectroscopists (Ravanbakhsh et al., 2015). For the less experienced users of the Web-based BAYESIL interface, a protocol is available that details the steps on how to submit NMR data for spectral analysis (Lipfert et al., 2019). BAYESIL is a great stand-alone tool, but it can be combined with other NMR programs such as BATMAN (Mediani et al., 2017) to increase data output and to compare/confirm the identification and quantification of metabolites.

The most common and well-known software tool for NMR metabolomics is ChenoMX (<http://www.chenomx.com/software/>), a commercial, patented software package (Izquierdo-Garcia et al., 2021). ChenoMX contains reference libraries for many types of organic compounds (e.g., amides, amines, and metabolites involved in carbohydrate metabolism), which are available for different pH values (4–9) and NMR field strengths (400–800 MHz). A handy feature of ChenoMX is its ability to automatically adjust the Reference Library to reflect the sample and acquisition conditions (e.g., pH and NMR field strength) by “fitting” the Reference Library to the experimental NMR spectra, which then enables

identification and quantification of the metabolites. ChenoMX can be used to identify metabolites in biofluids (Rosewell & Vitols, 2006) such as urine (Vitols & Fu, 2006; Weljie et al., 2006), blood serum (Vitols & Weljie, 2006), and plasma (Vitols & Weljie, 2006). ChenoMX has also been used to measure drug metabolism in cell culture, and to discover new biomarkers in the food industry (see <http://www.chenomx.com/software/>). Clearly, ChenoMX is a tool of interest in metabolite quantification and identification.

Perhaps most familiar to NMR spectroscopists are the programs TopSpin (<https://www.bruker.com/products/Mr/nmr/software/topspin>) and MestreNova (<https://mestrelab.com>) (Willcott, 2009), both of which are used to analyze and process NMR data, as well to prepare NMR spectra for publication. TopSpin has a user-friendly interface, which is useful for both novice and experienced users. It also provides tools to assist in NMR data acquisition and to make the structural elucidation of small molecules (e.g., metabolites) efficient. MestreNova, on the other hand, can open raw FID files in multiple formats (Varian, Bruker, JEOL, etc.), and show the spectra immediately with no operator intervention (e.g., Fourier transform). It can also open MS files, which can further aid in metabolite identification and quantification, as MS is much more sensitive than NMR. MestreNova has a powerful, user-friendly interface to visualize, process, analyze, and report 1D and 2D NMR spectra. Both TopSpin and MestreNova are commercial software, and licenses are available to businesses and universities alike.

The main difficulties in identifying and quantifying peaks in NMR derive from the shifts of the peaks that, especially for urine samples are not fixed. A novel approach to deal with these variations using an automated and accurate prediction of chemical shifts in urine was recently introduced (Takis et al., 2017). This algorithm is driven by the chemical shifts of five “navigator” signals from which, using a regression-based approach, the position of many other signals can be accurately predicted to the 3rd–4th decimal of ppm.

As with NMR databases, it would be helpful to have a uniform file format that any program (open source or commercial) can open and process. Efforts in this direction have been performed with international initiatives (Salek et al., 2015). Though not strictly necessary, it would aid in analyzing data that multiple collaborators and researchers are working on, thus facilitating the identification and quantification of metabolites in biofluids. More updates to metabolomics software tools are underway (Misra & Mohapatra, 2019; Misra & van der Hooft, 2016; Spicer et al., 2017) and are likely to continue in the future (Wishart, 2019).

Advantages of nuclear magnetic resonance spectroscopy

As discussed above, NMR is a versatile analytical instrument with superior advantages including its high reproducibility, nondestructive nature, and nonbiased approach. Furthermore, NMR is an inherently quantitative method for both

the identification and quantification of different molecules in sample mixtures. NMR is also a rapid and high-throughput technique allowing the acquisition of multiple samples per hour. Standard biological fluid sample preparation is relatively fast, and with minimal sample handling and workup. Finally, despite the high costs of NMR spectrometers and professional staffing, the cost per sample (with economies of scale) can be lower (e.g., 10–20 euros per sample) than other methods.

Reproducibility

A fundamental advantage of NMR spectroscopy over MS is high reproducibility. In 1988 a repeatability factor of over 0.8% and a confidence interval of 0.25% were determined during a study across 15 European laboratories, and the reproducibility when using different NMR systems was on the order of 2%–3%. This was improved to 1% with more homogeneous spectrometer systems (Guillou et al., 1988). In contrast, Bauer et al. (1998) observed significant laboratory effects in the integration of complex signal during the course of an interlaboratory study (5 centers), investigating the quantitative use of NMR. In 1999, a German interlaboratory study on quantitative NMR (Malz & Jancke, 2005) found results to differ up to 100% among participating sites; differences were attributed mostly to the individual and independent setup of the measurements, the data processing, and the evaluation procedure of each single laboratory (Gallo et al., 2015; Malz & Jancke, 2005). In a follow up study involving 33 sites (instruments operating from 200 to 600 MHz) (Malz & Jancke, 2005), a common protocol for the experimental setup and data processing was adopted and the determination of mole ratios of different compounds showed a measurement uncertainty of 1.5% for a confidence level of 95%. In another large-scale ring-test (Gallo et al., 2015), the different participants were able to produce NMR spectra of a given mixture that were statistically equivalent in terms of relative intensities of the signals with respect to the internal standard.

In a metabolomics context, Dumas et al. (2006) observed a >98% multivariate analytical reproducibility, with most of the inaccuracies originating from sample handling. Keun, Ebbels et al. (2002) assessed the analytical reproducibility of metabolomics protocols at two sites with spectrometers operating at 500 and 600 MHz and found that the relative concentrations of citrate, hippurate, and taurine were in >95% correlation (r^2) between the two instruments, with an analytical error comparable to normal physiological variation in concentration (4%–8%).

Ward et al. (2010) analyzed plant-derived samples by ^1H -NMR spectroscopy across five different sites using instruments with different probes and magnetic field strengths (400, 500 and 600 MHz). They found exceptional comparability of the data sets obtained from different laboratories and reported that field strength differences can be adjusted for in the data preprocessing. They concluded that

¹H-NMR fingerprinting is the ideal technique for large-scale plant metabolomics data collection requiring the participation of multiple laboratories.

Further, a large scale assessment of technical reproducibility of NMR metabolomics (Dumas et al., 2006) proved that NMR spectroscopy of biofluids combined with multivariate pattern recognition is a robust and precise approach for metabolomics studies, outperforming other “-omics” technologies in terms of reproducibility. The coefficient of variation for signals in repeated NMR spectra is reported in the range of 0%–10%.

However, all studies stressed the necessity of using standardized protocols for sample collection, handling, and measurement for precise quantitative and fingerprinting applications (Emwas, Luchinat et al., 2015), which render NMR spectroscopy highly reproducible. It has been our experience that volatility in physical sample handling (i.e., acquisition protocols, storage, shipping, and preparation) generates a far larger error in results than the NMR instrumentation itself (Sokolenko et al., 2013).

Challenges and limitations

Although NMR spectroscopy enjoys several advantages as summarized above, a few limitations still represent daunting challenges that need to be appreciated and overcome in order to improve sensitivity, spectral resolution, and both accurate and precise metabolite identification. The main limitation of NMR spectroscopy is intrinsic low sensitivity concerning metabolites below $\sim 10 \mu\text{M}$ or even into submicromolar levels. Without substantially extended experimental times, these metabolites remain undetectable by NMR spectroscopy. For a specific example, many essential metabolites such as hormones are usually below this detection limit. Additionally, secondary metabolites could be the focus of many targeted metabolomics studies, however the detection limit makes these impractical. Despite several developments including ultra-high field NMR magnet, cryoprobes, and using hyperpolarizations methods (Emwas, Roy et al., 2019; Wishart, 2019), many metabolites are still unobtainable.

Spectral overlap is another challenge in NMR spectroscopy, mainly in 1D spectra (Emwas, Alghrably et al., 2019; Mohammed et al., 2020; Naser et al., 2019). This presents a major difficulty in metabolic identification and subsequent quantification. For example, the typical spectra width of the entire possible 1D ¹H NMR spectra (i.e., in D₂O/H₂O under typical conditions) is less than 11 ppm. Indeed, the vast majority of ¹H NMR signals are mainly obtained between 1 and 4.4 ppm, and the aromatic region (~ 6.5 –8 ppm). As most NMR studies employ ¹H NMR, overlap is still one of the main problems in data analyses. While the spectral redundancy problem can be resolved using higher and higher magnetic fields, there is a limit (~ 1200 MHz or 28.2 Tesla costing ~ 12.5 M euros) to what researchers can access. Other techniques such as multidimensional (e.g., 2D)

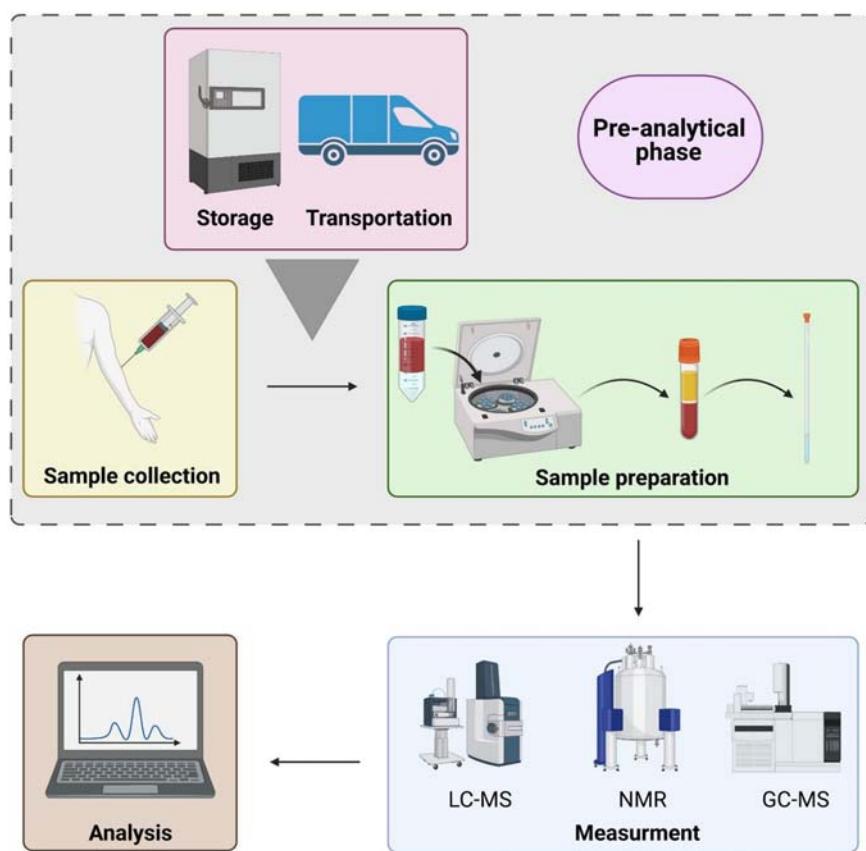
and/or multinuclear (e.g., ^1H , ^{13}C) NMR spectroscopy, pure shift NMR, and projection of JRES spectroscopy (Emwas, Roy et al., 2019) can all assist, at the cost of additional time and experimental complexity.

Another issue is the lack of fully developed NMR spectral libraries designed specifically for metabolomics (e.g., calibrated for common field strengths/instruments, gathered in standard format, and freely available). Although projects such as the Human Metabolomics DataBase (e.g., HMDB 4.0) exist and gather extensive information about metabolites (including NMR spectra), the number of experimental data sets provided for NMR (3840) is far behind those for MS/MS (22198) or GC-MS (7418) (Wishart et al., 2018). This aspect was addressed by the NIH with a call to develop a new, sustainable, and open natural products NMR spectral database containing NMR spectra of natural products for NMR metabolomics researchers (Wishart, 2019).

Lastly, the practical aspects related to the infrastructure and operational procedure seem to impede the popularization of NMR. NMR requires highly skilled and educated operators to control the instrument and evaluate and interpret the results (the problem increases even more when the lack of an appropriate NMR metabolic database is included). When combined with the fact that NMR magnets have extensive laboratory footprints (especially the new 1.2 GHz models), require special vibration, magnetic field and radio interference free spaces, and a constant supply of liquid helium and nitrogen to maintain the superconducting magnet, the utilization of NMR is therefore obviously hindered by the financial costs and huge equipment footprint (Emwas, Roy et al., 2019; Wishart, 2019). The use of “benchtop,” relatively inexpensive, nonsuperconducting magnets is attractive, however the present limit to 80 MHz and below makes the overlap problems described above even more prominent (Edgar et al., 2021; Percival et al., 2019; Wishart, 2019).

Sample preparation

The initial stage of any metabolomic investigation (Takis et al., 2019; Tenori et al., 2007; Vignoli et al., 2019) includes the fundamental steps of the collection of the biological samples, their handling, transportation, and storage. All these steps collectively constitute the so-called “preanalytical” phase (see Fig. 5.5). In a complex biological mixture, such as a biofluid, the concentrations of the different analytes could change according to the different conditions or procedures employed for the preanalytical phase. These changes are mostly induced by enzymatic reactions, chemical reactions, and exposure to air and light. Thus it is of utmost importance to establish optimal standard operating procedures (SOPs) for the proper collection and handling of the biological samples. This is important also for biobank activities. In principle, dedicated SOPs need to be developed for each different sample type. This can only be achieved by performing a systematic evaluation of all the possible effects that potentially could influence the

**FIGURE 5.5**

NMR samples acquisition overview. Overview of NMR sample acquisition, measurement, and analysis.

Created with BioRender.com.

composition of the sample, and analyzing the sample under different preanalytical conditions, in a process that can be dub “evidence-based” SOPs development. Thorough evaluations (Kirwan et al., 2018) exist for some of the most useful biofluids, such as urine, serum, plasma, saliva and CSF.

As an example, it was observed that the main source of preanalytical changes in urine samples is due to the presence of human or bacterial cells that may break upon water crystal formation after freezing (Bernini et al., 2011). Indeed, cell breaking releases enzymes that ignite uncontrolled reactions in the sample. For this reason, if cells are eliminated by filtration (and/or mild centrifugation) before NMR sample preparation or long-term storage, these undesirable effects are much reduced. This is preferred to the addition of stabilizers (such as enzyme

inhibitors) that introduce unwanted signals in the NMR spectra and may also induce changes in the sample. After monitoring the urine metabolic profile over time, it was concluded that urine needed to be processed within 2 hours of collection and maintained at 4°C between collection and processing. For the long-term storage of the samples, liquid nitrogen vapor or -80°C are the recommended conditions.

A similar investigation was performed for serum and plasma samples, and the stability of these biofluids was controlled with respect to time and storage temperature before separation from the whole blood (Bernini et al., 2011). They observed time-dependent and temperature-dependent changes both for serum and plasma samples: in both cases storage at 25°C caused deeper changes in the NMR profile, in particular for glucose, lactate, and pyruvate, with a decrement of glucose concentration and a connected increment in lactate.

Thus the optimal SOPs developed for serum and plasma collection for NMR-based metabolomic studies prescribe:

1. the serum/plasma separation within 2 hours after blood collection
2. the maintenance of the blood samples at 4°C during this delay
3. the immediate freezing of the samples after separation
4. the long-term storage at -80°C

The suitability of these recommendations was also monitored for long-term storage (Ghini et al., 2019).

These results obtained for urine, serum, and plasma derive from the activities performed under the EC founded project SPIDIA (FP7, #222916), and have been the basis for the production of technical specifications for the preanalytical processes for NMR metabolomics in urine, venous blood serum, and plasma published by CEN (CEN/TS 16945:2016) (Bernini et al., 2011; Molecular in Vitro Diagnostic Examinations—Specifications for Pre-Examination Processes for Metabolomics in Urine, Venous Blood Serum, & Plasma., 2016). This document is now in the process of being translated into ISO/IS specifications (ISO/DIS 23118) (Molecular in Vitro Diagnostic Examinations—Specifications for Pre-Examination Processes in Metabolomics in Urine, Venous Blood Serum, & Plasma, 2020) as part of the activities of the EC founded project SPIDIA4P (H2020, #733112), and these recommendations have been adopted by some biobanks (Carotenuto et al., 2015; Marcon & Nincheri, 2014).

Saliva is an interesting biofluid that, although less exploited than blood and urine, has demonstrated potentially useful applications in recent years (Aimetti et al., 2012; Romano et al., 2018, 2019). For this reason, a recent paper (Duarte, Castro et al., 2020) reported an untargeted NMR metabolomics study to assess the effects of different storage temperatures and times on saliva composition. The authors showed that, after collection, saliva may be kept at 22°C or 4°C for up to 6 hours, after which some metabolite levels start to change, especially at 22°C. For longer periods, saliva is stable at -20°C for at least 4 weeks (Duarte, Castro et al., 2020).

The composition of CSF, a secretion product of the central nervous system, indirectly reflects the biochemical processes occurring in the brain (Albrecht et al., 2020). Thus CSF metabolomics has received much attention in the research of neurological disorders including Alzheimer's disease (Vignoli et al., 2020), Parkinson's disease, multiple sclerosis and brain injury, amongst others. Optimal procedures for the collection and biobanking of CSF for clinical purposes have been proposed by Teunissen et al. (2009). These procedures are not specifically developed or validated for NMR metabolomics. However, according to a paper of the Metabolomic Society Initiative (Kirwan et al., 2018), these recommendations could be used for metabolomics with minor changes, including a centrifugation step (2000 g for 10 minutes at 4°C) before storage to remove and prevent cell lysis on thawing. In any case, further systematic validations are still needed.

Unfortunately, for many other potentially interesting biofluids (e.g., sweat, tears, sperm, vaginal fluid, synovial fluid, breast milk) specific SOPs for metabolomics investigations are not yet established. In these cases, the recommendation (Kirwan et al., 2018) is that, in the absence of specific studies and until proper validations, the procedures should be based on existing clinical protocols, reasonably adapted following similar metabolomics research.

Summary and future perspectives

The Human Genome Project (1993–2003) was termed as one of “the great feats of exploration in history.” Even if officially “complete,” it is still active in thousands of laboratories around the world delivering new data with an estimated 60 million human genomes available in the coming years (Birney et al., 2017; Langmead & Nellore, 2018). Genomics, which depicts in detail genotypes, gave birth to transcriptomics, proteomics, metabolomics, and single-cell-omics techniques that define phenotypes. Presently high-resolution imaging, electronic health and medical records, “big-data” analytics, and numerous internet-connected health devices have the potential to combine all into a far clearer picture of a “human being.”

Innovative cloud systems maintain and allow access to “-omics,” collaborating with other relevant clinical data sources in a secure manner. A standardized meta/data format could vastly simplify data sharing and increase the findability (Aarestrup et al., 2020) of key results. NMR, with its high reproducibility and standard data output, fits perfectly with the cloud data-storage concept. Matching genotypes with phenotypes and environmental factors is a complicated process, in which metabolomics has a fundamental role. In the future, the analysis of metabolites related to the aging process, virus and bacteria pathologies, ingested drugs and food, and climate changing factors (e.g., pCO₂, temperature) will define better pathologies and therapies.

Applying these ideas with information and communication technology-based medicine will permit the monitoring of the changes in the metabolome and the deviations from an individual baseline. Medical doctors and supporting staff can

follow the development of diseases and subsequent response(s) to developed therapies. Of course, genomic and other –omic information will need to be recorded and integrated in the “virtual patient” description (Bertini et al., 2012). Because of its simplicity and minimal sample processing, NMR is ideally suited for systems medicine applications (Dos Santos et al., 2020), profiling blood and tissue samples collected in the operating theater can provide almost real time information to the surgeons and to clinicians (Nicholson et al., 2012).

Although NMR spectroscopy has the limitation of intrinsically low sensitivity, the continuous development of NMR methods, probes, and ultra-high magnetic fields along with new hyperpolarization methods may alleviate many of NMR’s limitations. Moreover, recent developments of computer power, related molecular identification software, and metabolomics databases will facilitate more applications of NMR-based metabolomics in a wider range of research areas.

References

- Aarestrup, F. M., Albeyatti, A., Armitage, W., Auffray, C., Augello, L., Balling, R., Benhabiles, N., Bertolini, G., Bjaalie, J., & Black, M. (2020). Towards a European health research and innovation cloud (HRIC). *Genome Medicine*, 12(1), 1–14.
- Abd Ghafar, S. Z., Mediani, A., Maulidiani, M., Rudyantoro, R., Ghazali, H. M., Ramli, N. S., & Abas, F. (2020). Complementary NMR-and MS-based metabolomics approaches reveal the correlations of phytochemicals and biological activities in *Phyllanthus acidus* leaf extracts. *Food Research International*, 136, 109312.
- Abdul-Hamid, N. A., et al. (2019). 1H-NMR-based metabolomics to investigate the effects of Phoenix dactylifera seed extracts in LPS-IFN- γ -induced RAW 264.7 cells. *Food Research International* (Ottawa, Ont.), 125, 108565. Available from <https://doi.org/10.1016/j.foodres.2019.108565>.
- Abdul Jameel, A. G., Alquaity, A. B. S., Campuzano, F., Emwas, A.-H., Saxena, S., Sarathy, S. M., & Roberts, W. L. (2021). Surrogate formulation and molecular characterization of sulfur species in vacuum residues using APPI and ESI FT-ICR mass spectrometry. *Fuel*, 293, 120471. Available from <https://doi.org/10.1016/j.fuel.2021.120471>.
- Acciardo, S., et al. (2020). Metabolic imaging using hyperpolarized (13)C-pyruvate to assess sensitivity to the B-Raf inhibitor vemurafenib in melanoma cells and xenografts. *Journal of cellular and molecular medicine*, 24(2), 1934–1944. Available from <https://doi.org/10.1111/jcmm.14890>.
- Adams, R. W. (2007). Pure shift NMR spectroscopy. *Emagres*, 295–310.
- Agrawal, P. (2020). NMR spectroscopy in drug discovery and development. *Materials and Methods*. Available from <https://doi.org/10.13070/mm.en.4.599>.
- Aimetti, M., Cacciatore, S., Graziano, A., & Tenori, L. (2012). Metabonomic analysis of saliva reveals generalized chronic periodontitis signature. *Metabolomics: Official Journal of the Metabolomic Society*, 8(3), 465–474.
- Alahmari, F., Davaasuren, B., Emwas, A.-H., & Rothenberger, A. (2018). Thioaluminogermanate M (AlS₂)(GeS₂)₄ (M = Na, Ag, Cu): Synthesis, crystal structures, characterization, ion-exchange and solid-state ²⁷Al and ²³Na NMR spectroscopy. *Inorganic Chemistry*, 57(7), 3713–3719.

- Albrecht, B., Voronina, E., Schipke, C., Peters, O., Parr, M. K., Díaz-Hernández, M. D., & Schlörer, N. E. (2020). Pursuing experimental reproducibility: An efficient protocol for the preparation of cerebrospinal fluid samples for NMR-based metabolomics and analysis of sample degradation. *Metabolites*, 10(6), 251.
- Aljuhani, M. A., Zhang, Z., Barman, S., El Eter, M., Failvane, L., Ould-Chikh, S., Guan, E., Abou-Hamad, E., Emwas, A.-H., & Pelletier, J. D. (2019). Mechanistic study of hydroamination of alkyne through tantalum-based silica-supported surface species. *ACS Catalysis*, 9(9), 8719–8725.
- Alonso, J., Arús, C., Westler, W. M., & Markley, J. L. (1989). Two-dimensional correlated spectroscopy (COSY) of intact frog muscle: Spectral pattern characterization and lactate quantitation. *Magnetic Resonance in Medicine*, 11(3), 316–330.
- Alsiary, R. A., Alghrably, M., Saoudi, A., Al-Ghamdi, S., Jaremko, L., Jaremko, M., & Emwas, A.-H. (2020). Using NMR spectroscopy to investigate the role played by copper in prion diseases. *Neurological Sciences*, 1–18.
- Altes, T. A., & Salerno, M. (2004). Hyperpolarized gas MR imaging of the lung. *Journal of Thoracic Imaging*, 19(4), 250–258.
- Altieri, A., Miller, K., & Byrd, R. (1996). A comparison of water suppression techniques using pulsed field gradients for high-resolution NMR of biomolecules. *Magn Res Rev*, 17, 27–82.
- Apperley, D. C., Harris, R. K., & Hodgkinson, P. (2012). *Solid-state NMR: Basic principles and practice*. Momentum Press.
- Ashbrook, S. E., Griffin, J. M., & Johnston, K. E. (2018). Recent advances in solid-state nuclear magnetic resonance spectroscopy. *Annual Review of Analytical Chemistry*, 11, 485–508.
- Aue, W., Karhan, J., & Ernst, R. (1976). Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *The Journal of Chemical Physics*, 64(10), 4226–4227.
- Babgi, B. A., Alsayari, J., Alenezi, H. M., Abdellatif, M. H., Eltayeb, N. E., Emwas, A.-H. M., Jaremko, M., & Hussien, M. A. (2021). Alteration of anticancer and protein-binding properties of gold(I) alkynyl by phenolic Schiff bases moieties. *Pharmaceutics*, 13(4), 461. Available from <https://doi.org/10.3390/pharmaceutics13040461>.
- Banci, L., Barbieri, L., Calderone, V., Cantini, F., Cerofolini, L., Ciofi-Baffoni, S., Felli, I. C., Fraga, M., Lelli, M., & Luchinat, C. (2019). Biomolecular NMR at 1.2 GHz. ArXiv Preprint ArXiv:1910.07462.
- Barskiy, D. A., Coffey, A. M., Nikolaou, P., Mikhaylov, D. M., Goodson, B. M., Branca, R. T., Lu, G. J., Shapiro, M. G., Telkki, V.-V., & Zhivonitko, V. V. (2017). NMR hyperpolarization techniques of gases. *Chemistry (Weinheim an Der Bergstrasse, Germany)*, 23(4), 725.
- Barskiy, D. A., Shchepin, R. V., Coffey, A. M., Theis, T., Warren, W. S., Goodson, B. M., & Chekmenev, E. Y. (2016). Over 20% ¹⁵N hyperpolarization in under one minute for metronidazole, an antibiotic and hypoxia probe. *Journal of the American Chemical Society*, 138(26), 8080–8083.
- Bartel, J., Krumsiek, J., & Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal*, 4 (5), e201301009.
- Bauer, M., Bertario, A., Boccardi, G., Fontaine, X., Rao, R., & Verrier, D. (1998). Reproducibility of ¹H-NMR integrals: A collaborative study. *Journal of Pharmaceutical and Biomedical Analysis*, 17(3), 419–425.

- Bax, A., & Summers, M. F. (1986). Proton and carbon-13 assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. *Journal of the American Chemical Society*, 108(8), 2093–2094.
- Beckonert, O., Coen, M., Keun, H. C., Wang, Y., Ebbels, T. M., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2010). High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nature Protocols*, 5(6), 1019–1032.
- Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11), 2692–2703. Available from <https://doi.org/10.1038/nprot.2007.376>.
- Beirnaert, C., Meysman, P., Vu, T. N., Hermans, N., Apers, S., Pieters, L., Covaci, A., & Laukens, K. (2018). speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification. *PLoS Computational Biology*, 14(3), e1006018.
- Berg, J. M., Tymoczko, J. L., & Stryer, L. Biochemistry. 5th edition. New York: W. H. Freeman; 2002. Section 16.1, Glycolysis Is an Energy-Conversion Pathway in Many Organisms. Available from <https://www.ncbi.nlm.nih.gov/books/NBK22593/>.
- Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schažfer, H., Schužtz, B., Spraul, M., & Tenori, L. (2009). Individual human phenotypes in metabolic space and time. *Journal of Proteome Research*, 8(9), 4264–4271.
- Bernini, P., Bertini, I., Luchinat, C., Nincheri, P., Staderini, S., & Turano, P. (2011). Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. *Journal of Biomolecular NMR*, 49(3), 231–243.
- Bertelli, D., Lolli, M., Papotti, G., Bortolotti, L., Serra, G., & Plessi, M. (2010). Detection of honey adulteration by sugar syrups using one-dimensional and two-dimensional high-resolution nuclear magnetic resonance. *Journal of Agricultural and Food Chemistry*, 58(15), 8495–8501.
- Bertini, I., Luchinat, C., & Tenori, L. (2012). Metabolomics for the future of personalized medicine through information and communication technologies. *Personalized Medicine*, 9(2), 133–136.
- Bhinderwala, F., Evans, P., Jones, K., Laws, B. R., Smith, T. G., Morton, M., & Powers, R. (2020). Phosphorus NMR and its application to metabolomics. *Analytical Chemistry*, 92(14), 9536–9545.
- Bhinderwala, F., Lonergan, S., Woods, J., Zhou, C., Fey, P. D., & Powers, R. (2018). Expanding the coverage of the Metabolome with nitrogen-based NMR. *Analytical Chemistry*, 90(7), 4521–4528.
- Bhinderwala, F., Wase, N., DiRusso, C., & Powers, R. (2018). Combining mass spectrometry and NMR improves metabolite detection and annotation. *Journal of Proteome Research*, 17(11), 4017–4022.
- Bingol, K. (2018). Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods. *High-Throughput*, 7(2), 9.
- Birney, E., Vamathevan, J., & Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. *BioRxiv*, 203554.
- Blaive, B., Pietri, S., Miollan, M., Martel, S., Le Moigne, F., & Culcasi, M. (2000). Alpha- and beta-phosphorylated amines and pyrrolidines, a new class of low toxic highly sensitive 31P NMR pH Indicators. Modeling of pKa and chemical shift values as a

- function of substituents. *Journal of Biological Chemistry*, 275(26), 19505–19512. Available from <https://doi.org/10.1074/jbc.M001784200>.
- Blasco, H., Corcia, P., Moreau, C., Veau, S., Fournier, C., Vourc'h, P., Emond, P., Gordon, P., Pradat, P.-F., & Praline, J. (2010). 1H-NMR-based metabolomic profiling of CSF in early amyotrophic lateral sclerosis. *PLoS One*, 5(10), e13223.
- Blasco, H., Nadal-Desbarats, L., Pradat, P.-F., Gordon, P. H., Antar, C., Veyrat-Durebex, C., Moreau, C., Devos, D., Mavel, S., & Emond, P. (2014). Untargeted 1H-NMR metabolomics in CSF: Toward a diagnostic biomarker for motor neuron disease. *Neurology*, 82(13), 1167–1174.
- Blümich, B. (2019). Low-field and benchtop NMR. *Journal of Magnetic Resonance*, 306, 27–35.
- Bodenhausen, G., & Ruben, D. J. (1980). Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1), 185–189.
- Boersma, M., Solyanikova, I., Van Berkel, W., Vervoort, J., Golovleva, L., & Rietjens, I. (2001). 19F NMR metabolomics for the elucidation of microbial degradation pathways of fluorophenols. *Journal of Industrial Microbiology and Biotechnology*, 26(1–2), 22–34.
- Bosc, C., Broin, N., Fanjul, M., Saland, E., Farge, T., Courdy, C., Batut, A., Masoud, R., Larrue, C., & Skuli, S. (2020). Autophagy regulates fatty acid availability for oxidative phosphorylation through mitochondria-endoplasmic reticulum contact sites. *Nature Communications*, 11(1), 1–14.
- Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., & Liu, P. (2013). The human urine metabolome. *PLoS One*, 8(9), e73076.
- Braunschweiler, L., & Ernst, R. (1983). Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. *Journal of Magnetic Resonance* (1969), 53(3), 521–528.
- Brinson, R. G., Arbogast, L. W., Marino, J. P., & Delaglio, F. (2020). Best practices in utilization of 2D-NMR spectral data as the input for chemometric analysis in biopharmaceutical applications. *Journal of Chemical Information and Modeling*, 60(4), 2339–2355.
- Buchli, R., Meier, D., Martin, E., & Boesiger, P. (1994). Assessment of absolute metabolite concentrations in human tissue by 31P MRS in vivo. Part II: Muscle, liver, kidney. *Magnetic Resonance in Medicine*, 32(4), 453–458.
- Bunescu, A., Garric, J., Vollat, B., Canet-Soulas, E., Graveron-Demilly, D., & Fauville, F. (2010). In vivo proton HR-MAS NMR metabolic profile of the freshwater cladoceran *Daphnia magna*. *Molecular Biosystems*, 6(1), 121–125.
- Cady, E., Dawson, M. J., Hope, P., Tofts, P., Costello, A., de, L., Delpy, D., Reynolds, E., & Wilkie, D. (1983). Non-invasive investigation of cerebral metabolism in newborn infants by phosphorus nuclear magnetic resonance spectroscopy. *The Lancet*, 321 (8333), 1059–1062.
- Cao, Q., Liu, H., Zhang, G., Wang, X., Manyande, A., & Du, H. (2020). 1H-NMR based metabolomics reveals the nutrient differences of two kinds of freshwater fish soups before and after simulated gastrointestinal digestion. *Food & Function*, 11(4), 3095–3104.
- Carlstrom, L., Weis, J., Johansson, L., Korsgren, O., & Ahlstrom, H. (2017). Pre-transplantation 31P-magnetic resonance spectroscopy for quality assessment of human pancreatic grafts – A feasibility study. *Magnetic Resonance Imaging*, 39, 98–102.

- Carotenuto, D., Luchinat, C., Marcon, G., Rosato, A., & Turano, P. (2015). The Da Vinci European BioBank: A metabolomics-driven infrastructure. *Journal of Personalized Medicine*, 5(2), 107–119.
- Castañar, L. (2017). Pure shift ^1H NMR: What is next? *Magnetic Resonance in Chemistry*, 55(1), 47–53.
- Cavallari, E., et al. (2020). In-vitro NMR Studies of Prostate Tumor Cell Metabolism by Means of Hyperpolarized [^{13}C]Pyruvate Obtained Using the PHIP-SAH Method. *Frontiers in Oncology*, 10(497). Available from <https://doi.org/10.3389/fonc.2020.00497>.
- Cavanagh, J., Fairbrother, W. J., Palmer, A. G., III, & Skelton, N. J. (1995). *Protein NMR spectroscopy: Principles and practice*. Elsevier.
- Chan, E. C. Y., Koh, P. K., Mal, M., Cheah, P. Y., Eu, K. W., Backshall, A., Cavill, R., Nicholson, J. K., & Keun, H. C. (2009). Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *Journal of Proteome Research*, 8(1), 352–361.
- Chandra, K., Al-Harthi, S., Sukumaran, S., Almulhim, F., Emwas, A.-H., Atreya, H. S., Jaremko, Ł., & Jaremko, M. (2021). NMR-based metabolomics with enhanced sensitivity. *RSC Advances*, 11(15), 8694–8700.
- Chiechio, S., Canonico, P. L., & Grilli, M. (2018). L-Acetylcarnitine: A mechanistically distinctive and potentially rapid-acting antidepressant drug. *International Journal of Molecular Sciences*, 19(1), 11.
- Chisca, S., Duong, P., Emwas, A.-H., Sougrat, R., & Nunes, S. P. (2015). Crosslinked copolyazoles with a zwitterionic structure for organic solvent resistant membranes. *Polymer Chemistry*, 6(4), 543–554.
- Chorao, C., Traïkia, M., Besse-Hoggan, P., Sancelme, M., Bligny, R., Gout, E., Mailhot, G., & Delort, A. (2010). In vivo ^{31}P and ^{13}C NMR investigations of *Rhodococcus rhodochrous* metabolism and behaviour during biotransformation processes. *Journal of Applied Microbiology*, 108(5), 1733–1743.
- Chu, S., Maltsev, S., Emwas, A.-H., & Lorigan, G. A. (2010). Solid-state NMR paramagnetic relaxation enhancement immersion depth studies in phospholipid bilayers. *Journal of Magnetic Resonance*, 207(1), 89–94.
- Claridge, T. D. (2016). *High-resolution NMR techniques in organic chemistry* (Vol. 27). Elsevier.
- Clendinen, C. S., Lee-McMullen, B., Williams, C. M., Stupp, G. S., Vandeborne, K., Hahn, D. A., Walter, G. A., & Edison, A. S. (2014). ^{13}C NMR metabolomics: Applications at natural abundance. *Analytical Chemistry*, 86(18), 9242–9250.
- Clendinen, C. S., Stupp, G. S., Ajredini, R., Lee-McMullen, B., Beecher, C., & Edison, A. S. (2015). An overview of methods using ^{13}C for improved compound identification in metabolomics and natural products. *Frontiers in Plant Science*, 6, 611.
- Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., & Holmes, E. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Analytical Chemistry*, 77(5), 1282–1289.
- Cui, G., Liew, Y. J., Li, Y., Kharbatia, N., Zahran, N. I., Emwas, A.-H., Eguiluz, V. M., & Aranda, M. (2019). Host-dependent nitrogen recycling as a mechanism of symbiont control in *Aiptasia*. *PLoS Genetics*, 15(6), e1008189.

- Cui, J., Zhu, D., Su, M., Tan, D., Zhang, X., Jia, M., & Chen, G. (2019). The combined use of 1H and 2D NMR-based metabolomics and chemometrics for non-targeted screening of biomarkers and identification of reconstituted milk. *Journal of the Science of Food and Agriculture*, 99(14), 6455–6461.
- Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., Westler, W. M., Eghbalnia, H. R., Sussman, M. R., & Markley, J. L. (2008). Metabolite identification via the Madison metabolomics consortium database. *Nature Biotechnology*, 26 (2), 162–164.
- Čuperlović-Culf, M. (2012). *NMR metabolomics in cancer research*. Elsevier.
- Darpolar, M. M., et al. (2014). The aspartate metabolism pathway is differentiable in human hepatocellular carcinoma: transcriptomics and 13C-isotope based metabolomics. *NMR in Biomedicine*, 27(4), 381–389. Available from <https://doi.org/10.1002/nbm.3072>.
- da Silva, R. R., Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41), 12549–12550.
- de Brouwer, H., & Stegeman, G. (2011). A LEAN approach Toward automated analysis and data processing of polymers using proton NMR spectroscopy. *JALA: Journal of the Association for Laboratory Automation*, 16(1), 1–16.
- De Livera, A. M., Olshansky, M., & Speed, T. P. (2013). *Statistical analysis of metabolomics data. Metabolomics tools for natural product discovery* (pp. 291–307). Springer.
- Dey, A., Charrier, B., Martineau, E., Deborde, C., Gandriaux, E., Moing, A., Jacob, D., Eshchenko, D., Schnell, M., Melzi, R., Kurzbach, D., Ceillier, M., Chappuis, Q., Cousin, S. F., Kempf, J. G., Jannin, S., Dumez, J.-N., & Giraudeau, P. (2020). Hyperpolarized NMR metabolomics at natural 13C abundance. *Analytical Chemistry*, 92(22), 14867–14871. Available from <https://doi.org/10.1021/acs.analchem.0c03510>.
- Dahri, M., Sioud, S., Dridi, R., Hassine, M., Boughattas, N. A., Almulhim, F., Al Talla, Z., Jaremko, M., & Emwas, A.-H. M. (2020). Extraction, characterization, and anticoagulant activity of a sulfated polysaccharide from *Bursatella leachii* viscera. *ACS Omega*, 5(24), 14786–14795.
- Diaz, G., Miranda, I. L., & Diaz, M. A. N. (2015). Quinolines, isoquinolines, angustureine, and congeneric alkaloids—occurrence, chemistry, and biological activity. *Phytochemicals-Isolation, Characterisation and Role in Human Health*.
- Diaz, O., Fernandez, M., De Fernando, G. D. G., de la Hoz, L., & Ordoñez, J. A. (1997). Proteolysis in dry fermented sausages: The effect of selected exogenous proteases. *Meat Science*, 46(1), 115–128.
- Dona, A. C., Kyriakides, M., Scott, F., Shephard, E. A., Varshavi, D., Veselkov, K., & Everett, J. R. (2016). A guide to the identification of metabolites in NMR-based metabolomics/metabolomics experiments. *Computational and Structural Biotechnology Journal*, 14, 135–153.
- Dos Santos, V. A. M., Hardt, C., Skrede, S., & Saccenti, E. (2020). *Systems and precision medicine in necrotizing soft tissue infections*. *Necrotizing Soft Tissue Infections* (pp. 187–207). Springer.
- Dossey, A. T., Walse, S. S., Conle, O. V., & Edison, A. S. (2007). Parectadial, a monoterpenoid from the defensive spray of *Parectatosoma mocquerysi*. *Journal of Natural Products*, 70(8), 1335–1338.
- Duarte, C. M., Jaremko, Ł., & Jaremko, M. (2020). Hypothesis: Potentially systemic impacts of elevated CO₂ on the human proteome and health. *Frontiers in Public Health*, 8.

- Duarte, D., Castro, B., Pereira, J. L., Marques, J. F., Costa, A. L., & Gil, A. M. (2020). Evaluation of saliva stability for NMR metabolomics: Collection and handling protocols. *Metabolites*, 10(12), 515.
- Duarte, I. F., Diaz, S. O., & Gil, A. M. (2014). NMR metabolomics of human blood and urine in disease research. *Journal of Pharmaceutical and Biomedical Analysis*, 93, 17–26.
- Dudka, I., et al. (2020). Comprehensive metabolomics analysis of prostate cancer tissue in relation to tumor aggressiveness and TMPRSS2-ERG fusion status. *BMC cancer*, 20 (1), 437. Available from <https://doi.org/10.1186/s12885-020-06908-z>.
- Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., Nicholson, J. K., Stamler, J., Elliott, P., & Chan, Q. (2006). Assessment of analytical reproducibility of ¹H NMR spectroscopy based metabolomics for large-scale epidemiological research: The INTERMAP Study. *Analytical Chemistry*, 78(7), 2199–2208.
- Dumez, J.-N., et al. (2015). Hyperpolarized NMR of plant and cancer cell extracts at natural abundance. *Analyst*, 140(17), 5860–5863. Available from <https://doi.org/10.1039/C5AN01203A>.
- Eckhardt, B. J., Gulick, R. M., Cohen, J., Powderly, W. G., & Opal, S. M. (2017). 152—Drugs for HIV Infection (pp. 1293–1308). Elsevier e2. Available from <https://doi.org/10.1016/B978-0-7020-6285-8.00152-0>.
- Edgar, M., Percival, B. C., Gibson, M., Jafari, F., & Grootveld, M. (2021). Low-field bench-top NMR spectroscopy as a potential non-stationary tool for point-of-care urinary metabolite tracking in diabetic conditions. *Diabetes Research and Clinical Practice*, 171, 108554.
- Edison, A. S., Le Guennec, A., Delaglio, F., & Kupče, Ě. (2019). Practical guidelines for ¹³C-based NMR metabolomics. In G. A. N. Gowda & D. Raftery, (Eds.), *NMR-Based Metabolomics: Methods and Protocols* (pp. 69–95). Springer: New York.
- Ellinger, J. J., Chylla, R. A., Ulrich, E. L., & Markley, J. L. (2013). Databases and software for NMR-based metabolomics. *Current Metabolomics*, 1(1), 28–40.
- Emwas, A.-H., Alghrably, M., Al-Harthi, S., Gabriel Poulsom, B., Szczepski, K., Chandra, K., & Jaremko, M. (2019). *New advances in fast methods of 2D NMR experiments. Nuclear magnetic resonance*. IntechOpen.
- Emwas, A.-H., Luchinat, C., Turano, P., Tenori, L., Roy, R., Salek, R. M., Ryan, D., Merzaban, J. S., Kaddurah-Daouk, R., & Zeri, A. C. (2015). Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: A review. *Metabolomics: Official Journal of the Metabolomic Society*, 11(4), 872–894.
- Emwas, A.-H., Roy, R., McKay, R. T., Ryan, D., Brennan, L., Tenori, L., Luchinat, C., Gao, X., Zeri, A. C., & Gowda, G. N. (2016). Recommendations and standardization of biomarker quantification using NMR-based metabolomics with particular focus on urinary analysis. *Journal of Proteome Research*, 15(2), 360–373.
- Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G., Raftery, D., Alahmari, F., Jaremko, L., & Jaremko, M. (2019). NMR spectroscopy for metabolomics research. *Metabolites*, 9(7), 123.
- Emwas, A.-H., Saccenti, E., Gao, X., McKay, R. T., Dos Santos, V. A. M., Roy, R., & Wishart, D. S. (2018). Recommended strategies for spectral processing and post-processing of 1D ¹H-NMR data of biofluids with a particular focus on urine. *Metabolomics: Official Journal of the Metabolomic Society*, 14(3), 1–23.
- Emwas, A.-H., Saunders, M., Ludwig, C., & Günther, U. (2008). Determinants for optimal enhancement in ex situ DNP experiments. *Applied Magnetic Resonance*, 34(3–4), 483–494.

- Emwas, A.-H. M. (2015). *The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research*. *Metabonomics* (pp. 161–193). Springer.
- Emwas, A.-H. M., Al-Talla, Z. A., & Kharbatia, N. M. (2015). *Sample collection and preparation of biofluids and extracts for gas chromatography–mass spectrometry*. *Metabonomics* (pp. 75–90). Springer.
- Emwas, A.-H. M., Salek, R. M., Griffin, J. L., & Merzaban, J. (2013). NMR-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics: Official Journal of the Metabolomic Society*, 9(5), 1048–1072.
- Emwas, A. M., Al-Talla, Z. A., Guo, X., Al-Ghamdi, S., & Al-Masri, H. T. (2013). Utilizing NMR and EPR spectroscopy to probe the role of copper in prion diseases. *Magnetic Resonance in Chemistry*, 51(5), 255–268.
- Enderle, J. D. (2012). *Biochemical reactions and enzyme kinetics. Introduction to biomedical engineering* (pp. 447–508). Elsevier.
- Esmaeili, M., Bathen, T. F., Engebråten, O., Mælandsmo, G. M., Gribbestad, I. S., & Moestue, S. A. (2014). Quantitative ³¹P HR-MAS MR spectroscopy for detection of response to PI3K/mTOR inhibition in breast cancer xenografts. *Magnetic Resonance in Medicine*, 71(6), 1973–1981.
- Euceda, L. R., Giskeødegård, G. F., & Bathen, T. F. (2015). Preprocessing of NMR metabolomics data. *Scandinavian Journal of Clinical and Laboratory Investigation*, 75(3), 193–203.
- Fan, T. W.-M., Lane, A. N., & Higashi, R. M. (2012). Stable isotope resolved metabolomics analysis of ribonucleotide and RNA metabolism in human lung cancer cells. *Metabonomics*, 8(3), 517–527. Available from <https://doi.org/10.1007/s11306-011-0337-9>.
- Fan, T. W.-M., & Lane, A. N. (2016). Applications of NMR spectroscopy to systems biochemistry. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 92, 18–53.
- Fan, T. W.-M., Lane, A. N., & Higashi, R. M. (2016). Stable isotope resolved metabolomics studies in ex vivo tissue slices. *Bio-protocol*, 6(3).
- Farag, M. A., Mahrous, E. A., Lübken, T., Porzel, A., & Wessjohann, L. (2014). Classification of commercial cultivars of *Humulus lupulus* L. (hop) by chemometric pixel analysis of two dimensional nuclear magnetic resonance spectra. *Metabonomics: Official Journal of the Metabolomic Society*, 10(1), 21–32.
- Farjon, J. (2017). How to face the low intrinsic sensitivity of 2D heteronuclear NMR with fast repetition techniques: Go faster to go higher!. *Magnetic Resonance in Chemistry*, 55(10), 883–892.
- Felsenfeld, A. J., & Levine, B. S. (2015). *Pathophysiology of calcium, phosphorus, and magnesium in chronic kidney disease*. *Chronic renal disease* (pp. 391–405). Elsevier.
- Féraud, B., Leenders, J., Martineau, E., Giraudeau, P., Govaerts, B., & De Tullio, P. (2019). Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics. *Metabonomics: Official Journal of the Metabolomic Society*, 15(4), 1–14.
- Féraud, B., Martineau, E., Leenders, J., Govaerts, B., de Tullio, P., & Giraudeau, P. (2020). Combining rapid 2D NMR experiments with novel pre-processing workflows and MIC quality measures for metabolomics. *Metabonomics: Official Journal of the Metabolomic Society*, 16(4), 42. Available from <https://doi.org/10.1007/s11306-020-01662-6>.
- Flores, A., Manfron Schiefer, E., Sassaki, G., Menezes, L., Fonseca, R., Cunha, R., Canziani, M. E., Guedes, M., Moreno-Amaral, A. N., & Souza, W. (2020). P1057 untargeted 1h NMR-based serum metabolic profile analysis of patients treated with

- high volume hemodiafiltration (HDF). *Nephrology Dialysis Transplantation*, 35(Suppl. 3), gfaa142–P1057.
- Flores-Sanchez, I. J., Choi, Y. H., & Verpoorte, R. (2012). *Metabolite analysis of Cannabis sativa L. by NMR spectroscopy*. *Functional Genomics* (pp. 363–375). Springer.
- Foxall, P., Parkinson, J., Sadler, I., Lindon, J., & Nicholson, J. (1993). Analysis of biological fluids using 600 MHz proton NMR spectroscopy: Application of homonuclear two-dimensional J-resolved spectroscopy to urine and blood plasma for spectral simplification and assignment. *Journal of Pharmaceutical and Biomedical Analysis*, 11(1), 21–31.
- Freeman, C. D., Klutman, N. E., & Lamp, K. C. (1997). Metronidazole. *Drugs*, 54(5), 679–708.
- Frydman, L., Scherf, T., & Lupulescu, A. (2002). The acquisition of multidimensional NMR spectra within a single scan. *Proceedings of the National Academy of Sciences*, 99(25), 15858–15862.
- Gallo, A., Farinha, A. S., Dinis, M., Emwas, A.-H., Santana, A., Nielsen, R. J., Goddard, W. A., & Mishra, H. (2019). The chemical reactions in electrosprays of water do not always correspond to those at the pristine air–water interface. *Chemical Science*, 10(9), 2566–2577.
- Gallo, V., Intini, N., Mastorilli, P., Latronico, M., Scapicchio, P., Triggiani, M., Bevilacqua, V., Fanizzi, P., Acquotti, D., & Airoldi, C. (2015). Performance assessment in fingerprinting and multi component quantitative NMR analyses. *Analytical Chemistry*, 87(13), 6709–6717.
- García-García, A. B., Lamichhane, S., Castejón, D., Cambero, M. I., & Bertram, H. C. (2018). ¹H HR-MAS NMR-based metabolomics analysis for dry-fermented sausage characterization. *Food Chemistry*, 240, 514–523.
- Gargallo-Garriga, A., et al. (2020). (31)P-NMR Metabolomics Revealed Species-Specific Use of Phosphorous in Trees of a French Guiana Rainforest. *Molecules*, 25(17). Available from <https://doi.org/10.3390/molecules25173960>.
- Gattineni, J., & Friedman, P. A. (2015). Regulation of hormone-sensitive renal phosphate transport. *Vitamins & Hormones*, 98, 249–306.
- Gebregiworgis, T., & Powers, R. (2012). Application of NMR metabolomics to search for human disease biomarkers. *Combinatorial Chemistry & High Throughput Screening*, 15(8), 595–610.
- Geier, F. M., Leroi, A. M., & Bundy, J. G. (2019). ¹³C labeling of nematode worms to improve metabolome coverage by heteronuclear nuclear magnetic resonance experiments. *Frontiers in Molecular Biosciences*, 6, 27.
- Ghini, V., Quaglio, D., Luchinat, C., & Turano, P. (2019). NMR for sample quality assessment in metabolomics. *New Biotechnology*, 52, 25–34.
- Gil, A., Duarte, I., Cabrita, E., Goodfellow, B., Spraul, M., & Kerssebaum, R. (2004). Exploratory applications of diffusion ordered spectroscopy to liquid foods: An aid towards spectral assignment. *Analytica Chimica Acta*, 506(2), 215–223.
- Giraudeau, P. (2020). NMR-based metabolomics and fluxomics: Developments and future prospects. *Analyst*, 145(7), 2457–2472.
- Giraudeau, P., Guignard, N., Hillion, E., Baguet, E., & Akoka, S. (2007). Optimization of homonuclear 2D NMR for fast quantitative analysis: Application to tropine–nortropine mixtures. *Journal of Pharmaceutical and Biomedical Analysis*, 43(4), 1243–1248.
- Giraudeau, P., Silvestre, V., & Akoka, S. (2015). Optimizing water suppression for quantitative NMR-based metabolomics: A tutorial review. *Metabolomics: Official Journal of the Metabolomic Society*, 11(5), 1041–1055.

- Gogiashvili, M., Nowacki, J., Hergenröder, R., Hengstler, J. G., Lambert, J., & Edlund, K. (2019). HR-MAS NMR based quantitative metabolomics in breast cancer. *Metabolites*, 9(2), 19.
- Gout, E., Bligny, R., Douce, R., Boisson, A., & Rivasseau, C. (2011). Early response of plant cell to carbon deprivation: In vivo ^{31}P -NMR spectroscopy shows a quasi-instantaneous disruption on cytosolic sugars, phosphorylated intermediates of energy metabolism, phosphate partitioning, and intracellular pHs. *New Phytologist*, 189(1), 135–147.
- Griffiths, L., & Horton, R. (1998). Optimization of LC–NMR. III—Increased signal-to-noise ratio through column trapping. *Magnetic Resonance in Chemistry*, 36(2), 104–109.
- Gueniec, A. L., Giraudieu, P., & Caldarelli, S. (2014). Evaluation of fast 2D NMR for metabolomics. *Analytical Chemistry*, 86(12), 5946–5954.
- Guillou, C., Trierweiler, M., & Martin, G. (1988). Repeatability and reproducibility of site-specific isotope ratios in quantitative ^2H NMR. *Magnetic Resonance in Chemistry*, 26 (6), 491–496.
- Halabalaki, M., Vougogiannopoulou, K., Mikros, E., & Skaltsounis, A. L. (2014). Recent advances and new strategies in the NMR-based identification of natural products. *Current Opinion in Biotechnology*, 25, 1–7.
- Halse, M. E. (2016). Perspectives for hyperpolarisation in compact NMR. *TrAC Trends in Analytical Chemistry*, 83, 76–83.
- Händel, H., Gesele, E., Gottschall, K., & Albert, K. (2003). Application of HRMAS ^1H NMR spectroscopy to investigate interactions between ligands and synthetic receptors. *Angewandte Chemie International Edition*, 42(4), 438–442.
- Haner, R. L., & Keifer, P. A. (2007). Flow Probes for NMR spectroscopy. *EMagRes*.
- Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. (2012). BATMAN—An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics (Oxford, England)*, 28(15), 2088–2090.
- Harris, R. K., Becker, E. D., Cabral De Menezes, S. M., Granger, P., Hoffman, R. E., & Zilm, K. W. (2007). Further conventions for NMR shielding and chemical shifts (IUPAC Recommendations 2008). *EMagRes*.
- Hasanpour, M., Saberi, S., & Iranshahi, M. (2020). Metabolic profiling and untargeted ^1H -NMR-based metabolomics study of different Iranian pomegranate (*Punica granatum*) ecotypes. *Planta Medica*, 86(03), 212–219.
- Haviland, J. A., et al. (2013). NMR-based metabolomics and breath studies show lipid and protein catabolism during low dose chronic T1AM treatment. *Obesity*, 21(12), 2538–2544. Available from <https://doi.org/10.1002/oby.20391>.
- Heude, C., Lemasson, E., Elbayad, K., & Piotto, M. (2015). Rapid assessment of fish freshness and quality by ^1H HR-MAS NMR spectroscopy. *Food Analytical Methods*, 8(4), 907–915.
- Hill, D. K., Mariotti, E., & Ekykyn, T. R. (2018). *Imaging metabolic processes in living systems with hyperpolarised ^{13}C magnetic resonance* (pp. 280–309). The Royal Society of Chemistry Chapter 11. Available from <https://doi.org/10.1039/9781782627937-00280>.
- Holmes, E., Wilson, I. D., & Nicholson, J. K. (2008). Metabolic phenotyping in health and disease. *Cell*, 134(5), 714–717.
- Hong, S.-M., Choi, C.-H., Magill, A. W., Shah, N. J., & Felder, J. (2018). Design of a quadrature $^1\text{H}/^{31}\text{P}$ coil using bent dipole antenna and four-channel loop at 3T MRI. *IEEE Transactions on Medical Imaging*, 37(12), 2613–2618.

- Hoult, D. (1976). Solvent peak saturation with single phase and quadrature Fourier transformation. *Journal of Magnetic Resonance* (1969), 21(2), 337–347.
- Howarth, A., Ermanis, K., & Goodman, J. M. (2020). DP4-AI automated NMR data analysis: Straight from spectrometer to structure. *Chemical Science*, 11(17), 4351–4359.
- Hunt, C. T., Boulanger, Y., Fesik, S. W., & Armitage, I. M. (1984). NMR analysis of the structure and metal sequestering properties of metallothioneins. *Environmental Health Perspectives*, 54, 135–145.
- Izquierdo-Garcia, J. L., Comella-Del-Barrio, P., Campos-Olivas, R., Casanova, F., Dominguez, J., & Ruiz-Cabello, J. (2019). *Benchtop NMR-based metabolomic analysis as a diagnostic tool for tuberculosis in clinical urine samples*. Eur Respiratory Soc.
- Izquierdo-Garcia, J. L., Padro, D., Villa, P., Fadon, L., & Cifuentes, A. (2021). *2.25—NMR-based metabolomics* (pp. 353–369). Elsevier. Available from <https://doi.org/10.1016/B978-0-08-100596-5.22909-0>.
- Jiang, C., Yang, K., Yang, L., Miao, Z., Wang, Y., & Zhu, H. (2013). A ^1H NMR-based metabonomic investigation of time-related metabolic trajectories of the plasma, urine and liver extracts of hyperlipidemic hamsters. *PLoS One*, 8(6), e66786.
- Johnson, K., Barrientos, L. G., Le, L., & Murthy, P. P. (1995). Application of two-dimensional total correlation spectroscopy for structure determination of individual inositol phosphates in a mixture. *Analytical Biochemistry*, 231(2), 421–431.
- Johnson, S. R., & Lange, B. M. (2015). Open-access metabolomics databases for natural product research: Present capabilities and future potential. *Frontiers in Bioengineering and Biotechnology*, 3, 22.
- Jordan, K., Adkins, C., Su, L., Halpern, E., Mark, E., Christiani, D., & Cheng, L. (2010). Comparison of squamous cell carcinoma and adenocarcinoma of the lung by metabolomic analysis of tissue–serum pairs. *Lung Cancer (Amsterdam, Netherlands)*, 68(1), 44–50.
- Jung, Y., Lee, J., Kwon, J., Lee, K.-S., Ryu, D. H., & Hwang, G.-S. (2010). Discrimination of the geographical origin of beef by ^1H NMR-based metabolomics. *Journal of Agricultural and Food Chemistry*, 58(19), 10458–10466.
- Kaebisch, E., Fuss, T. L., Vandergift, L. A., Toews, K., Habbel, P., & Cheng, L. L. (2017). Applications of high-resolution magic angle spinning MRS in biomedical studies I—cell line and animal models. *NMR in Biomedicine*, 30(6), e3700.
- Kaluarachchi, M., et al. (2018). A comparison of human serum and plasma metabolites using untargeted ^1H NMR spectroscopy and UPLC-MS. *Metabolomics*, 14(3), 32. Available from <https://doi.org/10.1007/s11306-018-1332-1>.
- Kamal, A., Shaik, A. B., Kumar, C. G., Mongolla, P., Rani, P. U., Krishna, K. V. S. R., Mamidyalu, S. K., & Joseph, J. (2012). Metabolic profiling and biological activities of bioactive compounds produced by *Pseudomonas* sp. strain ICTB-745 isolated from Ladakh, India. *Journal of Microbiology and Biotechnology*, 22(1), 69–79. Available from <https://doi.org/10.4014/jmb.1105.05008>.
- Kanamori, K. (2017). In vivo $\text{N}-15$ MRS study of glutamate metabolism in the rat brain. *Analytical Biochemistry*, 529, 179–192.
- Kanwal, S., Ann, N.-u., Fatima, S., Emwas, A.-H., Alazmi, M., Gao, X., Ibrar, M., Zaib Saleem, R. S., & Chotana, G. A. (2020). Facile synthesis of NH-free 5-(hetero)aryl-pyrrole-2-carboxylates by catalytic C–H borylation and Suzuki coupling. *Molecules (Basel, Switzerland)*, 25(9), 210. Available from <https://doi.org/10.3390/molecules25092106>.
- Karaman, I. (2017). Preprocessing and pretreatment of metabolomics data for statistical analysis. *Metabolomics: From Fundamentals to Clinical Applications*, 145–161.

- Karaman, I., Ferreira, D. L., Boulangé, C. L., Kaluarachchi, M. R., Herrington, D., Dona, A. C., Castagné, R., Moayyeri, A., Lehne, B., & Loh, M. (2016). Workflow for integrated processing of multicohort untargeted ^1H NMR metabolomics data in large-scale metabolic epidemiology. *Journal of Proteome Research*, 15(12), 4188–4194.
- Keeler, J. (2011). *Understanding NMR spectroscopy*. John Wiley & Sons.
- Keifer, P. (2007). Flow techniques in NMR spectroscopy. *Annual Reports on NMR Spectroscopy*, 62, 1–47.
- Keifer, P. A. (2003). Flow injection analysis NMR (FIA–NMR): A novel flow NMR technique that complements LC–NMR and direct injection NMR (DI–NMR). *Magnetic Resonance in Chemistry*, 41(7), 509–516.
- Keifer, P. A., Smallcombe, S. H., Williams, E. H., Salomon, K. E., Mendez, G., Belletire, J. L., & Moore, C. D. (2000). Direct-injection NMR (DI-NMR): A flow NMR technique for the analysis of combinatorial chemistry libraries. *Journal of Combinatorial Chemistry*, 2(2), 151–171.
- Kern, S., Wander, L., Meyer, K., Guhl, S., Mukkula, A. R. G., Holtkamp, M., Salge, M., Fleischer, C., Weber, N., & King, R. (2019). Flexible automation with compact NMR spectroscopy for continuous production of pharmaceuticals. *Analytical and Bioanalytical Chemistry*, 411(14), 3037–3046.
- Keun, H. C., Beckonert, O., Griffin, J. L., Richter, C., Moskau, D., Lindon, J. C., & Nicholson, J. K. (2002). Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies. *Analytical Chemistry*, 74(17), 4588–4593.
- Keun, H. C., Ebbels, T. M., Antti, H., Bolland, M. E., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E., & Lindon, J. C. (2002). Analytical reproducibility in ^1H NMR-based metabonomic urinalysis. *Chemical Research in Toxicology*, 15(11), 1380–1386.
- Khasaeva, F., Parshikov, I., & Zaraisky, E. (2016). Biodegradation of 4-methylpyridine by *Arthrobacter* sp. *Asian Journal of Microbiology, Biotechnology and Environmental Sciences*, 18(1), 75–77.
- Kijewska, M., Sharafaldin, A. A., Jaremko, Ł., Cal, M., Setner, B., Siczek, M., Stefanowicz, P., Hussien, M. A., Emwas, A.-H., & Jaremko, M. (2021). Lossen rearrangement of p-toluenesulfonates of N-oxyimides in basic condition, theoretical study, and molecular docking. *Frontiers in Chemistry*, 9, 189.
- Kim, H. K., Choi, Y. H., & Verpoorte, R. (2010). NMR-based metabolomic analysis of plants. *Nature Protocols*, 5(3), 536–549.
- Kirwan, J. A., Brennan, L., Broadhurst, D., Fiehn, O., Cascante, M., Dunn, W. B., Schmidt, M. A., & Velagapudi, V. (2018). Preanalytical processing and biobanking procedures of biological samples for metabolomics research: A white paper, community perspective (for “Precision Medicine and Pharmacometabolomics Task Group”—The Metabolomics Society Initiative). *Clinical Chemistry*, 64(8), 1158–1182.
- Kogler, H., Sørensen, O., Bodenhausen, G., & Ernst, R. (1983). Low-pass J filters. Suppression of neighbor peaks in heteronuclear relayed correlation spectra. *Journal of Magnetic Resonance*, 55, 157–163.
- Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., & Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics: Official Journal of the Metabolomic Society*, 8(1), 146–160.
- Komoroski, R. A., Pearce, J. M., & Mrak, R. E. (2008). ^{31}P NMR spectroscopy of phospholipid metabolites in postmortem schizophrenic brain. *Magnetic Resonance in*

- Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 59(3), 469–474.
- Kono, H. (2013). ¹H and ¹³C chemical shift assignment of the monomers that comprise carboxymethyl cellulose. *Carbohydrate Polymers*, 97(2), 384–390.
- Koskela, H., et al. (2018). pH-Dependent Piecewise Linear Correlation of ¹H,³¹P Chemical Shifts: Application in NMR Identification of Nerve Agent Metabolites in Urine Samples. *Analytical Chemistry*, 90(14), 8495–8500. Available from <https://doi.org/10.1021/acs.analchem.8b01308>.
- Kovacs, H., Moskau, D., & Spraul, M. (2005). Cryogenically cooled probes—a leap in NMR technology. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2(46), 131–155.
- Kövér, K. E., & Batta, G. (1987). Strong coupling effects and their suppression in two-dimensional heteronuclear NOE experiments. *Journal of Magnetic Resonance (1969)*, 74(3), 397–405.
- Kovtunov, K. V., et al. (2014). Propane-d6 Heterogeneously Hyperpolarized by Parahydrogen. *The Journal of Physical Chemistry C*, 118(48), 28234–28243.
- Kovtunov, K. V., Pokochueva, E. V., Salnikov, O. G., Cousin, S. F., Kurzbach, D., Vuichoud, B., Jannin, S., Chekmenev, E. Y., Goodson, B. M., & Barskiy, D. A. (2018). Hyperpolarized NMR spectroscopy: d-DNP, PHIP, and SABRE Techniques. *Chemistry—An Asian Journal*, 13(15), 1857–1871.
- Krikken, E., et al. (2019). Early detection of changes in phospholipid metabolism during neoadjuvant chemotherapy in breast cancer patients using phosphorus magnetic resonance spectroscopy at 7T. *NMR in Biomedicine*, 32(6), e4086. Available from <https://doi.org/10.1002/nbm.4086>.
- Krishnamurthy, K. (2013). CRAFT (complete reduction to amplitude frequency table) – Robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR. *Magnetic Resonance in Chemistry*, 51(12), 821–829.
- Kupče, Ė., & Freeman, R. (2003). Frequency-domain Hadamard spectroscopy. *Journal of Magnetic Resonance*, 162(1), 158–165.
- Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., Ogiwara, A., Meissen, J., Showalter, M., & Takeuchi, K. (2018). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods*, 15(1), 53–56.
- Lalaleo, L., et al. (2020). Differentiating, evaluating, and classifying three quinoa ecotypes by washing, cooking and germination treatments, using ¹H NMR-based metabolomic approach. *Food Chemistry*, 331, 127351. Available from <https://doi.org/10.1016/j.foodchem.2020.127351>.
- Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews. Genetics*, 19(4), 208.
- Larson, P. E. Z., et al. (2018). Investigation of analysis methods for hyperpolarized ¹³C-pyruvate metabolic MRI in prostate cancer patients. *NMR in Biomedicine*, 31(11), e3997. Available from <https://doi.org/10.1002/nbm.3997>.
- Laserna, A. K. C., Lai, Y., Fang, G., Ganapathy, R., Atan, M. S. B. M., Lu, J., Wu, J., Uttamchandani, M., Moochhala, S. M., & Li, S. F. Y. (2020). Metabolic profiling of a porcine combat trauma-injury model using NMR and multi-mode LC-MS metabolomics—A preliminary study. *Metabolites*, 10(9), 373.
- Lawson, I. J., Ewart, C., Kraft, A., & Ellis, D. (2020). Demystifying NMR spectroscopy: Applications of benchtop spectrometers in the undergraduate teaching laboratory. *Magnetic Resonance in Chemistry*, 58(12), 1256–1260.

- Le Guennec, A., Dumez, J., Giraudeau, P., & Caldarelli, S. (2015). Resolution-enhanced 2D NMR of complex mixtures by non-uniform sampling. *Magnetic Resonance in Chemistry*, 53(11), 913–920.
- Le Guennec, A., Tea, I., Antheaume, I., Martineau, E., Charrier, B., Pathan, M., Akoka, S., & Giraudeau, P. (2012). Fast determination of absolute metabolite concentrations by spatially encoded 2D NMR: Application to breast cancer cell extracts. *Analytical Chemistry*, 84(24), 10831–10837.
- Lee, W., Ko, B. J., Sim, Y. E., Suh, S., Yoon, D., & Kim, S. (2019). Discrimination of human urine from animal urine using 1H-NMR. *Journal of Analytical Toxicology*, 43(1), 51–60.
- Levin, Y. S., Albers, M. J., Butler, T. N., Spielman, D., Peehl, D. M., & Kurhanewicz, J. (2009). Methods for metabolic evaluation of prostate cancer cells using proton and 13C HR-MAS spectroscopy and [3-13C] pyruvate as a metabolic substrate. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(5), 1091–1098.
- Levine, J., Panchalingam, K., McClure, R., Gershon, S., & Pettegrew, J. (2003). Effects of acetyl-L-carnitine and myo-inositol on high-energy phosphate and membrane phospholipid metabolism in Zebra Fish: A 31P-NMR-Spectroscopy Study. *Neurochemical Research*, 28(5), 687–690.
- Lewis, I. A., Karsten, R. H., Norton, M. E., Tonelli, M., Westler, W. M., & Markley, J. L. (2010). NMR method for measuring carbon-13 isotopic enrichment of metabolites in complex solutions. *Analytical Chemistry*, 82(11), 4558–4563.
- Lewis, I. A., Schommer, S. C., Hodis, B., Robb, K. A., Tonelli, M., Westler, W. M., Sussman, M. R., & Markley, J. L. (2007). Method for determining molar concentrations of metabolites in complex solutions from two-dimensional 1H–13C NMR spectra. *Analytical Chemistry*, 79(24), 9385–9390.
- Lewis, I. A., Schommer, S. C., & Markley, J. L. (2009). rNMR: Open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance in Chemistry*, 47(S1), S123–S126.
- Li, X., Luo, H., Huang, T., Xu, L., Shi, X., & Hu, K. (2019). Statistically correlating NMR spectra and LC-MS data to facilitate the identification of individual metabolites in metabolomics mixtures. *Analytical and Bioanalytical Chemistry*, 411(7), 1301–1309.
- Li, Y., Wang, C., Li, D., Deng, P., Shao, X., Hu, J., Liu, C., Jie, H., Lin, Y., & Li, Z. (2017). 1H-NMR-based metabolic profiling of a colorectal cancer CT-26 lung metastasis model in mice. *Oncology Reports*, 38(5), 3044–3054.
- Liang, Y.-S., Kim, H., Lefeber, A., Erkelens, C., Choi, Y., & Verpoorte, R. (2006). Identification of phenylpropanoids in methyl jasmonate treated Brassica rapa leaves using two-dimensional nuclear magnetic resonance spectroscopy. *Journal of Chromatography A*, 1112(1–2), 148–155.
- Lipfert, M., Rout, M. K., Berjanskii, M., & Wishart, D. S. (2019). *Automated tools for the analysis of 1D-NMR and 2D-NMR spectra. NMR-based metabolomics* (pp. 429–449). Springer.
- Lloyd, L. S., Adams, R. W., Bernstein, M., Coombes, S., Duckett, S. B., Green, G. G., Lewis, R. J., Mewis, R. E., & Sleigh, C. J. (2012). Utilization of SABRE-derived hyperpolarization to detect low-concentration analytes via 1D and 2D NMR methods. *Journal of the American Chemical Society*, 134(31), 12904–12907.
- Lopez, J. M., Cabrera, R., & Maruenda, H. (2019). Ultra-clean pure shift 1H-NMR applied to metabolomics profiling. *Scientific Reports*, 9(1), 1–8.

- Lown, J. W., & Hanstock, C. C. (1985). High field ^1H -NMR analysis of the 1:1 intercalation complex of the antitumor agent mitoxantrone and the DNA duplex [d (CpGpCpG)]. *Journal of Biomolecular Structure and Dynamics*, 2(6), 1097–1106.
- Ludwig, C., Easton, J. M., Lodi, A., Tiziani, S., Manzoor, S. E., Southam, A. D., Byrne, J. J., Bishop, L. M., He, S., & Arvanitis, T. N. (2012). Birmingham Metabolite Library: A publicly accessible database of 1-D ^1H and 2-D ^1H -resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics: Official Journal of the Metabolomic Society*, 8(1), 8–18.
- Ludwig, C., Marin-Montesinos, I., Saunders, M. G., Emwas, A.-H., Pikramenou, Z., Hammond, S. P., & Günther, U. L. (2010). Application of ex situ dynamic nuclear polarization in studying small molecules. *Physical Chemistry Chemical Physics*, 12 (22), 5868–5871.
- Ludwig, C., & Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: Review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*, 21(1), 22–32.
- Luke, T. D., Pryce, J. E., Wales, W. J., & Rochfort, S. J. (2020). A tale of two biomarkers: Untargeted ^1H NMR metabolomic fingerprinting of BHBA and NEFA in early lactation dairy cows. *Metabolites*, 10(6), 247.
- Lundberg, P., & Lundquist, P.-O. (2004). Primary metabolism in N₂-fixing *Alnus incana*–*Frankia* symbiotic root nodules studied with ^{15}N and ^{31}P nuclear magnetic resonance spectroscopy. *Planta*, 219(4), 661–672. Available from <https://doi.org/10.1007/s00425-004-1271-0>.
- Lutz, N. W., & Hull, W. E. (1999). Assignment and pH dependence of the ^{19}F -NMR resonances from the fluorouracil anabolites involved in fluoropyrimidine chemotherapy. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 12(4), 237–248.
- Lutz, N. W., Maillet, S., Nicoli, F., Viout, P., & Cozzzone, P. J. (1998). Further assignment of resonances in ^1H NMR spectra of cerebrospinal fluid (CSF). *FEBS Letters*, 425(2), 345–351.
- Macura, S., Kumar, N. G., & Brown, L. R. (1983). Combined use of COSY and double quantum two-dimensional NMR spectroscopy for elucidation of spin systems in polymyxin B. *Biochemical and Biophysical Research Communications*, 117(2), 486–492.
- Madrid-Gambin, F., Brunius, C., Garcia-Aloy, M., Estruel-Amades, S., Landberg, R., & Andres-Lacueva, C. (2018). Untargeted ^1H NMR-based metabolomics analysis of urine and serum profiles after consumption of lentils, chickpeas, and beans: An extended meal study to discover dietary biomarkers of pulses. *Journal of Agricultural and Food Chemistry*, 66(27), 6997–7005.
- Mahrous, E. A., & Farag, M. A. (2015). Two dimensional NMR spectroscopic approaches for exploring plant metabolome: A review. *Journal of Advanced Research*, 6(1), 3–15.
- Malloy, C. R., Maher, E., Marin-Valencia, I., Mickey, B., Deberardinis, R. J., & Sherry, A. D. (2010). *Carbon-13 nuclear magnetic resonance for analysis of metabolic pathways. Methodologies for Metabolomics: Experimental Strategies and Techniques* (pp. 415–445). Cambridge university Press.
- Malz, F., & Jancke, H. (2005). Validation of quantitative NMR. *Journal of Pharmaceutical and Biomedical Analysis*, 38(5), 813–823.
- Marcon, G., & Nincheri, P. (2014). The multispecialistic da Vinci European BioBank. *Open Journal of Bioresources*, 1.

- Marion, D. (2013). An introduction to biological NMR spectroscopy. *Molecular & Cellular Proteomics*, 12(11), 3006–3025.
- Markley, J. L., Anderson, M. E., Cui, Q., Eghbalnia, H. R., Lewis, I. A., Hegeman, A. D., Li, J., Schulte, C. F., Sussman, M. R., & Westler, W. M. (2007). *New bioinformatics resources for metabolomics*. *Biocomputing 2007* (pp. 157–168). World Scientific.
- Markley, J. L., Bax, A., Arata, Y., Hilbers, C., Kaptein, R., Sykes, B. D., Wright, P. E., & Wüthrich, K. (1998). Recommendations for the presentation of NMR structures of proteins and nucleic acids (IUPAC Recommendations 1998). *Pure and Applied Chemistry. Chimie Pure et Appliquée*, 70(1), 117–142.
- Markley, J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., & Wishart, D. S. (2017). The future of NMR-based metabolomics. *Current Opinion in Biotechnology*, 43, 34–40.
- Martineau, E., & Giraudeau, P. (2019). *Fast quantitative 2D NMR for untargeted and targeted metabolomics*. *NMR-based metabolomics* (pp. 365–383). Springer.
- Mattar, S. M., Emwas, A. H., & Calhoun, L. A. (2004). Spectroscopic studies of the intermediates in the conversion of 1, 4, 11, 12-tetrahydro-9, 10-anthraquinone to 9, 10-anthraquinone by reaction with oxygen under basic conditions. *The Journal of Physical Chemistry A*, 108(52), 11545–11553.
- McKay, R. T. (2009). Recent advances in solvent suppression for solution NMR: A practical reference. *Annual Reports on NMR Spectroscopy*, 66, 33–76.
- Mckay, R. T. (2011). How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: An examination of the pulse sequence components and evolution. *Concepts in Magnetic Resonance Part A*, 38(5), 197–220.
- McNally, D. J., Lamoureux, M., Li, J., Kelly, J., Brisson, J.-R., Szymanski, C. M., & Jarrell, H. C. (2006). HR-MAS NMR studies of 15N-labeled cells confirm the structure of the O-methyl phosphoramidate CPS modification in *Campylobacter jejuni* and provide insight into its biosynthesis. *Canadian Journal of Chemistry*, 84(4), 676–684.
- Mediani, A., Khatib, A., Ismail, A., Hamid, M., Lajis, N. H., Shaari, K., & Abas, F. (2017). Application of BATMAN and BAYESIL for quantitative ¹H-NMR based metabolomics of urine: Discriminant analysis of lean, obese, and obese-diabetic rats. *Metabolomics: Official Journal of the Metabolomic Society*, 13(11), 1–14.
- Meier, S., Jensen, P. R., Karlsson, M., & Lerche, M. H. (2014). Hyperpolarized NMR probes for biological assays. *Sensors*, 14(1), 1576–1597.
- Mercier, P., Lewis, M. J., Chang, D., Baker, D., & Wishart, D. S. (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. *Journal of Biomolecular NMR*, 49(3), 307–323.
- Mestrelab Research S.L.—Analytical Chemistry Software. (2021). <https://mestrelab.com>.
- Michel, N., & Akoka, S. (2004). The application of the ERETIC method to 2D-NMR. *Journal of Magnetic Resonance*, 168(1), 118–123.
- Miccheli, A., et al. (2015). Urinary ¹H-NMR-based metabolic profiling of children with NAFLD undergoing VSL#3 treatment. *International Journal of Obesity*, 39(7), 1118–1125. Available from <https://doi.org/10.1038/ijo.2015.40>.
- Misra, B. B., & Mohapatra, S. (2019). Tools and resources for metabolomics research community: A 2017–2018 update. *Electrophoresis*, 40(2), 227–246.
- Misra, B. B., & van der Hooft, J. J. (2016). Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis*, 37(1), 86–110.

- Mobli, M., & Hoch, J. C. (2014). Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 83, 21–41.
- Moe, S. M., & Daoud, J. R. (2014). *Disorders of mineral metabolism: Calcium, phosphorus, and magnesium*. In National Kidney Foundation Primer on Kidney Diseases (pp. 100–112). Elsevier.
- Mohammed, S. A. A., Khan, R. A., El-Readi, M. Z., Emwas, A.-H., Sioud, S., Poulsom, B. G., Jaremko, M., Eldeeb, H. M., Al-Omar, M. S., & Mohammed, H. A. (2020). Suaeda vermiculata aqueous-ethanolic extract-based mitigation of CCl₄-induced hepatotoxicity in rats, and HepG-2 and HepG-2/ADR cell-lines-based cytotoxicity evaluations. *Plants*, 9(10). Available from <https://doi.org/10.3390/plants9101291>.
- European Committee for Standardization. Molecular in vitro diagnostic examinations—Specifications for pre-examination processes for metabolomics in urine, venous blood serum and plasma. (2016). CEN Standard CEN/TS 16945 2016. <https://shop.bsigroup.com/ProductDetail?pid=00000000030339067>.
- International Organization for Standardization. Molecular in vitro diagnostic examinations—Specifications for pre-examination processes in metabolomics in urine, venous blood serum and plasma. (2020). ISO/DIS 23118. <https://www.iso.org/obp/ui#iso:std:iso:23118:dis:ed-1:v1:en>.
- Mroue, K. H., Emwas, A.-H. M., & Power, W. P. (2010). Solid-state ²⁷Al nuclear magnetic resonance investigation of three aluminum-centered dyes. *Canadian Journal of Chemistry*, 88(2), 111–123.
- Mulder, F. A., Tenori, L., & Luchinat, C. (2019). Fast and quantitative NMR metabolite analysis afforded by a paramagnetic co-solute. *Angewandte Chemie International Edition*, 58(43), 15283–15286.
- Nadal-Desbarats, L., et al. (2014). Combined ¹H-NMR and ¹H–¹³C HSQC-NMR to improve urinary screening in autism spectrum disorders. *Analyst*, 139(13), 3460–3468. Available from <https://doi.org/10.1039/C4AN00552J>.
- Nagana Gowda, G., & Raftery, D. (2017). Recent advances in NMR-based metabolomics. *Analytical Chemistry*, 89(1), 490–510.
- Nasca, C., Xenos, D., Barone, Y., Caruso, A., Scaccianoce, S., Matrisciano, F., Battaglia, G., Mathé, A. A., Pittaluga, A., & Lionetto, L. (2013). L-Acetylcarnitine causes rapid antidepressant effects through the epigenetic induction of mGlu2 receptors. *Proceedings of the National Academy of Sciences*, 110(12), 4804–4809.
- Naser, N., Abdul Jameel, A. G., Emwas, A.-H., Singh, E., Chung, S. H., & Sarathy, S. M. (2019). The influence of chemical composition on ignition delay times of gasoline fractions. *Combustion and Flame*, 209, 418–429. Available from <https://doi.org/10.1016/j.combustflame.2019.07.030>.
- Nemets, B., Fux, M., Levine, J., & Belmaker, R. (2001). Combination of antidepressant drugs: The case of inositol. *Human Psychopharmacology: Clinical and Experimental*, 16(1), 37–43.
- Nicholson, J. K., Holmes, E., Kinross, J. M., Darzi, A. W., Takats, Z., & Lindon, J. C. (2012). Metabolic phenotyping in clinical and surgical environments. *Nature*, 491(7424), 384–392.
- Nikolaou, P., Goodson, B. M., & Chekmenev, E. Y. (2015). NMR hyperpolarization techniques for biomedicine. *Chemistry—A European Journal*, 21(8), 3156–3166.
- Nishikawa, Y., et al. (2003). Vitamin C metabolomic mapping in experimental diabetes with 6-deoxy-6-fluoro-ascorbic acid and high resolution ¹⁹F-nuclear magnetic

- resonance spectroscopy. *Metabolism*, 52(6), 760–770. Available from [https://doi.org/10.1016/S0026-0495\(03\)00069-6](https://doi.org/10.1016/S0026-0495(03)00069-6).
- Nizioł, J., Ossoliński, K., Tripet, B. P., Copié, V., Arendowski, A., & Ruman, T. (2020). Nuclear magnetic resonance and surface-assisted laser desorption/ionization mass spectrometry-based serum metabolomics of kidney cancer. *Analytical and Bioanalytical Chemistry*, 412(23), 5827–5841.
- Ogunade, I. M., & Jiang, Y. (2019). PSIX-7 1H NMR-based plasma metabolomics reveals a potential biomarker of aflatoxin ingestion in dairy cows. *Journal of Animal Science*, 97(Suppl 3), 395–396. Available from <https://doi.org/10.1093/jas/skz258.789>.
- O'Sullivan, A., Avizonis, D., German, J. B., & Slupsky, C. M. (2007). Software tools for NMR metabolomics. *EMagRes*.
- Okazaki, Y., & Saito, K. (2012). Recent advances of metabolomics in plant biotechnology. *Plant Biotechnology Reports*, 6(1), 1–15.
- Orfali, R., & Perveen, S. (2019). Secondary metabolites from the Aspergillus sp. in the rhizosphere soil of Phoenix dactylifera (Palm tree). *BMC Chemistry*, 13(1), 103. Available from <https://doi.org/10.1186/s13065-019-0624-5>.
- Otto, A., Porzel, A., Schmidt, J., Wessjohann, L., & Arnold, N. (2015). A study on the biosynthesis of hygrophorone B12 in the mushroom *Hygrophorus abieticola* reveals an unexpected labelling pattern in the cyclopentenone moiety. *Phytochemistry*, 118, 174–180.
- Padayachee, T., Khamiakova, T., Louis, E., Adriaensens, P., & Burzykowski, T. (2019). The impact of the method of extracting metabolic signal from 1H-NMR data on the classification of samples: A case study of binning and BATMAN in lung cancer. *PLoS One*, 14(2), e0211854.
- Palmnas, M. S., & Vogel, H. J. (2013). The future of NMR metabolomics in cancer therapy: Towards personalizing treatment and developing targeted drugs? *Metabolites*, 3 (2), 373–396.
- Pan, Q., Mustafa, N. R., Verpoorte, R., & Tang, K. (2016). *13C-isotope-labeling experiments to study metabolism in Catharanthus roseus*. *Metabolomics—Fundamentals and applications*. InTechOpen.
- Park, B. K., Kitteringham, N. R., & O'Neill, P. M. (2001). Metabolism of fluorine-containing drugs. *Annual Review of Pharmacology and Toxicology*, 41(1), 443–470.
- Park, J. M., & Park, J. H. (2001). Human in-vivo 31P MR spectroscopy of benign and malignant breast tumors. *Korean Journal of Radiology*, 2(2), 80.
- Parsons, H. M., Ludwig, C., Günther, U. L., & Viant, M. R. (2007). Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, 8(1), 1–16.
- Pathan, M., Akoka, S., Tea, I., Charrier, B., & Giraudeau, P. (2011). “Multi-scan single shot” quantitative 2D NMR: A valuable alternative to fast conventional quantitative 2D NMR. *Analyst*, 136(15), 3157–3163.
- Pawlowski, P. H., Szczęsny, P., Rempoła, B., Poznańska, A., & Poznański, J. (2019). Combined in silico and 19F NMR analysis of 5-fluorouracil metabolism in yeast at low ATP conditions. *Bioscience Reports*, 39(12).
- Pearson, D., et al. (2019). 19F-NMR-based determination of the absorption, metabolism and excretion of the oral phosphatidylinositol-3-kinase (PI3K) delta inhibitor leniolisib (CDZ173) in healthy volunteers. *Xenobiotica*, 49(8), 953–960. Available from <https://doi.org/10.1080/00498254.2018.1523488>.

- Pendland, S. L., Piscitelli, S. C., Schreckenberger, P. C., & Danziger, L. H. (1994). In vitro activities of metronidazole and its hydroxy metabolite against *Bacteroides* spp. *Antimicrobial Agents and Chemotherapy*, 38(9), 2106–2110.
- Percival, B. C., Grootveld, M., Gibson, M., Osman, Y., Molinari, M., Jafari, F., Sahota, T., Martin, M., Casanova, F., Mather, M. L., Edgar, M., Masania, J., & Wilson, P. B. (2019). Low-field, benchtop NMR spectroscopy as a potential tool for point-of-care diagnostics of metabolic conditions: validation, protocols and computational models. *High Throughput*, 8(1), 2.
- Price, W. S. (1999). Water signal suppression in NMR spectroscopy. *Annual Reports on NMR Spectroscopy*, 38, 289–354.
- Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., & Gautam, B. (2011). The human serum metabolome. *PLoS One*, 6(2), e16957.
- Puchades-Carrasco, L., Palomino-Schätzlein, M., Pérez-Rambla, C., & Pineda-Lucena, A. (2016). Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. *Briefings in Bioinformatics*, 17(3), 541–552.
- Qiao, H., Zhang, X., Zhu, X.-H., Du, F., & Chen, W. (2006). In vivo 31P MRS of human brain at high/ultrahigh fields: A quantitative comparison of NMR detection sensitivity and spectral resolution between 4 T and 7 T. *Magnetic Resonance Imaging*, 24(10), 1281–1286.
- Qiu, X., Redwine, D., Beshah, K., Livazovic, S., Canlas, C. G., Guinov, A., & Emwas, A.-H. M. (2019). Amide vs amine ratio in the discrimination layer of reverse osmosis membrane by solid state 15N NMR and DNP NMR. *Journal of Membrane Science*, 581, 243–251.
- Quinn, R. H. (2012). *Rabbit colony management and related health concerns. The laboratory rabbit, guinea pig, hamster, and other rodents* (pp. 217–241). Elsevier.
- Radjursoga, M., et al. (2018). Nutritional Metabolomics: Postprandial Response of Meals Relating to Vegan, Lacto-Ovo Vegetarian, and Omnivore Diets. *Nutrients*, 10(8). Available from <https://doi.org/10.3390/nu10081063>.
- Raji, M., Amad, M., & Emwas, A. (2013). Dehydrodimerization of pterostilbene during electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, 27(11), 1260–1266.
- Ramirez, B., Durst, M. A., Lavie, A., & Caffrey, M. (2019). NMR-based metabolite studies with 15 N amino acids. *Scientific Reports*, 9(1), 1–5.
- Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., & Luchinat, C. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One*, 10(5), e0124219.
- Reineri, F., et al. (2010). Para-hydrogenated Glucose Derivatives as Potential 13C-Hyperpolarized Probes for Magnetic Resonance Imaging. *Journal of the American Chemical Society*, 132(20), 7186–7193. Available from <https://doi.org/10.1021/ja101399q>.
- Reineri, F., Boi, T., & Aime, S. (2015). ParaHydrogen Induced Polarization of 13C carboxylate resonance in acetate and pyruvate. *Nature Communications*, 6(1), 5858. Available from <https://doi.org/10.1038/ncomms6858>.
- Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 11(6), 1492–1513.

- Renault, S., et al. (2010). Uranyl nitrate inhibits lactate gluconeogenesis in isolated human and mouse renal proximal tubules: A ¹³C-NMR study. *Toxicology and Applied Pharmacology*, 242(1), 9–17. Available from <https://doi.org/10.1016/j.taap.2009.09.002>.
- Renault, M., Cukkemane, A., & Baldus, M. (2010). Solid-state NMR spectroscopy on complex biomolecules. *Angewandte Chemie International Edition*, 49(45), 8346–8357.
- Richardson, P. M., Parrott, A. J., Semenova, O., Nordon, A., Duckett, S. B., & Halse, M. E. (2018). SABRE hyperpolarization enables high-sensitivity ¹H and ¹³C benchtop NMR spectroscopy. *Analyst*, 143(14), 3442–3450.
- Riegel, S. D., & Leskowitz, G. M. (2016). Benchtop NMR spectrometers in academic teaching. *TrAC Trends in Analytical Chemistry*, 83, 27–38.
- Roberts, L. D., Souza, A. L., Gerszten, R. E., & Clish, C. B. (2012). Targeted metabolomics. *Current Protocols in Molecular Biology*, 98(1), 30–32.
- Roberts, M. J., Schirra, H., Lavin, M. F. Martin, F., & Gardiner Robert, A. (2014) NMR-based metabolomics: global analysis of metabolites to address problems in prostate cancer. *Cervical, Breast and Prostate Cancer*. Tokwawan, Kowloon, Hong Kong. iConcept Press. 1–43. Available from <https://espace.library.uq.edu.au/view/UQ:319160>.
- Rocha, C. M., Barros, A. S., Gil, A. M., Goodfellow, B. J., Humpfer, E., Spraul, M., Carreira, I. M., Melo, J. B., Bernardo, J., & Gomes, A. (2010). Metabolic profiling of human lung cancer tissue by ¹H high resolution magic angle spinning (HRMAS) NMR spectroscopy. *Journal of Proteome Research*, 9(1), 319–332.
- Rocha, C. M., Carrola, J., Barros, A. S., Gil, A. M., Goodfellow, B. J., Carreira, I. M., Bernardo, J., Gomes, A., Sousa, V., & Carvalho, L. (2011). Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. *Journal of Proteome Research*, 10(9), 4314–4324.
- Romano, F., Meoni, G., Manavella, V., Baima, G., Mariani, G. M., Cacciatore, S., Tenori, L., & Aimetti, M. (2019). Effect of non-surgical periodontal therapy on salivary metabolic fingerprint of generalized chronic periodontitis using nuclear magnetic resonance spectroscopy. *Archives of Oral Biology*, 97, 208–214.
- Romano, F., Meoni, G., Manavella, V., Baima, G., Tenori, L., Cacciatore, S., & Aimetti, M. (2018). Analysis of salivary phenotypes of generalized aggressive and chronic periodontitis through nuclear magnetic resonance-based metabolomics. *Journal of Periodontology*, 89(12), 1452–1460.
- Romero, J. A., Kazimierczuk, K., & Gołowicz, D. (2020). Enhancing benchtop NMR spectroscopy by means of sample shifting. *Analyst*, 145(22), 7406–7411.
- Rosewell, R., & Vitols, C. (2006). *Identifying metabolites in biofluids*. Edmonton, AB: Chonox Inc.
- Rouger, L., Gouilleux, B., & Nantes, F. P. G. (2017). Fast n-dimensional data acquisition methods. *Laetitia ROUGER*, 23.
- Rubtsov, D. V., Jenkins, H., Ludwig, C., Easton, J., Viant, M. R., Günther, U., Griffin, J. L., & Hardy, N. (2007). Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics: Official Journal of the Metabolomic Society*, 3(3), 223–229.
- Rutar, V. (1984). Suppression of long-range couplings in heteronuclear two-dimensional J spectroscopy. Effects of nonuniform one-bond couplings. *Journal of Magnetic Resonance* (1969), 58(1), 132–142.

- Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 10(3), 361–374.
- Sahoo, N. K., Tejaswini, G., Sahu, M., & Muralikrishna, K. (2020). An overview on NMR spectroscopy based metabolomics. *International Journal of Pharmaceutical Sciences and Developmental Research*, 6(1), 016–020.
- Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., Correa, E., Reijmers, T., Rosato, A., & Tenori, L. (2015). COordination of Standards in MetabOlonicS (COSMOS): Facilitating integrated metabolomics data access. *Metabolomics: Official Journal of the Metabolomic Society*, 11(6), 1587–1597.
- Sanders, J. K. M., & Hunter, B. K. (1993). Modern NMR spectroscopy - a guide for chemists; Oxford University Press: New York.
- Sands, C. J., Coen, M., Maher, A. D., Ebbels, T. M., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2009). Statistical total correlation spectroscopy editing of ^1H NMR spectra of biofluids: Application to drug metabolite profile identification and enhanced information recovery. *Analytical Chemistry*, 81(15), 6458–6466.
- Sarfraz, M. O., Myers, R. P., Coffin, C. S., Gao, Z.-H., Shaheen, A. A. M., Crotty, P. M., Zhang, P., Vogel, H. J., & Weljie, A. M. (2016). A quantitative metabolomics profiling approach for the noninvasive assessment of liver histology in patients with chronic hepatitis C. *Clinical and Translational Medicine*, 5(1), 1–13.
- Satake, M., et al. (2003). Vitamin C Metabolomic Mapping in the Lens with 6-Deoxy-6-fluoro-ascorbic Acid and High-Resolution ^{19}F -NMR Spectroscopy. *Investigative Ophthalmology & Visual Science*, 44(5), 2047–2058. Available from <https://doi.org/10.1167/iovs.02-057>.
- Schilling, F., Warner, L. R., Gershenson, N. I., Skinner, T. E., Sattler, M., & Glaser, S. J. (2014). Next-generation heteronuclear decoupling for high-field biomolecular NMR spectroscopy. *Angewandte Chemie International Edition*, 53(17), 4475–4479.
- Seitz, J. D., et al. (2015). Design, synthesis and application of fluorine-labeled taxoids as ^{19}F NMR probes for the metabolic stability assessment of tumor-targeted drug delivery systems. *Journal of Fluorine Chemistry*, 171, 148–161. Available from <https://doi.org/10.1016/j.jfluchem.2014.08.006>.
- Sekiyama, Y., Chikayama, E., & Kikuchi, J. (2011). Evaluation of a semipolar solvent system as a step toward heteronuclear multidimensional NMR-based metabolomics for ^{13}C -labeled bacteria, plants, and animals. *Analytical Chemistry*, 83(3), 719–726.
- Separovic, F., & Sani, M.-A. (2020). *Solid-state NMR. Applications in biomembrane structure*. IOP Publishing. Available from <https://doi.org/10.1088/978-0-7503-2532-5>.
- Serkova, N. J., & Niemann, C. U. (2006). Pattern recognition and biomarker validation using quantitative ^1H -NMR-based metabolomics. *Expert Review of Molecular Diagnostics*, 6(5), 717–731.
- Sethi, S., Pedrini, M., Rizzo, L. B., Zeni-Graiff, M., Dal Mas, C., Cassinelli, A. C., Noto, M. N., Asevedo, E., Cordeiro, Q., & Pontes, J. G. (2017). ^1H -NMR, ^1H -NMR T 2-edited, and 2D-NMR in bipolar disorder metabolic profiling. *International Journal of Bipolar Disorders*, 5(1), 1–9.
- Shah, P. K., Ye, F., Liu, M., Jayaraman, A., Baligand, C., Walter, G., & Vandeborne, K. (2014). In vivo ^{31}P NMR spectroscopy assessment of skeletal muscle bioenergetics after spinal cord contusion in rats. *European Journal of Applied Physiology*, 114(4), 847–858.

- Shanaiah, N., et al. (2007). Class selection of amino acid metabolites in body fluids using chemical derivatization and their enhanced ^{13}C NMR. *Proceedings of the National Academy of Sciences*, 104(28), 11540–11544. Available from <https://doi.org/10.1073/pnas.0704449104>.
- Sharma, R., Gogna, N., Singh, H., & Dorai, K. (2017). Fast profiling of metabolite mixtures using chemometric analysis of a speeded-up 2D heteronuclear correlation NMR experiment. *RSC Advances*, 7(47), 29860–29870.
- Shchepin, R. V., et al. (2014). Parahydrogen Induced Polarization of 1-13C-Phospholactate-d2 for Biomedical Imaging with >30,000,000-fold NMR Signal Enhancement in Water. *Analytical Chemistry*, 86(12), 5601–5605. Available from <https://doi.org/10.1021/ac500952z>.
- Shchepin, R. V., et al. (2016). 15N Hyperpolarization of Imidazole-15N2 for Magnetic Resonance pH Sensing via SABRE-SHEATH. *ACS Sensors*, 1(6), 640–644. Available from <https://doi.org/10.1021/acssensors.6b00231>.
- Shchepin, R. V., Birchall, J. R., Chukanov, N. V., Kovtunov, K. V., Koptyug, I. V., Theis, T., Warren, W. S., Gelovani, J. G., Goodson, B. M., & Shokouhi, S. (2019). Hyperpolarizing concentrated metronidazole 15NO2 group over six chemical bonds with more than 15% polarization and 20 minute lifetime. *Chemistry (Weinheim an Der Bergstrasse, Germany)*, 25(37), 8829.
- Shchepin, R. V., & Chekmenev, E. Y. (2014). Toward hyperpolarized molecular imaging of HIV: Synthesis and longitudinal relaxation properties of 15N-Azidothymidine. *Journal of Labelled Compounds and Radiopharmaceuticals*, 57(10), 621–624.
- Sheedy, J. R., Ebeling, P. R., Gooley, P. R., & McConville, M. J. (2010). A sample preparation protocol for 1H nuclear magnetic resonance studies of water-soluble metabolites in blood and urine. *Analytical Biochemistry*, 398(2), 263–265.
- Silva, C. L., Olival, A., Perestrelo, R., Silva, P., Tomás, H., & Câmara, J. S. (2019). Untargeted urinary 1H NMR-based metabolomic pattern as a potential platform in breast cancer detection. *Metabolites*, 9(11), 269.
- Silver, J., Naveh-Many, T., & Kronenberg, H. M. (2002). *Parathyroid hormone: Molecular biology*. In *Principles of bone biology* (pp. 407–422). Elsevier.
- Singh, A., et al. (2017). 1H NMR Metabolomics Reveals Association of High Expression of Inositol 1, 4, 5 Trisphosphate Receptor and Metabolites in Breast Cancer Patients. *PLOS ONE*, 12(1), e0169330. Available from <https://doi.org/10.1371/journal.pone.0169330>.
- Smolinska, A., Blanchet, L., Buydens, L. M., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*, 750, 82–97.
- Sokolenko, S., McKay, R., Blondeel, E. J., Lewis, M. J., Chang, D., George, B., & Aucoin, M. G. (2013). Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1 H-NMR spectra in synthetic mixtures and urine with additional insights on choice of pulse sequences and robotic sampling. *Metabolomics: Official Journal of the Metabolomic Society*, 9(4), 887–903.
- Song, C., Jiao, C., Jin, Q., Chen, C., Cai, Y., & Lin, Y. (2020). Metabolomics analysis of nitrogen-containing metabolites between two *Dendrobium* plants. *Physiology and Molecular Biology of Plants*, 26(7), 1425–1435.
- Spicer, R., Salek, R. M., Moreno, P., Cañuelo, D., & Steinbeck, C. (2017). Navigating freely-available software tools for metabolomics analysis. *Metabolomics: Official Journal of the Metabolomic Society*, 13(9), 1–16.

- Straadt, I. K., et al. (2010). Oxidative Stress-Induced Metabolic Changes in Mouse C2C12 Myotubes Studied with High-Resolution ^{13}C , ^1H , and ^{31}P NMR Spectroscopy. *Journal of Agricultural and Food Chemistry*, 58(3), 1918–1926. Available from <https://doi.org/10.1021/jf903505a>.
- Sterin, M., Cohen, J. S., Mardor, Y., Berman, E., & Ringel, I. (2001). Levels of phospholipid metabolites in breast cancer cells treated with antimitotic drugs: A ^{31}P -magnetic resonance spectroscopy study. *Cancer Research*, 61(20), 7536–7543.
- Stilbs, P. (1987). Fourier transform pulsed-gradient spin-echo studies of molecular diffusion. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 19(1), 1–45.
- Stringer, K. A., Puskarich, M. P., Finkel, M. A., Karnovsky, A., & Jones, A. E. (2014). *L-Carnitine treatment impacts amino acid and energy metabolism in sepsis as detected by untargeted ^1H -nuclear magnetic resonance (NMR) pharmacometabolomics. C15. Central nervous system and motor impairment in critical illness* (p. A3932) American Thoracic Society.
- Sykes, B. D. (2007). Urine stability for metabolomic studies: Effects of preparation and storage. *Metabolomics: Official Journal of the Metabolomic Society*, 3(1), 19–27.
- Taggi, A. E., Meinwald, J., & Schroeder, F. C. (2004). A new approach to natural products discovery exemplified by the identification of sulfated nucleosides in spider venom. *Journal of the American Chemical Society*, 126(33), 10364–10369.
- Taglienti, A., Tiberini, A., Ciampa, A., Piscopo, A., Zappia, A., Tomassoli, L., Poiana, M., & Dell'Abate, M. T. (2020). Metabolites response to onion yellow dwarf virus (OYDV) infection in 'Rossa di Tropea' onion during storage: A ^1H HR-MAS NMR study. *Journal of the Science of Food and Agriculture*, 100(8), 3418–3427.
- Takis, P. G., Ghini, V., Tenori, L., Turano, P., & Luchinat, C. (2019). Uniqueness of the NMR approach to metabolomics. *TrAC Trends in Analytical Chemistry*, 120, 115300.
- Takis, P. G., Schäfer, H., Spraul, M., & Luchinat, C. (2017). Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nature Communications*, 8(1), 1–12.
- Tarachiwin, L., Ute, K., Kobayashi, A., & Fukusaki, E. (2007). ^1H NMR based metabolic profiling in the evaluation of Japanese green tea quality. *Journal of Agricultural and Food Chemistry*, 55(23), 9330–9336.
- Tasic, L., Lacerda, A. L. T., Pontes, J. G. M., da Costa, T. B. B. C., Nani, J. V., Martins, L. G., Santos, L. A., Nunes, M. F. Q., Adelino, M. P. M., Pedrini, M., Cordeiro, Q., Bachion de Santana, F., Poppi, R. J., Brietzke, E., & Hayashi, M. A. F. (2019). Peripheral biomarkers allow differential diagnosis between schizophrenia and bipolar disorder. *Journal of Psychiatric Research*, 119, 67–75. Available from <https://doi.org/10.1016/j.jpsychires.2019.09.009>.
- Tasic, L., Pontes, J. G. M., Carvalho, M. S., Cruz, G., Dal Mas, C., Sethi, S., Pedrini, M., Rizzo, L. B., Zeni-Graiff, M., Asevedo, E., Lacerda, A. L. T., Bressan, R. A., Poppi, R. J., Brietzke, E., & Hayashi, M. A. F. (2017). Metabolomics and lipidomics analyses by ^1H nuclear magnetic resonance of schizophrenia patient serum reveal potential peripheral biomarkers for diagnosis. *Schizophrenia Research*, 185, 182–189. Available from <https://doi.org/10.1016/j.schres.2016.12.024>.
- Tavel, L., Fontana, F., Garcia Manteiga, J. M., Mari, S., Mariani, E., Caneva, E., Sitia, R., Camnasio, F., Marcatti, M., & Cenci, S. (2016). Assessing heterogeneity of osteolytic lesions in multiple myeloma by ^1H HR-MAS NMR Metabolomics. *International Journal of Molecular Sciences*, 17(11), 1814.

- Tayyari, F., Gowda, G. N., Gu, H., & Raftery, D. (2013). *15N-cholamine: A smart isotope tag for combining NMR-and MS-based metabolite profiling.* *Analytical Chemistry*, 85(18), 8715–8721.
- Tebben, P. J., Berndt, T. J., & Kumar, R. (2013). *Phosphatonins. Osteoporosis* (pp. 373–390). Elsevier.
- Tenori, L., Turano, P., & Luchinat, C. (2007). Metabolic profiling by NMR. *EMagRes*, 199–204.
- Teunissen, C., Petzold, A., Bennett, J., Berven, F., Brundin, L., Comabella, M., Franciotta, D., Frederiksen, J., Fleming, J., & Furlan, R. (2009). A consensus protocol for the standardization of cerebrospinal fluid collection and biobanking. *Neurology*, 73(22), 1914–1922.
- Thebault, M., Pichavant, K., & Kervarec, N. (2009). *31P nuclear magnetic resonance measurements of phosphate metabolites and intracellular pH in turbot Psetta maxima red blood cells using a novel flow method.* *Journal of Fish Biology*, 75(3), 747–754.
- Theis, T., et al. (2015). Microtesla SABRE enables 10% nitrogen-15 nuclear spin polarization. *Journal of the American Chemical Society*, 137(4), 1404–1407. Available from <https://doi.org/10.1021/ja512242d>.
- Timári, I., Wang, C., Hansen, A. L., Costa dos Santos, G., Yoon, S. O., Bruschweiler-Li, L., & Bruschweiler, R. (2019). Real-time pure shift HSQC NMR for untargeted metabolomics. *Analytical Chemistry*, 91(3), 2304–2311.
- Tiret, B., Brouillet, E., & Valette, J. (2016). Evidence for a “metabolically inactive” inorganic phosphate pool in adenosine triphosphate synthase reaction using localized *31P* saturation transfer magnetic resonance spectroscopy in the rat brain at 11.7 T. *Journal of Cerebral Blood Flow & Metabolism*, 36(9), 1513–1518.
- Tokumaru, O., Kuroki, C., Yoshimura, N., Sakamoto, T., Takei, H., Ogata, K., Kitano, T., Nisimaru, N., & Yokoi, I. (2009). Neuroprotective effects of ethyl pyruvate on brain energy metabolism after ischemia-reperfusion injury: A *31P*-nuclear magnetic resonance study. *Neurochemical Research*, 34(4), 775–785.
- Tomah Al-Masri, H., Emwas, A.-H. M., Al-Talla, Z. A., & Alkordi, M. H. (2012). Synthesis and characterization of new N-(diphenylphosphino)-naphthylamine chalcogenides: X-ray structures of (1-NHC10H7)P(Se)Ph₂ and Ph₂P(S)OP(S)Ph₂. *Null*, 187(9), 1082–1090. Available from <https://doi.org/10.1080/10426507.2012.668985>.
- Tredwell, G. D., Bundy, J. G., De Iorio, M., & Ebbels, T. M. (2016). Modelling the acid/base *1H* NMR chemical shift limits of metabolites in human urine. *Metabolomics: Official Journal of the Metabolomic Society*, 12(10), 1–10.
- Tsujimoto, T., et al. (2018). *13C-NMR-based metabolic fingerprinting of Citrus-type crude drugs.* *Journal of Pharmaceutical and Biomedical Analysis*, 161, 305–312. Available from <https://doi.org/10.1016/j.jpba.2018.08.044>.
- Tynkkynen, T., Tiainen, M., Soininen, P., & Laatikainen, R. (2009). From proton nuclear magnetic resonance spectra to pH. Assessment of *1H* NMR pH indicator compound set for deuterium oxide solutions. *Analytica Chimica Acta*, 648(1), 105–112.
- Uday, S., Höglér, W., Huhtaniemi, I., & Martini, L. (2019). *Rickets and osteomalacia* (pp. 339–354). Academic Press. Available from <https://doi.org/10.1016/B978-0-12-801238-3.65426-0>.
- Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., & Miller, Z. (2007). BioMagResBank. *Nucleic Acids Research*, 36(Suppl. 1), D402–D408.

- van Beek, T. A. (2021). Low-field benchtop NMR spectroscopy: Status and prospects in natural product analysis. *Phytochemical Analysis*, 32(1), 24–37.
- Van, Q. N., Issaq, H. J., Jiang, Q., Li, Q., Muschik, G. M., Waybright, T. J., Lou, H., Dean, M., Uitto, J., & Veenstra, T. D. (2008). Comparison of 1D and 2D NMR spectroscopy for metabolic profiling. *Journal of Proteome Research*, 7(2), 630–639.
- Vassilev, N. G., Simova, S. D., Dangalov, M., Velkova, L., Atanasov, V., Dolashki, A., & Dolashka, P. (2020). An ^1H NMR-and MS-based study of metabolites profiling of garden snail helix aspersa mucus. *Metabolites*, 10(9), 360.
- Vauclare, P., Bligny, R., Gout, E., & Widmer, F. (2013). An overview of the metabolic differences between *Bradyrhizobium japonicum* 110 bacteria and differentiated bacteroids from soybean (*Glycine max*) root nodules: An *in vitro* ^{13}C -and ^{31}P -nuclear magnetic resonance spectroscopy study. *FEMS Microbiology Letters*, 343(1), 49–56.
- Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., Turano, P., & Luchinat, C. (2019). High-throughput metabolomics by 1D NMR. *Angewandte Chemie International Edition*, 58(4), 968–994.
- Vignoli, A., Paciotti, S., Tenori, L., Eusebi, P., Biscetti, L., Chiasserini, D., Scheltens, P., Turano, P., Teunissen, C., & Luchinat, C. (2020). Fingerprinting Alzheimer's disease by ^1H nuclear magnetic resonance spectroscopy of cerebrospinal fluid. *Journal of Proteome Research*, 19(4), 1696–1705.
- Vinci, G., et al. (2018). An alternative to mineral phosphorus fertilizers: The combined effects of *Trichoderma harzianum* and compost on *Zea mays*, as revealed by ^1H NMR and GC-MS metabolomics. *PLOS ONE*, 13(12), e0209664. Available from <https://doi.org/10.1371/journal.pone.0209664>.
- Viola, R., Tucci, A., Timellini, G., & Fantazzini, P. (2006). NMR techniques: A non-destructive analysis to follow microstructural changes induced in ceramics. *Journal of the European Ceramic Society*, 26(15), 3343–3349.
- Vitols, C., & Fu, H. (2006). Targeted profiling of common metabolites in urine. *Edmonton, AB: Chonox Inc.*
- Vitols, C., & Weljie, A. (2006). Identifying and quantifying metabolites in blood serum and plasma. *Chonox Inc.*
- Vitorge, B., Bieri, S., Humam, M., Christen, P., Hostettmann, K., Muñoz, O., Loss, S., & Jeannerat, D. (2009). High-precision heteronuclear 2D NMR experiments using 10-ppm spectral window to resolve carbon overlap. *Chemical Communications*, 8, 950–952.
- Walker, T. G., & Happer, W. (1997). Spin-exchange optical pumping of noble-gas nuclei. *Reviews of Modern Physics*, 69(2), 629.
- Wang, J., Pu, S., Sun, Y., Li, Z., Niu, M., Yan, X., Zhao, Y., Wang, L., Qin, X., & Ma, Z. (2014). Metabolomic profiling of autoimmune hepatitis: The diagnostic utility of nuclear magnetic resonance spectroscopy. *Journal of Proteome Research*, 13(8), 3792–3801.
- Wang, Z. J., Ohliger, M. A., Larson, P. E., Gordon, J. W., Bok, R. A., Slater, J., Villanueva-Meyer, J. E., Hess, C. P., Kurhanewicz, J., & Vigneron, D. B. (2019). Hyperpolarized ^{13}C MRI: State of the art and future directions. *Radiology*, 291(2), 273–284.
- Ward, J. L., Baker, J. M., Miller, S. J., Deborde, C., Maucourt, M., Biais, B., Rolin, D., Moing, A., Moco, S., Vervoort, J., Lommen, A., Schäfer, H., Humpfer, E., &

- Beale, M. H. (2010). An inter-laboratory comparison demonstrates that [1H]-NMR metabolite fingerprinting is a robust technique for collaborative plant metabolomic data collection. *Metabolomics: Official Journal of the Metabolomic Society*, 6(2), 263–273. Available from <https://doi.org/10.1007/s11306-010-0200-4>.
- Watanabe, N., & Niki, E. (1978). Direct-coupling of FT-NMR to high performance liquid chromatography. *Proceedings of the Japan Academy, Series B*, 54(4), 194–199.
- Watts, A. (2005). Solid-state NMR in drug design and discovery for membrane-embedded targets. *Nature Reviews. Drug Discovery*, 4(7), 555–568.
- Webb, A. (2006). Advances in probe design for protein NMR. *Annual Reports on NMR Spectroscopy*, 58, 1–50.
- Wei, L., Liao, P., Wu, H., Li, X., Pei, F., Li, W., & Wu, Y. (2009). Metabolic profiling studies on the toxicological effects of realgar in rats by 1H NMR spectroscopy. *Toxicology and Applied Pharmacology*, 234(3), 314–325.
- Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, 78(13), 4430–4442.
- Weybright, P., Millis, K., Campbell, N., Cory, D. G., Singer, S., (2005). Gradient, high-resolution, magic angle spinning ¹H nuclear magnetic resonance spectroscopy of intact cells. *Magnetic Resonance in Medicine*, 39, 337–345.
- Wijnen, J. P., van der Kemp, W. J., Luttje, M. P., Korteweg, M. A., Luijten, P. R., & Klomp, D. W. (2012). Quantitative 31P magnetic resonance spectroscopy of the human breast at 7 T. *Magnetic Resonance in Medicine*, 68(2), 339–348.
- Willcott, M. R. (2009). *MestRe Nova*. ACS Publications.
- Winter, G., & Krömer, J. O. (2013). Fluxomics—connecting ‘omics analysis and phenotypes. *Environmental Microbiology*, 15(7), 1901–1916.
- Wishart, D. S. (2019). NMR metabolomics: A look ahead. *Journal of Magnetic Resonance*, 306, 155–161.
- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., & Karu, N. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617.
- Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., & Dong, E. (2012). HMDB 3.0—The human metabolome database in 2013. *Nucleic Acids Research*, 41(D1), D801–D807.
- Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., & Bouatra, S. (2009). HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(suppl_1), D603–D610.
- Wishart, D. S., Mandal, R., Stanislaus, A., & Ramirez-Gaona, M. (2016). Cancer metabolomics and the human metabolome database. *Metabolites*, 6(1), 10.
- Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., & Sawhney, S. (2007). HMDB: The human metabolome database. *Nucleic Acids Research*, 35(Suppl 1), D521–D526.
- Wu, R., Chen, J., Zhang, L., Wang, X., Yang, Y., & Ren, X. (2021). LC/MS-based metabolomics to evaluate the milk composition of human, horse, goat and cow from China. *European Food Research and Technology*, 1–13.
- Wüthrich, K. (1986). NMR with proteins and nucleic acids. *Europhysics News*, 17(1), 11–13.

- Yang, W., Wang, Y., Zhou, Q., & Tang, H. (2008). Analysis of human urine metabolites using SPE and NMR spectroscopy. *Science in China Series B: Chemistry*, 51(3), 218–225.
- Ye, T., Mo, H., Shanaiah, N., Gowda, G. N., Zhang, S., & Raftery, D. (2009). Chemoselective ¹⁵N tag for sensitive and high-resolution nuclear magnetic resonance profiling of the carboxyl-containing metabolome. *Analytical Chemistry*, 81(12), 4882–4888.
- Yilmaz, A., Geddes, T., Han, B., Bahado-Singh, R. O., Wilson, G. D., Imam, K., Maddens, M., & Graham, S. F. (2017). Diagnostic biomarkers of Alzheimer's disease as identified in saliva using ¹H NMR-based metabolomics. *Journal of Alzheimer's Disease*, 58 (2), 355–359.
- Zacharias, H. U., Altenbuchinger, M., & Gronwald, W. (2018). Statistical analysis of NMR metabolic fingerprints: Established methods and recent advances. *Metabolites*, 8(3), 47.
- Zacharias, N. M., Chan, H. R., Sailasuta, N., Ross, B. D., & Bhattacharya, P. (2012). Real-time molecular imaging of tricarboxylic acid cycle metabolism in vivo by hyperpolarized ¹–¹³C diethyl succinate. *Journal of the American Chemical Society*, 134(2), 934–943.
- Zacharias, N. M., McCullough, C. R., Wagner, S., Sailasuta, N., Chan, H. R., Lee, Y., Hu, J., Perman, W. H., Henneberg, C., & Ross, B. D. (2016). Towards real-time metabolic profiling of cancer with hyperpolarized succinate. *Journal of Molecular Imaging & Dynamics*, 6(1).
- Zambon, A., et al. (2019). Nucleoside 2',3'-Cyclic Monophosphates in Aphanizomenon flos-aquae Detected through Nuclear Magnetic Resonance and Mass Spectrometry. *Journal of Agricultural and Food Chemistry*, 67(46), 12780–12785. Available from <https://doi.org/10.1021/acs.jafc.9b05991>.
- Zanger, K. (2015). Pure shift NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 86, 1–20.
- Zanger, K., & Sterk, H. (1997). Homonuclear broadband-decoupled NMR spectra. *Journal of Magnetic Resonance*, 124(2), 486–489.
- Zhang, G., Emwas, A.-H., Hameed, U. F. S., Arold, S. T., Yang, P., Chen, A., Xiang, J.-F., & Khashab, N. M. (2020). Shape-induced selective separation of ortho-substituted benzene isomers enabled by cucurbit [7] uril host macrocycles. *Chem*, 6(5), 1082–1096.
- Zhang, M.-H., Jia-Qing, C., Hui-Min, G., Rui-Ting, L., Yi-Qiao, G., Yuan, T., Zhang, Z.-J., & Huang, Y. (2017). Combination of LC/MS and GC/MS based metabolomics to study the hepatotoxic effect of realgar nanoparticles in rats. *Chinese Journal of Natural Medicines*, 15(9), 684–694.
- Zheng, A., Liu, S.-B., & Deng, F. (2017). ³¹P NMR chemical shifts of phosphorus probes as reliable and practical acidity scales for solid and liquid catalysts. *Chemical Reviews*, 117(19), 12475–12531.
- Zheng, H., Chen, M., Lu, S., Zhao, L., Ji, J., & Gao, H. (2017). Metabolic characterization of hepatitis B virus-related liver cirrhosis using NMR-based serum metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*, 13(10), 1–9.

Targeted metabolomics

6

Michele Costanzo^{1,2}, Marianna Caterino^{1,2}, and Margherita Ruoppolo^{1,2}

¹Department of Molecular Medicine and Medical Biotechnology, University of Naples
“Federico II”, Naples, Italy

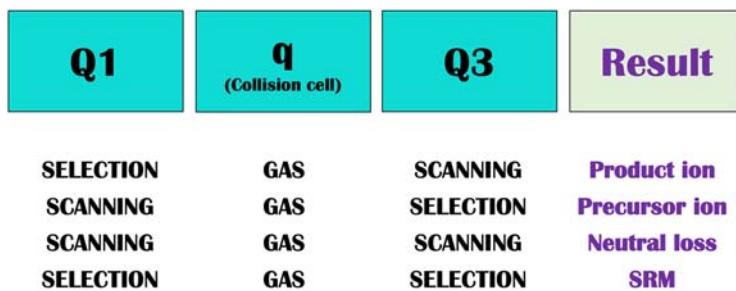
²CEINGE—Biotecnologie Avanzate s.c.ar.l., Naples, Italy

Targeted metabolomics

In the revolutionary context of the omics era, the metabolome is configured as the terminal downstream result of the genome expression and proteome activity. Metabolites concentrations may change very quickly and significantly even following minute perturbations within the cell, influencing the entire phenotype of an organism. Accordingly, considering the complexity of the interactions and the highly intricated metabolic networks that take place in a cell, and the reactivity of the metabolic response to a stimulus, it is not possible to individualize a single technique that allows the comprehensive profiling of the entire metabolome. The profiling of the global metabolome is made more convoluted by the size of the metabolome as the number of metabolites identified, which may range from ~600 in *Saccharomyces cerevisiae* (Förster et al., 2003), to ~110,000 in humans (Wishart et al., 2018) and ~200,000 metabolites in plants (Fiehn, 2001). Also the high dynamic range of metabolite species and classes is a counteracting factor for a comprehensive profiling, showing different physico-chemical properties and varying from small polar volatile to large hydrophobic lipid molecules. In fact, with regards to the latter class, the big complexity of the lipidome (that is defined as the sum of all the lipid species in a sample) is due to the huge number of distinct molecules that are endowed with distinct chemical properties. This, in combination with the mounting evidence for the roles of lipids in transcriptional and translational regulation, cellular signaling, and cell-cell interaction, has led to widen the knowledge of the metabolome, considering the lipidome as a significant portion of it. Thus, even though the lipidome is a subfraction of the metabolome, lipidomics techniques are being separately developed as for “the full characterization of lipid molecular species and of their biological roles with respect to expression of proteins involved in lipid metabolism and function, including gene regulation”. Furthermore, there is a high dynamic range in terms of metabolite concentrations, which can extend throughout large orders of magnitude ranging from femtomolar to millimolar concentrations, influencing

quantitative studies (Roberts et al., 2008; Spener et al., 2003). In the field of metabolomics, two main analytical strategies can be employed to accomplish an experimental design. In particular, untargeted and targeted metabolomic approaches can be configured as two faces of the same coin, allowing the comprehensive and precise metabolic characterization of a biological system. Untargeted metabolomics aims at simultaneously measuring the highest number of metabolites in a biological sample. Being hypothesis-free (or hypothesis-generating), untargeted metabolomics holds the unbiased perspective of omics studies. The main purpose of untargeted metabolomics is the global characterization of the metabolome, in order to compare metabolite patterns that are subjected to qualitative and/or quantitative variation in response to physiological or exogenous stimuli, such as the exposure to drugs or toxic substances, stressing events, genetic alterations, or pathological disturbances, such as the onset or the worsening of a pathology. The characterization of the whole metabolome of a sample, which may lead to the distinction between different metabolomic patterns, is to be considered as the metabolic “fingerprint” of that sample. For these main reasons, untargeted metabolomic approaches are mostly used in the discovery phase of an experiment. On the other hand, a valuable complementation in metabolomics techniques to the metabolic profiling is provided by targeted metabolomics, focusing the analysis on specific subsets of compounds. With the application of a targeted metabolomics approach the coverage of the metabolome is condensed on a defined group of biochemically characterized metabolites, related to a pathway or a class of homogeneous compounds, or specific analytes that can be biomarkers of a pathology or substrates of a precise reaction of the metabolism. In targeted metabolomics experiments, hypothesis-driven experimental schemes impose *a priori* the choice of the metabolites to characterize and/or quantify in a sample. The usefulness of targeted metabolomics can also be appreciated in the experimentations following the discovery phases for metabolites or metabolic pathways validation. Targeted metabolomics strategies have the main advantage to reduce the dynamic range in terms of chemical classes, thus lowering the likelihood of analytical artifacts due to the global metabolome analysis via untargeted approaches. Targeted techniques are sensitive and accurate, being able to diminish the interference created by the selection of high abundance species in the mass spectrometer, in example via the common data-dependent acquisition (DDA) mode that normally does not allow the identification and quantification of very low abundant species. Furthermore, the use of isotopically-labeled standards to be spiked in the sample in known concentrations allows for the absolute quantification of metabolite species. The gold value of targeted metabolomics strategies relies, in fact, on the capability to perform accurate quantitative experiments. Several analytical platforms have been developed for the identification and quantification of specific metabolites. The best analytical technologies for an ideal metabolomic analysis should provide a combination of high sensitivity, high specificity, high resolution, and high throughput. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are the two most popular techniques

used in metabolomics studies, each with advantages and disadvantages. Here, we will focus only on the employment of MS-based techniques in targeted metabolomics studies and their application. MS offers the possibility of conducting quantitative analyses with high specificity and sensitivity in a picomolar and nanomolar dynamic range. In addition, the most commonly used MS-based approaches rely on the combination with a chromatographic separation technique such as gas chromatography (GC), liquid chromatography (LC), or capillary electrophoresis, allowing to reduce the complexity of the metabolites mixture and make the interpretation of the data easier. The disadvantages of MS techniques are related to their destructive nature and the fact that different compounds have different ionization capacities with a consequent bias in the quantitative analysis. Usually, the process of metabolites identification is performed by tandem mass spectrometry (MS/MS) after chromatographic separation or direct injection or flow injection analysis of the sample within the mass spectrometer. The quantitative step is performed downstream metabolite identification and is based on the use of synthetic standards that are spiked in the samples in a known concentration prior the MS analysis, allowing absolute quantification of metabolites. Being their mass different respect to that of the molecules originally present in the sample, synthetic standards (i.e. isotopic heavy-labeled molecules) are distinguishable in the MS spectrum, producing different and non-overlapping fragment ions. A typical LC-MS/MS platform adopted to perform targeted metabolomics analyses is constituted by a LC system coupled with a mass spectrometer carrying an electrospray ionization (ESI) source and a triple quadrupole (QqQ) analyzer. The sample is injected in the chromatographic system and then passes from the LC to the ESI source where the metabolites are transformed into ionic species. Subsequently, these charged molecules called “precursor ions” or “parent ions” are separated in the first quadrupole analyzer (Q1) on the basis of their m/z ratio. Accordingly, the MS/MS system allows the selection of the precursor ions, and induces their fragmentation into the collision cell (commonly called q or Q2) by generating a collision with a neutral gas (nitrogen, argon, or helium). Depending on the collisional energy applied, two main modalities are used: CID (collision-induced dissociation) and HCD (high collision dissociation). Then, the third quadrupole analyzer (Q3) separates the fragments produced, called “product ions” or “daughter ions.” A detector is always positioned downstream the Q3 analyzer in order to detect the mass signals and produce the mass spectra. The main advantage of such LC-MS/MS system can be associated to its very high specificity; in fact, the analyte that undergoes fragmentation will present a fragmentation pattern exclusively characteristic of that metabolite. In more detail, four MS/MS operating modes are available on a QqQ system to perform targeted metabolomic analyses, and they are reported below ([Fig. 6.1](#)). Within these operating modes, Q1 and Q3 can be set to work in a “scanning” mode, in order to scan molecules within a specific mass range, or in a “selection” mode, allowing the only selection of ions having a specific value of m/z .

**FIGURE 6.1**

Triple quadrupole mass spectrometer operative model. Schematic representation of MS/MS operative methods in a QqQ triple quadrupolar mass spectrometer.

- **Product Ion Scan:** allows the identification of the product ions that are obtained by fragmentation of a specific precursor ion. In the Q1, working in the selection mode, a precursor ion is selected. This latter is fragmented in Q2 and, then, all the resultant masses of the product ions are scanned in the Q3 analyzer and detected. The result is a spectrum showing the m/z of the product ions obtained from the selected precursor.
- **Precursor Ion Scan:** aims at identifying all the possible precursor ions that generate a specific product ion by fragmentation; in this case, the Q1 works in scanning mode, while the Q3 in selection mode. The Q1 works by scanning all the precursor ions of a given m/z range. Then, selecting in the Q3 the specific product ions, it is possible to identify the precursor ion that has generated them. This operating mode is useful for the identification of metabolites with similar chemical-physical characteristics that give all the same fragment ion.
- **Neutral Loss Scan:** allows to monitor and identify all the precursor ions that lose a specific neutral fragment during fragmentation. Both the Q1 and the Q3 work in scanning mode, with the Q3 that scans with an offset from the first mass analyzer. This offset corresponds to a neutral loss fragment that is commonly observed within a class of compounds, and only those compounds that give a fragment having that specific loss are detected. In fact, similarly to the precursor ion scan, neutral loss scan is employed for the identification of metabolites belonging to the same class of compounds or with similar chemical-physical characteristics.
- **Single Reaction Monitoring (SRM):** in the case of SRM experiments, both the Q1 and the Q3 mass analyzers are set in the selection mode. In particular, the Q1 selects a specific precursor ion of a particular mass and the Q3 selects a specific product ion derived from the fragmentation in the Q2. The reaction of fragmentation that from a precursor ion generates a product ion (the precursor and product ion pair) is called SRM “transition” ([Lange et al., 2008](#)). **Multiple reaction monitoring (MRM)** is an advanced application of

SRM in which multiple product ions from one or more precursor ions are selected and detected. Parallel reaction monitoring (PRM) is an additional application of SRM in which the parallel detection of all transitions is performed in a single analysis using a high resolution mass spectrometer (Peterson et al., 2012). Being such experiments extremely selective, SRM and its applications are used for the identification and the quantitative analysis of specific analytes.

Generally, the data derived from a targeted metabolomics experiment can be used as validation of previous discovery experiments, helping building or implementing databases that can be integrated with information retrieved from genomics, transcriptomics or proteomics experiments. With its ability of identifying and quantifying selected classes of compounds, targeted metabolomics is the main strategy used for the in-depth investigation of altered metabolic pathways, enlarging the knowledge about the functional role of the metabolites in such pathways. Being the metabolome the biological endpoint of the genome, metabolomics has an intrinsic advantage with respect to the other omics sciences. In fact, it gives access to the quickest response of an organism to a stimulus (endogenous or exogenous) providing precious information regarding the real-time response of enzymes activity, thus allowing to monitor the metabolic status of an individual. For its nature, metabolomics encompasses the comprehensive identification of the regulatory metabolic networks that may result unbalanced in a particular status, reducing the gap between genotype and phenotype, especially in disease status. In fact, differential metabolomics performed by targeted approaches is regularly involved in research plans to achieve a comprehensive characterization of large panels of biochemical metabolites that may qualitatively and/or quantitatively change in one or more compared conditions. The analysis of the differential metabolome may reveal complex metabolite-phenotype relationships that are not easily deducible managing such complex big data. Particularly, the complexity of high throughput metabolomics data can be virtually reduced by applying data analysis and statistical techniques adopted from other omics sciences. Accordingly, univariate statistics (i.e. *t*-test or ANOVA) are normally employed to display the strongest response of the quantitative difference of the measured metabolites from several conditions. Unfortunately, univariate techniques fail to discriminate among groups because its major application is for summarizing only one variable at the time. Actually, when the analysis involves the observation of more than one variable, the application of multivariate statistics is more appropriated. Multivariate methods are able to examine the concurrent effect of multiple variables, defining the correlation between the most important metabolite molecules. The most used techniques of multivariate analysis are the principal component analysis (PCA) and the partial least squares regression comprising the discriminant analysis (PLS-DA). From a supervised analysis it is possible to find the set of important variables on the projection (VIP). These VIP can be considered the main metabolites responsible for separation among the groups (Noto et al., 2016).

Before to find application in the medical fields especially for diagnostic purposes, metabolomics techniques have grown over the time reaching high technical standards in terms of sensitivity, specificity, and accuracy. This has been possible through the advancement in the MS platforms, which are able nowadays to offer the highest throughput possible, reducing the costs and the time of analysis, and improving the number of metabolites detectable in a single-run experiment. However, very rigorous experimentations and studies have been conducted so far in order to let physicians and scientists use targeted MS/MS-based techniques for diagnostics reasons. This is of great value and importance especially in the medical field, for drug discovery or disease biomarker identification, where one or more metabolites together can characterize the diagnostic signature of a disease and can be identified and quantified in order to perform diagnosis. The analysis of amino acids and acylcarnitines in different biological matrices such as cells, tissues or biological fluids is an example of such kind of experiments in the targeted metabolomics field ([Caterino et al., 2021](#); [De Pasquale et al., 2020](#); [Giacco et al., 2019](#); [Ruoppolo et al., 2018](#)). Currently, targeted metabolomic approaches are being employed daily to investigate the quantitative variations of metabolites in every field of science. Without any doubt, the greatest success of targeted metabolomics is represented by the newborn screening (NBS) of genetic disorders such as the inborn error of metabolism (IEM), being able to detect and monitor biomarker metabolites in patients' blood samples. This is possible thanks to the national NBS programs that allow the early diagnosis in newborns since the first days of their life. The early diagnosis is extremely important because permits pediatricians to immediately apply the therapeutic plans available, in order to avoid the irreversible consequence of a genetic metabolic disorder ([Scolamiero et al., 2015](#); [Villani et al., 2017](#)). In the following sections, an overview on the classification of IEM and the application of targeted metabolomics to NBS is given.

Inborn errors of metabolism

For the first time, Sir Archibald E. Garrod used the term “inborn errors of metabolism” in 1908 to describe four genetically determined diseases characterizing some patients, namely alkapturia, albinism, cystinuria, and pentosuria. Garrod's main hypothesis behind the first definition of IEM foresaw that the defect or the block of a specific metabolic pathway were triggered by the reduced activity or complete lack of a given enzyme. He also suggested that the expression and functioning of each enzyme were controlled by a single gene but, being still unknown at that time that enzymes were protein species, the hypotheses of this scientist remained underrated for several decades. Thus, the discovery and the definition of numerous other IEM grew very slowly over the time. This was not helped by the consideration of these diseases as

extremely rare and, thus, not “clinically relevant.” Despite this, rare disorders are nowadays clinically considered and studied, and the mortality of patients and their life expectancy have been significantly ameliorated (Mussap et al., 2018). The amount of known IEM has extensively grown and, to date, it counts around 1000 different genetically and biochemically well characterized disorders. Around 100 additional less characterized disorders are considered borderline because not meeting strict criteria for being included in the current classification of IEM (Ferreira et al., 2019). If taken singularly, the IEM that individually affects a person is extremely rare (i.e. the incidence of methylmalonic acidemia is around 1:100,000 in Europe), but all the group of IEM presents an overall incidence of 1:1000 individuals. The majority of IEM consist in defects of enzymes or transport proteins, which cause accumulation of a toxic substrate proximal to the metabolic block, a lack of the product of the reaction, or a deviation of the substrate toward an alternative metabolic pathway. Since IEM are characterized by a broad heterogeneity, they can be cataloged considering many different criteria that include the clinical phenotype, the pathophysiological mechanism, the onset, the enzyme or the metabolic way impaired or the system involved, etc. Despite a single universal classification is not realizable, the Society for the Study of Inborn Errors of Metabolism has categorized the IEM according to the principle of homogeneity (Zschocke, 2014). Each category includes disorders that are part of the same biochemical pathway, identified using the same diagnostics technique, and monitored following similar approved procedures and protocols for the treatment of patients. Actually, a more recent classification based on clinical features allocates IEM into two big classes:

1. IEM that affect only one functional system or one organ;
2. IEM that affect one metabolic pathway common to a large number of cells or organs, or IEM that are limited to one organ but inducing systemic disturbance.

Additionally, this latter group has been subdivided into the following three distinct classes:

- disorders of intermediary metabolism affecting small molecules;
- disorders involving primarily energy metabolism;
- disorders involving complex molecules. Table 6.1 (Saudubray & Garcia-Cazorla, 2018).

Application of targeted metabolomics to the newborn screening of inborn errors of metabolism

NBS is a routinely service managed by the national public health system that screens each neonate for IEM by testing panels of metabolites extracted from the

Table 6.1 Inborn errors of metabolism.

Disorder class	Inherited disorders
Disorders of intermediary metabolism affecting small molecules	Aminoacidopathies; Organic acidurias/acidemias; Urea cycle defects; Galactose and fructose defects; Metal disorders; Neurotransmitter synthesis; Porphyrias
Disorders involving primarily energy metabolism	Mitochondrial defects: Fatty acid oxidation disorders, Glucose oxidation defects, Respiratory chain disorders; Cytoplasmic defects: Glycolysis defects, Glycogen metabolism and gluconeogenesis defects; Glucose transporter defects
Disorders involving complex molecules	Lysosomal storage disorders; Peroxisomal disorders; Carbohydrate-deficient glycoprotein syndrome; Purine and pyrimidine metabolism defects; Cholesterol and bile acid synthesis defects; Intracellular triglycerides, phospholipids, glycosphingolipids synthesis and remodeling defects

Classification of inborn errors of metabolism of the category 2 according to [Saudubray and Garcia-Cazorla \(2018\)](#).

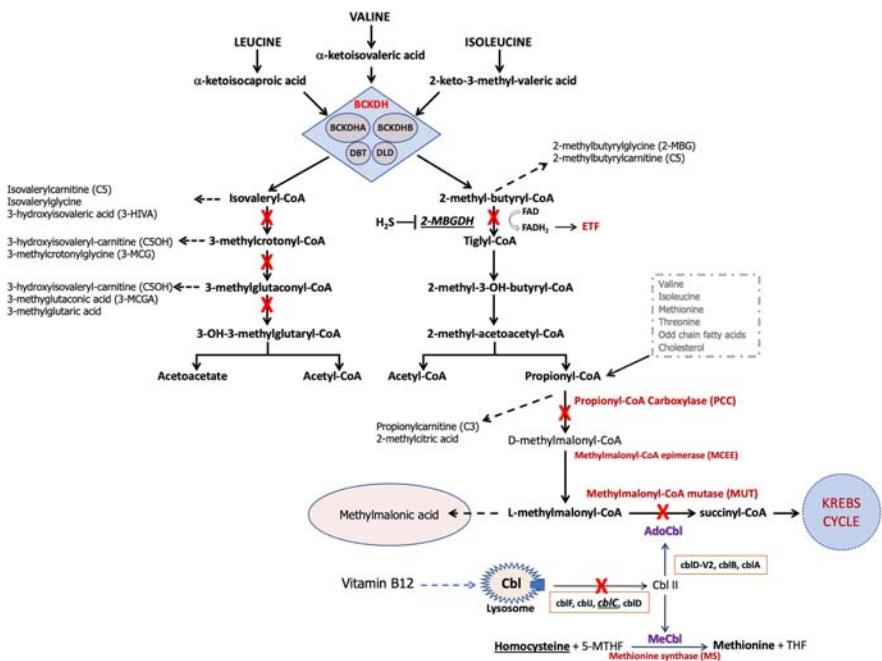
dried blood spots (DBS). DBS are collected within 48–72 hours of birth by picking blood drops from newborns' heels and sent to NBS laboratories. Since metabolic disorders are due to defects of enzymes or transport proteins that cause accumulation or lack of particular metabolites, the profiles and the variations in the metabolome of the biological fluids can reflect perturbations in the metabolic status of patients. Diagnostic delay of many IEM may have a poor outcome for the patient, resulting in acute metabolic decompensation, progressive neurologic damage, or death. Actually, timely NBS, coupled with genetic analysis, is pivotal for the accurate recognition and early diagnosis of IEM, allowing presymptomatic treatment of the patients ([Costanzo et al., 2017](#); [Ferreira et al., 2019](#); [Mussap et al., 2018](#); [Saudubray & Garcia-Cazorla, 2018](#); [Scolamiero et al., 2015](#); [Villani et al., 2017](#); [Zschocke, 2014](#)). Since 1980, MS/MS-based platforms were employed as reference methodologies for NBS, because of the ability of MS/MS to screen a large variety of IEM, which were previously unscreened, by a single test performed on a DBS. Currently, MS/MS techniques are routinely used for the neonatal screening of IEM, since they are fast, high-sensitive and specific, requiring very low sample volume and providing a high throughput. The current diagnostic toolbox for NBS laboratories includes a panel of targeted analyses based on a variety of MS-based instrumentations that include LC-MS/MS and GC-MS systems. Since their introduction in the public health care, only few metabolic disorders, those with the highest incidence in the population, were screened.

Nowadays, the NBS programs in many countries have been enhanced (called expanded NBS) in order to detect more than 50 inherited metabolic disorders, comprising aminoacidopathies, organic acidemias, fatty acid oxidation disorders, and lysosomal storage disorders. The diagnosis of these IEM is based on the metabolomic profiling via targeted MS/MS analysis of plasma amino acids, acylcarnitines, carbohydrates and urinary organic acids. The definitive diagnosis is obtained by complementation with enzyme activity assays, functional assays, and genetic mutational analyses (Coene et al., 2018; Costanzo et al., 2017; Ferreira et al., 2019; Mussap et al., 2018; Saudubray & Garcia-Cazorla, 2018; Zschocke, 2014). In the following section, many examples of the application of targeted metabolomics techniques to the diagnosis of IEM through the NBS programs are reported, with particular focus on some organic acidemias and aminoacidopathies, and their corresponding biomarkers Table 6.2. A schematic view of the metabolic pathways that link the amino acid metabolism and organic acidemias is reported in Fig. 6.2.

Table 6.2 Inborn errors of metabolism markers.

Inborn error of metabolism	Primary markers	Secondary makers	Ratios
MMA Mut	C3	C4DC, Gly	C3/C2, C3/C16
MMA CblC	C3	C4DC, Gly	C3/C2, C3/C16
Propionic acidemia	C3	Gly	C3/C2, C3/C16
Glutaric acidemia type I	C5DC		
Glutaric acidemia type II	C4, C5, C6, C8, C10, C14, C16, C18		
Isovaleric acidemia	C5		C5/C3, C5/C4
Phenylketonuria	Phe		Phe/Tyr
Tyrosinemia type I	Succinylacetone		
Tyrosinemia type II	Tyr		
Tyrosinemia type III	Tyr		
MSUD	Xle		

Primary and secondary markers of inborn error of metabolism and their ratios in diseases diagnosed in the newborn screening program. C2 = acetylcarnitine; C3 = propionylcarnitine; C4 = butyrylcarnitine; C5 = iso-/valerylcarnitine/2-methylbutyrylcarnitine; C6 = hexanoylcarnitine; C8 = octanoylcarnitine; C10 = decanoylcarnitine; C14 = tetradecanoylcarnitine; C16 = palmitoylcarnitine; C18 = stearoylcarnitine; C4DC = methylmalonylcarnitine; C5DC = glutarylcaritine; Gly = glycine; Phe = phenylalanine; Tyr = tyrosine; Xle = isoleucine, leucine, alloisoleucine, or OH-proline.

**FIGURE 6.2**

Amino acid metabolism pathways. Schematic view of the metabolic pathways that involve amino acid metabolism and organic acidemias. The red crosses indicate the metabolic blocks that are commonly causative of most of the IEM described in this chapter.

Examples of inborn error of metabolism diagnosed by the newborn screening

Methylmalonic acidemias

In the group of organic acidemias, methylmalonic acidemias (MMA) are classified as heterogeneous autosomal recessive inherited disorders caused by defects in propionyl-CoA catabolism. MMA are rare IEM that cause severe neurometabolic impairment and pleiotropic dysfunctions and can be detected in newborns through the NBS programs. Amongst these, isolated MMA is caused by deficiency of the enzyme methylmalonyl-CoA mutase (MUT), which can be characterized by reduced (mut^- phenotype) or absent (mut^0 phenotype) enzymatic activity. MUT is a key enzyme involved in the catabolism of the branched-chain amino acids (BCAA) leucine, isoleucine, and valine, as well as methionine and threonine, odd-chain fatty acids and the side chain of the cholesterol, by catalyzing the biochemical conversion of methylmalonyl-CoA to succinyl-CoA as substrate for the Krebs cycle (Costanzo et al., 2018, 2020). Other forms of MMA, defined as combined MMA with homocystinuria, are due to defects in

the transport or synthesis of MUT cofactor (adenosylcobalamin) or methionine synthase cofactor (methylcobalamin), involving mutations in the complementation genes *cblA*, *cblB*, *cblC*, *cblD*, *cblE*, *cblF*, *cblG*, *cblH*, and *cblJ*. An additional extremely rare form of MMA can also present deficiency of methylmalonyl-CoA epimerase. The biochemical hallmarks of MMA comprise high levels of propionylcarnitine (C3) in DBS that can be quantified by targeted metabolomics. Collaborative projects on NBS have highlighted the use of secondary markers (C4DC and Gly) and ratios (C3/C2 and C3/C16) in the identification of the diseases as reported in [Table 6.2](#). On the DBS, second-tier tests ([Yıldız et al., 2019](#)) are performed to measure methylmalonic acid, 3-OH-propionic acid, propionylglycine, methylcitric acid, and homocysteine. The second-tier tests result in an increase of methylmalonic acid for isolated MMA. On the other hand, increased levels of methylmalonic acid and homocysteine are detectable in patients affected by combined MMA with homocystinuria. Confirmatory biochemical tests for the diagnosis employ targeted LC-MS/MS or untargeted GC-MS to quantify methylmalonic acid, 2-methylcitric acid and 3-OH-propionic acid in urine samples.

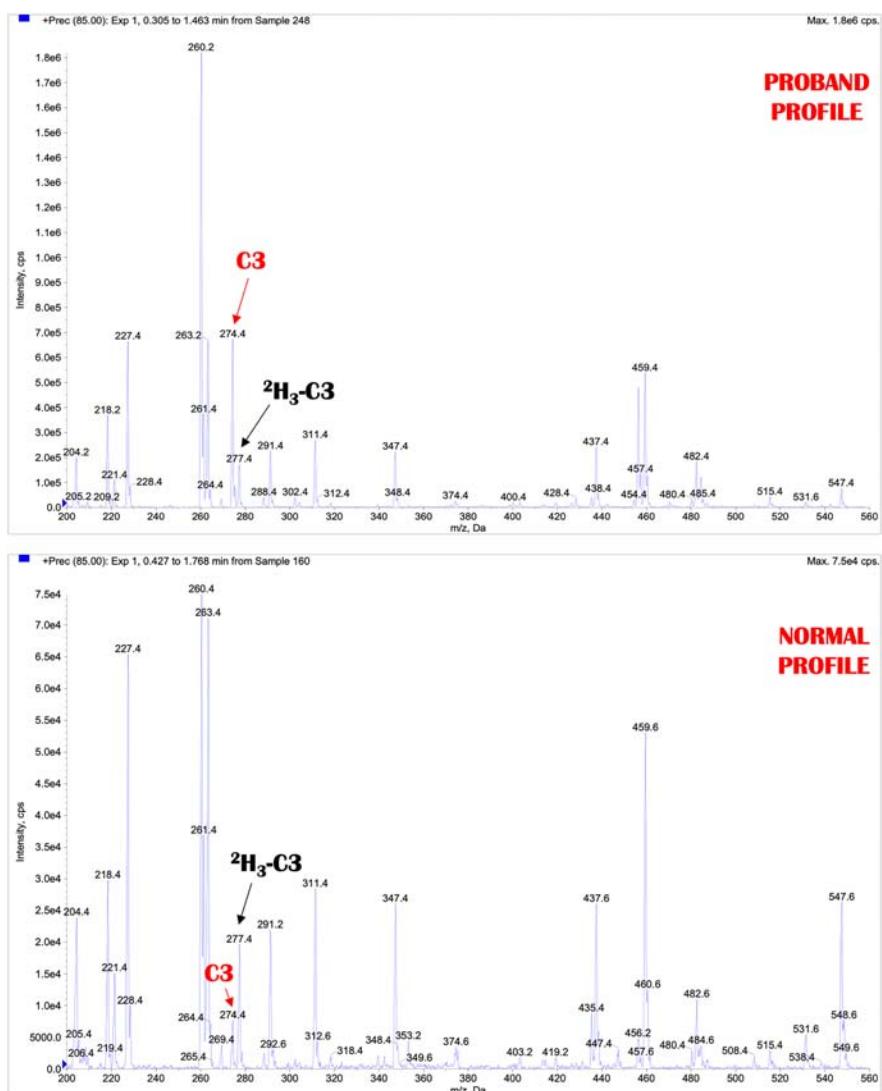
Examples of the identification of metabolites biomarkers for MMA performed on DBS of real newborn patients are reported in [Figs. 6.3](#) and [6.4](#). Respectively, the figures show a MS spectrum relative to acylcarnitines profiling and a chromatogram that reports the presence of a peak for the methylmalonic acid. According to NBS reference ranges, a proband (suspected/positive) patient profile and a normal (healthy) patient profile are reported in both figures.

Propionic acidemia

Propionic acidemia (PA) is an autosomal recessive IEM caused by deficiency of propionyl-CoA carboxylase, a mitochondrial biotin-dependent enzyme that catalyzes the reaction upstream that of MUT enzyme, by converting the propionyl-CoA into methylmalonyl-CoA. The metabolomic characterization of PA patients includes accumulation of C3 as main biomarker in DBS with secondary marker (Gly) and ratio (C3/C2 and C3/C16) as reported in [Table 6.2](#). On the DBS, a second-tier test is performed to measure methylmalonic acid, 3-OH-propionic acid, propionylglycine, methylcitric acid and homocysteine. The second-tier test results in an increase of 3-OH-propionic acid, propionylglycine and/or methylcitric acid for PA. Confirmatory biochemical tests for the diagnosis employ targeted LC-MS/MS or untargeted GC-MS to quantify 3-OH-propionic acid, methylcitric acid, tiglylglycine, propionylglycine, and 2-methylbutyrylglycine in urine samples ([Villani et al., 2017](#)).

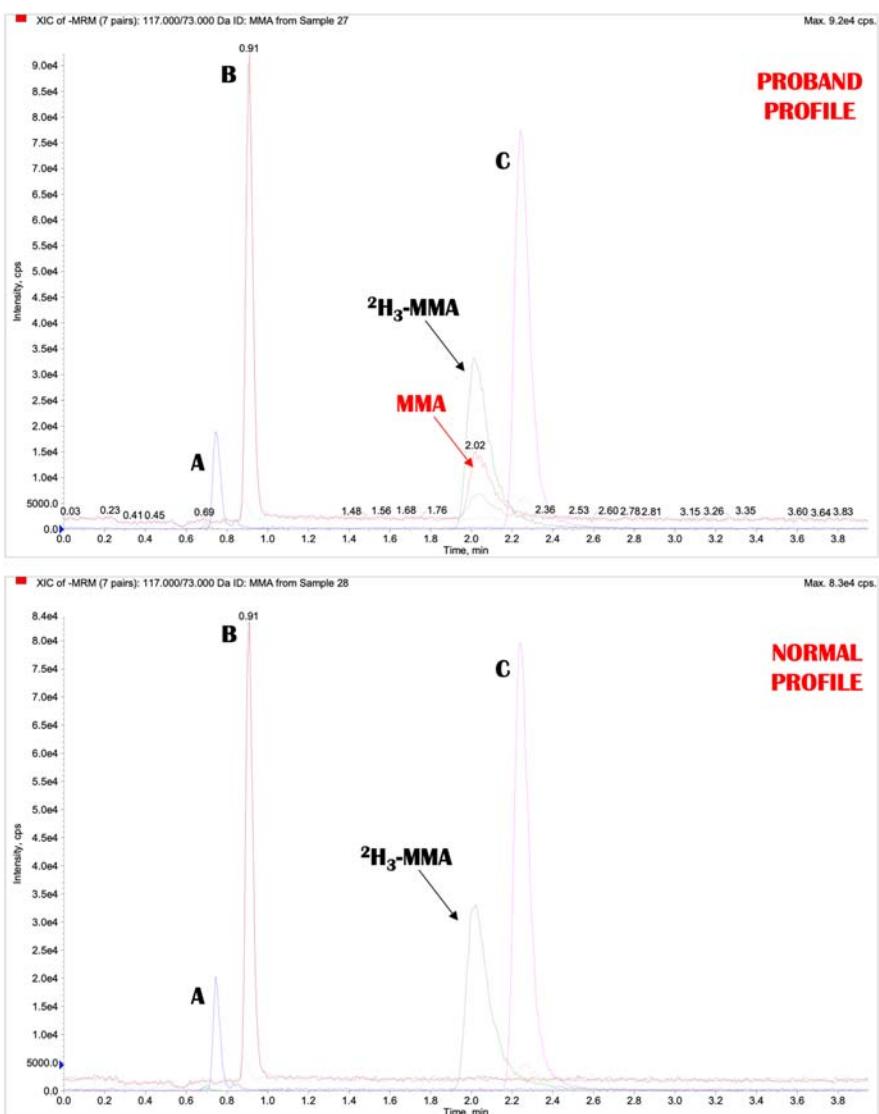
Glutaric acidemia

Glutaric acidemias are autosomal recessive IEM with severe neurometabolic impairment. Mutations causative of glutaric acidemia type 1 (GA-I) fall in the

**FIGURE 6.3**

Metabolic profiling of acylcarnitines obtained by Precursor Ion Scan on DBS from two patients. C3 refers to propionylcarnitine, ${}^2\text{H}_3\text{-C3}$ is a stable-isotope heavy-labeled standard used for the quantification of C3. The proband profile is suspected for a form of methylmalonic acidemia.

glutaryl-CoA dehydrogenase (*GCDH*) gene, which is involved in the degradation of the amino acids lysine, hydroxylysine, and tryptophan through the catalysis of the oxidative decarboxylation of glutaryl-CoA to crotanyl-CoA and carbon

**FIGURE 6.4**

Quantification of methylmalonic acid (MMA) in the DBS from two patients performed by Multiple Reaction Monitoring (MRM) analysis. $^{2\text{H}_3}\text{-MMA}$ is a stable-isotope heavy-labeled standard used for the quantification of MMA. The proband profile is suspected for a form of methylmalonic aciduria. The normal profile does not show presence of a MMA peak, which is mostly undetectable in healthy newborns. In addition, the current MRM method allows quantification of the following metabolites, reported as A = lactic acid, B = succinic acid, C = $^{2\text{H}_3}\text{-ethylmalonic acid}$ internal standard.

dioxide. Targeted metabolomics in the DBS is able to perform diagnosis of GA-I in the NBS program through the detection of glutarylcarnitine (C5DC) as the primary biomarker of this disorder. As effect of deficiency in *GCDH* gene, the metabolites glutaric acid and 3-hydroxyglutaric acid accumulate in body fluids, especially in urine. On the other hand, a second defect known as glutaric aciduria type 2 (GA-II) can be diagnosed in newborns. GA-II is caused by mutations in the genes codifying for the electron transfer flavoprotein (ETF) subunit alpha and beta (*ETFA*, *ETFB*) and ETF-dehydrogenase (*ETFDH*). The disorder is also known as multiple acyl-CoA dehydrogenation deficiency, because dysfunctions of either ETF or ETFDH flavoproteins lead to the compromission of the fatty acid oxidation and amino acid degradation, with alteration of the processes of energy production as well as the synthesis of energy molecules and their storage. Metabolomic profiling of GA-II patients include increased levels of C4 accompanied by elevated C5, C6, C8, C10, C14, C16, and C18 acylcarnitines. GA-II biochemical presentation is accompanied by elevated levels of glutaric acid in urine (Jacob et al., 2018; Yildiz et al., 2019).

Isovaleric acidemia

Isovaleric acidemia (IVA) has been the first organic aciduria reported in humans. IVA is an autosomal recessive disorder caused by deficiency of the mitochondrial flavoenzyme isovaleryl-CoA dehydrogenase (IVD) that participates in the third step of leucine catabolism, the dehydrogenation of isovaleryl-CoA with formation of 3-methylcrotonyl-CoA. When IVD is deficient, isovaleryl-CoA cannot be oxidized and accumulates in blood and urine, being conjugated with many compounds. Particularly, the diagnosis of IVA can be made in newborns by finding elevations of the biomarker metabolite C5 in DBS together with ratios C5/C3 and C5/C4 (Table 6.2). Since C5 can include a mixture of isomers (isovalerylcarnitine, 2-methylbutyrylcarnitine, and pivaloylcarnitine), second-tier tests are mandatory. Increased levels of 3-OH-isovaleric acid and isovalerylglycine in urine measured by GC-MS are used a confirmatory biochemical test for the diagnosis of IVA (Vockley & Ensenauer, 2006). Moreover, since isovaleryl-CoA inhibits the N-acetylglutamate synthetase, there is a reduced synthesis of N-acetylglutamate with impairment of the urea cycle and, finally, hyperammonemia (Villani et al., 2017).

Phenylketonuria

Phenylketonuria (PKU) is one of the most frequent IEM in Europe, showing high variability according to the country, with an incidence that ranges from 1:850 in (Russia) to 1:112,000 live births in Finland (Hillert et al., 2020). PKU is caused

by defects of the hepatic enzyme phenylalanine hydroxylase (PAH), which through the breakdown of phenylalanine (Phe) produces tyrosine (Tyr), required for the synthesis of neurotransmitters like epinephrine, norepinephrine, and dopamine. Thus, the result of lacking PAH is the accumulation of Phe and excretion of its catabolites (phenylpyruvic acid, phenyllactic acid, and phenylacetic acid). In the screening for PKU, the phenylalanine/tyrosine ratio has a greater clinical sensitivity and specificity than Phe concentration alone. This is one of the most informative ratio since, in condition of phenylalanine hydroxylase deficiency, phenylalanine is accumulated and tyrosine, the downstream product of the block, is decreased ([Table 6.2](#)).

Hereditary tyrosinemas

Hereditary tyrosinemas are IEM caused by impaired breakdown of the amino acid Tyr, leading to its accumulation in blood as the main hallmark of both type I and type II tyrosinemas (TYR-I, TYR-II). Type I tyrosinemia results from deficiency of fumarylacetoacetate hydrolase, the last enzyme involved in the Phe/Tyr degradation pathway. As consequence, fumarylacetoacetate accumulates producing succinylacetone, detected in DBS as the main biomarker of the TYR-I. On the other hand, TYR-II is due to deficiency of tyrosine transaminase that is involved in the conversion of Tyr to 4-hydroxyphenylpyruvic acid. This disorder is characterized by high levels of Tyr detected on dried blood spots. The compounds 4-hydroxyphenylpyruvic acid, 4-hydroxyphenyllactic acid, and 4-hydroxyphenylacetic acid detected in urine by GC-MS are used as confirmatory biochemical test ([Peña-Quintana et al., 2017](#)). Finally, a type III tyrosinemia is configured as the rarest of the three conditions, with only a few cases ever reported, and resulting from a mutation in the 4-hydroxyphenylpyruvate dioxygenase gene ([Heylen et al., 2012](#)). This disease is identified by an increase of Tyr on DBS. Genetic tests may discriminate between the different forms.

Maple syrup urine disease

Maple syrup urine disease (MSUD) is included in the group of aminoacidopathies and results from recessive mutations in one of the lipoic acid-dependent enzymes, namely branched-chain keto acid dehydrogenase-E1 alpha polypeptide, branched-chain keto acid dehydrogenase-E2 beta polypeptide, dihydrolipoamide branched-chain transacylase, and dihydrolipoamide dehydrogenase. The protein complex formed by these four enzymes is crucial for the catabolism of the BCAA. Mutations in one of these genes prevent the breakdown of BCAA, leading to accumulation of valine, leucine, isoleucine, alloisoleucine, and their corresponding branched-chain ketoacids (BCKA) in body fluids, including α -ketoisocaproic

acid, α -keto- β -methylisovaleric acid, and α -ketoisovaleric acid. Expanded NBS by targeted metabolomics is able to detect several isobaric species of different amino acids (isoleucine, leucine, alloisoleucine, and OH-proline) combined in one analytical signal named “X-Leu” (Xle). Second-tier test on DBS is necessary to quantify alloisoleucine, a pathognomonic marker of MSUD disease. In healthy newborns alloisoleucine is undetectable, thus affected babies can be clearly distinguished. Furthermore, the profiling of urinary organic acid analysis to identify BCKA also delivers supporting evidence for the diagnosis of MSUD (Blackburn et al., 2017).

Conclusion

In the last years, targeted metabolomic approaches based on LC-MS/MS methodologies have been universally established as the gold-standard technique for the analysis of DBS (and other body fluids) in NBS programs, for both diagnosis or follow-up procedures. Nowadays, with the improvement of targeted metabolomics and the application of several targeted techniques, many laboratories dispose of multiple panel assays for the detection, identification, and quantification of amino acids, acylcarnitines, carbohydrates, organic acids, and other metabolites. With the improvement of LC-MS/MS platforms, these panels may include a larger number of analytes, to screen the highest number of IEM possible in one single analysis.

References

- Blackburn, P. R., Gass, J. M., e Vairo, F. P., Farnham, K. M., Atwal, H. K., Macklin, S., Klee, E. W., & Atwal, P. S. (2017). Maple syrup urine disease: Mechanisms and management. *The Application of Clinical Genetics*, 10, 57.
- Caterino, M., Gelzo, M., Sol, S., Fedele, R., Annunziata, A., Calabrese, C., Fiorentino, G., D'Abbraccio, M., Dell'Isola, C., & Fusco, F. M. (2021). Dysregulation of lipid metabolism and pathological inflammation in patients with COVID-19. *Scientific Reports*, 11 (1), 1–10.
- Coene, K. L., Kluijtmans, L. A., van der Heeft, E., Engelke, U. F., de Boer, S., Hoegen, B., Kwast, H. J., van de Vorst, M., Huigen, M. C., & Keularts, I. M. (2018). Next-generation metabolic screening: Targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *Journal of Inherited Metabolic Disease*, 41(3), 337–353.
- Costanzo, M., Caterino, M., Cevenini, A., Jung, V., Chhuon, C., Lipecka, J., Fedele, R., Guerrera, I. C., & Ruoppolo, M. (2020). Proteomics reveals that methylmalonyl-CoA mutase modulates cell architecture and increases susceptibility to stress. *International Journal of Molecular Sciences*, 21(14), 4998.
- Costanzo, M., Cevenini, A., Marchese, E., Imperlini, E., Raia, M., Del Vecchio, L., Caterino, M., & Ruoppolo, M. (2018). Label-free quantitative proteomics in a

- methylmalonyl-CoA mutase-silenced neuroblastoma cell line. *International Journal of Molecular Sciences*, 19(11), 3580.
- Costanzo, M., Zacchia, M., Bruno, G., Crisci, D., Caterino, M., & Ruoppolo, M. (2017). Integration of proteomics and metabolomics in exploring genetic and rare metabolic diseases. *Kidney Diseases*, 3(2), 66–77.
- De Pasquale, V., Caterino, M., Costanzo, M., Fedele, R., Ruoppolo, M., & Pavone, L. M. (2020). Targeted metabolomic analysis of a mucopolysaccharidosis IIIB mouse model reveals an imbalance of branched-chain amino acid and fatty acid metabolism. *International Journal of Molecular Sciences*, 21(12), 4211.
- Ferreira, C. R., van Karnebeek, C. D., Vockley, J., & Blau, N. (2019). A proposed nosology of inborn errors of metabolism. *Genetics in Medicine*, 21(1), 102–106.
- Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2(3), 155–168.
- Förster, J., Famili, I., Fu, P., Palsson, B. Ø., & Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2), 244–253.
- Giacco, A., Delli Paoli, G., Senese, R., Cioffi, F., Silvestri, E., Moreno, M., Ruoppolo, M., Caterino, M., Costanzo, M., & Lombardi, A. (2019). The saturation degree of fatty acids and their derived acylcarnitines determines the direct effect of metabolically active thyroid hormones on insulin sensitivity in skeletal muscle cells. *The FASEB Journal*, 33(2), 1811–1823.
- Heylen, E., Scherer, G., Vincent, M.-F., Marie, S., Fischer, J., & Nassogne, M.-C. (2012). Tyrosinemia Type III detected via neonatal screening: Management and outcome. *Molecular Genetics and Metabolism*, 107(3), 605–607.
- Hillert, A., Anikster, Y., Belanger-Quintana, A., Burlina, A., Burton, B. K., Carducci, C., Chiesa, A. E., Christodoulou, J., Đorđević, M., & Desviat, L. R. (2020). The genetic landscape and epidemiology of phenylketonuria. *The American Journal of Human Genetics*, 107(2), 234–250.
- Jacob, M., Malkawi, A., Albast, N., Al Bougha, S., Lopata, A., Dasouki, M., & Rahman, A. M. A. (2018). A targeted metabolomics approach for clinical diagnosis of inborn errors of metabolism. *Analytica Chimica Acta*, 1025, 141–153.
- Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: A tutorial. *Molecular Systems Biology*, 4(1), 222.
- Mussap, M., Zaffanello, M., & Fanos, V. (2018). Metabolomics: A challenge for detecting and monitoring inborn errors of metabolism. *Annals of Translational Medicine*, 6(17).
- Noto, A., Fanos, V., & Dessì, A. (2016). Metabolomics in newborns. *Advances in Clinical Chemistry*, 74, 35–61.
- Peña-Quintana, L., Scherer, G., Curbelo-Estévez, M., Jiménez-Acosta, F., Hartmann, B., La Roche, F., Meavilla-Olivas, S., Pérez-Cerdá, C., García-Segarra, N., & Giguère, Y. (2017). Tyrosinemia type II: Mutation update, 11 novel mutations and description of 5 independent subjects with a novel founder mutation. *Clinical Genetics*, 92(3), 306–317.
- Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., & Coon, J. J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & Cellular Proteomics*, 11(11), 1475–1488.
- Roberts, L. D., McCombie, G., Titman, C. M., & Griffin, J. L. (2008). A matter of fat: An introduction to lipidomic profiling methods. *Journal of Chromatography B*, 871(2), 174–181.

- Ruoppolo, M., Caterino, M., Albano, L., Pecce, R., Di Girolamo, M. G., Crisci, D., Costanzo, M., Milella, L., Franconi, F., & Campesi, I. (2018). Targeted metabolomic profiling in rat tissues reveals sex differences. *Scientific Reports*, 8(1), 1–12.
- Saudubray, J.-M., & Garcia-Cazorla, Á. (2018). Inborn errors of metabolism overview: Pathophysiology, manifestations, evaluation, and management. *Pediatric Clinics*, 65(2), 179–208.
- Scolamiero, E., Cozzolino, C., Albano, L., Ansalone, A., Caterino, M., Corbo, G., di Girolamo, M. G., Di Stefano, C., Durante, A., & Franzese, G. (2015). Targeted metabolomics in the expanded newborn screening for inborn errors of metabolism. *Molecular BioSystems*, 11(6), 1525–1535.
- Spener, F., Lagarde, M., Géloën, A., & Record, M. (2003). What is lipidomics? Wiley Online Library.
- Villani, G. R., Gallo, G., Scolamiero, E., Salvatore, F., & Ruoppolo, M. (2017). “Classical organic acidurias”: Diagnosis and pathogenesis. *Clinical and Experimental Medicine*, 17(3), 305–323.
- Vockley, J., & Ensenauer, R. (2006). Isovaleric acidemia: New aspects of genetic and phenotypic heterogeneity. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, Vol. 142(Issue 2), 95–103.
- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., & Karu, N. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617.
- Yıldız, Y., Talim, B., Haliloglu, G., Topaloglu, H., Akçören, Z., Dursun, A., Sivri, H. S., Coşkun, T., & Tokathlı, A. (2019). Determinants of riboflavin responsiveness in multiple acyl-CoA dehydrogenase deficiency. *Pediatric Neurology*, 99, 69–75.
- Zschocke, J. (2014). *SSIEM classification of inborn errors of metabolism. Physician's guide to the diagnosis, treatment, and follow-up of inherited metabolic diseases* (pp. 817–830). Springer.

Approaches in untargeted metabolomics

7

Jacopo Troisi^{1,2,3}, Sean M. Richards^{4,5}, Giovanni Scala², and Annamaria Landolfi¹

¹*Department of Medicine, Surgery and Dentistry, “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy*

²*Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy*

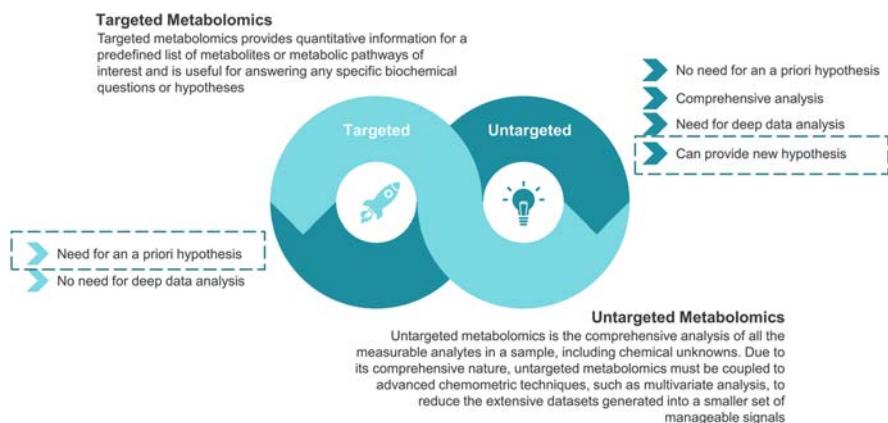
³*Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy*

⁴*Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States*

⁵*Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States*

Introduction

Metabolomics is usually divided into two separate approaches: targeted metabolomics, discussed in the previous chapter and untargeted metabolomics. The difference, as the name indicates, is that the targeted method is focused on specific metabolites whereas with untargeted methods there are no clear and predetermined study metabolites. Targeted studies are aimed at those cases where a hypothesis exists and must be confirmed. For example, speculating that a certain enzyme is involved in a given disease, substrates and products of that enzyme could be studied to test this hypothesis. In the untargeted approach, on the other hand, there is no hypothesis about a certain condition. In this case, studies should be directed to as many metabolites as possible. Through this untargeted approach, we attempt to identify metabolites that should be useful to form a hypothesis. Thus, the aim of a targeted metabolomics study is to confirm (or not) a given hypothesis, while the aim of an untargeted metabolomics study is to generate a hypothesis (Fig. 7.1). Of course, these two approaches are not mutually exclusive because a hypothesis generated in an untargeted metabolomics experiment could be confirmed using a targeted approach (targeted metabolomics or a targeted study from another omic technique).

**FIGURE 7.1**

Comparison of targeted and untargeted metabolomic approaches. The two approaches may be considered as a continuum because the hypotheses arising from untargeted studies could be confirmed using a targeted approach.

Local and nonlocal metabolomics effects

Untargeted metabolomics is defined as the comprehensive study of metabolites both in terms of presence and quantification within a cell (Dunn & Ellis, 2005). By means of untargeted metabolomics, researchers can outline the framework of metabolic reactions characterizing a specific condition or cell status, such as a disease, drug or environmental exposition, genetic predispositions, etc. (Nicholson et al., 1999). These frameworks can be used both to train mathematical algorithms building discriminating systems applicable to recognize these conditions (Troisi et al., 2019; Troisi, Autio et al., 2020; Troisi, Cavallo et al., 2020; Troisi, Landolfi, et al., 2018; Troisi, Pierris, et al., 2017; Troisi, Raffone et al., 2020; Troisi, Sarno, et al., 2018, 2017) (see also Chapter 9: Data Analysis in Metabolomics: From Information to Knowledge), or to generate new hypotheses regarding the mechanistic processes underlying the disease onset or the specific phenotype observed (Quanbeck et al., 2012; Schrimpe-Rutledge et al., 2016).

Despite both approaches being well explored and currently active research fields, limitations of these approaches are well documented. Specifically, diagnostic approaches of metabolomics fingerprints, as well as other omics, show two principal limitations. Above all, the use of large feature cohorts (metabolites in metabolomics) requires large sample sizes to build accurate and not-overfitted models. Secondly, an independent, blind, and large clinical validation is a crucial phase to translate the proposed models in clinical practice (Pinu et al., 2019), but few published models completed this important step.

Regarding the hypothesis generating role of metabolomics, this arises from the study of the metabolites primarily involved in the metabolomic signature of the

studied condition and of the metabolic route in which they are involved. These metabolites can be selected by means of several approaches as reported in [Chapter 10](#), Relevant Metabolites' Selection Strategies. One is a univariate analysis of each metabolite taken individually. In metabolomics (like in other omic sciences) statistical significance (in terms of p-value) is often conjugated with fold change evaluation in order to individuate the features with a higher variance among the studied conditions (see [Chapter 9](#) for further details). This combined assessment is generally made using “volcano plot” graphs, also known as “smile plots” ([Kumar et al., 2018](#)). The metabolites selected by means of this approach show large concentration differences among the studied conditions that could reflect different scenarios. Large quantitative differences could be inherent to molecules that are not endogenous metabolites (e.g., drugs or xenobiotic molecules). These molecules are often involved only in one condition, showing large quantitative differences. Another possibility is that these molecules are endogenous metabolites involved in metabolic routes far from complex regions.

Complexity is defined as the behavior of a system whose components interact in multiple ways and follow consistent patterns ([Chu, 2011](#)). Complexity can be disrupted by intrinsic or extrinsic factors, resulting in local and nonlocal changes in metabolomic pathways and resultant metabolomes ([Fig. 7.2](#)).

To better understand these different scenarios, an example based on a hypothetical metabolic pathway is provided in [Fig. 7.2](#). The yellow box highlights the

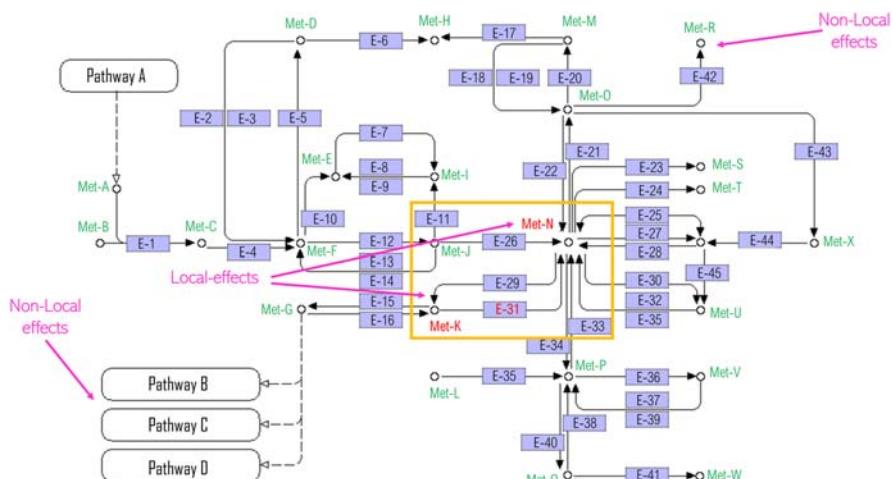


FIGURE 7.2

Hypothetical metabolic pathway illustrating local and nonlocal effects of metabolic pathway perturbation. Purple boxes indicate enzymes (E-1—E-45), circles represent the metabolites (Met-A—Met-X), arrows indicate the reactions and metabolic pathways. Yellow box highlights an example of a reaction where an enzyme (E-31) was altered generating local and nonlocal effect.

conversion of metabolite Met-K into the metabolite Met-N catalyzed by the enzyme E-31. Let hypothesize that the gene encoding E-31 mutates or is epigenetically downregulated; the lack of the full working E-31 enzyme could be easily linked to the accumulation of the enzyme substrate Met-K and the lack of the reaction product Met-N. We called this option “first scenario”. Moreover, as illustrated in Fig. 7.2, Met-K is not only a substrate of E-31 but also of other enzymes, resulting in a cascade of effects due to the increase in concentration of Met-K. Likewise, Met-N is produced by other reactions. The reduction of E-31 activity and the alteration of subsequent substrates/products, could cause a re-equilibration of the other reactions involving these metabolites according to Le Chatelier’s principle (Purich & Allison, 1999). Indeed, according to this principle: “*When a steady-state system is disturbed, it will adjust to diminish the change that has been made to it.*” For this reason, despite the lack of E-31 activity, its substrate and products could show concentrations similar to the ones registered in a cell with a full working E-31. Regardless, these compensations are not without effects. Indeed, the enzyme impairment leads to effects far away from the E-31 specific reaction. These may be referred to as “nonlocal effects.” These effects can be conspicuous or inconspicuous. In another situation (that we could call “second scenario”) the equilibrium change amplifies the original effect bringing high concentration changes far from the metabolic position where it was generated. On the contrary, a “third scenario” may produce minimal effects at great distance from the original perturbation. The third scenario is the most frequent, because biological function redundancy specifically acts to decrease the effects of a single enzymatic impairment. These masked differences are the hardest to be discovered: indeed, they cannot be found by means of mathematical univariate analyses such as volcano plot evaluations. In metabolomics, these effects are often analyzed by means of multivariate approaches.

Le Chatelier’s principle was developed observing steady-state equilibrium; however, none of the reactions in a living system can be considered as steady-state. For this reason, not-equilibrated reactions like the ones occurring in living cells need more complicated mathematical treatments to be understood and this makes the in-silico prediction of the effects of an enzymatic impairment less certain and correctly evaluable.

Untargeted metabolomics application

As stated above, in untargeted metabolomics studies there are no a priori hypotheses to be tested, so studies focus on as many metabolites as possible to try to describe the overall picture tracing the events that generated the observed phenotype or disease. This means that the whole history of each human (and we could extend the same concept for all unicellular or multicellular entities) is written in the concentration of its metabolites. The condition in which studied subjects’

kidneys, heart, brain, lungs and so on are, as well as drugs he/she has taken, the environment in which he/she has lived, the accidents he/she has had etc., all are written in his/her metabolome. For this reason, the metabolome could be used to train mathematical algorithms to associate this metabolic signature to an observed condition. Moreover, analyzing this metabolites' pattern, it is also possible, sometimes, to speculate about the future. It is possible to develop hypotheses based on the direction of the metabolomic changes and on future probable scenarios. In medicine, the description of a condition is called "diagnosis", while the inference on its evolution is called "prognosis" (Fig. 7.3).

The untargeted metabolomic approach was proposed more than 65 years ago, when [Dalglish \(1956\)](#) reported the use of paper chromatography to analyze indols and their metabolites in urine of human subjects to investigate several diseases, including bladder cancer and vitamin deficiencies. At that time, DNA structure had been reported by Watson and Crick only 3 years before ([Watson & Crick, 1953](#)), and genomics and transcriptomics were not yet theorized. Despite the foresight of the Dalglish work, a paper of the two-time Nobel laureate Linus Pauling is traditionally considered the first untargeted metabolomics experiment ([Pauling et al., 1971](#)). In this paper, published in 1971, he reported a gas-chromatographic method to separate 250 metabolites from human breath and 280 volatile metabolites from urine. He also understood the importance of gut

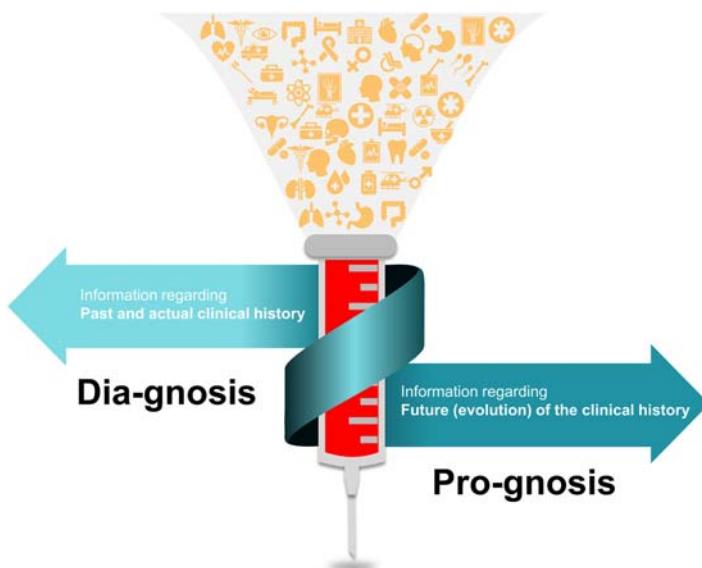


FIGURE 7.3.

Diagnosis and prognosis. Diagnosis, which is based on the Greek words "by means" (dia) "knowledge" (gnosis), and prognosis meaning "before" (pro) "knowledge" (gnosis). Both can be speculated using untargeted metabolomics.

microbiome in the metabolomics profile of such samples and performed his analysis after a specific diet able to reduce the gut-resident flora. This was a true omics approach. That study was developed in the context of the principles of the orthomolecular medicine that is currently considered as an alternative medical practice devoid of scientific evidence.

Studies about metabolomics are consistently increasing. In the last 20 years the number of published papers reported by Pubmed increased as reported in Fig. 7.4. The main research fields are related to human health and to plant metabolomics.

Metabolomics profiling

Metabolome profiling has several uses. Among these, the possibility of using these profiles as a diagnostic tool is the application that is arousing the greatest interest. In Fig. 7.5 a diagnostic pipeline using metabolomic profiling is illustrated. Samples from phenotype-identified subjects were analyzed and their metabolomic profiles were used to train a classification model able to discriminate between the studied conditions (i.e., presence or absence of a specific disease).

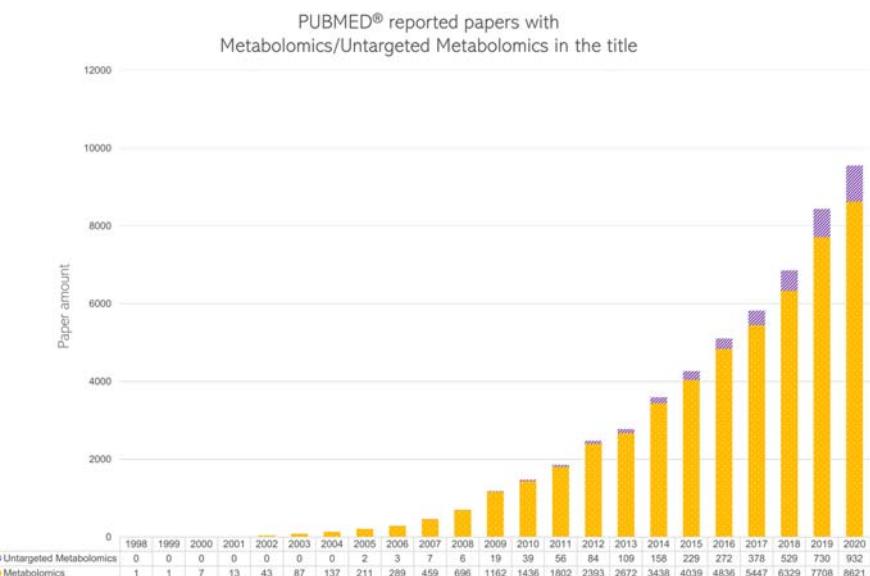


FIGURE 7.4

PUBMED reported manuscripts with Metabolomics/Untargeted Metabolomics in the title. The amount of manuscripts published in the last 20 years is consistently increasing.

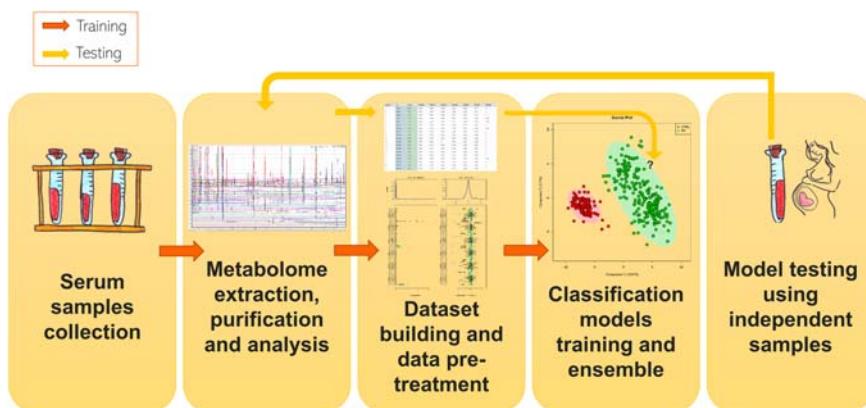


FIGURE 7.5

Metabolomic profiling as a diagnosis tool. Metabolome profiles acquired from samples collected from subjects with a known condition can be used (after data pretreatments) to train one or several classification models. These models can be then used to predict the class label of an independent sample collected from a subject with an unknown condition. For example, this approach was used to get information about the developmental status of fetuses.

Once such an algorithm is trained it can be used to assign the correct class for samples taken from subjects without a diagnosis (see Fig. 7.5).

Many human pathologies are intrinsically complex, because they derive from a series of events whose combination triggers a cascade of occurrences that culminates with the pathological phenotype. In such a scenario, as we have already highlighted, a reductionist approach that seeks to identify the cause of the disease from a single impairment is doomed to failure. In the same context, a single biomarker, understood as a molecule able, alone, to represent this complex network of events is unrealistic. A lot of evidence is now accumulating about the need of a large representation of the dynamicity of the living systems, and this could be considered the basis of the current success of the omics sciences, especially metabolomics.

The most common approach for the diagnostic applications of metabolomics remains the search for candidate biomarkers and their subsequent validation on large cohorts of patients. In 2014, [Medina et al. \(2014\)](#) presented a review of all biomarkers that had been identified by means of metabolomic approaches. Since then, this kind of research has continued successfully, and the biomarker list today can be considered even more extensive.

Nevertheless, due to the intrinsic complexity of metabolic systems, as already pointed out, the designation of a single representative for a pathological condition that required several actors to establish itself is often insufficient. Indeed, as [Evans and colleagues illustrated \(Evans et al., 2020\)](#) models trained using

information from all metabolites collected in a dataset, generally show classification performances that exceed those obtained from models trained with only a few metabolites. Using many metabolites results in high predictive performance in different experimental and pathological contexts suggesting that even metabolites whose significance seems not relevant play an important role. Moreover, it is often possible to train models and achieve good predictive performance even using only metabolic information deemed insignificant. This precisely suggests that, even in the absence of statistical significance, the information contained in a dataset can still be exploited to properly train classification systems because these contain information that are consistent with the health status of the subjects from which that metabolome derives. The work of [Evans et al. \(2020\)](#) highlights... the potential for disease diagnosis using metabolomic profiles by means of a task-driven approach and not just using candidate biomarkers.

The advantage in terms of diagnostic performance is also well represented by the growing interest of the human diagnostics industry for metabolomics profiling. In fact, several patents have recently been requested for diagnostic systems based on this technology ([Fraser et al., 2016](#); [Slupsky, 2012](#); [Troisi, Scala, et al., 2017, 2018](#)).

Moreover, metabolic profiling can also be used as a hypothesis generator system (see [Fig. 7.6](#)). Indeed, the metabolites that emerged as the most relevant either in a univariate statistical evaluation or using multivariate and machine learning approaches could be analyzed in the framework of their metabolic pathways in order to develop a hypothesis for the metabolic imbalance that generated the observed profile (see [Chapter 11: Pathway Analysis](#)).

A complete discussion of the applications and the results that have been obtained using metabolomics profiling is too broad and beyond the scope of this chapter. For purely illustrative purposes, only some of the results recently emerged in the field of cardiovascular and neurodegenerative diseases, two of the most impacting and disabling conditions affecting humankind, are reported.

Cardiovascular disease

Cardiovascular diseases (CVD) are the leading cause of death worldwide (except Africa). In 2015, CVD resulted in 17.9 million deaths, more than one third of the total deaths, while in 1990, CVD was responsible for about 12 million, about one quarter ([Abubakar et al., 2015](#); [Wang et al., 2016](#)). Incidence is higher in men who also receive a diagnosis about ten years earlier as compared to women. The lower CVD risk for woman seems be related to the fertile age; after menopause, incidence in men and women tends to be equal.

It is estimated that more than 90% of CVD could be prevented ([McGill et al., 2008](#)). Healthy eating, exercise, avoidance of tobacco smoke and limiting alcohol intake are the most effectives strategies to mitigate the risk. Moreover, other well

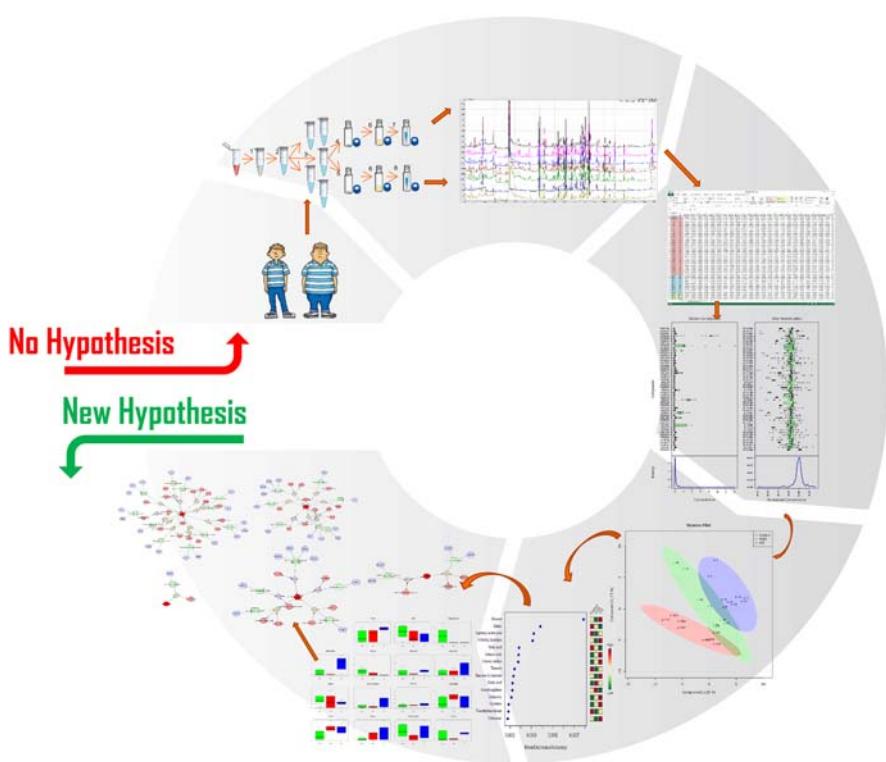


FIGURE 7.6

Metabolomic profiling as a mechanistic hypothesis generator. The figure illustrates the chain of events for using a case-control approach to develop a hypothesis for disease pathology. In this example, metabolome profiles acquired from samples collected from subjects with a known condition can be used (after metabolome extraction and purification and data pretreatments) to train one or several classification models. Metabolites with a higher impact in model building could be investigated in the context of the known metabolic pathways and an impairment(s) could be investigated as the driving force responsible for the observed profile.

known risk factors, such as high blood pressure, blood lipids and diabetes play a pivotal role. Unfortunately, the exact mechanisms linking the risk factors to the disease onset are not completely elucidated. Metabolomics profiling, defining changes in both global and cardiac-specific metabolism that occur across a spectrum of CVD is contributing to better understanding of the metabolic changes that occur in these conditions.

Plasma metabotype was used by (Wang et al. 2011) to investigate metabolites that could predict the increased risk for CVD. They enrolled patients undergoing elective coronary angiography who subsequently had a major adverse cardiac

event including nonfatal acute myocardial infarct, cerebrovascular accident or death and compared their profiles with the ones from age-matched healthy controls. From the 2000 metabolites identified, 18 were different in the observed cohort and validated in a smaller independent group. Three of the most promising metabolites (namely choline, trimethylamine N-oxide, and betaine) were also tested in a mouse model proving their ability to increase the CVD risk to develop atherosclerosis. These molecules are not produced by human cell metabolism while are produced by bacteria living in gut. One of the great merits of this work was to demonstrate the pivotal role in disease onset of the host-gut microbiota interactions. This is a great example of how untargeted metabolomics provided a hypothesis (gut microbiome metabolites involvement in CVD risk) which was validated in a mouse model.

The significance of these results is twofold. First, they indicate the complexity of the health and disease mechanisms. Metabolomics profiles, derived from human biochemistry, gut microflora, diet, pollutants are intricately complex and only an accurate data analysis process could glean the right information from it. Moreover, the delicate equilibrium between human cells and microbiota is a crucial aspect of health and disease making it a new target for pharmacological treatments (see also [Chapter 14: Gut Microbiota-derived Metabolites in Host Physiology](#)).

CVD are also important triggers for neurological disability leading to up to 20% of the stroke events. [Jové et al. \(2015\)](#) investigated plasma metabotypes of patients with transient ischemic attack with stroke recurrence or not, and found 94 metabolites (from 2081 analyzed) that significantly differentiated these subjects. Most of these metabolites were lysophosphatidylcholines, including LysoPC (16:0) and LysoPC(20:4), which are also proposed as potential biomarkers of atherosclerosis. The most important finding from this study was that pattern recognition using appropriate multivariate statistical approaches could differentiate subjects who present early (less than three months) stroke recurrence. This illustrates the utility of metabotyping as a robust prognostic tool.

Lysophosphatidylcholines were also selected from more than 10,000 features detected in a large study by [Ganna et al. \(2014\)](#) which studied 3668 subjects from 3 different cohorts to evaluate the relationship between metabotypes and CVD incidence. In particular, LysoPC(18:1), LysoPC(18:2), and sphingomyelin(28:1) were found to be associated with lower risk, while high monoglyceride(18:2) levels predicted a higher risk. These metabolites allow to increase of 10% the net reclassification index if included in the Framingham heart study risk score. Moreover, they propose new insights about the mechanism of the disease onset.

Most of the plasma monoglycerides are produced by lipoprotein lipase and hepatic lipase, which catalyze the hydrolysis of triglycerides to provide nonesterified fatty acids and monoglycerides for tissue needs. Subsequently, monoglycerides are converted into free fatty acids and glycerol by monoglyceride lipase. Moreover, monoglycerides are used to resynthesize diglycerides and triglycerides via monoacylglycerol pathway before being transported in lymph to the liver. These steps occur within the intestinal wall.

A large mendelian analysis (Do et al., 2013) revealed an association of several single nucleotide polymorphisms (in the PCSK9, HHIPL1, PLG, ApoE/ApoC1, COL4A1/COL4A2 regions) to both CVD risk and monoglyceride(18:2) levels revealing the possible role of this metabolite in the CVD onset. Furthermore, monoglyceride(18:2) blood concentrations better correlated with the CVD risk compared to total circulating triglycerides values.

To investigate the metabolic change in patients suffering myocardial ischemia, Sabatine et al. (2005) studied the plasma metabolomic profiles of subjects immediately before, immediately after, and 4 hours after stress testing. Focusing the attention on patients showing a clear-cut inducible ischemia or not, they reported that although most of the analyzed metabolites displayed concordant changes in cases and controls, six metabolites including γ -aminobutyric acid, uric acid, citric acid, and three metabolites for which it was not possible to get a structural identification, were found to be associated with inducible ischemia. A mathematical model built using these 6 metabolites showed a c-statistic (the probability for a randomly selected patient who experienced inducible ischemia had a higher risk score than a subject who had not experienced the event) of 0.83. Moreover, pathway analysis of annotated metabolites showed an involvement of citric acid pathway with a significant overrepresentation of some metabolites from this pathway that could in the long run serve as targets for therapeutic intervention.

Neurodegenerative disease

Neurodegenerative diseases are a multifaced set of central nervous system diseases, united by a chronic and selective process of neuron death. Depending on the type of disease, neuronal deterioration can lead to cognitive deficits, dementia, motor impairments, behavioral and psychological disturbances. Alzheimer (AD) and Parkinson (PD) disease are the most diffuse neurodegenerations.

It is predicted that by the year 2050 more than 115 million people around the world will be affected by AD. Unfortunately, to date no effective treatments or validated diagnostic and/or prognostic biomarkers are known. In both AD and PD, predictive biomarkers for disease diagnosis could greatly help medical management because accurate clinical diagnosis needs long observational time after the beginning of the pathological process.

Nevertheless, accumulating evidence indicates that differences in lipid and sugar metabolism, is well defined between AD and healthy subjects. It seems that these impairments are correlated with the role of ApoE4, a cholesterol carrier supporting the lipid trafficking, in the increasing of AD risk as well as other age-related cognitive decline during normal aging.

Mapstone et al. (2014) using an untargeted metabolomics approach, identified a blood-based biomarker panel including phosphatidylinositol (18:0/0:0),

proline-asparagine dipeptide, glycoursoodeoxycholic acid and malic acid with very high accuracy for detecting preclinical AD. Metabolic profiling also showed prognostic ability, robustly identifying cognitively normal individuals who, on average, will develop a mild cognitive impairment or AD within 2–3 years.

Glucose metabolism impairment in brain is also a consistent pathophysiological feature in AD. Several studies of glucose metabolism aberrations indicated that these may precede cognitive dysfunction, leading to a correlation of AD with diabetes (Cunnane et al., 2011; Wilkins & Trushina, 2018).

Understanding of AD mechanism and therapeutic options improved through the identification of impairments in several metabolic networks, including lipid and amino acid metabolism, and metabolic pathways involved in glucose and energy substrate utilization. These evidences, combined with the discovery of further risk factors, such as type 2 diabetes, boost the perception that AD has a metabolic base (Demetrius & Driver, 2013).

Parkinson's disease is another important neurodegenerative disease whose incidence ranges from 5 to >35 new cases per 100,000 individuals per year worldwide. PD leads to motor impairment (bradykinesia) variably associated to tremor and/or rigidity. Symptoms usually begin gradually and get worse over time. As the disease progresses, people may have difficulty walking and talking. They may also have mental and behavioral changes, sleep problems, depression, memory difficulties, and fatigue.

Mortality increases in the advanced stages of PD. Nevertheless, the progress in health care is providing an increase of patient survival. This increased survival means an increasing prevalence of Parkinson's disease and consequently, PD is expected to represent an increasing burden in our society in the next future. Because sporadic PD is widely considered to have a multifactorial origin, that is, it is caused by the interaction between genetic and environmental factors, the use of metabolomics is increasingly seen as a useful tool in the study of PD.

PD is due to the dopaminergic neuronal loss in the substantia nigra pars compacta and the progressively widespread intracellular pathologic accumulation of unfolded α -synuclein. Data about metabolite concentration, obtained from several biological matrices of PD-affected subjects, were used to train several classification algorithms (Troisi et al., 2019). Trezzi et al. (2017) proposed a logistic regression model based on the concentration of threonic acid, mannose and fructose in cerebrospinal fluid to differentiate drug-naïve PD patients from matched healthy controls. This result was further validated using an independent cohort of more advanced PD patients and controls. Ultimately this provided tools for both early PD diagnosis and to increase the insight about the disease mechanisms.

Cerebrospinal fluid was also collected postmortem to conduct a metabolomic profiling study on pathologically confirmed PD patients and controls using a support vector machine (LeWitt et al., 2017). In this study several N-acetylated amino acids showed a significant discrimination ability to distinguish between the two classes. The levels of 3-Hydroxykynurenone and the ratio of 3-hydroxykynurenone and kynurenic acid concentrations were significantly increased in PD subjects while

glutathione levels were lower in PD patients. These results enabled the authors to speculate that an imbalance in the concentration of 3-hydroxykynurenone and kynurenic acid, two tryptophan metabolites with opposite actions on brain oxidative stress, may act as possible biomarkers for Parkinson's disease and could have a role in disease onset.

Limitations

Although the advantages of untargeted approaches in metabolomics are numerous, some limitations must also be taken into consideration. First of all, profiling is generally more complicated and expensive than a single biomarker assay. Indeed, profiling requires specialized personnel and the use of sophisticated and expensive equipment (e.g., MS, NMR). Furthermore, metabolomic profiling has not been explored relative to how the information translates to clinical application. While this is a complicated issue to overcome, because it requires large investments and the involvement of large cohorts of subjects to be enrolled, some interesting results have already been reported. In particular, we have shown that a metabolomic signature obtained by mass spectrometry coupled with gas chromatography is useful in diagnosing endometrial cancer in postmenopausal women and that this signature is also useful for discriminating against some conditions that have a very similar clinical presentation (i.e., benign metrorrhagia) ([Troisi, Sarno, et al., 2018](#)). This approach was recently applied to a population screening campaign that demonstrated how this can be successfully applied and therefore the method is widely translatable ([Troisi, Raffone et al., 2020](#)). This and similar studies in progress seem to indicate a possible future way to overcome the approach based on single biomarkers.

There are other considerations that must be taken into account regarding the possibility of using metabolomic profiles as a database for training artificial intelligence systems. First, as already reported in other chapters, this will require the adoption of standardized procedures (SOPs) for both the collection, extraction and purification of the metabolomes, but also for instrumental investigations and for data analysis procedures (both in terms of preprocessing and in terms of classifiers training). Continuous progress in this area suggests that all these aspects are not yet quite mature and so there is still a long way to go to achieve such standardization.

Furthermore, metabolomic profiles are extremely dynamic. However, this represents a great advantage in terms of information content, indeed even small variations in pathological-specific conditions are quickly and constantly "written" in the metabolome and therefore can be "read" through its analysis. Indeed, variation in the metabolome is what allows us to differentiate (through trained algorithms) case from control cohorts. At the same time, however, this extreme speed of adaptation also represents a major source of fluctuation that makes the training

process more difficult. In the next chapter we will see in detail that some aspects of machine learning training, especially in its supervised version, can help in mitigating these fluctuations. As we gain experience in training models and interpreting the data, these dynamisms will be a strength of metabolomics, rather than a limitation.

Sources of metabolome variability

Medicine, since its origins, has had a mainly androcentric approach, relegating the interests of female health to aspects connected to the peculiarity of the reproductive system and the period of pregnancy. Even the studies conducted in the clinical and pharmacological field have historically been carried out considering, above all, male subjects and then adapting the results to the women. The evaluation of female biology, with the anatomical, functional and hormonal peculiarities that characterize and can influence, sometimes decisively, the development and progression of diseases, has not been considered until recently. For example, medical studies have historically been designed around an adult man weighing 70 kg. Only recently has gender medicine proposed to deepen the study of the impact of “gender” (and of all the variables that characterize it, not only biological but also environmental, cultural and socio-economic) on physiology and pathophysiology, with the aim of understanding the mechanisms through which gender-related differences act on the state of health and on the development of pathologies ([Oertelt-Prigione & Regitz-Zagrosek, 2011](#)). Of course, these approaches affect the way the metabolome was investigated between men and women.

[Krumsieck et al. \(2015\)](#) also investigated the gender influence in metabolomics profiles. They investigated a large cohort including 1756 fasting serum samples (903 females and 853 males) derived from the KORA F4 cohort. They performed an untargeted metabolomics evaluation resulting in the identification of 507 metabolites, (of which 318 were known identity and 189 unknown). This is a particularly valuable investigation because their results were further confirmed using another large population-based cohort from SHIP study containing 1000 fasting plasma samples, of 561 females and 439 males. The metabolomics profiling for both cohorts were obtained using the same platform. One hundred and eighty of the 507 measured metabolites (35.0%) were significantly different between male and female. A total of 88 of these 180 were confirmed in the independent cohort (SHIP). Both the KORA and SHIP cohort were recruited in Germany, and this reduced the metabolomics variability due to similar food and social habits of the enrolled people. For this reason, we can speculate that a different confirmation cohort could offer different and less coherent results. Regardless, the principal message is that a large and consistent difference exists between male and female metabolomics profiles. This is exemplified by a

metabolomics pathway analysis showing large and consistent differences in different diseases with gender-specific susceptibility, namely, CVD. This evidence could represent a concrete starting point for further development of gender-specific personalized health care as well as new insight in the disease onset mechanisms. These results should be taken in great consideration in untargeted metabolomic profiling studies.

In addition to gender, several other conditions showed great influence in metabolomics profiling. [Sato et al. \(2018\)](#) analyzed the blood and muscle tissue of 8 obese men and built metabolomic profiles for both the morning and in the evening. Among the 1063 analyzed blood metabolites, more than 50% exhibited time-of-day differences. Principal component analysis identified a good separation of samples collected in the morning and the ones collected the evening. Moreover, they studied the impact of high fat and high carbohydrate diets in metabolomics profiles. Respectively, the levels of about 18% and 16% of total serum metabolites differed at morning and evening measures, five days after high fat diets. Approximately 18% and 17% of total serum metabolites decreased after high carbohydrate diets in the morning and in the evening.

To further illustrate the complexity of the metabolome, diet, especially over time, deeply influences the equilibrium among the microbes in human gut. Microbiome-metabolite interactions have a profound influence in the human body. Recent studies revealed that the microbiome has a heavy impact on the metabolome both near and at distant body sites ([Lee-Sarwar et al., 2020](#)). In addition, the metabolome also influences the microbiome composition and behaviors. It is now clear that crosstalk between the microbiome and metabolome has an impact on many human diseases through a variety of mechanisms. This knowledge can be used to develop microbe- and metabolite-targeted treatments and preventive strategies, including probiotics, prebiotics and other dietary modifications, supplementing or inhibiting microbial-derived metabolites, and fecal microbiome transplants ([Troisi et al., 2019; Troisi et al., 2021; Troisi, Landolfi, et al., 2018; Troisi, Pierri, et al., 2017; Troisi, Raffone, et al., 2020; Troisi, Sarno, et al., 2018, 2017](#)). Nevertheless, the role of microbiome composition in metabolomic profiles results is pivotal, contributing to further increases in the metabolomics variability in human biofluids and, in turn, increasing the mathematical efforts to measure the disease/phenotype relevant information.

[Bucaciuc Mracica et al. \(2020\)](#) investigated the role of aging in the metabolomics profiles. They proposed MetaboAge (<http://www.metaboage.info>), a freely accessible database hosting 408 fully annotated metabolites with their biological and chemical information, and more than 1515 ageing-related variations. Beyond the specific aspects and the usefulness of these results in understanding the basic mechanisms of physiological aging, the main message is that age is a determining factor for the variability of metabolomic profiles. Therefore, this must be carefully evaluated in experimental design and subject recruitment and enrolling. In fact, in a case-control approach, for example, an imbalance in the average age in the two populations could lead to a false-positive indicating that a specific

metabolic pathway is due to a disease while, in reality, the specific difference is simply due to the different average age of the two populations.

Another key factor influencing the metabolomic profiles is the nutritional state both in terms of habits and in timing. This is a trivial observation, and this variability could be managed by standardizing the biofluid collection timing and enrolling subjects from populations with similar diets. However, Agueusop and colleagues ([Agueusop et al., 2020](#)), exploited this concept in a very elegant untargeted metabolomics experiment to highlight the profiling difference in normal, prediabetics and type II diabetes (T2DM) affected people. Indeed, at baseline (fasting) blood, 49 out of 1438 (3.4%) investigated metabolites were significantly different between healthy subjects and T2DM patients while 1 hour after eating, the difference affected 6.5% of all metabolites. Thus, collecting after a meal, results in a better profile clustering and an improved ability of classification system to discriminate between the studied conditions. We believe that this is an excellent example that teaches how to exploit apparent difficulties (variability induced by nutritional status) as strengths (increase in visibility of the effect due to the studied condition).

Several other factors contribute to the metabolomic profile diversity related to lifestyle choices (e.g., housing context, smoking, physical activity), genetic background, drugs, etc. A recent study ([Bar et al., 2020](#)) illustrated that the most influential aspects are diet and microbiome composition. Collectively, about 76% of the known variability can be attributed to the factors listed above; however, 24% of the variability is unknown.

Key trends in untargeted metabolomics

Metabolome coverage

The human genome, as well as the genomes of other species are fully described. This was possible because genes are codified in DNA which has a definite and known dimension in terms of nucleotide base pairs. Unfortunately, this is not the situation for the metabolome. Indeed, we have no reference to recognize if a metabolome is complete or not. Metabolites show a large spectrum of chemical characteristics in term of mass range, polarity, solubility and so on. For these reasons, a complete metabolome description is still lacking.

The first attempt to describe the human serum metabolome was provided 10 years ago, when Wishart and coworkers proposed a pool of 4229 known and probable metabolites for serum metabolome ([Psychogios et al., 2011](#)). In the last 10 years this number only slightly increased to 4651 metabolites. It is safe to assume that although the exact spectrum of serum metabolome is not yet reached the actual estimation are probably close to the true.

The Wishart group study required five different analytical approaches to capture the whole metabolome, because no technique, taken alone, was able to

separate and quantify all these metabolites. This is a crucial aspect in untargeted metabolomics and a cumbersome limitation that other omic techniques do not suffer. Indeed, since metabolomic profiling required the acquisition of as many metabolites as possible it is obvious that a single analytical approach could not be enough. In addition, the increase of the number of analytical techniques requires a parallel increase of the experimental costs. Although it is clear that no speculation is possible about metabolites or metabolic pathways that have not been investigated, it is also clear that experimental costs represent a crucial aspect of modern research. For this reason, the experimental design should choose the most efficient analytical approach.

Recently an interesting possible solution to the choice between completeness and experimental cost was proposed using chemical isotope labeling liquid chromatography mass spectrometry (CIL-LCMS) (Zhao et al., 2019). Unlike conventional LCMS, CIL-LCMS analyzes chemical-group-based submetabolomes and uses the combined results to represent the whole metabolome. Zhao et al. showed that detection of most of the metabolome is theoretically achievable by analyzing only the H (hydroxyls), A (amines and phenols), C (carboxyls), and K (ketones/aldehydes) submetabolomes by means of a CIL-LCMS approach. However, this study was related only to the polar metabolome, while, as reported by the human serum metabolome project (Psychogios et al., 2011), most of the serum metabolites are lipids (3381 out of 4229 total metabolites). These lipids can be analyzed after separating the lipid classes individually using a thin-layer chromatography with a subsequent GCMS analysis after hydrolyzing the lipids into their constituent acyl chains. This is a useful approach, moreover GCMS is also able to identify 99 polar metabolites, of which 70 were unique with respect the other investigated techniques. This evidence, taken together with the relatively low cost and the robustness made the GCMS a very diffuse technique and useful platform in untargeted metabolomics. Several LCMS based shotgun (without chromatography) approaches were also proposed to separate and quantify the lipidomic fraction. These approaches are less time-consuming compared to the TLC/GCMS and are quickly becoming the most diffuse approach.

Moving metabolomics from laboratories to clinics

Several challenges translating metabolomic biomarkers from discovery to application in population health studies are still present and represent the main goals to be addressed in the future. Metabolome coverage and profiling affecting factors are key aspects that should be considered in addition to metabolomics pipeline standardization and accurate sample sizing in discovery experiments. Further, the need to further validate the obtained results in an independent cohort are pivotal aspects to be considered.

Metabolomics pipeline standardization

Extensive quality control and quality assurance protocols are not yet diffuse in metabolomics studies although there is a consensus about their important role in clinical translating of untargeted metabolomics. Standard operation procedure (SOP) implementation could reduce the preanalytical variation and batch to batch variability of data in metabolomics workflows. Preanalytical variation is rooted in sample collecting, storage, and shipping. For example, there is no accepted procedure to limit the impact of hemolysis on plasma and serum samples. Significant degradation of metabolites can occur from sample collection until analysis despite the application of the best practices. For this reason, protocol standardization is mandatory, especially in metabolomics profiling. Currently, biofluid harvest and storage protocols are standardized in most clinics and hospitals; however, the procedures vary between countries, regions, and often between clinics. Data analytical manipulation further increase the variability of the results and are also critical aspects to standardize. This need is well represented by the formation of at least two different groups that are working to harmonize the untargeted pipelines: the Metabolomics Quality Assurance & Quality Control Consortium (<https://epi.grants.cancer.gov/Consortia/mQACC/>) and the Metabolome Standard Initiative (Spicer et al., 2017; Sumner et al., 2007).

Sample size

Sample size is a key feature potentially affecting the results of all experimental studies. In untargeted metabolomics it is particularly critical both because the experimental costs significantly increase with the sample size and also because the right samples size is not so easy to calculate *a priori*. Moreover, large-scale studies also required long analytical times that increase the chromatography drift, requiring an invasive approach in chromatographic alignment. Dunn et al. (2015), in a large-scale study involving clinical mass spectrometry-based metabolomics profiling, reported that underestimated samples size greatly affect results reproducibility. They reported that, median accuracy increased, and variation decreased, as sample size approached 600, a sample size generally considered as representative of the whole sample population. This is only a general representation because all experimental settings require an accurate and unique sample size evaluation.

Independent cohort to validate the results

Clinical translatability of findings from metabolomics studies, obtained using a discovery cohort, requires a validation step. The validation involves another

population with respect to the discovery one. There are two options to get such a population. First, it could be obtained by splitting samples from the discovery cohort in two independent populations: the discovery and the validation cohorts. The second choice is to choose a new cohort that is so systematically different from the discovery one. In the first scenario, beyond the considerations about the sample size, according to which, the presplitting population design must be taken into account, the postsplit statistical power calculation in both branches (discovery and validation) must be conducted. Such a validation cohort cannot be considered totally independent because it was obtained from a common origin. Moreover, the two cohorts are homogenous in several terms (age, sex distribution, diet- and social-habits, etc.). Furthermore, this kind of validation cohort generally includes the samples' label (that is mandatory in the discovery step) making the validation phase not blinded.

In the second scenario of selecting an independent validation population, a careful harmonization of the phenotype definitions and other experimental conditions across the discovery and validation samples is mandatory. To reduce systematic variations and bias, validation cohort characteristics, in terms of features able to influence the metabolomic profiles (e.g., sex, age, diet) as well as data analysis pipeline, should mirror those used in the discovery stage. Moreover, special caution should be used for validation across different ethnic groups. In genomics there are several obligate evaluations to be taken, while in metabolomics the importance of race/ethnic distributions in discovery and validation populations is still debated. Indeed, differences in genetic profiles influence several traits of human physiology and disease, and accounting for such differences is mandatory to generalize the results. A different approach supports the study of each ethnicity separately to achieve the best diagnostic performance in each of them although using different metabolites panels/classification models training.

Regardless of the above-mentioned considerations, validation is a pivotal step to allow the clinical deployment of a metabolomics-based discovery on a specific population-based study. Nevertheless, this important step is often neglected, making many metabolites and metabolomics-based prediction algorithms orphans of potential practical use. There are several reasons behind this apparent waste, and these involve different aspects of modern scientific research. First of all, the validation studies are complex because they require to take into consideration various aspects about the intrinsic variability of the metabolomic profiles which, if incorrectly evaluated or underestimated, can affect the entire process. Furthermore, these studies, especially if conducted blindly and on relatively rare conditions, require the recruitment of very large cohorts (accounting for thousands of subjects); this obviously represents a challenge both for logistical reasons and, above all, for the enormous costs to be incurred. Furthermore, the analysis of such a large number of samples also represents a huge technical challenge for metabolomic profiling. In fact, in this case it is necessary to use different analytical platforms that work in parallel and many different chromatographic columns (in the case of hyphenated techniques), in this scenario the pretreatment of the data

represents an extremely complex step. The algorithms traditionally used, for example for the alignment of signals, are in fact very reliable in the management of small deviations, while, on the contrary, they often fail in the management of large deviations (unavoidable on long-term projects) especially in the case of sample/analysis providing many signals. Moreover, although mass-produced, mass spectrometers and nuclear magnetic resonance spectrometers are not exactly alike. Small variations in magnetic fields or in the size of the MS analyzers components generate slightly different response factors for the various metabolites. This is another aspect that must be carefully managed in this type of studies. Validation studies of untargeted metabolomic profiles on independent and blinded cohorts are therefore expensive, extremely complex, time-consuming and required expert and specialized researchers to manage different technical aspects. It is therefore not surprising that many groups give up this step which, as mentioned, is crucial to finalize the efforts made in the discovery phase and bring the benefits of research to the bedside. Furthermore, the current research evaluation model in different countries is often based on quantitative rather than qualitative parameters and this discourages the development of long-term studies which produce few results in the short term. In the near future, the national and internationals bodies that finance the research should promote and enhance the studies that complete a validation phase and therefore are able to make the results obtained usable.

Cause/effects disambiguation

As we have already pointed out, metabolomics in its untargeted version does not provide certain answers, and it only generates hypotheses. The amount of hypotheses that can be obtained by analyzing such a mass of information is limited only by the imagination and speculation skills of the researchers who analyze such data.

For example, patients with severe COVID-19 (the disease induced by SARS-CoV-2 infection) have been reported to often have elevated serum lactate levels ([Troisi, Cavallo, et al., 2020](#)). T-lymphocytes during the transition from naïve to activated state have a metabolic switch toward anaerobic metabolism that leads them to produce and excrete large quantities of lactate. Curiously, many cells in the process of neoplastic transformation do the same (see [Chapter 16: Metabolomics for Oncology](#)), and this seems to be one of the strongest repressors of the immune response against these cells. This ongoing evidence of COVID-19 can be speculated as originating from a strong activation of T-lymphocytes. This hypothesis is also supported by the evidence regarding the crucial role of the immune system in the genesis of the symptomatology of this disease. In fact, COVID-19 seems to be mainly due to the cytochemical storm or, in general, to the damage induced by the immune response rather than by the direct action of the virus. Furthermore, a down-regulation of the immune response, such as the

good results obtained with the treatment of patients with severe symptoms with tocilizumab, an inhibitor of the interleukin-6 receptor (Luo et al., 2020), have shown to be an effective strategy. From all these facts, it follows that this increase in lactic acid can be regarded as further evidence of the overactivation of the immune system (in particular the T-mediated immunity) in the course of pathology with severe symptoms.

Conversely, it is also true that the COVID-19 affected subjects almost always have respiratory difficulties due to the severe interstitial pneumonia caused by this condition. This generates a reduction in the circulating oxygen concentration and consequently many cells are induced, by means of an HIF-mediated mechanism, to a switch toward an anaerobic metabolism that naturally generates lactate. From this simple example it is possible to understand that the same evidence may have at least two possible explanations, one as the cause of the condition (phenotype or presence of severe symptoms), while the other because of it (hypoxia/oxygenation).

To attempt to disambiguate between these it is necessary to conduct further experiments and often, a complete change of the experimental design is necessary. This type of ambiguity is, in fact, often generated by a case-control experimental design, a design in which subjects with a given phenotype (i.e., presence of disease) are compared with healthy subjects, albeit matched for other conditions (age, sex, eating habits, etc.). In this case it is often difficult, if not impossible, to understand if the differences found in metabolomic profiles were at the basis of the development of that condition or are a mere result of it. This makes it particularly difficult to understand the overall picture for those pathologies for which, to date, there are no solid mechanistic bases, for example, psychiatric pathologies.

In this regard, the current state of research on autism spectrum disorders (ASD) is particularly emblematic. This is a condition whose incidence seems to be constantly increasing and which currently affects about 1 in 54 children in the USA (Maenner et al., 2020) with a ratio between males to females of about 4:1. Although the pathophysiological basis of this condition has been speculated for over a century, to date, both the basic mechanisms and the reasons for this increase in incidence and the imbalance between males and females remain elusive. This is due to several reasons. ASD seems to be a typical example of a multifactorial condition in which one or more events act on the basis of genetic predisposition, which in turn trigger a cascade of events of a metabolic, epigenetic, proteomic nature and, above all, the imbalance of the intestinal microbiota and in turn of the circulating metabolome (both host- and gut-derived) that hesitate in the development of pathology. Studying such a large and complex picture is obviously a challenge. Case control studies (which represent almost all the studies conducted to date) have limited usefulness in this context. Indeed, establishing that the balance of the intestinal microbiota or metabolic profiles are different between ASD subjects and neurotypical controls does not help to understand how this condition arose. Only a longitudinal study that allows

following a group of subjects from birth until the onset of the disease and compare these with subjects in which the disease did not appear, can help to understand how the normal development path deviates to pathological and what are the triggering events—ultimately to understand how the chain of events that leads to the development of the pathology is generated. This type of study is obviously very complex, as well as long and expensive, but this challenge has recently been accepted by a Euro-American consortium that will follow the development of 600 children at risk of developing this pathology from birth for 5 years, trying to reconstruct the entire dynamic of the events leading to the development of ASD (Troisi, Autio, et al., 2020).

This long-term paradigm could be followed for many other conditions. The most important message is that the observation of evidence, in metabolomics, does not necessarily mean that it is responsible for the development of the observed phenotype. However, it could represent a hypothesis that must be further investigated.

Conclusion

In conclusion, untargeted metabolomics is a promising tool allowing the investigation of conditions whose mechanism of onset is not yet elucidated. It facilitates a hypothesis generation that should be further verified using targeted approaches. Metabolomic profiling could also be used to characterize the complexity of the biological system and train mathematical algorithms to recognize specific conditions in the general population. To do this, some critical aspects must be carefully managed. First, the population features (sex, age, diet, gut microbiome composition, lifestyles, ethnicity, etc.) that largely affect the profiles should be taken into account. Of course, the sample size should be always accurately evaluated. If such standards are uniformly adopted, and with the validation of results using a large, blind, and independent population, these important data can ultimately be used in the clinical setting.

References

- Abubakar, I., Tillmann, T., & Banerjee, A. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 385(9963), 117–171.
- Agueusop, I., Musholt, P. B., Klaus, B., Hightower, K., & Kannt, A. (2020). Short-term variability of the human serum metabolome depending on nutritional and metabolic health status. *Scientific Reports*, 10(1), 16310. Available from <https://doi.org/10.1038/s41598-020-72914-7>.

- Bar, N., Korem, T., Weissbrod, O., Zeevi, D., Rothschild, D., Levitan, S., Kosower, N., Lotan-Pompan, M., Weinberger, A., Le Roy, C. I., Menni, C., Visconti, A., Falchi, M., Spector, T. D., Vestergaard, H., Arumugam, M., Hansen, T., Allin, K., Hansen, T., & The IMI DIRECT consortium. (2020). A reference map of potential determinants for the human serum metabolome. *Nature*, 588(7836), 135–140. Available from <https://doi.org/10.1038/s41586-020-2896-2>.
- Bucaciuc Mracica, T., Anghel, A., Ion, C. F., Moraru, C. V., Tacutu, R., & Lazar, G. A. (2020). MetaboAge DB: A repository of known ageing-related changes in the human metabolome. *Biogerontology*, 21(6), 763–771. Available from <https://doi.org/10.1007/s10522-020-09892-w>.
- Chu, D. (2011). Complexity: Against systems. *Theory in Biosciences = Theorie in Den Biowissenschaften*, 130(3), 229–245. Available from <https://doi.org/10.1007/s12064-011-0121-4>.
- Cunnane, S., Nugent, S., Roy, M., Courchesne-Loyer, A., Croteau, E., Tremblay, S., Castellano, A., Pifferi, F., Bocti, C., & Paquet, N. (2011). Brain fuel metabolism, aging, and Alzheimer's disease. *Nutrition (Burbank, Los Angeles County, Calif.)*, 27(1), 3–20.
- Dalgleish, C. E. (1956). Two-dimensional paper chromatography of urinary indoles and related substances. *The Biochemical Journal*, 64(3), 481–485. Available from <https://doi.org/10.1042/bj0640481>.
- Demetrius, L. A., & Driver, J. (2013). Alzheimer's as a metabolic disease. *Biogerontology*, 14(6), 641–649.
- Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., & Chen, J. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, 45(11), 1345–1352.
- Dunn, W. B., & Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4), 285–294. Available from <https://doi.org/10.1016/j.trac.2004.11.021>.
- Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelenka, E., Vaughan, A. A., Halsall, A., Harding, N., & Knowles, J. D. (2015). Molecular phenotyping of a UK population: Defining the human serum metabolome. *Metabolomics: Official Journal of the Metabolomic Society*, 11(1), 9–26.
- Evans, E. D., Duvallet, C., Chu, N. D., Oberst, M. K., Murphy, M. A., Rockafellow, I., Sontag, D., & Alm, E. J. (2020). Predicting human health from biofluid-based metabolomics using machine learning. *Scientific Reports*, 10(1), 1–13.
- Fraser, D. D., Bartha, R., Brown, A., Stewart, T. C., Daley, M., Dekaban, G. A., Doherty, T., Fischer, L., Holmes, J., & Menon, R. (2016). Metabolomics profiling of central nervous system injury. Google Patents.
- Ganna, A., Salihovic, S., Sundström, J., Broeckling, C. D., Hedman, Å. K., Magnusson, P. K., Pedersen, N. L., Larsson, A., Siegbahn, A., & Zilmer, M. (2014). Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. *PLoS Genetics*, 10(12), e1004801.
- Jové, M., Mauri-Capdevila, G., Suárez, I., Cambray, S., Sanahuja, J., Quílez, A., Farré, J., Benabdellah, I., Pamplona, R., & Portero-Otín, M. (2015). Metabolomics predicts stroke recurrence after transient ischemic attack. *Neurology*, 84(1), 36–45.
- Krumsiek, J., Mittelstrass, K., Do, K. T., Stückler, F., Ried, J., Adamski, J., Peters, A., Illig, T., Kronenberg, F., & Friedrich, N. (2015). Gender-specific pathway differences

- in the human serum metabolome. *Metabolomics: Official Journal of the Metabolomic Society*, 11(6), 1815–1833.
- Kumar, N., Hoque, M. A., & Sugimoto, M. (2018). Robust volcano plot: Identification of differential metabolites in the presence of outliers. *BMC Bioinformatics*, 19(1), 128. Available from <https://doi.org/10.1186/s12859-018-2117-2>.
- Lee-Sarwar, K. A., Lasky-Su, J., Kelly, R. S., Litonjua, A. A., & Weiss, S. T. (2020). Metabolome–microbiome crosstalk and human disease. *Metabolites*, 10(5), 181.
- LeWitt, P. A., Li, J., Lu, M., Guo, L., & Auinger, P. (2017). Metabolomic biomarkers as strong correlates of Parkinson disease progression. *Neurology*, 88(9), 862–869.
- Luo, P., Liu, Y., Qiu, L., Liu, X., Liu, D., & Li, J. (2020). Tocilizumab treatment in COVID-19: A single center experience. *Journal of Medical Virology*, 92(7), 814–818.
- Maenner, M. J., Shaw, K. A., & Baio, J. (2020). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016. *MMWR Surveillance Summaries*, 69(4), 1.
- Mapstone, M., Cheema, A. K., Fiandaca, M. S., Zhong, X., Mhyre, T. R., MacArthur, L. H., Hall, W. J., Fisher, S. G., Peterson, D. R., & Haley, J. M. (2014). Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine*, 20(4), 415–418.
- McGill, H. C., Jr, McMahan, C. A., & Gidding, S. S. (2008). Preventing heart disease in the 21st century: Implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) study. *Circulation*, 117(9), 1216–1227.
- Medina, S., Dominguez-Perles, R., Gil, J., Ferreres, F., & Gil-Izquierdo, A. (2014). Metabolomics and the diagnosis of human diseases-A guide to the markers and pathophysiological pathways affected. *Current Medicinal Chemistry*, 21(7), 823–848.
- Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). “Metabonomics”: Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica; the Fate of Foreign Compounds in Biological Systems*, 29(11), 1181–1189.
- Oertelt-Prigione, S., & Regitz-Zagrosek, V. (2011). Sex and gender aspects in clinical medicine. *Springer Science & Business Media*.
- Pauling, L., Robinson, A. B., Teranishi, R., & Cary, P. (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proceedings of the National Academy of Sciences of the United States of America*, 68(10), 2374–2376. Available from <https://doi.org/10.1073/pnas.68.10.2374>.
- Pinu, R. F., Goldansaz, A. S., & Jaine, J. (2019). Translational metabolomics: Current challenges and future opportunities. *Metabolites*, 9(6). Available from <https://doi.org/10.3390/metabo9060108>.
- Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., & Gautam, B. (2011). The human serum metabolome. *PLoS One*, 6(2), e16957.
- Purich, D. L., & Allison, R. D. (1999). *Handbook of biochemical kinetics: A guide to dynamic processes in the molecular life sciences*. Elsevier.
- Quanbeck, S. M., Brachova, L., Campbell, A. A., Guan, X., Perera, A., He, K., Rhee, S. Y., Bais, P., Dickerson, J. A., Dixon, P., Wohlgemuth, G., Fiehn, O., Barkan, L., Lange, I., Lange, B. M., Lee, I., Cortes, D., Salazar, C., Shuman, J., & Nikolau, B. J. (2012). Metabolomics as a hypothesis-generating functional genomics tool for the annotation of *Arabidopsis thaliana* genes of “Unknown Function”. *Frontiers in Plant Science*, 3, 15. Available from <https://doi.org/10.3389/fpls.2012.00015>, 15.

- Sabatine, M. S., Liu, E., Morrow, D. A., Heller, E., McCarroll, R., Wiegand, R., Berri, G. F., Roth, F. P., & Gerszten, R. E. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, 112(25), 3868–3875.
- Sato, S., Parr, E. B., Devlin, B. L., Hawley, J. A., & Sassone-Corsi, P. (2018). Human metabolomics reveal daily variations under nutritional challenges specific to serum and skeletal muscle. *Molecular Metabolism*, 16, 1–11.
- Schrime-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies-challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12), 1897–1905. Available from <https://doi.org/10.1007/s13361-016-1469-y>.
- Slupsky, C. (2012). Methods for diagnosis, treatment and monitoring of patient health using metabolomics. Google Patents.
- Spicer, R. A., Salek, R., & Steinbeck, C. (2017). A decade after the metabolomics standards initiative it's time for a revision. *Scientific Data*, 4(1), 1–3.
- Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., & Griffin, J. L. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics: Official Journal of the Metabolomic Society*, 3(3), 211–221.
- Trezzi, J., Galozzi, S., Jaeger, C., Barkovits, K., Brockmann, K., Maetzler, W., Berg, D., Marcus, K., Betsou, F., & Hiller, K. (2017). Distinct metabolomic signature in cerebrospinal fluid in early parkinson's disease. *Movement Disorders*, 32(10), 1401–1408.
- Troisi, J., Autio, R., Beopoulos, T., Bravaccio, C., Carraturo, F., Corrivetti, G., Cunningham, S., Devane, S., Fallin, D., & Fetissov, S. (2020). Genome, Environment, Microbiome and Metabolome in Autism (GEMMA) study design: Biomarkers identification for precision treatment and primary prevention of autism spectrum disorders by an integrated multi-omics systems biology approach. *Brain Sciences*, 10(10), 743.
- Troisi, J., Belmonte, F., Bisogno, A., Pierri, L., Colucci, A., Scala, G., Cavallo, P., Mandato, C., Di Nuzzi, A., Di Michele, L., Delli Bovi, A. P., Guercio Nuzio, S., & Vajro, P. (2019). Metabolomic salivary signature of pediatric obesity related liver disease and metabolic syndrome. *Nutrients*, 11(2). Available from <https://doi.org/10.3390/nu11020274>.
- Troisi, J., Cavallo, P., Masarone, M., Sepe, I., Scala, G., Campiglia, P., De Caro, F., Boccia, G., Ciacci, C., & Poto, S. (2020). Serum metabolomic profile of symptomatic and asymptomatic SARS-CoV-2 infected patients. *Research Square*.
- Troisi, J., Cavallo, P., Richards, S., Symes, S., Colucci, A., Sarno, L., Landolfi, A., Scala, G., Adair, D., Ciccone, C., Maruotti, G., Martinelli, P., & Guida, M. (2021). Non-invasive screening for congenital heart defects using a serum metabolomics approach. *Prenatal Diagnosis*. Available from <https://doi.org/10.1002/pd.5893>.
- Troisi, J., Landolfi, A., Sarno, L., Richards, S., Symes, S., Adair, D., Ciccone, C., Scala, G., Martinelli, P., & Guida, M. (2018). A metabolomics-based approach for non-invasive screening of fetal central nervous system anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 14(6), 77. Available from <https://doi.org/10.1007/s11306-018-1370-8>.
- Troisi J, Landolfi A, Vitale C, Longo K, Cozzolino A, Squillante M, Savanelli MC, Barone P, Amboni M. (2019). A metabolomic signature of treated and drug-naïve patients with Parkinson's disease: a pilot study. *Metabolomics*. 15(6), 90. <https://doi.org/10.1007/s11306-019-1554-x>. PMID: 31183578.

- Troisi, J., Pierri, L., Landolfi, A., Marciano, F., Bisogno, A., Belmonte, F., Palladino, C., Guercio Nuzio, S., Campiglia, P., & Vajro, P. (2017). Urinary metabolomics in pediatric obesity and NAFLD identifies metabolic pathways/metabolites related to dietary habits and gut-liver axis perturbations. *Nutrients*, 9(5), E485. Available from <https://doi.org/10.3390/nu9050485>, pii:..
- Troisi, J., Raffone, A., Travaglino, A., Belli, G., Belli, C., Anand, S., Giugliano, L., Cavallo, P., Scala, G., Symes, S., Richards, S., Adair, D., Fasano, A., Bottiglieri, V., & Guida, M. (2020). Development and validation of a serum metabolomic signature for endometrial cancer screening in postmenopausal women. *JAMA Network Open*, 3(9), e2018327. Available from <https://doi.org/10.1001/jamanetworkopen.2020.18327>.
- Troisi, J., Sarno, L., Landolfi, A., Scala, G., Martinelli, P., Venturella, R., Di Cello, A., Zullo, F., & Guida, M. (2018). Metabolomic signature of endometrial cancer. *Journal of Proteome Research*, 17(2), 804–812. Available from <https://doi.org/10.1021/acs.jproteome.7b00503>.
- Troisi, J., Sarno, L., Martinelli, P., Di Carlo, C., Landolfi, A., Scala, G., Rinaldi, M., D'Alessandro, P., Ciccone, C., & Guida, M. (2017). A metabolomics-based approach for non-invasive diagnosis of chromosomal anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 13(11), 140. Available from <https://doi.org/10.1007/s11306-017-1274-z>.
- Troisi, J., Scala, G., & Guida, M. (2017). Non-invasive diagnostic method for the early detection of fetal malformations. Google Patents.
- Troisi, J., Scala, G., Campiglia, P., Zullo, F., & Guida, M. (2018). Method for the diagnosis of endometrial carcinoma. Google Patents.
- Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., Feldstein, A. E., Britt, E. B., Fu, X., & Chung, Y.-M. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, 472(7341), 57–63.
- Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., & Coates, M. M. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053), 1459–1544.
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738. Available from <https://doi.org/10.1038/171737a0>.
- Wilkins, J. M., & Trushina, E. (2018). Application of metabolomics in Alzheimer's disease. *Frontiers in Neurology*, 8, 719.
- Zhao, S., Li, H., Han, W., Chan, W., & Li, L. (2019). Metabolomic coverage of chemical-group-submetabolome analysis: Group classification and four-channel chemical isotope labeling LC-MS. *Analytical Chemistry*, 91(18), 12108–12115.

SECTION

Data analysis

2

This page intentionally left blank

Techniques for converting metabolomic data for analysis

8

Jacopo Troisi^{1,2,3}, Sean M. Richards^{4,5}, Giovanni Troisi², and Giovanni Scala²

¹*Department of Medicine, Surgery and Dentistry “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy*

²*Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy*

³*Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy*

⁴*Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States*

⁵*Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States*

Introduction

The main purpose of metabolomics is to illustrate the complexity and variability of metabolic networks and chemical reactions occurring in and out of cells. These interactions represent the basis of functional or dysfunctional mechanisms of life matter. The previous chapters describe how to efficiently design a metabolomic experiment starting from the correct experimental design to identification of the most useful approach and through the collection of the biological samples and metabolome extraction ([Chapter 2: Experimental Design in Metabolomics](#)), the separation of metabolites ([Chapter 3: Separation Techniques](#)), and the subsequent quantitation ([Chapter 4: Mass Spectrometry in Metabolomics](#) and [Chapter 5: Nuclear Magnetic Resonance in Metabolomics](#)).

The result of these analyses is generally summarized in a matrix, (i.e., dataset), which is a table in which each column represents the concentration of a specific metabolite and each row represents all the information regarding a specific sample that belongs to a certain class (label). The dataset can be considered the middle land between analytical chemistry and bioinformatics, allowing effective communication between these different fields, both pivotal for metabolomics.

In most cases, the dataset is not usable as it is. For example, in order to train classification models, there is a need to pretreat this matrix in order ensure that this matrix contains as much information as possible and that this information is focused on the analyzed samples with a very low amount of noise (e.g., analytical or sampling variability).

Data analysts often say: “Garbage in—garbage out,” meaning poor quality data entry leads to unreliable results (Kilkenny & Robinson, 2018). For this reason, the data need to be highly defined. All the data preprocessing presented in this chapter, as well as other important strategies reported in Chapter 9, Data Analysis in Metabolomics: From Information to Knowledge, are integral in this sense. According to Sarih et al. (2019) data analysts spend more than 60% of their time reducing noise and organizing data. This illustrates the importance of proper data pretreatment.

The procedures that structure the dataset and pretreat it to make it suitable for data analysis can be divided into three phases.

1. Data preprocessing
2. Normalization and scaling
3. Transformation

Data preprocessing

Data preprocessing is a set of operations that are performed on the raw data in order to build the dataset. This is a critical step because the choices made in this phase represent the greatest source of variability in metabolomics experiments (Gross et al., 2018). Many efforts are being made to try to create a standard for these operations but this is challenging because the analytical instrumentation, metabolite extraction and purification methods, as well as the experimental design evolves very rapidly (Sumner et al., 2007). Improving results is possible through use of new algorithms and data analysis strategies; however, this is a difficult process to standardize.

The first step in data preprocessing is to divide according to detection and quantitation methods. For example, divide data into specific pretreatments for analyses that come from mass spectroscopy investigations and pretreatments for investigations that come from nuclear magnetic resonance.

Mass spectrometry-based experiments

Data obtained by means of mass spectroscopy require three preliminary modifications in order to be adequately structured for data analysis (Perez de Souza et al., 2017):

1. Peak picking, smoothing, and deconvolution (peak identification)
2. Alignment
3. Gap filling

Peak picking and Smoothing

As reported in Chapter 2, Experimental Design in Metabolomics, and Chapter 3, Separation Techniques, mass spectrometry is generally coupled to a chromatographic system. The analysis of the metabolome by means of such a combined system produces a chromatogram, which can be considered a two-dimensional graph in which

the time is shown on the *x*-axis, while the *y*-axis shows a quantitative variable that is attributable to the number of ions that pass the detector at a specific time. The number of ions produced at a specific time depends on the amount of metabolite that is reaching the detector but also on the amount of fragments that have been generated by each single molecule. For this reason, the area underlying each chromatographic peak that is attributable to a specific metabolite is not an absolute mirror of its concentration but can be converted into absolute concentration by means of a calibration curve (see Chapter 2, Experimental Design in Metabolomics).

In target metabolomics, the calibration curve can be built quite easily because the metabolites to be analyzed are known. On the contrary, in an untargeted metabolomics setting, the conversion in absolute terms of the area into concentration is more complicated. This is both because the number of metabolites that are studied is much higher, and because the list of these metabolites is not known.

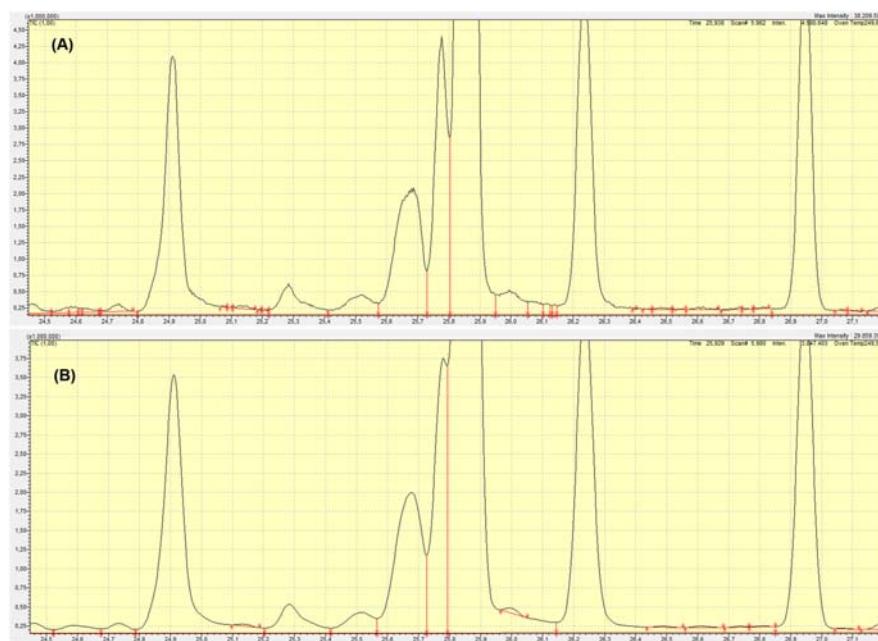
In the construction of the dataset, each single cell of the spreadsheet can be populated either with the absolute concentration of each metabolite for each single sample (if known) or, alternatively, with the area of the peak of that metabolite in each sample, which can be used as a concentration surrogate.

To determine the area of the peak, it is necessary to identify the start and end points, that is, the values in the time domain where the baseline changes from flat to an upward slope to begin forming the peak. This operation is known as peak picking (Bauer et al., 2011). Unfortunately, the signal that generates the chromatogram is subject to noise (elevation of the baseline of the chromatogram that is unrelated to metabolite concentration). Noise has different origins (e.g., the electronics of the detection systems, the thermal diffusions of the solvents, etc.) and can never be completely eliminated. Therefore, to optimize the identification of true metabolite peaks, it is good practice to apply a smoothing filter (Fig. 8.1).

There are different algorithms on which the various smoothing filters are based, but they all operate with a similar logic, reducing the background noise by mediating the contiguous signals. This operation also makes it possible to make the curve that draws the chromatographic peaks more homogeneous, making the calculation of the areas more reproducible. The most used filter is the Savitzky-Golay filter (Bromba & Ziegler, 1981). The figure (Fig. 8.1) reports a chromatographic plot before and after the Savitzky-Golay filter smoothing application. The application of the filter results in a softer peaks' shape. Moreover, the start and end peaks' points are estimated with greater precision, allowing a more accurate estimate of the areas. Furthermore, smoothing reduce noise reducing the identification of small area peaks.

Deconvolution

In biological samples there are many and varied metabolites. However, many belong to the same class (e.g., sugars, fatty acids, amino acids), so it is not uncommon for different metabolites in the same class to have very similar characteristics and therefore very similar chromatographic behavior (i.e., they are not adequately separated by chromatography). Sometimes these peaks partially or totally overlap (coeluted). The quantitation of such a peak would introduce an

**FIGURE 8.1**

Example of signal smoothing. (A) Original signal. (B) Smoothed chromatogram: peaks are more homogeneous and background noise was reduced. The three peaks in the middle from 25.6 to 26.0 were changed such that the resultant area under the curve has changed significantly.

error in the dataset. Fortunately, mass spectrometry can help to solve these drawbacks through fragmentation. Fragmentation occurs when the metabolite is passing through the ionization chamber of the mass spectrometer and is hit by a stream of electrons which removes an electron from the metabolite. This ionized metabolite is unstable and will subsequently spontaneously split into at least two pieces. Indeed, the coeluted metabolites, despite showing a similar chromatographic behavior, do not have the same molecular mass or, in any case, do not show the same spectrum after fragmentation. By analyzing signals from specific fragments, it is therefore possible to obtain the profile of the peaks corresponding to the different metabolites, even when those metabolites are not chromatographically separated. This is known as spectral deconvolution. There are many software options for deconvolution of entire chromatograms very quickly and efficiently, allowing for semiquantitation of a large number of metabolites ([Aksenov et al., 2021](#)) (Fig. 8.2).

Alignment

Another crucial operation of data preprocessing is alignment. The elution time of a molecule within a chromatographic system is based on molecular characteristics

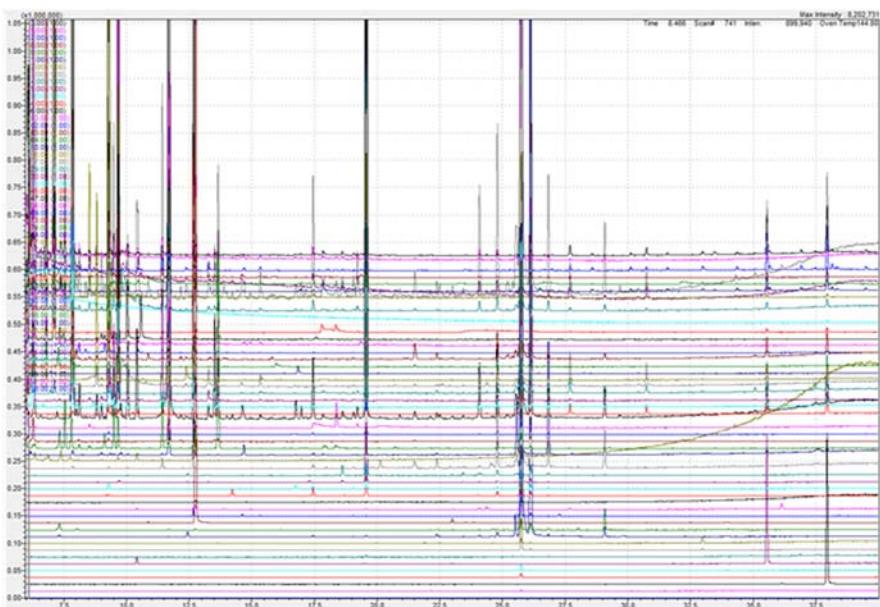


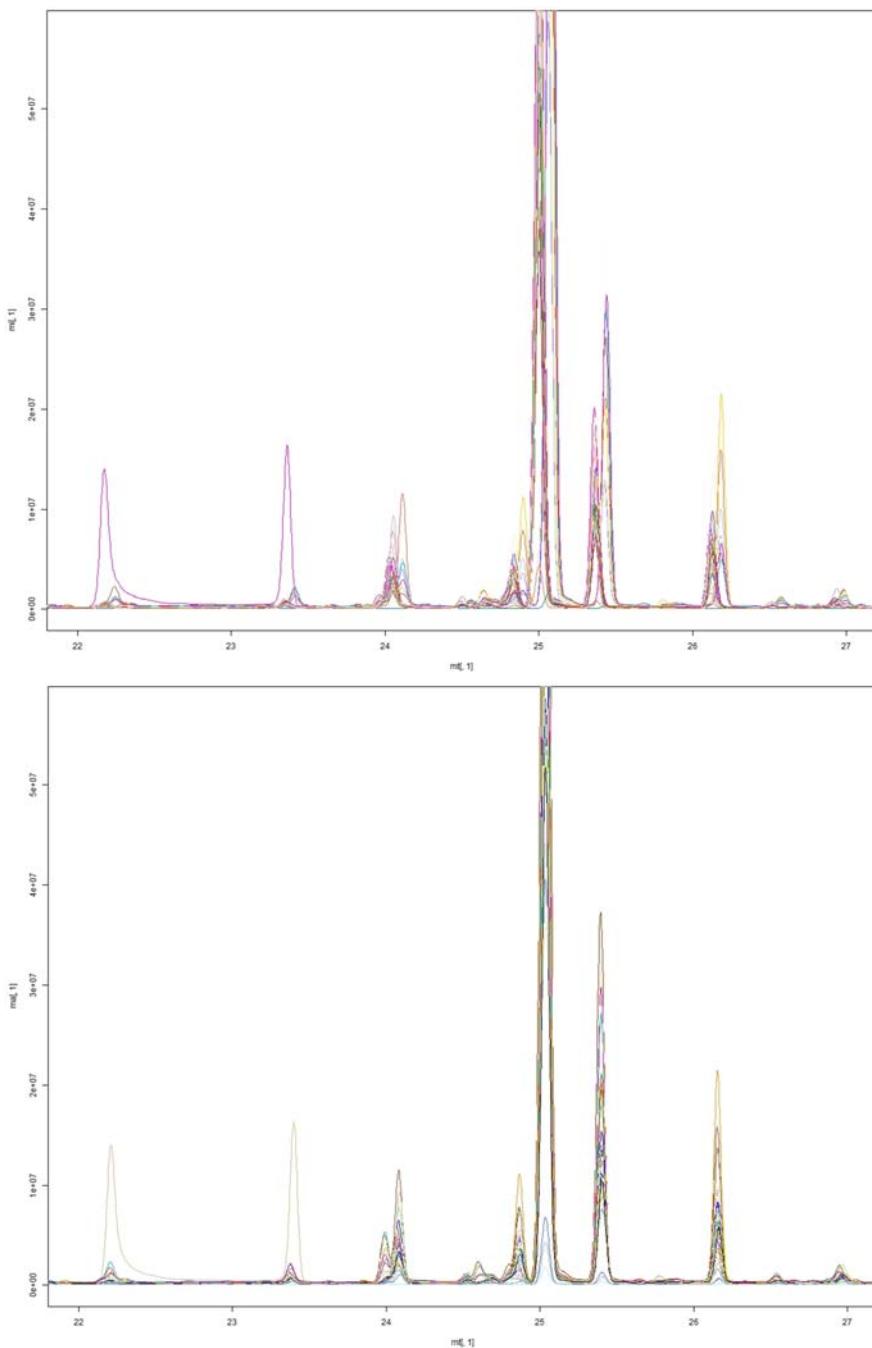
FIGURE 8.2

Chromatographic peak deconvolution. Specific mass to charge (m/z) signals were used to extract the peak profiles for each fragmented metabolite allowing the differentiation of the metabolites that were coeluted.

of that specific molecule and should be consistent. This is true in theory, but in practice, the chromatographic elution time is subject to error. Thus, a specific metabolite will not always have an identical elution time across samples. In addition to the measurement error, there are other mechanisms that increase the variability of the chromatographic elution time:

1. drift process, due for example to an imperfect synchronization between the injection of the sample onto the column and the start of the time count;
2. shrink process, due for example to the deterioration of the chromatographic column over time. As column efficiency deteriorates, its ability to retain molecules tends to decrease. This causes a decrease in elution times, both due to a lower interaction between metabolites and the column matrix and to a lower separation efficiency.

To align the chromatograms and ensure that the same metabolite produces the same time signal, the chromatograms must be subjected to an alignment process (Fig. 8.3). This is a mathematical operation that uses algorithms to correct for

**FIGURE 8.3**

Peaks alignment. Chromatographic signals from different samples are time-aligned using the Parametric Time Warping algorithm. Starting, ending and maximum peaks' points varied across samples, while alignment changes the time domain perfectly aligning these points.

drift. Thus, the algorithms take into account all variables that cause the metabolite to time-drift between samples and adjusts detection time to achieve a uniform elution time of a specific metabolite.

The most used algorithm is called “Parametric Time Warping” (PTW) (Wehrens et al., 2015). This identifies, among all the samples analyzed, the one that has an average behavior in terms of elution times compared to the other samples. This sample is called “Best Reference” and is used as a reference for subsequent alignment. At this point, all the other samples undergo a transformation of the time axis with respect to the time axis of the best referring sample. This compensates for the shift and compression effect. This process, indispensable in the preprocessing of data for mass spectrometry experiments, can also be used on data obtained by nuclear magnetic resonance.

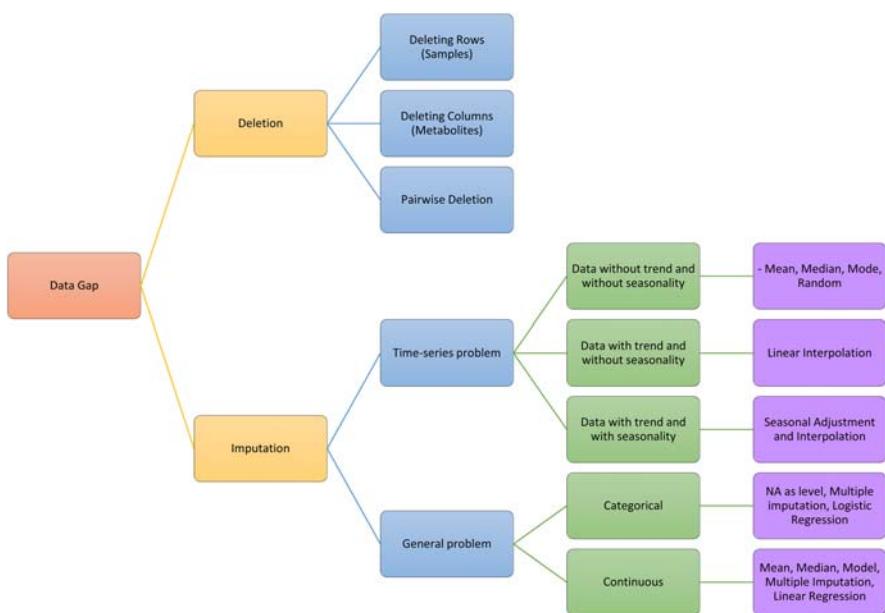
Gap filling

The alignment process can generate gaps in some chromatograms, as in the case of a sample whose chromatographic run started before the best reference sample. The alignment process, moving all the chromatograms of this sample forward, introduces a data gap in the first moments of acquisition. These missing points cannot be sent to the subsequent data analysis process because any algorithm would have difficulty in interpreting this gap. For this reason, these missing data must be managed and purged from the matrix before subjecting it to further investigation.

The missing data treatment process can follow two paths: **deletion** and **imputation** (see the decision tree, Fig. 8.4) (Karp et al., 2018). In deletion, the missing data are deleted. It is possible to delete the rows (the samples with missing data) or the columns (the missing signals, eliminating them from all the samples); alternatively, both can be deleted. This operation is protective for the information quality contained in the dataset but obviously eliminates part of the information acquired. A more conservative alternative is imputation, which is a strategy that estimates missing data. This estimation can be based on the mean, on the median, or on the data evaluation by means of a logistic regression by analyzing the data adjacent to the missing ones. The choice between deletion and imputation is often not simple. The consequence of reducing information must always be weighed against the consequence of including imperfect data.

Nuclear magnetic resonance

Nuclear magnetic resonance investigations give rise to a signal similar to the chromatographic one. Indeed, NMR signals can be represented by means of a Cartesian graph in which the resonance frequency is on the x axis and the intensity is on the y one. For this reason, some aspects of data preprocessing we have seen for MS-coupled techniques are applicable to NMR signals.

**FIGURE 8.4**

Missing value management. This decision tree shows the strategy to manage the missing data. Both deletion and imputation are viable alternatives.

However, there are some specific operations NMR signals need, in particular:

1. Water signal elimination
2. Chemical shift calibration
3. Binning

Water signal elimination

As it has been largely illustrated in [Chapter 5](#), Nuclear Magnetic Resonance in Metabolomics, most of the NMR applications are related to investigations on the ^1H nucleus. Since all biological samples are water-based, the water hydrogen signal is the highest signal in any NMR profile of metabolomic interest. This signal abundance presents a twofold problem. First of all, a signal of this intensity tends to mask all the other signals due to a scale issue (see [Fig. 8.5](#)).

Indeed, since the water signal is a few orders of magnitude more intense than the ^1H signals coming from the metabolites, the latter tend to disappear on a display scale. Unfortunately, a simple change of scales (zooming in the area of interest out of the H_2O signal) does not solve this issue, because the water tends to absorb all the signal generated to perturb the nuclear orientation induced by the magnetic field, thus not allowing the generation of ^1H signals of the metabolites.

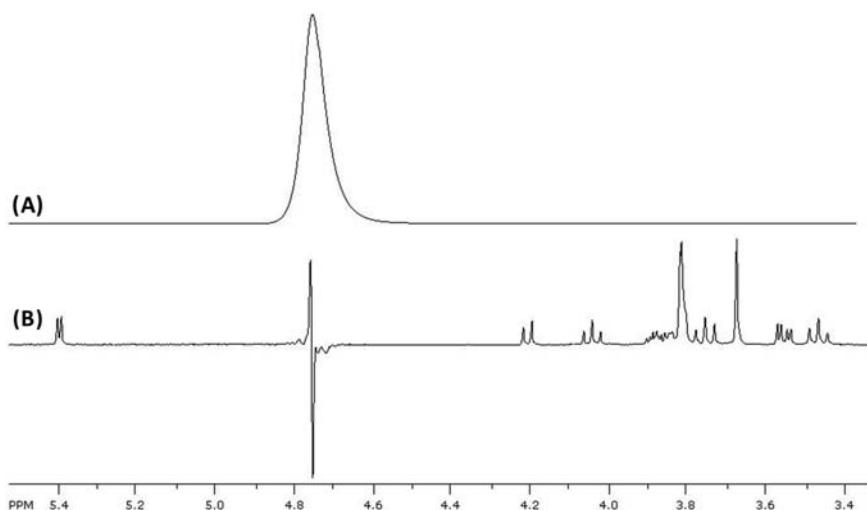


FIGURE 8.5

NMR-water signal elimination. (A). NMR spectra illustrating how a water signal may fully hide metabolite signals. (B). NMR spectra illustrating water signal presaturation. Metabolite signals appear while water signal is eliminated.

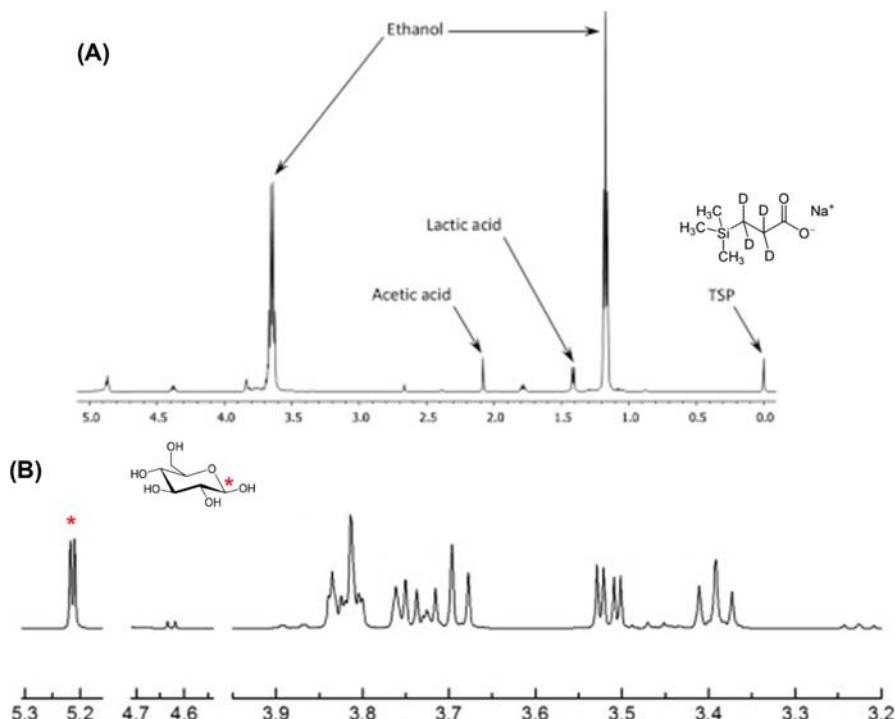
To overcome this drawback, it is possible to proceed with a presaturation of the signal. In other words, the sample is flooded with an orthogonal signal that saturates the absorption of water hydrogens which will therefore no longer be able to immediately generate a signal. In this fraction of time, in which a sort of “shocking” remains due to the presaturation of the signal, the signals of metabolomic interest are scanned. The effects of the signal presaturation are visible in Fig. 8.5 “B”

Chemical shift calibration

Every single ^1H of each metabolite generates a signal within a specific range of the spectrum. A reference must be used in order to assign an exact value to these specific field values. Typically, in NMR investigations, a molecule (trimethylsilylpropanoic acid or TSP) is added to each sample (Fig. 8.6 “A”). TSP generates a single signal to which a zero value is assigned. Metabolomic experiments also allow the use of a different strategy; indeed, almost all biological samples, from which the metabolome is extracted, contain glucose. Glucose generates a double-peaked signal at approximately 5.23 ppm (Fig. 8.6 “B”). This signal can then be used as an internal reference to calibrate the other shifts.

Binning

Binning is essentially a division of an NMR spectrum into several smaller areas and the subsequent building of the dataset using an estimation of the average behavior of the spectrum in a certain interval called the binning interval (Emwas

**FIGURE 8.6**

Chemical shift calibration in NMR spectra (A). TPS (trimethylsilylpropanoic acid) hydrogen signal could be used to mark the 0 ppm value. (B). In metabolomics experiments the double signal of the hydrogen marked with the red star could be used.

et al., 2018). The datasets that come from mass spectrometry investigation, as mentioned above, are set up in such a way as to report the quantitative or semi-quantitative information of a metabolite in each column. This can be expressed in terms of the area underlying the chromatographic peak, thus representing a single value per metabolite. However, in nuclear magnetic resonance, the entire profile obtained from a sample is used as a spectrum. In other words, it is as if each single value determined along this spectrum should represent a separate cell in the dataset matrix. This would generate very unbalanced, large datasets that contain a huge number of columns and a small number of rows (the observed samples). These very large datasets, as we will discuss in Chapter 9, Data Analysis in Metabolomics: From Information to Knowledge, often give rise to overfitting phenomena, that is, they tend to over-train the classification models and therefore tend to be inaccurate in prediction. To try to mitigate this effect, the information of the NMR spectra can be compressed. Through the binning process, the spectrum is separated into many more or less small pieces and only the integral value

(the area) under the spectrum in this interval is introduced in the dataset. This reduces the size of the dataset by reducing the number of columns. This procedure, in addition to reducing the dimensionality of the data entering the dataset, also contributes to a reduction of the misalignment effects due to small variations in pH existing in different samples derived from the same matrix.

Normalization

Normalization is a process of attenuating the mean difference of the metabolites in the samples (De Livera et al., 2012). For example, urine samples can be concentrated (e.g., early in the morning) or diluted (e.g., if obtained from a subject who has drunk a lot of water before collection) (see Fig. 8.7). This variation in the global concentration of metabolites is irrelevant because in metabolomics we are interested in the relationships that metabolites establish among themselves rather than in their absolute concentration. However, the metabolite absolute concentrations affected as a function of external conditions must be normalized to minimize bias.

There are several strategies for performing normalization. In the case of urine, creatinine or osmolarity can be used as a parameter by which to divide the area of each single metabolite to normalize it with respect to a parameter strictly dependent on the degree of hydration of the subject at the time of sampling.

In the case of metabolomic investigations on cell cultures, normalization can be conducted by dividing by the number of analyzed cells, or by the quantity of DNA or proteins of the analyzed sample.

Internal standard normalization

Chromatographic investigations, especially those conducted in gas chromatography, also have a further source of uncertainty regarding the total quantity of metabolites



FIGURE 8.7

Normalization. The concentration of urine samples could be due to a subject hydration. Normalization minimizes these differences.

detected. This is due to the unknown variability and efficiency of injecting the sample into the separatory column and quantifying the samples (see [Chapter 2: Experimental Design in Metabolomics](#) and [Chapter 3: Separation Techniques](#)). To mitigate this effect, the internal standard technique is often used. This is facilitated by introducing a known quantity of a molecule known not to be present in the sample. A reduction in the detected concentration of sample (due to error or unknown reasons) results in a smaller area underlying the peak of each metabolite. This can be considered a false-negative because less of the molecule is detected than is actually present. However, the area under the peak of an internal standard will also be reduced by the same percentage. Because the amount of internal standard is known, the investigator can adjust the falsely low areas under all metabolite peaks according to the internal standard. Each area can be normalized (divided) by the area of the internal standard, effectively reducing variability and correcting for the falsely low (or high) area. (see [Fig. 8.8](#)).

Normalization with respect to the total area, estimated as the sum of all detected metabolites is another method of using an internal standard to address the variability caused by the inconsistent quantities of sample introduced into the chromatographic system. This inconsistency could be a consequence of diluted vs. concentrated urine, for example, or differing cell quantity when investigating cell culture metabolites. This strategy, although simple to apply and generally reliable, suffers from estimation errors when the different chromatograms do not show the same number of peaks, for example, when some metabolites are present only in some samples. There are other more sophisticated normalization tools; in

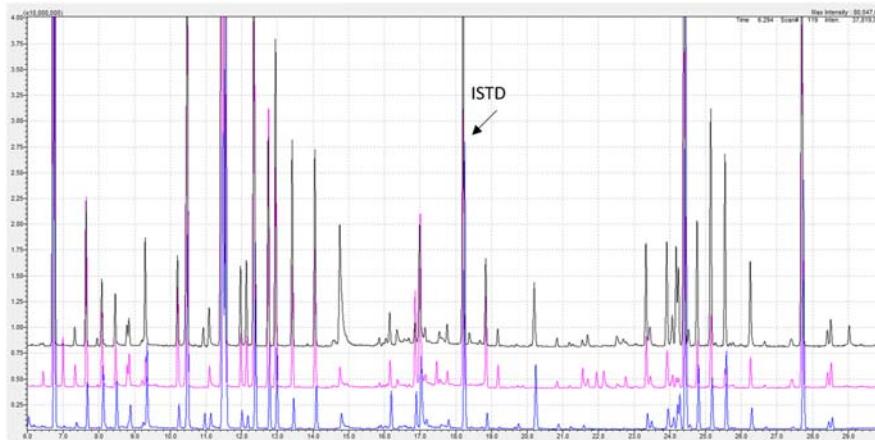


FIGURE 8.8

Internal standard normalization. A known amount of a molecule(s) not naturally occurring in the biological samples is injected with the biological sample. If the area under the internal standard is erroneous, all metabolite areas are adjusted by dividing by the area under the internal standard.

particular, two are widely used in metabolomics: probabilistic quotient normalization (PQN) and normalization on the quantile (see below).

Probabilistic quotient normalization

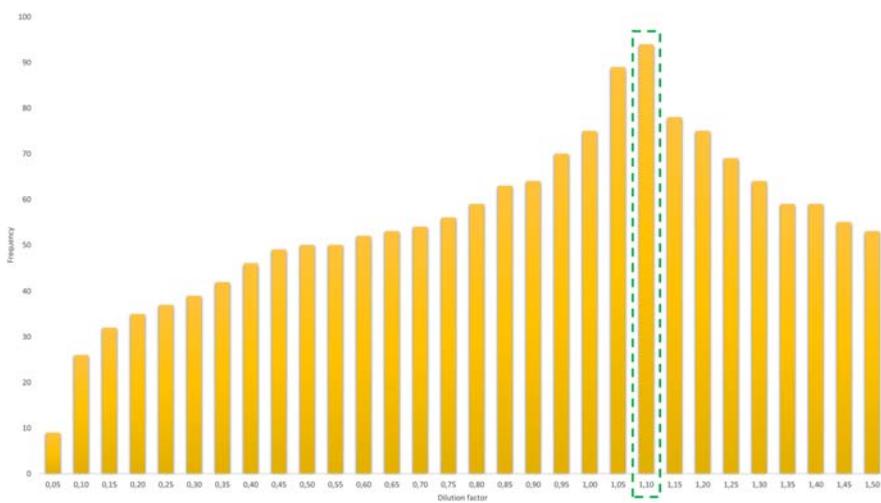
PQN accounts for different dilutions of samples by adjusting the chromatogram to the same virtual overall concentration (Dieterle et al., 2006). This method considers the distribution of the quotients of the amplitudes of a test spectrum and calculates a most probable dilution factor by comparing the test spectrum to a reference spectrum (Dieterle et al., 2006). This estimate is not made by comparing the total areas, but areas for each metabolite, and the most probable factor is then estimated as the one with the highest mode (frequency of occurrence). Specifically, the steps to be taken are the following:

1. Find the best reference in terms of signal value. To do this, a robust solution is to create a synthetic signal estimated as the average of each single peak or sampling point. The best reference is represented by the sample whose intensities are closest to this synthetic sample;
2. Estimate the dilution factors for each individual peak or signal. The ratio between each peak of a specific chromatogram/NMR spectrum and the corresponding peak of the best reference is calculated;
3. Estimate the best dilution quotient for each sample. For each sample, among the various calculated quotients, the most probable is estimated by determining the mode of the different quotients. To do this, it might be useful to aggregate the quotients. See the image Fig. 8.9
4. The normalized matrix is created by multiplying each peak/signal by the most probable quotient.

This strategy is more robust than normalizing on a total area because the less abundant signals tend not to change linearly with the total increase in metabolite concentrations. There are several reasons for this, but the most important is due to baseline fluctuations which can interfere with reading small minor peaks. Similarly, very abundant signals can be subject to undersizing due to signal saturation. The PQN method minimizes the error in estimating the dilution factor due to these extremes.

Quantile normalization

Quantile normalization forces all samples in a sample set to have identical peak intensity/area distribution. The difference of quantile normalization from the PQN previously described is that there is no estimated normalization factor for each sample. First, each row of the data table is sorted from smallest to greatest metabolite concentration. Thereafter, the mean/median of each column is calculated. This provides an estimation of the range of the new distribution (see Fig. 8.10). All rows of the data table are replaced with the estimate means. Finally, the data table is restored into its original order before sorting. The rows of the new data table are composed of the normalized samples. This method can be problematic with high-concentration metabolites in the data table because they can dramatically differ from sample to sample.

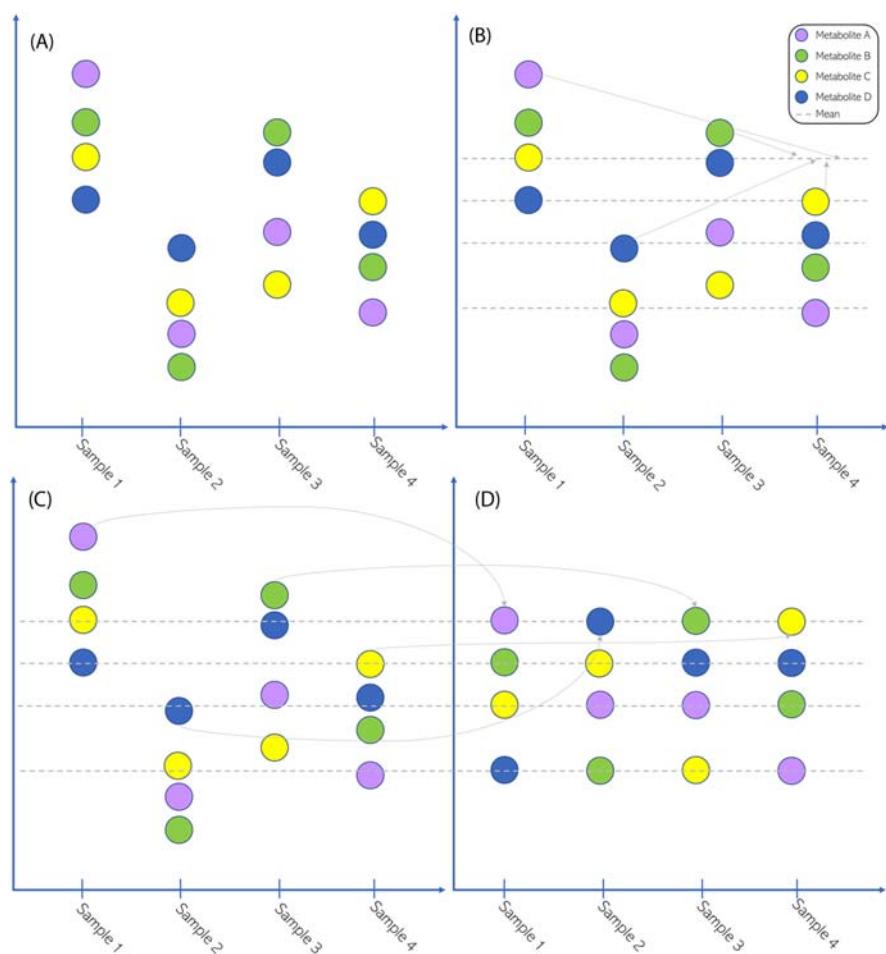
**FIGURE 8.9**

Probabilistic quotient normalization. This method accounts for different dilutions of samples. For each sample the dilution factor was estimated for each metabolite, comparing its concentration with the concentration of the same metabolite in a reference sample. The dilution factor with the highest mode was selected to normalize that sample.

It is important to note that the relative amount of each metabolite for each sample is not affected by the normalization. Indeed, all the signals are still reported in the original order. QN is a valuable normalization technique used in several omics strategies. Nevertheless, it is prone to class-effect and batch-effect due to the proportion of class-correlated variables in a dataset, causing higher false-positive and false-negative classifications. [Zhao et al. \(2020\)](#) recently suggested to first split data by sample class before performing quantile normalization independently on each split. This class-specific strategy effortlessly outperforms whole-data quantile normalization and is robust.

Data pretreatment

Machine learning, in particular, classification model training, is one of the most important steps in metabolomic data analysis. The usefulness of these models is not only related to their potential diagnostic role, that is, their ability to identify a pathological condition or a specific phenotype using the global metabolomic profile, but these algorithms, by means of analyzing mechanisms, can give insight that captures the complexity underlying the development of specific conditions. Indeed, machine learning can ultimately identify relevant metabolites to subject to further investigations.

**FIGURE 8.10**

Quantile normalization. The metabolite concentrations of each sample are first ordinated (A), then the mean of the most concentrated metabolites among the several samples is calculated (B). Each metabolite raw concentration is replaced with the mean value (C and D). The process is then repeated for the next most concentrated metabolite until all metabolites have been normalized.

Classification algorithms are prone to overestimate the role of metabolites whose concentration or variance is high. The concentration of the metabolites in biological fluids is extremely variable. Some are present in concentrations of several grams per liter, while others have concentrations on the order of a few micrograms per liter. Consequently, the variance of these concentrations is generally high. To manage this effect, and therefore to equalize the chance of all

metabolites to be selected as relevant by the classification algorithms, regardless of their absolute concentration, the data synthesized in the datasets should undergo two important preprocessing methods:

1. Centering
2. Scaling

These two processes are similar to normalization but, unlike normalization, centering and scaling do not differentially affect samples, only the different metabolites. Indeed, while normalization is carried out on the rows, tending to standardize the global concentrations of the various samples, centering and scaling works on columns, thus influencing the average and the distribution of semiquantitative data of the different metabolites in the entire dataset.

Centering

As the name suggests, centering is preprocessing that centers data with respect to a certain value. The most useful value on which to center the concentrations of the different metabolites is zero. To center the concentration of a metabolite, it is sufficient to estimate the mean (or the median, or another estimator of the central tendency) of that metabolite in the entire dataset and subtract this value from each single measurement of that metabolite. This operation has no effect on the variance but only on the mean. The figure (Fig. 8.11) shows the effect on the data distribution of this preprocessing.

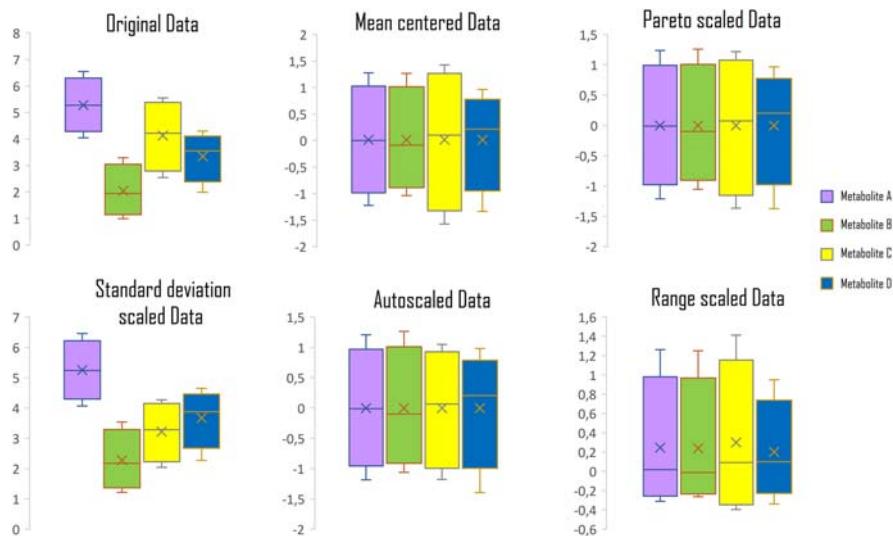


FIGURE 8.11

Centering and scaling. The figure illustrates the effect of the most used centering and scaling algorithms in metabolomics.

Scaling

The scaling process acts on the variance of the concentration distribution (or signal intensity) data. There are several ways to scale data. The figure (Fig. 8.11) shows the effect of the most common methods used in metabolomics experiments.

Scaling methods use a dispersion measure for scaling; the simplest scaling system consists in dividing each value related to the semiquantitative data of the different metabolites by the standard deviation of all these values.

Autoscaling is the most common scaling system in metabolomics experiments. This includes both a centering around zero and a scaling on standard deviation at the same time. Thus, not only the averages but also the variances are standardized. Autoscaled data always have zero mean and unit variance.

Pareto scaling is also based on mean centering, while scaling is based on the square root of the standard deviation. Range scaling, on the other hand, uses the range (maximum value minus minimum value) as a normalizer.

Generally, it is not easy to choose which scaling and centering system to use. Different pretreatment methods emphasize different aspects of the data and each pretreatment method has its own advantages and drawbacks (see Table 8.1). The choice for a pretreatment method generally depends on the biological question to be answered, the properties of the data set and the data analysis method selected. Nevertheless, often the choice is made to try different systems and then selecting the one maximizing the experimental aim. To that end, [van den Berg et al. \(2006\)](#) tested several centering and scaling processes on real-life metabolomic datasets, showing autoscaling and range scaling performed better than the other pretreatment methods. Moreover, according to their results, range scaling and autoscaling were able to remove the dependence of the range of the metabolites on the average concentration and the magnitude of the fold changes and showed biologically sensible results after PCA (principal component analysis).

Transformation

Despite the efforts to center the data and to equalize the variances, biological data are often not symmetrically distributed around the mean. In mathematical terms, they are said to be not-normal distributed. There are several reasons for this. To simplify, it is sufficient to think that it is more likely that metabolites have very low concentrations rather than very high so that this distribution of concentrations is skewed to the left. Furthermore, there are also reasons more closely related to the analytical process that make the data skewed ([van den Berg et al., 2006](#)).

This represents an impediment to data analysis for several reasons. Under conditions of non-normality, for example, it is not possible to use parametric statistical methods which are often more robust than nonparametric ones. Moreover, in biology, relationships between variables are not necessarily additive but can also be multiplicative ([Rohlf & Sokal, 1981](#)), and data distribution affects the ability

Table 8.1 Data preprocessing strategies—overview of the most used pretreatment in metabolomic studies.

Pretreatment	Formula	Aim	Advantages	Disadvantages
Centering	$\check{x}_i = x_{ij} - \bar{x}_i$	Remove the offset from the data	Allows an unbiased comparison between metabolites with a large concentration difference	Does not correct not-normal data distribution as well as heteroscedastic
Autoscaling	$\check{x}_i = \frac{x_{ij} - \bar{x}_i}{\sqrt{\frac{\sum_{j=1}^J (x_{ij} - \bar{x}_i)^2}{J-1}}}$	Normalize means and variations of metabolites	Allows comparison of metabolites with different means and variations, facilitating the application of methods based on correlations	Susceptible to outliers and analytical errors
Range scaling	$\check{x}_i = \frac{x_{ij} - \bar{x}_i}{x_{max} - x_{min}}$	Normalize means and range of metabolite concentration	Scaling is based on biological variability	Only 2 values (minimum and maximum) directs the scaling, making it extremely vulnerable to outliers
Pareto scaling	$\check{x}_i = \frac{x_{ij} - \bar{x}_i}{\sqrt{\frac{\sum_{j=1}^J (x_{ij} - \bar{x}_i)^2}{J-1}}}$	Normalize means and variations of metabolites	Reduces the relative importance of large values, keeping data structure closer to the original	Sensitive to large fold changes
Log transformation	$\check{x}_i = \log x_{ij}$	Reduce heteroscedasticity, increase data normality.	Correct for heteroscedasticity, adding a pseudo-scaling effect	Not able to manage zero values
Generalized log transformation	$\check{x}_i = \log(x_{ij} + \sqrt{x_{ij}^2 + \lambda})$	Reduce heteroscedasticity, increase data normality	Correct for heteroscedasticity, adding a pseudo-scaling effect. Able to manage zero values	Difficulties with values with large relative standard deviation
Power transformation	$\check{x}_i = \sqrt{x_{ij}}$	Reduce heteroscedasticity, increase data normality	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

to investigate such relationships. Data transformation is then necessary to investigate correlations using linear techniques.

Data transformation is a mathematical operation aimed at converting a population of not-normally distributed data (i.e., not conforming to the Gaussian distribution) into a population of normally distributed data. There are different transformation techniques, but the most used in metabolomics are log transformations, power transformations, and generalized log transformations (see Fig. 8.12).

In addition to the effect on the distribution of data which, as mentioned, tends to increase symmetry with respect to the average, the transformation operations have a pseudo-scaling effect due to the reduction of high values into relatively smaller values. However, this effect is not determined by a multiplication with a scale factor (as in real scaling process) but is due to a natural compression effect that is introduced by the transformation.

For this reason, this pseudo-normalization effect is generally not sufficient to completely regulate the differences in concentration magnitude, and scaling remains a crucial operation in the pretreatment of data deriving from metabolomic investigations. The combined effect of the different pretreatments is not easily predictable, so it is good practice to test the effectiveness of these processes not only independently but also in a synergistic manner.

The most used transformation is the logarithmic transformation that is also generally very effective in reducing heteroskedasticity if the relative standard deviation is constant (Kvalheim et al., 1994). Unfortunately, this is rare in experimental situations. Moreover, a serious drawback of the logarithmic transformation

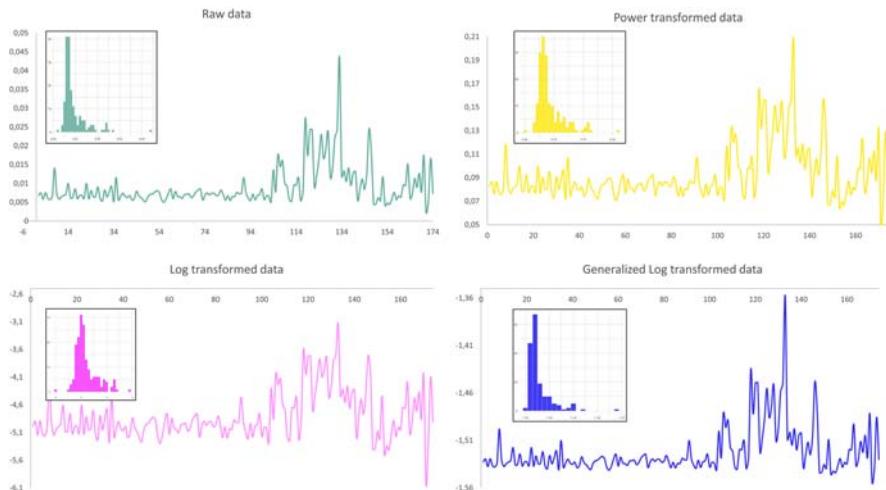


FIGURE 8.12

Data transformation—the figure illustrates the effect of the most used transformation processes used in metabolomics.

lies in its inability to handle null values. This can be managed in two ways: either by using a generalized transformation function that introduces an arbitrary lambda value that reduces this risk, or by acting on the preprocessing operations so as not to have zero values that can be replaced by estimators of the instrumental detection limit. A further drawback of the logarithmic transformation emerges when operating on values with a large analytical relative standard deviation. This is often found in the data of metabolites with a relatively low concentration, as these deviations are emphasized. These problems occur because the log transformation approaches negative infinity as the value to be transformed approaches zero.

Power transformation does not exhibit these drawbacks and also has positive effects on heteroskedasticity. The power transformation shows a transformation pattern similar to the logarithmic transformation, although it is often less effective.

Conclusion

Metabolomic data pretreatment represents a critical step in information mining. It directly influences the biomarker discovery, the learning ability of the classification models and, in general, the ability to generate knowledge from these data. In this chapter, the main techniques for data pretreatment are described. The main aim of data pretreatment is focalizing data analysis on the biologically relevant information. This step is generally time consuming. Moreover, it is pivotal to carefully evaluate each choice as well as each methodology used in this phase since, due to its early position in the metabolomic experiment pipeline, data pretreatments will have an influence on all the following steps (data analysis, relevant metabolites selection, pathways analysis, and so on).

References

- Aksenov, A. A., Laponogov, I., Zhang, Z., Doran, S. L. F., Belluomo, I., Veselkov, D., et al. (2021). Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nature Biotechnology*, 39(2), 169–173. Available from <https://doi.org/10.1038/s41587-020-0700-3>.
- Bauer, C., Cramer, R., & Schuchhardt, J. (2011). Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods in Molecular Biology* (Clifton, N.J.), 696, 341–352. Available from https://doi.org/10.1007/978-1-60761-987-1_22.
- Bromba, M. U. A., & Ziegler, H. (1981). Application hints for Savitzky-Golay digital smoothing filters. *Analytical Chemistry*, 53(11), 1583–1586. Available from <https://doi.org/10.1021/ac00234a011>.
- De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., et al. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry*, 84(24), 10768–10776. Available from <https://doi.org/10.1021/ac302748b>.

- Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ¹H NMR metabonomics. *Analytical Chemistry*, 78(13), 4281–4290. Available from <https://doi.org/10.1021/ac051632c>.
- Emwas, A.-H., Sacceti, E., Gao, X., McKay, R. T., Dos Santos, V. A. P. M., Roy, R., & Wishart, D. S. (2018). Recommended strategies for spectral processing and post-processing of 1D (¹H)-NMR data of biofluids with a particular focus on urine. *Metabolomics: Official Journal of the Metabolomic Society*, 14(3), 31. Available from <https://doi.org/10.1007/s11306-018-1321-4>, 31.
- Gross, T., Mapstone, M., Miramontes, R., Padilla, R., Cheema, A. K., Maciardi, F., et al. (2018). Toward reproducible results from targeted metabolomic studies: Perspectives for data pre-processing and a basis for analytic pipeline development. *Current Topics in Medicinal Chemistry*, 18(11), 883–895. Available from <https://doi.org/10.2174/1568026618666180711144323>.
- Karp, P. D., Weaver, D., & Latendresse, M. (2018). How accurate is automated gap filling of metabolic models? *BMC Systems Biology*, 12(1), 73. Available from <https://doi.org/10.1186/s12918-018-0593-7>.
- Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: “Garbage in—garbage out.” *Health Information Management Journal*, 47(3), 103–105. Available from <https://doi.org/10.1177/1833358318774357>.
- Kvalheim, O. M., Brakstad, F., & Liang, Y. (1994). Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66(1), 43–51.
- Perez de Souza, L., Naake, T., Tohge, T., & Fernie, A. R. (2017). From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web resources for mass spectral plant metabolomics. *GigaScience*, 6(gix037). Available from <https://doi.org/10.1093/gigascience/gix037>.
- Rohlf, F. J., & Sokal, R. R. (1981). *Biometry: The principles and practice of statistics in biological research*. New York: Freeman.
- Sarih, H., Tchangani, A., Medjaher, K., & PERE, E. (2019). Data preparation and preprocessing for broadcast systems monitoring in PHM framework. Available from <https://doi.org/10.1109/CoDIT.2019.8820370>.
- Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics: Official Journal of the Metabolomic Society*, 3(3), 211–221. Available from <https://doi.org/10.1007/s11306-007-0082-2>.
- van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7(1), 142. Available from <https://doi.org/10.1186/1471-2164-7-142>.
- Wehrens, R., Bloemberg, T. G., & Eilers, P. H. C. (2015). Fast parametric time warping of peak lists. *Bioinformatics (Oxford, England)*, 31(18), 3063–3065. Available from <https://doi.org/10.1093/bioinformatics/btv299>.
- Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports*, 10(1), 15534. Available from <https://doi.org/10.1038/s41598-020-72664-6>.

This page intentionally left blank

Data analysis in metabolomics: from information to knowledge

9

Jacopo Troisi^{1,2,3}, Giovanni Troisi², Giovanni Scala², and Sean M. Richards^{4,5}

¹Department of Medicine, Surgery and Dentistry “Scuola Medica Salernitana”, University of Salerno, Baronissi, Salerno, Italy

²Theoreo Srl—Spin-off Company of the University of Salerno, Montecorvino Pugliano, Salerno, Italy

³Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano, Salerno, Italy

⁴Department of Biological and Environmental Sciences, University of Tennessee-Chattanooga, Chattanooga, TN, United States

⁵Department of Obstetrics and Gynecology, College of Medicine, University of Tennessee Health Science Center, Chattanooga, TN, United States

Introduction

We are living in the big data epoch. Omics techniques allow biologists to join the big-data club (Marx, 2013).

The processing and structuring of data, as described in the previous chapter, leads to the construction of information. This, however, alone, is not useful for unraveling the mysteries of life that still fascinate researchers. Science needs knowledge. Knowledge is a combination of information, experience, and intuition (see Fig. 9.1). Data analysis is the process for converting data into information first and then into knowledge. In metabolomics, this knowledge extraction process is generally structured in 3 steps:

- Exploratory analysis;
- Unsupervised multivariate analysis;
- Supervised multivariate analysis.

Exploratory analysis

Before they can be used for any type of analysis, data generated by a metabolomics experiment must be structured in a dataset. This is a representation of the data in the form of a relationship matrix. In other words, the dataset corresponds to the content of a single table or matrix, in which each column of the table represents a particular variable (quantitative or semi-quantitative data of a

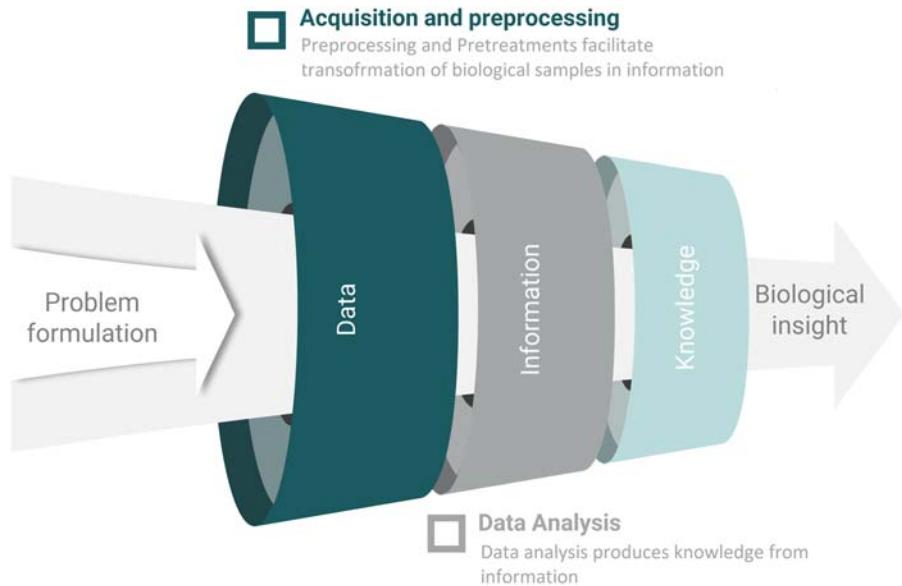


FIGURE 9.1 The knowledge creation process.

Data acquisition and pretreatments facilitate transformation of biological samples in information. Data analysis produces knowledge from information.

single metabolite), and each row corresponds to a certain member of the studied population (biological sample analyzed) (see also [Chapter 8: Techniques for Converting Metabolomic Data for Analysis](#)). Generally, the first column contains the sample identification, while the second column contains the sample label. These are two very important sample parameters used to univocally identify it (samples ID) and to assign it to a specific condition also known as “class.” A simple example of class separation is “healthy subjects” and “disease affected subjects” (aka “control subjects” and “case subjects”). Diseases show several forms (acute/chronic, symptomatic/asymptomatic, etc.); therefore, different labeling systems could be used during the dataset building and these choices will direct subsequent data analysis.

Exploratory data analysis (EDA) is an approach of analyzing datasets to summarize their main characteristics. Several strategies can be used to perform EDA ([Fig. 9.2](#)).

EDA was proposed by John W. Tukey in 1962 defining it as: “A procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data” (Tukey, 1962).

EDA is a precious resource in metabolomics data analysis because although it was not specifically designed to catch the complexity hidden in these kinds of data, it allows for data analysis beyond the formal modeling or hypothesis testing or extraction (Tukey, 1977). Moreover, curiously, Tukey’s work to develop EDA principles and strategies pushed the development of S, a statistical computing packages at Bell Labs. The S programming language inspired the systems

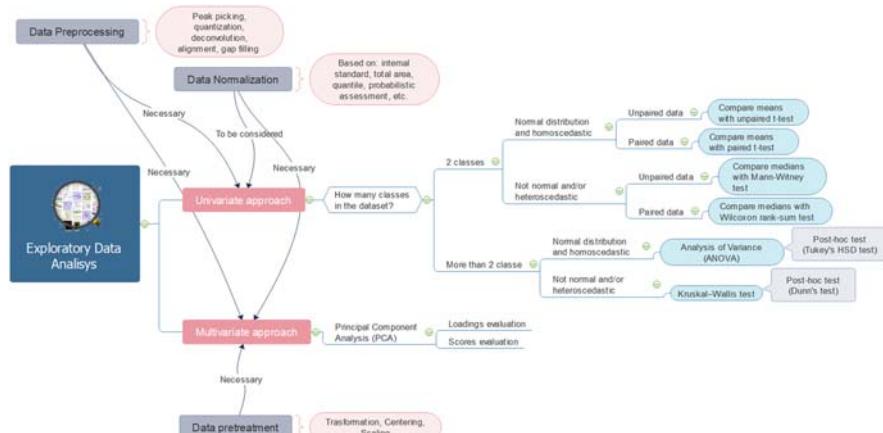


FIGURE 9.2 Exploratory analysis of metabolomics data.

Exploratory analysis in metabolomics can be performed following two separate routes: univariate and multivariate analysis. The flowchart of the two routes is represented.

S-PLUS and R. Currently R ([RDevelopment CORE TEAM, 2008](#)) is the more utilized tool for the analysis of metabolomics data both in terms of EDA and in terms of machine learning (ML).

Generally, in metabolomics, two separate routes are followed to perform EDA: univariate and multivariate.

- The univariate approach allows analysis of each metabolite, taken in a unique and independent manner from the others. Although it is a useful and very simple method to implement, it is not the most natural approach to analyzing omics data. In fact, as reported in [Chapter 1, System Biology](#), and [Chapter 7, Approaches in Untargeted Metabolomics](#), the omics sciences, especially in their untargeted versions, differ from the sciences from which they are derived because they tend to investigate the nature of the relationships that are established between the different elements (in metabolomics these elements are the metabolites).
- The multivariate approach allows analysis of the metabolites in their social context, thus taking into account the relationships they establish with each other. This approach, in turn, has several branches. However, the most used technique in metabolomics is the principal components analysis (PCA).

Univariate approach

The first thing that needs to be evaluated to apply the univariate approach is the number of classes of which the dataset is composed. The classes are labels that must be attributed to the analyzed samples to group them into homogeneous structures. In particular, the datasets containing only 2 classes (for example healthy controls or disease-affected patients) need a different approach than those with 3 or more classes (for example healthy subjects, symptomatic, and asymptomatic patients).

Another aspect of great importance is the evaluation of the distribution of data within the dataset. To find differences in the concentration of a specific metabolite between two or more classes, different statistical tests can be used which can be divided into parametric and nonparametric tests (see [Fig. 9.2](#)).

Parametric tests are easier to apply and generally give more robust results. The most common parametric test is the Student's *t*-test ([Owen, 1965](#)). This test can be applied only to data populations that have normal (or Gaussian) and homoscedastic (with similar variance) distribution ([Fig. 9.3A and C](#)).

If the concentrations of a specific metabolite meet these conditions of normality of distribution and homoskedasticity, then a parametric test (such as the Student's *t*-test) can be applied. Metabolites that do not meet this condition require a nonparametric test such as the U-test ([McKnight & Najab, 2010](#)) ([Fig. 9.3B](#)).

In addition to normality and homoskedasticity, another feature to be taken into consideration when choosing the most suitable statistical approach for exploratory analysis is the relationship between the different samples, in particular whether

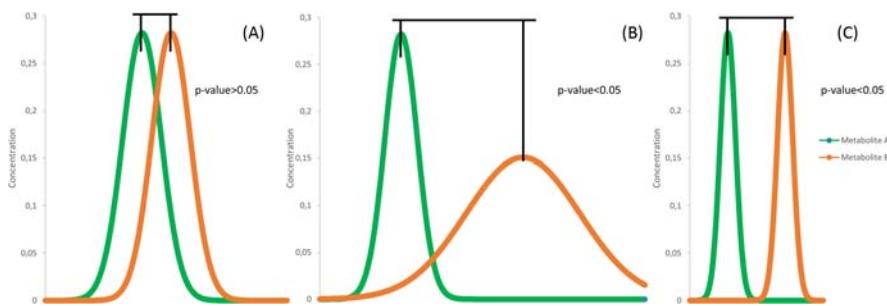


FIGURE 9.3 Univariate analysis to evaluate the differences in metabolite concentration.

(A) Concentration data of metabolites A and B are normally distributed and homoscedastic. Curves are partially overlapped so p-value is greater than 0.05. (B) Nonhomoscedastic data must be analyzed using nonparametric tests. (C) Concentration of metabolites A and B do not overlap resulting in a p-value less than 0.05.

the data are paired or unpaired. The data obtained from unpaired samples come from different subjects who have different conditions (for example, healthy subjects and disease-affected patients). On the other hand, experiments with paired-type samples are those that involve multiple observations on the same subjects; for example to evaluate the metabolic impact of a certain therapy, sick subjects can be analyzed before and after the administration of a certain therapy. In this case, there is a reduced metabolic variation in the different samples; therefore, these require specific statistical approaches for paired data.

These evaluations are important to choose the most appropriate statistical test to be applied to evaluate the differences in concentration between subjects of the same population with different labels (classes). Student's *t*-test can be used in two variants, one exclusively designed for paired data and one for unpaired data. Conversely, if the data are not normal or are not homoscedastic, they require different tests. For unpaired data, the "U-test," also known as the "Mann-Whitney test", is generally used; while for paired data the "Wilcoxon rank-sum test" is the most used alternative.

Experimental settings that involve the observation of more than two classes of subjects, can, in the same way, be analyzed using different approaches depending on the distribution of the data and the pairing. Data distributed in a normal and homoscedastic way are generally evaluated with the analysis of variance (ANOVA). This is a kind of *t*-test but applied to datasets with more than two classes. The result of this survey is simply a p-value that shows if there is a difference in metabolite concentration between one class and another, but it does not indicate exactly which class is different from the others. To have this type of answer it is necessary to apply a "post-hoc" test which is a test that is carried out on the ANOVA and indicates exactly which classes differ from another. In the case of nonnormal data, the variance test (ANOVA) is replaced by the

“Kruskal - Wallis” test, which is a test designed specifically to treat these type of data and also in this case it is necessary to apply a “post-hoc” test to identify exactly which classes differ from others.

Tests to investigate metabolite concentration differences

The *t*-test, applied to data distributed in a normal and homoscedastic way, evaluates whether a difference between the average concentrations of a given metabolite between two classes is statistically true, that is, is it representative or not. The test result is a “*p-value*.” This value is generally considered statistically significant if it is lower than a certain cut-off (e.g., 0.05), which means that there is a 95% chance that the two means are actually different and therefore that these two classes of subjects have concentrations of that metabolite that are actually different. On contrary, a *p*-value greater than 0.05 means that these two differences, although reported, are somehow hidden by the variability within the two classes of data and therefore represent a random effect for which repeating that experiment could have opposite results. Indeed, this evaluation, in addition to being based on the mean is also based on the standard deviation. The more the data are dispersed, the more it means that the two curves describing the distribution of the data can have a wider overlap (Fig. 9.3A and C). The greater the overlap, the less likely this difference is to be relevant.

In the case of nonparametric tests such as Mann–Whitney, the basic mechanism is the same; however, rather than analyzing the mean and the standard deviation, the medians and the rank are evaluated. The rank is the position that a value (e.g., metabolite concentration) occupies within the values ordered in ascending order.

In the case of paired evaluation, samples come from the same subjects observed under different conditions. Therefore, rather than simply evaluating the means, the differences in the means between the first and second observation or the difference in the medians are used.

In metabolomics, the number of parameters evaluated (metabolites) is generally large. A single metabolomics experiment can provide information on thousands of metabolites that can be analyzed independently of each other.

In the 1930s, Prof. Bonferroni, of the University of Florence, observed that if more hypotheses are verified, the possibility of observing a rare event increases and, therefore, the probability of mistakenly rejecting a null hypothesis increases. In other words, as the number of comparisons and therefore the metabolites analyzed increase, the probability of obtaining differences in concentrations with a *p*-value <0.05 increase proportionally, regardless of the biological significance of these evidences. For this reason, he suggested, in the case of multiple testing, to correct the reference cut-off by dividing it by the number of comparisons made. The new *p*-value, although it is lower than the original 0.05, retains significance at 95%.

Although this system is effective, it is very conservative. Indeed, this system makes it possible to attribute statistical significance only to those differences that actually have a great impact, such that these differences cannot be the result of

chance. On the other hand, less significant differences, which are still relevant, could be cut out by this correction system due to the significant decrease in the cut-off as the number of metabolites analyzed increases.

A different system, less conservative and widely used in metabolomics, is based on False Discovery Rate (FDR). This system is based on the idea of Benjamini and Hochberg (Benjamini & Hochberg, 1995), who proposed, in the case of multiple comparisons, to evaluate the various p-values of the different comparisons taken individually and independently of each other and then order them in a increasing manner by identifying the ranks of each p-value and then evaluating the correct p-value. This is determined by means of the equation:

$$p\text{-value}_{corr} = Q \frac{i}{m} \quad (9.1)$$

where Q is the rate of false discoveries to be accepted, i is the rank of the p-value and m the number of comparisons made. By accepting, as normal, the 95% confidence which corresponds to a raw p-value of 0.05, the corrected p-value corresponding to the raw p-value of 0.05 is identified and this is used as a new cut-off. This simple and computationally undemanding procedure has been very successful in the field of omics sciences and is now considered the gold standard in the case of multiple comparisons.

Regarding ANOVA, there is a process relatively similar to that of the analysis of the *t*-test done on two classes, but it concerns multiclass experiments. In this case the standard deviation, as well as the mean, plays a key role in understanding whether the concentration values of the metabolites in the different classes are different or not. In the ANOVA, the variance is evaluated both internally in each single population of data (each metabolite) and between the different classes (Fig. 9.4).

ANOVA, similar to Student's *t*-test, returns a p-value, which, by evaluating the overlap of the different Gaussians, indicates the significance of the difference in the means. However, ANOVA alone is unable to assess which class differs significantly from the others. To obtain this type of information, "post-hoc" tests are

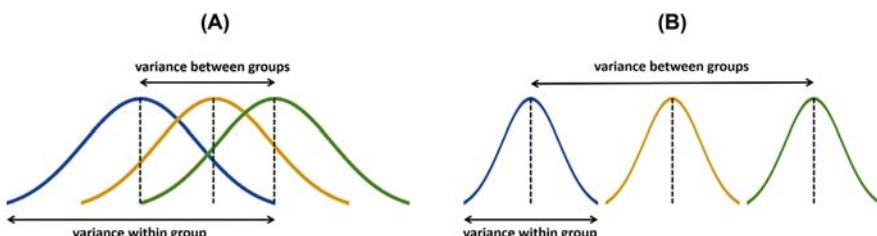


FIGURE 9.4 Analysis of Variance (ANOVA).

(A) Partially overlapped curves showing metabolite group's concentration difference without significance. (B) The ANOVA within and between groups allows the highlighting of the significance of the concentration differences in multiclass experiments.

used. In the case of ANOVA, the most used post-hoc test is the Tukey's test which acts as a kind of split of the p-value in the various possible comparisons, highlighting the pairs of classes with a significant difference.

Another interesting parameter to evaluate the difference in concentration of a metabolite with respect to the classes under study is the “fold change”. This too can only be applied when comparing between two classes. It is a simple parameter that is evaluated by making the ratio of the average of the concentration value of a metabolite in one class with respect to another. Fold change (FC) is an interesting parameter because it provides quantitative information about the relevance of the variation in a given metabolite between the two classes. In metabolomics, as in many other omics disciplines, a combined evaluation of p-value and fold change (FC) is commonly illustrated using the volcano plot or smile plot (Fig. 9.5).

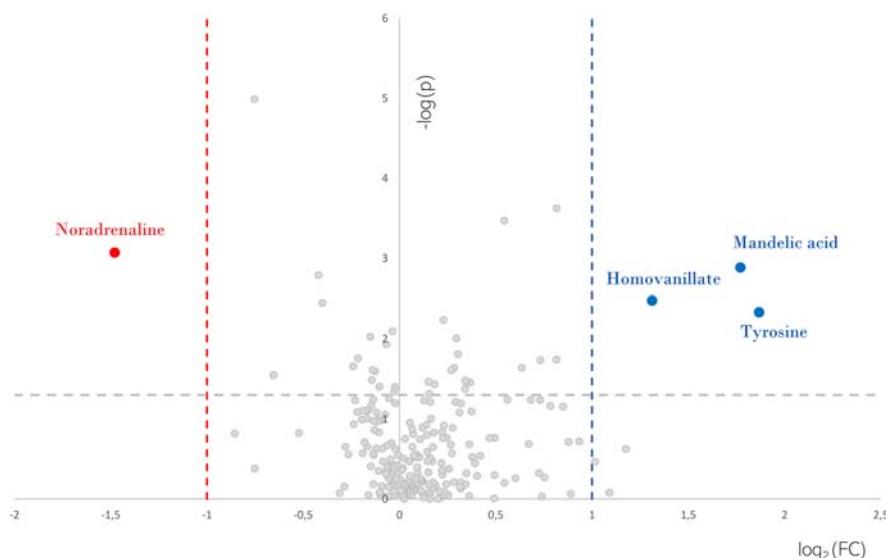


FIGURE 9.5 Volcano plot.

In the context of exploratory analysis, metabolite concentration can be investigated by means of volcano plot (aka smile plot) that is constructed by plotting the negative log of the p-value on the y-axis. This results in data points with low p-values (highly significant) appearing toward the top of the plot (red and blue points). The x-axis is the log of the fold change (FC) between two conditions. The log of the FC is used so that changes in both directions appear equidistant from the center. Plotting points in this way results in two regions of interest: those points that are found toward the top of the plot that are far either to the left- or right-hand sides. These represent values that display large magnitude FCs (hence being left or right of center) as well as high statistical significance (hence being toward the top). Here, mandelic acid, homovanillate, and tyrosine are examples of metabolites most significantly different than the cohort of metabolites.

The Cartesian graph shown in Fig. 9.5 shows both the FC (often represented in logarithmic form) and the co-logarithm of the p-value. In this graph each point represents a metabolite. The distance from the origin of the axes along the x-axis increases with increasing the absolute FC value, while metabolites with greater statistical significance (lower p-value) are plotted upwards. For this reason, the most significant metabolites (lower p-value and higher FC) are plotted in the upper right or upper left of the plot. They seem almost ejected from the mouth of a volcano, hence the name “volcano plot”.

Discovering the difference in concentration of the various metabolites analyzed between two classes is a possibility of EDA. A further interesting aspect to investigate in the exploratory analysis phase is the relationship that the metabolites have with each other. In particular, it is possible to study whether these have a positive, negative or zero correlation. Correlation, in this context, means the concordance of the values found experimentally with respect to a hypothetical fixed law that links these two concentrations. This law can be expressed graphically by means of a correlation curve.

As shown in Fig. 9.6, the strength of the correlation can also be measured through a parameter generally indicated with the letter “R”, based on this parameter correlation can be:

- Positive ($R > 0$), which indicates that as the concentration of one metabolite increases, the concentration of the other metabolite also increases;
- Negative correlation ($R < 0$), in which as the concentration of one metabolite increases, that of the other metabolite decreases;
- Lack of correlation ($R = 0$).

The more the parameter “R” approaches 1, the more the observed behavior is similar to the expected theoretical behavior. On the contrary, the smaller this value is, the more it means that these data are actually dispersed with respect to this correlation law, to the extent that in $R = 0$ the data are equally dispersed and therefore there is no type of correlation.

The correlation analysis is also affected by the distribution of the starting data. Data normally distributed and homoscedastic can be evaluated about their correlation using the Pearson equation. Conversely, data that are distributed in a

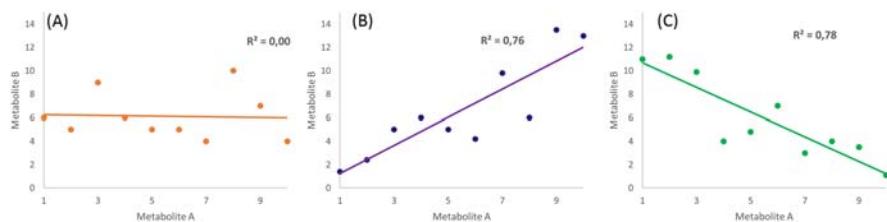


FIGURE 9.6 Metabolites' concentration correlation.

(A) Data not correlated. (B) Data positively correlated. (C) Data negatively correlated.

nonGaussian and/or are heteroskedastic must be evaluated in terms of correlation by means of Spearman's rank correlation (Winter, Gosling, & Potter, 2016).

Both the FC and the correlation, which can be analyzed for each individual metabolite, can also be displayed (although they have been evaluated individually for each metabolite) in an organic way on the entire set of metabolites or on a subset. Different graphical representations can be used to simultaneously visualize the correlations of different metabolites, of which the Heatmap is the most frequent (Fig. 9.7).

Similar representations can also be used to compare the FCs of different metabolites (Fig. 9.8). The intensity of the color is linked to the value of the represented parameter.

The colors that are used to represent negative and positive values (FC in Fig. 9.8), are generally warm and cold colors. In this case, red, which is a warm color, is used for positive FC and blue for negative, and the intensity of these colors is an indicator of how big the FC value is. Intense red cells, indicate large and positive FC, while intense blue cells indicate a negative and high FC. Such a heatmap helps to visualize those metabolites that have consistent positive or negative relationships with each other, so it helps to group these metabolites.

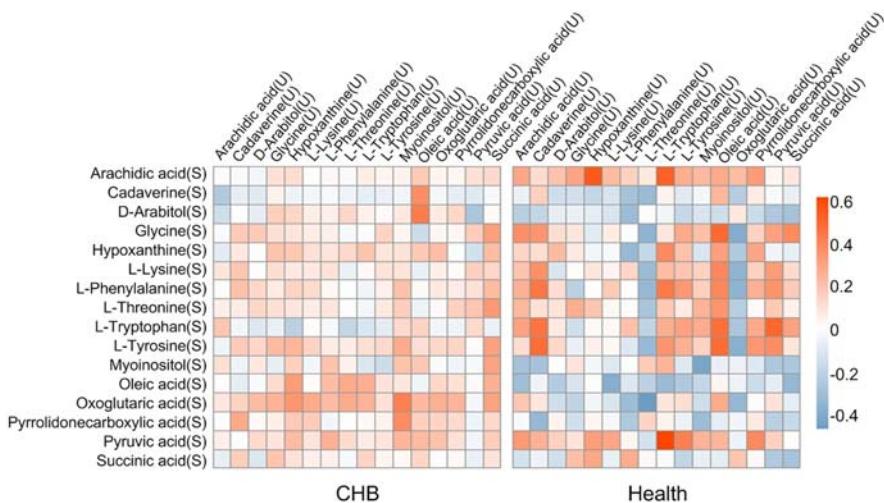


FIGURE 9.7 Inter-correlations between serum and urinary metabolites.

Correlation between metabolite concentrations in serum and urine of healthy subjects and of patients with chronic hepatitis B (CHB). “U” means metabolites in urine and “S” means metabolites in serum. Each lattice denotes correlation within one metabolite pair in terms of R. The darker the red, the more positive the correlation. The darker the blue, the more negative the correlation. White pairs are without correlation.

Adapted from Yang, L., Yang, X., Kong, X., Cao, Z., Zhang, Y., Hu, Y., & Tang, K. (2016). Covariation analysis of serumal and urinary metabolites suggests aberrant glycine and fatty acid metabolism in chronic Hepatitis B. PLoS One, 11(5), e0156166. <https://doi.org/10.1371/journal.pone.0156166>.

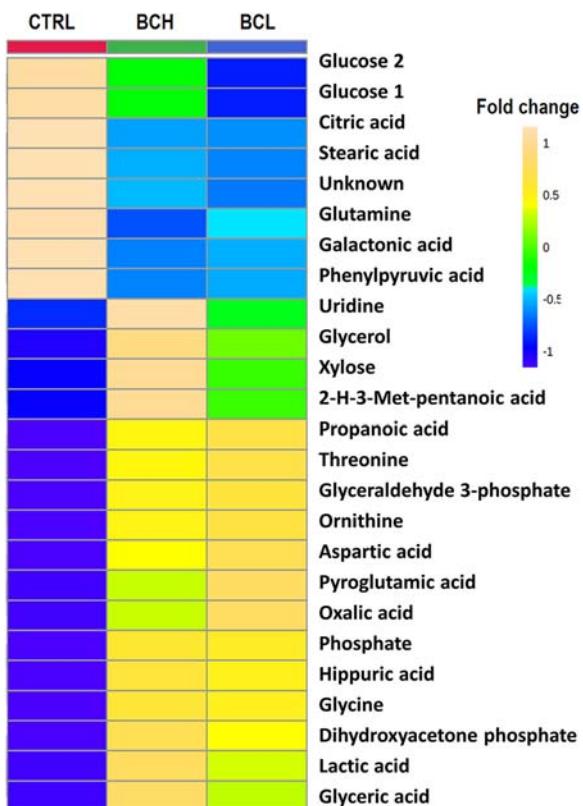
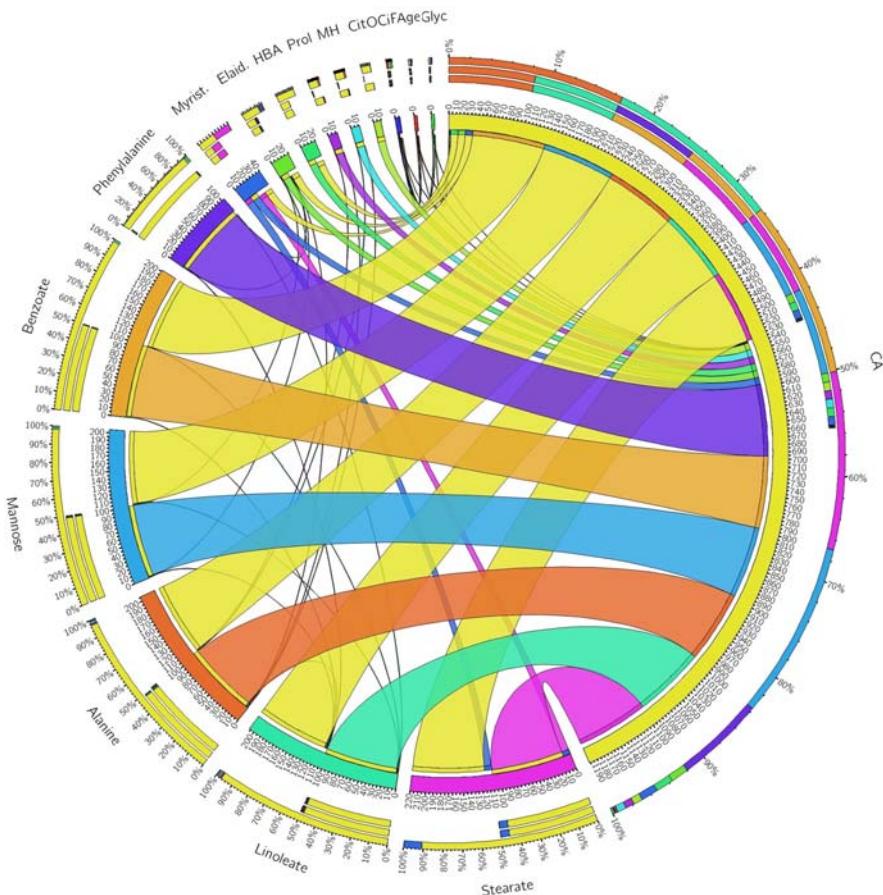


FIGURE 9.8 Heatmap representing the fold changes of selected metabolites.

Serum fold changes (FC) of metabolite concentration between healthy subjects (CTRL) and patients with high grade (BCH) and low grade (BCL) bladder cancer. Yellow indicates metabolites showing increased concentration ($FC > 0$), while blue metabolites represent decreased concentration ($FC < 0$). Green metabolites represent no significant fold change between the analyzed conditions. Glucose 1 and Glucose 2 represent two different derivatization products obtained by the silanization process.

Adapted from Troisi, J., Colucci, A., Cavallo, P., Richards, S., Symes, S., Landolfi, A., Scala, G., Maiorino, F., Califano, A., & Fabiano, M. (2021). A serum metabolomic signature for the detection and grading of bladder cancer. *Applied Sciences*, 11(6), 2835.

Another system of representation is the “circle plot”. By means of this type of representation, the relationship that a specific metabolite establishes with other metabolites or with specific conditions can be highlighted. The graphical representation involves the use of lines of variable thickness to link two objects (for example the concentration of a given metabolite and the sex or other characteristics of the studied subjects). The thickness of the link is often a function of the correlation coefficient. The example shown in Fig. 9.9, relates to a study of the

**FIGURE 9.9 Circle plot.**

Circle plot uses a circular composition to show connections between objects or between positions, which are difficult to visually organize when the underlying layout is linear. Concentration of several metabolites (stearate, linoleate, alanine, mannose, benzoate, phenylalanine, myristate, elaidate, hydroxy-butyrate, proline, methyl histidine, glycerol), as well as clinical conditions (maternal age, familiarity for chromosomal anomalies) were correlated with the actual diagnosis of fetal chromosomal anomalies (CA). Lines widths represent the correlation strength.

Adapted from Troisi, J., Sarno, L., Martinelli, P., Di Carlo, C., Landolfi, A., Scala, G., Rinaldi, M., D'Alessandro, P., Ciccone, C., & Guida, M. (2017). A metabolomics-based approach for noninvasive diagnosis of chromosomal anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 13(11), 140. <https://doi.org/10.1007/s11306-017-1274-z>.

untargeted metabolomic profile of pregnant women who had been diagnosed with a fetal chromosomal abnormality (Troisi et al., 2017). The graph shows that the presence of the fetal chromosomal abnormality is strongly correlated to the presence of stearate, while it is weakly correlated to the concentration of glycerol. The greater the thickness of the line, the greater is the relevance of that metabolite to that specific phenotype.

All the exploratory evaluations reported so far analyze metabolites singly or in pairs. Although this approach provides useful information, it is still quite far from fully utilizing the power of metabolomics which is based on the analysis of all metabolites simultaneously. For this reason, data from a true omics perspective is achieved through multivariate analysis techniques, in which not just a single metabolite is examined, but all the metabolites together.

Multivariate approach

Principal component analysis (PCA) is one of the most useful and versatile tools for multivariate data analysis. PCA is used often and in various contexts in metabolomics. This technique dates back to 1901 and was first formulated by Pearson, but Pearson actually described its general principles. It was Prof. Harold Hotelling of Stanford University, who perfected this technique, to formalize its analytical aspects so much that today we know it as PCA or as **Hotelling transform** (Jolliffe, 2005).

PCA is a **dimensional reduction** system that synthesizes the information dispersed in a large number of metabolites into a few new variables called principal components (PCs). PCA is also an intermediate step for many multivariate techniques. PCA consists of a process of rotation of the original data defined by a matrix X of size $n \times p$, where n represents the number of rows (generally in metabolomics one row is used for each sample) and p the number of columns (generally equal to the number of analyzed metabolites) carried out in such a way that the first new axis is oriented in the direction of maximum variance of the data, the second is perpendicular to the first and is in the direction of the subsequent maximum variance of the data, and so on for all the new axes.

In Fig. 9.10, an example is shown. It represents a three variable (metabolites) dataset. As can be seen from the figure, the first principal component (PC1) is in the direction of maximum variance (dispersion) of the original points while its origin is located in the mean value of the variable. The residual variance is represented by the second main component (PC2), in the direction perpendicular to the first component. Because in this case we have only three variables, the two components almost entirely describe the initial data. Each of the two components is a linear combination of the three original variables.

The mathematical procedure for the determination of the PCs consists in the computation of the eigenvalues and eigenvectors of the covariance (or correlation)

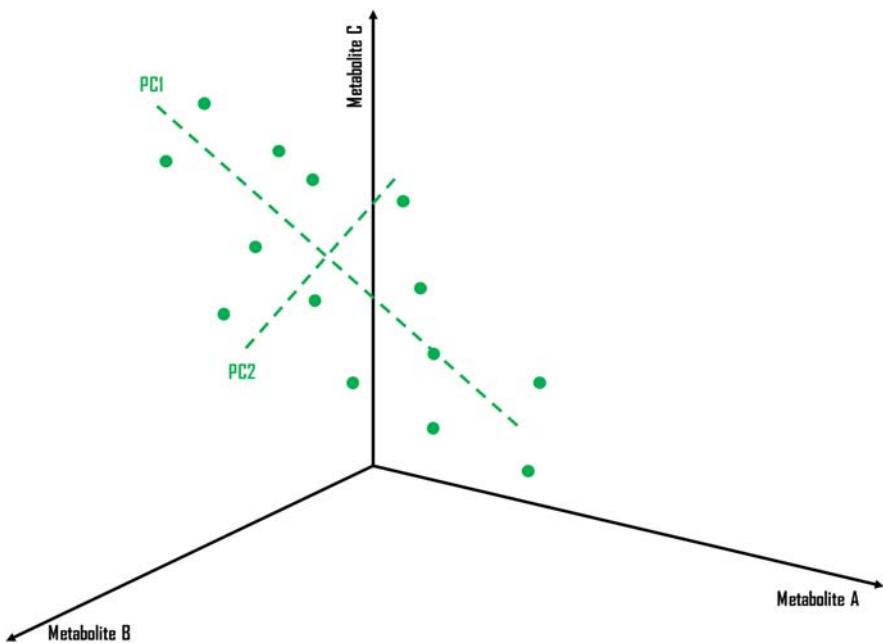


FIGURE 9.10 Principal component analysis.

A 3-dimensional dataset could be reduced using 2 orthogonal principal components (PC1 and PC2) explaining almost all the variance contained in the original dataset.

matrix of the data X , that is, in the diagonalization of the covariance matrix S of X , defined as:

$$\text{diag}(S) = \text{diag} \left[\frac{X_C^T X_C}{n-1} \right] \quad (9.2)$$

where X_C is the centered data matrix.

The variance is the square of the standard deviation. Therefore, it concerns a single variable (in metabolomics a single metabolite) and represents the dispersion of the concentrations of that metabolite in the different samples analyzed with respect to the mean concentration. The covariance, on the other hand, represents the reciprocal change of two metabolites (Eq. 9.3).

$$\text{Cov}(A, B) = \frac{1}{n-1} \sum_1^n ([A_i] - \bar{A}) ([B_i] - \bar{B}) \quad (9.3)$$

Starting from the evaluation of the covariance, it is possible to calculate the covariance matrix (Table 9.1). The covariance matrix is a symmetric matrix whose diagonal represents the variances of the individual metabolites.

Table 9.1 Covariance matrix: Covariance matrix between metabolites A and B is a square matrix giving the covariance between each pair of elements. It is symmetric and positive, its main diagonal contains variances (i.e., the covariance of both metabolites A and B with itself).

	Metabolite A	Metabolite B
Metabolite A	Var(A)	Cov(A, B)
Metabolite B	Cov(B, A)	Var(B)

PCA is generally applied to autoscaled data (see [Chapter 8: Techniques for Converting Metabolomic Data for Analysis](#)). This is a precaution applied to avoid that those metabolites with high average concentration or variance values monopolize the process of identifying the PCs. The covariance matrix, which is the main tool for managing the process, corresponds to the correlation matrix when the data are autoscaled.

The diagonalization of the covariance matrix involves the determination of a diagonal matrix, called the eigenvalue matrix, whose diagonal elements are the eigenvalues, ordered in descending order, and of a loadings matrix, whose columns are the eigenvectors of the covariance matrix, that is, each column contains the coefficients of the corresponding eigenvector whose number is, in general, less than or equal to the number of metabolites. The eigenvectors are the versors in the new space. The axes of the new space (PCs, also called factors or eigenvectors) are the axes relative to the directions of maximum variance, in decreasing order. This makes it possible to represent the matrix of the original autoscaled data in a new orthogonal space, according to the following relationship:

$$T_{n,M} = X_{n,p}L_{p,M} \quad (9.4)$$

where L has the function of a rotation matrix and T is called the scores matrix. In the case where $M = p$, the operation consists of a simple rotation of the original data in a new coordinate system, without any modification of the overall information initially contained in the X data matrix.

Since the eigenvalues represent the variance associated with each eigenvector (principal component), it is generally probable that the smaller eigenvalues are associated with variability due to noise or insignificant information. In these cases, it is possible to eliminate this part of the variability of the data by taking into consideration only a number M of components less than p . This aspect of PCA is of the greatest importance in metabolomics and many methods have been proposed to determine the number M of significant PCs.

Loadings and scores in principal components analysis

The loading matrix L is the matrix whose columns represent the eigenvectors of the covariance (or correlation) matrix; the rows represent the original variables. This means that, once an eigenvector is selected, in each row we find the numerical coefficients

that represent the importance of each original variable (metabolite) in that eigenvector. The loadings are standardized linear coefficients, that is, the sum of the squares of the loadings of an eigenvector is equal to 1 or the eigenvectors have unit variance. An eigenvalue close to 1 in absolute value indicates that the m-th component is mainly represented by the j-th original variable; conversely, a value close to zero indicates that the j-th variable is not represented (it is not important) in the m-th component.

The value of the scores is the result of a linear combination, in which the variables are the original variables (generally scaled) and whose multiplicative coefficients are the loadings of the m-th component. Unlike the loadings whose values are limited between ± 1 , the scores have an average value equal to zero, but they can assume any numeric values. The scores represent the new coordinates of the objects (samples) in the principal component space.

A very important aspect in the study of multivariate problems concerns the possibility of representing the data graphically (Fig. 9.11). PCA provides an

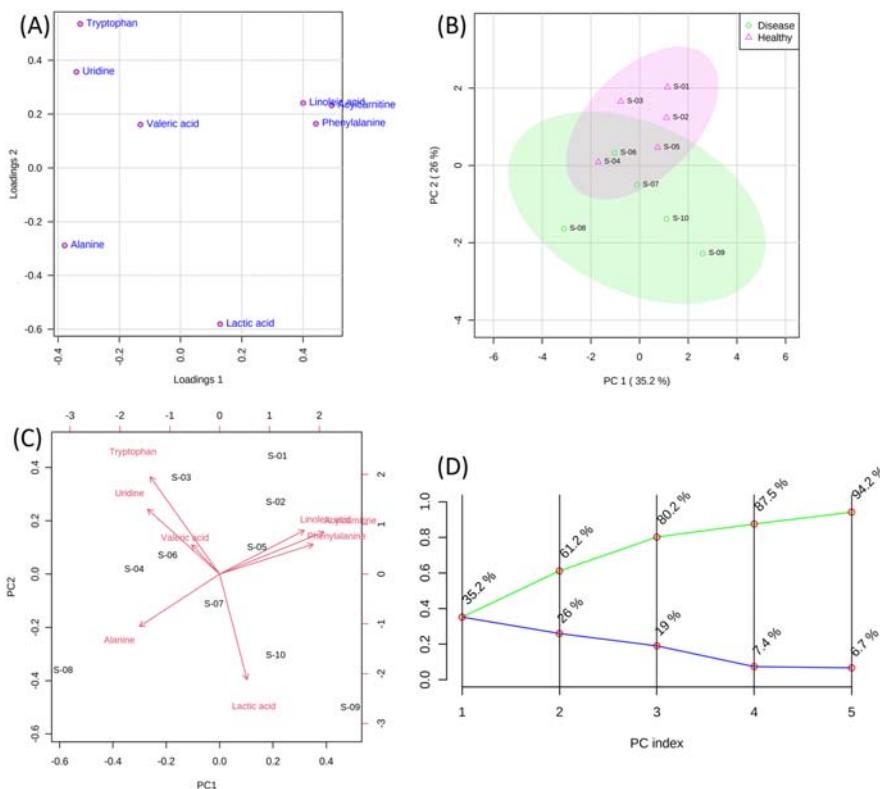


FIGURE 9.11 Principal Component Analysis graphical representation.

(A) Loading plot; (B) Score plot. The percentage of explained variance is reported in parentheses on each axis of the score plot; (C) Biplot; (D) Scree plot.

algebraic solution that also allows very effective graphical representations both of objects only (scores plots) and variables only (loadings plots) and of objects and variables simultaneously (biplot).

For each pair of PCs, the amount of total variance represented in the graph is given by the sum of the variances explained by the individual components. The loadings graph (Fig. 9.11A) illustrates the role of each variable in the different components, their direct and inverse correlations, and their importance. Once the two components that make up the coordinate axes have been chosen, the coordinates of each variable are defined by the pair of loadings (matrix L of the loadings) that each variable has in the two components considered. For this type of graph, each variable will always necessarily be between -1 and $+1$ (the range of definition of the loadings). For example, variables that are near the center (the 0.0 point) are not relevant to either component; variables that are located at the extremes of one of the components are instead important variables in this component (large loadings in absolute value). Hence, large positive or negative loadings for some variables indicate that these variables are significantly represented in the component. Groups of variables that appear close together in the loadings graph indicate that, limited to the information carried by these components, they carry common or similar information (i.e., they are correlated). If this happens for all the components considered as a model, it is possible to represent the information content carried by this group of variables with only one of them. This also applies to variables that appear opposite each other with respect to the origin (the point 0.0). In this case these variables are inversely correlated.

The scores plot (Fig. 9.11B) illustrates the behavior of the objects (analyzed samples) in the different components and their similarities. Once the two components that make up the coordinate axes have been chosen, the coordinates of each object are defined by the pair of scores (matrix T of the scores) that each object has for the two components considered. The score graph allows analysis of the behavior of the objects in the light of the components considered, that is, in the light of their meaning and the values of the variables that most adequately characterize them. In this way it is possible to detect groupings of similar objects (clusters), the presence of outliers, and the manifestation of particular regularities and distributions.

The biplot graph (Fig. 9.11C) illustrates objects and variables at the same time for evaluating the relationships between them. This allows the analysis of the PCs, to investigate the position of the samples in the score plot and the values of the important variables in the corresponding loading plot. It corresponds to the overlap of loading and score plots.

Significative components

One of the fundamental problems that the PCA brings is the determination of the number M of significant components (factors), with $M < p$. The search for the number of significant components is known as rank analysis. In fact, if the data contain an information structure (i.e., not random), the separation between the

variability due to experimental noise or spurious information and useful information occurs by means of an appropriate delimitation of the number of significant PCs. Any procedure for selecting a small number of significant components assumes that the variability of useful information is greater than the variability associated with experimental noise or secondary information. Because each eigenvalue represents the variance associated with the corresponding principal component and the sum of all the eigenvalues coincides with the total variance present in the data, the percentage variance explained by the PC1 (EV1%) with respect to the total variance is given by

$$EV_1\% = \frac{\lambda_1}{\sum_1^p \lambda_m} 100 \quad (9.5)$$

where λ represent the eigenvalues.

The number M of eigenvalues to be used can be evaluated based on the graphical analysis of the eigenvalues reported against the number of factors. In this type of graph, known as a scree plot (Fig. 9.11D), the number of components is shown on the abscissa axis and the corresponding eigenvalues on the ordinate axis. The first M factors are chosen for which the residual variance reduction is more pronounced. This is a substantial graphical-visual method, advisable only for the simplest cases. If the eigenvalues have very different values from each other, with the first eigenvalues being very large, the graph is often constructed using the logarithm of the eigenvalues. Alternatively, all the components that explain at least 95% of the variance can be considered. There are other strategies such as the mean eigenvalue criterion and the Malinowski indicator function, but these are not commonly used in metabolomics.

Conclusion

PCA is one of the fundamental techniques used in metabolomics for multivariate data analysis. It can be considered a step of exploratory analysis as well as an unsupervised ML algorithm because of its ability to be trained from data and to discover cluster aggregation (see below). Using this technique, it is possible to evaluate the correlations between the metabolites and their relevance, illustrate relevance and significance of metabolites (identification of outliers, classes, etc.), summarize the description of the data (elimination of noise or spurious information), reduce the dimensionality of the data, search for main properties and define a data representation model in an orthogonal space. Moreover, PCA is an intermediate step for many multivariate techniques.

A further important aspect of PCA concerns the interpretation in the sense of understanding the meaning of the PCs. In fact, this interpretation bridges the gap from a description-knowledge of the system in terms of original variables, each clearly known in its meanings, to a description at a higher semantic level that we could define as metadescription, whose role can allow us to grasp the emergent properties of the system. These are new properties with respect to the original

knowledge, which emerge from a holistic view of the system or from the synergistic or antagonistic effects of the original variables that describe it.

The metabolites that are selected (in terms of loadings) and joined in a linear combination that generates a certain PC (selected because it describes a good amount of initial information) are linked by a bond that cannot be deduced from the simple initial observation of the variables and constitutes the essence of an emergent property of the system. In a certain sense, we have a new operational definition of metabolic pathways: it no longer corresponds to any of the typical biochemical pathways, but it is a linear combination of these that represents the emergence of a new semantic plane, at a higher level of complexity and more general, but at the same time highly specific for the experimental data to which we applied it. For all practical purposes, these new variables (the PCs), whether interpreted or not, constitute a new synthetic description of the system. An important feature of the PCs is the fact that each of them links only the most informative part of each metabolite that contribute to the linear combination. This means that only the systematic variation of the original variables defines the most significant PCs: the irrelevant variation or the variation caused by experimental noise or the systematic variation not correlated with the component are not represented. In this sense, PCA can also be considered a powerful information filtering tool.

Unsupervised machine learning analysis

Introduction

ML is one of the branches of artificial intelligence (AI) that deals with creating systems that learn or improve performance based on the data they are given. AI is a generic term and refers to systems or machines that mimic human intelligence. The terms ML and AI are often used together and interchangeably, but they don't mean the same thing. ML is used everywhere today. When we interact with banks, online shopping, or use social media or cable television, ML algorithms are used to make our experience efficient, easy and safe. Algorithms are the engines that power the ML. The two main types of ML algorithms currently used are supervised and unsupervised ML algorithms. The difference between these two types is defined by how each algorithm uses the available data to make predictions.

In **supervised algorithms**, ML has the purpose of being trained to perform a specific task, which generally consists in recognizing a label or class to which it belongs. The model is built by assigning a certain label to samples used for training, the model, therefore, learns to recognize this class based on the data it receives. Imagine collecting a certain number of healthy subjects, obtaining biological samples, extracting the metabolome and then obtaining information relating to the concentration of a certain number of metabolites. Collect the same data from an equal number of sick subjects. All the collected information

(concentration of the metabolites analyzed), including belonging to the healthy or sick subject class, is summarized in the dataset and used to train the ML algorithm, which learns from the information contained in the various metabolites to assign the correct label to any unknown sample. This is a typical supervised ML experiment. This type of data analysis strategy is extremely useful in metabolomics because it allows to create a “diagnostic” tool that is useful for making predictions on future unknown samples but above all because, by means of the analysis of the steps that allowed the training, it is possible to gain new insights into the relationships between the metabolites that underlie that specific disease state.

In unsupervised algorithms, the class of the analyzed samples is not pre-set. In other words, the model simply receives the information relating to the semi-quantitative or quantitative data of the different metabolites analyzed. The training process consists of the discovery phase of an aggregation system of the analyzed samples, in other words a natural occurrence of similarity is sought between the different samples to identify common or aggregating behaviors.

The choice of the ML approach generally depends on factors related to the structure and volume of data but above all on the objectives for which it is being applied. For example, if you are investigating a pathology that can occur in different forms (asymptomatic, or medium severity and very severe; or acute and chronic), an unsupervised algorithm is used to evaluate whether there is a natural aggregation coherent with a pathological subtype and therefore to investigate the metabolites as a function of their loadings. Moreover, unsupervised systems also allow metabolites to be aggregated according to similar trends (all those that increase or decrease in a certain condition, for example). Furthermore, unsupervised systems can identify outliers, that is, samples that differ from the population of origin; these differences can then be investigated in detail.

On the contrary, supervised systems allow one to structure algorithms capable of identifying a certain condition in samples whose class attribution is not known *a priori*. By means of these tools, in fact, various screening tests based on metabolomics have been developed (Troisi et al., 2017, 2020; Troisi, Colucci, et al., 2021; Troisi, Cavallo, et al., 2021; Troisi, Landolfi, et al., 2018; Troisi, Scala, et al., 2018). Furthermore, the supervised approach has also proved useful for biomarker discovery (see Chapter 7: Approaches in Untargeted Metabolomics).

Cluster analysis

The main objective of cluster analysis is the search for groupings within datasets. Clustering is the process of reducing the number of objects in a few groups with similar behaviors. In metabolomics it can be applied in two different ways: by trying to group samples or by grouping metabolites.

There are different clustering systems, and these are generally divided into hierarchical and nonhierarchical algorithms. The former includes methods that can be agglomerative (such as single linkage, average linkage, complete linkage) or divisive; while the others include methods whose strategies are much more

different from each other, such as, for example, the Jarvis-Patrick and k-means methods (Fig. 9.12).

The clusters that each method identifies are characterized by their position in the p-dimensional space by a centroid, defined as the vector of the averages of the variables calculated for the objects assigned to the cluster, or by a centrototype, defined as the most representative object among the objects assigned to the cluster (as a rule, the closest to the centroid). Unlike the centroid, the centrototype is always an object present in the data.

In hierarchical systems there is a classification hierarchy, these systems can be agglomerative or divisive. These methods stem from choosing prototypes and aggregating with respect to these prototypes, while the divisive ones work separating from the group one sample at a time. Agglomerative models can work according to different logics (see Figs. 9.12 and 9.13). Agglomerative patterns are generally represented by means of a dendrogram. Nonhierarchical systems are not based on agglomerative or divisive logic starting from a prototype.

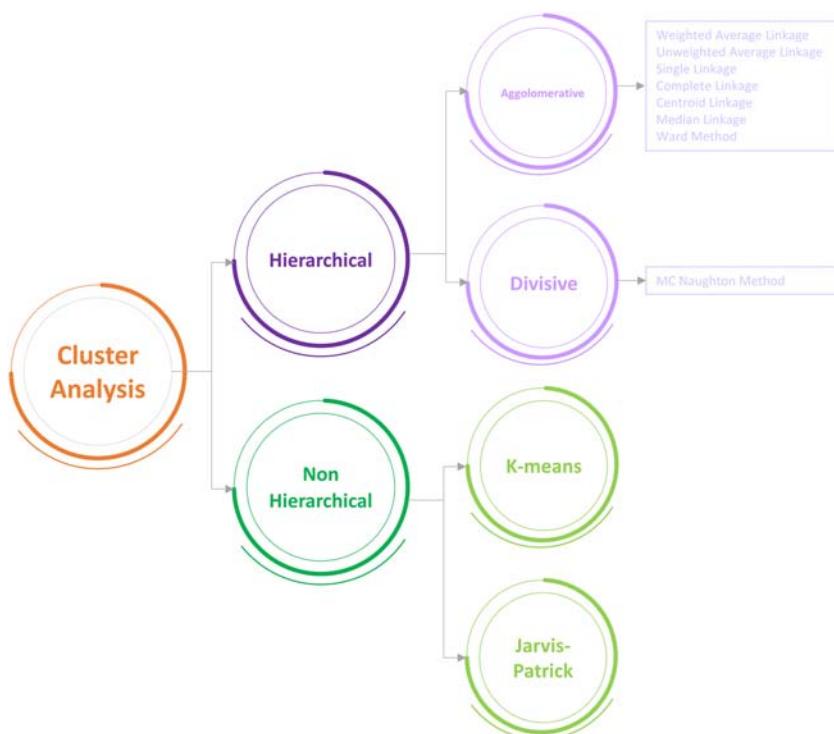
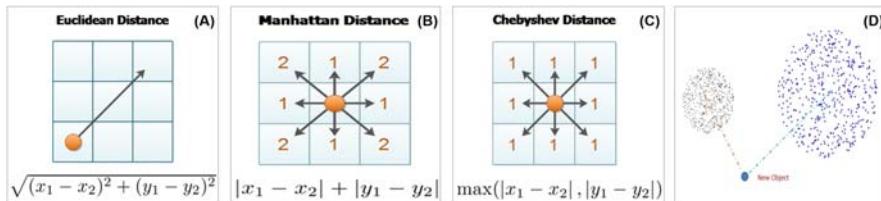


FIGURE 9.12 Cluster analysis.

Cluster analysis can be performed using hierachal and nonhierachal methods. The former can be agglomerative or divisive while the latter follows different logics, for example, the Jarvis-Patrick and k-means methods.

**FIGURE 9.13 Distances.**

Different distance metrics used in cluster analysis. (A) Euclidean distance, (B) Manhattan distance, (C) Lagrange (or Chebychev) distance, (D) Mahalanobis distance.

In general, cluster analysis methods stem from the evaluation of a type of distance (for example, the Euclidean distance), between the studied objects. The distance matrix is then calculated and from this, eventually, the similarity matrix. By applying the clustering algorithm, the final partition of the objects into clusters is obtained. The concept of cluster is not superimposable with that of class which is used in the case of supervised algorithms. Only through the interpretation of each cluster is it possible to identify classes.

Distance analysis is therefore the determining element of cluster analysis. The concept of similarity is a concept of great scientific and practical importance. Indeed, it is the mathematical equivalent of the analogy, a concept that human beings use to recognize, distinguish, classify. From the mathematical point of view, the concept of similarity is the complement of the concept of dissimilarity. To quantitatively measure these, it is possible to use the concept of distance: the more two objects are “distant”, with respect to a frame of reference in which many other objects appear, the more dissimilar they are. Conversely, the more they are “close”, the more they can be considered similar.

There are different types of distance that could be used for clustering, but all of these must respect at least these 3 parameters:

- **Symmetry:** the distance between a point a and a point b must be equal to the distance between point b and point a ;
- **Positivity:** the distance must always be a positive number;
- **Triangular inequality:** the distance between point a and point b is always less than the sum of the distances of a with any arbitrary point c and the distance between b and c .

Some of the most commonly used distances in metabolomics are (see Fig. 9.13) the following:

- **Euclidean distance:** estimated using an n -dimensional Cartesian plane by means of the Pythagorean theorem. It is the simplest and most common form of distance.

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (9.6)$$

- **Manhattan distance:** This type of distance is based on the architectural principle of all cities that have a Roman-like structural organization, in which the streets are distributed in intersecting mode, with angles of 90 degrees. In a geometry of this type, the distance between one point and another in the city cannot be estimated by means of the Euclidean distance. From a mathematical point of view this is the formula:

$$d_M(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n| = \sum_{i=1}^n |x_i - y_i| \quad (9.7)$$

- **Lagrange (or Chebychev) distance:** This is also called the distance of the chessboard because it can be equated to the estimate of the number of movements that the king can make to move along the chessboard. In the metabolomic field (as well as in other omics disciplines) it is a distance that is often used to evaluate the intrinsic variability of an experimental set. It is similar to the distance between the highest point of that experimental dataset and the lowest point. It can also be considered an estimator of dispersion with respect to the mean.

$$d_L(x, y) = \max_i |x_i - y_i| \quad (9.8)$$

- **Mahalanobis distance:** The estimate of this distance starts from the assumption that the sample points are distributed within a hypersphere around the centroids. If the distribution is not spherical (for example hyperellipsoidal), it would be natural to expect that the probability of the point under consideration to belong to the set depending on not only on the distance from the centroid, but also on the direction. When the hyperellipsoid has a shorter axis, the point under examination must be closer to be considered as belonging to the set, when the axis is longer, the point under examination is a greater distance away. Developing all this in mathematical terms, the hyperellipsoid that best represents the set of probabilities can be estimated through the covariance matrix of the samples. The Mahalanobis distance, therefore, is simply the distance of the point under examination from the centroid normalized with respect to the amplitude of the hyperellipsoid in the direction of the point under examination.

$$d_M(x, y) = \sqrt{(x - y)(x - y)^T S^{-1}} \quad (9.9)$$

- **Minkowski distance:** A generalized form of distance that includes Euclidean and Manhattan distance as special cases. Minkowski understood that by evaluating the distance using the sum of the difference in terms of the absolute value of the coordinates of the two points whose distance is to be measured elevated to any index and below the root of the same index, this

distance can be applied in general to any type of topology. For $m = 1$ this distance is transformed into the Manhattan distance, while for $m = 2$ into Euclid's:

$$d_{Mk}(x, y) = \sqrt[m]{|x_1 - y_1|^m + |x_2 - y_2|^m + \dots + |x_n - y_n|^m} = \sqrt[m]{\sum_{i=1}^n |x_i - y_i|^m} \quad (9.10)$$

Hierarchical clustering

As reported above, hierarchical methods fall into two broad categories: divisive hierarchical methods and agglomerative hierarchical methods. The first group of methods (rarely used in metabolomics) is based on strategies that start from a set that includes all the initial data and gradually separate the data that differ most from the others. On the contrary, the agglomerative hierarchical methods are the most commonly used in metabolomics and start from a number of clusters equal to the number of objects, gradually merging them into clusters of ever greater size.

Agglomerative hierarchical methods

Agglomeration methods usually require the following preliminary steps:

- definition of the metric to be used (distance type);
- calculation of the matrix of the distances between objects;
- calculation of the corresponding similarity matrix.

Once the similarity matrix is computed, the matrix is analyzed and reduced. First the two most similar clusters are identified (in the first phase, the most similar objects), then two clusters (or two objects) are merged into a single new cluster, at a given level of similarity. At this point, for the new cluster its similarity level with respect to the remaining clusters (or objects) is calculated, according to criteria that differ from method to method. This operation consists of eliminating the rows or columns related to the two clusters (or objects) that are placed together from the similarity matrix and adding a row or a column relating to the similarities of the new cluster with all the remaining clusters (or objects).

To identify the most similar objects or clusters, various methods can be followed, including single linkage, complete linkage, centroid linkage, and the Ward's method. As shown in Fig. 9.14, the single linkage method makes the distance between the two clusters coincide with the minimum distance between the objects; on the contrary, with the complete linkage method the two clusters are attributed a distance equal to the maximum distance between the objects; for the centroid linkage method the two clusters are attributed a distance equal to that of their centroids.

This type of procedure therefore starts from the analysis of the similarity matrix (symmetric, of size equal to the number of data n). After each merging of

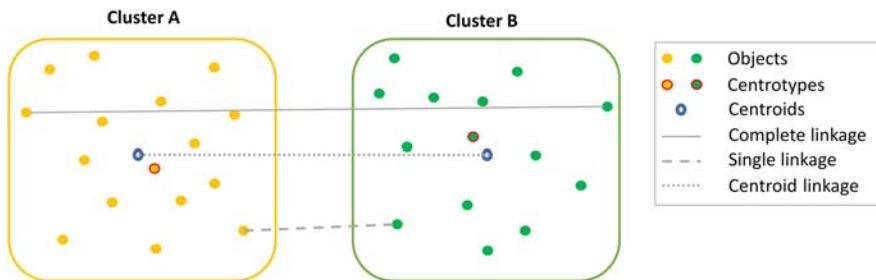


FIGURE 9.14 Hierarchical clustering methods.

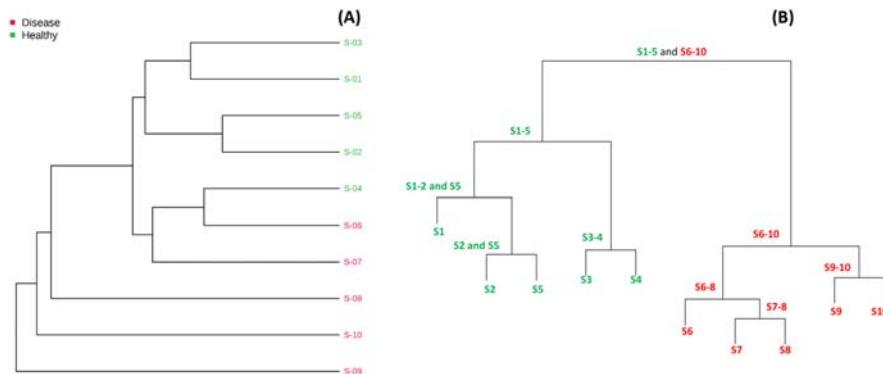
Distance between cluster (A) and (B) can be evaluated following different strategies such as complete linkage, single linkage, and centroid linkage.

two clusters, the columns (or rows) relating to the clusters that are joined are deleted and a column (or a row) relating to the similarities of the new cluster is added with all the others remaining. By doing so, the size of the similarity matrix is reduced by one with each step. The result of this procedure is normally represented by a graph called a dendrogram, which allows a highly informative visual analysis of the hierarchy of similarities between the objects considered. A graphic example of the result obtained by applying an agglomerative hierarchical clustering method is shown in Fig. 9.15A.

Examining the graph (Fig. 9.15A) from the left to right, we can observe the pairs of samples that are most similar to each other: samples S-09 and S-10 are the most similar to each other because they join first; subsequently the S-08 sample is added to this cluster and so on. Using different methods different groupings could be obtained. Clearly, the choice of a certain level of similarity is subjective and typical of cluster analysis methods.

Divisive hierarchical methods

As the name indicates, this method works with an inverse logic to the agglomerative methods (Fig. 9.15B). In particular, in the most used Mac Naughton method (Macnaughton-Smith et al., 1964), initially the set of all objects is divided into two subsets, which in turn are subsequently subdivided into two subsets, until each subset contains only one object. At each step, the most dissimilar object from all the others is selected and it will be the one whose sum of the distances from all the others is maximum. Isolating the first object, a comparison is then made between the average distance calculated between each of the objects of the remaining set and the distance of each of these objects with the isolated object. The object of the still undivided set is thus identified, more similar to the object (less distant from) previously isolated. The procedure continues by comparing the distances between the still undivided objects and their average distances with the objects already selected, until the starting set is exhausted. The procedure is

**FIGURE 9.15 Hierarchical clustering.**

(A) Agglomerative clustering method. (B) Divisive clustering method.

repeated separately on each of the two subsets obtained, until subsets containing a single object are obtained.

Nonhierarchical clustering

The methods for nonhierarchical clustering are based on techniques that are very different from each other and therefore difficult to schematize. A group of methods is based on techniques generally called relocation techniques, according to which, after an initial partition of the data, these are moved from one cluster to another until a predetermined criterion is met. Of these methods, the best known and widely used in metabolomics is the k-means method.

K-means method

The k-means method (MacQueen, 1967) is a method whose data relocation algorithm is based on the comparison of the distances of each object from the centroid of each cluster. The number of clusters is fixed in advance by the user as well as the metric (distance type) to be used. The algorithm is as follows: the objects are arranged in a number k of clusters, with k chosen *a priori*. For each cluster the centroid is determined, the distance between each object and each centroid is calculated and each object is assigned to the closest cluster. If in this process at least one object is moved to another cluster, the centroids are re-determined and continued repeatedly (Fig. 9.16A). New centroids are recalculated either after defining the relocation of all objects or whenever an object needs to be relocated to another cluster.

The k-means method is simple to apply and computationally undemanding, so it is widely used in metabolomics and many other omics disciplines. Nonetheless, it has some important vulnerabilities. In particular, the results obtained with this method depend on the partition of the initial objects, on the k parameters (number

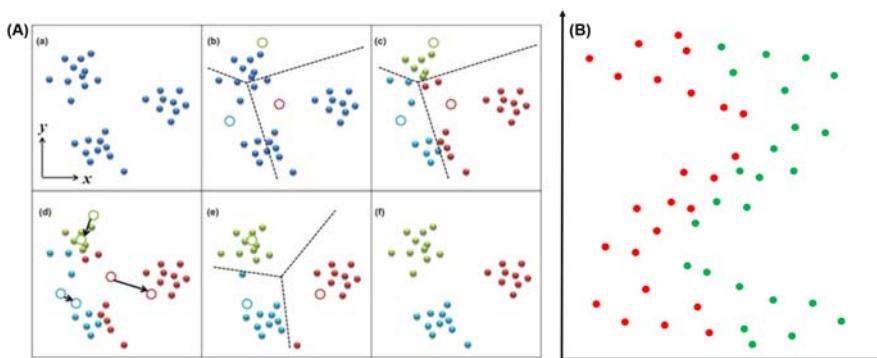


FIGURE 9.16 Nonhierarchical clustering.

(A) A schematic illustration of the K-means algorithm for two-dimensional data clustering. (a) The data points (solid blue circles) to be clustered in a 2D feature space. (b) For random locations of the cluster centers (aqua, green, and red hollow circles), each data point can be associated with the closest center. (c) The 2D space is divided into three regions through three decision boundaries (black dashed lines). (d) Each center moves to the centroid of the data points currently assigned to it (movements shown by the black arrows). (e) The updated cluster assignments of the data points are obtained according to the new center locations. The steps in (c) and (d) are repeated until convergence is achieved. (f) The final cluster assignments. (B) Healthy (red) and disease affected (green) subject clustered using the Jarvis—Patrick method.

Reproduced with permission from Chen, Y.-Z., & Lai, Y.-C. (2018). Sparse dynamical Boltzmann machine for reconstructing complex networks with binary dynamics. Physical Review E, 97 (3), 032317. <https://doi.org/10.1103/PhysRevE.97.032317>.

of pre-imposed clusters) and on the choice of the metric (distance type) with which to measure the object-centroid distances (like all other clustering systems).

Jarvis-Patrick method

Jarvis—Patrick’s method (Jarvis & Patrick, 1973) (Fig. 9.16B) is very simple, although a little more sophisticated than k-means, which makes it a well performing method even if rarely used in metabolomics except in specific cases. Unlike k-means, it does not require one to impose a number of clusters at first. Its strong point is its ability to cluster objects that occur in nonglobular agglomerations.

This method is based on the calculation of all the distances between objects and on the analysis of a matrix of neighborhoods derived from the matrix of the distances, of dimension (n, L) , where L is a parameter, generally fixed between 20 and 30. For each row, the elements of this matrix represent the closest L objects (integers between 1 and n) to each object in the row. The algorithm is as follows: select the distance to be used, define the length of the list of closest neighborhoods L , define the number k of common neighborhoods ($k < L$) and calculate the distance matrix. Then, a matrix of neighborhoods is constructed, of size

(n, L) , in which a list of the closest neighborhoods L is associated with each object.

The construction algorithm of the different clusters is based on the analysis of the neighborhood matrix. Two objects a and b are placed in the same cluster if object a appears in the list of L closest to object b , object b appears in the list of the L closest to the object a , or if a certain number N of objects (different from a and b) are common to both lists. An increase of N means that a more restrictive condition is required to place objects in the same cluster, which results in an increase in the final number of clusters. Current practice suggests using one third and one quarter of the total objects as L and N , respectively. The main feature of this method is the fact that the number of clusters is calculated by the method itself, and that their dimensionality can be very variable and, above all, their distribution can be nonglobular.

Conclusion

The cluster analysis in metabolomics can be used to aggregate the different metabolites analyzed in a set of samples, and this method attempts to interpret the data by aggregating the metabolites on the basis of their coherence. An alternative is the use of cluster analysis to aggregate the different samples looking for a consistency between them that can be interpreted and generate different classes.

Owing to the great variability that the results obtained by applying cluster analysis methods can have and to the inevitable subjectivity in evaluating the results, it is essential to adopt some general criteria. In particular, if the objective is to restructure the data by assigning each object into a class (not known previously), the most important aspect is that of being able to interpret the resulting groups, to give them a coherent meaning with the problem under consideration. It does not matter, therefore, that few or many groups have been found, what matters is that it is possible to give them meaning.

Another concern of cluster analysis is the need to reduce redundancy of information present in the data. This happens when we have a large number of objects, many of which represent situations that are very similar to each other: in this case the different situations are represented in an unbalanced way and the information linked to the less represented situations could be masked by the others. In situations of this type, it is not as important to interpret the groups that result from the application of a cluster analysis technique, as it is to consider several groups large enough to sample all the space in the most representative and exhaustive possible way.

Supervised machine learning

Introduction

The unsupervised ML algorithms we have discussed before are not directed to recognize pre-existing classes, but are trained to recognize the natural occurrence

of clusters within a population of data. For this reason, they are also referred to as “data driven” algorithms. If these clusters, recognized by the learning tools, are then validated by an external operator, on the basis of previous knowledge, they can be labeled within coherent classes. Conversely, in supervised models, the occurrence class is an essential element of the training process. Based on classes, indeed, the algorithm is trained to recognize the samples on the basis of a task that is assigned to it (generally the recognition of the class of the unknown samples). For this reason, these are also defined as “task driven” algorithms. Supervised ML algorithms of this type are also referred to as classification algorithms. If the label of each sample is not a discrete characteristic but a continuous one, these algorithms are defined as regressive (Fig. 9.17).

There are several ML algorithms and they are acquiring an increasingly important role. The most relevant aspect of ML algorithms is represented by the ability to learn on the basis of the data provided to them.

What does it mean to “learn”? How can a computer be taught to perform a certain function through a mathematical algorithm? Human beings learn on the basis of experience (their own or coming from others). How it is possible to teach to a computer instead? These questions have fascinated scientists for many years. The current trend is to make these algorithms work following more or less the same logic as humans. Obviously, a computer cannot learn from its own experience, but it can be taught to learn. To learn, it must experience. The problem then shifts to the complex operation of generating an experience. This can be achieved through the combination of two mechanisms. First of all, experience generation is achieved through the “splitting of the datasets”. This consists of a partition of the total information available (the entire dataset) into two parts, one of which is used

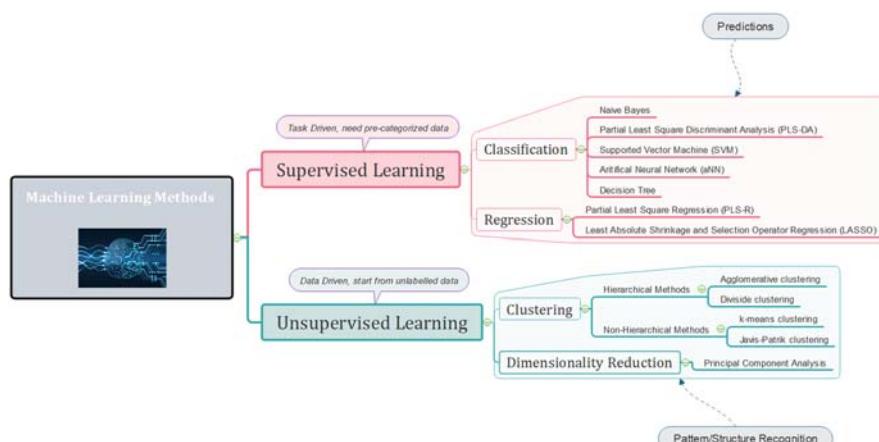


FIGURE 9.17 Machine Learning algorithms.

ML algorithms can be divided in supervised (or task driven) and unsupervised (or data driven).

to train the model and the other to test learning. This learning and testing mechanism can be reiterated many times by refining the algorithm's ability and thus increasing the accuracy in the execution of the assigned task. Another fundamental aspect of the experience generation process is filtering the initial information. It is common experience that some things strike our attention in a particular way, for example arousing our curiosity or a sense of danger, while other events are perceived as normal and are somehow excluded from our active attention. The discernment between these two types of occurrences is the fruit of our previous experience. Similarly, the classification algorithms must learn to focus attention on those elements (metabolites) that are particularly relevant to the assigned task (classification or regression) by not using computing power on those aspects (noise) that are not very or not at all relevant. As shown with regard to PCA, loadings assessment can be a useful tool for this aspect of supervised ML training, as well as metabolite clustering. However, there are other, more sophisticated tools (such as the genetic algorithms), that we will describe below that can optimize this phase of the learning process.

The reiteration of the learning process on the training sub-dataset and of the testing of predictive capabilities on the testing sub-dataset requires a definition of the concept of learning in ML and then a metric to evaluate the level of learning. In other words, the algorithm needs to measure how it has learned and how well it has learned.

A supervised ML system should be considered capable of learning if its classification capacity is greater than that which can be achieved using simple class occurrence. For example, if a dataset contains information on an equal number of subjects from two classes, the probability that a sample belongs to one class is 50%. The algorithm is trained if its classification metric (e.g., accuracy) is greater than 50%.

There are several metrics to evaluate learning but the one most used in metabolomics is accuracy. Learning accuracy indicates the ability of a system to correctly place a sample with respect to a pre-assigned class. In other words, it expresses the system's ability not to make classification errors. All ML algorithms can be structured both in a modeling and a nonmodeling way ([Fig. 9.18](#)).

Let's imagine we have a dataset with two classes (for example healthy subjects and disease affected subjects as reported in [Fig. 9.18](#)). By means of this dataset, an ML algorithm is trained and can then be queried to try to predict the class of a new sample (not used for training). Class membership can be established by the following two different rules. Once the point relating to the unknown sample has been placed in the graph representing the classification algorithm, the distance from the centroids of the two classes can be evaluated. In nonmodeling systems ([Fig. 9.18B](#)), the class of the unknown sample is established by assigning the sample to the class with the closest centroid. On the contrary, in modeling systems for each class a reference area is established (generally an ellipse that includes all the samples of that class). The unknown sample can be

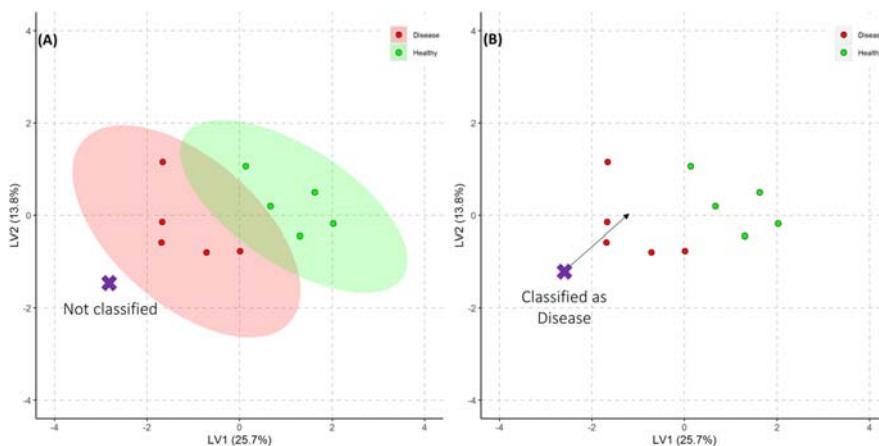


FIGURE 9.18 Modeling and non-modeling ML algorithms.

(A) A modeling ML algorithm permits the classification of an unknown sample only if it falls in the confidential ellipse of a certain class. (B) Nonmodeling ML algorithms always allows the classification of unknown samples based on proximity to a cluster centroid.

classified as belonging to a certain class only if it falls within the space defined by the reference ellipse.

The choice between a modeling and a nonmodeling approach to make the classification depends on the purpose of the training. Modeling algorithms are more conservative in class prediction and therefore are the algorithms of choice when classification accuracy is a determining element, such as when the trained system is used to make a diagnosis or to predict the response to a certain treatment. The downside of these systems is that due to the narrowness of the classification criteria some samples cannot be attributed to any class. For these samples there is therefore no answer, and this must be investigated with alternative tools.

On the contrary, nonmodeling algorithms always allow a class attribution and therefore are more permissive although the answers are in some cases riskier. These systems, however, unlike the modeling ones, are less affected by the population of data from which the training was drawn.

Decision trees

The decision tree (DT) is the simplest of the supervised ML algorithms and although it has many limitations, it is computationally undemanding and therefore often used in metabolomics especially as a basis for its ensemble version [i.e., the random forest (RF), see below for further details]. It is an algorithm that is structured in the form of a tree for which on each branch there is a decision. As a function of this decision, the tree is subdivided into further branches up to terminals that are called “leaves”. In the terminal leaf there is the model output.

Because DT is generally used as a classification system, the leaves express the class to which the sample belongs.

The datasets deriving from metabolomics experiments need some modifications to be used to train systems based on DTs. To describe the functioning mechanism of a DT training we will examine a very simple public dataset that contains the data of a series of people who have applied for a bank loan. The labeling system (class attribution) is based on whether the loan was approved or not.

The data contained in this dataset are related to age, work situation (whether the investigated subjects have a job or not), and whether or not they have a house and credit ranking. The dataset ([Table 9.2](#)) contains 15 cases (15 lines) of which 9 belong to the “Yes” class (approved loan) and 6 to the “No” class (not approved loan).

Regardless of the training of the model, the probability of approval for any loan applicant is 60%, as 9 of the 15 cases reported in the dataset were approved. The learning process can be considered valid if the trained algorithm is able to make predictions with an accuracy greater than 60%. [Fig. 9.19](#) shows several DTs trained using the data in [Table 9.2](#).

According to the tree shown in panel A, by evaluating the age of the applicants, 3 paths can be followed: one for young people, one for middle-aged and one for elderly. Young people should be assessed if they have a job or not: if they have a job then the loan will be approved (2 cases out of 2). With middle-aged ones the attention should be focused on the house ownership. With the

Table 9.2 Loan dataset: Loan dataset contains 15 cases and 4 features representing the characteristics of the people requesting a loan.

ID	Age	Job	Own house	Credit ranking	Class
1	Young	False	False	Fair	No
2	Young	False	False	Good	No
3	Young	True	False	Good	Yes
4	Young	True	True	Fair	Yes
5	Young	False	False	Fair	No
6	Middle	False	False	Fair	No
7	Middle	False	False	Good	No
8	Middle	True	True	Good	Yes
9	Middle	False	True	Excellent	Yes
10	Middle	False	True	Excellent	Yes
11	Old	False	True	Excellent	Yes
12	Old	False	True	Good	Yes
13	Old	True	False	Good	Yes
14	Old	True	False	Excellent	Yes
15	Old	False	False	Fair	No

Class membership is based on the loan appropriation.

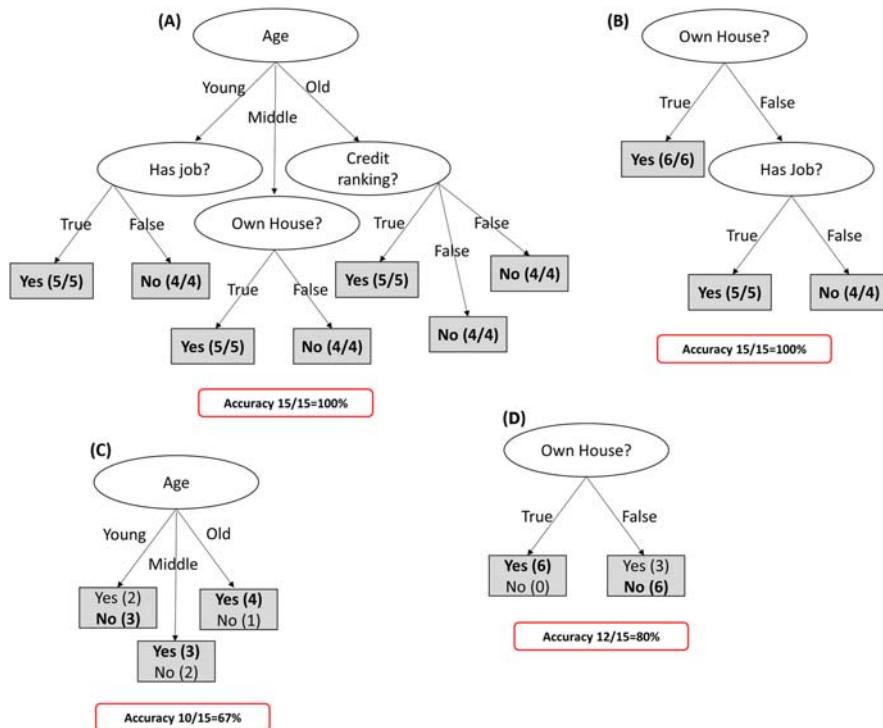


FIGURE 9.19 Decision trees.

Several decision trees trained using the dataset reported in Table 9.2. A and B trees showed large dimensions as well as 100% accuracy; on the contrary C and D trees are simpler and showed lower prediction accuracy.

elderly, on the other hand, it is necessary to evaluate the credit rating. The analysis of this tree helps us to understand the logic according to which loans are approved or not by the bank. Specifically, a young person has better opportunities to repay the loan if he/she have a job and therefore the ability to pay the mortgage payment. On the other hand, if he/she does not have a job and does not have the appropriate financial reserve, the loan cannot be repaid. For a middle-aged person, the possession of a job is less relevant, indeed the lack of work can be a temporary event, but it is more likely that this person has accumulated credits from previous jobs or has other sources of income that do not necessarily have to be linked to his/her main job. What discriminates, however, is the possession of a house; because, in this case the risk of foreclosure discourages nonrestitutions. The decisional analysis of the nodes of which the tree is constituted not only allows decision making, but also gives insight about the process behind this decision.

Migrating this logic to metabolomics experiments, this trained DT could be useful to understand whether or not a person has a certain disease or condition for what the algorithm was trained, but it is also possible to investigate the mechanisms that led to the onset of that disease/condition. As shown in Fig. 9.19, DTs are not unique, starting from a certain dataset it is possible to train different tree classifiers.

There are several systems for evaluating the effectiveness of these trees. Obviously, the classification accuracy is an important parameter but so is the tree dimension and complexity. In general, the smaller a tree is, the simpler it is to apply and the more generalizable it is, that is, it is less subject to overfitting. The tree shown in Fig. 9.19B is smaller and with fewer choices than the one shown in Fig. 9.19A while maintaining the same accuracy. This is possible because the young age choice can be included in the home parameter because from the analysis of the dataset shown in Table 9.2 only one young person owns a house and this is labeled as approved, while all the other young people do not own a house. While this is just one example it makes it clear that DTs can be structured in many different ways, choosing different nodes to start from and several ways to branch them. In order to obtain a compact tree with a good accuracy it is therefore important to carefully choose the nodes. To do this, there are several strategies used in metabolomics to estimate the effectiveness of a node:

- Information gain (IG)
- Gain ratio (GR)
- Gini index (GI)

These tools allow the evaluation of information gained from a certain type of choice. They are based on the assessment of entropy (H) according to the concept derived from the theory of information which defines it as:

$$H = \sum_{j=1}^{|C|} p(c_j) \log_2 p(c_j) \quad (9.11)$$

where $p(c_j)$ is the probability of the class c_j .

A decision-making algorithm must decrease the entropy with each decision node. The IG, in fact, expresses this reduction of entropy following a decision node. Each node can be evaluated based on its ability to create order by reducing entropy and thus allowing the separation of the different classes (Eq. 9.12).

$$IG = H(D) - H(D|a) \quad (9.12)$$

where $H(D)$ is the entropy of the system prior to the split introduced by a certain node a and $H(D|a)$ is the entropy after the split.

The GR was introduced in the late 1980s by John Ross Quinlan of the New South Wales Institute of Technology in Sydney, who theorized the *Iterative*

Dichotomiser 3 algorithm, more commonly known as ID3, for generating DTs (Quinlan, 1986). It is based on logic similar to IG but corrects this value for the intrinsic value (IV).

$$GR = \frac{IG}{IV} \quad (9.13)$$

$$IV = - \sum_{i=1}^V \frac{p_i + n_i}{p + n} \log_2 \frac{p_i + n_i}{p + n} \quad (9.14)$$

where IV represents a correction value to take into account the number and size of the branches while p and n are the objects of class P and N .

The GI (Eq. 9.15), on the other hand, is an older instrument developed in the early 20th century by the Italian mathematician Corrado Gini (Gini, 1912) which measures the inequality of a distribution. It is often used as a tool to measure inequality in the distribution of income in a country, however, it has also had a great influence in DT theory.

$$GI = \sum_{i=1}^n p_i(1 - p_i) \quad (9.15)$$

Using these different criteria (IG, GR and GI) it is possible to estimate the effectiveness of a node and then select the nodes that maximize these parameters more effectively by reducing total entropy.

The effectiveness of DTs also depends on simplicity. Over-structured trees are more prone to overfitting. This is a phenomenon that makes a classification model hyper-trained and is subsequently not very effective in its predictive action due to its poor generalizability. Overfitting is a risk that is always present in the training phase of any classification algorithm. To minimize this risk, it is crucial to keep DTs as simple as possible, that is, with the lowest number of nodes, even if this would sacrifice a little the accuracy of the model. DTs are therefore “pruned”, meaning that once a DT has been structured, some branches are eliminated from the tree by aggregating them together to make it easier/smaller. Although this, as mentioned, often involves a decrease in the overall accuracy of the model, in the testing phase using an independent dataset the accuracy/predictive ability will increase.

The pruning operation can be done either during the building of the tree and in this case, it is called “pre-pruning”. In pre-pruning the branching is terminated if, during the DT building, it is determined that the choices are no longer reliable or if the increase in accuracy resulting from a further branching is below a certain threshold. However, pruning can also be carried out after the DT construction. In this case it is defined as “post-pruning” and the nodes are aggregated assuring that there is not a reduction in the classification accuracy above a given threshold.

The dataset used for the examples reported so far contains categorical data, while the metabolomics datasets generally contain continuous data

(concentrations of metabolites). In this case, the DT type classification system can still be applied but needs to be modified. In particular, it is necessary to aggregate the quantitative data in packets, effectively making the entire dataset discrete. This operation can also be performed using the entropy reduction estimators described above. This complication makes the application of DTs on metabolomics datasets computationally more demanding than the application on naturally discrete ones although it is a valid choice, especially considering the recent advances in computing power.

Naïve Bayesian

Another interesting, supervised ML algorithm is the **Bayesian classifier**. It is very simple and computationally undemanding, often showing particularly high performance. Nevertheless, this system is rarely used in metabolomics but has a potential of great interest especially in the diagnostic applications. Unfortunately, Bayesian classifiers provide almost no information that can help to understand the role of the different metabolites and thus produce mechanistic hypotheses.

They are also called **Naïve Bayes (NB)** or idiot's Bayes classifiers and have an essentially probabilistic operating mechanism based on the evaluation of conditional probability. These classifiers are based on Bayes' theorem (Eq. 9.16). In NB classifiers the relationships between predictors (metabolites in metabolomics datasets) are ignored, that is, all metabolites are treated as uncorrelated.

$$p(g|x_i) = \frac{P_g f(x_i|g)}{\sum_k P_k f(x_i|k)} \quad (9.16)$$

where P_g is the *a priori* probability of class g and $f(x_i|g)$ is the probability density that a class g contains the object x_i .

To describe the operating mechanism of an NB classifier we could use the dataset reported in Table 9.2 which describes 15 cases of loan requests whose acceptance is based on the evaluation of 4 features.

On the basis of the frequencies reported in Table 9.3, it is possible to estimate the probability of occurrence of each of the two classes for any combination of features that a case may have. Assuming a new loan request from a middle-aged person, without a job or a house but with an excellent ranking, it can be determined:

$$\begin{cases} P(Yes|X) = P(Middle_{age}|Yes) \cdot P(False_{job}|Yes) \cdot P(False_{house}|Yes) \cdot P(Excellent_{CR}|Yes) \cdot P(Yes) \\ P(No|X) = P(Middle_{age}|No) \cdot P(False_{job}|No) \cdot P(False_{house}|No) \cdot P(Excellent_{CR}|No) \cdot P(No) \end{cases}$$

$$\begin{cases} P(Yes|X) = \frac{3}{5} \cdot \frac{4}{10} \cdot \frac{6}{6} \cdot \frac{4}{4} \cdot \frac{9}{15} = 0.144 \\ P(No|X) = \frac{2}{5} \cdot \frac{6}{10} \cdot \frac{0}{6} \cdot \frac{0}{4} \cdot \frac{6}{15} = 0 \end{cases}$$

Table 9.3 Frequency table: number of cases and relative frequencies for the cases reported in Table 9.2.

Features	Value	Loan approved (yes)	Loan not approved (no)	Total
Age	Young	2/5	3/5	5/15
	Middle	3/5	2/5	5/15
	Old	4/5	1/5	5/15
	<i>Total</i>	9/15	6/15	
Job	True	5/5	0/5	5/15
	False	4/10	6/10	10/15
	<i>Total</i>	9/15	6/15	
Own house	True	3/9	6/9	9/15
	False	6/6	0/6	6/15
	<i>Total</i>	9/15	6/15	
Credit ranking	Fair	1/5	4/5	5/15
	Good	4/6	2/6	6/15
	Excellent	4/4	0/4	4/15
	<i>Total</i>	9/15	6/15	

By normalizing the partial results, the percentages are obtained that estimate the occurrence of each of the two possible eventualities (approval or nonapproval):

$$\begin{cases} P(Yes|X) = \frac{0.144}{0.144 + 0} = 1 (100\%) \\ P(No|X) = \frac{0}{0.144 + 0} = 0 (0\%) \end{cases}$$

From which it appears that the loan will certainly be approved.

The NB algorithm, although based exclusively on a probabilistic evaluation, is still an ML algorithm because it is trained on the basis of data provided as an example (the training set). As these vary, in fact, it provides different predictions.

One of the limitations of NB algorithms consists in the assumption that the features are independent from each other. In metabolomics this is not true, there is a strict interdependence of the concentrations that metabolites can assume. Nevertheless, NB classifiers have shown good efficacy in solving classification problems (Wang et al., 2007) and also effectively managing datasets in which the number of metabolites measured is greater than the sample size because they are not very prone to overfitting (Ghosh et al., 2020).

As with DTs, NB algorithms can be adapted to situations in which the data contained in the dataset are continuous rather than discrete (like the metabolomics datasets). The solutions illustrated for the DTs (IG, GR and GI) to discretize the

data, may also be useful in this case, but the most widely used strategy is a simple division into quartiles. This also maintains the computational lightness of the algorithm.

Discriminant analysis

Linear Discriminant Analysis (LDA) and **Quadratic Discriminant Analysis** (QDA) are two classification methods widely used in metabolomics as well as in other omics disciplines (including the study of the intestinal microbiota) and have a solid and widely developed statistical foundation. These are Bayesian methods and therefore classify objects according to the following rule: an object x_i is classified in class g if

$$p(g|x_i) > p(k|x_i) \quad (9.17)$$

where $p(g|x_i)$ is the posterior probability that the object x_i belongs to class g . This probability is calculated according to Bayes' rule (Eq. 9.16). The probability density is generally unknown and must be estimated from the objects in the training set. Using Bayes' rule, Eq. (9.17) can therefore be expressed as

$$f(x_i|g)P_g > f(x_i|k)P_k \quad (9.18)$$

An x_i object is classified in class g if the Eq. (9.19) is minimal

$$D_g(x_i) = (x_i - \bar{x}_g)^T S_g^{-1} (x_i - \bar{x}_g) + \ln |S_g| - 2 \ln P_g \quad (9.19)$$

where $D_g(x_i)$ is called discriminant score, S_g^{-1} is the inverse of the covariance matrix of class g and \bar{x}_g is the corresponding class centroid.

The value $(x_i - \bar{x}_g)^T S_g^{-1} (x_i - \bar{x}_g)$ is the square of the Mahalanobis distance (see Eq. 9.9). The fundamental hypothesis underlying this classification approach is that the variables are normally distributed. Despite this requirement for normality of the variables, the discriminant analysis provides good results even in the presence of deviations from normality.

If the covariance matrices are different from each other, all covariance matrices of class S_g , are used. In this case, the hypersurfaces that separate the classes are quadratic and the analysis is called QDA.

Artificial neural network

Introduction

Artificial neural networks (ANNs) are models whose implementation was impossible until a few decades ago. Artificial neural circuits are the basis of sophisticated forms of AI, increasingly evolved, able to learn by exploiting mechanisms similar (at least in part) to those of human intelligence, obtaining results that are not possible for other algorithms.

ANNs can be applied to different areas of AI. The use as supervised classification models in the field of ML is the most widespread application in the metabolomic field. The prototype of ANNs are biological neural networks. The neural networks of the human brain are the foundation of the ability to recognize, process, interpret, and respond to the surrounding environment and its changes. The brain consists of sets of closely interconnected neurons; the constituent elements are:

- the *neuronal somes*, that is, the bodies of neurons. They receive and process information; if the action potential at the input exceeds a threshold, they in turn generate impulses capable of propagating a signal;
- *neurotransmitters*, chemical compounds of different nature (amines, peptides, amino acids), synthesized in the somes and responsible for the translation of electric impulses across a synapse;
- *axons*: the path of communication out of a neuron cell body; generally there is one per neuron;
- the *dendrites*: the main way of communication into a neuron cell body; there are multiple for each neuron, forming the so-called dendritic tree;
- *synapses*, or synaptic junctions: highly specialized functional sites where information is passed between neurons. Each neuron has thousands of them. Depending on the action exerted by the neurotransmitters, the synapses have an excitatory function, facilitating the transmission of the nerve impulse, or inhibitory, tending to dampen it. The transmissions take place when the neurotransmitter is released into the synaptic space, thus reaching the receptors of the post-synaptic membranes (i.e., the next neuron), and, by altering their permeability, transmitting the nerve impulse.

Each neuron can simultaneously receive signals from several synapses. One of its intrinsic abilities, is to measure the total electrical potential of these signals, thus establishing whether the activation threshold has been reached to generate a nerve impulse in turn. This property is also implemented in artificial networks.

The synaptic configuration within each biological neural network is dynamic. This is a determining factor in their efficiency. The number of synapses can increase or decrease depending on the stimuli that the network receives. The more numerous, the more synaptic connections are created, and vice versa. In this way, the adaptive response provided by neural circuits is more calibrated, and this is also a feature implemented in ANNs.

The first theoretical model of a rudimentary artificial neuron was developed in 1943 by McCulloch and Pitts ([McCulloch & Pitts, 1943](#)). They described an apparatus capable of receiving a series of binary input data in each of its elements, followed by a single output data for each. This machine was able to work only on elementary Boolean functions. A few years later, [Hebb \(1949\)](#) hypothesized the possibility of instructing machines with a learning that emulates the one at the basis of human intelligence. These ideas were then solidified in 1958 with Rosenblatt's proposal of the first neural network: the Perceptron ([Rosenblatt,](#)

1958) which is based on a layer of input nodes (artificial neurons) and a singular output node.

Synaptic weights (a weight indicates the strength of a connection between two nodes) are dynamic, allowing the machine to learn, in a roughly similar, though much more basic, way to that of biological neural networks. The model is feed-forward, that is, the impulses only propagate forward.

In the 1970s the multilayer Perceptron and how to set up its learning was described (Fukushima, 1975). This algorithm contains, between the input and output nodes, an intermediate layer, called hidden, where the processing of information from the input layer takes place and then sent to the output node. This is also a feed-forward network but is not linear: the connections to and from each single node are multiple.

One of the most important historical moments in the structuring of the ANN theory was the definition of the error back-propagation (EBP) algorithm (Rumelhart, Hinton, & Williams, 1986) which emulates one of the most important mechanisms underlying the learning of biological networks. EBP has proven to be so effective that it is still used today, indeed this allows optimizing the ML of a neural network in subsequent stages. It is implemented by changing the weights of the connections between nodes until optimal output is obtained. No less important, in this sense, was the introduction of reciprocal connections between neurons whose weight must increase only in case of convergence between the two pre- and post-synaptic values.

In addition to the feedforward impulse propagation mechanism, Hopfield in 1982 proposed a feedback architecture (since then also referred to as Hopfield networks) (Hopfield, 1982). In these networks, information between nodes travels in any direction: forward, backward and between neurons in the same layer.

Artificial neural networks training

In artificial networks the automatic learning process is obviously simplified compared to that of biological networks. There are no analogs of neurotransmitters, but the pattern of functioning is similar. The nodes receive input data, process them and, depending on the quantity and type of input, are able to send the information to other neurons by activating them (or blocking them). Through more or less numerous cycles of input-processing-output, in which the inputs have different variables, the nodes become able to generalize and provide correct outputs associated with inputs that are not part of the training set (Fig. 9.20).

Specifically, the ANNs emulating biological neuronal networks are equipped with a series of decision-making elements deferentially called neurons, which are linked together through connections. Each neuron has a so-called activation function (which is a mathematical equation), whereby when the neuron receives a series of stimuli, from other neurons, exceeds a certain threshold of the activation value, generates an impulse which is sent to another decision-making element (another neuron). So, the characteristic elements of an artificial neural networks are these:

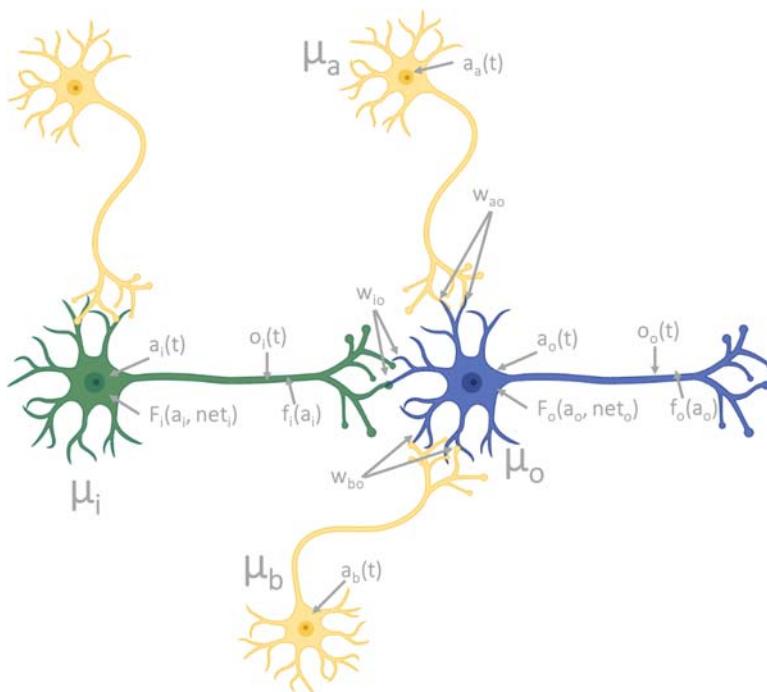


FIGURE 9.20 Artificial neural network.

μ_i represents the input neuron, while μ_o represents the output neuron. μ_a and μ_b are neurons of the second layer. $a_i(t)$ and $a_o(t)$ represent the value of the activation function of the input and output unit, $F_i(a_i, net_i)$ and $F_o(a_o, net_o)$ represent the activation functions of the input and output neuron, $f_i(a_i)$ and $f_o(a_o)$ are the input and output functions, $o_i(t)$ and $o_o(t)$ represent the outputs of the input and output neuron, respectively and w_{io} , w_{ao} and w_{bo} represent the weight of the connection between the units represented by the input neuron and the second layer, respectively.

1. a set of processing units (neurons);
2. a state of activation;
3. an output function for each unit;
4. the connections between the units;
5. a propagation rule to allow the transfer of output values through the network of connections;
6. an activation function to combine the inputs with the current activation value and produce a new activation level;
7. a learning rule to modify connections.

Learning allows the generation (and modification) of connections and the setting of activation thresholds based on the data contained in the training dataset. The human cerebral cortex has 6 layers of neurons, ANNs can be created with

any number of layers. The minimal amount of layers is two, one containing the neurons that collect information from the training dataset and one containing the neurons that generate the output (the classification). Such a network is called a linear associator. This type of neural network is very simple, but also inefficient. The neural network is more efficient the better the internal mechanism that propagates information is. Layers other than the input and output are called hidden layers. The number of internal layers is a hyper-parameter of the algorithm so it represents a value that cannot be learned but must be chosen *a priori*.

The ANN neuron has an internal function similar to biological responses to excitatory or inhibitory neurotransmitters. The connections, generally indicated with the letter w , as with biological neural networks, can be excitatory, inhibitory, or can be absent. A connection is excitatory when it inputs a positive value to that neuron's activation function ($w > 0$). Conversely, a connection that enters a negative parameter to that connection is inhibitory ($w < 0$). Connection absence could also be represented by a $w = 0$ connection weight.

Thus, the input of negative values tends to decrease the activation value, on the contrary, a positive value tends to activate the neuron and continue this information toward the next node. The training process plastically modulates the weights and the number of these connections.

The activation state is the result of the activation function and is the value that determines whether or not that particular neuron will propagate the signal to the next neuron. The activation state is generally of the boolean type (on/off). Biological neural networks have much finer transmission mechanisms with signal propagations at different levels. ANNs have recently been developed that have similar acute propagations. Obviously, these networks are computationally very demanding and more difficult to train.

With regard to the activation functions, several alternatives have been proposed. The most used are the sigmoidal functions, which allow almost constant activation values up to a certain level where a rapid switch of the activation value occurs. As already described, the training process creates connections between neurons and estimates their strength. This is determined and synthesized by means of the "association matrix" in which the weight (w) with respect to that given neuron is described for each individual input parameter, which can be positive, negative or null (indicating the absence of connections).

Conclusions

Neural networks are extremely powerful ML algorithms that are receiving increasing interest in many areas of use, including metabolomics. In supervised learning, the network is provided with a set of inputs to which known outputs correspond (training set). By analyzing them, the network learns the link that unites them. In this way, ANN learns to generalize, that is, to calculate new correct input-output associations by processing inputs external to the training set.

As the machine processes output, it proceeds to improve its responses by varying the weights between the neurons' connections. Obviously, the weights that

determine the correct outputs increase and those that generate invalid values decrease.

The use of the various types of neural networks arises from the important advantages they have, including:

- high parallelism, which allows large amounts of data to be processed relatively quickly;
- noise tolerance, that is, the ability to operate correctly despite inaccurate or incomplete inputs in many cases;
- adaptive evolution: a well implemented neural network is capable of self-updating in the presence of changes in the dataset.

However, ANN have limitations, the most important are:

- “black box” type operation. A notable handicap of ANNs is the fact that their computation cannot be fully analyzed. By this, we mean that they are able to provide correct or sufficiently correct outputs, but they do not allow us to examine the single stages of processing that determine them. In the metabolomic field it means not being able to investigate the relationship mechanisms of the different metabolites and/or the reconstruction of the specific pathways alteration;
- it is not possible to have *a priori* certainty that a problem will be solved;
- the learning time is often very high. The necessary iterations depend on factors such as number and complexity of the input variables, algorithm used, etc. Recently, important progress has been made in this area, and it is reasonable to hypothesize that in the future the learning time will be further reduced. In general, however, ANN algorithms are computationally very demanding.

Support vector machine

Support vector machines (SVM) are a set of supervised learning methods often used in metabolomics as a classification algorithm. The development of these algorithms is due to Vladimir Vapnik and his team at the Bell AT&T laboratories, that worked on them assiduously during the 1990s ([Cortes & Vapnik, 1995](#)).

The data included in a training matrix can be thought of as elements of a multidimensional Cartesian graph in which the number of dimensions is equal to the number of analyzed metabolites. In reality, this cannot be represented graphically, but conceptually, it can be considered the same as a graph in 2 or 3 dimensions. Within such a hyperspace, the data can be linearly separable or not. In [Fig. 9.21](#), a two-dimensional example is shown for simplicity.

The lines (the hyperplanes in multidimensional representations) that separate objects belonging to different classes are not unique ([Fig. 9.21C–E](#)), on the contrary they are potentially infinite. SVM algorithms determine which of these infinite hyperplanes is the best choice. The objective is to maximize the margin

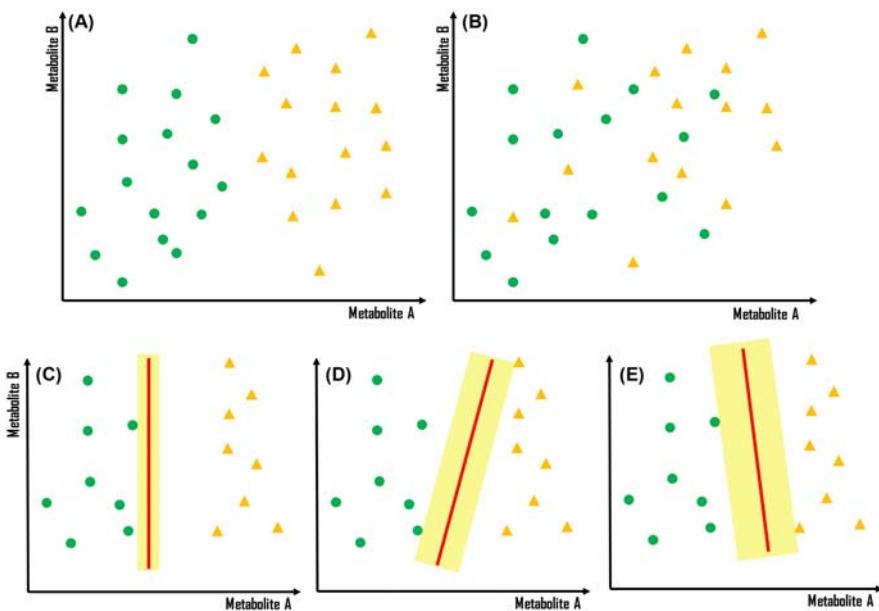


FIGURE 9.21 Separating hyperplanes.

(A) Linear separable data. (B) Nonlinear separable data. C-E Lines (in red) and margins (in yellow) separating samples of the two classes. SVM algorithms find the separating hyperplane maximizing the margins.

thereby increasing the classification accuracy. The margin is defined as the distance between the hyperplane and the objects of the two classes closest to it. These are also referred to as support vectors.

The separation hyperplane can be described with the equation $w \cdot x = 0$, where all the x_n , are the data closest to this plane. Two considerations are possible to simplify the subsequent calculations:

- w can be normalized making $|w \cdot x_n| = 1$. In particular, this product (considered in its absolute value) will take value 1 if x_n is positive – 1 if it is negative;
- The term ω_0 is taken out of w such that it becomes $w = (\omega_1, \dots, \omega_d)$. Moreover, let define $\omega_0 = b$. The equation of the plane then becomes: $w \cdot x + b = 0$.

The cornerstone of the SVM algorithm is the estimate of the distance between x_n (the support vectors) and the plane $w \cdot x + b = 0$ where $|w \cdot x_n + b| = 1$. To find this distance it is important to consider that the vector w is perpendicular to the plane in the input data space.

To calculate the distance of a point x_n from the plane, any point x belonging to the plane can be chosen to determine the projection of $x_n - x$ on w . To find

this projection the versor of w (which is equal to $\hat{w} = w/\|w\|$) is multiplied by the vector $x_n - x$ (Eq. 9.20).

$$\text{distance} = |\hat{w}(x_n - x)| = \frac{w}{\|w\|}(x_n - x) = \frac{1}{\|w\|}|wx_n - wx| = \frac{1}{\|w\|} \quad (9.20)$$

In the last step we considered that $wx_n = 1$ and $wx = 0$, while the absolute value was used because it is assumed that the distance is always positive.

The idea about maximizing the margin was introduced by Vapnik in 1992 (Boser, Guyon, & Vapnik, 1992). In its simplified version, SVM algorithms look for a separating hyperplane between the two sets of points to maximize the distance between this hyperplane and the closest points of the classes to be divided. The hyperspace in which all the data of the training set can be represented is then divided in two by the hyperplane. This is selected by maximizing the margin (the distance) between it and the support vectors (the objects closest to the hyperplane).

This strategy is relatively simple and effective but has 2 limitations. The cases in which the samples are not linearly separable, that is, those cases in which no hyperplane is able to divide the hyperspace to aggregate all the samples belonging to a class in a different fraction, require a modification of the algorithm. In particular, there are two possible strategies for dealing with these data (see below). Furthermore, the algorithm is designed specifically for two-class problems. Several algorithms deriving from the original SVM have recently been proposed to deal with multiclass problems (Hsu & Lin, 2002), although this is still a developing field of study because the performances of these algorithms are, on average, lower than those obtained with other MLs that routinely naturally deal with multiclass problems.

Nonlinear separable data

Nonlinear separable data cannot be addressed with SVM. In fact, applying the algorithm for linear separable data to a case with nonlinear separable data, no solution will be found. For example, if trying to draw a line to separate the yellow objects from the green ones represented in Fig. 9.21B, the line margins would be violated by several elements. The simplest strategy for resolving this violation is to allow it by making the margin more flexible. This can be done by introducing a new variable (ξ_n , also called slack variable), changing the assumptions to inviolability, in particular:

$$\begin{cases} x_n w + b \geq 1; \text{ for the class } y = 1 \\ x_n w + b \leq 1; \text{ for the class } y = -1 \end{cases} \quad \begin{cases} x_n w + b \geq 1 - \xi_n \\ x_n w + b \leq 1 + \xi_n \end{cases} \quad (9.21)$$

If a generic datum x_i is classified incorrectly, then it corresponds to a slack $\xi_i > 1$, because to be incorrectly classified it must necessarily lie beyond the separator hyperplane. We can therefore consider the total violation as $\sum_n \xi_n$. We are faced with a new optimization problem, in many ways similar to that seen in the case of separable data (Eq. 9.22):

$$\min \frac{1}{2} w^2 + C \left(\sum_{n=1}^N \xi_n \right)^2 \quad (9.22)$$

The value of C is chosen by the user and represents the weight to be attributed to errors (a large value of this hyperparameter corresponds to a high error penalty and it will therefore be preferable to have the least possible error).

An alternative to linear separation with the introduction of the slack function is the search for a nonlinear solution. Some data that are nonlinearly separable in a certain hyperspace can be separated by means of a linear solution in a hyperspace with a greater number of dimensions. It is therefore possible to attempt a dimensional expansion of the data matrix to search for a new linear solution. The function that transforms data into a dimensionality that makes it linearly separable is called kernel. The mathematics behind this operation is quite complex and beyond the scope of this text. Conceptually, however, the operation can be represented as in Fig. 9.22.

The class represented by the yellow samples in Fig. 9.22, is encapsulated within the population represented by the green samples. Any line that tried to separate these samples, however the support vectors were chosen, would allow at least 3 classification errors. The kernel operates in this case by multiplying the values assumed by the individual samples by themselves and placing this result on a new axis (a new dimension). In this way, the encapsulated samples build the bottom of a parabola in the new reference space. These, therefore, can be separated from the others by means of a line that uses 3 support vectors to maximize the margins (Fig. 9.22A–B). A similar approach can also be followed for a system consisting of the observation of two variables by expanding the space in 3 dimensions by means of the kernel function. A similar procedure can be generalized to any other dimensionality of the input data.

Regressive models

So far, we have described supervised ML algorithms that learn to recognize distinct classes of samples starting from characteristics inherent in the quantitative or semi-quantitative data of the complex of metabolites that describe each sample. We have defined these ML algorithms as classification algorithms. In addition, there are ML algorithms capable of recognizing and determining continuous quantitative characteristics of the different samples. These algorithms are referred to as regressive algorithms.

The most used regressive algorithms in metabolomics are:

- Partial Least Square Regression (PLS-R)
- RIDGE Regression
- Least Absolute Shrinkage and Selection Operator (LASSO) regression

The PLS-R algorithm dates back to 1982 thanks to the work of Prof. Hermann Wold (Jöreskog & Wold, 1982). The LASSO algorithm, on the other hand, was

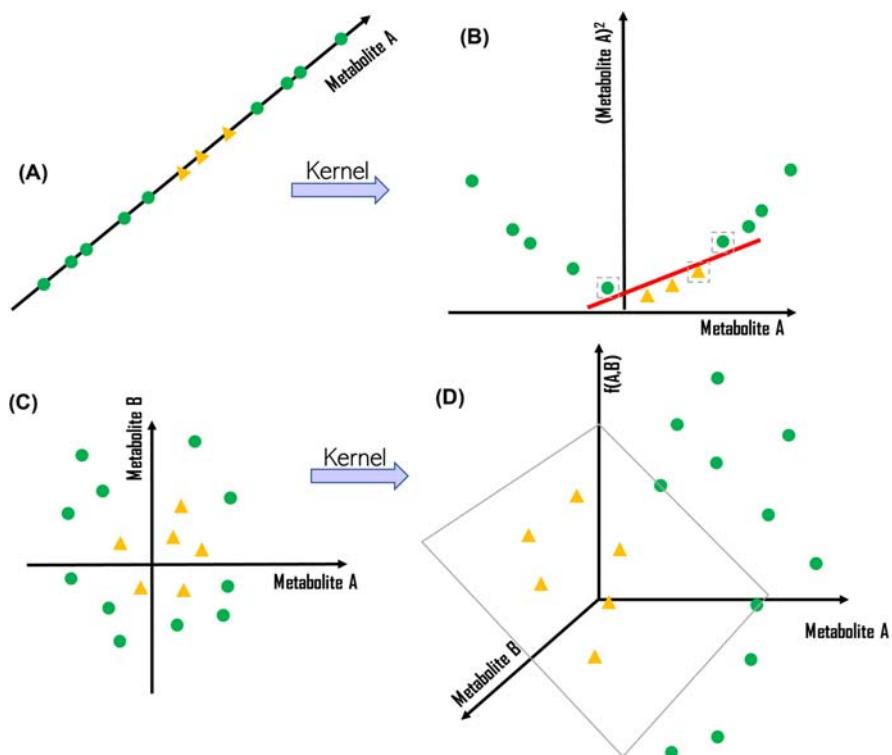


FIGURE 9.22 Kernel-based dimensional expansion.

Mono-dimensional (A) and bi-dimensional (C) nonlinear separable data can be represented in a higher dimensional space (B and D) applying a kernel function. This allows the use of a linear function to separate the data belonging to the two different classes.

formalized for the first time in 1996 by Prof. Robert Tibshirani at the University of Toronto in Canada (Tibshirani, 1996). This can be considered a modified version of the RIDGE regression algorithm, described in 1970 by Hoerl and Kennard (1970).

Partial least square regression

Partial Least Square (PLS) algorithms are currently the most important and widely used tools in metabolomics. They can be considered the supervised version of PCA. The regressive variant, known as PLS-R (from Partial Least Square Regression), was proposed in the early 1980s, by Prof. Hermann Wold, a teacher of statistics and economics at the University of Gothenburg (Jöreskog & Wold, 1982). The regressive version of the PCA, like its classification counterpart, is based on the covariance matrix between the features contained in the dataset. The

supervision of the PLS is carried out by creating the covariance matrix between the features of the dataset (X) and a matrix of the responses (Y) which contains, for each analyzed sample, the quantitative data that represent the training element (the task). In this way, the PCs of Y are rotated to maximize their correlation with the PCs of X .

The rotated components no longer coincide with the PCs, because the selection logic is different (no longer based on the ANOVA but on the correlation with the response matrix) and are called latent variables (LV). PLS techniques generally show better performance than PCA-based techniques because of the supervision mechanism in the training phase. The PCs, indeed, represent the directions of maximum variance, these directions do not depend on the “purpose” of the training, which is instead represented by the Y matrix, but are uniquely determined by the X matrix. The PCs, therefore, are not significant, per se, for the construction of a model that estimates Y .

The PLS-R algorithm is an analysis technique that combines the characteristics of the PCA with linear regression. This technique becomes very useful when it is necessary to predict a set of dependent variables starting from a large number of independent variables (also called predictors, that in metabolomics are of course metabolites). The process can be described by considering the following general conditions.

Let us consider an experiment characterized by M observations. For each of these observations, the values of the dependent variables (Y) and of the independent variables (X) are recorded. The goal of the PLS is to predict the Y variables starting from the knowledge of the independent variables X alone. When the data contained in the X matrix are colinear, the matrix will be singular and the solution to the problem cannot be solved with linear regression. Several approaches have been developed to solve this type of problem. One approach is the principal component regression which reduces the set of predictors, using instead of X , the PCs obtained from the same matrix by applying PCA. The choice of PCs eliminates the problem of co-linearity, but does not solve the problem of choosing the optimal subset of predictors of the matrix X . Indeed, the PCs of X may not be sufficient to find the maximum correlation with Y .

The PLS, on the other hand, finds that set of components of X , (the LVs), which represent a decomposition of the same matrix which at the same time maximize their correlation with the dependent variables Y . In formulas, X is decomposed as follows:

$$X = T \cdot P^T + E \text{ with } I = P \cdot P^T \quad (9.23)$$

where I is the identity matrix, T is the scores matrix, P the matrix of loadings and E the matrix of residuals. The matrix T contains the new coordinates of X in the new space described by the LVs.

In the same way we can decompose Y using the same matrix of the scores T of X :

$$X = T \cdot C^T + F \text{ with } I = C \cdot C^T \quad (9.24)$$

where C is the matrix of the loadings and F the matrix of the residuals. The purpose of the PLS is to find the matrices T and C to maximize the covariance between X and Y .

Geometric interpretation of the partial least square regression

The PLS-R is a projection method. Indeed, this technique can be considered as a projection of the vectors of the matrix X into a subspace whose dimension is defined by the k LVs that maximize the covariance between X and Y . The coordinates of this projection represent good predictors of Y as indicated in Fig. 9.23.

The prediction error in partial least square regression

Contrary to what happens for PCA-based regression (called Principal Component Regression or PCR), the PLS methods are prone to overfitting, that is, the error of the training set (in regression also called calibration set) is lower than the error

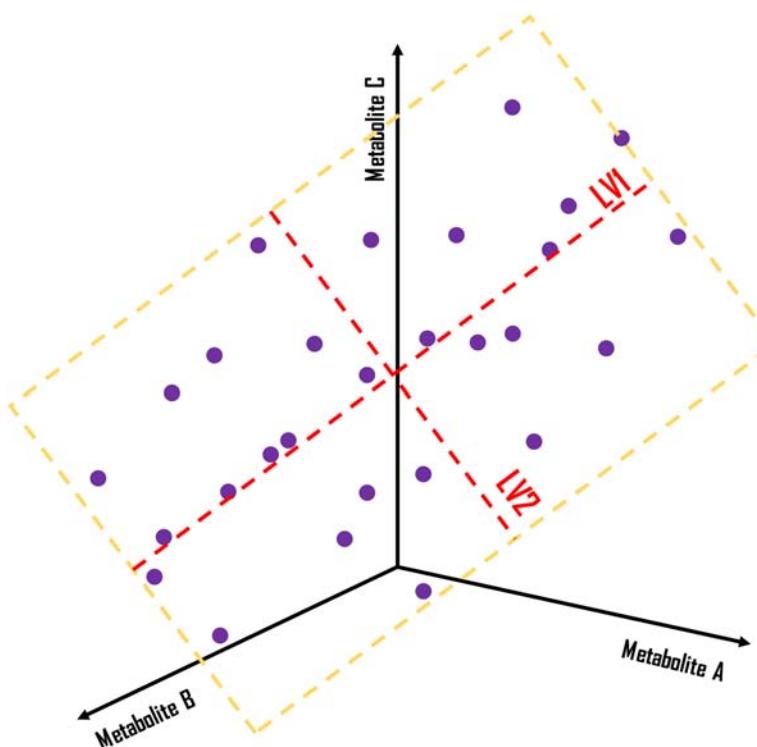


FIGURE 9.23 Partial least square regression.

Geometrical interpretation of the Partial Least Square Regression. The latent variables LV1 and LV2 represent the dimensions of the new subspace where the samples (purple dot) are projected.

related to validation data. In the case of PLS-R, overfitting is closely linked to the number of LVs used in the model. In particular, overfitting results from the decomposition process in which, proceeded by projections on orthogonal subspaces, the model is adapted to the training data, reducing the error.

To avoid overfitting, it is therefore necessary to optimize the number of LVs through a cross-validation process (see below). The goodness of the model obtained with the PLS-R can be evaluated by calculating the mean square error obtained during the training phase (Root Mean Square Error of Calibration: RMSEC) and during the validation phase (Root Mean Square Error of Cross Validation: RMSECV).

The calibration error decreases monotonically as the number of LVs increases, while the validation error has a minimal value. The monotonous trend of RMSEC indicates how by increasing the number of LVs used by the model, the noise of the data is also incorporated. RMSECV, on the contrary, provides the prediction error trend. In the prediction phase the amount of noise of the data will certainly be different, for this reason the part of the model that fits on the noise of the calibration data certainly will cause an error when used with data with different noise.

The minimum value of RMSECV precisely indicates the number of LVs for which the description of the deterministic part of the data is maximum and beyond which, the model begins to represent the noise of the calibration data. The number of LVs at which the minimum of RMSECV is obtained, therefore indicates the optimal LVs to consider for the model.

As with PCA, PLS loadings provide information on which metabolites (columns of the X matrix) contribute most to each single LV and consequently to the entire model. Based on this information, the number of variables can be reduced by reducing the complexity of the model while keeping the performance unchanged.

RIDGE regression

RIDGE regression can be considered a supervised ML version of ordinary least squares (OLS). This is also known as the least squares method. It consists of the creation of a regression function between the data, such as to minimize the distance between the points and the line represented by this function. In RIDGE regression, as in any other supervised ML algorithm, the data are initially divided into two groups, one will be used to create the regression function while the other will be used to test its effectiveness. In this case, the confusion matrix and accuracy are of no help in estimating the performance of the function obtained as the data is of a continuous and nondiscrete nature. A tool known as residuals sum of squares (RSS) is therefore used, which can be evaluated by means of Eq. (9.25):

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (9.25)$$

The regressive algorithms consist of the estimation of the β values such as to minimize the value of the RSS. On the contrary, the basic idea of RIDGE regression is that not necessarily a function that minimizes the RSS on the data used for training will be as effective in describing the relationships on data used to test the model. For this reason, a disturbing element is introduced in the process of selecting the β value that minimizes the RSS, called the “penalty term” or “tuning parameter” and indicated with the Greek letter λ (Eq. 9.26). Fig. 9.24 shows the graphic effect of this parameter:

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (9.26)$$

It is evident that having a $\lambda = 0$ means not having a penalty in the model. This produces the same estimates of β that would be obtained with Eq. (9.25), using the method of OLS. On the contrary, $\lambda \rightarrow \infty$ (i.e., very large) means having a high penalty effect, which will bring many coefficients to be close to zero but will not imply their exclusion from the model. It can also be noted that by increasing the tuning parameter (λ) there is less flexibility of the model, which will result in a smaller variance, but a greater bias. An optimal choice of λ finds the right compromise between bias and variance.

In other words, paradoxically, the regression function estimated with the RIDGE method produces a worse modeling performance on the training data than the OLS function, but it gains in predictive capacity toward the testing data and therefore in generalizability.

If the number of predictors is high, but less than the sample size ($p < n$), the use of OLS may encounter some difficulties due to the high variability. On the other hand, if the number of predictors is greater than the number of analyzed samples ($p > n$) (which represents the norm in metabolomics) the OLS cannot be used. Either way, the RIDGE method is effective. Again, the RIDGE regression method, never allows the exclusion of β estimated similar to 0 from the model. This limitation, from the point of view of the accuracy of the estimate, may not be a problem but it greatly reduces the interpretability of the resulting regressive function.

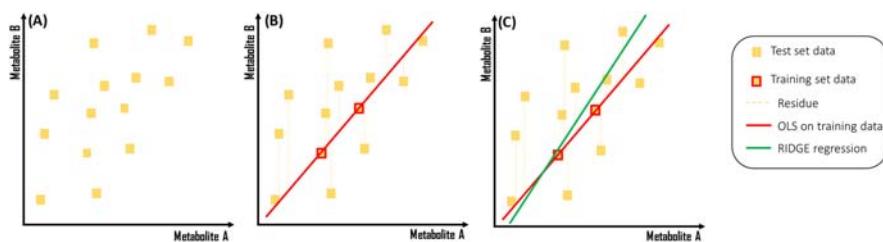


FIGURE 9.24 RIDGE regression.

RIDGE regression can be considered a variant of the OLS regression in which a penalty parameter allows to suboptimize the estimation of the parameter β , reducing the RSS evaluated on the validation set.

Least absolute shrinkage and selection operator regression

The LASSO algorithm overcomes the interpretative disadvantage of the RIDGE regression. It allows the coefficients of the estimated β to be excluded from the model when they are zero. It can be noted that the formula of the RIDGE regression (Eq. 9.26) is very similar to that of the LASSO (Eq. 9.27), the only difference consists in the structure of the penalty; for the LASSO it is necessary to consider the sum of the absolute value of the β .

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (9.27)$$

In the case of the LASSO regression, the penalty has the effect of forcing some of the coefficient estimates, those with minor contributions to the model, to be exactly equal to zero. This means that the LASSO is also inherently a feature selection tool (see below). This allows, on the one hand, to obtain simpler models because they are based on fewer variables and, on the other hand, to evaluate which are the most relevant features (metabolites) with respect to the phenomenon being studied (see also Chapter 10: Relevant Metabolites' Selection Strategies for further details).

In general, LASSO regression tends to perform better in situations where some of the predictors have high coefficients and the remaining predictors have very small coefficients. In other words, when a significant portion of the information is contained in a limited number of metabolites. On the contrary, when the information is widely dispersed, the RIDGE regression is more effective because the final model will be a function of many predictors, all with coefficients of comparable size.

Partial least square discriminant analysis

The most widely used ML algorithm in metabolomics in terms of classification is the Partial Least Square Discriminant Analysis (PLS-DA). This was introduced in 1987 by the Swedish chemist Svante Wold ([Ståhle & Wold, 1987](#)). This ingenious classification algorithm originates from the PLS-R, which was proposed 5 years earlier by Svante's father, Prof. Hermann Wold ([Jöreskog & Wold, 1982](#)).

The PLS method, as described in the previous paragraph, was born with regressive purposes and therefore to manage continuous variables. On the contrary, the PLS-DA version is built as a discriminant analysis performing classification tasks, thus managing output variables of a categorical type. In a binary classification problem, the Y matrix of the responses is recoded to consist of only two integers. Typically, 0 (or sometimes -1) and +1 are used to mean "out of group" and "in-group", respectively. These recodified class values are also known as dummy Y. The PLS-DA algorithms can also handle multiclass problems, in this case the response matrix is expanded in a sort of hierarchy,

dividing this problem into a series of two-class problems. This is possible because, as already mentioned, for PLS regression the response matrix for this type of algorithm does not necessarily have to be composed of a single column (Fig. 9.25).

The PLS-DA algorithm can be built using two different strategies called PLS1-DA and PLS2-DA. The former models one class at a time while the latter models multiple classes at the same time. Traditionally, binary classification problems preferentially use the PLS1-DA strategy. On the other hand, a multiclass problem is often modeled using the PLS2-DA algorithm, but this is not a forced choice, indeed different PLS1-DA algorithms, after hierarchizing the problem in a series of boolean problems, can lead to similar results.

The training of a PLS-DA model involves several stages. First, the weight vector (w) is estimated by maximizing the covariance between the matrix containing the data (X) and that of the responses (Y). Subsequently, the X-scores (t), the X-loadings (p) and the Y-loadings (q) are determined in sequence. Finally, the regression coefficient (b) is estimated using the resulting w , p and q .

Next, the first set of LVs and loadings are established. Then the residuals X (res_x) and Y (res_y) of the first LV become the input data (X) and the output data (Y), respectively, to construct the second LV. This procedure is repeated for all LVs needed to build the desired prediction model. The weight vector is

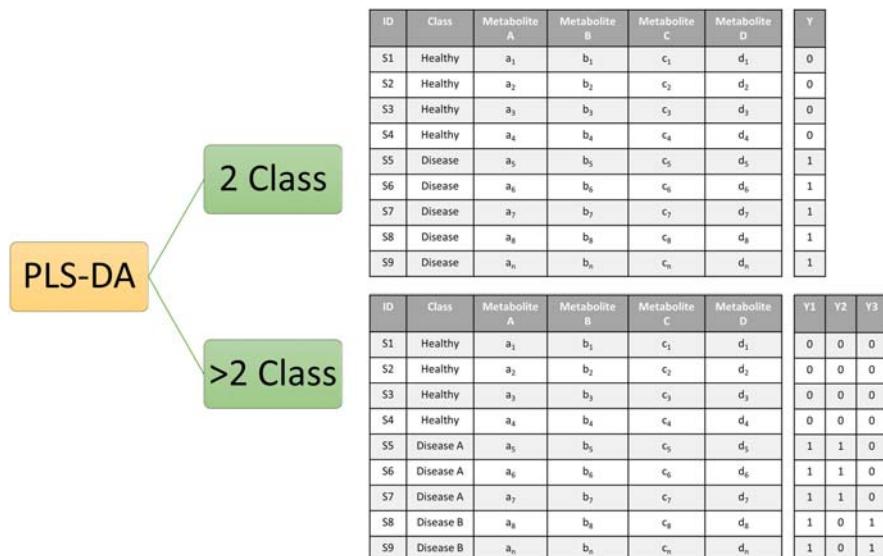


FIGURE 9.25 Partial least square discriminant analysis.

For two-class problems PLS-DA codifies the Y response matrix in a dummy Y containing a binomial value associated with the class. For multi-class problems Y matrix is split in several binomial problems in which 0 means “out-of-group” while 1 means “in-group.”

normalized in the second step. LVs, as with all supervised models, are constructed using training samples. At the same time, a regression coefficient matrix, (B), is also prepared for subsequent forecasting purposes.

For prediction purposes, the test set (X_{test}) is reduced to the new dimensions (the LVs) via B to produce the predicted values (Y_{pred}). Given a set of training data belonging to G classes, the PLS-DA algorithm produces Y values for each sample in the test dataset.

As has just been described, the correct class membership, that is, the expected value (Y_{pred}), should be “+1” or “0” to mean “in-group” or “out-group”, respectively. However, in practice, the resulting predicted values often take any value between 0 and 1 rather than an integer. For this reason, various decision rules have been proposed to translate the predicted value into meaningful class membership.

In practice, the different decision rules can be divided into two large groups: end point and fixed point. In the first group, the X-scores or Y_{pred} can be used together to determine the class membership of the samples in the test dataset. On the one hand, the X-scores could be converted to a particular distance metric (e.g., Euclidean or Mahalanobis distance), where the sample would be assigned to the class that presented the minimum distance. On the other hand, Y_{pred} could be used in its raw (naive) form or manipulated into a posterior probability form via a particular probability density function. Fixed point-based approaches use Y_{pred} only and may involve a fixed point (i.e., the cut-off value) or two fixed points (i.e., the boundary line). The optimum point or points can be determined arbitrarily or according to a particular diagnostic tool, for example, Y_{pred} graph, receiver operating characteristic (ROC) curve and probability density function. However, in most cases, the cut-off point is determined midway between the means of the two groups ($Y_{\text{pred}} = 0.5$).

The PLS-DA lacks a global function to be optimized to determine the fit of the model, for this reason it is essential to use validation tools to be able to evaluate its classification performance and the presence of overfitting (see below). Furthermore, being essentially based on a covariance matrix (such as PCA), although this is evaluated between the training data X and the response matrix Y , the PLS algorithms are dramatically affected by the mean value and variance of the individual metabolites. Indeed, metabolites that have higher averages and variances, naturally tend to drive and influence the training as well as the building of LVs in a decisive manner. To avoid this distortion, it is essential to always subject the data matrices to normalization, centering and scaling (or very often to auto-scaling) to standardize the means and variances of the individual metabolites, thus leaving the selection of variables to be based only on their covariance with Y and not on their absolute value of the mean and variance.

Latent variables

LVs are pivotal elements of PLS-based algorithms. We have already mentioned their similarities, and at the same time, their differences compared to the PCs

generated by the Hotelling transformation. The LVs are crucial elements not only for training and therefore in the optimization of diagnostic or forecasting performance (in the case of regressive algorithms) but also for the interpretation of models to generate new knowledge from them. An LV can be thought of as a feature not specifically registered, declared, or manifested in the original dataset. In other words, these are not clearly defined within the datasets although they represent a condensed amount of information. For this reason, they are also often referred to as “hidden variables”.

In order to understand its deeper meaning, let us give a very simple example. Think of the relationship that exists between damage caused by a fire and the number of firefighters involved in the management of the same fire. It might seem correct to speculate that the greater the number of firefighters involved in the management of a certain fire, the less damage from the fire. Indeed, a greater deployment of forces provides more energy for extinguishing and therefore helps to minimize the damage. On the contrary, the analysis of this type of relationship based on real observations shows an opposite relationship. In particular, the greater the number of firefighters involved in managing a certain fire, generally, the greater the damage reported. This apparent contradiction can be easily explained considering that the strategic management of resources tends to concentrate greater efforts (i.e., a greater number of firefighters involved) on larger fires which are inherently likely to cause greater damage.

Our initial analysis based on a dataset that contains only 2 features (x_1 = number of firefighters involved and x_2 = damage caused by the fire) does not allow us to take into account the size of the fire because this is not available data. From the analysis of the relationship, however, it is possible to infer that there is this additional feature (precisely an LV) which alone could explain both the observed features and also their relationship (thus condensing the information).

It is therefore evident that the analysis of the LVs and the features that generated them can provide a new point of view on a certain phenomenon and therefore allow for the generation of new knowledge. This is one of the most ambitious objectives that metabolomics proposes and the LVs are the best ally to be able to pursue it. In other terms, the LVs are the best tool to build a “metabolomics perspective.”

The variables important in the projection

One of the most interesting applications of the PLS-DA models is evaluating the impact that each single variable present in the original dataset (x_j) had in structuring the LVs and therefore ultimately in training the classifier by reconstructing the Y matrix. This evaluation does not constitute an element of selection as it happens for the LASSO models; it is a simple estimate of the impact of that variable in the training process. The calculation of this importance is done indirectly, because, on the one hand the contribution of the variable x_j to the building of the component t_l is measured by the weights w_{lj} , and on the other hand, the power of

the component t_l can be measured in the explanation of the set of the variables Y through the redundancy $Rd(Y; t_l)$ value. The same weight w_{lj} , indeed, indicates an explanatory power of the variable x_j on the matrix Y which is greater the higher the redundancy $Rd(Y; t_l)$ value is. This importance is estimated with a score known as VIP (variable importance in projection) which can be evaluated using the Eq. (9.28):

$$VIP_{\alpha j} = \sqrt{\frac{p}{\sum_{l=1}^{\alpha} \sum_{k=1}^q R^2(y_k; t_l)}} \cdot \sum_{l=1}^{\alpha} \left[\sum_{k=1}^q R^2(y_k; t_l) w_{lj}^2 \right] = \sqrt{\frac{p}{Rd(Y; t_1, \dots, t_\alpha)}} \sum_{l=1}^{\alpha} Rd(Y; t_l) w_{lj}^2 \quad (9.28)$$

where $Rd(Y; t_l)$ represents the arithmetic mean of the coefficients $R^2(Y; t_l)$ of the regressions of the y_k on t_l , which determine the fraction of the variability of Y that is attributable to the linear dependence on t .

By means of Eq. (9.28) it is possible to attribute to the variables x_j an explanatory power with respect to Y . The variables having a high VIP are the most important in the reconstruction of Y .

Because the mean of the squares of the VIP scores is equal to 1, it is to be considered in the construction of Y a variable x_j that has a VIP coefficient at least equal to 1 as an important variable. It is common practice in metabolomics to use more stringent cut-offs to evaluate a metabolite as relevant on the basis of its VIP score, so it is common to use 1.5 or 2 as a minimum VIP score.

Geometric interpretation of the partial least square discriminant analysis

It is possible to imagine the construction of a PLS-DA model, as a transformation operation understood as a geometric rotation. A dataset consisting of only two metabolites can be represented by a two-dimensional Cartesian graph in which each axis represents a metabolite and therefore the combination of these can describe the position of each individual sample. A hypothetical dataset composed of 3 metabolites could similarly be graphically represented by a three-dimensional graphic system. By analogy, a traditional dataset composed of n columns (metabolites) can be represented by an n -dimensional graphic system. Obviously, such a graph cannot be drawn, but with imagination it can be made real, and it would have the same characteristics as a graph in two or three dimensions.

The PLS-DA algorithm rotates this multidimensional graph until it searches for an observation point of view, such that objects belonging to the same class (same y) come to be as close as possible to each other, while objects belonging to different classes separate as much as possible. In other terms, the aim is to maximize what is identified as Within-Between (W/B) Ratio. Once this observation point has been obtained, the objects representing the different samples are projected into a lower dimensionality space which preserves the class separation. The new coordinates of this space represent the X-scores, while the new axes represent the LVs (which therefore synthesize the information contained in several

original variables) and are specifically structured to maximize the differentiation of objects belonging to different classes (Fig. 9.26).

Orthogonal partial least squares discriminant analysis

In 2002, Prof. Svante Wold proposed a new variant of PLS-DA called Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA) (Trygg & Wold, 2002). This algorithm uses the same statistical basis as the PLS-DA but takes advantage of a mathematical filter to remove systematic variance in the dataset not related to the sample class (or dummy variable Y). This minor change offers an incredible advantage for metabolomics. Consider, for example, an experiment for the evaluation of the metabolomic signature of patients suffering from a coronary disease. From epidemiological studies it is known that this pathology affects men more frequently than women (see Chapter 7: Approaches in Untargeted Metabolomics for further details), so a traditional PLS-DA model could link the condition under study (coronary heart disease) to the condition “sex” because there is a certain analogy between these in terms of occurrence. In other words, there is an orthogonal component of the variance concerning sex because men are more likely to develop the disease than women.

The OPLS-DA algorithm facilitates removal of this orthogonal component by focusing the training on the condition of interest.

Despite its enormous advantage in terms of interpretation of the resulting model, OPLS-DA is less widespread than the “traditional” PLS-DA in metabolomics, for two reasons:

1. The elimination of an orthogonal component and even more so, of several orthogonal components, significantly increases the risk of overfitting already heavily affecting PLS-DA models;
2. The model does not allow separation of more than two classes and therefore does not solve multiclass problems.

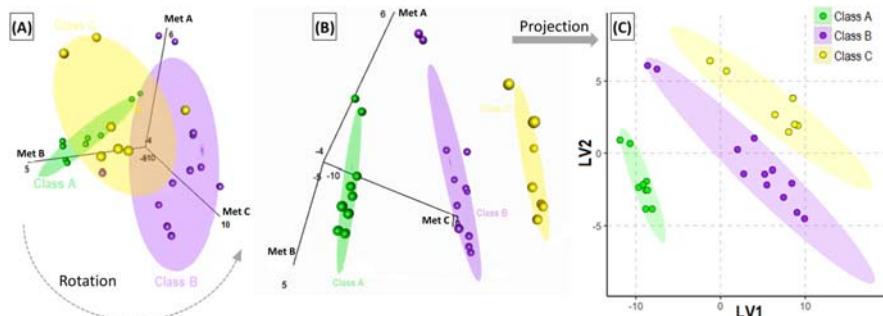


FIGURE 9.26 Partial least square discriminant analysis graphical interpretation.

The multidimensional representation of the original dataset simplified in a 3D-graph (A) is rotated searching a point of view that maximizes the Within-Between ratio (B). This freeze situation is projected in a reduced dimensionality using the LVs as a new representation space (C).

Classification model validation

The classification and regression models described so far play an essential role in the development of metabolomic field, as they summarize the state of knowledge with respect to a problem and can be used to “predict” future events. The description of the system defines the parameters and variables with which it is represented, and the available information content is intrinsically linked to the variability of the parameters that represent it. Indeed, if all the parameters that describe the system were constant, there would be no information content. This is important because the models we can build depend on the variables that describe the system. The usefulness of the models, however, is also closely linked to their ability to allow correct predictions. This ability cannot be estimated only from the training process but requires a further process called validation.

The validation of a model consists, in general, in seeking that structure of the model (its optimal complexity) that maximizes its predictive capacity. In other words, the model must have characteristics that make it sufficiently independent from the specific data used to build it.

The variance of the parameters that characterize a model is directly proportional to the complexity of the model. This is an extraordinarily important step in evaluating possible performance in the predictive phase, indeed, while an increase in the complexity of the model always increases its descriptive quality (fitting), an uncontrolled increase in complexity deteriorates its prediction performance. This phenomenon is known as **overfitting**.

As shown in Fig. 9.27, for example, an increase in the number of hidden layers in an ANN model improves the accuracy of the model (its ability to correctly classify the examples).

Indeed, an increase in model complexity increases the predictive capacity on an independent dataset, up to a certain point, beyond which its accuracy begins to decrease (yellow line in Fig. 9.27). There is therefore a maximum predictive power of the model. Further additions of hidden layers result in a deterioration of predictive power. The loss of fitting on the testing data and the divergence of the accuracy trend in the training and testing set are due to a better description of the training data trend corresponding to the increase in the number of hidden layers and a consequent loss of generalizability of the model and therefore the ability to make accurate predictions on new samples.

Therefore, the structure of the classification and regression models must always be checked by means of validation techniques evaluating the presence of overfitting, due to correlation, noise, the characteristics of the method used, the specificity of the sample, complexity of the model, etc. Validation procedures are important because if the ratio of cases studied/metabolites analyzed is low, it is possible to detect strong differences between the predictive ability of the model (it can go from 100% in fitting to 0% in prediction!). In addition, in subsequent moments, the model must be able to predict properties of new samples

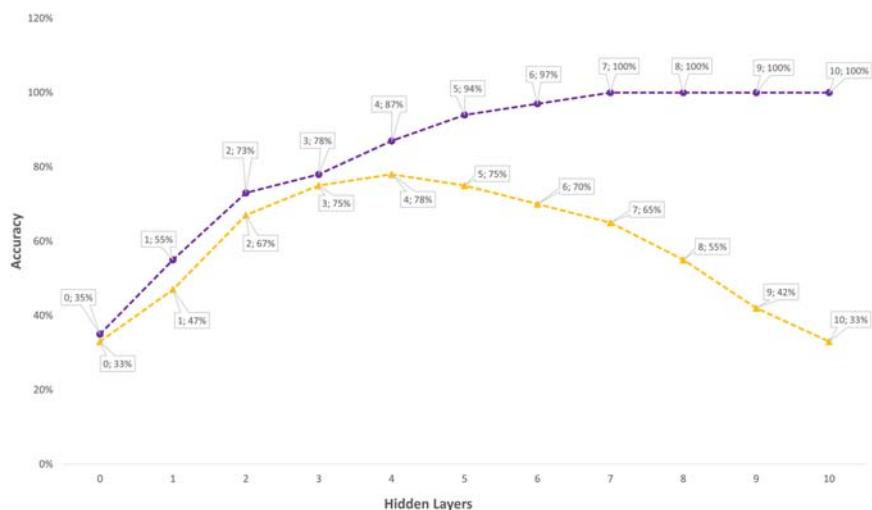


FIGURE 9.27 Overfitting.

An increase in the number of hidden layers of the ANN model improves the accuracy of the model (its ability to correctly classify the examples, *purple line*). The prediction accuracy evaluated on an independent dataset increases only up to a certain point (e.g., 4 hidden layers show 78% accuracy), beyond which accuracy begins to decrease (*yellow line*).

represented by parameters that are somehow related to the prediction problem. All the parameters calculated with procedures that do not take into account the predictive aspects have nothing to do with the performance of the same model in prediction.

From all this, it is evident that the validation process is crucial to produce a model that can provide accurate predictions even on unknown samples, and above all, to formulate hypotheses that are generalizable and not just related to the specific dataset used to train it. Choosing the most appropriate validation tool is not a simple task, indeed several validation strategies have been proposed. Among these, the one that offers the best guarantees is certainly the validation by means of an independent dataset. In other words, the best strategy for validating the data is to test the training on samples obtained from a different recruitment than the one from which the samples used to train the models were obtained. Indeed, this maximizes the generalizability of the models. This type of validation, although technically very valid, is rarely used in metabolomics due to the high costs of this type of experimentation which often involves many participants selected from different populations with different lifestyles and genetics.

Instead of an independent dataset validation, the most widely used strategy in metabolomics is the subdivision of the original dataset, obtained from a single population, into two sub-datasets called **training and testing sets**. This strategy is

also known as **cross-validation**. In this case the generalizability is slightly reduced because the subjects recruited are generally homogeneous in terms of population characteristics (diet, genetics, living environment, etc.). A further limitation of this strategy is inherent in the reduction of the number of elements available for the training phase. Indeed, the availability of samples is often the limiting element in metabolomics experiments (as in many other omics disciplines), while it is generally easier to obtain a lot of information for each individual sample. The subdivision of the samples into the two different datasets (the training and the test set), starting from a single observed population, involves a further reduction of the information on which the model can be trained.

Leave-one-out cross-validation

One of the solutions proposed to minimize the loss of information following the split of the data in cross-validation, is a technique known as leave-one-out validation (LOOCV), whereby the predictive characteristics of the model are evaluated globally by repeating the same procedure of calculation with the exclusion of one object at a time from the calculation (and therefore from the model) and evaluating the ability of the model to predict the considered property of the excluded object (Wong, 2015).

In other words, according to this procedure, given n objects, n models are built in each of which a single sample is excluded at a time: each model is trained with the $n - 1$ remaining objects and is used to predict the answer (be it a quantitative response of a regression model or qualitative response of a classification model). In order to estimate a parameter that is effectively related to the predictive power of the model, the difference between the experimental response and the predicted one is accumulated for all n objects which, in turn, are excluded from the model. As a rule, the actual final model is always calculated with all n data available, so that all the information available in all data can be exploited. This validation method is the one that involves the least perturbation as each model is always calculated using $n - 1$ data and is the one that provides the only way for a unique prediction comparison between models obtained with different methods and different descriptors. The weakness of this validation method is that it tends to make the estimated performance values in prediction converge with the value obtained on the entire dataset as the number n of samples increases. For this reason, the LOOCV method is used only in the case of low availability of samples as the predictive values become more liberal as the sample size increases.

Leave-k-out cross-validation

To overcome the limitation of LOOCV with large sample size, the leave-k-out cross-validation (LkOCV) model has been proposed (Wong, 2015). This method can be considered a modification of the LOOCV method. The data are divided into G -erasing groups such that k objects equal to n/G are assigned to the

evaluation set and therefore excluded from the training set. The model, estimated from a training set consisting of $n - k$ objects, is used to predict the response of the k objects excluded from the model. The k objects are excluded once each. Once excluded, they are reinserted into the model and other k objects are excluded. Each i -th object is then assigned to a deletion group (g), according to the equation:

$$g = \text{mod}(i - 1, G) + 1 \quad (9.30)$$

where $\text{mod}(i - 1, G)$ is the modulo operation, that is the residue of the division between the two elements $i - 1$ and G .

According to this modality, the placement of each object in the evaluation set occurs in a systematic way. The result depends on the order in which the objects are presented in the overall data matrix. An alternative to the systematic selection of objects is randomly selecting (only once) the k objects that constitute the evaluation set. In the context of the LkOCV method, the LOOCV method is an LkOCV method with $G = n$, that is, with a number of erasing groups equal to the number of objects and therefore $k = 1$. Compared to the LOOCV method, this cross-validation method is more conservative.

k-fold cross validation

A widely used alternative to LOOCV and LkOCV models in metabolomics is k -fold cross validation (Anguita et al., 2012). According to this method, starting from the n analyzed samples, k sets are generated, all n dimensional, constructed by randomly extracting n objects from the original set with repetition. This means that each data set is always n -dimensional, but each time some objects will appear multiple times in the same set, while others will not appear at all. The model obtained by the former is used to predict the responses of excluded objects. This method requires higher computational power compared to LOOCV and LkOCV in order to extract multiple n -dimensional samples. Generally, a few thousand iterations of the entire process are required to achieve stable estimates of the model's predictive performance.

Permutation test

Another tool, often used in metabolomics, to check for overfitting is the permutation test. This consists of a statistical procedure for hypothesis checking in which the value assumed by the test statistic is estimated on the observed data and on all the permutations of the same data (usually only a large sample of these permutations is used) to decide whether to accept or reject the null hypothesis. More precisely, the p-value of the test is calculated as the proportion of permutations in which the test statistic assumes a value no less than that assumed in the observed data.

In practice, it proceeds in this way: a model is trained on the basis of the data contained in the dataset using the class as a classification label (for example, healthy subject and disease-affected patient). A parameter is therefore estimated that evaluates the model's classification performance (for example, accuracy). At this point, on the basis of the data used for the first training, a new dataset is created, assigning a randomly chosen class to each sample. The model is then re-trained using this dataset with the randomly assigned classes and the classification accuracy of this second model is estimated. The process is repeated several times (usually a few thousand) and the difference between the original accuracy (that obtained with the dataset containing the correct labels) and those obtained after random assignment of the labels is estimated. Models not subject to overfitting provide much higher accuracy on original models than those obtained from datasets with permuted labels. If this difference is statistically significant (p -value <0.05) the model can be considered correctly trained; conversely, if the p -value >0.05 the model cannot be considered valid.

Class imbalance

Accuracy is often used as an estimator of the classification performance of an ML algorithm. Indeed, this is a widespread practice in metabolomics: however, there underlies the risk of gross errors when the population analyzed presents a class imbalance, that is, the condition in which, within the dataset used to train the classifier, most of the samples have the same label.

To better understand this risk, imagine we have a dataset containing 50 samples of which 45 have the “healthy subject” label and only 5 have the “disease-affected subject” label. A classification model trained on these data that makes only 5 errors out of the 50 analyzed samples would show an accuracy of 90%. This value is generally considered acceptable (if not good), but if these 5 errors had always been committed in the prediction of cases belonging to the “disease-affected subject” class, the model would be unable to recognize cases belonging to this category and therefore would be practically unusable. Paradoxically, a model that committed 10 classification errors but always on samples belonging to the “healthy subject” class that is most represented, despite having a lower accuracy (80%), would be more useful because it is capable of making a more realistic prediction by recognizing samples from both categories. The use of accuracy as an estimate of the classification performance of a model on a population of unbalanced classes, can therefore lead to paradoxical results, leading to the selection of ineffective models (Japkowicz, 2000).

For this reason, it is good practice to always operate on balanced datasets, which therefore have a homogeneous number of samples with respect to the labeling system. In reality, however, this is not always possible. For example, because rare conditions are being analyzed, a balanced dataset would be both very small and very distant from reality.

Several systems for the management of class imbalance have recently been proposed. These solutions can be divided in 3 general strategies:

1. Modification of the performance estimation metric;
2. Sampling methods;
3. Modification of ML algorithms.

Metrics to estimate the classification performances

The most important and widely used tool to calculate the performance estimators of the classification systems is called the **confusion matrix** and consists of a table that reports the number of objects labeled in the original dataset as positive and negative compared to those predicted by the model ([Table 9.4](#)).

In [Table 9.4](#) samples are divided into four possible outcomes: samples labelled positive and correctly classified by the model, known as true positives (TP), negative and correctly classified, called true negatives (TN), positive ones but erroneously classified as negatives called false negatives (FN) and finally the negative samples but erroneously classified as positive and therefore called false positives (FP). The latter two (FN and FP) represent classification errors. The accuracy represents the percentage of correctly assigned samples compared to the total number of samples contained in the dataset. In addition to accuracy, other widely used estimators are sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood value and negative likelihood value ([Table 9.5](#)).

By replacing one of these classification performance estimators, a different point of view can be obtained in evaluating the ability to make correct prediction from models trained by datasets with unbalanced classes. Furthermore, depending on the uses of the model, the different estimators can contribute to maximizing the predictive capacity of a model in the right direction. For example, classification models intended to be used as population screening systems for a certain disease (e.g., cancer) must exhibit high sensitivity. The risk that a person suffering from that pathology is not recognized should be minimized. On the contrary, specificity is not decisive in this case; indeed, a classification as positive of a healthy subject will be corrected by the second level investigations to which all

Table 9.4 Confusion matrix: samples are classified as true positive or true negative based on their original label and correctly classified by the classification model. On the contrary, false positive and false negative are samples wrongly classified by the classification model.

	Classified as positive	Classified as negative
Actual positive	True positive	False negative
Actual negative	False positive	True negative

Table 9.5 Classification performance estimation: Formulae and significance of the principal estimation systems to evaluate the performance of a classification model. Standard error formulae are also reported.

Parameter	Significance	Value	Standard Error
Sensitivity	The proportion of positive samples that were also classified as positive by the model	$SN = \frac{TP}{TP + FN}$	$\sqrt{\frac{SN(1 - SN)}{(TP + FN)}}$
Specificity	The proportion of negative samples that were also classified as negative by the model	$SP = \frac{TN}{TN + FP}$	$\sqrt{\frac{SP(1 - SP)}{(TN + FP)}}$
Positive predictive value	The proportion of samples classified as positive who truly were positive	$PPV = \frac{TP}{TP + FP}$	$\sqrt{\frac{PPV(1 - PPV)}{(TP + FP)}}$
Negative predictive value	The proportion of subjects classified as negative who truly were negative	$NPV = \frac{TN}{TN + FN}$	$\sqrt{\frac{NPV(1 - NPV)}{(TN + FN)}}$
Positive likelihood ratio	The probability of a positive sample classified as positive normalized by the probability of a negative sample to be classified as positive	$PLR = \frac{SN}{1 - SP}$	
Negative likelihood ratio	The probability of a negative sample classified as negative normalized by the probability of a positive sample to be classified as negative	$NLR = \frac{1 - SN}{SP}$	
Accuracy	The proportion of samples correctly classified by the model	$A = \frac{TP + TN}{TP + FP + TN + FN}$	$\sqrt{\frac{SN(1 - SP)}{(TP + FN)}}$

screening positives are generally submitted for diagnostic verification and to start the therapeutic process.

By joining the points that relate the proportion of TPs and FPs (the so-called coordinates), we obtain a curve called ROC. Another classification performance estimator widely used in metabolomics is the Area Under the Curve Receiver Operating Characteristics Curve (AUCROC). While sensitivity, specificity, negative and positive predictive power are evaluated, for classifying individuals as affected or not affected by a specific disease based on a predefined test value (threshold value), the ROC curve is constructed considering all possible test values and, for each of these, the proportion of TPs (the sensitivity) and the proportion of FPs are calculated. The proportion of FPs is calculated with the standard formula: $1 - specificity$. The area below the ROC curve is a measure of diagnostic power. If a test perfectly discriminated the sick from the healthy, the area of the ROC curve would have a value of 1, that is, 100% accuracy (Fig. 9.28A). In the event that the new test did not discriminate at all the sick from the healthy, the ROC curve would have an area of 0.5 (or 50%) which

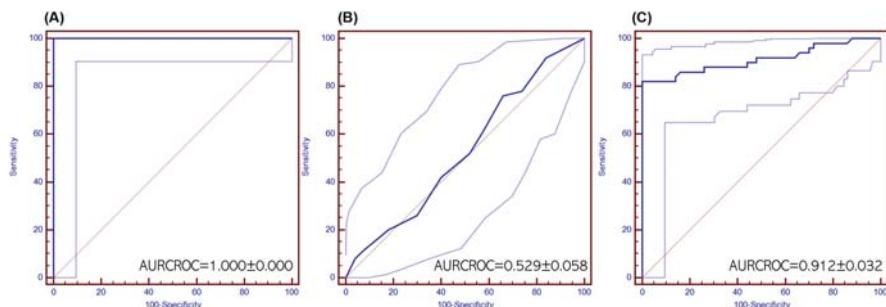


FIGURE 9.28 Area under the curve receiver operating characteristics curve.

(A) full discriminating model showing 100% accuracy and AUCROC = 1, (B) Model without any discriminant ability showing AUCROC = 0.53, (C) Good model with an AUCROC = 0.91.

would coincide with the area below the diagonal of the graph (Fig. 9.28B). In practice, a diagnostic test with an area under the curve $\geq 80\%$ is considered adequate (Fig. 9.28C). The area under the curve can assume values between 0.5 and 1.0. The greater the area under the curve (i.e., the closer the curve approaches the top of the graph), the greater the discriminating power of the test.

The ROC curve also allows to establish the conditions (threshold values or training conditions), which compensate for the FP and FN errors to maximize the overall diagnostic performance of the model. This operation consists in choosing the conditions that maximize the area under the ROC curve. Alternatively, sub-optimal conditions can also be chosen which minimize the least desired error, thus benefiting the specificity or sensitivity of a given model.

Sampling strategies

The most common strategy for dealing with the imbalance of classes in a dataset that is intended to be used to train a classifier is to use a sampling system that tends to balance the class representation. This strategy consists of the creation of a new dataset in which the class balance can be achieved through two paths, either by undersampling the most represented class (Fig. 9.29A), or by oversampling the least represented one (Fig. 9.29B).

Undersampling, as the name suggests, involves a random selection of samples from the most represented class equal to those of the least represented class. In this way, the new generated dataset will have a perfect symmetry in the distribution of the classes, but a drastically reduced number of samples compared to the original dataset. This is the principal weakness of this strategy. Sample size is always a critical element and therefore strategies that reduce it should be avoided.

Oversampling, on the other hand, increases the number of samples with the label corresponding to the less represented class and is the most common choice

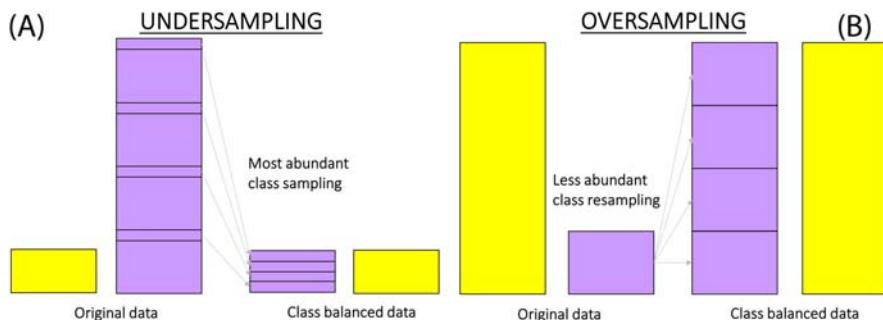


FIGURE 9.29 Sampling strategies to manage the class imbalance issue in machine learning.

(A) Undersampling consists of the creation of a new dataset containing all the less represented class samples and the same number of samples randomly selected from the most abundant class. (B) Oversampling consists in the creation of a new dataset containing all the most represented class samples and a copy of the less abundant class samples to achieve the class balance.

in metabolomics. This over-representation can be achieved simply by copying n -times the less represented class samples, or by introducing synthetic samples. The mere copy of the samples does not modify the information content of the dataset by not introducing new variance or entropy to the system, for this reason it is often considered a palliative but not very effective strategy. On the contrary, the genesis of synthetic samples also modifies the variance of the system by introducing new information content which, although plausible, is still forcibly introduced and therefore this strategy must always be carefully evaluated. Moreover, the results obtained by means of this technique must always be considered cautiously taking into account this limitation.

The most used technique for the generation of synthetic samples is known by the acronym **SMOTE** (Synthetic Minority Oversampling Technique). Let's imagine the case reported in Fig. 9.30A–B, in which the samples represented by the yellow squares show the most represented class, while the samples represented with purple spheres show the less represented one. SMOTE algorithm works by selecting a sample from the less represented class as well as a certain number of neighboring samples of the same class. The region of space delimited by these samples represents a domain in which the possibility of finding samples belonging to the same class is high. On the basis of this assumption, new synthetic samples are placed in this region. These samples show intermediate characteristics between samples actually observed and close to them. It should be noted that these samples have not been experimentally observed but are “likely”. As mentioned, this strategy increases the amount of total information about a generated dataset and therefore its entropy, minimizing the risk of introducing spurious information.

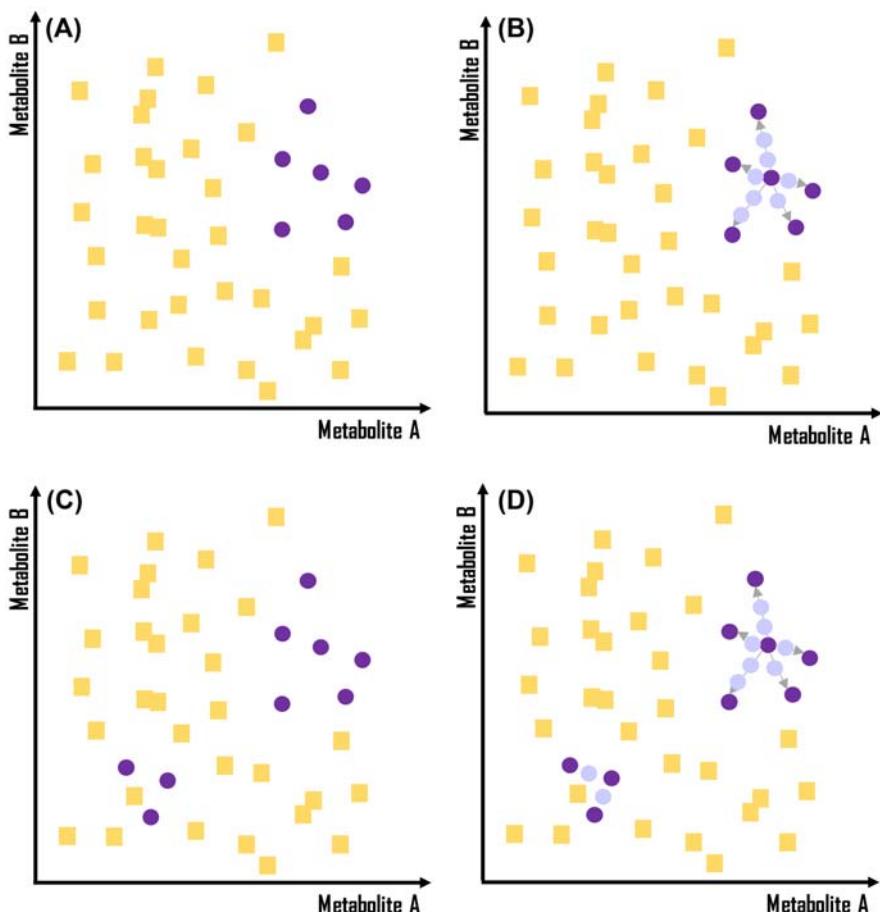


FIGURE 9.30 Synthetic Minority Oversampling Technique (SMOTE).

(A–B) Synthetic samples from less represented class (purple spheres) are created and placed between real samples. (C–D) Adaptive SMOTE algorithm for the creation of synthetic samples takes into account also samples in regions in which samples from the other class are present as nearest neighbors.

Recently, a slightly more sophisticated SMOTE system has been proposed, known as **adaptive SMOTE** (He et al., 2008) which takes into account not only the nearest neighbors samples with the same class but also the presence in the neighborhood of a “generation place” of samples with a different label. (Fig. 9.30C–D). This implementation is particularly relevant because if it is true that the presence of vicinal samples of different class makes that region of space less “safe” for the generation of synthetic samples, it is also true that those real samples represent a part of the observed reality that should be taken into account

in training. An under-representation of these samples could in turn introduce overfitting, where the sampling strategy is introduced precisely to minimize this risk.

Machine learning algorithms modification

Although supervised classification algorithms (DTs, discriminant analyses, Bayesian systems, etc.) perform a common function, they are based on very different operating mechanisms. Some algorithms are naturally more affected by class imbalance than others. In particular, DTs and their ensemble derivatives RFs (see below) are the classifiers that are least affected by this imbalance. The choice of the most suitable algorithm should therefore also take into consideration the class distribution.

Regardless of the classifier, an effective alternative, although still little used in metabolomics, is represented by the introduction of **cost matrices**. These are systems of penalization of the training steps that introduce classification errors on the less represented classes. In particular, cost matrices are generally built by introducing a penalty which is greater the lower the proportion of that class in the original dataset. In other words, the errors made in attributing a label corresponding to the less represented class have little influence on the accuracy but considerably influence the overall ability of the model to make predictions. For this reason, the introduction of this penalty tends to limit the introduction of errors on the less represented class. The idea of introducing cost matrices is due to Prof. Pedro Domingos ([Domingos, 1999](#)), who developed them in 1999 at the Higher Technical Institute of Lisbon.

Ensemble machine learning

AI systems in healthcare are constantly being used in new fields like clinical diagnosis, prognosis, therapy and so on. Just ten years ago it would have been unthinkable to imagine so much progress. Today it is not inconceivable that AI will eventually replace the doctor. Given the shortage of doctors, it is tempting to use computers to “automate” many medical and paramedical activities. In China, more than half of the population says they are ready to replace the general practitioner with an AI system, and the central government has launched several initiatives to push this transformation ([Kong et al., 2019](#)). In Europe, younger generations, born into a digital world (the famous “Y” and “Z” generations) will likely have few scruples about abandoning the “family doctor” for a computer, but the general populaces’ attitude about AI-based medical decision making is more cautious ([Mirbabaei, Stieglitz, & Frick, 2021](#)). Ultimately, while we are still a long way from completely leaving the medical decision to AI systems, it is obvious that the medical profession is changing, with an increasing number of AI-supported activities, even autonomously. However, a fundamental issue

remains: in case of error, who will be responsible? In AI systems that work independently, there are several sources of error. Is the data used for learning reliable and in sufficient quantity? Are patient data collected and integrated correctly? Is the algorithm reliable? Is the user interface functional (i.e., is there risk of misuse, incorrect display of results, etc.)?

Classification algorithms can make only a “false positive” and a “false negative”. As already pointed out, in the screening systems the first case is acceptable, to the extent that the patient will then consult a specialist who will eventually invalidate the result. But in the second case there is often no double check and therefore the consequences of this type of error are much more serious.

For this reason, currently, in Europe no medical decision can be made only on the basis of algorithmic processing. Europe is clearly in favor of a very controlled use of AI in medicine, perceived as helping doctors. Nonetheless, the health professional remains at the center of the medical and legal system, and, in the event of a medical error, the responsibility lies with the physician, not the machine.

One of the systems recently proposed to minimize this type of risk and make ML algorithms as efficient and robust as possible was the introduction of the so-called combined or “ensemble classification” (Kaur et al., 2020). The basic idea is that if it is true that a single model can produce FN and FP, it is less likely that different models, that work with different logics when queried about the classification of the same sample all produce the same type of error.

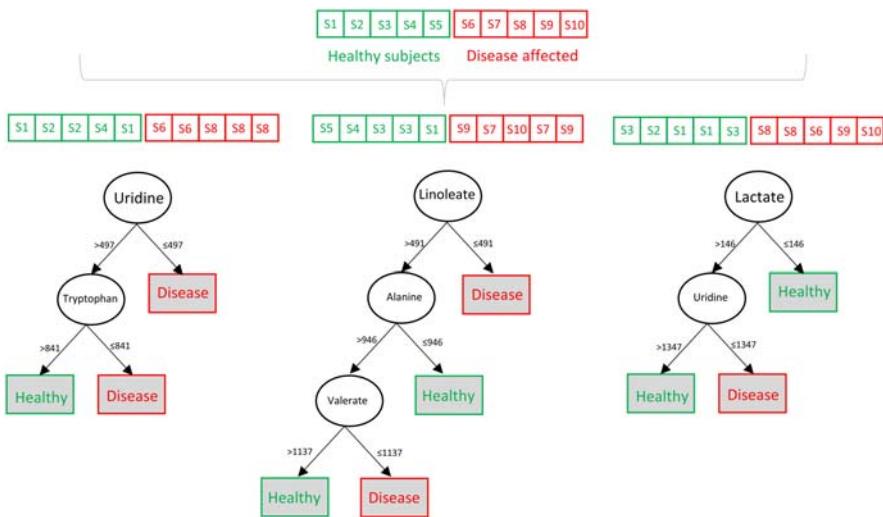
Therefore, rather than entrusting the classification proposal to a single classifier, it is possible to exploit the power of a set of classifiers and then mediate their responses to obtain a more secure classification. This system of using classifiers is called ensembling. There are several strategies for combining individual models into an ensemble model, the main ones being:

- Bagging
- Boosting
- Bayesian

Bagging

The ensembling systems based on **bagging** were first proposed in 1996 by Prof. Leo Breiman at the University of Barkley in California (Breiman, 1996). The term bagging comes from the synthesis of the words bootstrap and aggregating.

Bagging is a *meta-algorithm* that provides multiple trainings, using sub datasets, generated with repetition, starting from the original dataset, to obtain the same number of samples as the original dataset. At a superficial evaluation it might seem an ineffective strategy as the generated sub datasets contain only a fraction of the initial information and therefore train the individual classifiers less effectively. On the contrary, this tool is extremely effective especially in the case of classifiers unstable and strongly influenced by the distribution of input data

**FIGURE 9.31** Random forest.

Random forest is the bagging-based ensembling of DT algorithms.

such as the DTs. The ensemble ML build with DT using a bagging scheme is known as Random Forest (RF) (Qi, 2012) (Fig. 9.31). This is probably the most widespread and used ensemble model in metabolomics.

Boosting

Boosting *meta*-algorithms for ensembling classifiers were also introduced in 1996 (Freund & Schapire, 1996). There are several boosting strategies but the most famous and used is certainly the AdaBoost (Fig. 9.32).

The basic idea of this type of *meta*-algorithm is to use the entire database for each individual model but training several times the classifier to reinforce learning on samples whose classification is incorrect. In this way, each subsequent model represents a specialization of the previous one that corrects some of the errors produced. The synthesis of all these models will therefore constitute an ensemble that minimizes the errors. In other words, the different models generated constitute specialized improvements of the initial model able to classify the most critical samples. This is achieved by imposing a weight on the classification error. Initially all the samples have the same weight in building the model. Once generated, the samples placed incorrectly take on a greater weight and on the basis of this new weighted database, a second model is generated which therefore will not make that classification error. This might seem counterintuitive as models subsequent to the first usually have lower accuracies than the first. However, this

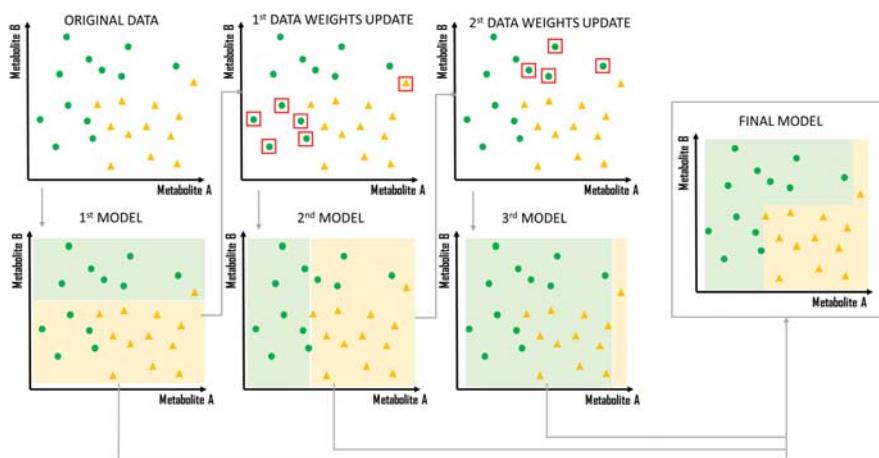


FIGURE 9.32 Boosting based ensembling.

Ensembling procedure based on boosting is based on the creation of several models minimizing the prediction errors of the previous one. The combination of all these models allows the building of a more accurate (although more prone to overfitting) ensemble model. Green points and yellow triangles represent the samples from 2 different groups. Green and yellow shading represent the hyperspace in which it is possible to classify green and yellow groups, respectively.

weakness is compensated by the mechanism for creating the final model that takes into account all models and therefore will be able to avoid all (or almost all) classification errors.

Boosting systems are very efficient as they use all the information initially contained in the dataset without dividing it (as happens with bagging systems); on the other hand, they are prone to overfitting, as they specialize the final model by stressing the training on the proposed samples. For this same reason, the boosting-produced ensemble models are strongly influenced by *outliers*, that is, samples that have provided anomalous responses. These in metabolomics are not rare, in fact, the articulated sequence of operations that generate the data (enrollment of the study subjects, obtaining the biological sample, extraction and purification of the metabolome, derivatization of metabolites, chemical analysis, pretreatment of data, etc.) offer various opportunities for errors arising that can result in the genesis of outliers. Boosting systems focus precisely on these samples. Knowledge of the operating mechanism of these *meta-algorithms*, however, allows the experimenter to pay particular attention to their weaknesses and therefore to carefully analyze the data management chain and to adopt effective overfitting analysis tools. Using these precautions, the boosting tools have provided excellent responses, significantly increasing the accuracy and robustness of many classification algorithms.

Both the bagging and boosting *meta*-algorithms involve the use of a single classification algorithm (trained with different subsets of the training dataset in the case of bagging and using specific cost matrices to minimize classification errors in the case of boosting). As we have already pointed out, however, the strength of ensemble systems is best expressed using different classification algorithms. The intrinsic difference in the classification logic of the various algorithms naturally minimizes the risk of repeating classification errors on the same samples. This evaluation is the basis of the **Bayesian ensembling** *meta*-algorithms.

There are different Bayesian ensembling versions such as the Bayesian model averaging (BMA) and the Bayesian model combination (BMC). The BMC is the most common in metabolomics and involves multiple trainings of different classifiers on the totality of the data contained in the original dataset (therefore without loss of information). The accuracy shown by the different models in the cross-validation phase is used to build a weighing matrix measuring the reliability of the different models. An elegant alternative is the introduction of a further weighing based on the distance between the sample to be classified and the centroid of the reference class. The votes (classification results) of the different classifiers, weighed by means of the confidence matrices, are then averaged and a unique classification is then issued. Several papers have shown that this kind of ensemble model is generally more accurate than each of the classifiers taken individually.

Our research group has proposed different population screening systems (some also validated on large cohorts of subjects recruited independently from the subjects used for the training of the individual classifiers) for different pathological conditions including endometrial cancer (Troisi et al., 2020; Troisi, Sarno, et al., 2018), bladder cancer (Troisi, Colucci, et al., 2021), hepatocellular carcinoma (Masarone et al., 2021) and several fetal malformations (Troisi et al., 2017; Troisi, Sarno, et al., 2018). This kind of ensemble offers several advantages. As already mentioned, it generally obtains higher accuracy than the individual classification models, and also represents a solution that naturally tends to minimize the effect of any overfitting of the individual models. All these aspects are useful for the potential applicability of these models for diagnostic purposes. In addition, the combined evaluation of the accuracy estimated in cross validation and the distance from the class centroid allows evaluation of each individual classification and therefore also the ensemble numerically. This score can simplify the adoption of thresholds that maximize the specificity or sensitivity of the model according to the intended use.

Ensembling of models can also offer advantages with respect to understanding the biological mechanisms underlying the development of a certain condition. For each single classification model, it is possible to evaluate which metabolites played a significant role in generating the separation between the classes (the nodes with the great IG in the DTs, or the metabolites with a higher VIP score in the PLS-DA models, for example). By combining different models, it is therefore possible to observe the same phenomenon from different observation points and

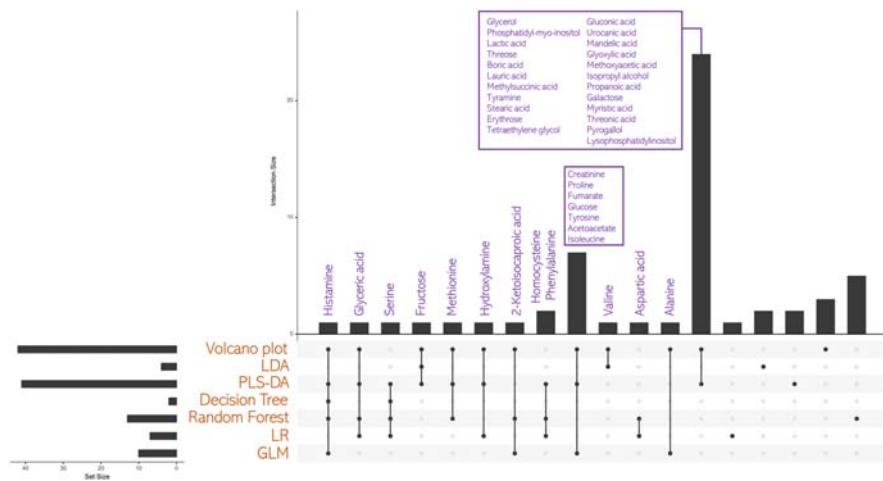


FIGURE 9.33 UpSet graph.

The UpSet graph improves upon the Venn diagram representation. The metabolites selected by each model are reported as the vertical bar. The horizontal bars represent the total number of selected metabolites from each model. The lines and dots represent the number of models selecting that metabolite combination.

then combine these with each other to develop a broader overall perspective. Furthermore, the metabolites that are selected from multiple models certainly play a key role from a biological point of view compared to the others. A tool often used in metabolomics to visualize this multiple participation in different models is the UpSet graph (Lex et al., 2014) (Fig. 9.33).

Features selection

The datasets from metabolomics experiments contain many more features (metabolites) than observations (samples). To think of it another way, these datasets are generally “fat”, that is, wider than tall. This imbalance is typical of investigations that come from the analysis of biological samples and does not usually occur in other areas of data analysis that use ML tools. For example, consider data analysts who produce classification models able to distinguish relevant emails from spam. In this case, there would be relatively few features on which to base the training (keywords, time of sending, recipient, etc.) and many observations (emails). This is just an example to show how the domain of data analysis application and the knowledge of the most delicate aspects to be taken into consideration are crucial to set up an adequate strategy.

The features/observations imbalance should always be managed as it has been shown that a reduction in the dimensionality of the data, in addition to favoring the

interpretation of the results obtained, allows the elimination of spurious information, making the training of classifiers and regression models more effective (Chen et al., 2020). Moreover, the data dimension reduction decreases the overfitting occurrence.

Data dimensionality reduction can be done by following several strategies. The PCA can be used; PCA allows synthesis of part of the initial information into new entities, the PCs. However, this is an unsupervised algorithm. The PLS-DA which can be considered its supervised counterpart can be useful for the same reason. Besides these, other techniques widely adopted in metabolomics are:

- Filtering systems
- Boruta's algorithm
- Genetic algorithms
- Embedded systems

Features filtering

Filtering systems are widely used in metabolomics to reduce the number of signals coming from mass spectrometers and nuclear magnetic resonance systems because they respond to a technical need relating to the use of these technologies. Indeed, a chromatogram, for example, could be considered a trace obtained by recording the intensity of the signal resulting from a mass spectrometry. It is a signal fluctuating in the time domain. This fluctuation, although thanks to current technologies, is very small, cannot be completely eliminated. It is therefore crucial, to reduce the amount of noise in the dataset, and consequently increase the information content, to minimize the amount of signals that do not have a clear chemical/biological nature but are merely the result of noise. To do this, it is possible, for example, to set a minimum area for a chromatographic signal below which that signal is no longer considered relevant and is therefore excluded from the dataset. This is a useful, often indispensable, but quite unrefined way of eliminating background noise, because metabolites showing low concentration have a chromatographic behavior that is often not very dissimilar to background noise. Various systems have therefore been proposed to make the filtering of data derived from untargeted metabolomics investigations more efficient.

A valuable tool increasingly used for the analysis of metabolomic data, proposed by the research group led by Prof. Xia at McGill University, and referred to as Metaboanalyst, allows users to filter characteristics based on mean/median value between samples, as well as variability between biological samples and samples prepared as quality controls (QC) (Chong et al., 2018, 2019) (see Chapter 2: Experimental Design in Metabolomics and Chapter 8: Techniques for Converting Metabolomic Data for Analysis for further details). While these are useful and often effective filtering metrics, most users don't determine appropriate thresholds for their specific data. Metaboanalyst suggests removing the k-percent of signals with the lowest value based on the size of the dataset (e.g., 20 or 30% lower) and relative standard deviation (e.g., using a 25% cut-off for LC-MS data). While these are useful

guidelines for selecting cutoffs, users often fail to verify whether they are appropriate for their data. An alternative strategy involves the use of technical and biological replicates on which to estimate the cut-offs starting from the idea that the noise is not constant while the concentration of metabolites is relatively stable and reproducible.

These strategies do not take into consideration the specific nature of the dataset and the task that the subsequent ML algorithm will have to perform. A more adaptive filtering pipeline has therefore been recently proposed, based on the analysis of signal abundances obtained from the analysis of the blank sample, proportions of missing values, and estimation of intra-class correlation coefficients (ICC) (Schiffman et al., 2019). Feature-filtering tools are often considered in the context of data pretreatment; therefore, in a much earlier stage than the training of ML systems (see Chapter 8: Techniques for Converting Metabolomic Data for Analysis in this regard). An alternative strategy, closer to the data analysis step, could be an assessment based on the outcomes of the exploratory analysis. For example, it could be decided to keep all the variables (metabolites) that show a difference in concentration between the classes under study that are statistically significant (for example evaluated through a *t*-test or a U-test or an ANOVA for multiclass problems). An alternative evaluation could be based on the fold change or on a combination of these two (as shown by the smile plot). These filtering strategies are effective because they focus the attention and ML training using metabolites with an important role in discriminating the different classes. The limitation of this approach, however, is twofold: it is very sensitive to the presence of outliers, and it represents a very rigid selection by eliminating many metabolites that can play a decisive role in differentiating the different classes. Indeed, as introduced in Chapter 7, Approaches in Untargeted Metabolomics, the relevant metabolites resulting from a certain pathology are not always described by metabolites that show large differences in concentration between the investigated classes. The strength of ML systems also lies in the ability to describe these nontrivial relationships. This type of selection, due to its severity, disrupts this possibility.

A further feature filtering strategy, slightly less severe, evaluates the AUC of the ROC curve for each single metabolite in the dataset, ordering them by increasing values and using only those included in the upper quartiles. In this case, the metabolites are selected on the basis of the greatest difference in concentration between the classes under study, but the threshold value can be chosen in a more arbitrary way. A further advantage of this filtering system is the ability to direct the choice of features using a different performance estimator for each single metabolite (for example, sensitivity or specificity) instead of AUCROC if chosen consistently with the optimization needs of the consequent classifier supporting the aim of the study.

Boruta's algorithm

Boruta's algorithm is a tool that is gaining importance and more use to manage features selection in metabolomics and, more in general, in the context of systems

biology. This features selection algorithm was devised in 2010 by three Polish-born researchers ([Kursa, et al., 2010](#)). It was conceived from the beginning as an R package ([Kursa & Rudnicki, 2010](#)), and this well describes how contemporary this algorithm is. As can be seen from the bibliographic reference, the name of the algorithm does not derive from the name of the researchers who conceived it, but rather dedicated to the demon Boruta, a mythological character from the Slavic sagas.

Boruta's algorithm is based on the training of a RF (a ML ensemble based on a bagging *meta*-algorithm of a series of DTs). The first brilliant idea introduced by Boruta's algorithm is represented by the strategy of not putting the features (metabolites) in competition with each other to be selected, but rather these are related to a randomized version of them also called shadow-features. In practice, starting from a certain dataset, another one is created that contains double the number of features of the original one. These duplicate (or shadow) features are generated by randomly permuting the values of the original variables. At this point, an RF is trained with all the features (both the original and the shadow ones). The importance of each original feature is evaluated using one of the tools generally used to evaluate the importance of the nodes of the DTs (IG, GI or information ratio, see above); these are selected based on a threshold value.

The selection of the threshold value is the second interesting solution introduced by this algorithm. It starts from the principle that the shadow features, having been generated using random permutations of the initial values, are of no value for classification purposes or for the training of the model. The highest value of significance (for example IG) estimated on the shadow features represents the value below which no features can play a decisive role. This is then used as a threshold. In practice, all the features that have an IG (or other selected importance estimator) higher than the highest value shown by the shadow variables will be selected.

Boruta's algorithm is simple in its applicability and computationally undemanding, moreover it has shown excellent increases in classification efficiency on ML algorithms trained only with the variables selected with this criterion. The only limitation of this algorithm is that it is entirely based on a specific classifier (the RF). As already stressed several times, classifiers work with different logics, and it is not certain that algorithms other than RF can take advantage of this type of selection. This is one of the main reasons that pushes many research groups to look for alternative and tightly encapsulated solutions in the specific used classifiers. These systems are also called wrapper methods, the most elegant and widely used of which are the genetic algorithms (GAs).

Genetic algorithm

By following simple and identical rules for different species, natural selection has managed to deliver the extraordinary biological diversity that we observe in the biosphere. GA are mathematical tools that are required to behave in a similar

way, that is, they are required to find solutions to problems with changing conditions following a finite series of standard steps. These represent the tools currently most used for the features selection in metabolomics. As with neural networks, GA are mathematical algorithms that mimic biological nature. Just as nature selects organisms using their fitness as a selection criterion, GAs have as their goal the selection of solutions that optimize a certain characteristic of the system, termed, in analogy with biology, fitness. In classification systems, fitness is generally represented by classification accuracy, but, sensitivity, specificity, AUCROC or other performance estimators can also be used depending on the aim.

GAs originate from the first theories of Ingo Rechenberg which date back to the late 1970s when he began to talk about “evolutionary strategies” within computer science (Rechenberg, 1978). The idea behind the GAs is to select the best solutions (pool of metabolites) and to recombine them with each other, in such a way that they evolve toward an optimum of fitness (accuracy). Fitness is estimated directly by training a specific classifier.

There is no strict definition for GAs, they are heuristic methods of search and optimization, inspired by the principle of natural selection by Charles Darwin. They are used both as adaptive algorithms and as computational models of natural evolutionary systems. Rather, with GA we mean a series of methods that have in common the characteristic of acting on a population of chromosomes that are selected on the basis of their fitness and are made to reproduce and change, obtaining evolution. Each chromosome is composed of genes, that is, a string of bits, each with multiple possibilities. If there are two possibilities, we call that an allele. Each chromosome can be thought of as a point in a space of candidate solutions (the size of this space will be 2^i where “*i*” represents the number of bits). GA modifies the populations of chromosomes, based on the value of their fitness function, that is, the degree of solvability of the classification problem.

When the genetic algorithm is initialized, a population is generated randomly. At each iteration, the algorithm provides for the measurement of the fitness value of all individuals. In practice, combinations of metabolites, that make up the different individuals, are randomly selected. The classifier to be used is then trained n-times as the many individuals are generated and the fitness (for example the accuracy of cross-validation) is estimated for each individual. The members with the best fitness are then selected and they are reproduced. Reproduction occurs essentially thanks to the crossing over mechanism, whereby parameters (number and position of the exchanged genes) are established *a priori*. The new individuals replace the previous individuals and the fitness of all second-generation individuals is established. As already mentioned, the new generation tends to keep the best genes, so fitness generally tends to improve with each generation.

In better emulating the biological process, in addition to the cross-over, a mutation mechanism can be used to introduce further variability between two successive generations. Of course, mutations can further increase or deteriorate fitness, so only positive changes tend to survive. The entire process of reproduction,

introduction of mutation, evaluation of fitness and selection of the best solutions is repeated several times until overall fitness tends to stabilize. Eventually, we expect to find a population of solutions that can adequately solve the problem posed.

Genetic algorithm operators

In its simplest form, a GA needs at least three elements:

- **Selection:** Select the chromosomes suitable for reproduction, the higher the fitness function, the more frequently it will be chosen for reproduction;
- **Crossing over:** This can be achieved through various strategies, including Single-point crossing over which is one of the simplest in which the selection operator cuts the genetic encoding strings of the selected elements in a random point with a certain probability established *a priori*. Thus, two heads and two tails are created, which are recomposed by crossing each other. An alternative is the standard crossover in which the fusion occurs in two different points of the chromosome allowing the exchange between two individuals of an internal chromosomal segment.
- **Mutation:** the mutation operator works by randomly modifying parts of a single gene. The mutation rate, as well as the probability of selecting a point for the crossover must be chosen *a priori*. The optimal choice of these values defined as hyperparameters will be the subject of the next paragraph. It is important to note that, regardless of fitness, all chromosomes must be subjected to mutation with the same probability (of generally low value). If changing were limited to the weak genes, fitness could end up in “holes”, creating very lucky and unrealistic cases. Furthermore, the mutation does not allow large displacements in the solution space. Crossing over, on the other hand, provides the possibility of mixing genes carrying properties that are isolated from the rest because the selection operation has already taken place.

In summary, the genetic algorithm evolves through the following process:

1. generation of an initial random population, made up of n m-bit chromosomes;
2. calculation of the fitness value for all individuals;
3. new population generation (repeats the following steps until n descendants are created, one per chromosome):
 - a. selection of a pair of parents;
 - b. crossing over with a certain probability, and replacement of the parents with the two new chromosomes. If there is no crossing, the two chromosomes are the exact copy of the parents;
 - c. mutation of the two descendants in each gene with *a priori* selected probability;
4. substitution of the new population for the old one;
5. check that the stopping criterion is met, otherwise go back to point 2.

There is no way to decide in advance whether the algorithm will actually be able to find an acceptable solution. Furthermore, among its limitations there is certainly the demand for a good computational effort. However, these algorithms allow the selection of variables using different criteria (fitness) and above all this can be the result of any classification algorithm. For this, this type of selection is highly focused and even more common in metabolomics.

Features generation

Because metabolomics datasets are generally “fat” that is, they are made up of many more features or columns (metabolites) than observations (or rows, or samples), then it would not make much sense to further increase the number of variables, generating additional features. Nevertheless, in some contexts it is possible to benefit from the introduction of artificial features for several reasons. First, not all metabolomics datasets are necessarily disproportionate in terms of width. It is possible, although rare, to have a large number of samples and to be able to obtain, from these, only information relating to a limited number of metabolites, in this case the features generation could help reduce the imbalance between rows and columns.

In the most common cases, where the number of samples is less than the number of features, the generation of artificial features starting from the real features could be an adequate dimensional reduction strategy. Indeed, this process generally adds new information to the model and makes it more accurate. Feature generation can improve model accuracy when there is an interaction between features. By adding new features that encapsulate the interaction of original features, new information becomes more accessible to the forecasting model ([van den Bosch, 2017](#)).

These artificial features, indeed, can be generated by applying different mathematical formulas to the values of two or more original features. In this way it is possible to grasp the relationships between the variables and consequently the latter can optimize the selection of the variables (through tools such as Boruta algorithm or genetic algorithm) further reducing the number of metabolites that capture the informative essence of the dataset. In turn, this allows better training of the classification or regression models.

A function $F(x)$ is said to exhibit an interaction between two of its variables x_j and x_k if the difference in the value of $F(x)$ as a result of changing the value of x_j depends on the value of x_k . For numeric variables, this can be expressed as:

$$E_x \left[\frac{\partial^2 F(x)}{\partial x_j \partial x_k} \right]^2 > 0 \quad (9.31)$$

To avoid generating meaningless features in the feature generating process, H-statistic can be used to detect feature interaction and assess its strength. The H-statistic was defined by [Friedman and Popescu \(2008\)](#) as a measure of

interaction strength. The idea is if two features x_j and x_k do not interact with each other, the partial dependence of $F(x)$ on the set of (x_j, x_k) , $F_{jk}(x_j, x_k)$ can be decomposed into the sum of the respective partial dependences on each variable separately:

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k) \quad (9.32)$$

where $F_j(x_j)$ is the partial dependence of a function $F(x)$, $F_{jk}(x_j, x_k)$ is the joint (x_j and x_k) partial dependence of $F(x)$. If the Eq. (9.32) is equal to zero, new features involving x_j and x_k can be created. If the features x_j and x_k do interact, another term should be added to the left of Eq. (9.32), expressing the effect of the interaction. Furthermore, if a given variable x_j does not interact with any other feature, then:

$$F(x) = F_j(x_j) + F_{-j}(x_{-j}) \quad (9.33)$$

where $F_{-j}(x_{-j})$ is the partial dependence of $F(x)$ on all features except x_j . If the Eq. (9.33) is equal to zero new features involving x_j can be created.

The partial dependence of a function $F(x)$ is the marginal effect one or more features has on the predicted outcome of a ML model. Given any subset x_s of the predictor feature, the partial dependence of a function $F(x)$ on x_s is defined as:

$$F_s(x_s) = E_{x_{-s}}[F(x_s, x_{-s})] = \int \hat{f}(z_s, x_{-s}) p_{-s}(z_{-s}) dz_{-s} \quad (9.34)$$

where x_s is a prescribed set of joint values for the variables in the subset, and the expected value is over the marginal (joint) distribution of all variables x_{-s} not represented in x_s . The marginal probability density of z_{-s} is indicated as $p_{-s}(z_{-s})$. Eq. (9.33) can be estimated from a set of training data by:

$$\hat{f}_s(z_s) = \frac{1}{n} \sum_{i=1}^n \hat{f}(z_s, z_{-s}) \quad (9.35)$$

where all the i values of z_s are the values of z_{-s} that occur in the training sample; that is, an average of the effects of all the other predictors in the model. Constructing partial dependence using Eq. (9.34) in practice is rather straightforward.

The properties of partial dependence functions are used to construct statistics to test for interaction effects of various types. From Eqs. (9.32) and (9.33) if there is a second order interaction between features j and k , and only second order interaction, $\Delta F_{jk}(x_j, x_k)$, is not zero.

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k) + \Delta F_{jk}(x_j, x_k) \quad (9.36)$$

$F_j(x_j)$ is the partial dependence of a function $F(x)$. For two-way interactions (second order) H_{jk}^2 is defined as:

$$H_{jk}^2 = \frac{\sum_{i=1}^N [\hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_j(x_{ij}) - \hat{F}_k(x_{ik})]^2}{\sum_{i=1}^N \hat{F}_{jk}^2(x_{ij}, x_{ik})} \quad (9.37)$$

where i is the number of observations in the data. The interaction strength H_{jk} is then calculated as $H_{jk} = (H^2 j_k)^{-2}$. The H-statistic is not restricted to two-way interactions and can be generalized to interaction effects of any order. Similarly, a statistic for testing whether a specified feature x_j interacts with any other feature would be:

$$H_{jk}^2 = \frac{\sum_{i=1}^N [F(x_i) - \hat{F}_j(x_{ij}) - \hat{F}_{-j}(x_{-j})]^2}{\sum_{i=1}^N \hat{F}^2(x_i)} \quad (9.38)$$

Here $F_{-j}(x_{-j})$ is the partial dependence of $F(x)$ on all features except x_j . The H-statistic is laborious to evaluate, because it iterates over all data points and at each point the partial dependence has to be evaluated which in turn is done with all N data points. In the worst case, we need $2N^2$ calls to the ML models predict function to compute the two-way H-statistic (j vs k , Eq. 9.37) and $3N^2$ for the total H-statistic (j vs all, Eq. 9.38). To speed up the computation, the N data can be sampled. This has the disadvantage of increasing the variance of the partial dependence estimates, which makes the H-statistic unstable. So, when using sampling to reduce the computational burden, one should sample enough data points.

An effective feature generation would be one that doesn't generate meaningless features that don't contribute to the prediction but still increases the complexity of the feature selection and calculation time. The H-statistics as a feature interaction detection algorithm is useful to generate less meaningful features and reduce calculation time.

Embedded methods

Embedded methods complete the features selection process within the construction of the ML algorithm itself. In other words, they perform feature selection during model training, which is why they are called embedded methods. The ML algorithm that best expresses this type of features selection is certainly the LASSO algorithm. During the training process of a LASSO model, a matrix system is generated that evaluates the various metabolites importance. The metabolites with lower weight are discarded. So, naturally, the LASSO algorithm selects the most important variables. In this case, there is no need to introduce a further feature selection step because this is already inherent in the ML algorithm.

Conclusions

Features selection is one of the most relevant steps to make supervised ML tools used in metabolomics more efficient. These strategies generally increase the accuracy, or more generally, the classification performance, of the trained models but can also be useful for biomarker discovery operations as well as to increase the

models' interpretability. As reported in [Chapter 7](#), Approaches in Untargeted Metabolomics, biomarker discovery indicates the search strategy of a single representative which can alone describe the presence or absence of a certain pathological condition. Candidate biomarkers are those metabolites or that combination of metabolites that show autonomously, given a certain threshold value, optimal classification efficiencies (in terms of accuracy or sensitivity or AUCROC). Such biomarkers must always be validated on large cohorts of patients in different conditions. Features selection is a pivotal step in biomarker discovery.

Hyperparameters optimization

Neil deGrasse Tyson, a famous and appreciated American astrophysicist and scientific communicator, a few years ago in one of his famous series of lectures used the incipit "*It seems one of the great challenges in this time is knowing enough about a subject to think you're right, but not enough about the subject to know you're wrong*". The time we are living in offers an incredible variety of stimuli and information never before seen in the history of humankind, so much so that it often does not allow for an adequate and necessary deep study. This trend is particularly dangerous in the context of ML. Indeed, these algorithms are surprisingly effective in the training process as well as in producing predictions. Moreover, for some of these, as we have already pointed out, the training mechanism is so complex and reiterative that their interpretability is strongly affected by it, ultimately transforming them into black boxes, that is, containers that perform tasks by means of a nonunderstandable mechanism. Furthermore, the growing availability of software tools and web services (see appendix for further details) that allow an extremely simple way to train a classifier using nonoptimized parameters, conditions and default choices represent an additional source for the generation of classifiers and decision makers whose control is almost impossible.

A thorough knowledge of the mechanisms inherent in the functioning of these algorithms and a great effort in the optimization and control of all the steps are a fundamental weapon to avoid being overwhelmed by data and predictions more or less meaningless and above all for which no confirmation or verification is possible.

Parameters and hyperparameters in machine learning

Before going into the strategies to tune the hyperparameters necessary for the functioning of the ML algorithms, it is important to define them and understand the differences with respect to the parameters used by the same algorithms. Although there is no formal definition of the parameters and hyperparameters, several criteria can be used to differentiate them effectively. Both are required by

ML algorithms to complete training and to make predictions. The parameters are estimated from the X matrix. Often the model parameters are estimated using an optimization algorithm, which is an efficient type of search through the possible values of the parameters.

Some examples of model parameters include:

- The weights in an artificial neural network
- The support vectors in a SVM
- The coefficients in a linear or logistic regression

Conversely, hyperparameters represent an external configuration to the model and whose value cannot be estimated from the data. They are often used in processes to help estimate model parameters and can be set using heuristics. The optimization of the hyperparameters is carried out through a tuning that takes into account the predictive modeling problem and its purposes.

The hyperparameters are many and varied; every single ML algorithm has its own, some examples of hyperparameters are:

- The number of hidden layers in an artificial neural network
- The type of neuronal activation function in an artificial neural network
- The kernel type of the SVM algorithms
- The penalty parameter C of the SVM algorithms
- The number of DTs in a RF system

In general, a good rule of thumb to understand if a certain term is a parameter or a hyperparameter is if you need to manually specify that model term, it is probably a model hyperparameter; on the contrary, if that term can be inferred from the data, it represents a simple parameter.

Hyperparameters tuning

The process of setting the combination of hyperparameters on which to base the entire training process of an ML algorithm must be done carefully and cannot be left to chance. One of the limitations of the automatic metabolomics data analysis services offered online or by means of simple software is the impossibility of fine-tuning these hyperparameters.

A rational tuning can be carried out by following different strategies including:

- Grid search
- Random search
- Bayesian optimization

Grid search

The grid search system is considered the traditional way of performing hyperparameter optimization. It is a relatively simple method and has been extensively

evaluated in the scientific literature. The basic idea is to identify the hyperparameters to be tuned and define a range of values to be scanned and an interval of variation between one estimate and the next. Afterward, all possible combinations of these values are used to train the classifier (or regressive model). The combination that gives the best results is selected as the best option. A grid search algorithm must be driven by a performance metrics: generally, accuracy for classifier is measured by cross-validation on the test set. One of the limitations of the grid search is the need to make a choice to manually impose the limits and the values to be evaluated. These, unfortunately, cannot be based on a completely rational evaluation. Indeed, grid search is computationally expansive since all the combinations of the hyperparameters have to be tested to determine the best option. For example, a typical soft edge of an SVM classifier equipped with an RBF Kernel has at least two hyperparameters that must be tuned to search for the best training environment: the regularization constant (C) and the kernel (γ) value. Both parameters are continuous, so to search the grid, a fine set of “reasonable” values for each hyperparameter has to be selected (see Fig. 9.34A).

Grid research trains an SVM with each pair (C, γ) and evaluates their performance on a validation set kept out (or by internal cross-validation on the training set, in which case multiple SVMs are trained per pair). Finally, the grid search algorithm returns the settings that achieved the best performance (for example in terms of accuracy) in the validation procedure.

Random search

Unlike the grid search, in which all the combinations of hyperparameters are evaluated (according to the range and the segmentation criterion chosen), in the random search only a certain number (*chosen a priori*) of hyperparameter combinations are selected. This reduces the number of trainings and therefore speeds the optimization mechanism (See Fig. 9.34B). The downside of this

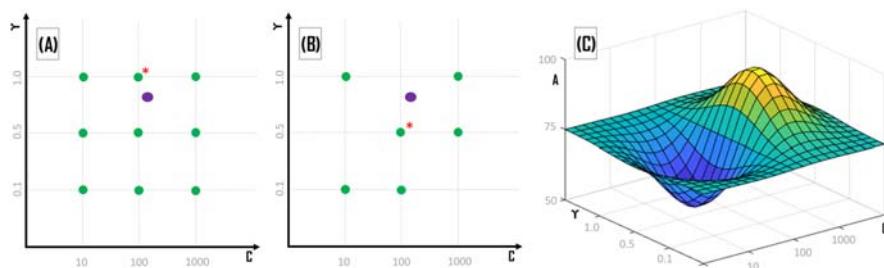


FIGURE 9.34 Hyperparameters optimization.

Two hyperparameters (C and γ) were optimized using grid search (A) testing all the hyperparameter combinations, random search (B) testing only a subset of all the combinations and Bayesian optimization (C) using the accuracy (A) as performance estimator.

acceleration is a reduced possibility of approaching the best combination of hyperparameters due to the reduced evaluations.

Both search systems based on the mechanism of the grid can generally provide solutions that are close to optimal but rarely identify the best solution. This limit is inherent in the discretization mechanism of the values to be probed.

Bayesian optimization

Bayesian optimization differs from random search and grid search because it improves search speed using past performance, while the other two methods are uniform (or independent) from past evaluations. In this sense, Bayesian optimization is like a manual search. Let's image manually optimizing the hyperparameter of a RF model. First, a series of parameters would be tried, then looking at the results, one of the parameters would be changed, rerun, and the results compared. This would indicate if the right direction has been followed. Bayesian optimization does a similar thing: the performance of past hyperparameters affects future decision. In comparison, random search and grid search do not take past performance into account when determining the new hyperparameters to evaluate. Therefore, Bayesian optimization is a much more efficient method.

As shown in Fig. 9.34C, the performance estimation function (for example accuracy) has a defined trend with respect to the hyperparameters considered. If unlimited resources were available, one could estimate every single point of the objective function to know its effective form. Obviously, this is generally not possible and, therefore, a number of combinations are identified to be tested to construct a surrogate model (also called a response surface model) to approximate the true objective function. A surrogate model by definition is “*the probabilistic representation of the objective function*”, which is essentially a model trained on hyperparameter combinations, relative to the actual objective function score.

Once a surrogate model has been constructed using a number of combinations of hyperparameters, an acquisition function (also called a selection function) is constructed to estimate which further combination to test. The combination of hyperparameters to be tested is then chosen as the one in which the acquisition function is maximized. This new measure updates the surrogate model making it closer to the reality and thus improving the estimate of the pair of hyperparameters that maximize classification efficiency (accuracy).

The most common acquisition function is the expected improvement (Eq. 9.39)

$$EI_{y^*}(x) = \int_{-\infty}^{\infty} \max(y^* - y, 0)p_M(y|x)dy \quad (9.39)$$

where $p_M(y|x)$ is the surrogate model, y is the true score of the objective function and x is the hyperparameter, y^* is the minimum score of the observed true objective function.

The expected improvement is based on the surrogate model, which means that a different surrogate model would result in different ways of optimizing this acquisition function.

Appendix

Given the complexity of the analysis of the data deriving from metabolomics experiments, as well as the need for a deep understanding of the different algorithms to implement effective systems, numerous tools have been developed to simplify this phase of the metabolomics pipeline. Some of these tools have been developed as independent software, some are free while for others the licenses are issued following the payment of a fee. These independent tools represent the easiest solutions to use because they do not require knowledge of coding and programming language and can also be used by personnel who are not particularly experienced in data analysis.

Without prejudice to the inherent risk of using any data analysis tool without in-depth knowledge of the topic, as we have already expressed in this discussion, these software programs offer the advantage of broadening the audience of potential users by simplifying the approach to analysis of these data. On the other hand, generally, these solutions are not very flexible with respect to the various optimizations necessary for an efficiency of the training mechanisms (features selection/generation, validation, ensemble, hyperparameters optimization, etc.).

Other solutions are released as packages for different programming languages (e.g., R, Python, and Matlab®). These are certainly more flexible in terms of available optimization chances and can be combined with each other to increase their functionality. On the other hand, these packages require knowledge of the programming languages for which they were designed and are more complex to use.

Table 9.6 summarizes some of the most popular software and packages used in the metabolomics field. This is not intended as a complete description of all available solutions, which would be difficult to report as these are constantly updated and new tools are developed frequently. For a review, updated to 2020, see the Misra paper published in *Metabolomics* (Misra, 2021).

More interesting, intermediate solutions are software not specifically designed for metabolomics, such as KNIME (Berthold et al., 2009), RapidMiner (Kotu & Deshpande, 2014) or Weka (Frank et al., 2004) which effectively manage data analysis and data mining. These software manage the different steps of the process in an extremely accurate way and are often even too thorough and rich in possibilities with respect to the needs of metabolomics data processing. These require a deep understanding of the dynamics and peculiarities of the different ML models but are generally very simple to use.

This book is accompanied by a dedicated website (<http://www.metabolomicsperspectives.com>) from which you can download an R package (*MetPer*) which is specifically developed to match the phases of data analysis as described in this chapter. The package also allows you to carry out all the pretreatments provided in Chapter 8, Techniques for Converting Metabolic Data for Analysis. The site contains a

Table 9.6 Principal bioinformatic tools used in metabolomics: Formulae and significance of the principal estimation systems to evaluate the performance of a classification model.

Name	Type	License	Specifications	Reference
Metaboanalyst	Web application and R package	Free	Allow for both data pretreatments and analysis	Chong et al. (2018), Chong, Yamamoto, and Xia (2019)
MetaboScape	Stand-alone software	For a fee	Specifically designed for Bruker high-resolution MS-platforms	Daltonics (n.d.)
MetaboPredict	Stand-alone software	For a fee	Allow for both data pretreatments and analysis	Theoreo srl (n.d.)
Workflow4metabolomics	Web application	Free	Virtual environment built upon the Galaxy web-based platform technology and maintained by French Bioinformatics Institute (IFB) and the French Metabolomics and Fluxomics Infrastructure (MetaboHUB)	Giacomoni et al. (2015)
Compound Discoverer	Stand-alone software	For a fee	Specifically designed for Thermo Scientific high-resolution MS-platforms	Comstock et al. (n.d.)
One omics suite	Stand-alone software	For a fee	Specifically designed for Sciex high-resolution MS-platforms	Antonoplis et al. (n.d.)
MetaboShiny	R package	Free	Allow for Direct Inject analysis data	Wolthuis et al. (2020)
Metabolite AutoPlotter	R-scripted code, wrapped into a shiny application	Free	A tool to process and visualize quantified metabolite data	Pietzke and Vazquez (2020)

(Continued)

Table 9.6 Principal bioinformatic tools used in metabolomics: Formulae and significance of the principal estimation systems to evaluate the performance of a classification model. *Continued*

Name	Type	License	Specifications	Reference
Metabolite-Investigator	Both web-tool and stand-alone Shiny-app	Free	Allow for both data pretreatments and analysis	Beuchel et al. (2021)
XCMS	Web application and R package	Free	Allow for both data pretreatments and analysis	Smith et al. (2006)
VIIME	Web application and R package	Free	Allow for both data pretreatments and analysis	Choudhury et al. (2020)
muma	R package	Free	Univariate and multivariate data analysis	Gaude et al. (2013)
Metaboverse	Stand-alone software	Free	A cross-platform app built to aid users in contextualizing their data on their model's metabolic and global reaction network	

Standard errors formulae are also reported.

detailed guide to use the package using the R language and also contains some datasets specially structured for practicing data analysis using the package *MetPer*.

References

- Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The ‘K’ in K-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* (pp. 441–446). [i6doc.com](#) publ.
- Antonoplis, A., Causon, J., & Hunter, C. (n.d.) Rapid analysis and interpretation of metabolomics SWATH acquisition data using a cloud-based processing pipeline. *Target*, 45 (50), 55.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME—the Konstanz information miner: Version 2.0 and beyond. *AcM SIGKDD Explorations Newsletter*, 11(1), 26–31.
- Beuchel, C., Kirsten, H., Ceglarek, U., & Scholz, M. (2021). Metabolite-investigator: An integrated user-friendly workflow for metabolomics multi-study analysis. *Bioinformatics (Oxford, England)*, 37(15), 2218–2220. Available from <https://doi.org/10.1093/bioinformatics/btaa967>.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152).
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123–140.
- Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1), 52. Available from <https://doi.org/10.1186/s40537-020-00327-4>.
- Chen, Y.-Z., & Lai, Y.-C. (2018). Sparse dynamical Boltzmann machine for reconstructing complex networks with binary dynamics. *Physical Review E*, 97(3), 032317. Available from <https://doi.org/10.1103/PhysRevE.97.032317>.
- Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., & Xia, J. (2018). MetaboAnalyst 4.0: Toward more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1), W486–W494.
- Chong, J., Yamamoto, M., & Xia, J. (2019). MetaboAnalystR 2.0: From raw spectra to biological insights. *Metabolites*, 9(3), 57.
- Choudhury, R., Beezley, J., Davis, B., Tomeck, J., Gratzl, S., Golzarri-Arroyo, L., Wan, J., Raftery, D., Baumes, J., & O'Connell, T. M. (2020). Viime: Visualization and integration of metabolomics experiments. *Journal of Open Source Software*, 5(54). Available from <https://doi.org/10.21105/joss.02410>.
- Comstock, K., Ding, C., Stratton, T., Wang, K., & Eiserberg, G. (n.d.). Rapid and Confident Metabolite Profiling and Identification using Bench-Top Orbitrap Q Exactive and Compound Discoverer. <http://www.unitylabservices.eu/content/dam/tfs/ATG/CMD/cmd-documents/sci-res/posters/ms/events/asms2014/PN-64125-Identification-Q-Exactive-ASMS2014-PN64125-EN.pdf>.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. Available from <https://doi.org/10.1007/BF00994018>.
- Daltonics, B. (n.d.). MetaboScape.
- Domingos, P. (1999). *MetaCost: A general method for making classifiers cost-sensitive*. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 155–164). San Diego, CA: ACM.
- Edoardo, G., Chignola, F., Spiliotopoulos, D., Spitaleri, A., Ghitti, M., Garcia-Manteiga, J. M., Mari, S., & Musco, G. (2013). muma, An R package for metabolomics univariate and multivariate statistical analysis. *Continued as Current Metabolomics and Systems Biology*, 1(2), 180–189. Available from <https://doi.org/10.2174/2213235X11301020005>.
- Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics (Oxford, England)*, 20(15), 2479–2481. Available from <https://doi.org/10.1093/bioinformatics/bth261>.
- Freund, Y., & R.E. Schapire 1996. *Experiments with a new boosting algorithm*. In *icml* (Vol. 96, pp. 148–156).

- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916–954.
- Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3), 121–136.
- Gaude, E., Chignola, F., Spiliotopoulos, D., Spitaleri, A., Ghitti, M., García-Manteiga, J. M., ... & Musco, G. (2013). muma, an R package for metabolomics univariate and multivariate statistical analysis. *Current Metabolomics*, 1(2), 180–189. 5.
- Ghosh, T., Zhang, W., Ghosh, D., & Kechris, K. (2020). Predictive modeling for metabolomics data. *Methods in Molecular Biology (Clifton, N.J.)*, 2104, 313–336. Available from https://doi.org/10.1007/978-1-0716-0239-3_16.
- Giacomoni, F., Corguillé, G. L., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., Duperier, C., et al. (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics (Oxford, England)*, 31(9), 1493–1495. Available from <https://doi.org/10.1093/bioinformatics/btu813>.
- Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica Ed. Pizetti E.*
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning (pp. 1322–1328). IEEE.
- Hebb, D. O. (1949). *The organisation of behaviour: A neuropsychological theory*. Science Editions New York.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558.
- Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415–425.
- Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies* (Vol. 56). Citeseer.
- Jarvis, R. A., & Patrick, E. A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, 100(11), 1025–1034.
- Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.
- Jöreskog, K. G., & Wold, H. O. A. (1982). Systems under indirect observation: Causality, structure, prediction (Vol. 139). North Holland.
- Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., El-Sappagh, S., Islam, M. S., & Islam, S. M. R. (2020). Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives. *IEEE Access*, 8, 228049–228069. Available from <https://doi.org/10.1109/ACCESS.2020.3042273>.
- Kong, X., Ai, B., Kong, Y., Su, L., Ning, Y., Howard, N., Gong, S., et al. (2019). Artificial intelligence: A key to relieve China's insufficient and unequally-distributed medical resources. *American Journal of Translational Research*, 11(5), 2632–2640.
- Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: Concepts and practice with rapidminer*. Morgan Kaufmann.
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13.
- Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta—a system for feature selection. *Fundamenta Informaticae*, 101(4), 271–285.

- Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1983–1992. Available from <https://doi.org/10.1109/TVCG.2014.2346248>.
- Macnaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*, 202 (4936), 1034–1035. Available from <https://doi.org/10.1038/2021034a0>.
- MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281–297).
- Marx, V. (2013). The big challenges of big data. *Nature*, 498(7453), 255–260. Available from <https://doi.org/10.1038/498255a>.
- Masarone, M., Troisi, J., Aglitti, A., Torre, P., Colucci, A., Dallio, M., Federico, A., Balsano, C., & Persico, M. (2021). Untargeted metabolomics as a diagnostic tool in NAFLD: Discrimination of steatosis, steatohepatitis and cirrhosis. *Metabolomics: Official Journal of the Metabolomic Society*, 17(2), 12. Available from <https://doi.org/10.1007/s11306-020-01756-1>.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- McKnight, P. E., & Najab, J. (2010). Mann-Whitney U test. *The Corsini encyclopedia of psychology*, 1.
- Mirbabaie, M., Stieglitz, S., & Frick, N. R. J. (2021). Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health and Technology*, 11(4), 693–731. Available from <https://doi.org/10.1007/s12553-021-00555-5>.
- Misra, B. B. (2021). New software tools, databases, and resources in metabolomics: Updates from 2020. *Metabolomics: Official Journal of the Metabolomic Society*, 17(5), 49. Available from <https://doi.org/10.1007/s11306-021-01796-1>.
- Owen, D. B. (1965). The power of Student's *t*-test. *Journal of the American Statistical Association*, 60(309), 320–333.
- Pietzke, M., & Vazquez, A. (2020). Metabolite AutoPlotter - An application to process and visualise metabolite data in the web browser. *Cancer & Metabolism*, 8, 15. Available from <https://doi.org/10.1186/s40170-020-00220-x>.
- Qi, Y. (2012). *Random forest for bioinformatics. Ensemble machine learning* (pp. 307–323). Springer.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81–106.
- RDevelopment CORE TEAM, R. (2008). *R: A language and environment for statistical computing*. R foundation for statistical computing Vienna, Austria.
- Rechenberg, I. (1978). *Evolutionsstrategien. Simulationsmethoden in der Medizin und Biologie* (pp. 83–114). Springer.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Schiffman, C., Petrick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rappaport, S., & Dudoit, S. (2019). Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics*, 20(1), 334. Available from <https://doi.org/10.1186/s12859-019-2871-9>.

- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3), 779–787. Available from <https://doi.org/10.1021/ac051437y>.
- Stähle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *Journal of chemometrics*, 1(3), 185–196.
- Theoreo srl. (n.d.). MetaboPredict. <http://www.theoreosrl.com/metabopredict>.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Troisi, J., Colucci, A., Cavallo, P., Richards, S., Symes, S., Landolfi, A., Scala, G., Maiorino, F., Califano, A., & Fabiano, M. (2021). A serum metabolomic signature for the detection and grading of bladder cancer. *Applied Sciences*, 11(6), 2835.
- Troisi, J., Landolfi, A., Sarno, L., Richards, S., Symes, S., Adair, D., Ciccone, C., Scala, G., Martinelli, P., & Guida, M. (2018). A metabolomics-based approach for non-invasive screening of fetal central nervous system anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 14(6), 77. Available from <https://doi.org/10.1007/s11306-018-1370-8>.
- Troisi, J., Raffone, A., Travaglino, A., Belli, G., Belli, C., Anand, S., Giugliano, L., et al. (2020). Development and validation of a serum metabolomic signature for endometrial cancer screening in postmenopausal women. *JAMA Network Open*, 3(9), e2018327. Available from <https://doi.org/10.1001/jamanetworkopen.2020.18327>.
- Troisi, J., Scala, G., Campiglia, P., Zullo, F., & Guida, M. (2018). *Method for the diagnosis of endometrial carcinoma*. Google Patents.
- Troisi, J., Sarno, L., Landolfi, A., Scala, G., Martinelli, P., Venturella, R., Di Cello, A., Zullo, F., & Guida, M. (2018). Metabolomic signature of endometrial cancer. *Journal of Proteome Research*, 17(2), 804–812. Available from <https://doi.org/10.1021/acs.jproteome.7b00503>.
- Troisi, J., Sarno, L., Martinelli, P., Di Carlo, C., Landolfi, A., Scala, G., Rinaldi, M., D'Alessandro, P., Ciccone, C., & Guida, M. (2017). A metabolomics-based approach for non-invasive diagnosis of chromosomal anomalies. *Metabolomics: Official Journal of the Metabolomic Society*, 13(11), 140. Available from <https://doi.org/10.1007/s11306-017-1274-z>.
- Troisi, J., Cavallo, P., Richards, S., Symes, S., Colucci, A., Sarno, L., Landolfi, A., Scala, G., Adair, D., & Ciccone, C. (2021) Non-invasive screening for congenital heart defects using a serum metabolomics approach. *Prenatal Diagnosis*, 41(6), 743–756.
- Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(marzo), 119–128. Available from <https://doi.org/10.1002/cem.695>.
- Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33 (1), 1–67.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.
- van den Bosch, S. (2017). Automatic feature generation and selection in predictive analytics solutions. *Master's thesis, Faculty of Science, Radboud University*, 3(1), 3–1.
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), 5261–5267.

- Winter, J. C. F. d., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3), 273.
- Wolthuis, J. C., Magnisdottir, S., Pras-Raves, M., Moshiri, M., Jans, J. J. M., Burgering, B., van Mil, S., & de Ridder, J. (2020). MetaboShiny: Interactive analysis and metabolite annotation of mass spectrometry-based metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 16(9), 99. Available from <https://doi.org/10.1007/s11306-020-01717-8>.
- Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), 2839–2846. Available from <https://doi.org/10.1016/j.patcog.2015.03.009>.
- Yang, L., Yang, X., Kong, X., Cao, Z., Zhang, Y., Hu, Y., & Tang, K. (2016). Covariation analysis of serumal and urinary metabolites suggests aberrant glycine and fatty acid metabolism in chronic hepatitis B. *PLoS One*, 11(5), e0156166. Available from <https://doi.org/10.1371/journal.pone.0156166>.

This page intentionally left blank

Relevant metabolites' selection strategies

10

Jos Hageman

Biometris, Applied Statistics, Wageningen University & Research, Wageningen, The Netherlands

Introduction

Metabolomics is the study of the metabolome or all small molecules or metabolites present in plants or any other organisms (Madsen et al., 2010; Weckwerth, 2003). The metabolome is dynamic and is the best reflection of the phenotype of the organism under study (Hageman, Hendriks, et al., 2008). A major goal of metabolomics studies is, besides gathering general knowledge of the study objects, finding biomarkers related to different kinds of traits (Hageman, van den Berg, et al., 2008; Hendriks et al., 2011). These traits can be very diverse ranging from biomarkers for (the onset of) disease states (Madsen et al., 2010) to predictors for crop ripening (Lombardo et al., 2011) or sensory perceptions (Lindinger et al., 2009). Metabolomics aims to simultaneously measure as many metabolites as possible. Currently, there is no single platform that can measure all metabolites, but by combining different platforms a comprehensive view of the metabolome is obtained. Metabolomics experiments typically measures hundreds of metabolites (Dettmer et al., 2007).

After obtaining a snapshot of the metabolome, a statistical model relating metabolites to the trait of interest is created. Statistical models are created to assess the association between metabolites (the predictors) and the trait of interest (the response variable). In statistical terms, the response variable is the variable whose variation depends on the values of the predictors (the metabolites) and is the outcome of statistical models. It can be a class membership (e.g., control vs infected), or a quantitative trait of interest (e.g., taste sweet). When creating these models, there are several moments and reasons why selection of metabolites is useful.

1. Usually, a limited number of metabolites is biochemically connected to the trait of interest. Therefor it is not expected that all metabolites contribute equally to the prediction of the trait of interest (Doeswijk et al., 2011; Hageman, van den Berg, et al., 2008; Saccenti et al., 2014). The (most) predictive metabolites in these models are typically identified after creation of the model.
2. Some statistical techniques suffer greatly from the presence of many metabolites (predictors in the model): it hinders the creation of the models, makes them prone to overfitting. Overfitting happens when statistical models

no longer learn the general trend of the data but learn the peculiarities of the data at hand. Large numbers of metabolites also slow down the creation and validation of the models. It is therefore important to limit the number of metabolites at the start of the modeling.

3. Another reason is that small metabolite sets are desirable as they make it easier to obtain insights into mechanisms (as compared to considering the whole observed metabolome). A small set of relevant metabolites connects better to research devoted to identifying biomarkers for the prediction of class membership (e.g., control vs infected) or any small set of metabolites that predict a quantitative trait the best ([Hageman, Hendriks, et al., 2008](#); [Saccenti et al., 2011](#)). The purpose of the biomarkers is typically to use them routinely in a targeted, single platform setup for fast classification. As metabolomics can measure hundreds of metabolites, variable selection is a must.

This chapter provides an overview of different strategies to reduce the high dimensional dataset derived from a metabolomics experiment. For ease of the overview, all variable selection methods and techniques have been divided into one of three categories ([He & Yu, 2010](#); [Shahrjooihaghghi et al., 2017](#)), see for an overview: [Table 10.1](#).

1. Low-level variable selection: deselect metabolites that are completely uninteresting. Low-level variable selection is primarily focused on removing non-informative or redundant variables ([Shahrjooihaghghi et al., 2017](#)). These approaches are sometimes referred to as filter methods as they act as a filter to separate the promising metabolites from the not-so-promising ones.
2. Medium-level variable selection: methods that intrinsically select metabolites. Methods in this category select important metabolites or deselect the ones that are not important as part of their inner workings. A typical result would be a statistical model that uses not all metabolites as predictors but is using a small subset of metabolites.
3. High-level variable selection: Assess explicit importance of metabolites. High-level variable selection entails methods that intrinsically select metabolites as part of the inner working of the modelling technique itself. This category also includes algorithms or heuristics that indicate the importance of metabolites through some importance criterion. So, methods in this category can possess innate variable selection properties but this is not necessary as this can also be obtained through external criteria. The latter typically results in a ranked list.

Low-level variable selection

Low-level variable selection or filter methods can be divided into two categories: supervised and unsupervised methods. The difference between the two is whether they consider the trait of interest in their decision to exclude metabolites or to retain them.

Table 10.1 Overview of the three different variable selection levels, including a description and example categories.

Variable selection level	Focus	Examples
Low	Removal of noninformative information	Unsupervised —percentage observed —variance based supervised —correlation with trait —fold change —hypothesis tests
Medium	Selection of important metabolites	Wrapper methods —stepwise regression —Global optimization techniques —genetic algorithms —simulated annealing —tabu search
High	Variable selection is intrinsically to the method or variable importance is indicated with an auxiliary criterion	Embedded techniques —regularized regression —latent variable techniques —tree based methods —support vector machines Heuristic approaches —bootstrapping —cross validation

Unsupervised low-level variable selection

In unsupervised variable selection, the response variable is not considered in the decision to select or deselect a metabolite. Metabolites are selected purely on the observed metabolic profile. The methods in this category have in common that they assess the observed variation in each metabolite. When the observed variation is deemed too low, the metabolite will be discarded.

Percentage observed

Not all metabolites are always observed in every sample. This can have different causes. Metabolites can be completely absent in some samples while being present in others. When a metabolite is present it can sometimes be present in a concentration that does not reach the detection limit. Another cause for not observing a metabolite could be of a more technical nature, like misalignments in processing of the measurements. All these cases lead to missing values for certain metabolites. When a metabolite has not been observed in a certain number of samples, it usually gets thrown out. Thresholds vary between 30%–60%.

Missing values in metabolomics is very common. Many statistical methods require a complete data set with no missing values. Imputing a modest amount of missing values is typically not a big problem. Metabolites that are not observed in a large part of the samples require too many values to be substituted. Metabolites

that have many imputed values may not be the most reliable ones to use in subsequent statistical analysis. Removing them from future statistical analysis reduces the chance of false positives.

Researchers should be very careful with removing metabolites using the percentage observed criterium. The absence of a metabolite (or rather being below the detection limit) in a group of samples could, in principle, correlate with different experimental factors making these metabolites the well sought biomarkers.

Variance based

Another criterion to consider in the selection of the metabolites fit for future analysis is assessing if the variation contained in metabolites is sufficient. It is expected that not all metabolites will respond to experimental or observational factors. In principle this means that the metabolite concentrations remain very much the same throughout all samples. Including these metabolites in statistical analysis does not make sense as they cannot be expected to function as biomarkers or otherwise as predictive metabolites. General values for a minimal threshold variance are difficult to give and are data set and probably even metabolite dependent. Thresholds are connected to the relative standard deviation of metabolites and general noise levels of the data.

Supervised low-level variable selection

Where unsupervised low-level variable selection methods for the most part deselect metabolites that do not have sufficient variation, supervised low-level variable selection methods select metabolites with sufficient relevant variation. This time the response variable is considered when deciding which variables to include or exclude as methods in this category assess the relationship between individual metabolites and the response. Metabolites that show a low degree of association can be discarded before any statistical model is even created. The idea is only to retain the metabolites that loosely show some association with the response. Based on the nature of the response, qualitative (like class membership) or quantitative (like taste sweet), different criteria can be used. All criteria calculate scores which are the basis on which to select the metabolites that will be used in subsequent statistical modeling.

Quantitative response

Several correlation metrics are available for expressing relatedness between a metabolite and a quantitative response.

Pearson's correlation coefficient

This expresses the linear association between two quantitative variables (in our case a metabolite and a quantitative response variable). The correlation coefficient is a dimensionless number between -1 and $+1$. Values close to -1 and $+1$ indicate a strong linear association, while outcomes close to 0 indicates that there

is no linear association. Correlation coefficient estimates can be tested using a *t*-test to test if the estimate is significantly different from 0. Selection of metabolites can be done in two ways, a threshold on the correlation coefficient estimate can be used, say the selection of metabolites with an absolute correlation coefficient of for example, >0.3 . Alternatively, metabolites for which the hypothesis test has shown the estimate to be significantly different from zero can be selected. If desired, partial correlation coefficients can be calculated. These are Pearson correlation coefficients but corrected for the influence of other metabolites. In the case of suspected outlying values, nonnormally distributed metabolites/traits or nonlinear relations between the two, it can be useful to use Spearman rank correlation. In this procedure, metabolite and traits values are first converted to rank numbers followed by the usual Pearson correlation coefficient.

Other metrics for establishing the degree of association between a metabolite and a quantitative trait are for example, distance related criteria such as Euclidian or Manhattan distance and Fisher's score.

Qualitative response

When the response variable is qualitative (like the classes *control* and *infected* but this is easily extended to more than two groups), we would like to know if metabolites appear to be up or down regulated in one of the groups. We would like to select metabolites for further analysis that show some degree of difference between our classes and thereby deselect the ones that do not show any relatedness with the treatments.

Fold change

A fold change describes how much the average metabolite concentration has changed between two groups ([van den Berg et al., 2006](#)). It is the ratio of the average metabolite concentrations of the groups. It is calculated for each metabolite separately. Fold change is connected to effect size, the bigger the fold change, the larger the difference between the two groups and thus the larger the effect size. The idea is to retain only metabolites that have a fold change above a certain threshold. Metabolites with large fold changes show clear differences between the two groups and could potentially be of interest. An important caveat is that metabolites showing a large fold change are not necessarily more important than metabolites showing a small fold change.

Hypothesis testing

Using hypothesis tests, it is investigated if metabolites show a significant difference between treatments groups. Hypothesis tests can be used to select metabolites showing a significantly different response between two groups (using two independent samples *t*-test) or more than two groups (using analysis of variance, Anova models). In contrast to regular hypothesis testing, a very modest confidence level must be used, say 50% (or even lower) to ensure not only the most significant metabolites are retained. That could easily lead to missing a

multivariate set of metabolites describing group differences. By using a modest confidence level, even loosely associated metabolites can be selected.

Fold changes and *t*-tests are connected in a sense. Where fold changes only consider the means between groups (in the form of a ratio), *t*-tests also consider the variability of these means (using so called standard errors). When fold changes are large, but the variability of the estimates is also large, hypothesis tests will likely indicate there is no significant difference. Selection using a fold change criterion could still happen as it ignores the variability of the estimates. Graphically fold changes and *t*-test can be connected in Volcano plots, showing the magnitude of the difference and the significance of this difference (Hur et al., 2013).

Medium-level variable selection

The previous section described mainly methods for reducing the number of metabolites to be used in subsequent modeling. Subsequent modeling could be for example, a prediction model using some form of regression or a classification model. The idea is to only use the most promising metabolites in these analyses. Metabolites with general insufficient variation or variation insufficiently related to the trait of interest are discarded not to hamper these analyses. However, this still does not mean that all available metabolites for these analyses are all relevant and have equally important predictive properties. At this point it is also likely that we'll still have more metabolites present than we have samples or observations. In general, when we have more variables than data points, we have an ill-posed, or undetermined problem. There is simply not one unique solution that gives the best model fit. There are several ways of addressing this problem. One way is to use variable selection methods. The other is to use modeling techniques that explicitly handle large number of variables and have an embedded way of selecting variables.

Variable selection or wrapper methods

These methods have in common that they evaluate multiple models using different subsets of the metabolites. The purpose of analyzing different subsets of metabolites is to identify metabolites with the most predictive properties (Saeys et al., 2007). By analyzing the results from these subsets and combining different metabolites, they come to a (near) optimal set of metabolites for the prediction of the trait under investigation. Wrapper methods are multivariate in nature, in the sense that they study the predictive properties of (small) sets of metabolites simultaneously. Sets of metabolites that are chosen form a combination that together predict the trait of interest the best. They supplement each other in their predictive behavior.

Variable selection methods use ordinary least squares regression [or multiple linear regression (MLR) as it is sometimes called] for creating a model predicting

the trait from the metabolites. However, since we have more metabolites than samples, we need to reduce the number of metabolites in the model. Variable selection techniques aim at selecting only the most predictive metabolites.

There are two important approaches of variable selection: local methods for variable selection like for example, stepwise regression or global optimization algorithms.

Stepwise regression

Stepwise regression starts with an “empty” model, a model with only an intercept. In subsequent iterations metabolites are added one at a time to the model. The metabolite that gives the largest improvement to the model, as measured using F-tests, will be added to the model in each iteration. This iterative procedure of adding variables is repeated until the model cannot be improved anymore (the model does not change significantly). It can happen that a set of metabolites makes another metabolite obsolete as its variance is contained in the other metabolites. A metabolite is removed from the model when the model is not significantly different from a model with that metabolite present. There are two variants of stepwise regression. One is forward stepwise regression, and it does not allow the removal of variables. A second one is backward elimination. It starts with a model with all variables present. For metabolomics data, we typically have more metabolites compared to the number observations meaning this model cannot be calculated.

The result from stepwise regression is a small and compact set of metabolites that predict the trait of interest the best. It is not guaranteed to be the best subset possible. Stepwise is a greedy algorithm and can get stuck in local optima. Adaptations to stepwise regression are possible, one is e.g., to limit the number of selected metabolites to a prefixed number. The idea is that the first few selected metabolites are most important and explain the large majority of variance. Later added metabolites do not explain a lot of variance.

Global optimization algorithms

Stepwise regression can get stuck in a local optimum, meaning that the obtained solution (a set of metabolites) is not the best one possible. One way of overcoming this is the use of global optimization algorithms such as genetic algorithm (GA’s) ([Wehrens & Buydens, 1998](#)), simulated annealing (SA) ([Kirkpatrick et al., 1983](#)) and tabu search (TS) ([Glover, 1990](#)). A complete discussion of these techniques would be beyond the scope of this chapter. In short, GA’s mimic evolution and work on a group of trial solutions, called a population ([Hageman, van den Berg, et al., 2008](#)). A trial solution in our context would be a subset of metabolites. These trial solutions are recombined with each other (a process called cross-over) and mutated into new trial solutions (called mutation). All trial solutions are evaluated, in our context that means their predictive properties are assessed. The best trials solutions, the ones with the best predictive properties, are kept and

serve as a starting point for the next generation. The process is repeated until, for a prespecified number of generations, no improvement has been encountered.

SA works on a single solution at a time. A solution in our context is a subset of metabolites. By taking small steps in the search space (and in this context steps in the search space are for example, changing one metabolite for another one) predictive properties of the model will change. Some metabolites will improve the model, whilst others will deteriorate the model. Changes that improve the model are kept, while changes that deteriorate the model are kept with a certain probability. During the optimization, this probability of accepting a worse solution is getting smaller and smaller. This decreasing probability mimics the cooling of a metal, hence the name simulated annealing. By allowing steps that deteriorate the model's performance, the algorithm can overcome local optima and eventually (or hopefully) reach the global optimum.

TS also works on a single solution at a time ([Hageman et al., 2000, 2003](#); [Hageman, Wehrens, et al., 2003](#)). By modifying this solution (modifying in this context is adding/removing/changing selected metabolites), the search space around the current solution is investigated. Parts of the search space that have been visited are stored in memory and are never revisited again (they are taboo, hence the name). By forcing the algorithm to always explore a new part of the search space, the algorithm will eventually overcome local optima and should be able to reach the global optimum.

All three of these algorithms are finicky to use, because of the many settings they have. When used correctly they can be very powerful and deliver a compact set of metabolites able to best predict a certain trait.

High-level variable selection

Embedded methods for the selection of variables

Many statistical methods have been devised to get around the problem of having more variables compared to the number of observations. They do this by implicitly or explicitly selecting variables used in the model or by the creation of a set of low dimensional latent variables (LVs) ([Gareth et al., 2013](#)).

Regularization techniques

When we have more metabolites compared to the number of samples, the problem is ill-conditioned. The use of MLR would result in many models that fit the data equally well and would highly overfit the data. It has become likely that MLR would achieve low errors by fitting random fluctuations in the metabolite data that do not represent the true relationships between metabolites and the response. One way to overcome ill-conditioning is using constraints or penalty functions on the regression coefficients. This is called regularization. The idea is that besides the original loss function from MLR an added regularization term or

penalty is added to the equation (see Eq. (10.1)). Here n is the number of observations and p the number of predictors.

$$\text{Error} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \text{penalty term on } \beta's \quad (10.1)$$

Two main regularization strategies are ridge regression and the lasso both differing in the definition of the penalty term. Ridge regression uses the so-called L2-norm, while the lasso using the L1-norm (Tibshirani, 1996). The L2-norm minimizes the sum of squared regression coefficients; the L1-norm minimizes the sum of absolute regression coefficients (see Eqs. (10.2) and (10.3)).

$$\text{Ridge penalty term (L2 norm)}: \lambda \sum_{j=1}^p \beta_j^2 \quad (10.2)$$

$$\text{Lasso penalty term (L1 norm)}: \lambda \sum_{j=1}^p |\beta_j| \quad (10.3)$$

So, regularization methods do not only optimize the fit of the data using the ordinary loss function, but they also minimize the regression coefficients themselves (Gareth et al., 2013). The lasso is of special interest for us since it has variable selection properties. While the L2-norm has the effect that many regression coefficients are shrunken toward zero, they never reach exactly zero. With the lasso and the L1-norm this is different, the L1-norm enables the deselection of variables by setting their corresponding regression coefficients exactly to zero. When regression coefficients reach exactly zero, they are effectively removed from the model, hence the variable selection properties of the lasso. The result of the lasso is a small subset of metabolites best able to predict the trait of interest. The balance between the loss function and the penalty term needs to be optimized by a meta-parameter (typically called λ) (Bujak et al., 2016).

Latent variable methods

A different strategy to deal with ill-conditioned problems is the use of LVs regression methods. Here, the metabolites are mathematically transformed into a low dimensional representation of the data. Each new dimension is to contain as much variation of the original the data. There are several ways these new dimensions can be constructed. With regression modeling in mind, two are of special interest, Principal Components Regression (PCR) and Partial Least Squares (PLS).

Principal component regression

Principal Component Analysis is a method for deriving dimension reduction by combining variables (metabolites in our case) into a small number of principal components (PCs) (Antonelli et al., 2019; Gareth et al., 2013). The PCs are constructed in such a way that each component describes as much of the variation of the data at hand. The components are ranked in decreasing order of explained

variance and are all uncorrelated (they are said to be orthogonal). Usually only a small number of PCs is needed to describe all relevant variation in a dataset; the remaining PCs describe very small amounts of variation usually associated with random noise. After the creation of the PCs, they are used in MLR as the explanatory variables, they take over the role of the metabolites.

Partial least squares

PCR involves the creation of PCs that describe as much of the variation as possible. In this step only the metabolites are involved, and the resulting PCs summarize these metabolites the best (Antonelli et al., 2019; Bujak et al., 2016). By focusing only on the variation in the metabolites, the important variation related to the prediction of the trait is not always retained in the first PCs. PLS solves this problem by the creation of LVs that do not only capture the variation of the metabolites but capture the variation in the metabolites that is relevant for the prediction of the response. PLS finds LVs that best summarize the metabolites and the response simultaneously.

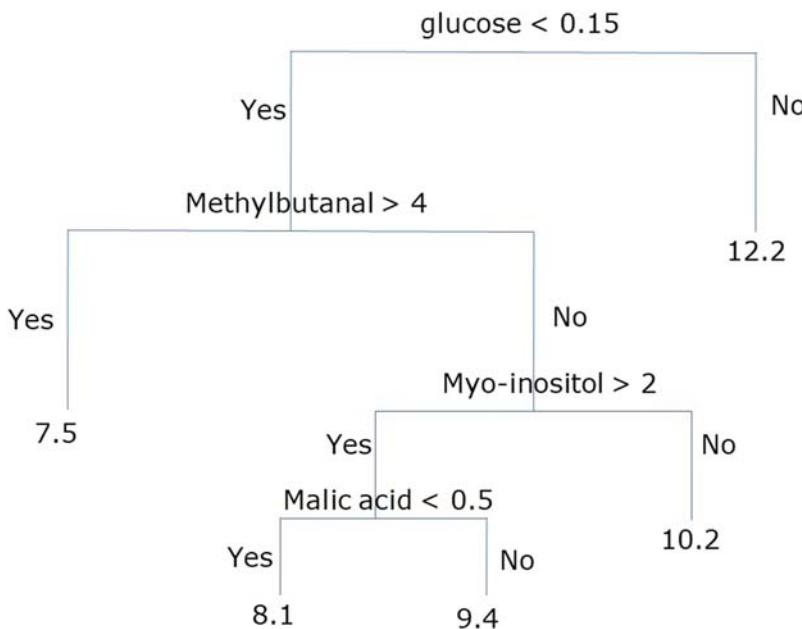
For both methods, the number of PCs or the number of LVs is a meta-parameter that needs to be optimized (Westerhuis et al., 2008).

PCs and LVs are typically difficult to interpret (Antonelli et al., 2019). They are a linear combination of all metabolites and as such all metabolites take part in the modeling. However, some metabolites are more important than other in the formation of the PCs/LVs. A metabolite's importance is expressed by so-called loadings. Loadings are weights that determine how much each individual metabolite contributes to a particular PC/LV. Inspection of the loadings will separate important metabolites from the unimportant ones; the higher the loading, the more important a certain metabolite is in that PC/LV. With PLS variable importance in projection (VIP) scores can be calculated. The importance of each variable is assessed, and VIP scores close to or greater to one are considered important in the PLS model.

There are several other modeling techniques that explicitly reduce or select variables to base their models on.

Decision trees

A decision tree is a flow chart like algorithm. It consists of nodes and branches where branches connect the nodes, and several branches emerge from a node. A node with no connection is typically called a leaf (Rokach, 2010). In decision trees, each node represents a test on a feature. Depending on the outcome of this test you traverse down different branches. This process is repeated several times until you reach a leaf (the end note). The leaf represents the predicted value. This can be a classification or a predicted variable in the case of a quantitative response. The test found on a nonend node relates to metabolites and contains a split point. It is a test on a single metabolite and depending on the outcome diverts to a certain branch. When following the node all the way to an end note, the end note contains a prediction for a certain object. An example tree is given in Fig. 10.1. Variable selection takes place at each

**FIGURE 10.1**

Decision tree. Example of decision tree. See text for details.

node since each node is a test on a single metabolite. Metabolites that do not appear in a node, are never selected, and do not influence the outcome. When training a tree, it is calculated how much each metabolite decreases the prediction error in a node. This is weighed by the chance to reach that node. The higher the value, called impurity, the more important a metabolite is.

Decision trees are easy to understand and interpret but are a bit unstable against small changes in the data (Gareth et al., 2013). They are also prone to overfitting which means they rather learn the specifics of the data set than generalize. This is something that can be reduced by pruning the tree which is the process of removing small noncritical branches.

Random forests

Random forests are an extension to decision trees. Random forests are an ensemble method and build and combines the output of many individual decision trees (Determan Jr, 2015; Gareth et al., 2013). Each tree in the forests is trained using a random subset of the original training set (a process called bootstrapping, see later in this chapter). As an addition, each time a tree is split, a random sample of the metabolites is chosen as split candidates. The number of metabolites available is typically the square root of the number of available metabolites. This has the

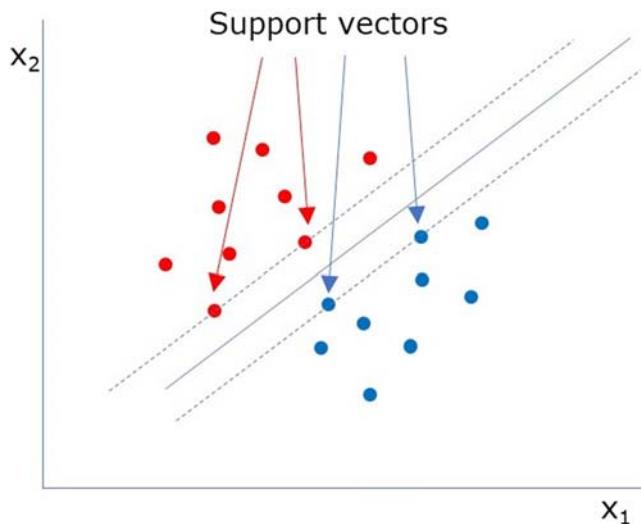
effect of forcing the decision tree to focus also on other important metabolites for predicting the response. It forces individual decision trees to look different from each other. This will make the trees uncorrelated and the average of the trees less variable and more reliable (Gareth et al., 2013). After training all trees in the forests (which is typically a large number, say 500–1000), the final prediction from the random forests is the average of all trees. The individual decision trees in random forests typically have many nodes and are not pruned. Variable selection in random forests is averaging all the impurity values from the individual decision trees. Alternative ways of variable selection in random forests are also present, one is the Boruta algorithm (Kursa & Rudnicki, 2010). Here, so called shadow features or randomized copies of the original variables are added to the data set. Random forests are trained, and a feature importance measure is applied. When the importance feature of actual metabolites is smaller compared to the ones from the randomized copies, these metabolites are removed. This process is repeated until convergence or a predefined number of forests.

Support vector machine

Super vector machine (SVM) is a machine learning algorithm typically used for classification but also able to predict quantitative traits (Grissa et al., 2016). SVMs are concerned with finding a hyperplane that best divides a dataset into two classes. Support vectors are the data points that are nearest to this hyperplane and thus define this hyperplane. Removal of these points would change the hyperplane. For a schematic overview see Fig. 10.2. The best hyperplane is the plane that separates the two classes while all data points are as far away as possible from the hyperplane (Vapnik, 1998). When no clear hyperplane exists, soft margins can be applied which allows for some misclassifications. When a nonlinear boundary is more appropriate, kernels can be used. Kernels allow for a transformation of the data after which a linear hyperplane can be found. Feature selection takes place by investigating the coefficients or weights of the model (Grissa et al., 2016). Large coefficients are deemed more important than smaller coefficients.

Heuristic approach

One recurring theme in variable selection with metabolomics is that the reported subset of metabolites is just one set out of many possible sets of metabolites that have a comparable fit. Owing to the correlated nature of many metabolites, metabolites have a certain interchangeable aspect with respect to their predictive properties. Methods with variable selection properties will include metabolites with high (or preferable the highest) explanatory properties in the model. Unfortunately, when sample sizes are low and the number of metabolites large (as is typically the case), the set of metabolites that is reported can show a

**FIGURE 10.2**

SVM hyperplane. Schematic overview of a separating (hyper)plane with several support vectors.

coincidental correlation with the trait of interest (Hageman et al., 2017; Westerhuis et al., 2008). This means that in the specific data set under study, metabolites appear to have important predictive properties but when the experiment and the statistical analysis is repeated, a (partly) different set of metabolites is reported. It is not always possible to repeat complete experiments, but we can mimic the process of analyzing different data sets using a computer. The central idea is that small changes in the data can result in substantial changes of the selected metabolites. Ideally, the set of selected metabolites would not change and would always be the same for every time we perturb our data set. Metabolites that are selected despite the perturbations of the data set are stable and, quite likely, the ones that carry the best predictive properties. On the other hand, metabolites that are not selected very often represent, most likely, coincidental correlations and probably do not hold up in future experiments.

Bootstrap and stability selection

Bootstrapping is a statistical procedure that resamples a single data set to create many simulated samples (Efron & Tibshirani, 1994). Bootstrapping allows you to calculate standard errors, confidence intervals and perform hypothesis tests for different kinds of sample statistics like for example, population mean, population difference of means etc. Under the correct assumptions these are sample statistics we could investigate using t distributions, but when we cannot make these

assumptions, bootstrapping can provide us with the required estimates. The power of the bootstrap lies in situations where there is no easy alternative available. We will not use the bootstrap only for estimation of for example, prediction quality but we will also study the metabolites that are selected during the bootstrap procedure.

During bootstrapping, a new and perturbed sample is generated by random drawing with replacement from the original sample. This approach allows different metabolites to be selected in the regression models and reveal their potential importance in each model.

The statistical analysis is performed on this newly created sample using a modeling method of choice. The only requirement for the statistical analysis is that it must perform some form of metabolite selection or be able to indicate the most important ones. The creation of a perturbed data set is repeated many times (e.g., 100 times) and each time the selected metabolites are stored. After the bootstrap procedure, we have 100 models. We can use the information from these models in two ways. First, we can use all 100 prediction errors to give us an idea of prediction variability, this is what the bootstrap is classically used for. Next, from the 100 models, we create an overview of the metabolites that have been selected the most. In follow up research we should devote our attention to the metabolites that have been selected the most. If no stable metabolites can be identified, the conclusion should be that, despite sets of metabolites being reported as predictive, none of them can be used reliably as small perturbations change the selected set.

Cross validation

Where bootstrapping perturbs the data randomly, we can also use a more systematic way for creating different “takes” on the existing data. One such mechanism is cross validation, sometimes called jackknife ([Rubingh et al., 2006](#)).

Cross validation is typically used to assess the prediction quality for objects that have not been used in the construction of the model ([Hageman et al., 2017](#); [Takahashi et al., 2020](#); [Westerhuis et al., 2008](#)). The rationale behind this is that when models predict truly unknown observations and not just observations used to train the model, we get the best indication of the prediction quality of the model. When data is scarce, we cannot sacrifice part of the data to serve as an independent test set. To solve this problem, we can divide the data into several groups, say k . What we do next is always leave out one group, build the model on the remaining groups and we predict the left-out group of observations. This gets repeated so that every group is left out once and serves as an independent test set to get predicted. The predictions of all left out objects is combined to give an impression of the prediction accuracy. This mechanism also allows us to study the variability between all the different k models with respect to the selected metabolites. Again, metabolites that are selected constantly in most of the models are the most worthwhile ([Wehrens et al., 2011](#)). Metabolites selected only

**FIGURE 10.3**

Schematic overview of fivefold cross validation. In each iteration one fifth, or one fold is left out the model building. Test error (here root mean square error) is calculated using the left out fold. Together with the prediction error, the most important metabolites (indicated by a letter) of each model in each iteration is stored. Metabolites most often encountered are the overall most important ones, in this example a, g and e.

incidentally probably represent accidental correlations and should not be pursued in follow up research. See for a schematic overview of this principle [Fig. 10.3](#).

The number of groups, k , is typically something like 5 or 10, allowing for the creation of 5 or 10 models. It is possible to create as many groups as there are objects. This is referred to as leave-one-out cross validation or jackknife.

Concluding remarks

Typical metabolomics data have inherent statistical difficulties. It contains many more metabolites compared to the number of objects while it is expected that most metabolites are not related or responding to the phenomena under investigation. This poses some modeling difficulties: regular regression techniques do not work unless variables are selected. Variable selection can be difficult to execute correctly. Deselecting uninformative metabolites can be a useful first step. Next, when a small, predefined number of metabolites is required, variable selection using wrapper techniques can provide a solution with a predefined number of metabolites. If the exact number of selected metabolites is not an issue, intrinsic variable selection methods, like lasso or random forests, are good candidates. Techniques like cross validation in concert with stability selection criteria can

provide even more detailed information. Besides an estimate of the prediction quality for unknown observations, it prevents, to a certain extent, the selection of spurious metabolites, focusing on metabolites that have shown repeatedly a relationship to the trait under investigation.

References

- Antonelli, J., Claggett, B. L., Henglin, M., Kim, A., Ovsak, G., Kim, N., Deng, K., Rao, K., Tyagi, O., & Watrous, J. D. (2019). Statistical workflow for feature selection in human metabolomics data. *Metabolites*, 9(7), 143.
- Bujak, R., Daghir-Wojtkowiak, E., Kaliszan, R., & Markuszewski, M. J. (2016). PLS-based and regularization-based methods for the selection of relevant variables in non-targeted metabolomics data. *Frontiers in Molecular Biosciences*, 3, 35.
- Determan, C. E., Jr (2015). Optimal algorithm for metabolomics classification and feature selection varies by dataset. *International Journal of Biology*, 7(1), 100.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78.
- Doeswijk, T., Smilde, A., Hageman, J., Westerhuis, J., & Van Eeuwijk, F. (2011). On the increase of predictive performance with high-level data fusion. *Analytica Chimica Acta*, 705(1–2), 41–47.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: With applications in R*. Springer.
- Glover, F. (1990). Tabu search: A tutorial. *Interfaces*, 20(4), 74–94.
- Grissa, D., Péterá, M., Brandolini, M., Napoli, A., Comte, B., & Pujos-Guillot, E. (2016). Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Frontiers in Molecular Biosciences*, 3, 30.
- Hageman, J., Wehrens, R., de Gelder, R., Leo Meerts, W., & Buydens, L. (2000). Direct determination of molecular constants from rovibronic spectra with genetic algorithms. *The Journal of Chemical Physics*, 113(18), 7955–7962.
- Hageman, J., Wehrens, R., Van Sprang, H., & Buydens, L. (2003). Hybrid genetic algorithm–tabu search approach for optimising multilayer optical coatings. *Analytica Chimica Acta*, 490(1–2), 211–222.
- Hageman, J., Streppel, M., Wehrens, R., & Buydens, L. (2003). Wavelength selection with Tabu search. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(8–9), 427–437.
- Hageman, J. A., Hendriks, M. M., Westerhuis, J. A., Van Der Werf, M. J., Berger, R., & Smilde, A. K. (2008). Simplivariate models: Ideas and first examples. *PLoS One*, 3(9), e3259.
- Hageman, J. A., Engel, B., de Vos, R. C. H., Mumm, R., Hall, R. D., Jwanro, H., Crouzillat, D., Spadone, J. C., & van Eeuwijk, F. A. (2017). *Robust and confident predictor selection in metabolomics* (pp. 239–257). Springer International Publishing. Available from https://doi.org/10.1007/978-3-319-45809-0_13.
- Hageman, J. A., van den Berg, R. A., Westerhuis, J. A., van der Werf, M. J., & Smilde, A. K. (2008). Genetic algorithm based two-mode clustering of metabolomics data. *Metabolomics*, 4(2), 141–149.

- He, Z., & Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4), 215–225.
- Hendriks, M. M., van Eeuwijk, F. A., Jellema, R. H., Westerhuis, J. A., Reijmers, T. H., Hoefsloot, H. C., & Smilde, A. K. (2011). Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry*, 30(10), 1685–1698.
- Hur, M., Campbell, A. A., Almeida-de-Macedo, M., Li, L., Ransom, N., Jose, A., Crispin, M., Nikolau, B. J., & Wurtele, E. S. (2013). A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Natural Product Reports*, 30(4), 565–583.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13.
- Lindinger, C., Pollien, P., de Vos, R. C., Tikunov, Y., Hageman, J. A., Lambot, C., Fumeaux, R., Voirol-Baliguet, E., & Blank, I. (2009). Identification of ethyl formate as a quality marker of the fermented off-note in coffee by a nontargeted chemometric approach. *Journal of Agricultural and Food Chemistry*, 57(21), 9972–9978.
- Lombardo, V. A., Osorio, S., Borsani, J., Lauxmann, M. A., Bustamante, C. A., Budde, C. O., Andreo, C. S., Lara, M. V., Fernie, A. R., & Drincovich, M. F. (2011). Metabolic profiling during peach fruit development and ripening reveals the metabolic networks that underpin each developmental stage. *Plant Physiology*, 157(4), 1696–1710.
- Madsen, R., Lundstedt, T., & Trygg, J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta*, 659(1), 23–33. Available from <https://doi.org/10.1016/j.aca.2009.11.042>.
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1), 1–39.
- Rubingh, C. M., Bijlsma, S., Derkx, E. P., Bobeldijk, I., Verheij, E. R., Kochhar, S., & Smilde, A. K. (2006). Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics*, 2(2), 53–61.
- Saccenti, E., Westerhuis, J. A., Smilde, A. K., van der Werf, M. J., Hageman, J. A., & Hendriks, M. M. (2011). Simplivariiate models: Uncovering the underlying biology in functional genomics data. *PloS One*, 6(6), e20747.
- Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3), 361–374.
- Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.
- Shahrjooihaghghi, A., Frigui, H., Zhang, X., Wei, X., Shi, B., & Trabelsi, A. (2017). An ensemble feature selection method for biomarker discovery. In *2017 IEEE international symposium on signal processing and information technology (ISSPIT)* (pp. 416–421). IEEE.
- Takahashi, Y., Ueki, M., Yamada, M., Tamiya, G., Motoike, I. N., Saigusa, D., Sakurai, M., Nagami, F., Ogishima, S., & Koshiba, S. (2020). Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Translational Psychiatry*, 10(1), 1–12.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

- van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7(1), 1–15.
- Vapnik, V. (1998). *The support vector method of function estimation. Nonlinear modeling* (pp. 55–85). Springer.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology*, 54(1), 669–689.
- Wehrens, R., & Buydens, L. M. (1998). Evolutionary optimisation: A tutorial. *TrAC Trends in Analytical Chemistry*, 17(4), 193–203.
- Wehrens, R., Franceschi, P., Vrhovsek, U., & Mattivi, F. (2011). Stability-based biomarker selection. *Analytica Chimica Acta*, 705(1–2), 15–23.
- Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., van Duijnhoven, J. P., & van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics*, 4(1), 81–89.

Pathway analysis

11

Rachel Cavill¹ and Jildau Bouwman²

¹*Data Science and Knowledge Engineering, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands*

²*Department of Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research (TNO), Zeist, The Netherlands*

Metabolites ontology

Introduction to ontologies

The word ontology originates from the Greek words Onto (what is) and Logica (logical discourse), and wikipedia therefore describes it as the study of being (<https://en.wikipedia.org/wiki/Ontology>). The main reason that this “study of being” is so essential in science, is that if we do not know what we study and what others are studying and how this concept relates to other concepts, it is hard to judge the outcomes. Therefore, concepts are described in detail in ontologies and also the relation of related concepts is made clear. This old concept of ontologies has gained more attention recently as analysis has become something that computers generally do. Automation of this work made it even more essential that the terminology was more sharply defined than if there is human interference in the analysis and interpretation of the data as computers are currently unable to consider context in the interpretation of terms. The development of the FAIR principles (Findable, Accessible, Interoperable and Reusable) further stimulated the ontology field. If talking about ontologies we also should describe the distinction between ontologies, taxonomies and vocabularies. Taxonomies are a classification system that is composed of hierarchical trees that minimize ambiguity. Because ontologies do not need to be hierarchically built, more relations are acceptable and a better representation of reality is possible. Vocabularies are lists of defined terms. Those vocabularies can be used as a bases of ontologies.

Ontologies for metabolites

For metabolomics, ontologies are required for the metabolites themselves: which metabolite did we measure and what are related metabolites? But also ontologies

are needed to specify how the metabolomics analysis was performed and how the study is executed. In this section ontologies for metabolites, metabolic pathways and study and metabolomics data analysis will be described. Metabolites can be defined by their chemical structure and those are well listed in databases such as pubchem (<https://pubchem.ncbi.nlm.nih.gov>), chemspider (<http://www.chemspider.com>) and for metabolites found in humans in the Human Metabolome Database (HMDB) (<https://hmdb.ca>). However, those databases do not describe the biological relation between and definition of the components. A chemical description does not always suffice for the biological definition of a metabolite. For example the chemical structure lactate is very well described in databases such as chemspider and pubchem. However, if lactate is found in a biological sample, the preparation of the sample can have changed lactic acid into lactate, just by changing the pH. Chemical databases describe those components separately and do not define this relationship, which makes it hard to connect outcomes of different studies if one reports that lactate (<http://www.chemspider.com/Chemical-Structure.82564.html?rid=f2d7625a-638d4b6f-940c-4696c4c1d420>) is measured and the other lactic acid (<http://www.chemspider.com/ChemicalStructure.592.html?rid=285cc1a4-d6dc-46ed-abea-63a38fc90be6pagenum=0>). Another example is glucose: if we find glucose in a human biological sample, most likely this is D – glucose (<https://pubchem.ncbi.nlm.nih.gov/compound/107526>) and in the original sample it is probably in equilibrium between the cyclic (<https://pubchem.ncbi.nlm.nih.gov/compound/5793>) and linear forms (<https://pubchem.ncbi.nlm.nih.gov/compound/107526>). Chemically these molecules look different and therefore they have different identifiers in chemical databases. However, describing finding glucose in the sample should be done in a more generic way. CheBi is a taxonomy that is specifically developed for this purpose. Glucose is here also defined as a generic component (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:1723>) and relates to D – glucose as an incoming relation and to aldohexose as an outgoing relation. This means that if one study is reporting that glucose was measured and the other that D – glucose was measured that those outcomes can be connected by the connection defined in CheBi.

For lipids the situation is even more complex: the current metabolomics methods are unable to specify the side – chains of, for instance, triglycerides. Chemical databases will have a separate identifier for every triglyceride while in practice sometimes a whole group of molecules are measured at once: TG (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:17855>), the group triacylglycerol 48:2 (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:85725>) or the specific molecule (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:140879>). As shown, CheBi is able to define all of those molecules and groups specifically and the relations between them are included. LipidMaps (<https://www.lipidmaps.org>) is another database that specifies lipid groups and molecules. A special type of relation between metabolites is the biochemical relationship between metabolites in pathways (further described below).

For metabolomics research the experimental (meta)data can also be described by ontologies or vocabularies. Experimental data includes data on the spectral details of a metabolite, helping identification of metabolites. Examples of databases that store spectral data are HMDB (<https://hmdb.ca>) and massbank (<https://massbank.eu/MassBank/>). Spectral information can only be interpreted if details about the experimental procedure are also stored. Standards for that are defined by the Metabolomics society in the MSI (Fiehn et al. 2007). These type of standards are structured in vocabularies for instance, for NMR in the NMR controlled vocabulary (<http://bioportal.bioontology.org/ontologies/NMR>) and for massspectrometry in the Mass Spectrometry Ontology (<http://bioportal.bioontology.org/ontologies/Ms>). For study data of studies that include metabolomics analyzes data can be stored in a structured way in Metabolights (<https://www.ebi.ac.uk/metabolights/>). This system makes use of several standards to structure the data. For study design details the tabular format of ISA – TAB is used. If all the above standards are used to describe metabolomics data and data collection then outcomes of studies can be more easily analyzed in an automated way. This will enable faster understanding about metabolism in biological systems and more straightforward comparisons between experiments.

Common metabolite databases

There are many common databases for metabolites, this section covers a few of the key databases. Importantly, databases which contain both metabolites and metabolic pathways are described in the next section.

Human metabolome database

The HMDB is now over 10 years old (Wishart et al., 2007, 2018). Its aim is to give metabolomics researchers working in humans a core resource, covering the metabolites which they may encounter in their samples in a consistent manner and more recently also providing access to tools related to assignment of these metabolites from spectra, in particular MS/MS and GC-MS.

LipidMaps

LipidMaps was developed from a consortium of metabolomics researchers in 2007 (Schmelzer et al., 2007). The aim was to focus on an often ignored class of metabolites—lipids. Whereas in many pathway databases these can be grouped into coarse groupings, HDL, VHDL etc., for metabolomics researchers measuring very specific changes in lipid types and chain-lengths, a finer grain of detail is necessary to capture useful information about their pathway involvement. Initially the main effort was toward a database of structures, to catalog the diversity of human lipids produced by our cells. This effort is still ongoing. Some of these lipids have been incorporated into 10 pathways, originally these were available through the lipidmaps consortium, however they have since been added to wikipathways.

CheBi

CheBi is hosted by the European Bioinformatics Institute (EBI), it is not restricted to any particular organism nor class of metabolites, but instead aims to contain “information about chemical entities of biological interest” (Hastings et al., 2016). As such it is one of the broader databases for metabolites. It also contains an ontology, as described earlier in this chapter.

Common pathway databases

Before we can analyze data to find which pathways are enriched, active, changed under particular circumstances given that data, we need definitions of the pathways themselves. This background knowledge was originally found in biochemistry textbooks, but has been expanded through publicly available databases which summarize this knowledge, linking together the entities involved in each pathway and recording the known (or suspected/predicted) interactions between them. There are an increasing number of these pathway databases available, in this section we highlight some of the most common ones, describing their important features.

Kyoto encyclopedia of genes and genomes

The Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa & Goto, 2000), is over 20 years old. KEGG is based around a generic biological network, and then for each organism the elements of that network which are present in that specific organism can be highlighted. In this way, the pathway variants present in many different organisms can be represented efficiently. For metabolomics data KEGG pathways are based around the metabolic network, linking the genomes of the organisms described to this network through the EC numbers of the enzymes present. More recently (Minoru et al., 2017), KEGG has expanded to include smaller subpathways, they call modules and disease pathways linking genes, drugs, pathways and diagnostic markers.

Small molecules pathway database

The Small Molecules Pathway Database (SMPDB) (Frolikis et al., 2010) was developed by David Wishart’s group, who also developed the well-regarded metaboanalyst tool discussed later. It links to the HMDB. This database was developed specifically with the needs of metabolomics researchers in mind, in particular clinical metabolomics, relating to human metabolism, human disease and drug therapy. Despite its roots as a human database, it has since been expanded to mice, *E. coli*, yeast and arabidopsis thaliana, and now contains almost 50,000 pathways for these organisms including over 20,000 disease pathways.

Consensus path database

With so many pathway databases available, the choices for a researcher or software developer of which pathway set to use can be tricky. Even pathways named the same may have differing boundaries, or even occasionally be completely non-overlapping, when examined in different databases. Thus, ConsensusPathDB was developed (Kamburov et al., 2011). The aim here was to take all publicly available open pathway sources and map them on top of each other. As well as allowing users to search pathways and interactions from 32 different public resources (including most of the others discussed in this section), they also allow visualization of these databases and the interactions contained within them. For each link in their network it is possible to visualize which database/s contain this link.

Reactome

Reactome (Joshi-Tope et al., 2005) is a human-centric pathway database. They group their 1803 pathways into 26 super-pathways (Jassal et al., 2020), and additionally included 484 disease versions of the super-pathways. More recently, they have extended their efforts to collaborate with experts in particular subject areas for annotation, into allowing users with ORCID records (Haak et al., 2012) to participate in the review process for their data.

Wikipathways

Wikipathways takes this idea of user annotation and review to another level. Treating their pathway database as a “wiki” in where interested biologists are encouraged to add, edit, annotate and curate the content (Pico et al., 2008). By crowd-sourcing pathway creation and curation to the experts it allows for academic communities to have ownership of the process. In 2018, they reported having had 634 individual curators involved in the process (Slenter et al., 2018). Wikipathways is increasingly useful for metabolomics research, as the annotated metabolites are increasing and they have been specifically focusing on adding new metabolic content to the database.

Metabolic pathway analysis

There have been a range of algorithms developed for pathway analysis, in this section we split them into three categories; Overrepresentation algorithms being the simplest methods, using sets of metabolites and looking for significant overlaps; Enrichment methods which take values for all measured metabolites and look for pathway sets which have unusual sets of values [see for instance (Ackermann & Strimmer, 2009) for a good classification of enrichment methods used with gene expression data]; and finally topological methods which also take into account the network architecture of the pathway or even the genome-wide metabolic network.

Overrepresentation

Overrepresentation (Beißbarth & Speed, 2004) takes as an input a list of interesting metabolites. This list could be those whose levels are different between two groups by any statistical test (for instance a *t*-test), or those which are correlated to a particular endpoint, those which have high loadings in a multivariate model (PLS-DA or PCA for instance), or any other list of interest which has been derived by applying some criteria to your data.

Overrepresentation takes this input list and compares it to the list of each pathway in turn, calculating the overlap and evaluating the probability of such an overlap occurring. A multiple test correction (typically Benjamini-Hochberg FDR) will then be applied to correct for comparing against multiple pathways and a list of significant pathways is returned.

To calculate the probability of the observed overlap a cumulative hypergeometric test is used (also known as a cumulative Fisher's exact test). It is very important to use the cumulative versions here, as, among other reasons, this allows for comparisons between pathways of different sizes. A Fisher's exact test tells us the probability that this pathway will overlap with our input list exactly x times, then larger pathways will tend to have smaller probabilities as there are more options possible. By using a cumulative calculation we look at the probability that the overlap is x or greater (looking at the tail of the distribution) which compensates for these differences and allows us to find pathways of any size with unusual overlaps.

There are several assumptions implicit in this *P*-value calculation which bear examination;

1. It is assumed that it is possible for every metabolite from the pathway to appear in the input list.
2. The probability of each entity appearing in the overlap is assumed to be independent.

Generally speaking these assumptions do not hold in our metabolomics data, so we need to examine why these assumptions are violated how we can mitigate their effects when performing overrepresentation analysis.

For assumption 1 to hold we need to have reliable measurements (and assignments) for all metabolites in all pathways. Given no technology can yet deliver this, with metabolomic data we are generally restricted to having measured a subset of the globally present metabolites. Some will be undetectable by the methods used, others will be present at an abundance lower than the limit of detection, or their signal may be obscured by that of a more abundant metabolite (overlapping peaks).

When we perform metabolite enrichment, ignoring this assumption, we can easily obtain biased results. The pathways output will primarily reflect the pathways which contain metabolites which can be measured and easily assigned using our platform of choice. Mathematically this assumption can be mitigated if we

provide a second input list to the algorithm, those metabolites which were measured in our experiment. This second list is generally termed a background list, and tells the algorithm which metabolites from the genome-wide pathway database should be counted as possible candidates to appear in the overlap. However, with a typical metabolomic workflow, even this mitigation is very difficult or even impossible to obtain. The amount of effort needed to reliably identify all metabolites in a dataset is unrealistic in most settings. Using a background list of just the assigned metabolites would suffice, if it were not for the fact that we generally put more effort into assigning those metabolites which are deemed “interesting” by our statistical analysis or modeling, and therefore a bias remains.

Therefore, where background lists contain known biases in the assigned metabolites, or are completely infeasible, we recommend a resampling procedure. Taking the list of measured and assigned metabolites, select x , where x is the number of metabolites present in your original input list and then performing the overrepresentation analysis on the new dummy list of metabolites. By doing this process repeatedly you will notice which pathways repeatedly appear in your output list and their significance each time. You can then take this bias into account when interpreting your real output list.

For assumption 2 we need to think about the types of metabolites that appear in pathways and the types of metabolites which are measured. Up till now we have assumed there is a one-to-one mapping between a measured metabolite and a pathway entity. However, this is often not the case. Take for instance, chiral metabolites, very often the pathway metabolite will have a specific chirality, whereas many platforms will measure only a pool of all forms of the metabolite.

At this point it is important to be aware of the unique identifiers which your pathway database uses to recognize metabolites. Using English names for metabolites is potentially fraught with confusions. Labeling your measured metabolites with a unique identifier which is recognized by your pathway database is thus crucial. If you label with a different set of identifiers to those used by your pathway database, then a mapping will need to be used to convert your labels into those used by the database, so that the overlap can be calculated. Due to differences in the level at which the databases assign metabolites these mappings are rarely (if ever) 1–1, meaning that a single labeled metabolite may map to multiple metabolites in the new scheme (Martijn et al., 2010). This is one place where assumption 2 can be easily violated, if these multiple mapped metabolites are in the same pathway, then whenever the single metabolite is in your input list, you will have multiple entities in the input list in the new labeling scheme. Meaning any pathway in which this metabolite appears will have an increased probability of appearing spuriously significant. The best advice to avoid these issues is to find out which metabolite identifiers are natively used by the tool/pathway database you wish to apply, and to label your metabolites with these from the start.

However, even this won’t completely rectify this complex mapping issue. As there will still be crucial mapping decisions to be made which may have unforeseen impacts on the enrichment results you obtain. For instance, if you decide to

use KEGG metabolite identifiers, and you have measured a mixed pool of Lactate then you have a choice of three KEGG identifiers C01432 which is for Lactate, C00186 for L-Lactate and C00256 for D-Lactate. In this case it may seem obvious that the Lactate identifier would be the best choice, however if you look at KEGG pathways, then you see that this identifier does not occur in any of them. Mapping to both will incur the problem detailed in the previous paragraph when both L- and D-lactate occur in the same pathway (as sometimes happens). Currently our best advice is to take care with your selection of metabolite identifiers for your assigned metabolites and to be aware of these issues. The better you know your pathway database, the more easily you can select appropriate identifiers that relate to the pathways you are interested in testing for significance.

When we move toward lipid pathways for instance, the issue can become more complex. Our data may be telling us that particular lipids are increased or decreased in a situation, but often the pathways represent lipids at a different level, talking about HDL, or VHDL rather than specific species. Lipid specific resources can help alleviate this issue.

Ontologies may help with this issue in the future, as they give us a structured way to link these identifiers together, although methods that can use these ontologies to enhance overrepresentation analysis are still under development.

Where ontologies have already been extensively used is as a replacement for the pathway databases with these methods. This is equivalent to GOterm enrichment analysis with transcriptomics data. Here instead of looking for metabolites which are present in a particular pathway, we are searching for sets of metabolites which are associated with a particular ontological term. There are several tools for this analysis, for instance BiNChE ([Moreno et al., 2015](#)) which looks for overrepresentation of metabolites associated with particular terms in the ChEBI ontology or ChemRich ([Kumar & Oliver, 2017](#)) which looks at compounds associated with MeSH terms.

Enrichment

Going beyond simple overrepresentation we have enrichment methods such as Metabolite Set Enrichment Analysis (MSEA) or the Kolmogorov–Smirnov test (K–S test). These methods are based not on sets of metabolites, but on real values (a single value per metabolite) indicating the relative importance of all measured, assigned metabolites. The exact details of the null hypothesis varies from method to method, but the general idea is to find pathways where the metabolites given from those pathways show an unusual (however that is defined) pattern. As all measured, assigned metabolites are given values, this means a background list is unnecessary for these analyzes, a potential important advantage given the difficulties associated with obtaining a good background list.

The values used for enrichment tests can vary, they could be fold changes between groups, loadings from multivariate models (e.g., [Wagner, 2015](#)) or

correlations to a phenotypic endpoint. There are also two variants possible, one where you are looking at absolute deviation, so metabolites with high values (whether positive or negative) are at the top of the list, and the other lists them numerically. In the first instance we are searching for the situation where the metabolites are changed, without specifying the direction of change. In the second instance, we specify that the direction of change should be consistent within the group.

In this section we will summarize some of the main algorithms used for metabolite enrichment analysis.

Metabolite set enrichment analysis

Confusingly MSEA ([Xia & Wishart, 2010](#)), uses a different enrichment approach to its namesake Gene Set Enrichment Analysis ([Subramanian et al., 2005](#)). For MSEA they chose the globaltest algorithm ([Jelle et al., 2003](#)) as the enrichment test for their approach. This algorithm builds a generalized linear regression model between the metabolites in each pathway and the outcome feature of interest. It looks for sets of metabolites which are predictive of the outcome value. One key advantage of this approach is that the phenotypic labels can be binary, multiclass or continuous. Secondly, the computation of the *P*-values can be handled efficiently without the need for the large number of permutations to obtain accurate *P*-values to rank pathways as seen in GSEA ([Keller et al., 2007](#)).

Kolmogorov–Smirnov test

The K–S test looks at the distribution of all metabolite values (one value per metabolite) vs the distribution of those metabolites contained in the examined pathway. It looks at the cumulative distribution, that is, what percentage of the values for metabolites are above any particular threshold. By comparing this distribution for the pathway metabolites with the distribution for all metabolites it assesses at which point the difference between these two distributions is largest. For instance if 30% of the pathway metabolites had a value > 10 , but only 20% of all the metabolites had such a value, this would give a difference between the distributions at this point of 10%. If this was the largest such difference, then the *P*-value would be generated from this difference. The test calculates how likely it would be that you would see such a difference by chance, if the two datasets were drawn from the same distribution. The advantages of the K–S test are that it is very fast and precise, unlike some other methods which rely on permutations and can only become precise when millions of permutations have been performed ([Keller et al., 2007](#)). However, by focusing on the point of largest difference between the distributions it may not always reflect the overall pattern.

Wilcoxon signed rank test

The Wilcoxon signed rank test requires that the values entered into the algorithm are distributed around 0. This could mean they are fold changes, mean differences in abundance between two groups, correlations or loadings from a multivariate model. The test is looking for whether the values follow a symmetric distribution around 0. So for any particular pathway, it will examine and test whether this assumption of a symmetric distribution around 0 holds true for the metabolites in that pathway. Note, that in the case where metabolites may be both increased and decreased in a particular pathway, this assumption may hold even though the pathway contains many altered metabolites. As it is a rank test, it is nonparametric, this means that the actual magnitude of the value for each metabolite is not important, only its position in a ranked list of the absolute values. The test statistic is generated by summing the multiple of the sign of each value with its rank in this list.

Topological methods

The central idea between topological methods is that not all metabolites in a pathway should have equal weight. A pathway is not an unordered set, but rather it is based on a network. Through this network we can estimate the importance of each metabolite to the pathway's function. The critical assumption is that metabolites which are hubs in the pathway, central and well-connected will have a more critical impact on the pathway's function when they change, than metabolites which are neither central nor well-connected. One tool which has integrated this into their analysis is Metaboanalyst, which has topological analysis in the MetPA tool (Jianguo Xia & Wishart, 2010). Like overrepresentation analysis, MetPA requires a list of metabolites which are changed in an experiment. MetPA then examines either the betweenness centrality of each metabolite, or the out-degree centrality and then calculates the pathway impact of the changed metabolites by summing the centrality scores for all changed metabolites, normalized to the total sum of the centrality scores for all metabolites in the pathway.

Tools for metabolomic pathway analysis

There are an increasing array of tools available for metabolomic pathway analysis. The main tools are shown in Table 11.1 where we can see the identifiers used, the pathway databases available, the pathway analysis methods available and the format of the interface (website or programming language).

Several of these tools have noteworthy additional features. The possibility to generate R code through the website of Metaboanalyst (Chong et al., 2019) is a powerful move toward reproducibility of the data analysis process. Another feature seen in IMPaLA (Kamburov et al., 2011), PaintOomics (García-Alcalde et al., 2011), 3omics (Kuo et al., 2013), MarVis (Alexander et al., 2015),

Table 11.1 Common pathway analysis tools for metabolomics data.

Tool	Identifiers for metabolite input	Pathway databases used	Methods available	Format	Integration with transcriptomics data?
ConsensusPathDB	KEGG, Chebi, Pubchem, CAS, HMDB	ChembI, Drugbank, Biocarta, EHMN, HumanCyc, INOH, KEGG, Netpath, PID, Reactome, SMPDB, TTD, Wikipathways	Overrepresentation, Wilcoxon	Website	
IMPaLA	KEGG, Chebi, Pubchem, CAS, HMDB	Biocarta, EHMN, BioCyc, INOH, KEGG, Netpath, PID, PharmGKB, SignalLink, Reactome, SMPDB, Wikipathways	Overrepresentation, Wilcoxon	Website	y
MetaboAnalyst	Names, HMDB, KEGG, Pubchem, ChEBI, METLIN	SMPDB, KEGG, Biocarta, SNP-associated metabolite sets, User provided sets	Overrepresentation, MSEA, topological analysis	Website, R package	y
PaintOmics	KEGG	KEGG	Overrepresentation	Website	y
MPINet	Pubchem	ChembI, Drugbank, Biocarta, EHMN, HumanCyc, INOH, KEGG, Netpath, PID, Reactome, SMPDB, TTD, Wikipathways	Overrepresentation	R package	y
3omics	Pubchem	KEGG, HumanCyc	Overrepresentation	Website	y
MarVis	IDs, Names, Accurate masses	Kegg, BioCyc	Overrepresentation, K-S test, Wilcoxon	Website	y
LIPEA	KEGG, HMDB, abbreviations, swissLipids, lipidMaps, ChEBI	KEGG	Overrepresentation	Website	
Lipid mini-on	Names	LipidMaps	Overrepresentation	website	

MPINet (Li et al., 2011) as well as Metaboanalyst is the ability to combine the analysis of metabolomics and transcriptomics data through pathways by mapping both to the same pathways [see (Cavill et al., 2016) for a broader review of transcriptomic/metabolomic integration]. A couple of tools are specifically aimed at a specific class of metabolites, for instance LIPEA (Acevedo et al., 2018) and Lipid Mini-on (Clair et al., 2019) target lipid classes.

Conclusions

The drive toward reproducible research has led to the development of ontologies, which can be used to describe our metabolomic data, making our results easier to transfer, more robust, and reproducible. These ontologies can be used in combination with tools for finding enrichment or overrepresentation in metabolic pathways. Although there are still many potential issues with these tools and their use, as highlighted in this chapter, we believe that their potential for adding interpretability to our results means that we as a community must continue to develop these tools and strive to use them in a transparent, robust, and reproducible manner.

References

- Acevedo, A., Durán, C., Ciucci, S., Gerl, M., & Cannistraci, C. V. (2018). LIPEA: Lipid pathway enrichment analysis.
- Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1), 47.
- Alexander, K., Manuel, L., Kirstin, F., Alina, M., Ingo, H., Burkhard, M., Ivo, F., & Peter, M. (2015). MarVis-Pathway: Integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 764–777. Available from <https://doi.org/10.1007/s11306-014-0734-y>.
- Beißbarth, T., & Speed, T. P. (2004). GOstat: Find statistically overrepresented gene ontologies with a group of genes. *Bioinformatics (Oxford, England)*, 20(9), 1464–1465. Available from <https://doi.org/10.1093/bioinformatics/bth088>.
- Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17(5), 891–901. Available from <https://doi.org/10.1093/bib/bbv090>.
- Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current Protocols in Bioinformatics*, 68(1), e86. Available from <https://doi.org/10.1002/cpb1.86>.
- Clair, G., Reehl, S., Stratton, K. G., Monroe, M. E., Tfaily, M. M., Ansong, C., & Kyle, J. E. (2019). Lipid Mini-On: Mining and ontology tool for enrichment analysis of lipidomic data. *Bioinformatics (Oxford, England)*, 35(21), 4507–4508. Available from <https://doi.org/10.1093/bioinformatics/btz250>.
- Feng Li, Yanjun Xu, Desi Shang, Haixiu Yang, Wei Liu, Junwei Han, Zeguo Sun, Qianlan Yao, Chunlong Zhang, Jiquan Ma, Fei Su, Li Feng, Xinrui Shi, Yunpeng Zhang, Jing Li, Qi Gu, Xia Li, Chunquan Li. (2014). "MPINet: Metabolite pathway identification via

- coupling of global metabolite network structure and metabolomic profile. *BioMed Research International*, 2014, Article ID 325697, 14. Available from <https://doi.org/10.1155/2014/325697>.
- Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Srivastava, S., & Wishart, D. S. (2010). SMPDB: The small molecule pathway database. *Nucleic Acids Research*, 38, D480–D487.
- García-Alcalde, F., García-López, F., Dopazo, J., & Conesa, A. (2011). Paintomics: A web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics (Oxford, England)*, 27(1), 137–139. Available from <https://doi.org/10.1093/bioinformatics/btq594>.
- Haak, L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers (pp. 259–264). Learned Publishing. Available from <https://doi.org/10.1087/20120404>.
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1), D1214–D1219. Available from <https://doi.org/10.1093/nar/gkv1031>.
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., ... Eustachio, P. D. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*.
- Jelle, J., Goeman, S. a, Geer, Kort, Houwelingen, H. C., van, Houwelingen, H. C., van, Houwelingen, H. C. van, & Houwelingen, van (2003). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics (Oxford, England)*, 20(1), 93–99.
- Joshi-Tope, G., Gillespie, M., Vastrik, I., Eustachio, P., Schmidt, E., Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., & Stein, L. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33, D428–D432.
- Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R., & Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics (Oxford, England)*, 27(20), 2917–2918. Available from <https://doi.org/10.1093/bioinformatics/btr499>.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. Available from <https://doi.org/10.1093/nar/28.1.27>.
- Keller, A., Backes, C., & Lenhof, H.-P. (2007). Computation of significance scores of unweighted gene set enrichment analyzes. *BMC Bioinformatics*. Available from <https://doi.org/10.1186/1471-2105-8-290>.
- Kumar, B. D., & Oliver, F. (2017). Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports*. Available from <https://doi.org/10.1038/s41598-017-15231-w>.
- Kuo, T. C., Tian, T. F., & Tseng, Y. J. (2013). 3Omics: A webbased systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*.
- Fiehn, O., Robertson, D., Griffin, J. et al. (2007). The metabolomics standards initiative (MSI). *Metabolomics* 3, 175–178. Available from <https://doi.org/10.1007/s11306-007-0070-6>.
- Martijn, P., Iersel, Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., & Evelo, C. T. (2010). The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1), 5.

- Minoru, K., Miho, F., Mao, T., Yoko, S., & Kanae, M. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, D353–D361. Available from <https://doi.org/10.1093/nar/gkw1092>.
- Moreno, P., Beisken, S., Harsha, B., Muthukrishnan, V., Tudose, I., Dekker, A., Dornfeldt, S., Taruttis, F., Grosse, I., Hastings, J., Neumann, S., & Steinbeck, C. (2015). BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics*. Available from <https://doi.org/10.1186/s12859-015-0486-3>.
- Pico, A., Kelder, T., van Iersel, M., Hanspers, K., Conklin, B., & Evelo, C. (2008). WikiPathways: Pathway editing for the people. *PLoS Biology*, e184. Available from <https://doi.org/10.1371/journal.pbio.0060184>.
- Schmelzer, K., Fahy, E., Subramaniam, S., & Dennis, E. A. (2007). The lipid maps initiative in lipidomics. *Methods in Enzymology*, 432, 171–183. Available from [https://doi.org/10.1016/S0076-6879\(07\)32007-7](https://doi.org/10.1016/S0076-6879(07)32007-7).
- Slenter, D., Kutmon, M., Hanspers, K., Ruitta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S., Dinges, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L., ... Willighagen, E. (2018). WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, D661–D667. Available from <https://doi.org/10.1093/nar/gkx1064>.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. Available from <https://doi.org/10.1073/pnas.0506580102>.
- Wagner, F. (2015). GO-PCA: An unsupervised method to explore gene expression data using prior knowledge. *PLoS One*, 10(11), e0143196. Available from <https://doi.org/10.1371/journal.pone.0143196>.
- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., ... Scalbert, A. (2018). HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617. Available from <https://doi.org/10.1093/nar/gkx1089>.
- Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., ... Querengesser, L. (2007). HMDB: The Human Metabolome Database. *Nucleic Acids Research*, 35(Suppl. 1), D521–D526. Available from <https://doi.org/10.1093/nar/gkl923>.
- Xia, J., & Wishart, D. S. (2010). MetPA: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics (Oxford, England)*, 26(18), 2342–2344. Available from <https://doi.org/10.1093/bioinformatics/btq418>.
- Xia, Jianguo, & Wishart, D. (2010). MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, W71–W77. Available from <https://doi.org/10.1093/nar/gkq329>.

SECTION

Application

3

This page intentionally left blank

Cell culture metabolomics and lipidomics

12

Irina Alecu^{1,*}, Carmen Daniela Sosa-Miranda^{2,3,*}, Jagdeep K. Sandhu^{2,4},
Steffany A.L. Bennett¹, and Miroslava Cuperlovic-Culf^{4,5}

¹*Neural Regeneration Laboratory, Ottawa Institute of Systems Biology, Brain and Mind Research Institute, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada*

²*Human Health Therapeutics Research Centre, National Research Council of Canada, Ottawa, ON, Canada*

³*Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada*

⁴*Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada*

⁵*Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada*

Introduction

Cell culture, the process of maintaining and growing cells *in vitro* under controlled conditions outside of living organisms, is one of the major tools of biology and biochemistry as well as medicinal and environmental chemistry, systems biology, and biotechnology. Both primary cell cultures, that is dispersed cells that are cultured directly from tissues and have limited lifespan, and cell lines, immortalized cells that can be cultured indefinitely, provide excellent models for studying cell physiology and biochemistry in health and disease or test drugs or toxins. Cell cultures are also used for the production of biologics, vaccine particles and gene therapy components, or for bioprocessing and bioremediation. In all of these and many other applications, metabolomics and lipidomics provide crucial molecular data for the optimization or modeling of cell growth, or analysis of effects of treatments or gene mutations. Metabolomics, the high throughput method measuring metabolites, provides a method for understanding biology, assessing cells' health, monitoring toxins or drugs, determining needs for growth or cell passaging. Metabolomics can also provide, on its own or in combination with other omics methods, data for predictive modeling of cell behavior. Analysis of metabolites gives the closest molecular data to phenotype, showing the outcome of

* Equal first authors.

combined genetic, epigenetic, and environmental effects on the behavior and characteristics of biological systems.

Interest in the application of metabolomics for analysis and utilization of cell cultures, possibly combined with other types of omics investigations has greatly increased due to a number of technological and analytical advances and a wide range of cell culture applications. Metabolomics has been utilized in many cell culture experiments with some examples including the analysis of microbial extracellular metabolite production ([Pinu & Villas-Boas, 2017](#)), test antimicrobial drugs in high throughput ([Campos & Zampieri, 2019](#)), explore human cell biotransformation of xenobiotics ([Flasch et al., 2020](#)) or define cancer cell characteristics ([Li et al., 2019](#)).

In addition to the major ethical and humane advantages of cell cultures compared to animal models, they can be utilized in a fully controlled environment thereby limiting sample variability and providing high statistical power even with small numbers of biological replicates. Additionally, cell cultures are generally a cost-effective method for initial or high throughput testing of, for example, drugs or toxins ([Flasch et al., 2020](#); [Muschet et al., 2016](#)). In spite of a number of advantages several possible issues need to be carefully considered in metabolomics examination of cell cultures including cell type and growth media and environment selection, mode of harvesting, quenching of metabolism, cell passage age, data preprocessing including metabolites assignment and quantification, data processing including normalization and finally selection of appropriate analysis tools ([Čuperlović-Culf et al., 2010](#)) [Table 12.1](#).

Broadly, application of cell culture metabolomics can be divided into:

1. monitoring of cell culture state and cell biology in single cultures or in different coculturing scenarios;
2. testing effect of a treatment including drugs, growth media, toxins or gene editing approaches;
3. analysis of metabolic processes including metabolic flux under different conditions possibly with isotopic labeling;
4. production of biological material (biologics, vaccines, gene therapy) or bioremediation;
5. cell therapy production.

The cell metabolism can be observed through analysis of: extracellular media (metabolic footprint analysis) which could include media or extracellular particles, cell or cell organelle extract either for bulk or single cell analysis (metabolic fingerprinting) which could include analysis of metabolic extracts or in-cell analysis which includes *in vivo* analysis of metabolism in cells. For each of the application modes, sample preparation procedures have to be optimized for the experimental goals, cell properties and analytical tools. Sample preparation protocol depends on the cell culture properties (adherent or suspension, 2D or 3D, single cell), cell type, chemical properties of metabolites of interest (in targeted approach) or aims to cover as many metabolites as possible (untargeted),

Table 12.1 Cell culture metabolomics experimentation steps.

Process	Major considerations	Solution examples
Study design	Cell type; Study type; Sample type	Cell culture type has to be well defined Treatment analyses or growth optimization Extra- intra-cellular; hydrophilic and lipophilic
Sample collection and storage	Standard operating procedure; Medium use and addition; Sample quantities	Compatibility between centers and during study Fed batch versus profusion Number of cells as well as amount of material Material storage
Sample preparation	Metabolite extraction possibly with Derivatization	Method selection; selection of metabolite groups; Changing of biochemical properties for measurement
Sample analysis	Method identification	NMR, LC-MS/MS, GC-MS/MS or another
Data analysis	Statistical, Unsupervised or Supervised Machine learning	Correlation; fold changes; Clustering or visualization; Feature selection; classification
Modeling	Mechanistic modeling Machine learning Hybrid methods	Correlation with other data; Pathway and network analysis; Predictive modeling

The most important steps in the cell culture metabolomics experimentation with some important considerations and possible general solutions.

detection techniques [e.g., Nuclear Magnetic Resonance—NMR spectroscopy or Mass Spectrometry (MS)] and type of analysis. In this chapter we will provide several examples of applications of cell culture metabolomics and lipidomics with detailed protocols for sample preparation particularly for mass spectrometric analysis of lipids and metabolites and analysis of extracellular vesicles (EVs). This will be followed by an introduction of approaches for metabolomics analysis for cell processes description, modeling and design.

Sample processing and experimentation for cell culture lipidomics and metabolomics

Methods for optimized metabolite and lipid extractions for cell culture analysis

The metabolome has a major chemical and physical diversity, including both highly hydrophobic lipids such as triglycerides and highly hydrophilic compounds such as sugars, with partition coefficient values spanning 40 orders of magnitude (Cajka & Fiehn, 2016). The huge diversity led to “divide and conquer”

approaches like metabolomics and lipidomics in order to be able to increase coverage of the metabolites measured, wherein water-soluble (polar) metabolites and water-insoluble (hydrophobic) lipids are extracted with different methods. It is important to note here that lipids will partition into the organic phase, whereas most metabolites will partition into the aqueous phase. Thus, either a highly optimized procedure for separation from the same sample or enough replicates are necessary for analysis of both fractions requiring large amounts of biological material for combined metabolomic and lipidomic analysis. One of the important benefits of using cell culture for the analysis of metabolomic/lipidomic changes in response to a variety of treatments, is the possibility for multiple replicates of the same condition as well as possibility for increase of sample size as needed. Exploration of functional and dysfunctional metabolic mechanisms and pathways simultaneously, requires unbiased measurement of a maximal number of lipids and metabolites (Dunn et al., 2005). This section will discuss various lipid and metabolite extraction protocols which can be used to maximize the efficiency for parallel lipidomics and metabolomics analysis of cells.

As an example of the power of parallel metabolite and lipid analysis, we use the study by Zhen et al. which utilized a cell culture model to determine the ecotoxicological effects of chemicals in the aquatic environment (Zhen et al., 2018). In this work zebrafish liver cells were exposed to wastewater treatment plant effluent collected at various distances from the discharging point. They then analyzed both hydrophilic metabolites and lipids. While the effects on the hydrophilic metabolome diminished with increasing distance from the discharge point, the effects on the lipidome increased. The study demonstrated the utility of cell-based systems as a tool to determine impact on both the metabolome and lipidome, as well as the importance of studying both hydrophilic and hydrophobic metabolites in order to be able to fully assess the biological effects of various treatments.

The first step which must be performed in order to be able to analyze changes or disturbances in metabolites and lipids is their extraction from cells and cell media. The general principle of lipid extraction is, simply, mixing of an aqueous solvent with an organic solvent and then separating the phases by centrifugations, wherein the lipids partition to the organic phase, while proteins, many hydrophilic metabolites, as well as many water-soluble contaminants partition to the aqueous phase. The solvent system needs to effectively extract the lipids of interest in an unbiased manner, without promoting the degradation of only specific lipids, and should not introduce contamination by other compounds (Xu et al., 2013). The two most commonly used lipid extraction methods are the ones described by Bligh and Dyer (1959) and Folch et al. (1957) more than 50 years ago involving the use of different ratios of chloroform, methanol and water, wherein the lipids partition to the lower chloroform phase. The methanol is added to the solvent system in order to disrupt the electrostatic forces and hydrogen bonding networks between the lipids and proteins. Other lipid extraction protocols which have been developed more recently use solvent mixtures such as butanol

(Hammad et al., 2010; Löfgren et al., 2012), methyl tert-butyl ether (MTBE) (Byeon et al., 2012; Graessler et al., 2009; Kosicek et al., 2010; Wiesner et al., 2009), and hexane (Hara & Radin, 1978). In general these alternative solvent systems do not show significant differences in the extraction efficiencies of the predominant lipid classes (Byeon et al., 2012; Iverson et al., 2001; Löfgren et al., 2012; Matyash et al., 2008). For example, the MTBE method has been reported to have very similar extraction efficiency to the Bligh and Dyer method in human plasma (Matyash et al., 2008). This method has become very popular for extracting sphingolipids in fluids (Hammad et al., 2010; Wiesner et al., 2009). Specific protocols for these methods, along with their associated disadvantages, are summarized in Table 12.2.

Following phase separation either the organic phase can be collected and transferred to a new tube, or the aqueous phase can be removed and discarded. If one is more interested in having an organic phase as clean of other contaminants as possible, but with the caveat of potentially losing some lipids, especially low abundance lipids, then the aqueous phase, typically the upper phase, can be removed and discarded, and the lower phase containing the lipids can be washed numerous times by multiple additions/removal of aqueous phase (Alecu, Tedeschi, et al., 2017). Alternatively, if the goal is to maximize the amount of extracted lipids, they can collect the lower organic phase containing the lipids, transfer to another tube, and repeatedly re-extract the aqueous phase by repeated additions of organic solvent, which are collected and pooled with the organic phase that has already been collected (Xu et al., 2013).

A variety of modifications to the chloroform-methanol extraction procedure have been made in order to maximize the ability to extract lipids of particular interest with high efficiency. Saunders and Horrocks used isopropanol-hexane (2:3 v/v) to extract lipids from bovine brain with a 12%–37% greater recovery of prostaglandins, compared with traditional chloroform-methanol extraction (Saunders & Horrocks, 1984). The Bligh and Dyer extraction has been modified by a number of groups to use acidified methanol (2% acetic acid) in order to increase the recovery of ether-linked glycerophospholipids, including platelet activating factors (PAFs) (Bonin et al., 2004; Liu et al., 2011; Weerheim et al., 2002; Whitehead et al., 2007). The acidified methanol is added directly to cells being extracted at the time of extraction; however, exposure of the sample to these acidic conditions should be minimized, as extended exposure could lead to the hydrolysis of glycerophospholipids, especially in aqueous solution (Ford et al., 1992; Kayganich & Murphy, 1992). For example, to extract lipids from adherent cells, the cells can be scraped off the plate directly into cold acidified methanol, followed by the addition of chloroform and 0.1M sodium acetate for a final ratio of 1:0.95:0.8 (methanol:chloroform:0.1M sodium acetate). Next, samples are vortexed and centrifuged at 800 × g for 2 minutes. The lower phase is then collected and the upper phase is back-extracted 3 more times by the addition of 2 mL of chloroform. The lower chloroform phase is collected each time and pooled with the other lower phases, and this is then evaporated under nitrogen gas. The dried

Table 12.2 Lipid extraction protocols.

Method	Time	Equipment	Advantages	Disadvantages
Ultracentrifugation, differential centrifugation 700, 2400, 10,000, and 100,000 × g	140–300 min	Ultracentrifugation equipment, rotors and tubes	Isolation from reasonable volumes (upto 1.5 L), low cost if access to UC equipment, sEV cargo, that is protein and RNA not affected	Equipment-dependent, laborious, time-consuming, non-EV contamination, low reproducibility, low yield, low purity, high centrifugation forces cause structural damage to sEVs, higher risk of contamination and low-throughput (only six samples fit in one UC spin)
Density gradient ultracentrifugation, sucrose or iodixanol density gradient after UC	280 min–2 days	Ultracentrifugation equipment, rotors and tubes. As well, sucrose and iodixanol density media	Pure sEVs population; No contamination with viral particles, high sEVs population purity and high separation efficiency after iodixanol UC	Equipment-dependence, low yield, laborious, time-consuming and low-scalability
Tangential flow filtration	110–150 min	Sterile hollow fiber polyethersulfone membrane filter with specific molecular weight cut-off	Pure sEVs population, high sEVs structural integrity, fast, higher reproducibility, better sterility, and large-scale stable production	Lack of method validation, risk of the sEVs being stuck in the membrane pores (filter-plugging), loss of sample, various factors affecting the filtration rate (e.g., temperature), and purified sEVs have small quantity of exosomal proteins

Lipid extraction protocols used for cell culture lipidomics.

lipids can be re-dissolved in 100% ethanol and stored at -80°C in amber glass vials under nitrogen to prevent lipid oxidation (Xu et al., 2013).

In the ecotoxicological study by Zhen et al. (2018), the authors used a modified chloroform/methanol extraction method where they kept both the aqueous and organic phases, wherein one fraction contained the hydrophilic metabolites and the other fraction contained lipids. Cells were homogenized in methanol using a tissue lyser, followed by the addition of 0.24 mL chloroform and further homogenization. The same volume of chloroform was added again to the resulting homogenate, followed by 0.22 mL of deionized water and further homogenization. To separate the phases the mixture was centrifuged at $3000 \times g$ for 15 minutes. The two phases were separated by pipetting and then dried down using a vacuum concentrator (Zhen et al., 2018). It should be noted here that it is important to perform optimization experiments with standards for all metabolites and lipids of interest in order to determine whether using such an extraction method for both hydrophilic metabolites and lipids is biasing the analysis towards specific species. Other such “double” biphasic extraction methods have recently been developed where both the aqueous and organic phases from the sample are used for MS analyses (Villaret-Cazadamont et al., 2020). Villaret-Cazadamont et al., compared the extraction efficiency of a classical water-soluble metabolite extraction with acetonitrile, methanol, and water acidified with formic acid and a lipid extraction with dichloromethane and methanol to a double extraction method. In the double extraction method, both hydrophilic metabolites and lipids were extracted using a quenching solution of cold methanol, acetonitrile, and milliQ water with 0.1% formic acid in a volume ratio of 2:2:1. Samples were centrifuged at $400 \times g$ and 2.5 mL of dichloromethane were added. The upper aqueous phase and lower organic phase were separated and dried down for metabolite and lipid analysis, respectively. Lipid internal standards were added to perform relative quantification of lipids and ^{13}C was added for absolute quantification of metabolites. The absolute concentration of the metabolites was found to be similar for the double extraction compared to the two separate extractions for the majority of polar metabolites. However, lower extraction efficiency was obtained for the amino acids methionine and phenylalanine and the following metabolites linked to energy metabolism: 6-phosphogluconate, pyridoxal-5-phosphate, cytidine diphosphate, α -ketoglutarate, guanosine diphosphate, uridine diphosphate acetyl-glucosamine and uridine 5'-monophosphate. The polarity of metabolites determines their distribution in aqueous and organic phases during liquid-liquid extractions (Houck et al., 2015), with polar metabolites preferentially partitioning into aqueous phases, hydrophobic metabolites migrating to organic phases (Humbert et al., 2014; Poole & Poole, 2010), and metabolites of intermediate polarity distributing between both phases. This can result in an underestimation of polar metabolites like the ones mentioned above when the concentration is assessed in the aqueous phase. For lipids, the classical extraction protocol showed significantly better extraction efficiency for triglycerides, phosphatidylethanolamines, phosphatidylcholines, and phosphatidylinositol, while the double

extraction protocol demonstrated higher extraction efficiency for ceramides and cholesterol. The relative distribution of different lipid molecular species in each lipid family was not affected by the extraction protocol used.

After extraction of lipids from cell samples, the most common analysis method for their separation and subsequent identification and quantification is liquid-chromatography MS (LC-MS) (Cajka & Fiehn, 2016; Fauland et al., 2011; Zhai & Reilly, 2002). Depending on the lipids of interest, either normal phase chromatography or reverse phase chromatography can be used. In normal-phase chromatography the column packing is polar and the mobile phase is nonpolar for example, hexane, ethyl acetate, etc., and lipids are separated based on their polar head groups. In reverse phase chromatography the column packing is hydrophobic (e.g., silica beads bonded to C18 chains) and the mobile phase is water (buffer) + water-miscible organic solvent (e.g., MeOH), and therefore lipids will separate based on the carbon chain length, double bonds, number of OH groups.

For the analysis of hydrophilic metabolites, gas chromatography-MS was widely used in earlier times and is still used today for the detection of organic acids and amino acids (Kvitvang et al., 2011; Milkovska-Stamenova et al., 2015; Tanaka et al., 1980). However, there are a number of drawbacks to using GC-MS analysis. It is not suitable to use for compounds that are unstable or have high boiling points, such as nucleotides and keto acids and often, complex derivatization methods are required, thereby restricting the range of hydrophilic metabolites which can be analyzed (Hu et al., 2020). Therefore, more and more analysis of hydrophilic metabolites is now also being performed by LC-MS.

The traditional reverse phase columns which are widely used for lipid analysis cannot retain hydrophilic metabolites, as the nonpolar stationary phase cannot form strong interactions with these metabolites. However, a variety of new strategies using different stationary phases and additives to mobile phases have been developed, allowing for broader and more in-depth analysis of these compounds. Wang et al., developed a 2D LC method using both a reverse phase C18 column and a T3 column to be able to separate short-chain, medium-chain, and long-chain Coenzyme A esters (Wang et al., 2017). The T3 column is composed of a trifunctional C18 alkyl phase at a low-ligand density, allowing the metabolites to more easily access the pore structure of the material and therefore greatly improving the retention of polar compounds. However, this interaction is still not able to retain small hydrophilic metabolites (Hu et al., 2020). Currently the column which can retain the largest number of hydrophilic metabolites is the hydrophilic interaction liquid chromatography, HILIC, which was initially proposed in 1990 by Andrew Alpert (1990). HILIC columns consist of polar silica gel (Hemström & Irgum, 2006) which can be modified with functional groups such diol, amide, aminopropyl, and zwitterionic compounds (Jandera & Janás, 2017; Periat et al., 2013), while the mobile phase is an organic solvent containing 2%–3% water. The metabolites partition into the aqueous component of the mobile phase which then forms a layer on the surface of the stationary phase, aiding retention (Jandera, 2008; Wikberg et al., 2011). Recent studies suggest that HILIC columns

modified with zwitterionic sulfobetaine allow for analysis of a wider array of metabolites as well as better chromatographic peak shape and resolution compared to underivatized HILIC columns (Sonnenberg et al., 2019). Even with these improvements in metabolite coverage and retention, poor peak shape and low sensitivity is still a problem in the analysis of phosphorylated metabolites and organic acids. This can be improved by reducing the chelation between these metabolites and metal ions by using medronic acid in the mobile phase (Hsiao et al., 2018).

It is clear that in order to analyze a broad range of both hydrophilic metabolites and hydrophobic lipids, even if a double extraction method is optimized, to be able to separate and identify the largest number of molecular species different chromatographic strategies need to be employed. Therefore, maximizing the amount of biological material analyzed in order to elucidate metabolic pathways and networks through the use of cell culture based models that closely reflect more complex *in vivo* models is necessary at least as a first step in identifying important nodes in pathways.

Analysis of metabolic processes including metabolic flux

Metabolic processes, pathways, and networks can be elucidated by tracking the metabolic flux of lipids and metabolites in cells of interest. This is critical for the understanding of dysregulation of these processes in pathological conditions and consequently the identification of therapeutic targets to correct these disturbances. Pulse-chase experiments can be used to track the fate of metabolites and lipids, as well as to discover novel metabolites. To do this, lipids tagged with a variety of functional groups such as fluorophores, different numbers of deuteriums or other natural isotopes such as ^{13}C , or alkyne lipids which can later be “clicked” with other functional groups can be used. As an example, the application of mammalian cell culture together with metabolic labeling approaches and differential metabolic analysis was used to discover a novel metabolic pathway for neurotoxic 1-deoxysphingolipids, thus elucidating the reason for increased levels of these lipids in pathological conditions like diabetic sensory polyneuropathy (Alecu, Tedeschi, et al., 2017). Pulse-chase experiments with deuterated 1-deoxysphingolipids led to the discovery of a novel metabolic pathway involving eight never-before measured lipid metabolites (Alecu, Othman, et al., 2017).

1-Deoxysphingolipids are cytotoxic atypical sphingolipids which are implicated in the pathology of the inherited neuropathy, hereditary sensory neuropathy type 1 (HSAN1) and diabetic sensory neuropathy. Due to their molecular structure it was always thought that they are “dead-end” metabolites with no metabolic exit point, thereby continuously accumulating to toxic levels. Alecu, Othman, et al. (2017) demonstrated that this was not the case by treating mouse embryonic fibroblasts with a pulse of d3-labeled 1-deoxysphinganine, an

upstream 1-deoxysphingolipid, for 2 hours. The “pulse” media was then replaced with fresh media without d3-deoxysphinganine for a chase period of 0, 1, 4, 8, 24, and 48 hours. Both the media and the cells were collected at these time points. It is important to note here that sample collection has to be performed as fast as possible to immediately block all enzymatic processes and prevent modifications of metabolites. Therefore, cells and media must be stored on ice/frozen as soon as they have been collected. The collected cell media should be lyophilized before lipid extraction. This step is necessary in order to maintain the appropriate aqueous/organic solvent proportion for lipid extraction without having to extract the media from one plate in multiple batches due to the large volume collected. After lyophilization, the media is re-suspended in the appropriate amount of aqueous phase for example, 200 µL. It is very important here to note the volume of the media lyophilized for normalization of lipid/metabolite levels. Once the lyophilized media is re-suspended, one can proceed with a metabolite or typical lipid extraction such as a Bligh and Dyer extraction described earlier.

The cells are harvested by trypsinization, followed by centrifugation, re-suspension of the pellet in PBS in order to wash off any media which could affect the levels of metabolites/lipids measured, and then cell counting. As with keeping track of the amount of media, cell counting is necessary for normalizing the amounts of metabolites/lipids. If cells are harvested by trypsinization plus addition of stop media, it is necessary to pellet the cells, remove trypsin + media, then wash/re-suspend in PBS and pellet again such that lipids in the cell media are not contributing to what is measured in the cells. Detachment of all cells from the plate should be visually confirmed before proceeding with the next steps. There are a variety of other options for cell harvesting, such as scraping adherent cells of the plate in PBS. However, one needs to consider the “harshness” of their harvesting method as it could lead to cell lysis, which would result in inaccurate cell counts and thereby less accurate final metabolite/lipid quantification. A “softer” harvesting technique would be the addition of 10 mM EDTA at 37°C for a total of 10 minutes ([Ziemanski et al., 2020](#)). A disadvantage of softer techniques is that all cells may not detach from the plate, thereby decreasing the amount of total lipids and metabolites. This could lead to the inability to measure or quantify these low abundance lipids if their quantities are below the lower limit of detection or quantification.

The next step following the collection of cells and media is the extraction of lipids and metabolites for analysis. [Alecu, Othman, et al. \(2017\)](#) chose to perform an acid-base hydrolysis lipid extraction on both the media and the harvested cells in order to remove the *N*-acyl fatty acid of the 1-deoxysphingolipids. This protocol would also remove the head group of endogenous lipids such as sphingolipids. The acid hydrolysis specifically breaks the *N*-acyl chain, whereas the base hydrolysis leads to a release of the *O*-linked phosphoester or carbohydrate head group. Five hundred microliters of methanol containing 200 pmol of internal standards (D7 labeled sphinganine and sphingosine, the 2 kinds of C18 sphingoid bases) were added to each sample (cell pellets or lyophilized medium re-suspended in

200 µL PBS). Internal standards are necessary in order to account for different extraction efficiencies and to monitor method-accuracy drifts of the MS method. Internal standards selected should not be endogenously present in the sample, and should be different from the labeled lipids/metabolites used for the metabolic flux experiments. If measuring levels of endogenous lipids, at least one internal standard should be used for each lipid subclass of interest.

For the acid hydrolysis, the sample was incubated with methanolic hydrochloric acid (1N HCl/10M water in methanol) for 12–15 hours at 65°C. Next, 40 µL KOH (5M) were added to neutralize the acid, followed by the addition of 4 volumes of 0.125M KOH in methanol for base hydrolysis, 1 volume of chloroform, then 0.5 mL of chloroform and 0.5 mL of alkaline water (Penno et al., 2010). The sample should be vortexed after each step. The aqueous and organic phases are then separated by centrifugation (12,000 × g, 5 minutes). The upper aqueous phase is aspirated, and the lower phase is washed 2 more times with alkaline water in order to remove any remaining contaminants from the organic phase (chloroform) containing the lipids. The chloroform phase is then evaporated under N₂, and the dried lipids should be stored at –80°C until analysis by LC-Ms.

In this case the authors chose to perform the acid-base hydrolysis because they were interested in the total sum of all the lipids with the deuterated 1-deoxysphingoid base backbone in order to be able to monitor the total amount of these lipids. The idea was that if these lipids are a metabolic dead-end, the total sum of exogenously added labeled 1-deoxysphingolipids should be constant. Without the acid-base hydrolysis, some of the low abundance lipids formed, such as a 1-deoxyceramide with an 18:1 *N*-acyl chain may be below the lower limit of detection/lower limit of quantification. If many of these low abundance species were missed, this would have made it impossible to monitor the total sum. Another instance where the user may choose to perform an acid-base hydrolysis lipid extraction would be studying host-pathogen interactions and determining which lipids are produced by the host and which are produced by the pathogen, wherein this protocol would allow for an in depth analysis of the sphingoid base backbone which could differ in length or branching in the pathogen vs the host and may elucidate a potential drug target in the lipid pathway which is host-specific (Lochnit et al., 1997).

There is also the option of simultaneous cell harvesting and extraction. Ziemanski et al. (2020) used the Folch extraction method, adding premixed chloroform–methanol (2:1 v/v, 3 mL), prechilled to –20°C, directly to the petri dish surface with adherent human meibomian gland epithelial cells, and the cells were then scraped off with a stainless steel scraper. Although the benefit of this strategy is that it is much higher throughput than first harvesting the cells, followed by extraction, there are a few factors which need to be considered if undertaking such a protocol. Firstly, the leaching of plastic upon addition of organic solvents to cell culture plates needs to be considered, which would interfere with the detection of metabolites/lipids by MS. Therefore, the cells would need to be

grown in glass dishes. Furthermore, cell counting would present an even bigger problem here than with only cell scraping, as chloroform has been shown to lead to rapid cell lysis ([Sapcariu et al., 2014](#); [Vellaichamy et al., 2010](#)).

Once lipids are extracted, they can be identified and quantified by LC-MS as described earlier. There is an optional step of derivatizing lipids by tagging them with specific functional groups before LC-MS analysis. This option is useful for lipids that cannot be efficiently ionized, which is necessary for detection by the MS, or if the lipids lack characteristic fragmentation patterns in tandem MS (MS/MS) analysis which would be necessary for identification of the lipid species ([Yang & Han, 2016](#)). In the current example Alecu et al., chose to derivatize the lipids with o-Phthalaldehyde (50 mg/mL in EtOH) in a 0.005:1:99 v/v/v with 3% boric acid and 2-mercaptoethanol in order to improve ionization, and consequently detection, of the lipids ([Alecu, Othman, et al., 2017](#)).

Alecu et al. wanted to monitor the time-dependent conversion of 1-deoxysphinganine to its downstream product 1-deoxysphingosine, and to determine whether the total amount of labeled 1-deoxysphingolipids remained constant with time ([Alecu, Othman, et al., 2017](#)). This would confirm that there was no further metabolic or catabolic processing of these lipids. The authors chose to collect the cell media in order to determine whether the cells were secreting any of the deuterated 1-deoxysphingolipids which elucidated whether the change in total labeled lipid levels was due to this and not to downstream metabolism. The user could also choose to collect cell media and extract metabolites/lipids in order to monitor changes in the secretome upon treatment with different compounds such as drugs or toxins, or upon mutations in genes coding for specific enzymes. If analyzing the levels of unlabeled, endogenously produced lipids secreted into the media, the use of synthetic serum-free media should be considered such that the amount of lipids already present in the media does not interfere with the identification/quantification of secreted lipids.

Alecu et al. found that the total levels of labeled 1-deoxysphingolipids decreased over time, while the amount of labeled lipids in the media was constant, indicating that further metabolic conversion of these lipids was occurring ([Alecu, Othman, et al., 2017](#)). In order to identify these unknown downstream 1-deoxysphingolipid metabolites, the authors used differential analysis of mass spectral data as well as visual analysis of the total ion chromatogram to identify new spectra appearing over the time course of the pulse-chase experiments. For this, the software Sieve from ThermoFisher was used, wherein multiple replicates of two conditions were compared in order to determine the appearance of new lipid molecular species. The basic workflow for this has been previously described ([Snyder et al., 2013](#)). A variety of different filters were applied to perform this analysis, including the m/z range and retention time expected for potential lipids of interest, odd-numbered m/z ratios which would indicate that the compound potentially carried the 3 deuterium label, as well as the criterium that the potential molecular formulas generated based on the m/z identified should contain one sulfur coming from the o-phthalaldehyde solution used for

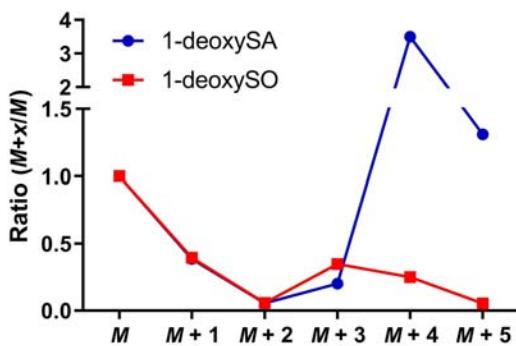
derivatization, since all lipids of interest would have been derivatized. The total ion chromatogram generated for each sample was also visually analyzed scan by scan using the same criteria in order to identify any novel peaks. Sieve software has now been updated to Compound Discoverer which has further features such as pathway analysis.

Another labeling approach was used to elucidate a different aspect of 1-deoxysphingolipid metabolism. In order to determine whether the newly identified downstream metabolites were also formed from de novo synthesized 1-deoxysphingolipids, and not just from those added exogenously, labeled substrates needed for 1-deoxysphingolipid biosynthesis were used (Alecu, 2016). Cells were treated with deuterated versions of both substrates necessary for the synthesis of 1-deoxysphingolipids, methyl-d3-palmitic acid and d4-alanine. Methyl-d3-palmitic acid was given as a 1:1 molar complex with fatty acid-free BSA in order to prevent the fatty acid from sticking to the cell culture dish. For this experiment, 1-deoxysphinganine and 1-deoxysphingosine with mass offsets for +6, +5, +4, +3, +2, +1, as well as the unlabeled mass M, were monitored in order to be able to analyze 1-deoxysphingolipids formed from conjugation of d3-palmitic acid (+3), d4-alanine (+3, as one deuterium is lost upon conjugation), +6 (when both labeled substrates were conjugated), as well as the natural isotopologues arising from this labeling. A similar kind of deuterium exchange assay was performed with 11,11,12,12-d4 palmitic acid in order to determine whether the position of the double bond which is inserted upon the conversion of 1-deoxysphinganine to 1-deoxysphingosine is C14, which is the double bond position in canonical sphingolipids (Alecu, 2016). 1-Deoxysphingolipids with mass offsets of +5, +4, +3, +2, +1 and the unlabeled mass M were analyzed in order to monitor for double bond insertions. Upon being conjugated with alanine, the product formed would be 13,13,14,14-d4-labeled 1-deoxysphinganine. If the double bond was inserted at C14, this would entail the loss of a deuterium at this position, resulting in a d3-labeled 1-deoxysphingosine (meaning a mass offset of +3), compared to the d4-labeled 1-deoxysphinganine (mass offset of +4), which was indeed the case Fig. 12.1.

As illustrated, metabolic labeling and flux experiments in cells can be used to analyze many different aspects of metabolism ranging from the structure of compounds, the substrates used to form metabolites and lipids, as well as novel metabolic pathways. In the bigger picture, this could help elucidate the reason for metabolic dysregulations in pathological conditions, or under different treatment conditions.

Methods and protocols for isolation and metabolomics of small extracellular vesicles from cell culture supernatants

EVs, including small extracellular vesicles (sEVs) or exosomes are naturally secreted in culture by almost all eukaryotic cells except mature red blood cells

**FIGURE 12.1**

The majority of de novo formed 1-deoxysphingosine does not have the double bond inserted at C4. The lines on the graph represent labeled upstream 1-deoxysphinganine (blue) and downstream 1-deoxysphingosine (red) after incorporation of 11,11,12,12-d4 palmitic acid. The highest relative amount of de novo formed 1-deoxysphinganine has a +4 label coming from the d4-palmitic acid. However, the highest relative amount of 1-deoxysphingosine has a +3 label, which indicates that one of the deuteriums was lost from the C12 of the palmitic acid (which would become C14 upon condensation with alanine in the de novo formation of the upstream 1-deoxysphinganine). This would only occur upon insertion of the double bond at C14, and not at C4. *1-deoxySA*, 1-deoxysphinganine; *1-deoxySO*, 1-deoxysphingosine.

under both physiological and pathological conditions (Pegtel & Gould, 2019). sEVs are a subpopulation of membrane-bound, 30–150 nm in diameter vesicles that are formed in the multivesicular bodies, which fuse with the plasma membrane to release sEVs in the extracellular milieu. These nanovesicles harbor a variety of bioactive cargo of cellular components such as nucleic acids (microRNA, mRNA, circular RNA and noncoding RNA), proteins (cytokines, chemokines, receptors and ligands), lipids and metabolites that represent distinct “molecular signatures” of their parental cells (Kalluri & LeBleu, 2020). Therefore, sEVs provide a snapshot of crucial molecular information about the health of its parental cell. Recently, sEVs have emerged as important intercellular communication vehicles exchanging crucial information not only between neighboring cells but also distant organs (Théry et al., 2009). They are stamped with “unique addresses” that dictate their cellular and organ specificity. Increasing evidence demonstrates that sEVs can selectively transfer their cargoes into recipient cells and contribute to the modulation of a wide range of biological processes, including pro-survival, antiinflammatory, antitumorigenic, regenerative and regulation of immune responses. Over the past decade, sEVs have gained clinical utility and are being harnessed for their intrinsic therapeutic properties and also being explored as nanodevices for drug delivery and biomarkers of disease (Andaloussi et al., 2013; Fais et al., 2016).

Under in vitro conditions, cells produce a heterogeneous population of EVs, such as sEVs or exosomes (30–150 nm), microvesicles or ectosomes (100–1000 nm) and apoptotic bodies (1–5 µm) which accumulate in cell culture supernatants (conditioned media) (Raposo & Stoorvogel, 2013). These three types of EVs not only vary in size but also differ in their biogenesis, cargo content and regulation of cellular mechanisms. Recently, it has been demonstrated that cells also release distinct subpopulations of sEVs with different biophysical properties as well as proteomic and RNA repertoires, further emphasizing the heterogeneity of EVs (Willms et al., 2016). Therefore, it is crucial that prior to any metabolomics or lipidomics studies, specific populations of EVs are purified from the biological sample using differential isolation methods. With recent advances in science and technology, many different techniques exploiting the unique physicochemical and biochemical characteristics of EVs, such as size, shape, mass, buoyant density, and molecules on EV surface have been developed for the isolation and purification of sEVs (Sidhom et al., 2020). Here we will describe two protocols that capitalize on the EV properties, such as size and buoyant density for the isolation of sEVs: (1) ultracentrifugation (UC) method, which employs differential centrifugation steps and still remains the gold-standard method of sEV isolation; (2) tangential flow filtration (TFF), an emerging new technique that is coupled to membrane filtration and flow to obtain clinical grade sEVs preparations with high yield, purity and integrity. Both methods are capable of processing large volumes of cell culture medium, for example, for UC several hundreds of liters and for TFF up to several thousand liters. The first part of this section describes the most common protocols used to isolate sEVs, and the second part describes different methods for characterizing and analyzing the purity of the isolated sEVs preparation.

Cell culture for isolation of small extracellular vesicles

Most of the mammalian cells are cultured in media supplemented with 10%–20% fetal bovine serum (FBS), a rich source of nutrients and growth factors, which is important for cell survival. Small EVs are found in almost all biological fluids, including serum (Lässer et al., 2011) and FBS contains many different types of bovine EVs. Since bovine EVs in culture media are bioactive and their presence can influence experimental results (Kornilov et al., 2018; Shelke et al., 2014), they are often removed from FBS prior to addition to the culture media (Théry et al., 2006). Thus far, no standardized protocol for EV-depletion of FBS exists and different laboratories use different depletion protocols. Briefly, cell culture media containing FBS is centrifuged for at least 2 hours at 100,000 × g to remove sEVs. The supernatant is filtered using a 0.22 µm vacuum bottle top filter. Some cell types can be grown in the absence of serum and culture medium without FBS can be used. Several commercial serum alternatives are available, however caution should be exercised in selecting these alternatives for metabolomic

studies as these may contain polyethylene glycol which can interfere with NMR spectra. Cells are usually grown in T-175 flasks at 70%–80% confluence (a total of $20\text{--}40 \times 10^6$ cells/sample or equivalent to 1–2 mg protein/sample) in the presence of serum and then media is changed to EV-depleted media or serum-free media for 24 hours. The user is recommended to include a culture media alone control incubated for 24 hours in the absence of cells. Following incubations, cell culture supernatant (conditioned medium) is collected and subjected to UC or TFF to isolate sEVs as described below.

Isolation of small extracellular vesicles using ultracentrifugation

The traditional EV isolation methods employing UC, namely differential UC and density gradient UC utilize EVs properties, such as size, mass and buoyant density for the separation and purification of sEVs (Romano et al., 2020). Table 12.3 provides the comparison of the inherent advantages and limitations of each method which are important to keep in mind while designing an experiment.

Differential ultracentrifugation

The common method of sEVs isolation is UC, which still remains the gold-standard technique for EV isolation. In brief, conditioned media is sequentially subjected to increasing centrifugal forces and duration to pellet cells at $700 \times g$, microvesicles at $2400 \times g$ and sEVs at $100,000 \times g$, as described (Čuperlović-Culf et al., 2020; Kuo & Jia, 2017; Romano et al., 2020; Théry et al., 2006; Witwer et al., 2013). The workflow for the purification of sEVs using differential centrifugation is presented in Fig. 12.2 (top panel). All centrifugations should be performed at 4°C . The low speed spins ($<10,000 \times g$) gradually remove particles with a high buoyant density such as cells, cell debris, apoptotic bodies, and proteins aggregates, while the high speed spin ($100,000 \times g$) sediments small EVs. The sEVs pellet is washed once with 1 mL sterile $1 \times$ PBS and the $100,000 \times g$ step is repeated to obtain an sEVs pellet that can be further purified as below.

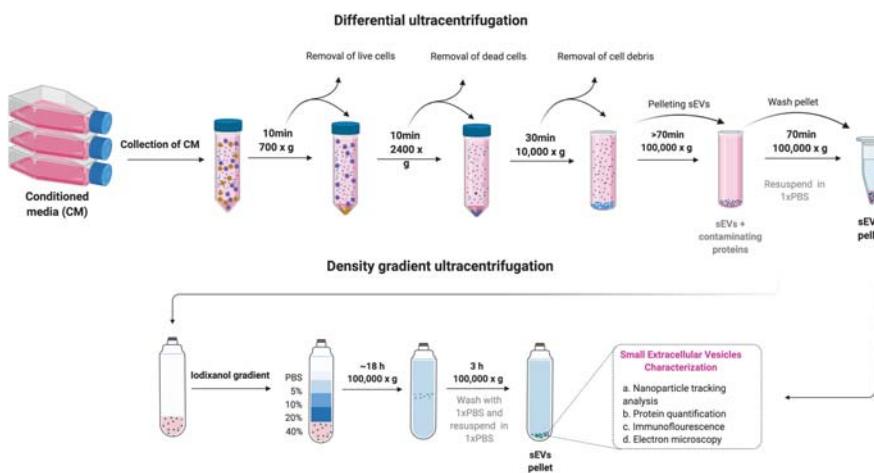
Density gradient ultracentrifugation

Although the differential UC EV isolation method provides a reasonably pure sEVs population, it can also coisolate contaminants, such as aggregated proteins and nucleic acids. Therefore, an extra step can be added using density gradient UC to improve the purity of sEVs population (Abramowicz et al., 2016; Zhang et al., 2014). Several gradient medias are available, however sucrose cushions and iodixanol (OptiPrep) gradients coupled with differential UC is most commonly used to isolate different EV populations based on their buoyant densities and

Table 12.3 Standard strategies for isolation of small extracellular vesicles (sEVs).

Method	Time	Equipment	Advantages	Disadvantages
Ultracentrifugation, differential centrifugation 700, 2400, 10,000, and 100,000 × g	140–300 min	Ultracentrifugation equipment, rotors and tubes	Isolation from reasonable volumes (upto 1.5 L), low cost if access to UC equipment, sEV cargo, that is protein and RNA not affected	Equipment-dependent, laborious, time-consuming, non-EV contamination, low reproducibility, low yield, low purity, high centrifugation forces cause structural damage to sEVs, higher risk of contamination and low-throughput (only six samples fit in one UC spin)
Density gradient ultracentrifugation, sucrose or iodixanol density gradient after UC	280 min–2 days	Ultracentrifugation equipment, rotors and tubes. As well, sucrose and iodixanol density media	Pure sEVs population; No contamination with viral particles, high sEVs population purity and high separation efficiency after iodixanol UC	Equipment-dependence, low yield, laborious, time-consuming and low-scalability
Tangential flow filtration	110–150 min	Sterile hollow fiber polyethersulfone membrane filter with specific molecular weight cut-off	Pure sEVs population, high sEVs structural integrity, fast, higher reproducibility, better sterility, and large-scale stable production	Lack of method validation, risk of the sEVs being stuck in the membrane pores (filter-plugging), loss of sample, various factors affecting the filtration rate (e.g., temperature), and purified sEVs have small quantity of exosomal proteins

Key advantages and disadvantages of the standard methods for the purification of sEVs are summarized.

**FIGURE 12.2**

Schematic representation of common strategies for the isolation and purification of small extracellular vesicles. Flow-chart for the isolation and purification of sEVs based on differential ultracentrifugation (Čuperlović-Culf et al., 2020) and density gradient ultracentrifugation. sEVs indicate small extracellular vesicles.

Illustration created in BioRender (www.BioRender.com).

mass (Araújo et al., 2008; Graham, 1999). The workflow for the purification of sEVs using density gradient UC is presented in Fig. 12.2 (bottom panel).

Isolation of small extracellular vesicles using tangential flow filtration

Despite the wide use of differential UC for sEVs isolation, this method has major limitations (Table 12.3). Therefore, improved techniques that increase sEVs yield, integrity, scalability and reproducibility have been adapted for sEVs isolation (Furi et al., 2017; Konoshenko et al., 2018). TFF is an emerging ultrafiltration technique that couples membrane filtration and fluid flow for efficient isolation and concentration of sEVs from large volumes of biological fluids (Fig. 12.3). In brief, clarified conditioned media (after 700 and 2400 × g centrifugation steps to remove cells and cell debris) is concentrated and filtered at the same time using a polyethersulfone hollow fiber filter with varying range of pore sizes or molecular weight cutoff cartridges (10, 50 and 500 kDa) (Lee et al., 2020). The clarified media is pumped using a peristaltic pump system. Multiple rounds of filtration leads to the isolation and concentration of specific sEVs populations. For instance, filtered culture media is concentrated ~10 fold and in the final step culture media is exchanged with 1x PBS and further concentrated ~5–10 fold.

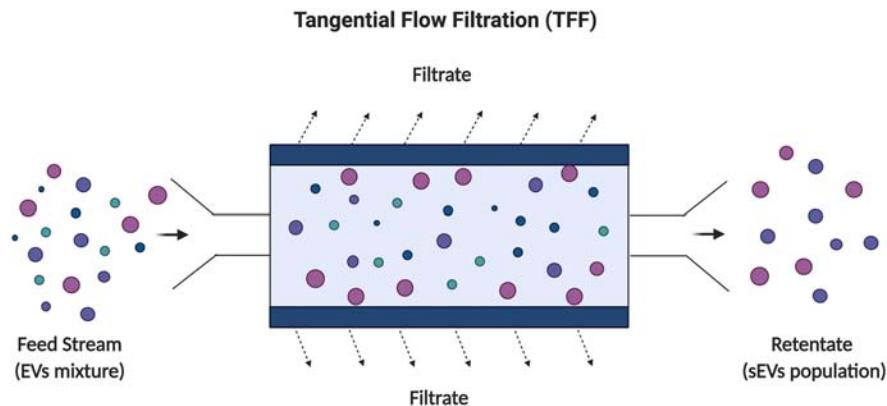


FIGURE 12.3

Tangential flow filtration. Schematic representation of the principle of tangential flow filtration for isolation and purification of small extracellular vesicles.

Illustration created in BioRender (www.BioRender.com).

Characterization of small extracellular vesicles

A large number of methods have been developed to assess the size, concentration, integrity and purify of sEVs. The most common techniques include nanoparticle tracking analysis (NTA), western blotting, scanning electron microscopy, transmission electron microscopy, cryo-electron microscopy, flow cytometry and fluorescence-activated cell sorting (Gurunathan et al., 2019; Noreldin et al., 2021). sEVs originating from a variety of different cell types share common structural and functional characteristics, such as exosomal proteins, tetraspanins (CD9, CD63, CD81), TSG101, Alix and flotillin-1, which can be detected using western blotting. The NTA method (NanoSight and ZetaView) allows real-time visualization and analysis of EVs based on the rate of Brownian motion of individual nanoparticles (EVs) in solution and their ability to scatter light (Bachurski et al., 2019). Hence, NTA allows the measurement of concentration and size distribution of EVs.

Metabolite extractions from cells and small extracellular vesicles

Extraction of polar metabolites using acetonitrile/water method. Metabolomics analysis of the content of sEVs can be performed using MS methods presented above or using approaches that were previously developed for cell and tissue analysis with NMR-based metabolic profiling (Beckonert et al., 2007; Belle et al., 2002; Lin et al., 2007). Since extraction parameters can influence the detection and quantification of metabolites, it is important to consistently adhere to the same extraction protocols to obtain optimum results and ensure experimental reproducibility. The protocols described here can be used as a guide for the

extraction of polar metabolites and the combined extraction of polar and lipophilic metabolites from cells and their sEVs harvested from the conditioned medium of the same culture.

Polar metabolites are extracted using acetonitrile from cells, media and sEVs as previously described (Čuperlović-Culf et al., 2020). Remove cell culture dishes from the 37°C incubator and place them on ice slurry to slow down the metabolism. Collect the conditioned medium which is then subjected to sEV isolation by UC or TFF as described above. An aliquot of medium can be saved at -80°C for the analysis of exometabolomics. All subsequent steps are carried out under ice-cold conditions and making sure that cells and media never approach room temperature. Harvest cells in ice-cold 5 mL of 1 × PBS (Ca^{2+} and Mg^{2+} free) by gentle scraping, transfer to 15 mL falcon tubes and centrifuge at $300 \times g$ for 5 minutes at 4°C. Place tubes on ice slurry and aspirate the 1 × PBS without disturbing the pellets. Wash the pellets once again with ice-cold 5 mL of 1 × PBS to remove any residual medium. Hold the cell pellets on ice slurry for 5 minutes to keep metabolic activity low. Subsequently, resuspend pellets in 1 mL of extraction solvent [50% acetonitrile/50% water (vol/vol) mixture, prechilled at -20°C overnight], which further quenches metabolism and lyses cells. Mix the suspension thoroughly by vortexing and transfer to eppendorf tubes. Centrifuge at $12,000 \times g$ for 10 minutes at 4°C. After centrifugation, the suspension separates into supernatants (contains polar metabolites) and a pellet of cellular proteins, lipids and debris. Transfer the supernatants to fresh eppendorf tubes and evaporate the solvents from the samples under a stream of nitrogen gas or using a SpeedVac concentrator. Alternatively, samples can also be freeze-dried/lyophilized overnight. A similar protocol should be followed for the extraction of intra-exosomal metabolites by adding 200 µL ice-cold acetonitrile/water mixture to the $100,000 \times g$ pellet. The dried samples can be stored at -80°C until NMR analysis.

Extraction of combined polar and lipophilic metabolites using methanol/chloroform/water method: Harvest cells as described above and resuspend pellets in ice-cold mixture containing 2 parts methanol/0.8 parts water to quench metabolic activity (Vuckovic, 2012). Vortex the suspension thoroughly to achieve good mixing. Place samples on ice slurry and sonicate 3–5 times for 1 seconds each time (Folch et al., 1957). Transfer suspension to glass tubes, add 1 part chloroform to a total solution of methanol/chloroform/water (2:1:0.8) and vortex again. Add 1 part chloroform and 1 part water for a final solution of methanol/chloroform/water (2:2:1.8) and vortex again. Hold the samples on an ice slurry for 15 minutes or at 4°C overnight. Centrifuge at $1000 \times g$ for 15 minutes at 4°C. After centrifugation, the suspension separates into three phases: an upper methanol/water phase (contains polar metabolites), an interface of protein/cellular debris (protein disk) and a lower chloroform phase (contains lipophilic metabolites). The protein disk can be saved for proteomics analysis. Transfer the upper and lower phases into fresh glass tubes and evaporate the solvents from the samples under a stream of nitrogen gas. The dried samples can be stored at -80°C. The upper phase is used for metabolomics and the lower phase is used for lipidomics studies. A similar

protocol should be followed for the extraction of intra-exosomal metabolites and lipids, adjusting the volumes accordingly.

Sample preparation and analysis with nuclear magnetic resonance spectroscopy

Samples are prepared in NMR buffer (50 mM sodium phosphate buffer, pH 7.4, in deuterium oxide, 0.1% 4,4-dimethyl-4-silapentane-1-sulfonic acid and 0.5 mM sodium azide) and an internal standard solution (NMR grade). The standard solution is added at 10% of the total sample. For dried samples, reconstitute in 160 µL NMR buffer containing 16 µL standard and for liquid samples, mix 100 µL media with 60 µL deuterium oxide and 16 µL standard. Vortex samples to mix thoroughly. Using gel loading tips, load approximately 10 µL of sample into 3 mm NMR tubes and proceed to NMR analysis.

Although a number of different nuclei can be measured in NMR metabolomics, including ^{13}C , ^{15}N and ^{31}P , ^1H NMR spectroscopy measurements are the most significant for general metabolomics profiling and thus far the only approach used for the analysis of sEVs. One dimensional (1D) ^1H (proton) NMR spectra with water suppression sequence (NOESY 1D) provides a good combination of speed, excellent water suppression and good lineshape for quantification. NMR experimental techniques and possible pulse sequences that are generally used in metabolomics have been previously reviewed ([Čuperlović-Culf et al., 2010](#); [Ranjan & Sinha, 2019](#)) and all the methods for NMR metabolomics described in Chapter 5, Nuclear Magnetic Resonance in Metabolomics, can be applied in this case as well.

Cell culture metabolomics and lipidomics data analysis

Data analysis method selection, application and interpretation depends on the level of background knowledge, metabolomics coverage and sample set size as well as specific goals of the study. Although the majority of analytical methods can be utilized for knowledge discovery, presentation or model development from metabolomics and lipidomics data regardless of the biological source, analysis of cell culture data provides some unique opportunities including a possibility for analysis of cells, organelles and media for the same system in a highly controlled environment, possibly with isotopic labeling and flux analysis. Metabolomics data analysis is described in Chapters 8–11. Here we will only show examples of analysis either specifically applied to cell culture metabolomics or lipidomics or methods that can provide some unique benefits to the cell culture application. This includes:

1. utilization of cell modeling for design and optimization of cell cultures including optimization of growth conditions or productivity in cell bioreactors;

2. utilization of cell culture metabolomics for the development of Artificial Intelligence (AI) methods for optimal design and prediction of behavior of cell and gene therapy modalities including cells and exosomes as therapy carriers.

Both of these groups of applications include analysis of metabolomics data, determination of major metabolic pathways and networks, simulation of cell metabolism and linking these models and data within machine learning models of cell metabolism and cell growth conditions and these steps will be described in some detail below.

Cell culture metabolomics and cell modeling for the design and optimization of cell culture applications

Application of cells for the production of biologics, vaccines or for bioprocessing has transformed therapy fabrication and provided avenues for synthetic biology utilization. A bioprocess is an extremely complex interplay of numerous factors that requires regulation and optimization while still lacking complete understanding. Traditionally around 10 biochemical molecules are monitored including oxygen, CO₂, glucose as well as some toxic by-products, for example, lactate, however these are insufficient to track cell metabolism, growth and productivity. Metabolomics and lipidomics can provide additional quantification of tens to hundreds of metabolites in media or cell extracts providing information about the cells' oxidative state, cell growth or death, metabolic needs or toxins, active pathways, etc. Accordingly, metabolomics can provide a way for finding the perfect media for each application; identify clonal instabilities early in the process and help provide continuous process monitoring for optimization of growth. As an example, the most popular mammalian cells used in bioprocessing are Chinese hamster ovary (CHO) cells and human embryonic kidney 293 cells (HEK293). Although CHO cells remain the most often used, they can result in nonhuman posttranslational protein. Thus, HEK293 is becoming a predominant cell line for expression of recombinant proteins and biologics providing appropriate human cell glycosylation and protein folding appropriate for *in vivo* use ([Dietmair et al., 2012](#); [Petiot et al., 2015](#)) making this cell line of particular relevance in biotechnology. HEK293 cells are explored as a possible way to provide production of difficult-to-express proteins as well as next-generation biologics including bispecific antibodies and weaponized antibodies. HEK293 cells grow easily in suspension serum-free culture, reproduce rapidly, and produce high levels of protein. In this context, metabolomics can be used for testing of gene editing methods, cell productivity and health as well as analysis and optimization of HEK293 growth for biomanufacturing. Metabolomics analysis of HEK293 cells have shown major influence of media on cell metabolism and the measured cell secretome ([Daskalaki et al., 2018](#)) necessitating optimization of media for specific application. Metabolomics combined with models of specific cell lines used in the bioprocessing can be directly used in this process however

modeling of metabolism in cell lines requires determination of significant pathways or metabolic interaction network, followed by the development of mechanistic, machine learning of hybrid models.

Determination of major metabolic pathways or network from metabolomics or fluxomics data

Cell function, growth or productivity is largely regulated through the metabolism and metabolites. Metabolic processes are driven through allosteric regulation, posttranslational modifications, inter-compartmental material balance, and signaling control (O'Brien et al., 2020). Mapping metabolomics data on its own or combined with other omics data onto cellular pathways or determination of data-driven interaction networks can be used to establish significant pathways and regulation mechanisms under different conditions. Several highly advanced, free-ware pathway analysis tools (Table 12.4) can be used to map metabolomics data and determine statistical significance of the representation of metabolic pathways by a selected metabolite set with or without concentration information.

Mapping of the omics data on the metabolic pathways provides a static representation of relevant processes. Although these methods do not provide any predictive power, some of these tools provide a sophisticated way to determine major metabolic differences between conditions or cell types. As an example Lilikoi analytical method (AlAkwa et al., 2018) provides metabolite ID matching, feature selection through information gain calculation, ML classification modeling and pathway deregulation score determination. It also transforms metabolite-sample matrix into pathway-sample matrix providing in this way personalized pathway mapping. In another example, MetExplore (Cottret et al., 2018), provides statistical information about the organism-specific metabolic network coverage and gives interactive visualization of metabolomics data on the whole metabolic network, selection of pathways or specific reactions. All pathway mapping methods, by their design only give mapping onto the known pathways included in one of many databases and in this way do not allow determination of novel interactions between biological molecules.

Network analysis in cell culture metabolomics

Cell culture metabolomics provides data that can be utilized for novel mechanistic insight about biological processes under different conditions, stimuli or phenotypes and this move away from the known processes mapping into a hypothesis generation can be a major advancement for omics and systems biology (Rosato et al., 2018). A powerful approach for the data-driven metabolic mechanism analysis can be accomplished through development of interaction, that is correlation, statistical or clustering networks. A biological network in this context is a graphic representation of features (metabolites or lipids)—nodes and their

Table 12.4 Pathway and Network analysis tools in cell cultures metabolomics.

Method	Application	Availability and references
Lilikoi	Group of applications in R for: mapping of metabolites to pathways, dimension transformation to personalized pathway-based profiles using pathway deregulation scores;, feature selection module, and classification and prediction module, which offers various machine learning classification algorithms.	https://github.com/lanagarmire/lilikoi (AIKwaa et al., 2018)
MetPA	Web-based tool for the analysis and visualization of metabolomic data within the biological context of metabolic pathways combining several advanced pathway enrichment analysis procedures with the analysis of pathway topological characteristics to help identify the most relevant metabolic pathways involved in a given metabolomic study. The results are presented in a network visualization system.	http://metpa.metabolomics.ca
IMPALA	Pathway overrepresentation and enrichment analysis with expression and/or metabolite data. Both gene and metabolite information can be either a list for overrepresentation analysis or values in different conditions for enrichment analysis.	http://impala.molgen.mpg.de/
MBRole	Overrepresentation, enrichment, analysis of categorical annotations for user provided sets of compounds. Provided categorical annotations correspond to biological and chemical information available in a number of public databases and software. Provided is also information about metabolite-protein interaction.	http://csbg.cnb.csic.es/mbrrole2/index.php
MetExplore	Web-based collection of interactive tools for metabolic network curation, network exploration and omics data analysis. In particular, it is possible to curate and annotate metabolic networks in a collaborative environment with the contextualization of metabolic elements in the network and the calculation of overrepresentation statistics.	https://metexplore.toulouse.inrae.fr/index.html/

Pathway and Network determination from metabolomics data can be applied to cell cultures metabolomics.

associations—edges. Network can be represented as an adjacency or connectivity matrix of interactions describing the strength of the relationships between any two nodes. In the cell metabolomics context network can represent interactions within cells, extracellularly or across the cell membrane and can be based on topology, stoichiometry, directionality or kinetics of the metabolome. Several reviews have recently described network analysis methods including applications to metabolomics in some detail (Perez De Souza et al., 2020; Rosato et al., 2018; Toubiana et al., 2019) and any of these methods can be applied to cell culture metabolomics data. Several examples that have shown utility in cell culture analysis are outlined in Table 12.5.

Metabolic networks based on correlation analysis can indicate rapid equilibrium between metabolites or presence of conserved chemical groups. They are particularly useful for analysis of changes between different conditions, treatments or phenotypes. Correlation network requires establishment of threshold of relevant correlation with several authors showing that correlation of 0.6 and P value of 0.01 provide good threshold levels indicating lower bound for weak correlations in metabolomics data (Camacho et al., 2005; Ghini et al., 2015; Saccenti et al., 2016). It is important to point out however that lack of strong correlation does not necessarily mean lack of proximity between metabolites in the metabolic pathways and that strong correlation can be observed for metabolites that are metabolically distant. Therefore, even in the context of cell culture analysis, correlation networks may not be sufficient for reverse engineering of metabolic pathways (Rosato et al., 2018). The Debiased Sparse Partial Correlation algorithm (DSPC) was developed as an attempt to regularize correlation methods. DSPC uses a desparsified graphical lasso modeling procedure and assumes that the number of real connections in the network is much smaller than what is determined from correlation analysis. DSPC is implemented in MetScape within Cytoscape (Basu et al., 2017; Perez De Souza et al., 2020).

A number of methods that were originally developed for gene and protein network determination and analysis are finding their place in metabolomics. As an example weighted gene correlation network analysis (WCGNA) can be used to determine modules, clusters of highly correlated features and finding “module eigengene” a representative feature summarizing module profile or an intramodular features that relates modules to one another and to sample trait (Langfelder & Horvath, 2008). WCGNA provides dissimilarity profiles through analysis of topological overlap matrix (TOM) that makes the network less sensitive to distant connections or connections that are missing due to noise. TOM is related to correlation between metabolite pairs. TOM and thus WCGNA assumes scale-free topology which does not apply to all metabolic networks (Broido & Clauset, 2019; Rosato et al., 2018). Use of WCGNA in metabolomics, often with low coverage, requires some modifications to the original method with the detailed tutorial provided for this application by Pei et al. (2017).

A number of other methods have also been developed for gene network determination with few examples of their use in metabolomics, albeit thus-far

Table 12.5 Interaction network methods in cell cultured metabolomics.

Network approach/availability	Method in brief	Advantages	Disadvantages
Correlation or relevance: Pearson, Spearman or Distance correlation methods Available in Python, R, etc.	Pearson—measure of the linear association between variables; Spearman—nonparametric measure of rank correlation; Distance—dependence of two random vectors of possibly different dimension	Inferred network is a good description of the physiological state of the system	Cannot be used to reverse engineer metabolic pathways
Weighted gene correlation network analysis—WCGNA Available as R package.	Calculates dissimilarity profiles based on topological overlap matrix based on pair correlations between metabolites.	Less sensitive to spurious connection and missing connections due to noise	Assumes power-law probability distribution for correlation and scale-free network design which does not apply to all networks
Context likelihood of relatedness—CLR Available in minet Bioconductor package	Uses Mutual Information (MI) to calculate similarity between pairs of variables and infers direct interactions by accounting for the local context for each interaction.	CLR does not require threshold as it prunes spurious interactions from network by its design.	High variability for smaller sample sets (<100 samples)
Algorithm for the reconstruction of accurate cellular networks—ARACNE Available in minet Bioconductor package	MI is calculated for each pair of nodes and the interactions are pruned by considering each triplet of edges and removing the weakest edge as it is considered as an indirect interaction.	Good at reconstruction of the backbone of association network	Produces very sparse network missing many significant associations
PCLRC available at: http://download.systemsbiology.nl/	Combination of CLR and iteratively sampling of dataset wherein each iteration a subset is chosen and a weighted adjacency matrix is determined using correlation calculations. Final network is calculated from the average of iterations.	Effective at discriminating between direct and indirect correlations	Requires thresholding by user (usually value of 0.9 is imposed) and larger number of samples

Examples of methods for determination of interaction network from data with some of their advantages and disadvantages in the analysis of metabolomics data.

rarely in cell culture metabolomics such as Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) and Probabilistic Context Likelihood of Relatedness Algorithm (PCLR) (Suarez-Diez & Saccenti, 2015). Based on detailed analysis of the network reconstruction performance Suarez-Diez and Saccenti (Suarez-Diez & Saccenti, 2015) have shown that as many as 100–400 samples may be necessary to obtain a stable network estimate making utilization of these methods in cell culture applications challenging. With the development of methods that can profile large numbers of cells (e.g., single cell metabolomics) (as described in Chapter 15: MALDI–Mass Spectrometry Imaging: The Metabolomics Visualization) or high throughput metabolomics or lipidomics, application of network design methods will likely become more relevant.

Mechanistic modeling for cell culture optimization, design, and information gathering

Pathway mapping as well as network determinations provide static representation of the metabolic interactions in the system without possibility to predict behavior under changing conditions. Metabolic models can provide a way to explore metabolic complexity and systematically investigate significant cellular properties for a variety of cell culture applications. Importantly, models can be used to infer processes in cells that were not directly measured. Longitudinal cell culture metabolomics and lipidomics can be utilized for the development and optimization of in silico models of cellular metabolism either aiming to explain observed effects, determine possibly significant targets or to provide predictive models for cell or media design. One of the most significant applications of cell metabolomics and metabolism modeling is the optimization of cell and gene therapies particularly through the optimization of bioreactor production (Selvarasu et al., 2012) or for the design of predictable, optimized cells.

Numerous publications have discussed in great detail metabolism modeling approaches based on either kinetic and ordinary differential equation models or genome scale metabolic models and the readers are referred to those (Almquist et al., 2014; Covert, 2017). Cell culture metabolomics provides a unique opportunity for time-course analysis of metabolites in intra or extracellular space as well as cellular or extracellular organelles possibly augmented with the utilization of isotopic labeling. Access to time-course information combined with the investigation of flux for isotopic labels can be used for the development of predictive models and simulations of metabolic pathways as well as genome scale networks. Feeding cells with isotopically labeled nutrients, measuring the isotopic labeling of extra and intracellular metabolites, and computationally inferring flux through the Metabolic Flux Analysis is the most direct approach for determining metabolic flux through metabolic network and pathways on a whole-cell level (Sauer, 2006; Wiechert, 2002). In this way cell culture metabolomics can

help in both bottomup or forward modeling as well as topdown or inverse approach by providing either data for the determination of model reaction constants, for model parameterization or for determination of network structures for the model.

Mathematical modeling of cell metabolism is an essential approach for gaining system-level understanding of cell behavior and development of predictions of cellular behavior, ultimately providing methods for the design of cells of desired properties or cell growth conditions. Mathematical modeling including deterministic kinetic modeling and stochastic and statistical modeling, have been widely used in the application of cell cultures (Richelle et al., 2020). In the mechanistic, mathematical modeling system functions and properties are described as the result of the interaction of the system elements within the cell and with the environment. Thus, mechanistic models can predict behavior of cellular systems or (metabolic) processes when elements of the model, their properties or interactions change (Stalidzans et al., 2020). Different mechanistic modeling approaches have been extensively utilized for the description and interpretation of cell culture metabolomics results with several methods and related freely available tools and some examples of their application listed in Table 12.6.

Changes in the flow through metabolic networks are a reflection of genetic, epigenetic and environmental factors. Measurement and the analysis of the network can be done through the analysis of the flow of a label from isotopically labeled precursors into metabolites (see above). Metabolic flux and concentration do not necessarily correlate as metabolic concentration increase can come from either increased production flux or decreased consumption flux. Thus, metabolite levels and fluxes provide complementary information. Fluxes can not be directly measured, but can be inferred from measurement of isotope tracers (Jang et al., 2018) with some examples of metabolic flux experiments described above. In addition to MS application for flux analysis, NMR spectroscopy can be used for a highly sensitive site-specific label quantification. In cell culture applications isotopic label tracing with NMR can be used for analysis of extracellular, intracellular or organel specific flow. Time-series measurements of metabolome are essential for the development and validation of dynamic models of metabolism (Judge et al., 2019; Sefer et al., 2016). In a typical cell culture metabolomics setting described above, information about the dynamic metabolome change would require significant resources, and sample material. Time-series sampling has to provide sufficient number of replicates, ensure sufficient experiment duration, and the time resolution. Sampling introduces extraction biases and the confounding of biological and analytical variance (Sitnikov et al., 2016; Tabatabaei Anaraki et al., 2018).

Different types of metabolic modeling have been presented for number of significant cell lines including for example for a model of a generic human cell (Brunk et al., 2018; Robinson et al., 2020) as well as number of specific cell types, including HEK293 cells (Quek et al., 2014), CHO cells (Lund et al., 2017;

Table 12.6 Freeware methodologies for mechanistic modeling.

Method	Software application	Examples of some application in cell culture metabolomics
Bayesian modeling	GRASP	Methionine cycle modeling using approximate Bayesian computation
Logical modeling	CellNetOptimizer (http://www.cellnopt.org) GINsim (http://ginsim.org)	Combination of cell line proteomics and metabolomics data logic mechanistic model modeling to explain heterogeneous drug response in cellular cholesterol regulation
Dynamic modeling through Ordinary differential equations	COPASI (Hoops et al., 2006) CellDesigner (Matsuoka et al., 2014) VCell	Many examples of COPASI's use in biotechnology cell modeling are reviewed in; recent example of hybrid cybernetic modeling that combines dynamic modeling between different metabolic states for CHO cells
Stochastic modeling	COPASI (Hoops et al., 2006) StochKit MaBOSs (http://maboss.curie.fr)	Theoretical foundation to study metabolism in conjunction with stochastic enzyme expression has been presented showing metabolic heterogeneity resulting from enzyme level stochasticity
Stoichiometric modeling	COBRA (Heirendt et al., 2019) CobraPy Raven 2.0 (Wang et al., 2018) Merlin	Genome-scale stoichiometric reconstructions and computational models of mammalian metabolism particularly for CHO cells coupled to protein secretion
Agent based modeling	ARCADE	Extensive review of agent based methods for cancer cell modeling

Several freeware methodologies for mechanistic modeling.

Robinson et al., 2020), iPSCs (Chandrasekaran et al., 2017; Shen et al., 2019), cancer cell lines (Ghaffari et al., 2015; Yizhak et al., 2015). Development of large combined models that can take advantage of advanced knowledge of some pathways and ability to simulate large networks continues to be an active area of research (Hameri et al., 2019; Jamshidi & Palsson, 2010; Opdam et al., 2017). Extreme high-throughput metabolomics and lipidomics of cell cultures, with an increasing coverage over time or flux provides information for parameter optimization. At the same time this data can be used for the development of data-driven, machine learning models and hybrid mechanistic-machine learning models with major potential in the design of optimal cells and cell environments.

Machine learning and hybrid models and artificial intelligence for cell design

Current mechanistic models, although increasingly detailed, still can not provide complete simulation and explanation of cellular processes possibly due to the self-regulatory nature of metabolic networks, posttranslational regulation and the topological organization of metabolism (Zampieri et al., 2019; Zelezniak et al., 2018) all making relationship between enzyme function and metabolites highly dynamic and multifactorial and therefore suboptimally covered with current mechanistic models.

In the development of safe, specific, and affordable gene and cell therapies the ability to design appropriate modalities with predictable behavior in different environments is of particular importance. Machine learning has been extensively used for the analysis of high throughput omics data as well as images of cell cultures with some examples presented above. AI systems that can describe and predict behavior of biological networks of cells will allow more accurate, faster and less expensive innovations in life sciences while at the same time ensuring predictable outcomes. In particular, the full potential for safe and efficient utilization of gene editing and live cell therapies requires an ability for controlled design of these modalities with simulations that allow testing and optimization under different conditions in both production and utilization. However, the current inability to predict the behavior of biological systems including predicting the phenotype from genotype and the inability to extrapolate large-scale or *in vivo* outcomes from small-scale, *ex vivo* experiments severely hampers progress of cell therapy development. The lack of sufficient quantity and quality of data hampers the direct use of machine learning for the development of predictive models of cellular systems. Simultaneously, the lack of biological knowledge as well as the extreme complexity of the system makes development of whole cell system mechanistic models impossible at this point.

Machine learning are algorithms that perform pattern formation and classification and establish rules and statistical structures from data without any explicit instructions. Machine learning is widely used in the analysis of cell culture data including analysis of “omics” data as well as metabolomics and lipidomics (Cuperovic-Culf, 2018; Pomyen et al., 2020). In cell culture applications there is an increasing abundance of data, both metabolomics/lipidomics as well as other types of omics and data for gene knock-out screens of protein inhibition and this resources can be now used to develop data-driven models without any mechanistic assumptions or inclusion of only very well defined theoretical knowledge. Several recent examples show the power of these approaches when linked to cell culture metabolomics (Zelezniak et al., 2018). Zelezniak et al. (2018) have used machine learning modeling of proteomics data to predict metabolite concentrations. The predicted concentrations correlated strongly with measured metabolomics data in yeast cell analysis. Different data transformation techniques and a

large number of different machine learning algorithms were tested and the quality of obtained models was ranked based on the correlation with the measured metabolite concentrations. This analysis has shown that machine learning approaches can provide some information about multifactorial relationships in metabolic networks without information from mechanistic models. In another example, Costello and Martin (2018) used longitudinal proteomics and metabolomics cell culture data to develop machine learning predictive metabolism models. In this approach authors used tree-based pipeline optimization tool to combine, through genetic algorithms, 18 different feature selection algorithms and 11 different machine learning regressors in order to find function f which satisfies: $\text{argmin} \sum \sum ||f(m^i[t], p^i[t]) - m^i(t)||^2$ where $m^i[t]$ and $p^i[t]$ are, respectively, metabolite and protein concentrations at time t . Metabolome and proteomics concentration measurements over time were the input variables into the machine learning model and $m^i(t)$ is metabolite time derivative (rate of change) is the output of the model.

Machine learning generally performs poorly in prognosis particularly when trained using sparse data. However, these methods can be combined with mechanistic models in order to provide a combination of knowledge-based and data-driven systems for modeling and design. Examples of metabolomics applications that were combining constraint-based metabolism modeling analysis and machine learning have been recently outlined (Zampieri et al., 2019). Method comparison has shown the possibility to link results from mechanistic models with further analysis with machine learning. Machine learning can also be enhanced with the integration of knowledge in the form of driving equations, constraints or boundary conditions in order to reduce the model search space improving handling of sparse, noisy data. Mechanistic models can benefit from machine learning in creating surrogate models, identify networks, system dynamics and parameters from data (Cuperlovic-Culf, 2018; Peng et al., 2020). Metabolomics and lipidomics investigation of cell culture in different applications provides uniquely rich data for creation of better cell models and AI tools for design of cell environment for optimal utilization as well as design of cells with optimal behavior. Cell culture metabolomics also provides a possibility for measurement of metabolite concentrations in whole cells, organelles, extracellular vesicles and media in static or flux mode. All this data can be combined in order to develop predictive cellular models that can be further linked with other information about the cells including other omics measurements or image analysis data and finally implemented in the cell design systems.

Large datasets primarily resulting from single-cell RNASeq analysis are driving development of a number of new AI methods for prediction and modeling of biological data and many of these approaches can be adapted to metabolomics and metabolism modeling in cell culture analysis. Recently published scGen (Lotfollahi et al., 2019) method combines variational autoencoders (consisting of an encoder and a decoder and able to generate new data points) and latent space vector arithmetics for modeling cell behavior from single cell gene expression

data. In another example Graph Convolutional Neural networks for Genes were developed for inferring gene-gene interactions from high throughput spatial gene expression data (Yuan & Bar-Joseph, 2020). This method, through its use of graph structure, can utilize both the gene expression values (in the original use or metabolite concentration in metabolomics) encoded in each node and relationship between cells expressing these genes or metabolites in order to predict extracellular interactions. Modeling methods used in cell culture design have to provide information about the functional outcome as well as mechanisms leading to the outcomes in order to aid in drug or gene editing making “black box” deep learning models in-appropriate for this application. Use of a “visible” neural network was attempted as a method for an interpretable NN for simulation of a basic eukaryotic cell (Ma et al., 2018). The resulting simulation DCell (<http://d-cell.ucsd.edu/>) provides an excellent simulation of cell growth and allows in silico investigation of the molecular mechanisms underlying genotype to phenotype relationship.

Theory inspired machine learning methods seek casualties by integrating prior knowledge and data. Wide range of information available for cell cultures as well as mechanistic models for a number of pathways as well as genome scale networks can be used to narrow the search space for machine learning models. At the same time, machine learning can be used to reduce the number of dynamic variables and unknown parameters present in mechanistic models while providing uncertainty quantification. Extensive consideration of the power and opportunities of theory inspired machine learning is provided in Alber et al. (2019). Some examples of biological knowledge inspired machine learning models include knowledge-primed neural networks (KPNN) (Fortelny & Bock, 2020) and simulation-based kernel ML (SimKernML) (Deist et al., 2019). In the KPNN method, a biological network is used to define a graph where each node corresponds to a protein or a gene, and each edge corresponds to a regulatory relationship obtained from biological databases or literature. Network that is designed in this way can be trained, that is optimized, with a much smaller dataset then is needed for example for artificial neural networks as fewer free parameters need to be optimized. Additionally, every node and every edge within a KPNN have a corresponding biological interpretation. SimKernML uses mechanistic simulations of biological processes to build machine learning kernel (e.g., support vector machine) and this improves the downstream machine learning performance for small training dataset.

Novel, faster and more accurate theory inspired and “white box” metabolism modeling methods, developed and trained using metabolomics and lipidomics measurements in cell cultures can be utilized for the design of cell growth conditions, explanation of different cell culture test results or design of predictable and safe cell therapies. Addition of novel methods for single cell metabolomics, in-cell analysis, 3D cell cultures and increased metabolite coverage will provide invaluable data for further development of improved applications of cell cultures.

References

- Abramowicz, A., Widlak, P., & Pietrowska, M. (2016). Proteomic analysis of exosomal cargo: The challenge of high purity vesicle isolation. *Molecular Biosystems*, 12(5), 1407–1419.
- AlAkwa, F. M., Yunits, B., Huang, S., Alhajaji, H., & Garmire, L. X. (2018). Lilikoi: An R package for personalized pathway-based classification modeling using metabolomics data. *GigaScience*, 7(12), giy136.
- Alber, M., Tepole, A. B., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., & Petzold, L. (2019). Integrating machine learning and multi-scale modeling—Perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digital Medicine*, 2(1), 1–11.
- Alecu, I. (2016). *Elucidating novel metabolic and trafficking pathways of 1-deoxysphingolipids*. University of Zurich.
- Alecu, I., Othman, A., Penno, A., Saied, E. M., Arenz, C., von Eckardstein, A., & Hornemann, T. (2017). Cytotoxic 1-deoxysphingolipids are metabolized by a cytochrome P450-dependent pathway. *Journal of Lipid Research*, 58(1), 60–71.
- Alecu, I., Tedeschi, A., Behler, N., Wunderling, K., Lamberz, C., Lauterbach, M. R., Gaebler, A., Ernst, D., Van Veldhoven, P. P., & Al-Amoudi, A. (2017). Localization of 1-deoxysphingolipids to mitochondria induces mitochondrial dysfunction. *Journal of Lipid Research*, 58(1), 42–59.
- Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., & Jirstrand, M. (2014). Kinetic models in industrial biotechnology—improving cell factory performance. *Metabolic Engineering*, 24, 38–60.
- Alpert, A. J. (1990). Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography A*, 499, 177–196.
- Andaloussi, S. E., Mäger, I., Breakefield, X. O., & Wood, M. J. (2013). Extracellular vesicles: Biology and emerging therapeutic opportunities. *Nature Reviews. Drug Discovery*, 12(5), 347–357.
- Araújo, M., Hube, L. A., & Stasyk, T. (2008). Isolation of endocytic organelles by density gradient centrifugation. *2D PAGE: Sample Preparation and Fractionation*, 317–331.
- Bachurski, D., Schuldner, M., Nguyen, P.-H., Malz, A., Reiners, K. S., Grenzi, P. C., Babatz, F., Schauss, A. C., Hansen, H. P., & Hallek, M. (2019). Extracellular vesicle measurements with nanoparticle tracking analysis—An accuracy and repeatability comparison between NanoSight NS300 and ZetaView. *Journal of Extracellular Vesicles*, 8 (1), 1596016.
- Basu, S., Duren, W., Evans, C. R., Burant, C. F., Michailidis, G., & Karnovsky, A. (2017). Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics (Oxford, England)*, 33(10), 1545–1553.
- Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11), 2692–2703.
- Belle, J. L., Harris, N., Williams, S., & Bhakoo, K. (2002). A comparison of cell and tissue extraction techniques using high-resolution ^1H -NMR spectroscopy. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance in Vivo*, 15(1), 37–44.

- Bligh, E. G., & Dyer, W. J. (1959). A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology*, 37(8), 911–917.
- Bonin, F., Ryan, S. D., Migahed, L., Mo, F., Lallier, J., Franks, D. J., Arai, H., & Bennett, S. A. (2004). Anti-apoptotic actions of the platelet-activating factor acetylhydrolase I α 2 catalytic subunit. *Journal of Biological Chemistry*, 279(50), 52425–52436.
- Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1), 1–10.
- Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Gonzalez, G. A. P., & Aurich, M. K. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, 36 (3), 272.
- Byeon, S. K., Lee, J. Y., & Moon, M. H. (2012). Optimized extraction of phospholipids and lysophospholipids for nanoflow liquid chromatography-electrospray ionization-tandem mass spectrometry. *Analyst*, 137(2), 451–458.
- Cajka, T., & Fiehn, O. (2016). Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical Chemistry*, 88(1), 524–545.
- Camacho, D., De La Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics: Official Journal of the Metabolomic Society*, 1(1), 53–63.
- Campos, A. I., & Zampieri, M. (2019). Metabolomics-driven exploration of the chemical drug space to predict combination antimicrobial therapies. *Molecular Cell*, 74(6), 1291–1303.
- Chandrasekaran, S., Zhang, J., Sun, Z., Zhang, L., Ross, C. A., Huang, Y.-C., Asara, J. M., Li, H., Daley, G. Q., & Collins, J. J. (2017). Comprehensive mapping of pluripotent stem cell metabolism using dynamic genome-scale network modeling. *Cell Reports*, 21 (10), 2965–2977.
- Costello, Z., & Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Systems Biology and Applications*, 4(1), 1–14.
- Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., & Poupin, N. (2018). MetExplore: Collaborative edition and exploration of metabolic networks. *Nucleic Acids Research*, 46(W1), W495–W502.
- Covert, M. W. (2017). *Fundamentals of systems biology: From synthetic circuits to whole-cell models*. CRC Press.
- Cuperovic-Culf, M. (2018). Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites*, 8(1), 4.
- Čuperlović-Culf, M., Barnett, D. A., Culf, A. S., & Chute, I. (2010). Cell culture metabolomics: Applications and future directions. *Drug Discovery Today*, 15(15–16), 610–621.
- Čuperlović-Culf, M., Khiu, N. H., Surendra, A., Hewitt, M., Charlebois, C., & Sandhu, J. K. (2020). Analysis and simulation of glioblastoma cell lines-derived extracellular vesicles metabolome. *Metabolites*, 10(3), 88.
- Daskalaki, E., Pillon, N. J., Krook, A., Wheelock, C. E., & Checa, A. (2018). The influence of culture media upon observed cell secretome metabolite profiles: The balance between cell viability and data interpretability. *Analytica Chimica Acta*, 1037, 338–350.
- Deist, T. M., Patti, A., Wang, Z., Krane, D., Sorenson, T., & Craft, D. (2019). Simulation-assisted machine learning. *Bioinformatics (Oxford, England)*, 35(20), 4072–4080.

- Dietmair, S., Hodson, M. P., Quek, L.-E., Timmins, N. E., Gray, P., & Nielsen, L. K. (2012). A multi-omics analysis of recombinant protein production in Hek293 cells.
- Dunn, W. B., Bailey, N. J., & Johnson, H. E. (2005). Measuring the metabolome: Current analytical technologies. *Analyst*, 130(5), 606–625.
- Fais, S., O'Driscoll, L., Borras, F. E., Buzas, E., Camussi, G., Cappello, F., Carvalho, J., Da Silva, A. C., Del Portillo, H., & El Andaloussi, S. (2016). Evidence-based clinical use of nanoscale extracellular vesicles in nanomedicine.
- Fauland, A., Köfeler, H., Trötzmüller, M., Knopf, A., Hartler, J., Eberl, A., Chitraju, C., Lankmayr, E., & Spener, F. (2011). A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. *Journal of Lipid Research*, 52(12), 2314–2322.
- Flasch, M., Bueschl, C., Woelflingseder, L., Schwartz-Zimmermann, H. E., Adam, G., Schuhmacher, R., Marko, D., & Warth, B. (2020). Stable isotope-assisted metabolomics for deciphering xenobiotic metabolism in mammalian cell culture. *ACS Chemical Biology*, 15(4), 970–981.
- Folch, J., Lees, M., & Stanley, G. S. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *Journal of Biological Chemistry*, 226(1), 497–509.
- Ford, D., Rosenbloom, K., & Gross, R. (1992). The primary determinant of rabbit myocardial ethanolamine phosphotransferase substrate selectivity is the covalent nature of the sn-1 aliphatic group of diradyl glycerol acceptors. *Journal of Biological Chemistry*, 267 (16), 11222–11228.
- Fortelny, N., & Bock, C. (2020). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biology*, 21(1), 1–36.
- Furi, I., Momen-Heravi, F., & Szabo, G. (2017). Extracellular vesicle isolation: Present and future. *Annals of Translational Medicine*, 5(12).
- Ghaffari, P., Mardinoglu, A., & Nielsen, J. (2015). Cancer metabolism: A modeling perspective. *Frontiers in Physiology*, 6, 382.
- Ghini, V., Saccenti, E., Tenori, L., Assfalg, M., & Luchinat, C. (2015). Allostasis and resilience of the human individual metabolic phenotype. *Journal of Proteome Research*, 14 (7), 2951–2962.
- Graessler, J., Schwudke, D., Schwarz, P., Herzog, R., Shevchenko, A., & Bornstein, S. (2009). *Lipidomic profiling reveals a deficiency of ether lipids in blood plasma of men with hypertension*. *Diabetologia* (Vol. 52, pp. S426–S427). New York: Springer.
- Graham, J. M. (1999). Purification of a crude mitochondrial fraction by density-gradient centrifugation. *Current Protocols in Cell Biology*, 4(1), 3–4.
- Gurunathan, S., Kang, M.-H., Jeyaraj, M., Qasim, M., & Kim, J.-H. (2019). Review of the isolation, characterization, biological function, and multifarious therapeutic approaches of exosomes. *Cells*, 8(4), 307.
- Hameri, T., Fengos, G., Ataman, M., Miskovic, L., & Hatzimanikatis, V. (2019). Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations. *Metabolic Engineering*, 52, 29–41.
- Hammad, S. M., Pierce, J. S., Soodavar, F., Smith, K. J., Al Gadban, M. M., Rembiesa, B., Klein, R. L., Hannun, Y. A., Bielawski, J., & Bielawska, A. (2010). Blood sphingolipidomics in healthy humans: Impact of sample collection methodology. *Journal of Lipid Research*, 51(10), 3074–3087.

- Hara, A., & Radin, N. S. (1978). Lipid extraction of tissues with a low-toxicity solvent. *Analytical Biochemistry*, 90(1), 420–426.
- Hemström, P., & Irgum, K. (2006). Hydrophilic interaction chromatography. *Journal of Separation Science*, 29(12), 1784–1821.
- Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., Magnusdóttir, S., Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Clancy, C. M., Modamio, J., Sauls, J. T., ... Fleming, R. M. T. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nature Protocols*, 14(3), 639–702.
- Houck, M. M., Siegel, J. A., Houck, M. M., & Siegel, J. A. (2015). *Chapter 6—Separation methods* (pp. 121–151). Academic Press. Available from <https://doi.org/10.1016/B978-0-12-800037-3.00006-6>.
- Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., & Kummer, U. (2006). COPASI—a COMplex PAthway SIMulator. *Bioinformatics*, 22(24), 3067–3074.
- Hsiao, J. J., Potter, O. G., Chu, T.-W., & Yin, H. (2018). Improved LC/MS methods for the analysis of metal-sensitive analytes using medronic acid as a mobile phase additive. *Analytical Chemistry*, 90(15), 9457–9464.
- Hu, Q., Tang, H., & Wang, Y. (2020). Challenges in analysis of hydrophilic metabolites using chromatography coupled with mass spectrometry. *Journal of Analysis and Testing*, 1–23.
- Humbert, L., Hoizey, G., & Lhermitte, M. (2014). *Drugs involved in drug-facilitated crimes (DFC): Analytical aspects: 1—blood and urine. Toxicological aspects of drug-facilitated crimes* (pp. 159–180). Elsevier.
- Iverson, S. J., Lang, S. L., & Cooper, M. H. (2001). Comparison of the Bligh and Dyer and Folch methods for total lipid determination in a broad range of marine tissue. *Lipids*, 36(11), 1283–1287.
- Jamshidi, N., & Palsson, B. Ø. (2010). Mass action stoichiometric simulation models: Incorporating kinetics and regulation into stoichiometric models. *Biophysical Journal*, 98(2), 175–185.
- Jandera, P. (2008). Stationary phases for hydrophilic interaction chromatography, their characterization and implementation into multidimensional chromatography concepts. *Journal of Separation Science*, 31(9), 1421–1437.
- Jandera, P., & Janás, P. (2017). Recent advances in stationary phases and understanding of retention in hydrophilic interaction chromatography. A review. *Analytica Chimica Acta*, 967, 12–32.
- Jang, C., Chen, L., & Rabinowitz, J. D. (2018). Metabolomics and isotope tracing. *Cell*, 173(4), 822–837.
- Judge, M. T., Wu, Y., Tayyari, F., Hattori, A., Glushka, J., Ito, T., Arnold, J., & Edison, A. S. (2019). Continuous *in vivo* metabolism by NMR. *Frontiers in Molecular Biosciences*, 6, 26.
- Kalluri, R., & LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. *Science (New York, N.Y.)*, 367(6478).
- Kayganich, K. A., & Murphy, R. C. (1992). Fast atom bombardment tandem mass spectrometric identification of diacyl, alkylacyl, and alk-1-enylacyl molecular species of glycerophosphoethanolamine in human polymorphonuclear leukocytes. *Analytical Chemistry*, 64(23), 2965–2971.

- Konoshenko, M. Y., Lekchnov, E. A., Vlassov, A. V., & Laktionov, P. P. (2018). Isolation of extracellular vesicles: General methodologies and latest trends. *BioMed Research International*, 2018.
- Kornilov, R., Puhka, M., Mannerström, B., Hiidenmaa, H., Peltoniemi, H., Siljander, P., Seppänen-Kaijansinkko, R., & Kaur, S. (2018). Efficient ultrafiltration-based protocol to deplete extracellular vesicles from fetal bovine serum. *Journal of Extracellular Vesicles*, 7(1), 1422674.
- Kosicek, M., Kirsch, S., Bene, R., Trkanjec, Z., Titlic, M., Bindila, L., Peter-Katalinic, J., & Hecimovic, S. (2010). Nano-HPLC–MS analysis of phospholipids in cerebrospinal fluid of Alzheimer's disease patients—A pilot study. *Analytical and Bioanalytical Chemistry*, 398(7), 2929–2937.
- Kuo, W. P., & Jia, S. (2017). *Extracellular vesicles: Methods and protocols*. Springer.
- Kvitvang, H. F., Andreassen, T., Adam, T., Villas-Bôas, S. G., & Bruheim, P. (2011). Highly sensitive GC/MS/MS method for quantitation of amino and nonamino organic acids. *Analytical Chemistry*, 83(7), 2705–2711.
- Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), 1–13.
- Lässer, C., Alikhani, V. S., Ekström, K., Eldh, M., Paredes, P. T., Bossios, A., Sjöstrand, M., Gabrielsson, S., Lötvall, J., & Valadi, H. (2011). Human saliva, plasma and breast milk exosomes contain RNA: Uptake by macrophages. *Journal of Translational Medicine*, 9(1), 1–8.
- Lee, J. H., Ha, D. H., Go, H., Youn, J., Kim, H., Jin, R. C., Miller, R. B., Kim, D., Cho, B. S., & Yi, Y. W. (2020). Reproducible large-scale isolation of exosomes from adipose tissue-derived mesenchymal stem/stromal cells and their application in acute kidney injury. *International Journal of Molecular Sciences*, 21 (13), 4774.
- Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., Souza, A., Pierce, K., Keskula, P., & Hernandez, D. (2019). The landscape of cancer cell line metabolism. *Nature Medicine*, 25(5), 850–860.
- Lin, C. Y., Wu, H., Tjeerdema, R. S., & Viant, M. R. (2007). Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*, 3(1), 55–67.
- Liu, S. T., Sharon-Friling, R., Ivanova, P., Milne, S. B., Myers, D. S., Rabinowitz, J. D., Brown, H. A., & Shenk, T. (2011). Synaptic vesicle-like lipidome of human cytomegalovirus virions reveals a role for SNARE machinery in virion egress. *Proceedings of the National Academy of Sciences*, 108(31), 12869–12874.
- Lochnit, G., Dennis, R. D., Zahringer, U., & Geyer, R. (1997). Structural analysis of neutral glycosphingolipids from *Ascaris suum* adults (Nematoda: Ascaridida). *Glycoconjugate Journal*, 14(3), 389–399.
- Löfgren, L., Ståhlman, M., Forsberg, G.-B., Saarinen, S., Nilsson, R., & Hansson, G. I. (2012). The BUME method: A novel automated chloroform-free 96-well total lipid extraction method for blood plasma. *Journal of Lipid Research*, 53(8), 1690–1700.
- Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8), 715–721.
- Lund, A. M., Kaas, C. S., Brandl, J., Pedersen, L. E., Kildegaard, H. F., Kristensen, C., & Andersen, M. R. (2017). Network reconstruction of the mouse secretory pathway applied on CHO cell transcriptome data. *BMC Systems Biology*, 11(1), 1–17.

- Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4), 290.
- Matsuoka, Y., Funahashi, A., Ghosh, S., & Kitano, H. (2014). Modeling, simulation using Cell Designer. *Methods in Molecular Biology*, 1164, 121–145.
- Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A., & Schwudke, D. (2008). Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research*, 49(5), 1137–1146.
- Milkovska-Stamenova, S., Schmidt, R., Frolov, A., & Birkemeyer, C. (2015). GC-MS method for the quantitation of carbohydrate intermediates in glycation systems. *Journal of Agricultural and Food Chemistry*, 63(25), 5911–5919.
- Muschet, C., Möller, G., Prehn, C., de Angelis, M. H., Adamski, J., & Tokarz, J. (2016). Removing the bottlenecks of cell culture metabolomics: Fast normalization procedure, correlation of metabolites to cell number, and impact of the cell harvesting method. *Metabolomics: Official Journal of the Metabolomic Society*, 12(10), 1–12.
- Noreldin, A. E., Khafaga, A. F., & Barakat, R. A. (2021). *Isolation and characterization of extracellular vesicles: Classical and modern approaches. Role of exosomes in biological communication systems* (pp. 1–25). Springer.
- O'Brien, C. M., Mulukutla, B. C., Mashek, D. G., & Hu, W.-S. (2020). Regulation of metabolic homeostasis in cell culture bioprocesses. *Trends in Biotechnology*.
- Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C., & Lewis, N. E. (2017). A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell Systems*, 4(3), 318–329.
- Pegtel, D. M., & Gould, S. J. (2019). Exosomes. *Annual Review of Biochemistry*, 88, 487–514.
- Pei, G., Chen, L., & Zhang, W. (2017). WGCNA application to proteomic and metabolic data analysis. *Methods in Enzymology*, 585, 135–158.
- Peng, G. C., Alber, M., Tepole, A. B., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., & Perdikaris, P. (2020). Multiscale modeling meets machine learning: What can we learn? *Archives of Computational Methods in Engineering*, 1–21.
- Penno, A., Reilly, M. M., Houlden, H., Laurá, M., Rentsch, K., Niederkofler, V., Stoeckli, E. T., Nicholson, G., Eichler, F., & Brown, R. H., Jr (2010). Hereditary sensory neuropathy type 1 is caused by the accumulation of two neurotoxic sphingolipids. *Journal of Biological Chemistry*, 285(15), 11178–11187.
- Perez De Souza, L., Alseekh, S., Brotman, Y., & Fernie, A. R. (2020). Network-based strategies in metabolomics data analysis and interpretation: From molecular networking to biological interpretation. *Expert Review of Proteomics*, 17(4), 243–255.
- Periat, A., Debrus, B., Rudaz, S., & Guillarme, D. (2013). Screening of the most relevant parameters for method development in ultra-high performance hydrophilic interaction chromatography. *Journal of Chromatography A*, 1282, 72–83.
- Petiot, E., Cuperlovic-Culf, M., Shen, C. F., & Kamen, A. (2015). Influence of HEK293 metabolism on the production of viral vectors and vaccine. *Vaccine*, 33(44), 5974–5981.
- Pinu, F. R., & Villas-Boas, S. G. (2017). Extracellular microbial metabolomics: The state of the art. *Metabolites*, 7(3), 43.
- Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal*.

- Poole, C. F., & Poole, S. K. (2010). Extraction of organic compounds with room temperature ionic liquids. *Journal of Chromatography A*, 1217(16), 2268–2286.
- Quek, L.-E., Dietmair, S., Hanscho, M., Martínez, V. S., Borth, N., & Nielsen, L. K. (2014). Reducing Recon 2 for steady-state flux analysis of HEK cell culture. *Journal of Biotechnology*, 184, 172–178.
- Ranjan, R., & Sinha, N. (2019). Nuclear magnetic resonance (NMR)-based metabolomics for cancer research. *NMR in Biomedicine*, 32(10), e3916.
- Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: Exosomes, microvesicles, and friends. *Journal of Cell Biology*, 200(4), 373–383.
- Richelle, A., David, B., Demaegd, D., Dewerchin, M., Kinet, R., Morreale, A., Portela, R., Zune, Q., & von Stosch, M. (2020). Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry: A process systems biology engineering perspective. *NPJ Systems Biology and Applications*, 6(1), 1–5.
- Robinson, J. L., Kocabas, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., & Billa, V. (2020). An atlas of human metabolism. *Science Signaling*, 13(624).
- Romano, E., Netti, P. A., & Torino, E. (2020). Exosomes in gliomas: Biogenesis, isolation, and preliminary applications in nanomedicine. *Pharmaceutics*, 13(10), 319.
- Rosato, A., Tenori, L., Cascante, M., Carulla, P. R. D. A., Dos Santos, V. A. M., & Saccenti, E. (2018). From correlation to causation: Analysis of metabolomics data using systems biology approaches. *Metabolomics: Official Journal of the Metabolomic Society*, 14(4), 1–20.
- Sacenti, E., Menichetti, G., Ghini, V., Remondini, D., Tenori, L., & Luchinat, C. (2016). Entropy-based network representation of the individual metabolic phenotype. *Journal of Proteome Research*, 15(9), 3298–3307.
- Sapcariu, S. C., Kanashova, T., Weindl, D., Ghelfi, J., Dittmar, G., & Hiller, K. (2014). Simultaneous extraction of proteins and metabolites from cells in culture. *MethodsX*, 1, 74–80.
- Sauer, U. (2006). Metabolic networks in motion: ¹³C-based flux analysis. *Molecular Systems Biology*, 2(1), 62.
- Saunders, R. D., & Horrocks, L. A. (1984). Simultaneous extraction and preparation for high-performance liquid chromatography of prostaglandins and phospholipids. *Analytical Biochemistry*, 143(1), 71–75.
- Sefer, E., Kleyman, M., & Bar-Joseph, Z. (2016). Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments. *Cell Systems*, 3(1), 35–42.
- Selvarasu, S., Ho, Y. S., Chong, W. P., Wong, N. S., Yusufi, F. N., Lee, Y. Y., Yap, M. G., & Lee, D. (2012). Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnology and Bioengineering*, 109(6), 1415–1429.
- Shelke, G. V., Lässer, C., Gho, Y. S., & Lötvall, J. (2014). Importance of exosome depletion protocols to eliminate functional and RNA-containing extracellular vesicles from fetal bovine serum. *Journal of Extracellular Vesicles*, 3(1), 24783.
- Shen, F., Cheek, C., & Chandrasekaran, S. (2019). *Dynamic network modeling of stem cell metabolism*. Computational stem cell biology (pp. 305–320). Springer.
- Sidhom, K., Obi, P. O., & Saleem, A. (2020). A review of exosomal isolation methods: Is size exclusion chromatography the best option? *International Journal of Molecular Sciences*, 21(18), 6466.

- Sitnikov, D. G., Monnin, C. S., & Vuckovic, D. (2016). Systematic assessment of seven solvent and solid-phase extraction methods for metabolomics analysis of human plasma by LC-MS. *Scientific Reports*, 6(1), 1–11.
- Snyder, N. W., Khezam, M., Mesaros, C. A., Worth, A., & Blair, I. A. (2013). Untargeted metabolomics from biological sources using ultraperformance liquid chromatography-high resolution mass spectrometry (UPLC-HRMS). *JoVE (Journal of Visualized Experiments)*, 75, e50433.
- Sonnenberg, R. A., Naz, S., Cougnaud, L., & Vuckovic, D. (2019). Comparison of underivatized silica and zwitterionic sulfobetaine hydrophilic interaction liquid chromatography stationary phases for global metabolomics of human plasma. *Journal of Chromatography A*, 1608, 460419.
- Stalidzans, E., Zanin, M., Tieri, P., Castiglione, F., Polster, A., Scheiner, S., Pahle, J., Stres, B., List, M., & Baumbach, J. (2020). Mechanistic modeling and multiscale applications for precision medicine: Theory and practice. *Network and Systems Medicine*, 3(1), 36–56.
- Suarez-Diez, M., & Saccenti, E. (2015). Effects of sample size and dimensionality on the performance of four algorithms for inference of association networks in metabolomics. *Journal of Proteome Research*, 14(12), 5119–5130.
- Tabatabaei Anaraki, M., Simpson, M. J., & Simpson, A. J. (2018). Reducing impacts of organism variability in metabolomics via time trajectory *in vivo* NMR. *Magnetic Resonance in Chemistry*, 56(11), 1117–1123.
- Tanaka, K., West-Dull, A., Hine, D. G., Lynn, T. B., & Lowe, T. (1980). Gas-chromatographic method of analysis for urinary organic acids. II. Description of the procedure, and its application to diagnosis of patients with organic acidurias. *Clinical Chemistry*, 26(13), 1847–1853.
- Théry, C., Amigorena, S., Raposo, G., & Clayton, A. (2006). Isolation and characterization of exosomes from cell culture supernatants and biological fluids. *Current Protocols in Cell Biology*, 30(1), 3–22.
- Théry, C., Ostrowski, M., & Segura, E. (2009). Membrane vesicles as conveyors of immune responses. *Nature Reviews Immunology*, 9(8), 581–593.
- Toubiana, D., Puzis, R., Wen, L., Sikron, N., Kurmanbayeva, A., Soltabayeva, A., Wilhelmi, M., del, M. R., Sade, N., Fait, A., & Sagi, M. (2019). Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Communications Biology*, 2(1), 1–13.
- Vellaichamy, A., Lin, C., Aye, T., Kunde, G., Nesvizhskii, A., Liu, E., & Sze, S. (2010). A chloroform-assisted protein isolation method followed by capillary nano LC-MS identify estrogen-regulated proteins from MCF7 cells. *Journal of Proteomics & Bioinformatics*, 3, 212–220.
- Vuckovic, D. (2012). Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Analytical and Bioanalytical Chemistry*, 403(6), 1523–1548.
- Villaret-Cazadumont, J., Poupin, N., Tournadre, A., Batut, A., Gales, L., Zalko, D., Cabaton, N. J., Bellvert, F., & Bertrand-Michel, J. (2020). An optimized dual extraction method for the simultaneous and accurate analysis of polar metabolites and lipids carried out on single biological samples. *Metabolites*, 10(9), 338.
- Wang, S., Wang, Z., Zhou, L., Shi, X., & Xu, G. (2017). Comprehensive analysis of short-, medium-, and long-chain acyl-coenzyme A by online two-dimensional liquid chromatography/mass spectrometry. *Analytical Chemistry*, 89(23), 12902–12908.

- Wang, L., & Maranas, C. D. (2018). MinGenome: an in silico top-down approach for the synthesis of minimized genomes. *ACS Synthetic Biology*, 7(2), 462–473.
- Weerheim, A. M., Kolb, A. M., Sturk, A., & Nieuwland, R. (2002). Phospholipid composition of cell-derived microparticles determined by one-dimensional high-performance thin-layer chromatography. *Analytical Biochemistry*, 302(2), 191–198.
- Whitehead, S. N., Hou, W., Ethier, M., Smith, J. C., Bourgeois, A., Denis, R., Bennett, S. A., & Figgeys, D. (2007). Identification and quantitation of changes in the platelet activating factor family of glycerophospholipids over the course of neuronal differentiation by high-performance liquid chromatography electrospray ionization tandem mass spectrometry. *Analytical Chemistry*, 79(22), 8539–8548.
- Wiechert, W. (2002). An introduction to ^{13}C metabolic flux analysis. *Genetic Engineering*, 215–238.
- Wiesner, P., Leidl, K., Boettcher, A., Schmitz, G., & Liebisch, G. (2009). Lipid profiling of FPLC-separated lipoprotein fractions by electrospray ionization tandem mass spectrometry. *Journal of Lipid Research*, 50(3), 574–585.
- Wikberg, E., Sparrman, T., Viklund, C., Jonsson, T., & Irgum, K. (2011). A ^2H nuclear magnetic resonance study of the state of water in neat silica and zwitterionic stationary phases and its influence on the chromatographic retention characteristics in hydrophilic interaction high-performance liquid chromatography. *Journal of Chromatography. A*, 1218(38), 6630–6638.
- Willms, E., Johansson, H. J., Mäger, I., Lee, Y., Blomberg, K. E. M., Sadik, M., Alaarg, A., Smith, C. E., Lehtiö, J., & Andaloussi, S. E. (2016). Cells release subpopulations of exosomes with distinct molecular and biological properties. *Scientific Reports*, 6(1), 1–12.
- Witwer, K. W., Buzás, E. I., Bemis, L. T., Bora, A., Lässer, C., Lötvall, J., Nolte-'t Hoen, E. N., Piper, M. G., Sivaraman, S., & Skog, J. (2013). Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *Journal of Extracellular Vesicles*, 2(1), 20360.
- Xu, H., Valenzuela, N., Fai, S., Figgeys, D., & Bennett, S. A. (2013). Targeted lipidomics—advances in profiling lysophosphocholine and platelet-activating factor second messengers. *The FEBS Journal*, 280(22), 5652–5667.
- Yang, K., & Han, X. (2016). Lipidomics: Techniques, applications, and outcomes related to biomedical sciences. *Trends in Biochemical Sciences*, 41(11), 954–969.
- Yizhak, K., Chaneton, B., Gottlieb, E., & Ruppin, E. (2015). Modeling cancer metabolism on a genome scale. *Molecular Systems Biology*, 11(6), 817.
- Yuan, Y., & Bar-Joseph, Z. (2020). GCNG: Graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*, 21(1), 1–16.
- Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Computational Biology*, 15(7), e1007084.
- Zeleznik, A., Vowinkel, J., Capuano, F., Messner, C. B., Demichev, V., Polowsky, N., Mülleider, M., Kamrad, S., Klaus, B., & Keller, M. A. (2018). Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts. *Cell Systems*, 7(3), 269–283.
- Zhai, D., & Reilly, P. J. (2002). Effect of FA chain length on normal-and reversed-phase HPLC of phospholipids. *Journal of the American Oil Chemists' Society*, 79(12), 1187–1190.

- Zhang, Z., Wang, C., Li, T., Liu, Z., & Li, L. (2014). Comparison of ultracentrifugation and density gradient separation methods for isolating Teca8113 human tongue cancer cell line-derived exosomes. *Oncology Letters*, 8(4), 1701–1706.
- Zhen, H., Ekman, D. R., Collette, T. W., Glassmeyer, S. T., Mills, M. A., Furlong, E. T., Kolpin, D. W., & Teng, Q. (2018). Assessing the impact of wastewater treatment plant effluent on downstream drinking water-source quality using a zebrafish (*Danio Rerio*) liver cell-based metabolomics approach. *Water Research*, 145, 198–209.
- Ziemanski, J. F., Chen, J., & Nichols, K. K. (2020). Evaluation of cell harvesting techniques to optimize lipidomic analysis from human meibomian gland epithelial cells in culture. *International Journal of Molecular Sciences*, 21(9), 3277.

Single cell metabolomics 13

Minakshi Prasad¹, Mayukh Ghosh², and Rajesh Kumar³

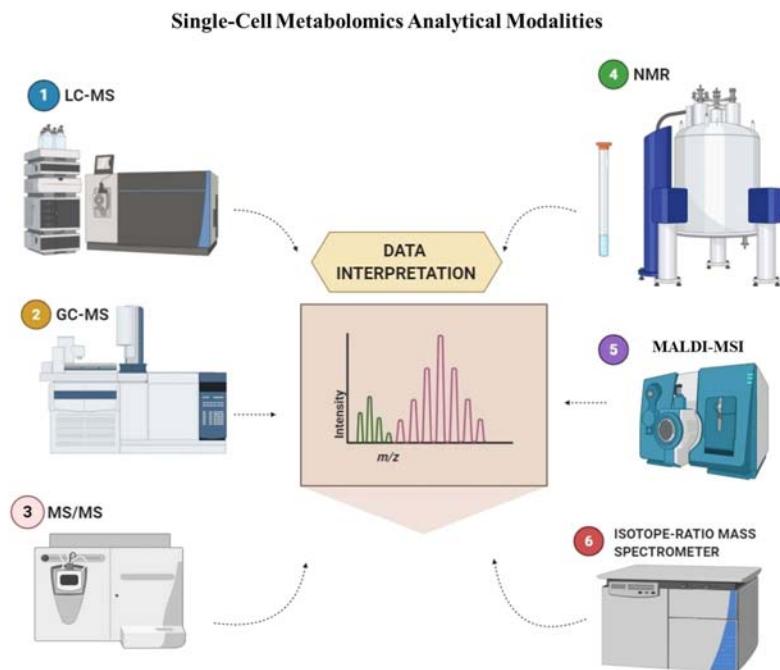
¹*Department of Animal Biotechnology, Lala Lajpat Rai University of Veterinary and Animal Sciences, Hisar, India*

²*Department of Veterinary Physiology and Biochemistry, RGSC, Banaras Hindu University, Mirzapur, India*

³*Department of Veterinary Physiology and Biochemistry, Lala Lajpat Rai University of Veterinary and Animal Sciences, Hisar, India*

Introduction

Metabolism encompasses an array of purposefully regulated chemical reactions, which renders the cells to adapt and dynamically respond to diverse pathophysiological stimuli, ultimately deciding the cell fate and phenotypic outcome. Metabolic alterations, reflected as metabolite turn-over, are far more rapid and exclusively related to the phenotypic cellular responses, as compared to genetic or epigenetic introspection. Evidently metabolite profiling has gained significant impetus as a penetrative analytical chemistry tool to unearth molecular basis of diverse cellular and biochemical mechanisms. Despite the inherent challenges of enormous chemical complexity and diversity, rapid turn-over, limitation of low concentration and unavailability of amplification methods, which have rendered “metabolomics” lagging behind the other “omics” techniques, the worth of information generated through metabolomics introspection and its vast applications has certainly propelled the developments in instrumentation, as well as in analytical tools to take place in metabolomics arena during the ongoing era. Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) spectroscopy are the two major vertices of the metabolomics analytical modalities along with chromatographic or electrophoretic separation methods, fluorescence-based techniques, Vibrational spectroscopy including Raman spectroscopy and Fourier transform–infrared spectroscopy, MS imaging (MSI) accompanied by several state-of-the-art bioinformatics tools and data analysis programs which have facilitated metabolic profiling from diverse biological samples (Fig. 13.1). Often, these modalities are employed to monitor the metabolite diversity from tissue samples or biological fluids as stimuli-induced systemic manifestation, which is actually a cumulative overlapping metabolic response of multiple cells or diverse cell types under certain patho-physiological signal. Such cumulative population effect

**FIGURE 13.1**

Cell isolation techniques for single-cell metabolomics.

usually masks the metabolic responses that arise from the cellular heterogeneity as it is now vividly established that individual cells, even within the same population, behave quite differently. Further, these variations in cellular responses bear significant influence over the health and disease outcome, more importantly during the disease establishment and early progression. Hence, high-throughput techniques capable of isolating and handling the single cells with minimum manipulations, and the analytical modalities robust as well as sensitive enough to achieve single cell level resolution from picoliter to femtoliter sample volume along with adept bioinformatics programs to manage the huge metabolomics data and interpret them in a meaningful way, are the prerequisites to unmask such heterogeneity of cellular metabolic responses at single cell level.

The importance of single-cell research has been well recognized by the National Institutes of Health, which have initiated funding of around United States \$2 million in 2014 among 60 prominent research groups throughout the world under a collaborative program to promote the single-cell profiling. Similarly, The Single Cell Surveyor Society, a Japan-based corporate and academic collaboration also provides research funding as well as organizes workshop and symposia to promote research as well as decipher knowledge in single-cell technology. Several other international initiatives such as The Human Cell Atlas,

which has been intended toward comprehensive mapping of all the different human cell types, or The International Human Epigenome Consortium, which has been targeted to generate reference epigenomes from diverse human cell types, enormously depend upon single-cell research, simultaneously justifying the significance of the current discussion. Besides putting more emphasis to the single-cell analysis by different international societies, devoted set up have been developed by different countries such as Single Cell Omics Germany, Single Cell Center by Chinese Academy of Sciences, The Sanger Institute-EBI Single-Cell Genomics Centre, Oxford Single Cell Biology Consortium, etc. toward customization of specialized instruments, data analysis programs or adept resources to facilitate cutting-edge single-cell research. The diligent effort to increase the robustness, sensitivity and throughput of the analytical modalities have paid off considerably as several state-of-the-art analytical platforms have evolved or enriched recently to meet up the challenges of single-cell metabolomics analyses. Capillary electrophoresis (CE) or chromatography coupled with MS based platforms such as CE- μ ESI-MS, Nano-HPLC-MS or GC-TOF-MS; several matrix-based ionization methods integrated with MS or MSI modality such armada-MSI, live nano-ESI-MS, image-guided micro-MS, secondary ion MS (SIMS) imaging, matrix-free laser desorption ionization (LDI) methods, Nanoparticle-assisted LDI-MSI, Desorption Electrospray Ionization (DESI)-MSI, laser ablation electrospray ionization (LAESI)-MSI, silicon-based methods like Desorption Ionization on Silicon, or nanostructure-initiator MS imaging, Silicon Nano post Array Mass Spectrometry Imaging, Microfluidic chips or microarray system-based MS platform like microarrays for mass spectrometry chips are some of the prominent MS-based single-cell metabolomics (SCM) analytical modalities which are getting established progressively, mostly through proof-of-concept researches. Although NMR is extremely robust and convenient method for fluid-phase metabolite analysis due to easy sample processing but limited sensitivity has restricted its application in SCM analysis. However, increasing the magnetic field from the conventionally operated 500–800 MHz frequency to 1.2 GHz level through advanced magnets can enhance the sensitivity of NMR significantly as well as improve the spectral dispersion for distinct separation of overlapping signals. Simultaneously, development of highly sensitive cryogenically cooled probes or micropores which enhances the signal-to-noise ratio along with employment of microfluidic platforms can be suitable to work under mass-limited condition (Giraudieu, 2020). Further, improvisation of 2D-COSY with microfluidic diamond quantum sensor and nitrogen-vacancy-doped diamond nanograting chip-based NMR are certain prominent developments toward alleviating the limitations of NMR for SCM analysis. Recently, Raman spectroscopy has also proved its worth in SCM analyses by elucidating druggable fatty acid synthesis pathway in melanoma cells, thus enriching the arsenal of single-cell research modalities (Du et al., 2020).

The aforesaid discussion has clearly outlined the enormous intensity of the efforts and extensive investment of resources already devoted and still ongoing to facilitate SCM research. This can only be justified by either extensive application

of SCM techniques in diversified area or extremely precise intervention in the niche area, which exclusively comes under the purview of SCM introspection. The field of SCM is relatively nascent; therefore most of the current applications are serving “the quest for precision and excellence” through proof-of-concept works rather than operated for commonplace interventions. However, SCM techniques may have a smooth-going in panoramic biological arena which includes but not limited to plant and animal diseases, their molecular pathogenesis, diagnosis and therapy, elucidation of therapeutic checkpoints through metabolic pathway introspection and drug discovery, understanding drug resistance issues, biomarker discovery, precision medicine, host-pathogen interaction, microbiology and development of antibacterial resistance, cell and developmental biology, distribution of bioenergetics within cells under different physiological and pathological conditions, reproductive biology, environmental science and enumerating the effects of different environmental determinants on single cell to multicellular organisms, food science, and agriculture; and plethora of other improvised applications ([Table 13.1](#)). The already proven as well as the potential SCM applications with an emphasis over the biomedical interventions will be delineated in the respective subsections of the current chapter.

However, wide diversity in cell size and intern in operational volume among plant cells sizing about 10–100 µm, animal cells of 10–20 µm and submicron size microbial cells renders difficulty in SCM analyses as no single analytical modality or method can accommodate such ample variations ([Fig. 13.2](#)). Further, not undermining the potential of SCM analyses, integration of SCM data with other existing single-cell-omics outcomes is desirable as it will certainly be helpful to plug-in several knowledge-gaps in explaining the molecular mechanisms of several patho-physiological processes. To be honest, now SCM techniques have already voyaged some distance and more emphasis should be directed toward validation which may not be so much enticing or glorifying but earnestly needed for successful transition of the SCM techniques from proof-of-concept work to clinical biomedical applications ([Table 13.1](#)).

Single-cell metabolomics in microbial technology

Microscopic organisms such as bacteria, archaea, protozoa or fungi possess panoramic diversity, probably the greatest among the living entity in the earth. Apart from the infamous pathogenic ones, the diverse microbial groups are significant contributors of the ecosystem. Microbial metabolism holds the repository of several key enzymes and metabolites of commercial importance. Evidently, insight into the microbial metabolic processes and pathways opens the avenue for identification of potential targets to maneuver through diverse industrial applications such as food processing and microbial bioprocess technology, textile industry, pharmaceutical industry, feed technology, etc. deciphering the benefits to broad

Table 13.1 Prominent SCM interventions in diverse biological applications.

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>In lower organisms</i>						
<i>Euglena viridis</i> , <i>Scenedesmus obliquus</i> , <i>Dunaliella salina</i> and <i>Chlamydomonas reinhardtii</i>	Electromigration	Electroporation	nESI-MS negative ion mode	YMDB and plant metabolic pathway databases (https://www.plantcyc.org/)	11 common metabolite in all organism, 47 metabolites in <i>C. reinhardtii</i>	Li et al. (2020)
<i>Saccharomyces cerevisiae</i>	Flow cytometer	Quenched with methanol at 408 C a and extract with help of ethanol boiling	MALDI-MS time-of-flight mass spectrometry	Spectra identification only	Attomoles level of sensitivity from a single cell having volume ranging from 4 nL and 390 pL	Amantonico et al. (2008)
<i>Escherichia coli</i> , <i>Akkermansia muciniphila</i> and <i>Bacteroides acidifaciens</i>	Raman microspectroscopy and D ₂ O	—	—	—	Used heavy water (D ₂ O) combined with Raman microspectroscopy for single sorting of live microorganism	Berry et al. (2014)
<i>Daphnia magna</i>	Solid-phase microextraction (SPME)	Probe electrospray ionization coupling SPME probe	nanoESI-MS	Principal component analysis (PCA)	1 μm tip for sampling single cell and studied bioaccumulation and kinetics of perfluorooctanesulfonic acid (PFOS) and perfluorooctanoic acid (PFOA)	Deng et al. (2015)
<i>Mycobacterium smegmatis</i>	Microfluidics	Live, no need of destruction	Time-lapse microscopy and FRET-based ATP biosensors		ATP levels were tracked in respect to time, antibiotic treatment.	Maglica et al. (2015)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>E. coli</i> , <i>Pseudomonas putida</i> , <i>Pseudomonas fluorescens</i> <i>Bacillus subtilis</i>	Deuterium isotope probing Plasmonic trap array	Live no need of destruction No need	Raman microspectroscopy Surface-enhanced Raman scattering (SERS)	LabSpec 5 (Horiba) SERS spectra software	Differential phenylalanine pathways in two species was observed Dipicolinic acid and paminothiophenol were measure at attomole level sensitivity	Xu et al. (2017) Yao et al. (2018)
<i>Synechococcus</i> sp., the chlorophyte <i>S. obliquus</i> , and the cryptophyte <i>C. ovata</i>	Culture cell by flow cytometry		NanoSIMS	Imaging software of instruments	Computation of size and uptake rates of metabolites performed in order to see volume based differentiation.	Nakashima et al. (2016)
<i>Tetrahymena</i>	From culture by silicon for cryogenic freezin	Freeze-fractured under ultrahigh vacuum	Mass spectrometric imaging	Mass spectrum	A high curvature lipid 2-aminoethylphosphonolipid are present in elevated amounts in fusion region indicating a heterogeneous redistribution of lipids	Ostrowski (2004)
<i>In plants</i>						
<i>Allium cepa</i>	Three-dimensional manipulator	≤ 1 μm probe ESI	Orbitrap MS	Plant Metabolic Network database: http://plantcyc.org/ ; or LIPID MAPS	Revealed various metabolites and reported internal epidermal cell have 6 types of fructans while only three type in outer.	Gong et al. (2014)
<i>Allium cepa internal epidermal cell</i>	Three-dimensional manipulator	≤ 1 μm probe ESI	Orbitrap MS	Plant Metabolic Network database: http://plantcyc.org/ ; or LIPID MAPS	Nucleus have only four type of fructose while cytoplasm have 6 type of fructose but concentration vary	Gong et al. (2014)

<i>Allium cepa epidermis and PC12 cell</i>	Micro manipulator with emitter tip for SPIESI	Polarization induced electrospray ionization (PI-ESI) with alternating current square wave voltage (AC-SWV)	Q-TOF mass spectrometer	METLIN metabolite database, LIPID MAPS, and Massbank database	Leucine, glutamine, histidine, dopamine histamine and fructosamine were detected uniquely in positive-ion mass whereas negative ion mass spectra detected malic acid, ascorbic acid, citric acid, ADP, UDP-HexNac, NAD exclusively from the sample indicating both mode is important for greater range Secreted material study was performed with Several hundreds of peaks were detected, Heterogeneity in trichome cells was found and some metabolites were specific in certain trichome cell.	Hu et al. (2016)
<i>Glandular Cells of Intact Trichomes of Solanum lycopersicum L</i>	Capillary tip	Capillary tip	IEC-PPESI-MS	Metlin, Plant Metabolite Network, and Solcyc	More than 19 metabolites identified Picoliter magnitude quantitative analysis was achieved with 100 of spectra	Nakashima et al. (2016)
<i>Outer epidermal cells of Allium cepa</i> <i>Holly leaf, Allium cepa inner epidermal cell</i>	Nanopipette under microscope Pressure-assisted microsampling probe under microscope	NanoESI Pressure-assisted microsampling probe with picoliter pump	LTQ/Orbitrap XL MS Hydrogen Flame Desorption Ionization triple-quadrupole MS	MS spectra	More than 19 metabolites identified Picoliter magnitude quantitative analysis was achieved with 100 of spectra	Yin et al. (2018) Zhao et al. (2019)
<i>Allium cepa, Citrus aurantium</i>	Laser ablation	Chemical	LAESI-MS	OPLS-DA	Metabolite cyaniding, the ion responsible for purple pigmentation, 300 peaks identified, potential biomarker identification	Shrestha et al. (2011)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>In animal and human subjects</i>						
Liver cells	Microfluidics	Fluorescence	Laser-induced fluorescence detection (MCE-LIFD)	In built software of PMT-2	Hydrogen peroxide, glutathione, and cysteine were determined accurately in ethanolic mouse showed variable response to stressor and antioxidant generation among cells.	Li et al. (2016)
MCF7 cells	SilacaTip with a tip diameter of $10 \pm 1 \mu\text{m}$	Chemical	Ion Trap Mass Spectrometer	Software tool based on ITCL (Ion Trap Control Language)	Repeat mode of analysis enhance identification by 2 to 22 folds hence very beneficial in low abundance metabolites identification had been archived in more than 7 such compounds	Si et al. (2017)
Mouse liver	PESI needle	Electrospray ionization	PESI- MS/MS and GC-MS	Multivariate analysis was executed using SIMCA-P + software, Welch's <i>t</i> -test	26 metabolites identified	Zaitsu et al. (2016)
Human Hepatocytes	Capillary microsampling with help of motorized micromanipulator	Electrospray ionization after sucking, also by chemical method	ESI-IMS-MS	DriftScope 2.8 software, Human Metabolome Database, METLIN metabolite database, and LIPID MAPS	Identified 22 metabolites and 54 lipids the technique hold good for animal cells which have 50 – 1000 lower volume than a plant cell	Zhang and Vertes (2015)
PC12 single cell	Reduced graphene oxide functionalized copper probe (rGO–Cu)	Electrospray ionization	rGO-Cu functional probe FPESI - UHPLC-Q-TOF MS	ChemBank, PubChem, MassBank,	AD related neurotransmitters and sixteen biomarkers metabolite, 19 other	Zheng et al. (2020)

<i>Allium cepa</i> and <i>HeLa</i> cell	Single cell manipulator platform and capillary	Electric field strength	ESI-MS	ChemSpider and SciFinder scholar Massbank database, MATLAB	metabolites identified and useful in rapid detection 1034 components and 656 components in two cell identified	Wei et al. (2015)
Developmental biology applications						
<i>Xenopus laevis</i> embryo	Manual	Chemical method	CE- μ ESI-MS	Tandem MS databases	Identified 40 metabolites expressed differentially between cells of 16 cell embryo indicating heterogeneity	Onjiko et al. (2015)
<i>Xenopus laevis</i> Single unfertilized and fertilized eggs, and 32-cell stage embryo cell	Manual	Chemical	LAESI-MS	METLIN, MetaCyc, Lipid Maps, HMDB, KEGG and NIST Isotope Calculator program (ISOFORM, Version 1.0	Subcellular Metabolite and Lipid analysis were performed to study fertilization and early embryonic development in vertebrates and identified more than 52 metabolites and other small molecules	Shrestha et al. (2014)
8-cell Embryo of <i>Xenopus laevis</i>	Microcapillary	Chemical	CE-ESI-MS	Compass DataAnalysis version 4.3, MetaboAnalyst 4.0	With a LOD of ~5–10 nM (50–100 amol) nearly 450 peaks with more than 80 metabolites identified.	Portero and Nemes (2019)
<i>Xenopus laevis</i>	Capillary microprobe	Chemical	CE-ESI-MS	Metlin and HMDB	8-to-32-cell <i>X. laevis</i> embryos we reanalyzed with 10 nL fluid extracted with LOD of 60-amol revealed metabolic differences between vegetative pole left and right cells.	Onjiko et al. (2017)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
16-cell <i>Xenopus laevis</i> embryo	Manual with help of with a stereomicroscope	Multiple solvent of differed pH.	CE-HRMS	Compass Data Analysis ver. 4.0, MetaboAnalyst3.0, Mass Frontier 7.0, MarvinSketch 16.1.11	Hundreds of different metabolites detected with differential abundance of metabolites between right and left cells.	Onjiko et al. (2016)
Skeletal muscle	Manual	Methanol/ chloroform/ water and water/methanol	FIA-MS/MS, GC-MS, LC-HRMS, and NMR		Detail insight of skeletal muscle metabolism with 132 metabolites and 58 pathways were identified.	Bruno et al. (2018)

SCM in studying aging, senescence of cell and stem cell biology applications

<i>S. cerevisiae</i>	Biotin-labeled isolation MACS	Chemical	GC-MS	SIMCA-P + program	Alteration of metabolomics profile during various generation, particularly accumulation of pyruvic acid and TCA cycle intermediates, depletion of most amino acids, particularly branched-chain amino acid	Kamei et al. (2014)
<i>Human Pluripotent Stem Cell-Derived Cardiomyocytes</i>	Flow cytometry	Chemical	NMR	ACD/1D NMR processor, HMDB metabohunter, targeted profiling by using ChenomX NMR Suite Profiler	Identifies Metabolic Markers of Maturation	Bhute et al. (2017)
<i>CD4 + T Cell</i>	Flow cytometry	Chemical	LC/MS.	e MAVEN software suite, SciKit-Learn machine learning library in Python (https://scikit-learn.org/stable/)	Studied activation of T cell which was attributed due to defective respiration and one-carbon metabolism in from aged mice it was decreased. more than 100 metabolites identified	Ron-Harel et al. (2018)

<i>Progenitor cell</i>	Manual	Chemical	Gas chromatography	GC solutions 2.3	Brain activity of learning depends upon Lipid Metabolism of neuron and their activity is further related with neuronal stem cell activity	Bowers et al. (2020)
<i>Yeast mutants</i>	Culture cells	Chemical	ESI-MS	MassLynx software	Metabolic footprinting' approach recognizes the significance of "overflow metabolism" in identifying mutants site	Allen et al. (2003)
<i>S. cerevisiae strains</i>	Culture cells	Chemical	NMR	NMR spectra	Faster method of silent gene identification within different strains	Raamsdonk et al. (2001)
<i>Induced pluripotent stem cells (hiPSCs)</i>	Tissue embedded	No need live	Single-cell Raman microspectroscopy (SCRIM)	Machine learning platform: k-Nearest Neighbor (kNN), SGB, random forest (RF), linear support vector machine (SVM), and SVM with radial basis function (RBF)	Faster biomarker based identification and differentiation of stem cell in live form	Su et al. (2020)
<i>Bone marrow mesenchymal cells</i>	Flow cytometry	Vortexed and chemical treatment	LCMS	MassHunter Qualitative Analysis B.08.00 software	Studied the effect of AST to induce osteogenic differentiation with 24 metabolite identified differentially and affecting 11 pathways	Zhao et al. (2020)
<i>Mesenchymal Stem Cells</i>	Flow cytometry or immunocitochemistry profile and karyotype	Various methods	NMR spectroscopy, MS, chromatography	NMR assignment software. T etc.	Secretion from such cell were identified	Ivanova et al. (2016)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>Human Pluripotent Stem Cell-Derived Cardiomyocytes</i>	Flow cytometry	Chemical	NMR	NOESYPR1D pulse sequence, ChenomX NMR Suite ProfilerMetaboLights database	Studied maturation profiling and revealed that fatty acid oxidation and metabolism are more profoundly altered during maturation	Bhute et al. (2017)
Cancer stem cells	FACS	Single-probe	LTQ Orbitrap XL mass spectrometer	Thermo Xcalibur Qual Browser; Geena 2 online software tool; Metaboanalyst 4.0; human metabolome database (HMDB)	Live Single were studied to reveal metabolite changes between cancer and normal stem cell particularly related with TCA cycle metabolites including unsaturated lipids; inhibiting the activities of stearoyl-CoA desaturase-1, NF-κB, and aldehyde dehydrogenases	Sun et al. (2018b)
<i>E. coli</i>	Microfluidic	—	—	—	Different cell response differentially in term of survival and growth when given differential environment and growth media.	Dal Co et al. (2019)

Environmental science applications

<i>MCF-7, A2780, 293, and 4T1 cells</i>	Droplet-based single-cell printing analysis system	Droplet-based single-cell printing analysis system	Mass spectrometry	METLIN and HMDB	30 to 40 times higher processing rate can be beneficial in environmental study, cell quality control, cell biology, cancer diagnosis, and prevention	Li et al. (2019)
---	--	--	-------------------	-----------------	--	------------------

<i>Various tomato and other plant cell including Arabidopsis thaliana and Vicia faba guard cells</i>			GC-TOF-MS, laser capture microdissection (LCM), LCM optionally coupled to laser pressure catapulting (LMPC) and RT-PCR GC-MS, UPLC-QTOF-MS	Various programs	How different plant cells respond to different stressors	Nourbakhsh-Rey and Libault (2016)
<i>B. japonicum</i>	Culture, manual	Chemical	LAESI-MS	MET-IDEA software	2610 root hair metabolites were identified under different n2 fixer conditions	Streeter (2003), Brechenmacher et al. (2010)
<i>Allium cepa and Narcissus pseudonarcissus bulb epidermis, as well as single eggs of Lytechinus pictus</i>	Laser ablation of a single plant cell		Synchrotron-based infrared microspectroscopy, fluorometry, FTIR	e NIST Isotope Calculator package (ISOFORM, version 1.02). The Plant Metabolic Network database	332 peak identified with 25 metabolites diversity depends on species, Different Age and environment, pigmentation	Shrestha and Vertes (2009)
<i>Acropora millepora</i>	Formalin-fixed cells	No	CE-ESI-MS	Unscrambler software package	Studied metabolomics under heat stress in relation to symbiotic behavior with alga	Petrou et al. (2018)
<i>Aplysia californica</i>	Manual	Chemical	ESI-MS	Mass spectra software	Metabolomic profile changes during analysis of fresh and culture neuron including arginine, glutamine, histamine etc.	Nemes et al. (2012)
<i>Circulating tumor cells</i>					Directly perform metabolomics from their native environment,	Hiyama et al. (2015)

Organ system study

<i>Various cell</i>	Pulsed laser beam	Pulsed laser beam	Mass spectrometry imaging (MSI) based on MALDI-MSI	MIRION	phospholipids, drug molecules, neuropeptides, and tryptic peptides identified	Römpf and Spengler (2013)
---------------------	-------------------	-------------------	--	--------	---	---------------------------

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>Aplysia californica metacerebral cell and R2 cell</i>	Manual by needles under visual control	Capillary electrophoresis	CE-ESI-MS	maXis tandem mass spectrometer	more than 100 compound identified	Lapainis et al. (2009)
<i>Aplysia californica B1 vs B2 neurons</i>	Manual	Chemical	CE-ESI-MS	Mass spectra software	300 distinctive cell-related signal with 35 metabolites identified	Nemes et al. (2012)
<i>rat thalamic cells</i>	Patch pipet	Capillary electrophoresis (CE) with patch clamp capillary electrophoresis	CE-MS	Metlin and HMDB	approximately 60 metabolites were identified	Aerts et al. (2014)
<i>Aplysia californica B1, B2, left pleural 1 (LP1), metacerebral cell (MCC), R2, and R15</i>	Manually	CE-ESI-MS	Visualization software package (O	144 ionic species showed heterogeneity among cell	Nemes et al. (2011)	
<i>Aplysia californica and Rattus norvegicus neuron</i>	Manual	Chemical	CEESI MS	Metlin, HMDB, LIPID Metabolites and Pathways Strategy resource (LIPID MAPS, LipidBank, MassBank ChemSpider, ChEBI and NeuroPred for neuropeptides	300 distinct compounds with LOD to about 100 amol heterogeneity with respect of cell	Nemes et al. (2013)
<i>Aplysia californica</i>	Microdissection	Chemical	MALDI/TOF-MS	Mass spectrum	Peptide and neurohormone profiling	Garden et al. (1996)
<i>Lymnaea brain</i>	Manual	Chemical	MALDI-MS	Mass spectrum	Various bioactive peptides and x prohormone were identified	Jiménez et al. (2008)
<i>Insecte; Periplaneta americana</i>	Manual	Air dried and dissolve in water	MALDI-MS	Mass spectrum	21 bioactive peptide identified	Neupert and Predel (2005)

<i>Human RBC</i>	Microfluidics	Buffer and electric field strength	ESI-MS	m/z spectra	RBC compounds were detected including form of hemoglobin	Mellors et al. (2010)
<i>HeLa (human cervical cancer)</i>	Single-probe	Single-probe	Live MS	Metlin	192 and 70 compounds were identified with two dicationic reagents	Pan et al. (2016)
<i>Wbc, ctc</i>	FACS		MS	Hmida metabolome data	More than 310 metabolites identified	Hiyama et al. (2015)

Sub-cellular metabolite analysis

<i>Cervical cancer cells</i>	Single-probe with one fused silica capillary and one nano-ESI emitter	Single-probe	CE-ESI-MS	MS spectra	Remove several steps and by live cell metabolome generated to understand cancer cell metabolomics and drug effects	Pan et al. (2014)
<i>Human T-cell</i>	MACS	Staining by barcoded antibody	(Cytometry by time of flight, CyTOF)	FlowSOM R package, the SCORPIUS and Slingshot algorithms, CDSE approach	Antibody based metabolomics profiling to understand metabolic regulome profiling (scMEP), and pathways	Hartmann et al. (2021)
<i>Rat Basophil Leukemia Cell</i>	Nano-ESI tip	Nano-ESI tip	Q-TOF MS		Tryptophan and Histidine Metabolites and their pathway studied by live single-cell video-mass spectrometry.	Mizuno et al. (2008)
<i>T cells</i>	Magnetic separation using CD3 Microbeads and FACS		Extracellular flux analyzer, Cytokine and chemokine analysis, Flowcytometer	FlowJo, chord plots1	Single-cell metabolic analysis showing heterogeneity in same type of immune cell	Ahl et al. (2020)
<i>Macrophage</i>	Flow cytometry	Chemical	LC-MS	XCMS mzMatch Xcalibur and IDEOM KEGG	Generated macrophage metabolome model	Rattigan et al. (2018)
<i>Red and blue, (yellow and green, blue and green) PC12 cell</i>	Micro-droplet, Fluorescent confocal	ESI	Micro-well array and mass spectrometry LTQ-Orbitrap mass spectrometer		Tyrosine, dopa, dopamine and epinephrine and other secondary metabolites were measured showing heterogeneity of function	Fujita et al. (2015)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>Aplysia californica</i> bag cell neurons	Microfluidics	C18 coated surface	Mass spectrometric imaging	Bruker FlexImaging, Biomap imaging software	Peptide released from neurons studied in potassium stimulation experiment and identified several peptides	Jo et al. (2007)
<i>Aplysia</i> bag cell neuron	Microfluidic	OTS functionalized surface	MALDI-MS	Fleximaging software	LOD was as low as 600 fmol for AP and 400 fmol for aBCP.	Zhong et al. (2012)
Neuron, macrophage etc	Direct insertion of nanoelctrode	No need	Nanoelectrode sensors		Standardized for live single cell to detect various biochemicals	McCormick and Dick (2021)
Hepatocytes of mice	Manual	Chemical	MS	LipidXplorer	17 lipid classes showing characteristic among different cell and glycerolipid metabolism with a LOD as low 0.2 fmol labeled lipid in 20 µl sample	Tebani and Bekri (2019)
<i>Chara australis</i> vacuole and cytoplasm	Manual	Microinjection	CE-MS	MS spectra	Used for study metabolomics in sub cellular component where 125 metabolites identified and revealed metabolomic changes in dark and light cycle at several time lapses.	Oikawa et al. (2011)
HEPG2 mitochondria	Live cell	Nanospray tip	LTQ-Orbitrap MS	MetFrag, Marker view, KEGG, LIPID MAP, MitoRed, Massbank	Identified 5000 metabolites peak with 1700 metabolites	Esaki and Masujima (2015)

<i>Catharanthus roseus</i> stem tissue in four different kinds of cells <i>[idioblast (specialized</i> <i>parenchyma cell),</i> <i>laticifer, parenchyma,</i> <i>and epidermal cells</i>	Bright field and epifluorescence microscopy	Chemical	LTQ-Orbitrap MS	Principal component analysis (PCA)	Detail description of metabolite localization particularly alkaloid.	Yamamoto et al. (2016)
<i>C. glutamicum</i> DM1919	Microfluidics	Capillary	DI-MRMS direct infusion magnetic resonance mass spectrometry (DI- MRMS)			Dusny et al. (2019)
<i>Human cervical cancer</i> <i>cells (HeLa)</i>	Microarrays	Micro-tip	Fluidic force microscopy and MALDI-TOF MS	EasyScan2 software; AxioVision software (ZEISS), openMS PeakPickerHighRes algorithm	Even sample volumes of 1 µL and less can be quantified, Microfluidics have potential to provide environment required so act as bioreactor. 20 metabolites identified from cell cytoplasm by non-destructive and quantitative analysis	Guillaume-Gentil et al. (2017)
<i>HeLa cells</i>	Laser ablation inductively coupled plasma (LA-ICP)	Electrospray	Micro-Lensed Fiber Laser Desorption Mass Spectrometry		Revealed subcellular localization of different drugs	Hang et al. (2020)
<i>HeLa cells, NIH3T3</i> <i>fibroblasts, human</i> <i>hepatocytes</i>	Cell-ablation	Electrospray	MALDI-imaging mass spectrometry and bright-field and 57 fluorescence microscopy	METASPACE cloud software	Give localization of various metabolites indicating human hepatocytes accumulation pattern of accumulation of long- chain phospholipids	Rappez et al. (2019)
<i>Allium cepa cell</i>	Tip	Tip direct desorption/ ionization	PESI-MS	Plant Metabolic Network database: http://plantcyc.org/ ; or LIPID MAPS	Detected metabolites at sub-cellular level	Gong et al. (2014)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

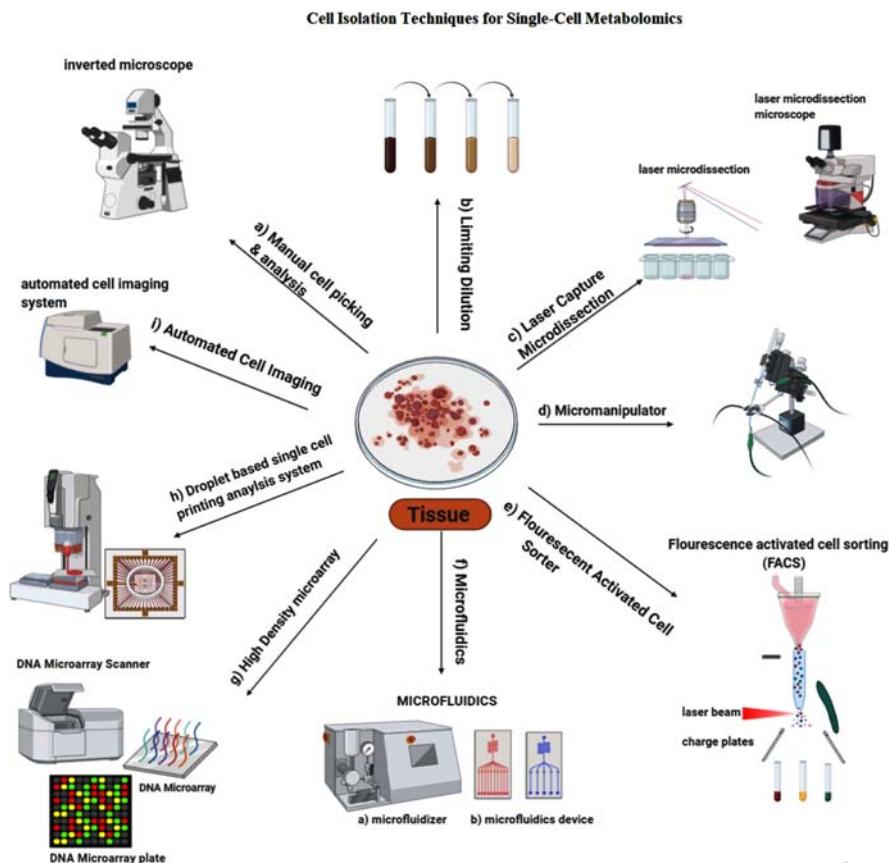
Cell	Isolation	Quenching	Platform	Software	Major finding	References
Clinical applications						
264.7 cell	Manual	Dried	MALDI/ESI-FTICR	MALDI IMS dataset, PCA analysis	670 distinct signals identified in cell and cell with LPS stimulation revealed PI(36:1), PI(36:2), PE(36:1), PE(36:2), and PA(36:2) are significantly increased and PI(40:5), PI (38:4), PE(38:4), and PE (36:4) significantly decreased in LPS cell	Yang et al. (2018)
Cancer						
HeLa cell	Fused silica capillary	Live	Nano-LCMS	Labsolution 5.91, MetaboAnalyst	18 hydrophilic metabolites profiling indicate heterogeneity in same culture	Nakatani et al. (2020)
Pancreatic and ovarian cancer	Dispersed magnetic Fe ₃ O ₄ @TiO ₂ nanocomposite particle	Phase microextraction	Extractive Electrospray Ionization Mass Spectrometry	Human Metabolome Database (http://www.hmdb.ca). Matlab	Lipid profiling which is highly sensitive and accurate for trace phospholipids in microvolume	Zhang et al. (2020)

<i>CTM of neuroblastoma</i>	Direct trapping	Nanospray tip, sonication	MS	Human Metabolome data base, Marker view, KEEG, Lipid MAPS	Generated lipid metabolomic profile of 216 peak	Hiyama et al. (2015)
<i>CTM of different origin</i>	Microfluidics	LIVE	LTQ Orbitrap	MarkerView software, Kyoto Encyclopedia of Genes,	Metabolomic profile, Human Metabolome Database,35 and LIPID MAPS structure database	Abouleila et al. (2018)
<i>U87, U251, HepG2, MCF-7, 293, Caco-2, HUVEC, and 3T3 cells</i> <i>HeLa cell</i>	Inkjet sampling from suspension Microscale multifunctional device, the T-probe	ESI T-probe	PESI-MS LTQ Orbitrap XL mass spectromete	Spectra analysis Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA), METLIN and HMDB	Identified cellular surface phospholipid composition variations. The LOD of PC(18:1/16:0) is 5 nM, got 100 s of peak, identified various metabolites changes due to drug effects.	Chen et al. (2016) Liu et al. (2018)
<i>Human Breast Cancer Cells</i>	Dilution and hemocytometer	Glass-bead-assisted method and ultrasonication cell	nanoLC-MS	Daltonics Data Analysis 4.3., IsoMS, dansyl standard library and MyCompoundID (MCID) libraries	Used radiolabeled metabolite culture to see its addition and changes in metabolomic maps	Luo and Li (2017)
<i>Breast cancer</i>	Flow cytometry	ESI	ESI-MS	HMDBand METLIN metabolite database	Label-free Mass Cytometry for Unveiling Cellular Metabolic Heterogeneity and identified Hundreds of metabolites	Yao et al. (2019)
<i>Breast cancer subtypes</i>	Droplet microextraction	ESI	n LTQ MS	MATLAB, HMDB database	Identified various metabolic differences in various breast cancer cell types	Wang et al. (2019)
<i>K562 cells</i>	Single-probe	Single-probe	MS/MS	MetaboAnalyst, Geena2	Accessed the effect of drug	Standke et al. (2019)

(Continued)

Table 13.1 Prominent SCM interventions in diverse biological applications. *Continued*

Cell	Isolation	Quenching	Platform	Software	Major finding	References
<i>Circulating tumor cells</i>			Radioluminescence Microscopy (RLM) and Single-Cell Autoradiography (SCAR)		Metastatic breast cancer	Sasportas et al. (2014)
Monitoring of drug action						
<i>Anticancer drug irinotecan effect on HepG2 cell</i>	Microfluidics, robotic liquid handling, and LPME techniques.	Microextraction	MALDI-TOF mass spectrometry	–	Detect, quantify and study metabolomic profile during 7-ethyl-10-hydroxyl-camptothecin (SN-38) formation, GH \$00 ng mL ⁻¹ with the limits of detection (LOD) and quantitation (LOQ) of 2.2 and 4.5 ng mL ⁻¹ for SN-38,	Sun et al. (2020)
<i>BRAFV600E mutant M397 cell</i>	Microchip	Microchip	Full barcode array and molecular probe	MATLAB code, FLOWMAPR R package (version 1.2.0)	Heterogeneity in drug response was observed with distinct changes in metabolism	Su et al. (2020)
<i>HeLa</i>	Live	Live and e ionization solvent lysis	Raman spectroscopy and LSC – MS	R, PDA etc.	Tamoxifen effect within cell with 410 spectra of single cells	Li et al. (2019)
<i>NS-1 cell</i>	Manual dilution	Liquid–liquid extraction and capillary electrophoresis	CE–LIF analysis		Doxorubicin metabolism and detection within cell with sensitivity of sub-nmol per cell	Anderson et al. (2002)
<i>Plasmodium falciparum, Hemiculter leucisculus cell, Paramecium caudatum</i>	Mesh copper grid	Microextraction	MPME/DBDI/MS	QuantAnalysis version 4.3 software	Pyrimethamine detection	Lu et al. (2020)

**FIGURE 13.2**

Single-cell metabolomics analytical modalities.

range of beneficiaries that includes plants, animals and most importantly the human subject.

Undoubtedly, the most remarkable contribution of microbes in pharmaceutical industry is the production of antibiotics, which is basically the product of microbial metabolism. Modulation of the microbial metabolic processes through genetic manipulations have already yielded generation of newer drugs with improved therapeutic potential, thus highlighting the importance of metabolomics introspection in such drug producing microbial population. Vaccine production and generation of bioactive compounds, steroids and immunomodulators are the other important contributions of the microbes in pharmaceutical industry which further strengthens the requirement of metabolomics insights into the drug-producing micro-organisms ([Lancini & Demain, 2013](#)). Glucoamylase, α -amylase, β -galactosidase, cellulases, xylanases, pectinases, glucose oxidase, laccase, catalase, peroxidase, superoxide

dismutase, proteases, lipases, esterases, asparaginase are some of the prominent microbial enzymes which have found diverse industrial applications such as manufacturing of dairy products, beverages, baked products, processed meat and meat products, animal feeds, pulp and paper industry, leather technology, cosmetics products, textile designing and finishing, manufacturing of textile dyes, etc. (Raveendran et al., 2018; Singh et al., 2016). Exploring microbial metabolism can be of enormous importance in industrial waste-management and securing environmental safety as it illuminates the scope of microbes-mediated bioremediation, degradation and scavenging of environmental pollutants or toxicants like heavy metals, hydrocarbons, radioactive wastes, various organic and inorganic pollutants which are usually released through diverse industrial wastes (Kumar & Kundu, 2020). Amidase, amylase, amyloglucosidase, lipase, nitrile, hydratase, manganese peroxidase, lignin peroxidase, oxygenase, protease, laccase, cutinase are some of the reputed microbial enzymes which have found utility in management of various industrial waste materials (Singh et al., 2016). Even microbial metabolic machinery holds the potential to be exploited toward generation of green energy as an alternative to the fossil fuel. Microbial fuel cell technology employing microbial bioremediation and transformation capability to harness electricity from the chemical energy of organic compounds present in the waste materials is such a promising approach toward this direction (Chaturvedi & Verma, 2016). Biofuels such as bioethanol and other alcohols, biodiesel, and methane generated through microbial metabolic processes can serve as sustainable and renewable sources of green energy (Ramanjaneyulu & Reddy, 2019). Lipases, proteases, laccase, pectinases, xylanases, cellobiohydrolase or exoglucanase, endoglucanase, and β -glucosidase are crucial microbial enzymes which have been employed for biofuel production from diverse industrial waste materials (Srivastava et al., 2020). Further, microbial biogeochemical cycles are indispensable for the conservation of biosphere and ecosystem as the metabolism of soil microbes is a crucial determinant for improving the soil health and maintenance of soil-plant-microbe interactions (Sahu et al., 2017). So, maneuvering the microbial metabolic processes can definitely extend significant benefit but at the same time, the great extent of heterogeneity displayed by the same microbial species and even observed within a monoclonal microbial population should be dealt adeptly for successful and consistent outcome (Calabrese et al., 2019). The heterogeneity may develop through acquired or inherent genetic or epigenetic attributes yielding variation in transcript or protein expression with ultimate manifestation in terms of qualitative or quantitative deviation in metabolite production among the microbes of the same colony or the same species under the same condition. Single-cell techniques can only gauge the extent of such alteration and SCM is probably the best to sought vivid depiction of the scenario as it directly enumerates the metabolite production in individual microbes under precise micro-environment and at a temporal framework. For instance, the metabolomic heterogeneity in different individual yeast cells grown under the same culture condition

have been enumerated through nESI-MS which have elucidated marked cell to cell variations in metabolite production involving several yeast metabolic pathways such as saccharides metabolism (hexose phosphate and hexose), metabolism of amino sugars (uridinediphosphate N-acetyl-hexosamineanduridinediphosphate hexose), aerobic respiration (succinic acid to citric acid), energy metabolism (ATP and AMP), free radical scavenging (GSH and GSSG), etc. Further, the heterogeneity in metabolism have also been depicted to be persistent under exogenous chemical stimuli induced byiodoacetamide treatment which acts as a protein inhibitor by binding with the sulphydryl group of enzymes involved in diverse metabolic pathways like glyceraldehyde 3-phosphate dehydrogenase enzyme of glycolytic pathway. Not only to observe yeast metabolism but also the analytical modality has been found to be equally effective in SCM analysis of other cell-wall containing unicellular microbial species such as *Chlamydomonas*, *Euglena*, *Dunaliella*, etc. as the individual variation in metabolic changes among the *Chlamydomonas reinhardtii* cells under light and nitrogen deprived condition have been clearly elucidated by employing the SCM modality (Li et al., 2020). Even, the correlation of metabolic rate with the cell size or volume can also be enumerated through SCM analysis as differential metabolomic profile has been successfully depicted in aquatic phytoplankton of different cell volume ranging between 1 and $10^3 \mu\text{m}^3$ by employing nano-scale secondary ion mass spectrometry (NanoSIMS) (Zaoli et al., 2019). Similarly, SCM-based monitoring of the ATP dynamicity as stimuli-induced response in microbes on single-cell basis can serve as an important indicator of cellular energy metabolism and in-tern cellular physiological status as FRET-based ATP biosensor coupled with microfluidic platform has been successfully employed for ATP tracking in antibiotic treated individual *Mycobacterium smegmatis* cells, vividly depicting the antimicrobial efficacy and mechanism of action, even better than conventional propidium iodide staining in a much sensitive and reliable manner (Maglica et al., 2015). Evidently, SCM approach can be helpful to introspect the mechanisms of acquiring MDR or XDR attributes among bacterial populations and assist to improvise better therapeutic strategies against such tricky microbes. Influence of gut microbiota in host metabolism is another crucial aspect of potential SCM intervention which extends the scope for proper dietary supplementation to improve the human as well as animal health. Single cell analysis of different gut microbes and their responses to dietary additives through modulation of metabolic pathways yielding altered metabolite profile can enrich the knowledge toward better utilization of fiber diet in the gut of respective hosts (Chijiwa et al., 2020). SCM-based identification of specific dietary additive responder microbes with precise functional attributes can be useful for dietary manipulation to manage certain human lifestyle-related enteric diseases like inflammatory bowel disease (Berry et al., 2014; Minakshi et al., 2020). Raman-DIP (Deuterium Isotope Probing) has emerged as one of the prominent nondestructive SCM modality for introspecting the metabolic machinery of such culture-independent gut microbes (Xu et al., 2017). Beyond the human subject,

gut microbial manipulation and dietary supplementation approach can be excellent to achieve production optima from the livestock animals too through proper nutritional management.

Single-cell metabolomics in plant science and agriculture

Plant metabolite repository is extremely vast and complex, yet deciphers plethora of valuable information crucial to enumerate plant growth and development, response and tolerance to the environmental stressors, identification of biomarkers for disease resistance and improved crop quality as well as quantity, tracing the geographical origin of the plants, which is a crucial determinant of the quality of herbal compounds and their safety status, evaluation of seed quality, analyses of plant-derived medicinal compounds, elucidation of plant functional genomics based upon the metabolite phenotype, thus holding significant impact in crop science and herbal medicine ([Sousa Silva et al., 2019](#)). Despite the diversified applications, plant metabolomics introspection usually suffers from the dilution or population effect due to sampling from organ or tissue possessing heterogeneous cell populations usually encountered in SCM analyses of all the multicellular organisms. Population-based metabolomics can be suitable to counter the stochastic biological effects but the deterministic metabolic effects which may arise from the mutants or underlies the basis for synthesis or overproduction of certain specialized metabolites usually gets overshadowed. The SCM analysis has paved the path for elucidating such deterministic effects in plant physiology and metabolism to exploit them toward improved production. As mentioned earlier that SCM is a late-mover than the other system biology single-cell omics techniques and thus bulk of the SCM introspection involving the plants is still carried-out toward proof-of-concept works and to establish the efficiency of the employed SCM analytical modality. For instance, probe ESI mass spectrometry (PESI-MS) has been successfully detected several metabolites including fructans, lipids, and flavone derivatives in single *Allium cepa* cells along with elucidating considerable metabolite diversity among diverse cells types of *A. cepa* bulb as well as metabolite profile of different subcellular components within the same cell. The introspection also depicted that the inner epidermal cells are many-fold richer in fructans as compared to the outer epidermal cells while the later ones are richer in lipid content. So, the observations vividly established the efficiency of PESI-MS platform in SCM analysis by depicting the cellular heterogeneity of metabolites in *A. cepa* cells and subcellular components ([Gong et al., 2014](#)). The same analytical modality has precisely elucidated the cellular heterogeneity of metabolite expression between the adjacent stalk and glandular cells of the same trichome unit in *Solanum lycopersicum L.* plants through SCM analysis. Significant metabolite diversity among different trichomes has also been unveiled. Diverse metabolites including flavonoids, organic acids, carbohydrates and amino acids have been

detected from the picoliter quantity of cell sap. Such cell-to-cell metabolic variations may illuminate its correlation with anatomical, physiological and stimulus-responsive functional attributes of the plants (Nakashima et al., 2016). Similarly synchronized polarization induced electrospray ionization method has been successfully employed for SCM analysis of *A. cepa* cells and PC-12 cells. The modality has the capability to analyze the samples simultaneously in both positive as well as negative ion mode which can enhance the coverage of metabolite identification. Several molecules including lipids, peptides, proteins and metabolites have been identified from single epidermis cells of *A. cepa* by using this modality thus establishing the aptness of the method in SCM analysis (Hu et al., 2016). Live single-cell video MS which directly infuses the aspirated sample into the MS ionization source with minimum prior handling and MSI have been employed to identify the cell-specific localization terpenoid indole alkaloids in the *Catharanthus roseus* stem tissue. The medicinal plant is well-known as the source of several terpenoid indole alkaloids including anticancer drugs vinblastine and vincristine. The SCM introspection has elucidated idioblast and laticifer cells of the plant stem tissue as the predominant site of synthesis and accretion of the alkaloids of therapeutic importance (Yamamoto et al., 2016). Elucidation of metabolic signatures employing matrix-free LDI-MS approach has facilitated taxonomic identification of planktons, which otherwise carried out by morphological analysis demanding comprehensive expertise (Baumeister et al., 2020). The aforementioned examples of SCM intervention in plant science and agriculture are conclusive enough to decipher steady transition of these high-end techniques from proof-of-concept work toward on-field applications.

Diversified animal applications

Animal cells are extremely heterogeneous, which render them interesting as well as tricky targets for SCM introspection, most commonly to elucidate the dynamics of metabolic profile in response to diverse micro-physiological stimuli. The response in terms of metabolic alterations can be elicited through propagation of nerve impulse, differentiation, cell growth and proliferation, intra and transcellular communication, stress due to environmental factors, production or reproduction, adaptation, disease establishment, progression and their therapeutic management, immune modulation, and so many other factors, which ultimately change the metabolite phenotype of individual cells (Evers et al., 2019). Such metabolite variations can be exploited as the signatures of specific pathophysiological conditions for potential application as disease-specific biomarkers, druggable checkpoints, identifier of stress response and tools for monitoring the developmental biology aspects. For instance, CE-ESI-MS platform has been employed for SCM analysis of different neuronal types from *Aplysia californica* which identified the common as well as distinct metabolites present in different

neuronal cell types according to their physiological significance (Nemes et al., 2011). The metabolite diversity between fresh and cultured neuronal cells from *A. californica* has been elucidated by the same analytical modality depicting the regulation in cellular metabolic responses to the changing environment (Nemes et al., 2012). NanoSIMS along with stable isotope labeling and multiisotope imaging to monitor the incorporation of ¹⁵N labeled thymidine in the genomic material of young transgenic C57Bl/6 mice has been employed to elucidate the mechanism of cardiomyocyte replacement during normal aging as well as in myocardial injury where preexisting cardiomyocytes serve as the primordial source material. The single-cell based introspection has illuminated the pivotal mechanism of myocardial homeostasis (Senyo et al., 2012). Nano-DESI-MS based SCM approach has facilitated high-throughput profiling and quantification of several amino acids, phospholipids and metabolites in cheek cells from buccal swab samples. The amount of phosphatidylcholine in a single cell has been enumerated to be 1.2 picomoles. Such SCM introspection bears a futuristic value for evaluation of drug effects, cell differentiation and other stimuli-response analyses (Bergman & Lanekoff, 2017). Cellular heterogeneity in drug metabolism has also been explored through SCM analysis of tafluprost-treated human hepatocytes by nano-ESI-MS modality. Tafluprost which is commonly employed for the treatment of glaucoma has observed to be metabolized differentially in individual hepatic cells to yield distinct metabolite profile in terms of tafluprost acid and its beta oxidation metabolite dinor-tafluprost acid production. Such kind of study extends the scope of SCM analysis for evaluation of drug pharmacokinetics and toxicity analysis (Fukano et al., 2012). Similarly SCM introspection of anticancer drug treated individual HeLa cells employing single-probe MS have elucidated several cellular metabolites such as AMP, ADP, ATP, phosphatidylcholines, sphingomyleins, diglycerides, and triglycerides along with different drug molecules like doxorubicin, paclitaxel, and OSW-1 and their derivatives such as hydroxylpaclitaxel, doxorubicinol, and deoxyadriamycinone, thus reinforcing the utility of SCM analysis in drug efficiency and toxicity evaluation (Pan et al., 2014). Further, nanoSIMS-based dual imaging method has efficiently documented the internalization, distribution and nucleolar targeting of ¹⁵N-labeled polynuclear Platinum antitumor drug Triplatinin individual human breast adenocarcinoma (MCF7) cells (Wedlock et al., 2013). Gas chromatography coupled with MS platform has also been used for single-cell analysis of the volatile organic compounds generated from the p53 mutant lung cancer cells depicting a distinct metabolite profile in comparison with the healthy lung cells thus providing the way for customization of future diagnostic strategy (Serasanambati et al., 2019). Capillary micro-sampling ESI MS coupled with ion mobility separation method has successfully identified several metabolites and lipids from individual single human hepatocytes (HepG2/C3A). The modality has been employed for elucidation of cellular response to metabolic modulators by enumerating the adenylate energy charge levels in control and rotenone treated hepatocytes. Depletion in ATP and GTP level along with the accretion of AMP and GMP has been observed in

rotenone-treated cells. The ratio of reduced glutathione/oxidized glutathione diminished considerably while uridinediphosphate N-acetylhexosamine/uridinediphosphate hexose ratio increased in the rotenone-treated cells which are indicative of rotenone-induced oxidative stress and corresponding cellular metabolic responses by the hepatic cells. Further, the elevated median value and broadening range of the stress markers in rotenone-treated cells depicted cellular heterogeneity in oxidative stress response by the hepatic cells (Zhang & Vertes, 2015). Probe electrospray ionization (PESI) coupled with triple quadrupole MS/MS has identified several metabolites such as amino acids, organic acids, and sugars in mice hepatocytes. The platform has also successfully identified the distinct metabolite profile between healthy and CCl₄-treated mice hepatocytes. Further, the method has also facilitated the scope for in-vivo real time metabolic analysis of integral liver cells from living mouse as it captured the alteration in α-ketoglutaric acid and fumaric acid as key tricarboxylic acid (TCA) cycle intermediates following pyruvic acid injection through portal vein (Zaitzu et al., 2016). Repeated enrichment by Ion Trap MS is an advanced variant of MS which help in identifying low abundance metabolites which is generally not recognized by the normal MS, further analysis in both positive and negative ion mode can enhance the range of spectral acquisition. Several metabolites including 5-methylcytosine, caffeine and ATP have been identified from MCF7 cell by this technique (Si et al., 2017). Various non-MS based platforms also carry potential for SCM analysis such as genetically encoded nano-sensors have found their applications in high-resolution imaging of brain energy metabolism. FLIP series for glucose, FLIPE/ GluSnFR for glutamate, ATeam and PercevalHR for ATP, Perodox and FREX for NADH/NAD⁺, Laconic for lactate, Pyronic for pyruvate are some of the nano-sensors applied for cellular metabolite enumeration hold the potential for SCM introspection in diverse cell types (San Martín et al., 2014). Similarly, a multicolor fluorescence detection-based microfluidic device (MFD-MD) has been employed for targeted SCM investigation of acute ethanol-stimulated mice liver cells which depicted elevation in hydrogen peroxide and reduction in glutathione and cysteine level in response to ethanol treatment (Li et al., 2016). Despite considerable robustness of SCM analytical platforms and reliability of the generated data, integration with other single-cell techniques such as single-cell RNA-seq, single-cell proteomics, etc. can facilitate simulation-based metabolic modeling and upscaling of SCM research in diversified biological arena (Zhang et al., 2020). An integrative approach, simultaneously employing electrical cell impedance assay for enumerating cell contractions, FRET-based single cell metabolite analysis, cell membrane permeabilization assay and western blotting for protein analysis along with deep learning for high-throughput data analysis has elucidated that intra-cytosolic release of actin-bound glycolytic enzyme aldolase A through RhoA/ROCK-1 dependent actin reorganization enhances glycolysis leading to contraction of endothelial cells during septic shock which underlines the molecular basis of inflammation and organ edema at such clinical conditions (Wu et al., 2019). The abovementioned examples are some of

the pioneering SCM interventions involving higher order animals including humans which emphasized the extensive scope of these state-of-the-art systemic biology tools in animal sciences, several more prominent and precise applications will be delineated under suitable subsections.

Single-cell metabolomics in developmental biology

Developmental biology is one of the fascinating arenas where SCM has already proved pivotal in unveiling the phenomena and molecular detail of the surrounding micro-milieu essential for the embryo to organ and organism development. SCM introspection to elucidate the factors introducing cellular heterogeneity and differential responses of diverse cellular layers to environmental conditions such as gravity, rotation, fluid flow, temperature, light etc. along with interaction with the extracellular matrix can yield plethora of information pivotal for development of embryo to full length healthy organism. Such introspection can decipher the genetic, epigenetic, cellular and molecular influences in solo or integrative manner over the acquisition of heterogeneous phenotypic attributes of diverse cells, organs or organisms. For instance, exploration of subcellular proteo-metabolic detail of *Xenopus laevis* embryos using the in-vivo high-resolution MS (HRMS) as a single analytical modality has unveiled crucial molecular interactions for improved apprehension of system biology events and developmental processes ([Lombard-Banek et al., 2021](#)). Regenerative and personalized medicine is the other domain where SCM intervention in developmental biology perspective carries enormous potential. Elucidation of the molecular events propelling the structural and cellular heterogeneity is primordial to operate successful organogenesis. Precise knowledge of directional cellular differentiation toward organ development at transcriptomic, proteomic and metabolomic levels complemented by the epigenetic influences needs to be comprehensively explored through multiomics integration which may pave the way for commercial exploitation of organogenesis to usher optimism in organ failure patients ([Clair, 2019](#); [Lancaster & Knoblich, 2014](#)). Thus, SCM intervention is evident in stem cell biology as embryonic stem cells and induced pluripotent stem cells serve as the major sources for organoid development. Even, SCM can be effective in novel reproductive biology applications such as providing the scope for extra-uterine development of embryos, particularly for the wildlife species which requires conservation; and development of artificial womb to support highly premature fetus is certainly a step toward this direction ([Partridge et al., 2017](#)). Even such system can be effective to sustain premature human infants too. In fact, the evidences of SCM intervention focusing the embryonic developmental processes, particularly in the small model animals are progressively enriching the experimental space currently. For instance, metabolomics profiling of 16-cell embryo of *Xenopus laevis* recognized nearly 40 metabolites through single-cell capillary electrophoresis-electrospray ionization

MS. These metabolites were integral components of various central metabolic pathways regulating the developmental processes and differentially expressed according to cellular heterogeneity. The spatiotemporal introspection enlightened that how metabolite variations in embryonic cells can lead to altered differentiation and migration pattern during early embryonic development. The detection threshold for the metabolites was recorded in the range of less than 10nM or 60attomole (Onjiko et al., 2015). The sensitivity of such MS-based metabolite detection can be further enhanced by capturing the MS spectra in both positive and negative ion mode. The dual cationic-anionic profiling approach enabled detection of over 450 molecular features with 84 identified metabolites in the 50–500 m z range⁻¹ from single cell of *X. laevis* embryo employing the same platform (Portero & Nemes, 2019). Similarly, MS-based subcellular metabolite and lipid analyses in different stages of embryo and egg to elucidate the molecular events of fertilization and early embryonic development in *X. laevis* revealed several metabolites and unique proteins important for various developmental networks. Further, the introspection shaded light over the metabolic changes occurring during transition from unfertilized to fertilized egg and ultimately to 32 cell embryo. The differential expression of metabolites in unfertilized and fertilized eggs, vegetal poles of the eggs and in respective animal has also been unfurled through this LAESI-based direct MS study (Shrestha et al., 2014). Such understanding of metabolomic remodeling during early embryogenesis is not only very helpful to apprehend developmental processes but also opens the window to exert exogenous control over the development process toward organogenesis which in turn may impart significant impact in the field of regenerative medicine and drug development to tackle several developmental diseases. Further, SCM introspection also provides the scope for deducing biomarkers of developmental disorders as well as markers of pharmacological efficiency or toxicity (Cezar et al., 2007). Not only prenatal development but also postnatal developmental issues such as reduction in pancreatic β cell proliferative capacity yielding faulty glucose homeostasis and the underlying molecular signatures of such aberration in terms of upregulated amino acid metabolism, mitochondrial activity, and nutrient responsive transcription factors can also be elucidated through SCM analysis (Zeng et al., 2017).

Single-cell metabolomics in aging and senescence study

Aging study is another area where SCM holds potential to solve some long-lasting mystery. Aging and regeneration capacity of cells vary in terms of organs and cell types. Strengthening of the knowledge database regarding the associated factors and their regulations will not only help to delay the aging process but also assist in curing the aging related disorders. This may also pay in terms of realizing how aging affects other disease outcomes such as viral or microbial attack

and the extent of host efficiency to combat them through age dependent immune responses. Several experimental evidences have vividly established strong linkage between metabolic alterations and aging, ultimately affecting the biological outcomes like longevity and senescence. Identification of such aging-related metabolic signatures at cellular and organ level may facilitate early assessment of biological fitness and customization of life-extending strategies targeting the metabolic networks (López-Otín et al., 2016; Robinson et al., 2020). Initial study in yeast depicted that considerable extent of gene regulation may take place with the passage of generation; even the expression of around 20% of genes may get altered after 4–6 generations along with significant alteration in metabolism. Further, the biosynthetic pathways were decelerated considerably whereas the catabolic pathways were accelerated significantly affecting the longevity and promoting senescence after 11th generation in budding yeast. This kind of introspection using SCM modalities integrated with other single-cell omics techniques may be helpful in designing strategy or drugs to exert metabolic control to combat the aging-related diseases. Even the complications of secondary infections or associated ailments can be handled in a better way through targeted metabolite maneuvering (He et al., 2020; Kamei et al., 2014). Stem cell based introspection of senescence-associated metabolic alterations have clearly established their reproducibility to exploit them for monitoring the status of cellular aging (Fernandez-Rebollo et al., 2020; Schüller et al., 2020).

Single-cell metabolomics in stem cell biology

Insinuation regarding the utility of SCM introspection in stem cell biology has been outlined in developmental biology section, yet it needs further elaboration. Regulation of metabolomic machineries of the stem cells may facilitate exploitation of the stemness ability of these cells for regenerative medicine development (Goodarzi et al., 2019). Comprehensive knowledge in this perspective will not only help to control the pluripotency or multipotent behavior of the stem cells by controlling the biochemical environment but also give insight about signaling and metabolic pathway alterations during activation and transformation of the stem cells into other cell lines. Even, the earlier concept regarding metabolic alterations as the consequences of stem cell differentiation events has entirely changed now. Currently, metabolic alterations are perceived to be deterministic in cell fate regulation; every such modulation exerts signals to decide proliferation, differentiation, quiescence or senescence (Shyh-Chang et al., 2013). SCM introspection of cancer stem cells have special value in oncology as the molecular signals either arising from or regulating the tumor development, metastasis or recurrence can be traced for early cancer diagnosis and customization of effective therapeutic intervention (Sun & Yang, 2018a). For instance, single-probe MS-based introspection has revealed distinct metabolic profile in colorectal cancer stem cells from nonstem cancer cells as the former was found to be richer in TCA cycle

metabolites which pointed toward difference in energy metabolic pathways between the two cancer cell types. Further, unsaturated lipid contents were higher in the cancer stem cells which can be down-regulated by several metabolic enzyme and pathway modulation that ultimately leads to reduction in stemness of the cancer cells (Sun & Yang, 2018a). Such investigation facilitates novel biomarker and therapeutic checkpoint identification. Selection of suitable platform is also an essential factor to obtain wider metabolic profiling range as evident by the fact that only 60 metabolites were identified from 10,000 cells when hematopoietic stem cells (HSCs) were isolated by Flow cytometry and paramagnetic MicroBeads followed by vortexing with 80% methanol before analysis by LC-MS/MS (Agathocleous et al., 2017) while 160 metabolites were identified from 10,000 HSCs through hydrophilic liquid interaction chromatography and high-sensitivity orbitrap MS-based modality (DeVilbiss et al., 2020). The former study further reported Vitamin C dependent reduction in HSC function and myelopoiesis leading to suppression of leukaemogenesis (Agathocleous et al., 2017). The later introspection depicted higher glycerophospholipid metabolites in HSCs in comparison to unfractionated bone marrow cells. Alteration in purine biosynthesis has also been observed in HSCs by methotrexate treatment. Purine synthesis has also been found to be curtailed during metastasis as evidenced by lower purine intermediates in circulating melanoma cells than subcutaneous tumors (DeVilbiss et al., 2020). Thus it is quite clear that SCM introspection certainly provides important clues regarding the cancer status and stem cell differentiation which is having diagnostic, therapeutic and prognostic significance.

Tissue stem cells are crucial determinants of aging and aging-related disorders. Several factors and interlinked pathways are influencing the stemness property of these cells such as accumulation of ROS hinders the renewal of the stem cells and progressively propels them toward aging ultimately affecting the tissue and organ integrity leading to disease pathogenesis (Rossi et al., 2008). Likewise, the orchestrated mechanism of reduction in tissue stem cells accompanied by curtailing of telomerase function and mitochondrial dysfunction inflicting aging and aging-related disorders leads to diverse metabolic consequences. Even, genotoxic signaling along with certain metabolic responses can also stimulate the aging process. Tracing of such metabolic alterations or stimulators through SCM introspection may facilitate identification of biomarkers of aging-related disorders as well as customization of metabolic maneuvering strategies to combat them (Sahin & DePinho, 2010). For instance, noninvasive single-cell Raman microspectroscopy has been successfully employed to differentiate between human induced pluripotent stem cells (hiPSCs) and hiPSC-derived neural cells based on the cellular biochemical profile and glycogen has been depicted as the biomarker of neuronal differentiation (Hsu et al., 2020). Several signaling pathways such as insulin-PI3K, Akt-FOXO, mTOR and AMPK pathways are found to be closely associated with stem cell-based balancing of aging, quiescence and proliferation processes (Chen et al., 2009; Jasper & Jones, 2010; Kalaitzidis et al., 2012; Kharas et al., 2009; Magee et al., 2012).

Single-cell metabolomics in functional genomics

Central dogma of life illuminates the dissemination of the functions harbored in the genomic DNA yielding cognate mRNA expression which transcribes to generate a protein repository performing as the working machinery at cellular level. Although the central dogma conspicuously omits metabolites but they are probably the most nascent phenotypic outcome of the effected functional signals from central dogma through transcriptomic and proteomic modulations at cellular and subcellular level (De Lorenzo, 2014). As spatiotemporal metabolite dynamicity and flux through metabolic pathways are encrypted in genomic DNA, thus metabolites are functional extension of central dogma in true sense. Thus, functional genomics study reflects how the metabolic profile of cellular microenvironment is spatiotemporally regulated through alteration in genomic DNA such as mutation in DNA, gene editing, transcriptional and posttranscriptional modulations, translational and posttranslational modifications, etc. The influence of functional genomics is usually exerted through various pathways modulation either in solo or often in a concerted manner through several interconnected pathway network. Any multicellular organism contains diversified cellular clusters depending upon the organ and tissue types; even a single tissue also possesses heterogeneous cell population which regulates every metabolic pathway differently under specific stimulus encrypted by the genomic DNA of each individual cells. Thus apparently similar cellular cluster responds to any particular stimuli with differential metabolite profile within each of the individual cells at a particular temporal point. Such subtle but consistent metabolite diversity which is usually masked in population metabolomics introspection can be elucidated through SCM analysis. As every individual cell is functionally directed by its own genomic DNA resulting into differential SCM phenotype, thus the actual influence of functional genomics can be most precisely enumerated through SCM analysis. But SCM must be complemented with other single cell techniques for elucidating the relationship between metabolomics and functional genomics comprehensively at a precise cellular context. For example, metabolism in the endothelial cells is modulated through transcriptomic regulations during pathological angiogenesis and the metabolic transcriptome heterogeneity can be enumerated by single-cell transcriptomic profiling of the individual endothelial cells which enables identification of crucial metabolic targets for antiangiogenic therapy (Rohlenova et al., 2020). Enumeration of stress responses through SCM analyses and establishment of genomic correlation may enable selecting animals and plants having better genetic make up to withstand various stressful conditions. Such introspection carries commercial value as better adaptability may reflect with less-affected production under stressful conditions. Functional genomic approaches can be used for metabolic engineering to produce specific target metabolites from plants or animals. For instance, jasmonate-induced genetic reprogramming of metabolism in *Nicotiana tabacum* L. cv. BY-2 cells has been demonstrated by cDNA-amplified

fragment length polymorphism-based transcript profiling followed by GC-MS analysis of alkaloids (Goossens et al., 2003). The application can be further extended for predicting the function of unknown genes after editing it in single cells and observing the phenotypic metabolic outcome (Tyagi et al., 2010). Likewise functional genomics-based metabolomics approach has enabled identification of several novel secondary metabolites from cyanobacteria (Baran et al., 2013). Several derivatives of histidine-betaine, and other uncommon oligosaccharides were identified by MS which ultimately detected 264 metabolites in cyanobacterium *Synechococcus* sp. PCC 7002 cells. Metabolite expression varied with different strains indicating genomic role behind the metabolite expression and regulation. Similarly ¹H-NMR spectroscopy-based metabolomics analysis of yeast cells have been performed to elucidate the function of apparently silent or unstudied genes exploiting comparative metabolite variation in different mutant variants. The mutant variants depicting similar growth rate phenotypes differed significantly in different glycolytic metabolite concentrations and ATP/ADP ratio, thus tracing the functions of the deleted apparently silent genes as well as their point of interaction in the cognate metabolic pathway (Raamsdonk et al., 2001). Thus the aforesaid examples are enough to confer that SCM introspection certainly has the potential to elucidate the functional genomics implications in cellular context.

Single-cell metabolomics in nutrition research

Supply of nutrients is one of the primary determinants influencing cellular metabolism, metabolite turnover and flux through diverse metabolic pathways essential for plethora of cellular functions. The fundamental cellular functions of dietary metabolites include supply of essential constituents for the synthesis of cellular macromolecules and structural components, delivery of energy sources and factors implicated in energy yielding pathways, and metabolites influencing signaling pathways (Astarita & Langridge, 2013). Moreover, some of the dietary metabolites performed specialized functions such as acting as antioxidants and immuno-modulators; for instance, vitamin A, vitamin E, and vitamin C are acting as antioxidants to influence free radical-mediated oxidative stress pathways and diverse cellular redox reactions (Astarita & Langridge, 2013). So, dietary antioxidant metabolites are evidently effective against different diseases inflicted by oxidative stress such as cancer, aging, neurodegenerative disorders, inflammation, cardiovascular diseases, etc. Thus, SCM introspection of the cellular metabolism under such stress response and the alleviating effect of nutritional maneuvering on cell to cell basis may add significant value in precision nutrition (Jones et al., 2012; Tebani & Bekri, 2019). Likewise, nutritional status is strongly correlated with cellular metabolism and immune cell function. Under-nutrition often leads to immunosuppression. Prebiotics, probiotics, micro-nutrients and several other food-derived metabolites are well documented to provide localized gut immunity

or generalized immunity and regulate crucial signaling functions mediated by the immune cells. Thus nutritional perspective of several autoimmune diseases and immune-mediated disorders can be explored through SCM introspection of the immune cell metabolism which may provide dietary metabolite-based therapeutic strategies to combat such diseases (Alwarawrah et al., 2018). Further, dietary metabolite-based stress management in animal as well as in plants may have commercial value to restore production under stressful conditions.

Single-cell metabolomics in environmental biology

Environmental determinants such as temperature, humidity, aeration, light color, intensity, radiation, pollutants and toxicants, nutrient adequacy, etc. are pivotal regulators of health, disease, production and reproduction status of plants as well as animals (Kumar, & Ghosh, Kumar, et al., 2020; Svatoš et al., 2020). Considerable focus has been extended toward evaluating the effects of biotic and abiotic stresses on animal and plant production system under the current context of climate change and global warming. The quest for better suitable animal and plant varieties which can withstand and grow optimally under specific environmental stress or development of potent all-weather suitable type variety is one of the focal research themes in the ongoing era. As spatiotemporal metabolite profile is more dynamic than the genomic, transcriptomic or even proteomic alterations and represents the most immediate cellular phenotype (Zenobi, 2013); evidently, SCM introspection can identify the cellular effects of environmental factors with utmost sensitivity and thus helpful in identifying such desirable resistant variety of organisms along with functional genomics analyses. For instance, the metabolite dynamicity in individual algal cell under environmental modulations such as variable light intensity and nitrogen perturbation has been enumerated through nano-ESI based single-probe MS approach (Sun et al., 2018b). The combined effect of heat stress and expulsion on the metabolic profile of hard coral algal symbionts has been explored through single-cell approach which revealed an opposing metabolic response, however, the combined effect was found to be additive (Petrou et al., 2018). The introspection not only depicted the metabolic variations in heat stress response, but also emphasized over the importance of illuminating cumulative effects of multiple stressors on biological system in an integrated manner. Differential metabolite profile due to seasonal variations has been depicted in genetically modified maize using ¹H-NMR approach, although the study lacks single-cell level penetration (Barros et al., 2010). Likewise, the content and types of active principles in various parts of the medicinal plants are significantly influenced by the environmental attributes which can be elucidated by SCM analyses of specific types of plant cells and tissues. SCM approach has also found to be useful in environmental microbiology study as microbial interaction with surrounding mineralogical contents, gaseous components of volcanic eruption including carbon dioxide, sulfur dioxide, water vapor, methane, etc. as

well as interaction among the adjacent microbial communities has been successfully introspected in volcanic fumarole sediments (Marlow et al., 2020). Further, SCM introspection of microbial metabolism and its purposeful modulation toward industrial waste-management and heavy-metal bioremediation can be of immense value in the context of ensuring environmental safety and mitigation of public health hazard from such toxic chemicals (Cui et al., 2021; Kumar & Kundu, 2020). Besides the microbes, plants and their rhizospheric microbes mediated phytoremediation of organic pollutants and heavy-metals can also be exploited through SCM approach toward improved soil health and eco-friendly environment. (Khatiwada et al., 2020; Ojuederie & Babalola, 2017).

Single-cell metabolomics in system biology

The conventional system for elucidating the biology of multicellular organisms usually follows the path of analyzing the molecular components such as metabolites, proteins, RNA transcripts or genes regulating them at tissue, organ or organism level. However, the cellular heterogeneity is masked due to population effect in such type of introspection and the deterministic effects of low-abundance molecules often remain obscure. The generated incomplete data-sets of diverse cellular components in population study leave several gaps to integrate them for system biology application. The focal theme of system biology is elucidation of the complex biological processes by integrating the spatiotemporal profile and dynamicity of all the interacting cellular components at a predefined subcellular, cellular, tissue, organ or organism level set-up and deciphering the knowledge to customize a biologically relevant functional network or computational model of the entire system. Combining the SCM along with the other single-cell techniques have the potential to capture all the interacting chemical features and their spatiotemporal dynamicity at a precise cellular or subcellular context, thus rendering these state-of-the-art techniques useful for system biology applications (Libault et al., 2017). Further, the precise metabolic phenotypes of some specific cell types present in the distinct tissues or organs executing dedicated functions apart from normal homeostasis can only be unmasked through SCM analyses which can be extremely beneficial for deciphering the interacting biomolecules and signaling networks associated with their unique functioning. SCM-based elucidation of all the metabolic features arising at the cellular level within a given timeframe under specific patho-physiological stimulus such as differentiation, growth, multiplication, apoptosis, signaling, stress, aging, infection, cancer, etc. ultimately reflected as the consolidated phenotypic response by different tissues or the entire organ can be helpful for the generation of reliable system biology model for organ system study, stress response analyses, disease modeling, metabolic pathway mapping for identification of therapeutic checkpoints, drug discovery and toxicity analysis, evaluating the impact of environmental determinants, etc. The variable susceptibility of diverse cell types in a bulk tissue

or organ toward drugs, toxicants and pollutants can also be revealed by SCM analysis. Single-cell type metabolomic integrated with transcriptomics analysis has been found several applications in plant science such as the metabolomic features of different types of cotton fibers have been compared during fiber elongation and cell wall differentiation (Tuttle et al., 2015). The similar approach has elucidated distinct metabolic attributes in the border cells of *Medicago truncatula* roots producing antimicrobial metabolites such as flavonoids (7, 4'-dihydroxyflavone) which are implicated in plant-microbe interaction, plant defense and drawing of symbiotic microbes (Watson et al., 2015). Metabolic introspection in chicken astrocyte revealed inhibition of pyruvate generation from both glucose and glycogen by iodoacetate leads to inhibition of memory process besides elucidating a dose-dependent memory inhibition by 1,4-dideoxy-1,4-imino-D-arabinitol (DAB) due to precise inhibition of glycogenolysis. Thus both the metabolic pathways: glycolysis and glycogenolysis are crucial for memory process. Further, glycogenolysis also found to supply glutamine to the neurons serving as the precursor of neuronal glutamate and GABA (Gibbs et al., 2006). Though the study was conducted on cell culture, but similar type of SCM introspection can elucidate metabolic underpinning of diverse physiological processes along with deciphering the importance of cellular heterogeneity in respective context. Experiment involving astrocytes from rat hippocampus has depicted crucial role of astrocytic lactate transporters for long term memory besides demonstrating the importance of glycolysis and glycogenolysis in memory formation (Suzuki et al., 2011). The metabolic interrelationship between different cell types of any organ is also stringently regulated; it may be cooperative or competitive. For instance, NMR-based metabolic introspection of juxtaposed astrocytes and neurons revealed the competition for glucose and lactate uptake as oxidative energy substrates between the two kinds of cells within brain. It was also evident that astrocytes preferentially utilize glucose while neurons are fond of lactate as energy substrate and astrocytes have less active oxidative metabolism than neurons (Bouzier-Sore et al., 2006). Such metabolomics profiling has been successfully effected in identifying several novel neurohormones from invertebrates like *A. californica*, *Lymnaea stagnalis*, *Periplaneta americana*, and *Cherax destructor* (Garden et al., 1996; Jiménez et al., 1998; Jiménez et al., 2008; Li et al., 2000; Neupert & Predel, 2005; Skiebe et al., 2002). So, the above discussion comprehensively justifies the importance of SCM analysis in system biology introspections.

Single-cell metabolomics in immunology

Immune system presents an adorable exhibition of diverse immune cell types having precise structural and functional contributions; even cellular heterogeneity within a specific cell type is enormous which contributes distinctly in cellular cross-talking during normal development as well as in disease-induced immune response. The immense cellular diversity is obviously reflected through diverse metabolic phenotypes which are spatiotemporally modulated by specific immune

response either as the outcome of a stimulated metabolic pathway network or as a determinant regulator of a downstream signaling process, mutually or exclusively. Derangement in such stringently regulated immunometabolic network leads to onset of immune-mediated disorders such as cancer, diabetes, obesity, atherosclerosis, rheumatoid arthritis, etc. SCM-based metabolic profiling of individual heterogeneous immune cells can overcome the limitation of bulk studies to reveal the biological implications of every immunometabolic modulations occurring at single-cell level in the context of disease establishment and progression. Further, the drastic variations in immunometabolism under in-vitro and in-vivo condition and wide inter-species variability also advocate for SCM introspection in immunology applications (Artyomov & Van den Bossche, 2020). In fact, comprehensive analyses by integrating SCM with other single-cell techniques can be more informative; for instance introspecting the expression of RNA-transcripts of key metabolic genes by single-cell RNA sequencing or quantitative proteomics analysis of the crucial metabolic enzymes under precise patho-physiological response or the metabolic response and pathway modulation arise from antibody-based immune-checkpoint blockade therapy can be of immense value in clinical immunology. Various analytical modalities have been exploited to introspect the spatio-temporal metabolic dynamicity of immune cells at single-cell level and tried to link them with immune functioning under normal or disease state to find diagnostic and therapeutic implications. For instance, mass cytometry by time of flight along with multiplexed ion beam imaging by time of flight has been employed to elucidate immunometabolic profile of human cytotoxic T cells with the help of 41 selected metabolic antibodies targeting several metabolic regulators such as metabolite transporters, rate-limiting enzymes and their regulators, mitochondrial modifiers, transcription factors and components of signaling pathways. Lineage-specific metabolic profile along with regulated expression of metabolic enzymes by different immune cell-types has been reliably identified by the approach, even in clinical samples of human colorectal carcinoma. Introspection of metabolic enzyme expression has successfully provided the clue regarding the metabolic pathway activity and metabolite flux through different pathways in certain immune cell types. The metabolic heterogeneity displayed by the activated T-cells in a temporal framework was also efficiently illuminated which was distinct from the bulk analysis and showed a sequential convergent pattern of metabolic reprogramming through diverse metabolic and cellular regulatory nodes involving modulation of several transcription factors as well as signaling molecules. Further, metabolic remodeling was in accordance with metabolic protein expression during early phase of T-cell activation which became divergent in later phase due to channelization of different metabolic intermediates through diverse metabolic pathways; most importantly, the metabolic reprogramming of immune cells was found to be closely linked with diverse cell cycle events. Differential tissue distribution of CD8⁺ T cells has been depicted in the study with more vibrant metabolic phenotypes of CD8⁺ T cells were observed inside the tissues in comparison to the peripheral blood, clearly justifying the significance of SCM

introspection in immunology research (Hartmann et al., 2021). A multiparameter flow cytometry based SCM modality employing antibodies against rate-limiting enzymes and proteins of different anabolic and catabolic pathways has been used to identify the metabolic reprogramming in human peripheral blood mononuclear cells which was closely associated with the activation status of specific T-cell subsets and immunological functions. The heterogeneity in metabolic flux through several pathways such as fatty acid synthesis, arginine metabolism, glycolysis, Krebs cycle, HMP pathway, oxidative phosphorylation, antioxidant response pathways, fatty acid oxidation, etc. in different immune-cell subsets under various stages of immune-function has been elucidated which is extremely valuable to introspect the phenotypic attributes of a precise immune cell specific and status-dependent immunometabolic response. Such introspection holds potential for exogenous metabolic modulation to influence the immune function as evidenced by the differentiation of a subset of inflammatory memory T cells driven by glucose restriction and metabolic rearrangement (Ahl et al., 2020). For instance, elucidation of the underpinning mechanisms of metabolic reprogramming undergoes in the activated macrophages during immune-mediated chronic inflammatory diseases such as systemic lupus erythematosus, rheumatoid arthritis, etc. can provide potential therapeutic targets to treat such autoimmune disorders (Jing et al., 2020). Similarly genome-scale modeling and multiomics introspection has successfully been carried out to identify the crucial metabolic attributes of macrophage activation during inflammatory response (Bordbar et al., 2012). Further, the effects of host factors and pathogen factors over immunometabolism, either separately or in combination can provide comprehensive information regarding the metabolic reprogramming during host-pathogen interaction (Rattigan et al., 2018). Similarly metabolomic profiling of effector CD4 T-cells and regulatory T-cells during autoimmune inflammatory diseases has depicted marked reprogramming in several crucial metabolic pathways, particularly in glutamine and serine metabolism, glycolysis, TCA cycle and nucleotide synthesis which provides the scope for identification of therapeutic checkpoints targeting the T cell metabolism (Andrejeva et al., 2018). Not only intracellular metabolites but also secreted metabolites from different cells during an immune response can also be analyzed. For instance, heterogeneity of secreted metabolites from single individual PC12 pheochromocytoma cells has been identified through micro-droplet based cell capturing coupled with LTQ-Orbitrap mass spectrometer (Fujita et al., 2015).

Single-cell metabolomics in detection of metabolite dynamicity and pathway modulation

Elucidation of metabolic pathway alterations at cellular and subcellular level induced by specific patho-physiological stimuli within a precise spatiotemporal framework is one of the core competencies of SCM introspection. Diverse SCM analytical modalities have been tried and tested to unearth such consistent

stimuli-responsive metabolite dynamicity and associated pathway modulations in the context of heterogeneous cellular metabolic phenotypes. Functionalized carbon nanoelectrodes, carbon nanopipettes, carbon nanowire electrodes employing silicon carbide semiconductor have been developed which can detect and quantify the metabolite dynamicity within live single-cells (McCormick & Dick, 2021). Noncarbon nanoelectrodes such as Platinum black nanoelectrodes has been employed for sensitive as well as selective detection of reactive oxygen and nitrogen species in activated murine macrophages during oxidative burst (Wang et al., 2012). A wireless asymmetric nanopore electrode having gold-coated interior sensing interface has successfully measured the NADH concentration within live single-cell at 1 picomolar detection limit facilitating in real-time monitoring of cellular redox metabolism. The electrode was found to be advantageous over the conventional wired electrodes as it yielded a high signal-to-noise ratio by amplifying the current signal from nanoamperes to picoamperes level. Further, the catechol-functionalized asymmetric nanopore electrode has also detected the reduction in NADH concentration induced by anticancer drug Taxol treatment in human breast cancer MCF-7 cells due to suppression of intracellular metabolism (Ying et al., 2018). Functionalized Tungsten nanoelectrode has been employed to enumerate the level of hydroxyl radical in RAW 264.7 murine macrophages under oxidative stress inflicted by amyloid β with the detection limit of 0.33 nM. Further, the modality has also revealed the cytoprotective effect of antioxidant Cordycepin by scavenging of hydroxyl radical through upregulation of antioxidant enzyme heme oxygenase-1 via PI3K/Akt pathway modulation. Such introspection carries enormous value in drug discovery and therapeutic target identification to ameliorate oxidative stress-induced inflammatory and neurodegenerative disorders like Alzheimer's disease (Ding et al., 2020). The concentrations and release dynamics of several cholinergic transmitters such as acetylcholine, serotonin, gamma-aminobutyric acid, dopamine, glutamate etc. have been enumerated in synaptic microenvironment of *A. californica* living neuron by employing a nano interface between two immiscible electrolyte solutions (nano-ITIES) electrode and nano-resolved scanning electrochemicalmicroscopy. The achieved signal to noise ratio of 6–130 through this modality clearly depicts the extreme sensitivity of the measuring method. The comprehensive and precise knowledge regarding synaptic neurotransmission carries immense implication in elucidating the nerve communication under various neuronal disorders (Shen et al., 2018). Similarly, dopamine release inside a single dopaminergic synapses has been measured by employing a carbon fiber nanoelectrode. Such introspection bears enormous value toward introspecting neurodegenerative disorders as progressive damage of dopaminergic neurons accompanied by reduced dopamine level in the midbrain are the cardinal signs of Parkinson's disease. Further, the neuroprotective effect of plant-derived natural compound Harpagide has also been elucidated as the compound enhanced dopamine release from synaptic vesicles and restored the secretion from damaged neurons in 6-hydroxydopamine-induced Parkinson's disease model in rat neuronal culture. The compound also protected the dopaminergic neurons by

preventing the reactive oxygen species-induced phosphorylation and aggregation of α -synuclein (Tang et al., 2019). Further development in different sensors along with intelligent detection system will certainly facilitate SCM intervention to detect stimuli-responsive metabolite dynamicity at living single-cell level on real time basis. The consistent deterministic effects in conjunction with artificial intelligence may be helpful to develop certain reliable model for clinical transition of SCM aftermaths in near future. Despite being a major macromolecule and an important component of cellular metabolism, fatty acids and lipid dynamics at single-cell level is relatively less explored area in health as well as disease context, mainly due to chemical complexity, huge diversity and technical difficulty. Recently, a MS-based method have successfully employed in tracing of alkyne-labeled lipids to study glycerolipid metabolism in hepatocytes. The technique extends subfemtomole level sensitivity, considerable robustness and the scope for multiplexing which facilitated parallel quantitative enumeration of more than hundred labeled lipid entities. The method will certainly ease the hurdles of single-cell lipidomics analysis in impending days (Thiele et al., 2019). Incessant technological development resulting into extreme detection sensitivity has now enabled penetration beyond single cell level to reach even up to single organelle level. For instance, a CE-MS based approach has effectively detected compartmentalization and dynamics of over hundreds of preexisting metabolites from the vacuole and cytoplasm isolated from a single *Chara australis* algal cell (Oikawa et al., 2011). The introspection also revealed that the spatiotemporal dynamicity of metabolites within different subcellular compartments and metabolite flux through different pathways are significantly influenced by developmental and environmental stimuli. Similarly, several mitochondrial metabolites including amino acids, TCA cycle intermediates, fatty acids and sterols have been detected in live HepG2 cells employing single-cell MS coupled with fluorescence imaging approach (Esaki & Masujima, 2015). Cellular and subcellular localization of lipids and metabolites in maize has also been elucidated by high-resolution MALDI-MSI platform (Dueñas et al., 2017). A nondestructive fluidic force microscopy along with MALDI MS analysis has successfully identified 20 cytosolic metabolites from individual HeLa cells (Guillaume-Gentil et al., 2017). The increased ^{13}C -labeled glucose uptake by the cancer cells due to Warburg effect resulting into accumulation ^{13}C -labeled metabolites was also depicted by the method. So, the current discussion leaves no ambiguity that SCM approach certainly brings special value over population metabolomics to elucidate cellular and subcellular metabolite dynamicity and pathway modulations in healthy as well as pathological context.

Single-cell metabolomics in clinical metabolism and disease perspective

The aforementioned discussion regarding the applications of SCM introspection has certainly established its potential in deciphering the clinical implications of

disease-associated metabolic pathway modulations inflicted by genetic, metabolic, developmental, microbial, parasitic or environmental etiology in both plant as well as animal kingdom. The unearthed information is not only effective for disease-specific metabolic biomarker discovery but also helpful in identifying of novel therapeutic targets. Further, monitoring the progression of disease and the effectiveness of an ongoing therapeutic regimen can also be enumerated through SCM analysis rendering it as a valuable tool in personalized medicine. Even identification of the common metabolic pathway alterations by a broad group of microbial pathogens can be helpful to customize a universal therapeutic strategy effective against wide-range of pathogens (Kumar, & Ghosh, Kumar, et al., 2020). Further, toxicological analysis of drugs yielding metabolic alterations at cellular and subcellular level can be exploited to develop preclinical drug toxicity testing models to hasten novel drug development. For instance, the drug-induced hepatotoxic effect of several hepatotoxic compounds on cellular metabolite profile and pathway modulations has been enumerated in HepG2 cells employing MS-based metabolomics approach. Oxidative stress modulating the glutathione and γ -glutamyl cycle, steatosis inflicting fatty acid β -oxidation with subsequent spiking in triacylglycerides synthesis, phospholipidosis due to inhibition of phospholipid degradation are depicted to be the predominant modes of hepato-cellular injury (García-Cañaveras et al., 2016). Similarly, the MS-based metabolomics analysis has also been employed to depict the synergistic toxic effect of anticancer drug dichloroacetate and pantothenate on different ovarian cancer cell lines linked with Coenzyme A biosynthesis; however, the noncancerous cells were spared from the toxic effect. The strategy was also worked successfully to mitigate other cancer lines too with varying degree of sensitivity. Several central metabolic pathways such as glycolysis, pentose phosphate pathway, oxidative phosphorylation, nucleotide metabolism, etc. were found to be modulated in drug-induced metabolic profiling (Dubuis et al., 2018). Such introspections can provide novel concepts and pivotal information in drug-discovery arena. Therefore SCM is progressively developing as an invincible technique in modern clinical research and commercial application for better understanding of pathophysiological processes and customization of appropriate countering strategies. Fourier transform ion cyclotron resonance mass spectrometer (FTICR MS) has been used to study for the stimulatory effect of Lipopolysaccharide (LPS) on RAW 264.7 cell as LPS is a conserved part of outer membrane of several gram negative bacteria which exerts toxic effects in animals. Abundance of less unsaturated phospholipids and reduction in lipids containing higher degree of unsaturation resulted from LPS stimulation. Sphingolipid profile was also altered under the influence of LPS along with modulation in several other lipid and metabolite profile (Yang et al., 2018). The changes observed in single-cell analyses may not be so pronounced or may be entirely masked in population study as individual cells within the body despite being genetically identical have divergent molecular attributes and responds differentially under various disease conditions due to variability in local micro-environmental factors and stochastic processes. Such variability in

cellular response has been explored through GC-MS analyses depicting distinct metabolite profile as the consequence of differential fatty acid biosynthesis and cholesterol metabolism in A549 and AGS two different cell lines under Influenza A virus infection (Kumar, & Ghosh, Kumar, et al., 2020). Thus SCM and single-cell type specific metabolomics analyses yield enormous precision with unparalleled sensitivity and can be carried out following both untargeted and targeted approaches. However, integration of SCM data with other omics observations is always desirable for comprehensive elucidation of any pathophysiological mechanism. Such an integrated approach involving proteomic and lipidomic profiling of Huh-7.5 cells infected with Hepatitis C virus revealed modulation of crucial host metabolic pathways such as glycolysis, pentose phosphate pathway, TCA cycle to facilitate viral replication and propagation. Several lipid components including phospholipids, sphingomyelins, ceramide, etc. were also modulated to favor the viral mechanisms (Diamond et al., 2010). Even the temporal variation in metabolite profile during viral infection cycle can assist to enumerate the disease status and infection stages. Time-specific influence of Influenza virus over cellular metabolite profile yielded similar metabolite profile between infected and mock-infected cells during the first 10–12 hours of infection; however, marked variation in metabolite flux through different pathways was documented between the two cell types after 12 hour of infection (Ritter et al., 2010). Another integrative transcriptomics and metabolomics approach has elucidated the molecular events leading to clinical tolerance in subsequent *Plasmodium vivax* infection in human subjects. Metabolomics introspection revealed prominent variations in methionine and cysteine metabolism, urea cycle and fatty acid metabolism between the naïve and semiimmune individuals. Elevated lipoxygenase, cyclooxygenase and lipid peroxidation was also evidenced in the subjects of later group. All such metabolic modulations converge into platelet activation, activation of innate immunity and T cell signaling to elicit tolerogenic response (Gardinassi et al., 2018). Thus, it can be easily conferred that integration of SCM analyses with other single-cell analytical platforms will certainly facilitate smooth transition of these high-end techniques toward clinical applications in a reliable and convincing manner.

Conclusion and future prospect

The enormous sensitivity, precision, and in-depth penetration of SCM analysis have rendered it suitable for versed biological applications more often than not in specialized purpose instead of common interventions. The reasons are manifold; predominantly SCM techniques are relatively naïve, skill and resource intensive, lacking in generalized standard protocols and instrumentations compatible for diverse sample types and chemically divergent analyte subjects, along with relatively deficient metabolite database. Further, the inherent limitations of metabolomics analysis such as rapid metabolite turnover, chemical complexity, lack of

amplification scope, etc. synergistically obstacle the routine utilization and smooth transition of these high-end SCM techniques toward clinical applications. Now, the focus should be toward improving the rate of “bench to beside” transition of these SCM techniques to convert their ample potential into on-field biological applications which will ensure a bright future prospect of SCM analyses and justify the massive input devoted to establish the SCM arsenal. The following measures will certainly assist in the process to render SCM analysis more scientific household in biological applications: 1. Standardization of the sampling and analytical procedures to customize general SCM workflow for divergent subjects. 2. Parallel developments in instrumentation as well as computational programs for high-throughput data analysis and bioinformatics tools for meaningful data interpretation. 3. Prominent distinction between the stochastic and deterministic variations. 4. Diligent pursuit for development of “Point of Care” type analytical modalities using microfluidics like platforms for on-spot applications. 5. Strengthening of metabolite database and facilitation of data sharing options. 6. Integration of multiple single-cell techniques for comprehensive and reliable introspection. 7. Generation of reliable model based upon compiled metabolomics data using advanced methods like artificial intelligence for diagnostic and therapeutic applications. Considering the age of SCM methods, it is not too late yet and the global scientific community nurturing this field is already eying to resolve the issues imminently.

References

- Abouleila, Y., Onidani, K., Ali, A., Shoji, H., Kawai, T., Lim, C. T., Kumar, V., Okaya, S., Kato, K., Hiyama, E., Yanagida, T., Masujima, T., Shimizu, Y., & Honda, K. (2018). Live single cell mass spectrometry reveals cancerspecific metabolic profiles of circulating tumor cells. *Cancer Science*. Available from <https://doi.org/10.1111/cas.13915>.
- Aerts, J. T., Louis, K. R., Crandall, S. R., Govindaiah, G., Cox, C. L., & Sweedler, J. V. (2014). Patch clamp electrophysiology and capillary electrophoresis—mass spectrometry metabolomics for single cell characterization. *Analytical Chemistry*, 86(6), 3203–3208. Available from <https://doi.org/10.1021/ac500168d>.
- Agathocleous, M., Meacham, C. E., Burgess, R. J., Piskounova, E., Zhao, Z., Crane, G. M., Cowin, B. L., Bruner, E., Murphy, M. M., Chen, W., Spangrude, G. J., Hu, Z., DeBerardinis, R., & Morrison, S. J. (2017). Ascorbate regulates haematopoietic stem cell function and leukaemogenesis. *Nature*. Available from <https://doi.org/10.1038/nature23876>.
- Ahl, J. P., Hopkins, R. A., Xiang, W. W., Au, B., Kalaperumal, N., Fairhurst, A., & Connoll, J. E. (2020). A novel strategy for single-cell metabolic analysis highlights dynamic changes in immune subpopulations. *bioRxiv*. Available from <https://doi.org/10.1101/2020.01.21.914663>.
- Allen, J., Davey, H. M., Broadhurst, D., Heald, J. K., Rowland, J. J., Oliver, S. G., & Kell, D. B. (2003). High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology*, 21(6), 692–696. Available from <https://doi.org/10.1038/nbt823>.

- Alwarawrah, Y., Kiernan, K., & MacIver, N. J. (2018). Changes in nutritional status impact immune cell metabolism and function. *Frontiers in Immunology*, 9. Available from <https://doi.org/10.3389/fimmu.2018.01055>.
- Amantonico, A., Oh, J. Y., Sobek, J., Heinemann, M., & Zenobi, R. (2008). Mass spectrometric method for analyzing metabolites in yeast with single cell sensitivity. *Angewandte Chemie International Edition*, 47(29), 5382–5385. Available from <https://doi.org/10.1002/anie.200705923>.
- Anderson, A. B., Gergen, J., & Arriaga, E. A. (2002). Detection of doxorubicin and metabolites in cell extracts and in single cells by capillary electrophoresis with laser-induced fluorescence detection. *Journal of Chromatography*, B769, 97–106.
- Andrejeva, G., Wolf, M. M., Johnson, M. O., Rutledge, A. C., Gabriela, S., Codreanu, G. S., Sherrod, S. D., Gutierrez, D., Rose, K. L., Norris, J. L., Schey, K. L., McLean, J. A., & Rathmell, J. C. (2018). Metabolomics analysis reveals differential T cell serine metabolism as a target in autoimmunity. *The Journal of Immunology*, 200(Suppl. 1), 167.7.
- Artyomov, M. N., & Van den Bossche, J. (2020). Immunometabolism in the single-cell era. *Cell Metabolism*. Available from <https://doi.org/10.1016/j.cmet.2020.09.013>.
- Astarita, G., & Langridge, J. (2013). An emerging role for metabolomics in nutrition science. *Journal of Nutrigenetics and Nutrigenomics*, 6(4–5), 181–200. Available from <https://doi.org/10.1159/000354403>.
- Baran, R., Ivanova, N., Jose, N., Garcia-Pichel, F., Kyrpides, N., Gugger, M., & Northen, T. (2013). Functional genomics of novel secondary metabolites from diverse cyanobacteria using untargeted metabolomics. *Marine Drugs*, 11(10), 3617–3631. Available from <https://doi.org/10.3390/mdl11103617>.
- Barros, E., Lezar, S., Anttonen, M. J., van Dijk, J. P., Röhlig, R. M., Kok, E. J., & Engel, K.-H. (2010). Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. *Plant Biotechnology Journal*, 8 (4), 436–451. Available from <https://doi.org/10.1111/j.1467-7652.2009.00487.x>.
- Baumeister, T. U. H., Vallet, M., Kaftan, F., Guillou, L., Svatoš, A., & Pohnert, G. (2020). Identification to species level of live single microalgal cells from plankton samples with matrix-free laser/desorption ionization mass spectrometry. *Metabolomics: Official Journal of the Metabolomic Society*, 16(3). Available from <https://doi.org/10.1007/s11306-020-1646-7>.
- Bergman, H.-M., & Lanekoff, I. (2017). Profiling and quantifying endogenous molecules in single cells using nano-DESI MS. *The Analyst*, 142(19), 3639–3647. Available from <https://doi.org/10.1039/c7an00885f>.
- Berry, D., Mader, E., Lee, T. K., Woebken, D., Wang, Y., Zhu, D., Palatinszky, M., Schintlmeistera, A., Schmidta, M. C., Hansona, B. T., Shterzere, N., Mizrahi, I., Rauchf, I., Decker, T., Bocklitzg, T., Poppg, J., Gibsoni, C. M., Fowleri, P. W., Huang, W. E., & Wagner, M. (2014). Tracking heavy water (D₂O) incorporation for identifying and sorting active microbial cells. *Proceedings of the National Academy of Sciences*, 112(2), E194–E203. Available from <https://doi.org/10.1073/pnas.1420406112>.
- Bhute, V. J., Bao, X., Dunn, K. K., Knutson, K. R., McCurry, E. C., Jin, G., Lee, W., Lewis, S., Ikeda, A., & Palecek, S. P. (2017). Metabolomics identifies metabolic markers of maturation in human pluripotent stem cell-derived cardiomyocytes. *Theranostics*, 7(7), 2078–2091. Available from <https://doi.org/10.7150/thno.19390>.
- Bordbar, A., Mo, M. L., Nakayasu, E. S., Schrimpe-Rutledge, A. C., Kim, Y.-M., Metz, T. O., Johnes, M. B., Frank, B. C., Smith, R. D., Peterson, S. N., Hyduke, D. R., Adkins, J. N., & Palsson, B. O. (2012). Model-driven multi-omic data analysis

- elucidates metabolic immunomodulators of macrophage activation. *Molecular Systems Biology*, 8. Available from <https://doi.org/10.1038/msb.2012.21>.
- Bouzier-Sore, A. K., Voisin, P., Bouchaud, V., Bezancon, E., Franconi, J. M., & Pellerin, L. (2006). Competition between glucose and lactate as oxidative energy substrates in both neurons and astrocytes: A comparative NMR study. *European Journal of Neuroscience*, 24(6), 1687–1694. Available from <https://doi.org/10.1111/j.1460-9568.2006.05056.x>.
- Bowers, M., Liang, T., Gonzalez-Bohorquez, D., Zocher, S., Jaeger, B. N., Kovacs, W. J., Röhrl, C., Cramb, K. M. L., Winterer, J., Kruse, M., Dimitrieva, S., Overall, R. W., Wegleiter, T., Najmabadi, H., Semenkovich, C. F., Kempermann, G., Földy, C., & Jessberger, S. (2020). FASN-dependent lipid metabolism links neurogenic stem/progenitor cell activity to learning and memory deficits. *Cell Stem Cell*. Available from <https://doi.org/10.1016/j.stem.2020.04.002>.
- Brechenmacher, L., Lei, Z., Libault, M., Findley, S., Sugawara, M., Sadowsky, M. J., Sumner, L. W., & Stacey, G. (2010). Soybean metabolites regulated in root hairs in response to the symbiotic bacterium *Bradyrhizobium japonicum*. *Plant Physiology*, 153 (4), 1808–1822. Available from <https://doi.org/10.1104/pp.110.157800>.
- Bruno, C., Patin, F., Bocca, C., Nadal-Desbarats, L., Bonnier, F., Reynier, P., & Blasco, H. (2018). The combination of four analytical methods to explore skeletal muscle metabolomics: Better coverage of metabolic pathways or a marketing argument? *Journal of Pharmaceutical and Biomedical Analysis*, 148, 273–279. Available from <https://doi.org/10.1016/j.jpba.2017.10.013>.
- Calabrese, F., Voloshynovska, I., Musat, F., Thullner, M., Schröder, M., Richnow, H. H., Lambrecht, J., Müller, S., Wick, L. Y., Musat, N., & Stryhanyuk, H. (2019). Quantitation and comparison of phenotypic heterogeneity among single cells of monoclonal microbial populations. *Frontiers in Microbiology*, 10. Available from <https://doi.org/10.3389/fmicb.2019.02814>.
- Cezar, G. G., Quam, J. A., Smith, A. M., Rosa, G. J. M., Piekarczyk, M. S., Brown, J. F., Gaze, F. H., & Muotri, A. R. (2007). Identification of small molecules from human embryonic stem cells using metabolomics. *Stem Cells and Development*, 16(6), 869–882. Available from <https://doi.org/10.1089/scd.2007.0022>.
- Chaturvedi, V., & Verma, P. (2016). Microbial fuel cell: A green approach for the utilization of waste for the generation of bioelectricity. *Bioresources and Bioprocessing*, 3(1). Available from <https://doi.org/10.1186/s40643-016-0116-6>.
- Chen, C., Liu, Y., Liu, Y., & Zheng, P. (2009). mTOR regulation and therapeutic rejuvenation of aging hematopoietic stem cells. *Science Signaling*, 2(98), ra75. Available from <https://doi.org/10.1126/scisignal.2000559>.
- Chen, F., Lin, L., Zhang, J., He, Z., Uchiyama, K., & Lin, J.-M. (2016). Single-cell analysis using drop-on-demand inkjet printing and probe electrospray ionization mass spectrometry. *Analytical Chemistry*, 88(8), 4354–4360. Available from <https://doi.org/10.1021/acs.analchem.5b04749>.
- Chijiwa, R., Hosokawa, M., Kogawa, M., Nishikawa, Y., Ide, K., Sakanashi, C., Takahashi, K., & Takeyama, H. (2020). Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome*, 8(5). Available from <https://doi.org/10.1186/s40168-019-0779-2>.
- Clair, G. (2019). A multi-omics zoom on the molecular networks of Lung development. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 317, L554–L555. Available from <https://doi.org/10.1152/ajplung.00364.2019>.

- Cui, J., Xie, Y., Sun, T., Chen, L., & Zhang, W. (2021). Deciphering and engineering photosynthetic cyanobacteria for heavy metal bioremediation. *Science of The Total Environment*, 761. Available from <https://doi.org/10.1016/j.scitotenv.2020.144111>.
- Dal Co, A., Ackermann, M., & van Vliet, S. (2019). Metabolic activity affects the response of single cells to a nutrient switch in structured populations. *Journal of The Royal Society Interface*, 16(156), 20190182. Available from <https://doi.org/10.1098/rsif.2019.0182>.
- De Lorenzo, V. (2014). From the selfish genetoselfish metabolism: Revisiting the central dogma. *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology*, 36(3), 226–235. Available from <https://doi.org/10.1002/bies.201300153>.
- Deng, J., Yang, Y., Xu, M., Wang, X., Lin, L., Yao, Z.-P., & Luan, T. (2015). Surface-coated probe nanoelectrospray ionization mass spectrometry for analysis of target compounds in individual small organisms. *Analytical Chemistry*, 87(19), 9923–9930. Available from <https://doi.org/10.1021/acs.analchem.5b03110>.
- DeVilbiss, A. W., Zhao, Z., Martin-Sandoval, M. S., Ubellacker, J. M., Tasdogan, A., Agathocleous, M., Mathews, T. P., & Morrison, S. J. (2020). Metabolomic profiling of rare cell populations isolated by flow cytometry from tissues. *bioRxiv*, 246900. Available from <https://doi.org/10.1101/2020.08.11.246900>
- Diamond, D. L., Syder, A. J., Jacobs, J. M., Sorensen, C. M., Walters, K.-A., Proll, S. C., McDermott, J. E., Gritsenko, M. A., Zhang, Q., Zhao, R., Metz, T. O., Camp, D. G., Water, K. M., Smith, R. D., Rice, C. M., & Katze, M. G. (2010). Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics. *PLoS Pathogens*, 6(1), e1000719. Available from <https://doi.org/10.1371/journal.ppat.1000719>.
- Ding, S., Li, M., Gong, H., Zhu, Q., Shi, G., & Zhu, A. (2020). Sensitive and selective measurement of hydroxyl radicals at subcellular level with tungsten nanoelectrodes. *Analytical Chemistry*, 92(3), 2543–2549. Available from <https://doi.org/10.1021/acs.analchem.9b04139>.
- Du, J., Su, Y., Qian, C., Yuan, D., Miao, K., Lee, D., Ng, A. H. C., Wijker, R. S., Ribas, A., Levine, R. D., Heath, J. R., & Wei, L. (2020). Raman-guided subcellular pharmaco-metabolomics for metastatic melanoma cells. *Nature Communications*, 11(1). Available from <https://doi.org/10.1038/s41467-020-18376-x>.
- Dubuis, S., Ortmayr, K., & Zampieri, M. (2018). A framework for large-scale metabolome drug profiling links coenzyme A metabolism to the toxicity of anti-cancer drug dichloroacetate. *Communications Biology*, 1(1). Available from <https://doi.org/10.1038/s42003-018-0111-x>.
- Dueñas, M. E., Feenstra, A. D., Korte, A. R., Hinnens, P., & Lee, Y. J. (2017). Cellular and subcellular level localization of maize lipids and metabolites using high-spatial resolution MALDI mass spectrometry imaging. *Methods in Molecular Biology*, 217–231. Available from https://doi.org/10.1007/978-1-4939-7315-6_13.
- Dusny, C., Lohse, M., Reemtsma, T., Schmid, A., & Lechtenfeld, O. J. (2019). Quantifying a biocatalytic product from a few living microbial cells using microfluidic cultivation coupled to FT-ICR-MS. *Analytical Chemistry*, 91, 7012–7018. Available from <https://doi.org/10.1021/acs.analchem.9b00978>.
- Esaki, T., & Masujima, T. (2015). Fluorescence probing live single-cell mass spectrometry for direct analysis of organelle metabolism. *Analytical Sciences*, 31(12), 1211–1213. Available from <https://doi.org/10.2116/analsci.31.1211>.

- Evers, T., Hochane, M., Tans, S. J., Heeren, R. M. A., Semrau, S., Nemes, P., & Mashaghi, A. (2019). Deciphering metabolic heterogeneity by single-cell analysis. *Analytical Chemistry*, 91(21), 13314–13323. Available from <https://doi.org/10.1021/acs.analchem.9b02410>.
- Fernandez-Rebollo, E., Franzen, J., Goetzke, R., Hollmann, J., Ostrowska, A., Oliverio, M., Sieben, T., Rath, B., Kornfeld, J., & Wagner, W. (2020). Senescence-associated metabolomic phenotype in primary and iPSC-derived mesenchymal stromal cells. *Stem Cell Reports*, 14(2), 201–209. Available from <https://doi.org/10.1016/j.stemcr.2019.12.012>.
- Fujita, H., Esaki, T., Masujima, T., Hotta, A., Kim, S. H., Noji, H., & Watanabe, T. M. (2015). Comprehensive chemical secretory measurement of single cells trapped in a micro-droplet array with mass spectrometry. *RSC Advances*, 5(22), 16968–16971. Available from <https://doi.org/10.1039/c4ra12021c>.
- Fukano, Y., Tsuyama, N., Mizuno, H., Date, S., Takano, M., & Masujima, T. (2012). Drug metabolite heterogeneity in cultured single cells profiled by pico-trapping direct mass spectrometry. *Nanomedicine: Nanotechnology, Biology, and Medicine*, 7(9), 1365–1374. Available from <https://doi.org/10.2217/nnm.12.34>.
- García- Cañaveras, J. C., Castell, J. V., Donato, M. T., & Lahoz, A. (2016). A metabolomics cell-based approach for anticipating and investigating drug-induced liver injury. *Scientific Reports*, 6(1). Available from <https://doi.org/10.1038/srep27239>.
- Garden, R. W., Moroz, L. L., Moroz, T. P., Shippy, S. A., & Sweedler, J. V. (1996). Excess salt removal with matrix rinsing: Direct peptide profiling of neurons from marine invertebrates using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, 31(10), 1126–1130.
- Gardinassi, L. G., Arévalo-Herrera, M., Herrera, S., Cordy, R. J., Tran, V., Smith, M. R., Johnson, M. S., Chako, B., Liu, K. H., DarleyUsmar, V. M., Go, Y. M., MaHPIC Consortium, Johnes, D. P., Galinski, M. R., & Li, S. (2018). Integrative metabolomics and transcriptomics signatures of clinical tolerance to *Plasmodium vivax* reveal activation of innate cell immunity and T cell signaling. *Redox Biology*, 17, 158–170. Available from <https://doi.org/10.1016/j.redox.2018.04.011>.
- Gibbs, M. E., Anderson, D. G., & Hertz, L. (2006). Inhibition of glycogenolysis in astrocytes interrupts memory consolidation in young chickens. *Glia*, 54(3), 214–222. Available from <https://doi.org/10.1002/glia.20377>.
- Giraudeau, P. (2020). NMR-based metabolomics and fluxomics: Developments and future prospects. *The Analyst*, 145(7), 2457–2472. Available from <https://doi.org/10.1039/d0an00142b>.
- Gong, X., Zhao, Y., Cai, S., Fu, S., Yang, C., Zhang, S., & Zhang, X. (2014). Single cell analysis with probe ESI-mass spectrometry: Detection of metabolites at cellular and subcellular levels. *Analytical Chemistry*, 86(8), 3809–3816. Available from <https://doi.org/10.1021/ac500882e>.
- Goodarzi, P., Alavi-Moghadam, S., Payab, M., Larijani, B., Rahim, F., Gilany, K., Bana, N., Tayyanloo-Beik, A., Foroughi Heravani, N., Hadavandkhani, M., & Arjmand, B. (2019). Metabolomics analysis of mesenchymal stem cells. *International Journal of Molecular and Cellular Medicine*, Winter, 8(Suppl. 1), 30–40. Available from <https://doi.org/10.22088/IJMCM.BUMS.8.2.30>.
- Goossens, A., Häkkinen, S. T., Laakso, I., Seppänen-Laakso, T., Biondi, S., De Sutter, V., Lammertyn, F., Nuutila, A. M., Söderlund, H., Zabeau, M., Inzé, D., & Oksman-Caldentey, K. M. (2003). A functional genomics approach toward the understanding of

- secondary metabolism in plant cells. *Proceedings of the National Academy of Sciences USA*, 100(14), 8595–8600. Available from <https://doi.org/10.1073/pnas.1032967100>.
- Guillaume-Gentil, O., Rey, T., Kiefer, P., Ibáñez, A. J., Steinhoff, R., Brönnimann, R., Dorwling-Carter, L., Zambelli, T., Zenobi, R., & Vorholt, J. A. (2017). Single-cell mass spectrometry of metabolites extracted from live cells by fluidic force microscopy. *Analytical Chemistry*, 89(9), 5017–5023. Available from <https://doi.org/10.1021/acs.analchem.7b00367>.
- Hang, W., Meng, Y., Cheng, X., Wang, T., Li, X., Nie, W., Li, x., Nie, W., Liu, R., Lin, Z., Hang, L., Yin, Z., Zhang, B., & Yan, X. (2020). Micro-lensed fiber laser desorption mass spectrometry imaging reveals subcellular distribution of drugs within single cells. *Angewandte Chemie*, 59, 17864. Available from <https://doi.org/10.1002/ange.202002151>.
- Hartmann, F. J., Mrdjen, D., McCaffrey, E., Glass, D. R., Greenwald, N. F., Bharadwaj, A., Khair, Z., Verberk, S. G. S., Baranski, A., Baskar, R., Graf, W., Van Valen, D., Van den Bossche, J., Angelo, M., & Bendall, S. C. (2021). Single-cell metabolic profiling of human cytotoxic T cells. *Nature Biotechnology*, 39, 186–197. Available from <https://doi.org/10.1038/s41587-020-0651-8>.
- He, X., Memczak, S., Qu, J., Belmonte, J. C. I., & Liu, G. H. (2020). Single-cell omics in aging: A young and growing field. *Nature Metabolism*, 2(4), 293–302. Available from <https://doi.org/10.1038/s42255-020-0196-7>.
- Hiyama, E., Ali, A., Amer, S., Harada, T., Shimamoto, K., Furushima, R., Emara, S., & Masujima, T. (2015). Direct lipido-metabolomics of single floating cells for analysis of circulating tumor cells by live single-cell mass spectrometry. *Analytical Sciences*, 31 (12), 1215–1217. Available from <https://doi.org/10.2116/analsci.31.1215>.
- Hsu, C. C., Xu, J., Brinkhof, B., Wang, H., Cui, Z., Huang, W. E., & Ye, H. (2020). A single-cell Raman-based platform to identify developmental stages of human pluripotent stem cell-derived neurons. *Proceedings of the National Academy of Sciences*, 202001906. Available from <https://doi.org/10.1073/pnas.2001906117>.
- Hu, J., Jiang, X. X., Wang, J., Guan, Q. Y., Zhang, P. K., Xu, J. J., & Chen, H. Y. (2016). Synchronized polarization induced electrospray: Comprehensively profiling biomolecules in single cells by combining both positive-ion and negative-ion mass spectra. *Analytical Chemistry*, 88(14), 7245–7251. Available from <https://doi.org/10.1021/acs.analchem.6b01490>.
- Ivanova, G., Pereira, T., Caseiro, A. R., Georgieva, P., & Maurício, A. C. (2016). Metabolomic and proteomic analysis of the mesenchymal stem cells' secretome. *Metabolomics—Fundamentals and Applications*. Available from <https://doi.org/10.5772/66101>.
- Jasper, H., & Jones, D. L. (2010). Metabolic regulation of stem cell behavior and implications for aging. *Cell Metabolism*, 12(6), 561–565. Available from <https://doi.org/10.1016/j.cmet.2010.11.010>.
- Jiménez, C. R., Li, K. W., Dreisewerd, K., Spijker, S., Kingston, R., Bateman, R. H., Burlingame, A. M., Smit, A. L., van Minnen, J., & Geraerts, W. P. M. (1998). Direct mass spectrometric peptide profiling and sequencing of single neurons reveals differential peptide patterns in a small neuronal network. *Biochemistry*, 37(7), 2070–2076. Available from <https://doi.org/10.1021/bi971848b>.
- Jiménez, C. R., van Veelen, P. A., Li, K. W., Wildering, W. C., Geraerts, W. P., Tjaden, U. R., & van der Greef, J. (2008). Rapid communication: Neuropeptide expression and processing as revealed by direct matrix-assisted laser desorption ionization mass

- spectrometry of single neurons. *Journal of Neurochemistry*, 62(1), 404–407. Available from <https://doi.org/10.1046/j.1471-4159.1994.62010404.x>.
- Jing, C., Castro-Dopico, T., Richoz, N., Tuong, Z. K., Ferdinand, J. R., Lok, L. S. C., Loudon, K. W., Banham, G. D., Mathews, R. J., Cader, Z., Fitzpatrick, S., Bashant, K. R., Kapllan, M. J., Kaser, A., Johnson, R. S., Murphy, M. P., Siegel, R. M., & Clatworthy, M. R. (2020). Macrophage metabolic reprogramming presents a therapeutic target in lupus nephritis. *Proceedings of the National Academy of Sciences USA*, 202000943. Available from <https://doi.org/10.1073/pnas.2000943117>.
- Jo, K., Heien, M. L., Thompson, L. B., Zhong, M., Nuzzo, R. G., & Sweedler, J. V. (2007). Mass spectrometric imaging of peptide release from neuronal cells within microfluidic devices. *Lab on a Chip*, 7(11), 1454. Available from <https://doi.org/10.1039/b706940e>.
- Jones, D. P., Park, Y., & Ziegler, T. R. (2012). Nutritional metabolomics: Progress in addressing complexity in diet and health. *Annual Review of Nutrition*, 32(1), 183–202. Available from <https://doi.org/10.1146/annurev-nutr-072610-145159>.
- Kalaitzidis, D., Sykes, S. M., Wang, Z., Punt, N., Tang, Y., Ragu, C., Sinha, A. U., Lane, S. W., Souza, A. L., Clish, C. V., Anastasiou, D., Gilliland, D. G., Scadden, D. T., & Armstrong, S. A. (2012). mTOR complex 1 plays critical roles in hematopoiesis and Pten-loss-evoked leukemogenesis. *Cell Stem Cell*, 11(3), 429–439. Available from <https://doi.org/10.1016/j.stem.2012.06.009>.
- Kamei, Y., Tamada, Y., Nakayama, Y., Fukusaki, E., & Mukai, Y. (2014). Changes in transcription and metabolism during the early stage of replicative cellular senescence in budding yeast. *Journal of Biological Chemistry*, 289(46), 32081–32093. Available from <https://doi.org/10.1074/jbc.m114.600528>.
- Kharas, M. G., Okabe, R., Ganis, J. J., Gozo, M., Khandan, T., Paktinat, M., Gilliland, D. J., & Gritsman, K. (2009). Constitutively active AKT depletes hematopoietic stem cells and induces leukemia in mice. *Blood*, 115(7), 1406–1415. Available from <https://doi.org/10.1182/blood-2009-06-229443>.
- Khatiwada, B., Sunna, A., & Nevalainen, H. (2020). Molecular tools and applications of *Euglena gracilis*-from biorefineries to bioremediation. *Biotechnology and Bioengineering*, 117, 3952–3967. Available from <https://doi.org/10.1002/bit.27516>.
- Kumar, R., Ghosh, M., Kumar, S., & Prasad, M. (2020). Single cell metabolomics: A future tool to unmask cellular heterogeneity and virus-host interaction in context of emerging viral diseases. *Frontiers in Microbiology*, 11. Available from <https://doi.org/10.3389/fmicb.2020.01152>.
- Kumar, R., & Kundu, S. (2020). Microbial bioremediation and biodegradation of hydrocarbons, heavy metals, and radioactive wastes in solids and wastewaters. In M. Shah (Ed.), *Microbial bioremediation & biodegradation*. Singapore: Springer. Available from https://doi.org/10.1007/978-981-15-1812-6_4.
- Lancaster, M. A., & Knoblich, J. A. (2014). Organogenesis in a dish: Modeling development and disease using organoid technologies. *Science (New York, N.Y.)*, 345(6194), 1247125. Available from <https://doi.org/10.1126/science.1247125>.
- Lancini, G., & Demain, A. L. (2013). Bacterial pharmaceutical products. *The Prokaryotes*, 257–280. Available from https://doi.org/10.1007/978-3-642-31331-8_28.
- Lapainis, T., Rubakhin, S. S., & Sweedler, J. V. (2009). Capillary electrophoresis with electrospray ionization mass spectrometric detection for single-cell metabolomics. *Analytical Chemistry*, 81(14), 5858–5864. Available from <https://doi.org/10.1021/ac900936g>.

- Li, L., Garden, R. W., & Sweedler, J. V. (2000). Single-cell MALDI: A new tool for direct peptide profiling. *Trends in Biotechnology*, 18(4), 151–160. Available from [https://doi.org/10.1016/s0167-7799\(00\)01427-x](https://doi.org/10.1016/s0167-7799(00)01427-x).
- Li, Q., Chen, P., Fan, Y., Wang, X., Xu, K., Li, L., & Tang, B. (2016). Multicolor fluorescence detection-based microfluidic device for single-cell metabolomics: Simultaneous quantitation of multiple small molecules in primary liver cells. *Analytical Chemistry*, 88(17), 8610–8616. Available from <https://doi.org/10.1021/acs.analchem.6b01775>.
- Li, Q., Tang, F., Huo, X., Huang, X., Zhang, Y., Wang, X., & Zhang, X. (2019). Native state single-cell printing system and analysis for matrix effects. *Analytical Chemistry*, 91(13), 8115–8122. Available from <https://doi.org/10.1021/acs.analchem.9b00344>.
- Li, Z., Wang, Z., Pan, J., Ma, X., Zhang, W., & Ouyang, Z. (2020). Single-cell mass spectrometry analysis of metabolites facilitated by cell electro-migration and electroporation. *Analytical Chemistry*, 92(14), 10138–10144. Available from <https://doi.org/10.1021/acs.analchem.0c02147>. Epub 2020.
- Libault, M., Pingault, L., Zogli, P., & Schiefelbein, J. (2017). Plant systems biology at the single-cell level. *Trends in Plant Science*, 22(11), 949–960. Available from <https://doi.org/10.1016/j.tplants.2017.08.006>.
- Liu, R., Pan, N., Zhu, Y., & Yang, Z. (2018). The T-probe: An integrated microscale device for online in situ single cell analysis and metabolic profiling using mass spectrometry. *Analytical Chemistry*, 90(18), 11078–11085. Available from <https://doi.org/10.1021/acs.analchem.8b02927>.
- Lombard-Banek, C., Li, J., Portero, E. P., Onjiko, R. M., Singer, C. D., Plotnick, D. O., Al Shabeeb, R. Q., & Nemes, P. (2021). In vivo subcellular mass spectrometry enables proteo-metabolomic single-cell systems biology in a chordate embryo developing to a normally behaving tadpole (*X. laevis*). *Angewandte Chemie International Edition in English*, 60, 12852. Available from <https://doi.org/10.1002/anie.202100923>.
- López-Otín, C., Galluzzi, L., Freije, J. M. P., Madeo, F., & Kroemer, G. (2016). Metabolic control of longevity. *Cell*, 166(4), 802–821. Available from <https://doi.org/10.1016/j.cell.2016.07.031>.
- Lu, Q., Lin, R., Du, C., Meng, Y., Yang, M., Zenobi, R., & Hang, W. (2020). Metal probe microextraction coupled to dielectric barrier discharge ionization-mass spectrometry for detecting drug residues in organisms. *Analytical Chemistry*, 92(8), 5921–5928. Available from <https://doi.org/10.1021/acs.analchem.0c00004>.
- Luo, X., & Li, L. (2017). Metabolomics of small numbers of cells: Metabolomic profiling of 100, 1000, and 10000 human breast cancer cells. *Analytical Chemistry*, 89(21), 11664–11671. Available from <https://doi.org/10.1021/acs.analchem.7b03100>.
- Magee, J. A., Ikenoue, T., Nakada, D., Lee, J. Y., Guan, K. L., & Morrison, S. J. (2012). Temporal changes in PTEN and mTORC2 regulation of hematopoietic stem cell self-renewal and leukemia suppression. *Cell Stem Cell*, 11(3), 415–428. Available from <https://doi.org/10.1016/j.stem.2012.05.026>.
- Maglica, Ž., Özdemir, E., & McKinney, J. D. (2015). Single-cell tracking reveals antibiotic-induced changes in mycobacterial energy metabolism. *mBio*, 6(1). Available from <https://doi.org/10.1128/mbio.02236-14>.
- Marlow, J. J., Colucci, I., Jungbluth, S. P., & Kallmeyer, J. (2020). Mapping metabolic activity at single cell resolution in intact volcanic fumarole sediment. *FEMS Microbiology Letters*, 367(1). Available from <https://doi.org/10.1093/femsle/fnaa031>.

- McCormick, H. K., & Dick, J. E. (2021). Nanoelectrochemical quantification of single-cell metabolism. *Analytical and Bioanalytical Chemistry*, 413(1), 17–24. Available from <https://doi.org/10.1007/s00216-020-02899-9>.
- Mellors, J. S., Jorabchi, K., Smith, L. M., & Ramsey, J. M. (2010). Integrated microfluidic device for automated single cell analysis using electrophoretic separation and electrospray ionization mass spectrometry. *Analytical Chemistry*, 82(3), 967–973. Available from <https://doi.org/10.1021/ac902218y>.
- Minakshi, P., Kumar, R., Ghosh, M., Brar, B., Barnela, M., & Lakhani, P. (2020). Application of polymeric nano-materials in management of inflammatory bowel disease. *Current Topics in Medicinal Chemistry*, 20(11), 982–1008. Available from <https://doi.org/10.2174/156802662066200320113322>.
- Mizuno, H., Naohiro, T., Sachiko, D., Takanori, H., & Tsutomu, M. (2008). Live single-cell metabolomics of tryptophan and histidine metabolites in a rat basophil leukemia cell. *Analytical Sciences*, 24, 1525–1527. Available from <https://doi.org/10.2116/analsci.24.1525>.
- Nakashima, T., Wada, H., Morita, S., Erra-Balsells, R., Hiraoka, K., & Nonami, H. (2016). Single-cell metabolite profiling of stalk and glandular cells of intact trichomes with internal electrode capillary pressure probe electrospray ionization mass spectrometry. *Analytical Chemistry*, 88(6), 3049–3057. Available from <https://doi.org/10.1021/acs.analchem.5b03366>.
- Nakatani, K., Izumi, Y., Hata, K., & Bamba, T. (2020). An analytical system for single-cell metabolomics of typical mammalian cells based on highly sensitive nano-liquid chromatography tandem mass spectrometry. *Mass Spectrom (Tokyo)*, 9(1), A0080. Available from <https://doi.org/10.5702/massspectrometry.A0080>.
- Nemes, P., Knolhoff, A. M., Rubakhin, S. S., & Sweedler, J. V. (2011). Metabolic differentiation of neuronal phenotypes by single-cell capillary electrophoresis–electrospray ionization-mass spectrometry. *Analytical Chemistry*, 83(17), 6810–6817. Available from <https://doi.org/10.1021/ac2015855>.
- Nemes, P., Knolhoff, A. M., Rubakhin, S. S., & Sweedler, J. V. (2012). Single-cell metabolomics: Changes in the metabolome of freshly isolated and cultured neurons. *ACS Chemical Neuroscience*, 3(10), 782–792. Available from <https://doi.org/10.1021/cn300100u>.
- Nemes, P., Rubakhin, S. S., Aerts, J. T., & Sweedler, J. V. (2013). Qualitative and quantitative metabolomic investigation of single neurons by capillary electrophoresis electrospray ionization mass spectrometry. *Nature Protocols*, 8(4), 783–799. Available from <https://doi.org/10.1038/nprot.2013.035>.
- Neupert, S., & Predel, R. (2005). Mass spectrometric analysis of single identified neurons of an insect. *Biochemical and Biophysical Research Communications*, 327(3), 640–645. Available from <https://doi.org/10.1016/j.bbrc.2004.12.086>.
- Nourbakhsh-Rey, M., & Libault, M. (2016). Decipher the molecular response of plant single cell types to environmental stresses. *BioMed Research International*, 2016, 1–8. Available from <https://doi.org/10.1155/2016/4182071>.
- Oikawa, A., Matsuda, F., Kikuyama, M., Mimura, T., & Saito, K. (2011). Metabolomics of a single vacuole reveals metabolic dynamism in an alga *Chara australis*. *Plant Physiology*, 157(2), 544–551. Available from <https://doi.org/10.1104/pp.111.183772>.
- Ojuederie, O., & Babalola, O. (2017). Microbial and plant-assisted bioremediation of heavy metal polluted environments: A review. *International Journal of Environmental Research and Public Health*, 14(12), 1504. Available from <https://doi.org/10.3390/ijerph14121504>.

- Onjiko, R. M., Moody, S. A., & Nemes, P. (2015). Single-cell mass spectrometry reveals small molecules that affect cell fates in the 16-cell embryo. *Proceedings of the National Academy of Sciences USA*, 112(21), 6545–6550. Available from <https://doi.org/10.1073/pnas.1423682112>.
- Onjiko, R. M., Morris, S. E., Moody, S. A., & Nemes, P. (2016). Single-cell mass spectrometry with multi-solvent extraction identifies metabolic differences between left and right blastomeres in the 8-cell frog (*Xenopus*) embryo. *The Analyst*, 141(12), 3648–3656. Available from <https://doi.org/10.1039/c6an00200e>.
- Onjiko, R. M., Portero, E. P., Moody, S. A., & Nemes, P. (2017). Microprobe capillary electrophoresis mass spectrometry for single-cell metabolomics in live frog (*Xenopus laevis*) embryos. *Journal of Visualized Experiments*, 130, 56956. Available from <https://doi.org/10.3791/56956>.
- Ostrowski, S. G. (2004). Mass spectrometric imaging of highly curved membranes during tetrahymena mating. *Science (New York, N.Y.)*, 305(5680), 71–73. Available from <https://doi.org/10.1126/science.1099791>.
- Pan, N., Rao, W., Kothapalli, N. R., Liu, R., Burgett, A. W. G., & Yang, Z. (2014). The single-probe: A miniaturized multifunctional device for single cell mass spectrometry analysis. *Analytical Chemistry*, 86(19), 9376–9380. Available from <https://doi.org/10.1021/ac5029038>.
- Pan, N., Rao, W., Standke, S. J., & Yang, Z. (2016). Using dicationic ion-pairing compounds to enhance the single cell mass spectrometry analysis using the single-probe: A microscale sampling and ionization device. *Analytical Chemistry*, 88(13), 6812–6819. Available from <https://doi.org/10.1021/acs.analchem.6b01284>.
- Partridge, E. A., Davey, M. G., Hornick, M. A., McGovern, P. E., Mejaddam, A. Y., Vrcenak, J. D., Mesas-Burgos, C., Olive, A., Caskey, R. C., Weiland, T. R., Han, J., Schupper, A. J., Connelly, J. T., Dysart, K. C., Rychik, J., Hedrick, H. L., Peranteau, W. H., & Flake, A. W. (2017). An extra-uterine system to physiologically support the extreme premature lamb. *Nature Communications*, 8, 15112. Available from <https://doi.org/10.1038/ncomms15112>.
- Petrou, K., Nielsen, D. A., & Heraud, P. (2018). Single-cell biomolecular analysis of coral algal symbionts reveals opposing metabolic responses to heat stress and expulsion. *Frontiers in Marine Science*, 5, 110. Available from <https://doi.org/10.3389/fmars.2018.00110>.
- Portero, E. P., & Nemes, P. (2019). Dual cationic–anionic profiling of metabolites in a single identified cell in a live *Xenopus laevis* embryo by microprobe CE-ESI-MS. *Analyst*, 144(3), 892–900. Available from <https://doi.org/10.1039/c8an01999a>.
- Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhof, H. V., Dam, K. V., & Oliver, S. G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19(1), 45–50. Available from <https://doi.org/10.1038/83496>.
- Ramanjaneyulu, G., & Reddy, B. R. (2019). Emerging trends of microorganism in the production of alternative energy. *Recent Developments in Applied Microbiology and Biochemistry*, 275–305. Available from <https://doi.org/10.1016/b978-0-12-816328-3.00021-0>.
- Rappez, L., Stadler, M., Triana, S., Phapale, P., Heikenwalder, M., & Alexandrov, T. (2019). Spatial single-cell profiling of intracellular metabolomes in situ. *bioRxiv*; Available from <https://doi.org/10.1101/510222>.

- Rattigan, K. M., Pountain, A. W., Regnault, C., Achcar, F., Vincent, I. M., Goodyear, C. S., & Barrett, M. P. (2018). Metabolomic profiling of macrophages determines the discrete metabolomic signature and metabolomic interactome triggered by polarising immune stimuli. *PLoS One*, 13(3), e0194126. Available from <https://doi.org/10.1371/journal.pone.0194126>.
- Raveendran, S., Parameswaran, B., Ummalyma, S. B., Abraham, A., Mathew, A. K., Madhavan, A., Rebello, S., & Mathew, A. K. (2018). Applications of microbial enzymes in food industry. *Food Technology and Biotechnology*, 56(1). Available from <https://doi.org/10.17113/fb.56.01.18.5491>.
- Ritter, J. B., Wahl, A. S., Freund, S., Genzel, Y., & Reichl, U. (2010). Metabolic effects of influenza virus infection in cultured animal cells: Intra- and extracellular metabolite profiling. *BMC Systems Biology*, 4(1), 61. Available from <https://doi.org/10.1186/1752-0509-4-61>.
- Robinson, O., Chadeau Hyam, M., Karaman, I., Climaco Pinto, R., Ala-Korpela, M., Handakas, E., Fiorito, G., Gao, H., Heard, A., Lewis, M., Pazoki, R., Polidoro, S., Tzoulaki, I., Wielscher, M., Elliott, P., & Vineis, P. (2020). Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. *Aging Cell*, 19(6), e13149. Available from <https://doi.org/10.1111/acel.13149>.
- Rohlenova, K., Goveia, J., García-Caballero, M., Subramanian, A., Kalucka, J., Treps, L., Falkenberg, K. D., de Rooij, L. P. M. H., Zheng, Y., Lin, L., Sokol, L., Teuwen, L., Geldhof, V., Taverna, F., et al. (2020). Single-Cell RNA sequencing maps endothelial metabolic plasticity in pathological angiogenesis. *Cell Metabolism*, 31(4). Available from <https://doi.org/10.1016/j.cmet.2020.03.009>, 862–877.e14.
- Römpp, A., & Spengler, B. (2013). Mass spectrometry imaging with high resolution in mass and space. *Histochemistry and Cell Biology*, 139(6), 759–783. Available from <https://doi.org/10.1007/s00418-013-1097-6>.
- Ron-Harel, N., Notarangelo, G., Ghergurovich, J. M., Paulo, J. A., Sage, P. T., Santos, D., Satterstrom, F. K., Gygi, S. P., Rabinowitz, J. D., Sharpe, A. H., & Haigis, M. C. (2018). Defective respiration and one-carbon metabolism contribute to impaired naïve T cell activation in aged mice. *Proceedings of the National Academy of Sciences USA*, 201804149. Available from <https://doi.org/10.1073/pnas.1804149115>.
- Rossi, D. J., Jamieson, C. H. M., & Weissman, I. L. (2008). Stems cells and the pathways to aging and cancer. *Cell*, 132(4), 681–696. Available from <https://doi.org/10.1016/j.cell.2008.01.036>.
- Sahin, E., & DePinho, R. A. (2010). Linking functional decline of telomeres, mitochondria and stem cells during aging. *Nature*, 464(7288), 520–528. Available from <https://doi.org/10.1038/nature08982>.
- Sahu, N., Vasu, D., Sahu, A., Lal, N., & Singh, S. K. (2017). Strength of microbes in nutrient cycling: A key to soil health. *Agriculturally Important Microbes for Sustainable Agriculture*, 69–86. Available from https://doi.org/10.1007/978-981-10-5589-8_4.
- San Martín, A., Sotelo-Hitschfeld, T., Lerchundi, R., Fernández-Moncada, I., Ceballo, S., Valdebenito, R., Baeza-Lehnert, F., Alegría, K., Contreras-Baeza, Y., Garrido-Gerter, P., Romero-Gómez, I., & Barros, L. F. (2014). Single-cell imaging tools for brain energy metabolism: A review. *Neurophotonics*, 1(1), 011004. Available from <https://doi.org/10.1117/1.nph.1.1.011004>.
- Saspotras, L., Turkcan, S., Pratx, G., & Gambhir, S. (2014). Single cell metabolomics in circulating tumor cells. *Journal of Nuclear Medicine*, 55(supplement 1), 8.

- Schüler, S. C., Gebert, N., & Ori, A. (2020). Stem cell aging: The upcoming era of proteins and metabolites. *Mechanisms of Aging and Development*, 111288. Available from <https://doi.org/10.1016/j.mad.2020.111288>.
- Senyo, S. E., Steinhauser, M. L., Pizzimenti, C. L., Yang, V. K., Cai, L., Wang, M., Wu, T. D., Guerquin-Kern, J.-L., Lechene, C. P., & Lee, R. T. (2012). Mammalian heart renewal by pre-existing cardiomyocytes. *Nature*, 493(7432), 433–436. Available from <https://doi.org/10.1038/nature11682>.
- Serasanambati, M., Broza, Y. Y., Marmur, A., & Haick, H. (2019). Profiling single cancer cells with volatolomics approach. *iScience*, 11, 178–188. Available from <https://doi.org/10.1016/j.isci.2018.12.008>.
- Shen, M., Qu, Z., DesLaurier, J., Welle, T. M., Sweedler, J. V., & Chen, R. (2018). Single synaptic observation of cholinergic neurotransmission on living neurons: Concentration and dynamics. *Journal of the American Chemical Society*, 140(25), 7764–7768. Available from <https://doi.org/10.1021/jacs.8b01989>.
- Shrestha, B., Patt, J. M., & Vertes, A. (2011). In situ cell-by-cell imaging and analysis of small cell populations by mass spectrometry. *Analytical Chemistry*, 83(8), 2947–2955. Available from <https://doi.org/10.1021/ac102958x>.
- Shrestha, B., Sripadi, P., Reschke, B. R., Henderson, H. D., Powell, M. J., Moody, S. A., & Vertes, A. (2014). Subcellular Metabolite and lipid analysis of *Xenopus laevis* eggs by LAESI mass spectrometry. *PLoS One*, 9(12), e115173. Available from <https://doi.org/10.1371/journal.pone.0115173>.
- Shrestha, B., & Vertes, A. (2009). In situ metabolic profiling of single cells by laser ablation electrospray ionization mass spectrometry. *Analytical Chemistry*, 81(20), 8265–8271. Available from <https://doi.org/10.1021/ac901525g>.
- Shyh-Chang, N., Daley, G. Q., & Cantley, L. C. (2013). Stem cell metabolism in tissue development and aging. *Development (Cambridge, England)*, 140(12), 2535–2547. Available from <https://doi.org/10.1242/dev.091777>.
- Si, X., Xiong, X., Zhang, S., Fang, X., & Zhang, X. (2017). Detecting low-abundance molecules at single-cell level by repeated ion accumulation in ion trap mass spectrometer. *Analytical Chemistry*, 89(4), 2275–2281. Available from <https://doi.org/10.1021/acs.analchem.6b03390>.
- Singh, R., Kumar, M., Mittal, A., & Mehta, P. K. (2016). Microbial enzymes: Industrial progress in 21st century. *3 Biotech*, 6(2), 174. Available from <https://doi.org/10.1007/s13205-016-0485-8>.
- Skiebe, P., Dreger, M., Meseke, M., Evers, J. F., & Hucho, F. (2002). Identification of orcokinins in single neurons in the stomatogastric nervous system of the crayfish, *Cherax destructor*. *The Journal of Comparative Neurology*, 444(3), 245–259. Available from <https://doi.org/10.1002/cne.10145>.
- Sousa Silva, M., Cordeiro, C., Roessner, U., & Figueiredo, A. (2019). Editorial: Metabolomics in crop research—current and emerging methodologies. *Frontiers in Plant Science*, 10, 1013. Available from <https://doi.org/10.3389/fpls.2019.01013>.
- Srivastava, N., Mishra, P. K., & Upadhyay, S. N. (2020). Significance of lignocellulosic biomass waste in the biofuel production process. In Srivastava (Ed.), *Industrial enzymes for biofuels production* (pp. 1–18). Elsevier, ISBN 9780128210109.
- Standke, S. J., Colby, D. H., Bensen, R. C., Burgett, A. W. G., & Yang, Z. (2019). Mass spectrometry measurement of single suspended cells using combined cell manipulation

- system and the single-probe device. *Analytical Chemistry*, 91(3), 1738–1742. Available from <https://doi.org/10.1021/acs.analchem.8b05774>.
- Streeter, J. G. (2003). Effect of trehalose on survival of *Bradyrhizobium japonicum* during desiccation. *Journal of Applied Microbiology*, 95(3), 484–491. Available from <https://doi.org/10.1046/j.1365-2672.2003.02017.x>.
- Su, Y., Ko, M. E., Cheng, H., Zhu, R., Xue, M., Wang, J., Lee, J. W., Frankiw, L., Xu, A., Wong, S., Robert, L., Takata, K., Yuan, D., Lu, Y., Huang, S., Ribas, A., Levine, R., Nolan, G. P., Wei, W., . . . Heath, J. R. (2020). Multi-omic single-cell snapshots reveal multiple independent trajectories to drug tolerance in a melanoma cell line. *Nature Communications*, 11(1). Available from <https://doi.org/10.1038/s41467-020-15956-9>.
- Sun, M., & Yang, Z. (2018a). Metabolomic studies of live single cancer stem cells using mass spectrometry. *Analytical Chemistry*, 91(3), 2384–2391. Available from <https://doi.org/10.1021/acs.analchem.8b05166>.
- Sun, M., Yang, Z., & Wawrik, B. (2018b). Metabolomic fingerprints of individual algal cells using the single-probe mass spectrometry technique. *Frontiers in Plant Science*, 9, 571. Available from <https://doi.org/10.3389/fpls.2018.00571>.
- Sun, W. H., Wei, Y., Guo, X. L., Wu, Q., Di, X., & Fang, Q. (2020). Nanoliter-scale droplet-droplet microfluidic microextraction coupled with MALDI–TOF mass spectrometry for metabolite analysis in cell droplets. *Analytical Chemistry*, 92(13), 8759–8767. Available from <https://doi.org/10.1021/acs.analchem.0c00007>.
- Suzuki, A., Stern, S. A., Bozdagi, O., Huntley, G. W., Walker, R. H., Magistretti, P. J., & Alberini, C. M. (2011). Astrocyte-neuron lactate transport is required for long-term memory formation. *Cell*, 144(5), 810–823. Available from <https://doi.org/10.1016/j.cell.2011.02.018>.
- Svatoš, A., Lee, Y. J., & Yang, Z. (2020). Editorial: Single plant cell metabolomics. *Frontiers in Plant Science*, 11, 161. Available from <https://doi.org/10.3389/fpls.2020.00161>.
- Tang, Y., Yang, X. K., Zhang, X. W., Wu, W. T., Zhang, F. L., Jiang, H., Liu, Y., Amatore, C., & Huang, W. H. (2019). Harpagide, a natural product, promotes synaptic vesicle release as measured by nanoelectrode amperometry. *Chemical Science*, 11(3), 778–785. Available from <https://doi.org/10.1039/c9sc05538j>.
- Tebani, A., & Bekri, S. (2019). Paving the way to precision nutrition through metabolomics. *Frontiers in Nutrition*, 6, 41. Available from <https://doi.org/10.3389/fnut.2019.00041>.
- Thiele, C., Wunderling, K., & Leyendecker, P. (2019). Multiplexed and single cell tracing of lipid metabolism. *Nature Methods*, 16, 1123–1130. Available from <https://doi.org/10.1038/s41592-019-0593-6>.
- Tuttle, J. R., Nah, G., Duke, M. V., Alexander, D. C., Guan, X., Song, Q., Chen, Z. J., Scheffle, B. E., & Haigler, C. H. (2015). Metabolomic and transcriptomic insights into how cotton fiber transitions to secondary wall synthesis, represses lignification, and prolongs elongation. *BMC Genomics*, 16(1). Available from <https://doi.org/10.1186/s12864-015-1708-9>.
- Tyagi, S., Raghvendra, U., Singh, T., Kalra, K., & Munjal. (2010). Applications of metabolomics—A systematic study of the unique chemical fingerprints: An overview. *International Journal of Pharmaceutical Sciences Review and Research*, 3, 83–86.
- Wang, R., Zhao, H., Zhang, X., Zhao, X., Song, Z., & Ouyang, J. (2019). Metabolic discrimination of breast cancer subtypes at single-cell level by multiple microextraction

- coupled with mass spectrometry. *Analytical Chemistry*, 91(5), 3667–3674. Available from <https://doi.org/10.1021/acs.analchem.8b05739>.
- Wang, Y., Noel, J. M., Velmurugan, J., Nogala, W., Mirkin, M. V., Lu, C., Collignon, M. G., Lemaître, F., & Amatore, C. (2012). Nanoelectrodes for determination of reactive oxygen and nitrogen species inside murine macrophages. *Proceedings of the National Academy of Sciences*, 109(29), 11534–11539. Available from <https://doi.org/10.1073/pnas.1201552109>.
- Watson, B. S., Bedair, M. F., Urbanczyk-Wojniak, E., Huhman, D. V., Yang, D. S., Allen, S. N., Li, W., Tang, Y., & Sumner, L. W. (2015). Integrated metabolomics and transcriptomics reveal enhanced specialized metabolism in *Medicago truncatula* root border cells. *Plant Physiology*, 167(4), 1699–1716. Available from <https://doi.org/10.1104/pp.114.253054>.
- Wedlock, L. E., Kilburn, M. R., Liu, R., Shaw, J. A., Berners-Price, S. J., & Farrell, N. P. (2013). NanoSIMS multi-element imaging reveals internalisation and nucleolar targeting for a highly-charged polynuclear platinum compound. *Chemical Communications*, 49(62), 6944. Available from <https://doi.org/10.1039/c3cc42098a>.
- Wei, Z., Xiong, X., Guo, C., Si, X., Zhao, Y., He, M., Yang, C., Xu, W., Tang, F., Fang, S., & Zhang, X. (2015). Pulsed direct current electrospray: Enabling systematic analysis of small volume sample by boosting sample economy. *Analytical Chemistry*, 87(22), 11242–11248. Available from <https://doi.org/10.1021/acs.analchem.5b02115>.
- Wu, D., Harrison, D., Mutlu, G. M., Huang, J., & Fang, Y. (2019). Single cell metabolism with deep learning reveals that a RhoA-mediated glycolytic burst drives endothelial cell contractions. C60. *Vascular Biology In Lung Disease*. Available from https://doi.org/10.1164/ajrccm-conference.2019.199.1_meetingabstracts.a741.
- Xu, J., Zhu, D., Ibrahim, A. D., Allen, C. C. R., Gibson, C. M., Fowler, P. W., Song, Y., & Huang, W. E. (2017). Raman deuterium isotope probing reveals microbial metabolism at the single-cell level. *Analytical Chemistry*, 89(24), 13305–13312. Available from <https://doi.org/10.1021/acs.analchem.7b03461>.
- Yamamoto, K., Takahashi, K., Mizuno, H., Anegawa, A., Ishizaki, K., Fukaki, H., & Mimura, T. (2016). Cell-specific localization of alkaloids in *Catharanthus roseus* stem tissue measured with Imaging MS and Single-cell MS. *Proceedings of the National Academy of Sciences, USA*, 113(14), 3891–3896. Available from <https://doi.org/10.1073/pnas.1521959113>.
- Yang, B., Patterson, N. H., Tsui, T., Caprioli, R. M., & Norris, J. L. (2018). Single-cell mass spectrometry reveals changes in lipid and metabolite expression in RAW 264.7 cells upon lipopolysaccharide stimulation. *Journal of the American Society for Mass Spectrometry*, 29(5), 1012–1020. Available from <https://doi.org/10.1007/s13361-018-1899-9>.
- Yao, H., Zhao, H., Zhao, X., Pan, X., Feng, J., Xu, F., Zhang, S., & Zhang, X. (2019). Label-free mass cytometry for unveiling cellular metabolic heterogeneity. *Analytical Chemistry*, 91(15), 9777–9783. Available from <https://doi.org/10.1021/acs.analchem.9b01419>.
- Yao, Y., Ji, J., Zhang, H., Zhang, K., Liu, B., & Yang, P. (2018). Three-dimensional plasmonic trap array for ultrasensitive surface-enhanced raman scattering analysis of single cells. *Analytical Chemistry*, 90(17), 10394–10399. Available from <https://doi.org/10.1021/acs.analchem.8b02252>.
- Yin, R., Prabhakaran, V., & Laskin, J. (2018). Quantitative extraction and mass spectrometry analysis at a single-cell level. *Analytical Chemistry*, 90(13), 7937–7945. Available from <https://doi.org/10.1021/acs.analchem.8b00551>.

- Ying, Y. L., Hu, Y.-X., Gao, R., Yu, R. J., Gu, Z., Lee, L. P., & Long, Y. T. (2018). Asymmetric nanopore electrode-based amplification for electron transfer imaging in live cells. *Journal of the American Chemical Society*, 140(16), 5385–5392. Available from <https://doi.org/10.1021/jacs.7b12106>.
- Zaitsu, K., Hayashi, Y., Murata, T., Ohara, T., Nakagiri, K., Kusano, M., Nakajima, H., Nakajima, T., Ishikawa, T., Tsuchihashi, H., & Ishii, A. (2016). Intact endogenous metabolite analysis of mice liver by probe electrospray ionization/triple quadrupole tandem mass spectrometry and its preliminary application to *in vivo* real-time analysis. *Analytical Chemistry*, 88(7), 3556–3561. Available from <https://doi.org/10.1021/acs.analchem.5b04046>.
- Zaoli, S., Giometto, A., Marañón, E., Escrig, S., Meibom, A., Ahluwalia, A., & Rinaldo, A. (2019). Generalized size scaling of metabolic rates based on single-cell measurements with freshwater phytoplankton. *Proceedings of the National Academy of Sciences, USA*, 116(35), 17323–17329. Available from <https://doi.org/10.1073/pnas.1906762116>.
- Zeng, C., Mulas, F., Sui, Y., Guan, T., Miller, N., Tan, Y., Liu, F., Jin, W., Carrano, A. C., Huising, M. O., Shirihai, O. S., Yeo, G. W., & Sander, M. (2017). Pseudotemporal ordering of single cells reveals metabolic control of postnatal β cell proliferation. *Cell Metabolism*, 25(5). Available from <https://doi.org/10.1016/j.cmet.2017.04.014>, 1160–1175.e11.
- Zenobi, R. (2013). Single-cell metabolomics: Analytical and biological perspectives. *Science (New York, N.Y.)*, 342(6163), 1243259. Available from <https://doi.org/10.1126/science.1243259>.
- Zhang, H., Lu, H., Huang, K., Li, J., Wei, F., Liu, A., Chingin, K., & Chen, H. (2020). Selective detection of phospholipids from human blood plasma and single cells for cancer differentiation using dispersed solid-phase microextraction combined with extractive electrospray ionization mass spectrometry. *The Analyst*. Available from <https://doi.org/10.1039/d0an01204a>.
- Zhang, L., & Vertes, A. (2015). Energy charge, redox state, and metabolite turnover in single human hepatocytes revealed by capillary microsampling mass spectrometry. *Analytical Chemistry*, 87(20), 10397–10405. Available from <https://doi.org/10.1021/acs.analchem.5b02502>.
- Zhao, G., Zhong, H., Rao, T., & Pan, Z. (2020). Metabolomic analysis reveals that the mechanism of astaxanthin improves the osteogenic differentiation potential in bone marrow mesenchymal stem cells. *Oxidative Medicine and Cellular Longevity*, 2020, 1–11. Available from <https://doi.org/10.1155/2020/3427430>.
- Zhao, J., Zhang, F., & Guo, Y. (2019). Quantitative analysis of metabolites at the single-cell level by hydrogen flame desorption ionization mass spectrometry. *Analytical Chemistry*, 91(4), 2752–2758. Available from <https://doi.org/10.1021/acs.analchem.8b04422>.
- Zheng, Y., Liu, Z., Xing, J., Zheng, Z., Pi, Z., Song, F., & Liu, S. (2020). In situ analysis of single cell and biological samples with rGO-Cu functional probe ESI-MS spectrometry. *Talanta*, 211, 120751. Available from <https://doi.org/10.1016/j.talanta.2020.120751>.
- Zhong, M., Lee, C. Y., Croushore, C. A., & Sweedler, J. V. (2012). Label-free quantitation of peptide release from neurons in a microfluidic device with mass spectrometry imaging. *Lab on a Chip*, 12(11), 2037. Available from <https://doi.org/10.1039/c2lc21085a>.

This page intentionally left blank

Gut microbiota-derived metabolites in host physiology

14

Francesco Strati and Federica Facciotti

Mucosal Immunology Lab, Department of Experimental Oncology, IEO—European Institute of Oncology IRCCS, Milan, Italy

Introduction

Estimates place the ratio between human and microbial cells in human body at 1:1 (Sender et al., 2016), with the gut microbiome contributing to the host genomic pool with >22 million genes, that is, exceeding the human genome by a factor of 1000 (Tierney et al., 2019). The output of this huge genomic diversity is reflected in the production of a diverse array of metabolites and other small molecules that affect host physiology.

The human gastrointestinal (GI) tract is inhabited by a complex and metabolically active ecological niche in which Archaea, Bacteria, Eukarya, and Viruses coexist in close association with the host (Huttenhower et al., 2012; Reyes et al., 2010; Strati et al., 2016a). This complex microbial community, known as gut microbiome, has coevolved with the host in a mutualistic relationship that influences many functions, from energy metabolism to immune system homeostasis and its metabolic activity is essential in maintaining host health (Maslowski & Mackay, 2011). Given the pivotal roles of the gut microbiota, changes in its composition may cause metabolic shifts that affect host phenotype. Indeed, alterations of the gut microbiome and its functionality have been mechanistically associated with an increasing number of pathological conditions such as metabolic disorders (e.g., diabetes, obesity) (Paun et al., 2019; Turnbaugh et al., 2009), blood pressure and heart disease (Abbasi, 2019; Holmes et al., 2008), autoimmune and CNS disorders (Huttenhower et al., 2012; Reyes et al., 2010; Strati et al., 2016b). The development of the gut microbiota starts at birth with colonization by a low number of species from the vaginal and fecal microbiota of the mother and is characterized by many shifts in composition during infancy (Palmer et al., 2007). From an initial gut microbiota composition characterized by low ecological richness and diversity during infancy, the adult intestinal microbiota becomes a complex ecosystem composed of thousands of species belonging principally to the bacterial phyla Firmicutes and Bacteroidetes. Actinobacteria, Fusobacteria, Proteobacteria and Verrucomicrobia are also present but in small proportions

(Huttenhower et al., 2012). Although the gut microbiome is highly malleable throughout lifespan, environmental factors such as diet (David et al., 2014) have the major impact on microbiota composition and are accountable for interindividual differences despite other important variables like age, sex and genetic factors (Lozupone et al., 2012; Sommer & Bäckhed, 2013). The switch from a high-fat/low-fiber diet to a low-fat/high fiber diet causes notable changes in the gut microbiota within 24 hours (David et al., 2014). Such changes reflect trade-offs between carbohydrate and protein fermentation. Production of microbial metabolites is therefore driven by a combination of dietary substrate availability and interindividual variability both within populations and across geographic or ethnic groups (Sonnenburg & Bäckhed, 2016). Cross sectional studies on different human populations have shown that there are evident differences between the gut microbiota of individuals from Western and non-Westernized countries (De Filippo et al., 2010; Yatsunenko et al., 2012). Although these differences may also be ascribed to the distinct genetic pools of these populations, cultural factors related to diet are critical in shaping the gut microbiome and consequently microbe-derived metabolites. Indeed, the populations having a diet rich in fibers tend to have a gut microbiota enriched in bacterial taxa bearing enzymatic repertoires for the hydrolysis of complex plant polysaccharides (De Filippo et al., 2010). The fermentation of such fibers produces relevant amounts of short chain fatty acids (SCFAs) which are important for colon health and are a primary energy source for colonic cells (Scheppach, 1994). By contrast, the gut microbiota of Western populations is deprived of these microorganisms and tends to produce less SCFAs (De Filippo et al., 2010) with important implication for host health. The reported effects of gut microbial metabolites on host health are numerous (and not limited to SCFA); in this chapter we describe the gut metabolome, the technics used to interrogate host–microbiota cometabolism of compounds, measurement of metabolites and integration of metabolomics with other multiomics data.

Metabolomics methods in host-microbiome studies

Metabolomics is the analysis of small molecules (<1500 Da) in biological specimens and a variety of high-throughput analytical techniques are suitable for microbiome research. Typically, these platforms are based on mass spectrometry (MS) or magnetic nuclear resonance (NMR) (see Chapter 4: Mass Spectrometry in Metabolomics and Chapter 5: Nuclear Magnetic Resonance in Metabolomics). With respect to DNA- and RNA-based sequencing, metabolomics allows the characterization of the actual molecular makeup of the microbiome, together with all the environmental and host small molecules present in the specimen, allowing to explore the effects of particular microbial modifiers on host-microbes interactions and host health (Holmes et al., 2011). In host-microbiome research, mass

spectrometry is becoming more common due to its high sensitivity, unbiased and high-throughput discovery, and the wide range of metabolites to be detected. Mass spectrometry is typically combined with an earlier separation method, such as gas or liquid chromatography, in order to reduce the high complexity of biological samples and improve resolution, specificity and quantification (see [Chapter 3: Separation Techniques](#)). Gas-chromatography mass spectrometry (GC-MS) is particularly suitable for the detection of volatile metabolites such as SCFA, but can be used also for the detection and quantification of simple sugars, amino acids and their derivates. For the study of both nonpolar metabolites such as bile acids (BAs) and lipids and polar metabolites, for example, purines, sugars, amino acids and vitamins, liquid chromatography mass spectrometry is more suited than GC-MS owing to the lower temperatures and weaker ionizations used during the process for the analysis of larger metabolites. NMR spectroscopy is also used in metabolomics, but typically it has poorer sensitivity than approaches focused on mass spectrometry. However, NMR allows the quantification of abundant metabolites with relatively easy sample preparation and provides structural and spatial information through magnetic resonance imaging. Furthermore, both MS and NMR can be used to trace metabolites and study the fate of microbiome metabolic activity through isotope labeling ([Berry & Loy, 2018](#); [Jang et al., 2018](#)), although understanding microbiota-specific metabolic flux and the origin of specific metabolites using this technique is challenging because of the shared metabolites by the microbiome and the host.

Metabolomics can be used to drive discovery or to test hypothesis and therefore, based on research needs, may use untargeted and targeted approaches, respectively (see [Chapter 6: Targeted Metabolomics](#) and [Chapter 7: Approaches in Untargeted Metabolomics](#)). Untargeted metabolomics, because of its “unbiased” approach targeting a wide variety of molecular classes, represents the election method for microbiome discovery programmes. However, this approach, yet powerful, shows important caveats. Indeed, different, nonstandardized, extraction methods for various molecular classes ([Mushtaq et al., 2014](#)), the variability linked to the diversity of biological matrixes, the limited number of molecular standards compared to the huge diversity of microbial products, many of which have not been previously characterized, its semiquantitative nature and lack of standardized bioinformatic pipelines and databases for annotation of spectral outputs hinder data integration at a systems biology level with other meta'omics data. On the contrary, targeted metabolomics, despite its inability to discover new associations outside of the molecular class of interest, offers optimized extraction protocols and known internal standards that allows better absolute quantification but require an experimental hypothesis to be tested.

At present, attempts to combine both targeted and untargeted approaches with metagenomics analysis of the gut microbiome are in place to identify microorganism–metabolite associations ([Melnik et al., 2017](#)). In this regard, upon identification of putative metabolite associations in health and disease, the use of germ-free mice and genetically manipulated microbes ([Dodd et al., 2017](#))

provides unique platforms to dissect mechanisms and investigate causality beyond merely correlation.

Fermentable metabolites and short chain fatty acids

Microbial metabolism of dietary nutrients results in the production of many metabolites. In particular, the complex dietary fibers that pass through the GI tract without undergoing significant transformations serve as substrate for gut microbes expanding the host's metabolic capacity (El Kaoutari et al., 2013). Up to 10% of the daily energy intake comes from dietary fibers but it can increase substantially upon higher fiber intake (McNeil, 1984) and in function of gut microbiota composition (El Kaoutari et al., 2013). The host indigestible carbohydrates that can be metabolized by the wide variety of microbial carbohydrate-active enzymes are known as microbiota-accessible carbohydrates (MAC). The major by-products of microbial fermentation of MACs are SCFAs (Cummings & Macfarlane, 1991) and much research has focused on their important implications in host physiology (Sonnenburg & Bäckhed, 2016). While most SCFAs are derived from dietary fiber, the colonic mucus layer is another source of MACs. Indeed, the highly *O*-glycosylated mucins, proteins present in the mucus and secreted by goblet cells in the small and large intestine, are a substrate for microbial exoglycosidases when dietary MACs are scarce fostering SCFA production (Desai et al., 2016; Martens et al., 2008). Different intestinal anaerobic bacteria belonging to Bacteroidetes, Firmicutes and some Clostridia genera such as *Faecalibacterium prausnitzii* and *Roseburia intestinalis*, harbor the capacity to convert MACs in SCFAs. Acetate, propionate and butyrate constitute the vast majority (>95%) of the produced microbial SCFA with the branched-chain fatty acids (BCFAs) isobutyrate, 2-methylbutyrate, isovalerate, succinate and lactate representing a smaller yet biologically active fraction of the SCFA pool (Koh et al., 2016). The gut microbiota may also ferment nutritional, host and microbial proteins, adding a limited amount to the overall SCFA reservoir from the branched-chain amino acids valine, leucine and isoleucine (Macfarlane et al., 1992).

SCFAs regulates several host physiological and biochemical functions affecting immune system homeostasis and response to inflammation in intestinal and extraintestinal diseases, modulating gut barrier integrity, host metabolism and acting as epigenetic regulators. SCFAs can be absorbed via passive transport by SLC5A8-dependent transit, or recognized by G-protein-coupled receptors (GPCRs) including GPR41 (FFAR3), GPR43 (FFAR2) and GPR109A (HCA2), among others (Husted et al., 2017; Koh et al., 2016).

SCFAs are important to modulate immune system homeostasis and inflammation in the gut acting on different immune cells. The differentiation of antiinflammatory Foxp3⁺T regulatory (Treg) cells can be modulated by SCFAs (Al Nabhani et al., 2019). Indeed, SCFA-mediated GPR43 signaling stimulates

interleukin-10 (IL-10)-producing Foxp3^+ Treg cell differentiation controlling barrier integrity (Macia et al., 2015) and protecting against colitis (Smith et al., 2013). This effect has been mechanistically illustrated in pivotal works demonstrating that different members of Clostridia induce Treg cells differentiation via SCFA production eliciting a TGF- β response in intestinal epithelial cells (Atarashi et al., 2011, 2013) or activating the MyD88/ROR γ t pathway in naïve Tregs (Abdel-Gadir et al., 2019). Butyrate interacts also on dendritic cells and macrophages through GPR109a fostering IL-10 production and boosting Treg development while controlling inflammation (Singh et al., 2014). The effects mediated by SCFAs are not linked only to GPCRs signaling. Indeed, SCFAs, in particular butyrate, can directly modulate histone posttranslational modifications acting as epigenetic regulators. Butyrate has been known to inhibit histone deacetylases (HDACs) since the 1970s (Riggs et al., 1977). The SCFA produced by Clostridia exert epigenetic control over TGF- β expression in Intestinal epithelial cells (IECs) by its HDAC activity (Martin-Gallausiaux et al., 2018). Similarly, acetate controls the differentiation of Th1 and Th17 cells to enhance immunity in a murine model of *Citrobacter rodentium* infection acting directly as an epigenetic regulator rather than through GPCRs (Park et al., 2015). Through the combination of GPCRs signaling and HDAC activity, SCFAs facilitate the secretion of mucins by goblet cells (Birchenough et al., 2016) and the production of antimicrobial peptides by IECs (Zhao et al., 2018). The effects exerted by SCFAs are not limited to the immune system and have been mechanistically linked to host metabolism. Indeed, acetate and butyrate may act preventing obesity by enhancing the production of the anorectic hormones GLP-1 and fasting peptide YY and therefore modulating food intake (Brooks et al., 2017; Chambers et al., 2015).

In addition to classical SCFAs (i.e., acetate, propionate and butyrate), microbial fermentation of dietary fibers can produce significant levels of succinate, a precursor of propionate metabolism and intermediate of the TCA cycle. Microbiota-produced succinate improves glycemic control through activation of intestinal gluconeogenesis (De Vadder et al., 2016). By acting through the succinate receptor GPR91 (or SUCNR1), succinate promotes robust type 2 immune responses to parasites (Schneider et al., 2018). Nevertheless, SUCNR1 is upregulated in the intestinal tissue of Crohn's disease patients (Macias-Ceja et al., 2019) and succinate acts as a significant pro-inflammatory signal in the host acting as a main mediator of macrophage response to lipopolysaccharide through IL-1 β (Mills et al., 2016). Lactate, which is also a propionate intermediate, harbors relevant metabolic and immunological properties and acts as an HDAC inhibitor and modulator of GPR81 signaling (Koh et al., 2016). Microbiota-derived lactate modulates intestinal CX3CR1 $^+$ cells to bind antigens in a GPR31-dependent manner activating adaptive immune response against *Salmonella* infection (Morita et al., 2019). High lactate levels, as a consequence of lactate-producing *Bifidobacterium* and *Lactobacillus* administration, stimulates stem cells-mediated epithelial development by Wnt/ β -catenin signals in Paneth cells and intestinal stromal cells via the lactate receptor GPR81 (Lee et al., 2018).

Collectively, SCFAs are the gut lumen's most abundant microbiome-derived metabolites and are endowed with the ability to regulate mucosal immunity and systemic metabolism primarily by triggering GPCRs or inducing HDACs to manipulate gene expression.

Secondary bile acids

Primary BAs are synthesized in the liver from cholesterol (dietary or endogenous). The human liver synthesizes only two primary BAs, cholic acid (CA) and chenodeoxycholic acid (CDCA). Prior to secretion, primary BAs are conjugated to taurine or glycine, passed into the gall bladder and later released in the small intestine to favor lipid digestion and absorption. More than 90% of primary BAs are reabsorbed through the enterohepatic circulation. Along the intestinal tract, primary BAs undergo several chemical transformations that are catalyzed by gut microbes, leading to the formation of secondary BAs. These enzymatic transformations deconjugate the amino acid moiety (i.e., the removal of the taurine/glycine moiety on the side chain) from primary BAs with microbial bile salt hydrolases; a narrower range of clostridial species can also convert primary to secondary BAs by the activity of $7\alpha/\beta$ -dehydroxylation enzymes increasing the diversity of the BA pool (Wahlström et al., 2016). This process results in the transformation of CA into deoxycholic acid (DCA) and CDCA into lithocholic acid (LCA), as well as other derivatives (Wahlström et al., 2016). Microbial BA transformations include deconjugation, oxidation and epimerization of hydroxyl groups, reduction of ketone groups, dehydroxylation (Edenharder, 1984), desulfation (Huijghebaert et al., 1984), esterification of hydroxyl groups (Kelsey et al., 1980) and the oxidation of a steroid ring to form unsaturated BAs (Prabha & Ohri, 2006). The effects exerted by BAs are mediated through the binding with two types of receptors: the farnesoid X receptor (FXR) and G-protein-coupled BA receptor 1 (also known as TGR5) although unconjugated BAs have also been shown to signal through pregnane X receptor, constitutive androstane receptor and vitamin D receptor (Jia et al., 2018). The unconjugated secondary BAs have a higher affinity for their receptors compared with the host liver-derived primary BAs (Wahlström et al., 2016).

BAs exert essential immunoregulatory and metabolic functions and have a strong influence on gut microbiota composition and density (Wahlström et al., 2016). DCA, LCA and ursodeoxycholic acid, through activation of FXR signaling, favor gut barrier integrity, intestinal crypt regeneration and wound repair (Jain et al., 2018; Wang et al., 2019). Secondary BAs also promote colonization resistance to *Clostridioides difficile* infection. Several studies have revealed that this effect is in part mediated by *Clostridium scindens*, a 7α -dehydroxylating gut bacterium, through the inhibitory role of 7α -dehydroxylated BAs on *C. difficile* spore germination and outgrowth (Buffie et al., 2015; Shen, 2015; Theriot et al., 2016).

The importance of secondary BAs is proved by the fact that dysbiosis can lead to an imbalance of primary and secondary BAs in the colon because of loss of BSH activity, transformation and desulfation of BAs (Ridlon et al., 2014). Intestinal inflammation, a consequence of dysbiosis, can lead to loss of absorption of BAs and increased colonic primary BAs. These imbalances have important repercussion toward inflammatory disorders such as IBDs and carcinogenesis, both sharing the common feature of chronic inflammation (Jia et al., 2018). BA–microbiome crosstalk has also been indicated in both hepatocellular carcinoma and colorectal cancer (CRC). For example, in response to production of secondary BAs by Clostridia, natural killer T cells accumulate in the liver inhibiting tumor growth and liver metastases, which are commonly derived from CRC (Ma et al., 2018). Thus, secondary BAs signaling have an important role in host physiology with distinct effects, from colonization resistance to inflammation and tumorigenesis, which are modulated by the composition of the gut microbiota.

Amino acids- and tryptophan-derived metabolites

Overall, between 6 and 18 g nitrogen containing compounds become available to the gut microbiota on a daily basis (Cummings & Macfarlane, 1991) as proteins and peptides. Indeed, roughly 5% of dietary proteins escape assimilation in the small intestine and therefore become available (Evenepoel et al., 1999) to the gut microbiota for assimilation or microbial secondary metabolism; otherwise proteins and peptides are excreted with feces. There is a trade-off between the fermentation of proteins vs carbohydrates by the gut microbiota, which is dictated by substrate availability. Carbohydrates fermentation drives bacterial growth and the subsequent production of SCFAs reduces luminal pH by limiting bacterial protease activity; thus, gut microbes tend to assimilate rather than ferment unabsorbed proteins in the colon. In case of low carbohydrate availability, bacterial fermentation of proteins occurs leading to reduced SCFA production and increased luminal pH that in turn alters the composition of the gut microbiota (Neis et al., 2015; Ratzke & Gore, 2018; Stephen & Cummings, 1980). SCFAs, BCFAs, indoles, amines, sulfides, ammonia, phenols and N-nitroso metabolites are produced by gut microbes upon fermentation of proteins (P.M. Smith et al., 2013). However, owing to the huge diversity of nitrogenous compounds, this chapter will focus only on bioactive amino acid derivates with known activity on host physiology.

Tryptophan is an essential amino acid commonly found in protein-rich foods such as egg, milk, beans and nuts. Because of its structure that can undergo several biochemical transformations and its relatively low abundance as free amino acid within cells, tryptophan derivates play a central role in interkingdom signalling between microbiota and the host. Tryptophan is the precursor for the synthesis of several important bioactive molecules, such as serotonin, melatonin and nicotinamide and its catabolism may follow one of three pathways in the gut. More than 90% of dietary tryptophan is metabolized through the kynurenine pathway in

mucosal and immune cells by the indoleamine 2,3-dioxygenase 1 enzyme (IDO1) and in the liver through the activity of the tryptophan 2,3-dioxygenase (Cervenka et al., 2017). IDO1 expression is also modulated by the gut microbiota (Atarashi et al., 2011). The serotonin pathway in enterochromaffin cells is controlled by the rate-limiting enzyme tryptophan hydroxylase 1 (TPH1) where 5-hydroxytryptamine (i.e., serotonin) is mainly produced (Côté et al., 2003). Finally, the gut microbiota may convert tryptophan in indole derivates (including indoleacetic acid, indole-3-acetaldehyde, indole-3-aldehyde, indoleacrylic acid and indole-3-propionic acid) that act as ligands of the aryl hydrocarbon receptor (AhR) in the host cells (Agus et al., 2018). A limited number of bacteria, including *Peptostreptococcus russellii* and members of *Lactobacillus*, are known to produce AhR agonists, and indole-3-propionic acid production has been characterized in *Clostridium sporogenes* (Wikoff et al., 2009). These indole derivates play a central role in immune system homeostasis. In response to dietary tryptophan, commensal lactobacilli produce AhR ligands (in particular the indole-3-aldehyde) inducing production of IL-22 by ILC3s cells (Zelante et al., 2013) that, in turn, controls *Candida* colonization in the gut. At the mucosal site, fungal colonization induces the production of IL-17 and IFN γ that is a strong activator of IDO1 (Bozza et al., 2005). The activation of the IFN γ -IDO1 axis harmonizes the tolerogenic response against the fungal microbiota allowing commensalism (Romani, 2011). Nevertheless, commensal fungi such as *Candida albicans*, may shift IDO's activity from kynurenine toward 5-hydroxytryptophan, an inhibitor of Th17 host responses (Cheng et al., 2010) promoting transition from yeast to hyphal morphology (Bozza et al., 2005), thus from commensalism to infection (Romani, 2011). CARD9 (caspase recruitment domain-containing protein 9, an IBD susceptibility gene)-knockout mice are characterized by impaired immune responses and susceptibility to colitis because of decreased levels of indole-3-acetic acid and insufficient IL-22 production by ILCs. The enhanced susceptibility of CARD9 $^{-/-}$ mice to colitis has been linked to the inability of the gut microbiome to convert tryptophan into AhR ligands. Supplementation of AhR-activating strains of the genus *Lactobacillus* reduces the severity of colitis (Lamas et al., 2016). In addition, the commensal *P. russellii* can reduce susceptibility to colitis by metabolizing tryptophan into indoleacrylic acid enhancing goblet cells differentiation and preventing inflammation (Włodarska et al., 2017). A similar mechanism controlling systemic inflammation has been described in response to indole-3-propionic acid, which is produced by *C. sporogens* (Jennis et al., 2018).

Microbiota-derived tryptophan metabolites modulates also host metabolism increasing the release of GLP-1 secretion. Administration of AhR ligands or a *Lactobacillus* strain with a high AhR ligand-production capacity leads to improvement of both dietary- and genetic-induced metabolic impairments (Chimerel et al., 2014; Natividad et al., 2018).

The gut microbiota has also an important role in modulating the serotonergic system since it can induce transcription of *Tph1* and subsequently serotonin production (Reigstad et al., 2015; Yano et al., 2015). Serotonin levels and related metabolites are altered in the striatum and hippocampus of germ-free mice

(Clarke et al., 2013; Heijtz et al., 2011). Colonization by the gut microbiota in turn increases the levels of plasma serotonin by circa threefold (Wikoff et al., 2009). This effect may be in part mediated by the production of microbiota-derived metabolites such as SCFAs (Reigstad et al., 2015) and the secondary BA deoxycholate (Yano et al., 2015) since there is no evidence of direct synthesis of 5-HT by the gut microbiota; nevertheless, different gut microbes harbor the genomic potential for 5-HT biosynthesis (Valles-Colomer et al., 2019).

In addition to tryptophan, microbial metabolism of glutamate, arginine, histidine, tyrosine and phenylalanine have been described with important effects on host metabolism. Some members of the gut microbiota, principally lactobacilli, bifidobacteria and *Bacteroides spp.*, can produce the neurotransmitter γ -aminobutyric acid (GABA) through different pathways (decarboxylation of glutamate, degradation of putrescine, or from arginine or ornithine) (Barrett et al., 2012; Lyte, 2013; Strandwitz et al., 2019) with implication for anxiety-like behaviors and depression (Bravo et al., 2011; Strandwitz et al., 2019).

Imidazole propionate, a by-product of histidine metabolism of *Streptococcus mutans* and *Escherichia coli*, impairs insulin signaling contributing to type 2 diabetes pathogenesis (Koh et al., 2018). Similarly, the gut microbiome can metabolize dietary tyrosine into the precursor phenyl sulfate through the enzyme tyrosine phenol-lyase. Together with the tryptophan derivative indoxyl sulfate, the tyrosine and phenylalanine derivatives p-cresol sulfate and phenylacetylglutamine, these microbiota-dependent metabolites contribute to pathogenesis and disease progression in renal diseases by inducing renal damage, inflammation and fibrosis (Devlin et al., 2016; Kikuchi et al., 2019; Vanholder et al., 2014).

Thus, amino acids and tryptophan derivatives from the gut microbiota exert pleiotropic activities within the host, similarly to SCFAs, from immunomodulatory to central nervous system functions.

Additional microbially derived metabolites

Fermentation of dietary substrates by the gut microbiota produces also bioactive polyphenols (Makki et al., 2018). Polyphenols are a family of structurally diverse organic compounds abundant in plants and characterized by multiple phenolic units. Based on structural similarity they are classified as phenolic acids, flavonoids, stilbenes, lignins, lignans and coumarins. Dietary polyphenols, because of their large structural diversity and extensive metabolism, have a poor bioavailability and only a fraction ($\sim 10\%$) is absorbed in the small intestine. The unabsorbed fraction passes through the large bowel undergoing extensive metabolism by gut microbes or being excreted (Cardona et al., 2013). Despite antiinflammatory, antioxidant and antimicrobial properties have been attributed to polyphenols, due to metabolism, rapid excretion and difficulty to assess their biological effects *in vivo*, polyphenols health claims are still undetermined (EFSA Panel on Dietetic Products, 2010).

Indeed, the lack of validated *in vivo* biomarkers to be used in long term studies and the unphysiological concentrations used in the *in vitro* tests undermine the health-promoting effects attributed to these dietary compounds (D'Archivio et al., 2010). However, it has been observed that polyphenols may influence the microbial community structure of the gut microbiota. Recent studies suggested also that polyphenols can exert measurable effects due to the metabolism of parental compounds by the gut microbiota (Cardona et al., 2013). *Citrus* flavonoids, for example, showed protection toward dysmetabolic conditions in a microbiota-dependent manner (Zeng et al., 2020). Desaminotyrosine, a microbiota-derived metabolite produced by *Clostridium orbiscindens* from flavonoids has been shown to boost type I IFN responses protecting from influenza infections (Steed et al., 2017).

One-carbon metabolism involves a variety of intertwined metabolic processes that include the methionine and folate cycles, supplying 1C units (methyl groups) for DNA, polyamines, amino acids, creatine, and phospholipids synthesis (Clare et al., 2019). Because of its universal and interkingdom nature, one-carbon metabolism is of central importance in host-microbiota crosstalk. Indeed, B vitamins are essential cofactors of this metabolic pathway. Host synthesis of B vitamins is not sufficient to maintain host health and therefore essential B vitamins are obtained from diet and from *de novo* synthesis by the gut microbiota. Gut microbiota-derived B vitamins are even more relevant of dietary intake in the case of folate (Aufreiter et al., 2009). Nearly the 50% of human gut microbes have the genomic potential to synthesize B vitamins and the majority of these genome-based predictions matches published experimental data (Magnúsdóttir et al., 2015).

Choline is an essential nutrient for humans. It must be obtained from the diet as choline or as choline phospholipids since its *de novo* production is not sufficient to maintain health. The microbial metabolism of choline and choline derivates, that is, phosphatidylcholine, produces high levels of trimethylamine (TMA) (Craciun & Balskus, 2012). Once absorbed from the gut, TMA circulates to the liver and is enzymatically oxidized by flavin-containing monooxygenases to TMA N-oxide (TMAO). TMAO has been associated with cardiovascular disease and atherosclerosis in mice and humans (Jain et al., 2018; K. Wang et al., 2019). TMA production serves as an excellent example of the interaction between the diet and the microbiota. Indeed, germ-free mice cannot produce TMA and antibiotic treatment of conventionally raised mice leads to reduced TMA formation. Moreover, transplantation of choline-converting bacteria to gnotobiotic mice can increase TMA production (Romano et al., 2015). Dietary L-carnitine, which is rich in red meat, is also metabolized by the gut microbiota into TMA. Accordingly, vegans are poor producers of TMA (Koeth et al., 2013).

Perspectives and future directions

The understanding that dysbiosis is in causal relationship with different diseases is stimulating translational research toward the use of microbiome-based

therapies. At present, microbiota-derived metabolites have already provided to be valuable biomarkers in cardiovascular disease, chronic kidney disease, and type 2 diabetes (Devlin et al., 2016; Koh et al., 2018; Tang et al., 2013) in addition to traditional risk factors. The attractiveness of microbiome-based therapies, that is, next generation probiotics and live biotherapeutic products also relies on their ability to deliver metabolites or their precursors at physiologically relevant concentrations directly in the gut since it would be otherwise pharmacologically challenging to deliver these metabolites via oral administration. The potential for targeted microbiome–metabolite therapeutics is provided, for example, by the improved resistance to *C. difficile* infection upon colonization with *C. scindens* through the conversion of CA to DCA by 7 α -dehydrogenation (Buffie et al., 2015; Francis et al., 2013). These studies provided the promise for the rational design of microbiota-derived metabolite therapies. Indeed, oral administration of the secondary BAs ursodeoxycholic acid has shown to be therapeutically effective in patients with *C. difficile* infection-associated pouchitis (Weingarden et al., 2016). A placebo-controlled, crossover trial is currently underway to determine the effect of oral 5-hydroxytryptophan administration on fatigue in patients with IBD (NCT03574948). Niacin (nicotinic acid) also known as vitamin B3, beneficially affects insulin sensitivity in patients by targeting the gut microbiome (Fangmann et al., 2018) and its application shows clinical promise in the treatment of Ulcerative Colitis patients (Li et al., 2017). The delivery of microbiota-derived tryptamine from tryptophan, a reaction catalyzed by the gut microbes *C. sporogenes* and *Ruminococcus gnavus*, accelerates GI transit time (Bhattarai et al., 2018; Williams et al., 2014) with important implications in pathologies such as irritable bowel syndrome.

These and other examples that have been treated in this chapter highlight the power of metabolite-based research strategies to move forward microbiome-based therapies in medicine. Indeed, metabolites are a more robust clinical end-point than microbial taxa, since carriage of metabolic pathways in the gut microbiome is stable among individuals despite variation in microbial community structure (Huttenhower et al., 2012).

References

- Abbasi, J. (2019). TMAO and heart disease: The new red meat risk? *Jama*, 321(22), 2149–2151.
- Abdel-Gadir, A., Stephen-Victor, E., Gerber, G. K., Rivas, M. N., Wang, S., Harb, H., Wang, L., Li, N., Crestani, E., & Spielman, S. (2019). Microbiota therapy acts via a regulatory T cell MyD88/ROR γ t pathway to suppress food allergy. *Nature Medicine*, 25(7), 1164–1174.
- Agus, A., Planchais, J., & Sokol, H. (2018). Gut microbiota regulation of tryptophan metabolism in health and disease. *Cell Host & Microbe*, 23(6), 716–724.

- Al Nabhani, Z., Dulauroy, S., Marques, R., Cousu, C., Al Bounny, S., Déjardin, F., Sparwasser, T., Bérard, M., Cerf-Bensussan, N., & Eberl, G. (2019). A weaning reaction to microbiota is required for resistance to immunopathologies in the adult. *Immunity*, 50(5), 1276–1288.
- Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., Fukuda, S., Saito, T., Narushima, S., & Hase, K. (2013). T reg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature*, 500(7461), 232–236.
- Atarashi, K., Tanoue, T., Shima, T., Imaoka, A., Kuwahara, T., Momose, Y., Cheng, G., Yamasaki, S., Saito, T., & Ohba, Y. (2011). Induction of colonic regulatory T cells by indigenous Clostridium species. *Science*, 331(6015), 337–341.
- Aufreiter, S., Gregory, J. F., III, Pfeiffer, C. M., Fazili, Z., Kim, Y.-I., Marcon, N., Kamalaporn, P., Pencharz, P. B., & O'Connor, D. L. (2009). Folate is absorbed across the colon of adults: Evidence from cecal infusion of ¹³C-labeled [6S]-5-formyltetrahydrofolic acid. *The American Journal of Clinical Nutrition*, 90(1), 116–123.
- Barrett, E., Ross, R., O'Toole, P. W., Fitzgerald, G. F., & Stanton, C. (2012). γ -Aminobutyric acid production by culturable bacteria from the human intestine. *Journal of Applied Microbiology*, 113(2), 411–417.
- Berry, D., & Loy, A. (2018). Stable-isotope probing of human and animal microbiome function. *Trends in Microbiology*, 26(12), 999–1007.
- Bhattarai, Y., Williams, B. B., Battaglioli, E. J., Whitaker, W. R., Till, L., Grover, M., Linden, D. R., Akiba, Y., Kandimalla, K. K., & Zachos, N. C. (2018). Gut microbiota-produced tryptamine activates an epithelial G-protein-coupled receptor to increase colonic secretion. *Cell Host & Microbe*, 23(6), 775–785.
- Birchenough, G. M., Nyström, E. E., Johansson, M. E., & Hansson, G. C. (2016). A sentinel goblet cell guards the colonic crypt by triggering Nlrp6-dependent Muc2 secretion. *Science*, 352(6293), 1535–1542.
- Bozza, S., Fallarino, F., Pitzurra, L., Zelante, T., Montagnoli, C., Bellocchio, S., Mosci, P., Vacca, C., Puccetti, P., & Romani, L. (2005). A crucial role for tryptophan catabolism at the host/Candida albicans interface. *The Journal of Immunology*, 174(5), 2910–2918.
- Bravo, J. A., Forsythe, P., Chew, M. V., Escaravage, E., Savignac, H. M., Dinan, T. G., Bienenstock, J., & Cryan, J. F. (2011). Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences*, 108(38), 16050–16055.
- Brooks, L., Viardot, A., Tsakmaki, A., Stolarczyk, E., Howard, J. K., Cani, P. D., Everard, A., Sleeth, M. L., Psichas, A., & Anastasovskaj, J. (2017). Fermentable carbohydrate stimulates FFAR2-dependent colonic PYY cell expansion to increase satiety. *Molecular Metabolism*, 6(1), 48–60.
- Buffie, C. G., Bucci, V., Stein, R. R., McKenney, P. T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., & Viale, A. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. *Nature*, 517(7533), 205–208.
- Cardona, F., Andrés-Lacueva, C., Tulipani, S., Tinahones, F. J., & Queipo-Ortuño, M. I. (2013). Benefits of polyphenols on gut microbiota and implications in human health. *The Journal of Nutritional Biochemistry*, 24(8), 1415–1422.
- Cervenka, I., Agudelo, L. Z., & Ruas, J. L. (2017). Kynurenines: Tryptophan's metabolites in exercise, inflammation, and mental health. *Science*, 357(6349).

- Chambers, E. S., Viardot, A., Psichas, A., Morrison, D. J., Murphy, K. G., Zac-Varghese, S. E., MacDougall, K., Preston, T., Tedford, C., & Finlayson, G. S. (2015). Effects of targeted delivery of propionate to the human colon on appetite regulation, body weight maintenance and adiposity in overweight adults. *Gut*, 64(11), 1744–1754.
- Cheng, S.-C., van de Veerdonk, F., Smeekens, S., Joosten, L. A., van der Meer, J. W., Kullberg, B.-J., & Netea, M. G. (2010). *Candida albicans* dampens host defense by downregulating IL-17 production. *The Journal of Immunology*, 185(4), 2450–2457.
- Chimerel, C., Emery, E., Summers, D. K., Keyser, U., Gribble, F. M., & Reimann, F. (2014). Bacterial metabolite indole modulates incretin secretion from intestinal enteroendocrine L cells. *Cell Reports*, 9(4), 1202–1208.
- Clare, C. E., Brassington, A. H., Kwong, W. Y., & Sinclair, K. D. (2019). One-carbon metabolism: Linking nutritional biochemistry to epigenetic programming of long-term development. *Annual Review of Animal Biosciences*, 7, 263–287.
- Clarke, G., Grenham, S., Scully, P., Fitzgerald, P., Moloney, R. D., Shanahan, F., Dinan, T. G., & Cryan, J. F. (2013). The microbiome-gut-brain axis during early life regulates the hippocampal serotonergic system in a sex-dependent manner. *Molecular Psychiatry*, 18(6), 666–673.
- Côté, F., Thévenot, E., Fligny, C., Fromes, Y., Darmon, M., Ripoche, M.-A., Bayard, E., Hanoun, N., Saurini, F., & Lechat, P. (2003). Disruption of the nonneuronal tph1 gene demonstrates the importance of peripheral serotonin in cardiac function. *Proceedings of the National Academy of Sciences*, 100(23), 13525–13530.
- Craciun, S., & Balskus, E. P. (2012). Microbial conversion of choline to trimethylamine requires a glycol radical enzyme. *Proceedings of the National Academy of Sciences*, 109(52), 21307–21312.
- Cummings, J., & Macfarlane, G. (1991). The control and consequences of bacterial fermentation in the human colon. *Journal of Applied Bacteriology*, 70(6), 443–459.
- D'Archivio, M., Filesi, C., Varì, R., Scazzocchio, B., & Masella, R. (2010). Bioavailability of the polyphenols: Status and controversies. *International Journal of Molecular Sciences*, 11(4), 1321–1342.
- David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., & Fischbach, M. A. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484), 559–563.
- De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poulet, J. B., Massart, S., Collini, S., Pieraccini, G., & Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences*, 107(33), 14691–14696.
- De Vadder, F., Kovatcheva-Datchary, P., Zitoun, C., Duchampt, A., Bäckhed, F., & Mithieux, G. (2016). Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metabolism*, 24(1), 151–157.
- Desai, M. S., Seekatz, A. M., Koropatkin, N. M., Kamada, N., Hickey, C. A., Wolter, M., Pudlo, N. A., Kitamoto, S., Terrapon, N., & Muller, A. (2016). A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. *Cell*, 167(5), 1339–1353.
- Devlin, A. S., Marcabal, A., Dodd, D., Nayfach, S., Plummer, N., Meyer, T., Pollard, K. S., Sonnenburg, J. L., & Fischbach, M. A. (2016). Modulation of a circulating uremic solute via rational genetic manipulation of the gut microbiota. *Cell Host & Microbe*, 20(6), 709–715.

- Dodd, D., Spitzer, M. H., Van Treuren, W., Merrill, B. D., Hryckowian, A. J., Higginbottom, S. K., Le, A., Cowan, T. M., Nolan, G. P., & Fischbach, M. A. (2017). A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. *Nature*, 551(7682), 648–652.
- Edenharder, R. (1984). Dehydroxylation of cholic acid at C12 and epimerization at C5 and C7 by *Bacteroides* species. *Journal of Steroid Biochemistry*, 21(4), 413–420.
- EFSA Panel on Dietetic Products, N. and A. (NDA). (2010). Scientific Opinion on the substantiation of health claims related to various food (s)/food constituent (s) and protection of cells from premature aging, antioxidant activity, antioxidant content and antioxidant properties, and protection of DNA, proteins and lipids from oxidative damage pursuant to Article 13 (1) of Regulation (EC) No 1924/2006. *EFSA Journal*, 8(2), 1489.
- El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D., & Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7), 497–504.
- Evenepoel, P., Claus, D., Geypens, B., Hiele, M., Geboes, K., Rutgeerts, P., & Ghoos, Y. (1999). Amount and fate of egg protein escaping assimilation in the small intestine of humans. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 277 (5), G935–G943.
- Fangmann, D., Theismann, E.-M., Türk, K., Schulte, D. M., Relling, I., Hartmann, K., Keppler, J. K., Knipp, J.-R., Rehman, A., & Heinsen, F.-A. (2018). Targeted microbiome intervention by microencapsulated delayed-release niacin beneficially affects insulin sensitivity in humans. *Diabetes Care*, 41(3), 398–405.
- Francis, M. B., Allen, C. A., Shrestha, R., & Sorg, J. A. (2013). Bile acid recognition by the *Clostridium difficile* germinant receptor, CspC, is important for establishing infection. *PLoS Pathogens*, 9(5), e1003356.
- Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., Hibberd, M. L., Forssberg, H., & Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences*, 108(7), 3047–3052.
- Holmes, E., Li, J. V., Athanasiou, T., Ashrafi, H., & Nicholson, J. K. (2011). Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends in Microbiology*, 19(7), 349–359.
- Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., & Veselkov, K. A. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453(7193), 396–400.
- Huijghebaert, S., Parmentier, G., & Eyissen, H. (1984). Specificity of bile salt sulfatase activity in man, mouse and rat intestinal microflora. *Journal of Steroid Biochemistry*, 20(4), 907–912.
- Husted, A. S., Trauelsen, M., Rudenko, O., Hjorth, S. A., & Schwartz, T. W. (2017). GPCR-mediated signaling of metabolites. *Cell Metabolism*, 25(4), 777–796.
- Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., & Fulton, R. S. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207.
- Jain, U., Lai, C.-W., Xiong, S., Goodwin, V. M., Lu, Q., Muegge, B. D., Christophi, G. P., VanDussen, K. L., Cummings, B. P., & Young, E. (2018). Temporal regulation of the bacterial metabolite deoxycholate during colonic repair is critical for crypt regeneration. *Cell Host & Microbe*, 24(3), 353–363.

- Jang, C., Hui, S., Lu, W., Cowan, A. J., Morscher, R. J., Lee, G., Liu, W., Tesz, G. J., Birnbaum, M. J., & Rabinowitz, J. D. (2018). The small intestine converts dietary fructose into glucose and organic acids. *Cell Metabolism*, 27(2), 351–361.
- Jennis, M., Cavanaugh, C., Leo, G., Mabus, J., Lenhard, J., & Hornby, P. (2018). Microbiota-derived tryptophan indoles increase after gastric bypass surgery and reduce intestinal permeability in vitro and in vivo. *Neurogastroenterology & Motility*, 30(2), e13178.
- Jia, W., Xie, G., & Jia, W. (2018). Bile acid–microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. *Nature Reviews Gastroenterology & Hepatology*, 15 (2), 111.
- Kelsey, M., Molina, J., Huang, S., & Hwang, K. (1980). The identification of microbial metabolites of sulfolithocholic acid. *Journal of Lipid Research*, 21(6), 751–759.
- Kikuchi, K., Saigusa, D., Kanemitsu, Y., Matsumoto, Y., Thanai, P., Suzuki, N., Mise, K., Yamaguchi, H., Nakamura, T., & Asaji, K. (2019). Gut microbiome-derived phenyl sulfate contributes to albuminuria in diabetic kidney disease. *Nature Communications*, 10 (1), 1–17.
- Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., & Li, L. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature Medicine*, 19(5), 576–585.
- Koh, A., De Vadder, F., Kovatcheva-Datchary, P., & Bäckhed, F. (2016). From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell*, 165 (6), 1332–1345.
- Koh, A., Molinaro, A., Ståhlman, M., Khan, M. T., Schmidt, C., Mannerås-Holm, L., Wu, H., Carreras, A., Jeong, H., & Olofsson, L. E. (2018). Microbially produced imidazole propionate impairs insulin signaling through mTORC1. *Cell*, 175(4), 947–961.
- Lamas, B., Richard, M. L., Leducq, V., Pham, H.-P., Michel, M.-L., Da Costa, G., Bridonneau, C., Jegou, S., Hoffmann, T. W., & Natividad, J. M. (2016). CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands. *Nature Medicine*, 22(6), 598–605.
- Lee, Y.-S., Kim, T.-Y., Kim, Y., Lee, S.-H., Kim, S., Kang, S. W., Yang, J.-Y., Baek, I.-J., Sung, Y. H., & Park, Y.-Y. (2018). Microbiota-derived lactate accelerates intestinal stem-cell-mediated epithelial development. *Cell Host & Microbe*, 24(6), 833–846.
- Li, J., Kong, D., Wang, Q., Wu, W., Tang, Y., Bai, T., Guo, L., Wei, L., Zhang, Q., & Yu, Y. (2017). Niacin ameliorates ulcerative colitis via prostaglandin D2-mediated D prostanoid receptor 1 activation. *EMBO Molecular Medicine*, 9(5), 571–588.
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415), 220–230.
- Lyte, M. (2013). Microbial endocrinology in the microbiome-gut-brain axis: How bacterial production and utilization of neurochemicals influence behavior. *PLoS Pathogens*, 9 (11), e1003726.
- Ma, C., Han, M., Heinrich, B., Fu, Q., Zhang, Q., Sandhu, M., Agdashian, D., Terabe, M., Berzofsky, J. A., & Fako, V. (2018). Gut microbiome–mediated bile acid metabolism regulates liver cancer via NKT cells. *Science*, 360, 6391.
- Macfarlane, G., Gibson, G., Beatty, E., & Cummings, J. (1992). Estimation of short-chain fatty acid production from protein by human intestinal bacteria based on branched-chain fatty acid measurements. *FEMS Microbiology Ecology*, 10(2), 81–88.

- Macia, L., Tan, J., Vieira, A. T., Leach, K., Stanley, D., Luong, S., Maruya, M., McKenzie, C. I., Hijikata, A., & Wong, C. (2015). Metabolite-sensing receptors GPR43 and GPR109A facilitate dietary fiber-induced gut homeostasis through regulation of the inflammasome. *Nature Communications*, 6(1), 1–15.
- Macias-Ceja, D. C., Ortiz-Masiá, D., Salvador, P., Gisbert-Ferrández, L., Hernández, C., Hausmann, M., Rogler, G., Esplugues, J. V., Hinojosa, J., & Alós, R. (2019). Succinate receptor mediates intestinal inflammation and fibrosis. *Mucosal Immunology*, 12(1), 178–187.
- Magnúsdóttir, S., Ravcheev, D., de Crécy-Lagard, V., & Thiele, I. (2015). Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. *Frontiers in Genetics*, 6, 148.
- Makki, K., Deehan, E. C., Walter, J., & Bäckhed, F. (2018). The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host & Microbe*, 23(6), 705–715.
- Martens, E. C., Chiang, H. C., & Gordon, J. I. (2008). Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host & Microbe*, 4(5), 447–457.
- Martin-Gallausiaux, C., Béguet-Crespel, F., Marinelli, L., Jamet, A., Ledue, F., Blottièvre, H. M., & Lapaque, N. (2018). Butyrate produced by gut commensal bacteria activates TGF-beta1 expression through the transcription factor SP1 in human intestinal epithelial cells. *Scientific Reports*, 8(1), 1–13.
- Maslowski, K. M., & Mackay, C. R. (2011). Diet, gut microbiota and immune responses. *Nature Immunology*, 12(1), 5–9.
- McNeil, N. (1984). The contribution of the large intestine to energy supplies in man. *The American Journal of Clinical Nutrition*, 39(2), 338–342.
- Melnik, A. V., da Silva, R. R., Hyde, E. R., Aksenen, A. A., Vargas, F., Bouslimani, A., Protsyuk, I., Jarmusch, A. K., Tripathi, A., & Alexandrov, T. (2017). Coupling targeted and untargeted mass spectrometry for metabolome-microbiome-wide association studies of human fecal samples. *Analytical Chemistry*, 89(14), 7549–7559.
- Mills, E. L., Kelly, B., Logan, A., Costa, A. S., Varma, M., Bryant, C. E., Tourlomousis, P., Däbritz, J. H. M., Gottlieb, E., & Latorre, I. (2016). Succinate dehydrogenase supports metabolic repurposing of mitochondria to drive inflammatory macrophages. *Cell*, 167(2), 457–470.
- Morita, N., Umemoto, E., Fujita, S., Hayashi, A., Kikuta, J., Kimura, I., Haneda, T., Imai, T., Inoue, A., & Mimuro, H. (2019). GPR31-dependent dendrite protrusion of intestinal CX3CR1+ cells by bacterial metabolites. *Nature*, 566(7742), 110–114.
- Mushtaq, M. Y., Choi, Y. H., Verpoorte, R., & Wilson, E. G. (2014). Extraction for metabolomics: Access to the metabolome. *Phytochemical Analysis*, 25(4), 291–306.
- Natividad, J. M., Agus, A., Planchais, J., Lamas, B., Jarry, A. C., Martin, R., Michel, M.-L., Chong-Nguyen, C., Roussel, R., & Straube, M. (2018). Impaired aryl hydrocarbon receptor ligand production by the gut microbiota is a key factor in metabolic syndrome. *Cell Metabolism*, 28(5), 737–749.
- Neis, E. P., DeJong, C. H., & Rensen, S. S. (2015). The role of microbial amino acid metabolism in host metabolism. *Nutrients*, 7(4), 2930–2946.
- Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., & Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biology*, 5(7), e177.
- Park, J., Kim, M., Kang, S. G., Jannasch, A. H., Cooper, B., Patterson, J., & Kim, C. H. (2015). Short-chain fatty acids induce both effector and regulatory T cells by

- suppression of histone deacetylases and regulation of the mTOR–S6K pathway. *Mucosal Immunology*, 8(1), 80–93.
- Paun, A., Yau, C., Meshkibaf, S., Daigneault, M. C., Marandi, L., Mortin-Toth, S., Bar-Or, A., Allen-Vercoe, E., Poussier, P., & Danska, J. S. (2019). Association of HLA-dependent islet autoimmunity with systemic antibody responses to intestinal commensal bacteria in children. *Science Immunology*, 4(32).
- Prabha, V., & Ohri, M. (2006). Bacterial transformations of bile acids. *World Journal of Microbiology and Biotechnology*, 22(2), 191–196.
- Ratzke, C., & Gore, J. (2018). Modifying and reacting to the environmental pH can drive bacterial interactions. *PLoS Biology*, 16(3), e2004248.
- Reigstad, C. S., Salmonson, C. E., R., J. F., III, Szurszewski, J. H., Linden, D. R., Sonnenburg, J. L., Farrugia, G., & Kashyap, P. C. (2015). Gut microbes promote colonic serotonin production through an effect of short-chain fatty acids on enterochromaffin cells. *The FASEB Journal*, 29(4), 1395–1403.
- Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., & Gordon, J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304), 334–338.
- Ridlon, J. M., Kang, D. J., Hylemon, P. B., & Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Current Opinion in Gastroenterology*, 30(3), 332.
- Riggs, M., Whittaker, R., Neumann, J., & Ingram, V. (1977). n-Butyrate causes histone modification in HeLa and Friend erythroleukaemia cells. *Nature*, 268(5619), 462–464.
- Romani, L. (2011). Immunity to fungal infections. *Nature Reviews Immunology*, 11(4), 275–288.
- Romano, K. A., Vivas, E. I., Amador-Noguez, D., & Rey, F. E. (2015). Intestinal microbiota composition modulates choline bioavailability from diet and accumulation of the proatherogenic metabolite trimethylamine-N-oxide. *MBio*, 6, 2.
- Scheppach, W. (1994). Effects of short chain fatty acids on gut morphology and function. *Gut*, 35(1 Suppl.), S35–S38.
- Schneider, C., O'Leary, C. E., von Moltke, J., Liang, H.-E., Ang, Q. Y., Turnbaugh, P. J., Radhakrishnan, S., Pellizzon, M., Ma, A., & Locksley, R. M. (2018). A metabolite-triggered tuft cell-ILC2 circuit drives small intestinal remodeling. *Cell*, 174(2), 271–284.
- Sender, R., Fuchs, S., & Milo, R. (2016). Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3), 337–340.
- Shen, A. (2015). A gut odyssey: The impact of the microbiota on *Clostridium difficile* spore formation and germination. *PLoS Pathog*, 11(10), e1005157.
- Singh, N., Gurav, A., Sivaprakasam, S., Brady, E., Padia, R., Shi, H., Thangaraju, M., Prasad, P. D., Manicassamy, S., & Munn, D. H. (2014). Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity*, 40(1), 128–139.
- Smith, P. M., Howitt, M. R., Panikov, N., Michaud, M., Gallini, C. A., Bohlooly-y, M., Glickman, J. N., & Garrett, W. S. (2013). The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science*, 341(6145), 569–573.
- Sommer, F., & Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. *Nature Reviews Microbiology*, 11(4), 227–238.
- Sonnenburg, J. L., & Bäckhed, F. (2016). Diet–microbiota interactions as moderators of human metabolism. *Nature*, 535(7610), 56–64.

- Steed, A. L., Christophi, G. P., Kaiko, G. E., Sun, L., Goodwin, V. M., Jain, U., Esaulova, E., Artyomov, M. N., Morales, D. J., & Holtzman, M. J. (2017). The microbial metabolite desaminotyrosine protects from influenza through type I interferon. *Science*, 357 (6350), 498–502.
- Stephen, A. M., & Cummings, J. H. (1980). Mechanism of action of dietary fiber in the human colon. *Nature*, 284(5753), 283–284.
- Strandwitz, P., Kim, K. H., Terekhova, D., Liu, J. K., Sharma, A., Levering, J., McDonald, D., Dietrich, D., Ramadhar, T. R., & Lekbua, A. (2019). GABA-modulating bacteria of the human gut microbiota. *Nature Microbiology*, 4(3), 396–403.
- Strati, F., Cavalieri, D., Albanese, D., De Felice, C., Donati, C., Hayek, J., Jousson, O., Leoncini, S., Pindo, M., & Renzi, D. (2016a). Altered gut microbiota in Rett syndrome. *Microbiome*, 4(1), 1–15.
- Strati, F., Di Paola, M., Stefanini, I., Albanese, D., Rizzetto, L., Lionetti, P., Calabò, A., Jousson, O., Donati, C., & Cavalieri, D. (2016b). Age and gender affect the composition of fungal population of the human gastrointestinal tract. *Frontiers in Microbiology*, 7, 1227.
- Tang, W. W., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., Wu, Y., & Hazen, S. L. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *New England Journal of Medicine*, 368(17), 1575–1584.
- Theriot, C. M., Bowman, A. A., & Young, V. B. (2016). Antibiotic-induced alterations of the gut microbiota alter secondary bile acid production and allow for *Clostridium difficile* spore germination and outgrowth in the large intestine. *MSphere*, 1(1).
- Tierney, B. T., Yang, Z., Luber, J. M., Beaudin, M., Wibowo, M. C., Baek, C., Mehlenbacher, E., Patel, C. J., & Kostic, A. D. (2019). The landscape of genetic content in the gut and oral human microbiome. *Cell Host & Microbe*, 26(2), 283–295.
- Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., & Affourtit, J. P. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228), 480–484.
- Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., Schiweck, C., Kurilshikov, A., Joossens, M., & Wijmenga, C. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature Microbiology*, 4(4), 623–632.
- Vanholder, R., Schepers, E., Pletinck, A., Nagler, E. V., & Glorieux, G. (2014). The uremic toxicity of indoxyl sulfate and p-cresyl sulfate: A systematic review. *Journal of the American Society of Nephrology*, 25(9), 1897–1907.
- Wahlström, A., Sayin, S. I., Marschall, H.-U., & Bäckhed, F. (2016). Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metabolism*, 24(1), 41–50.
- Wang, K., Liao, M., Zhou, N., Bao, L., Ma, K., Zheng, Z., Wang, Y., Liu, C., Wang, W., & Wang, J. (2019). Parabacteroides distasonis alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. *Cell Reports*, 26(1), 222–235.
- Weingarden, A. R., Chen, C., Zhang, N., Graizer, C. T., Dosa, P. I., Steer, C. J., Shaughnessy, M. K., Johnson, J. R., Sadowsky, M. J., & Khoruts, A. (2016). Ursodeoxycholic acid inhibits *Clostridium difficile* spore germination and vegetative growth, and prevents recurrence of ileal pouchitis associated with the infection. *Journal of Clinical Gastroenterology*, 50(8), 624.

- Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., & Siuzdak, G. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences*, *106* (10), 3698–3703.
- Williams, B. B., Van Benschoten, A. H., Cimermancic, P., Donia, M. S., Zimmermann, M., Taketani, M., Ishihara, A., Kashyap, P. C., Fraser, J. S., & Fischbach, M. A. (2014). Discovery and characterization of gut microbiota decarboxylases that can produce the neurotransmitter tryptamine. *Cell Host & Microbe*, *16*(4), 495–503.
- Włodarska, M., Luo, C., Kolde, R., d'Hennezel, E., Annand, J. W., Heim, C. E., Krastel, P., Schmitt, E. K., Omar, A. S., & Creasey, E. A. (2017). Indoleacrylic acid produced by commensal *peptostreptococcus* species suppresses inflammation. *Cell Host & Microbe*, *22*(1), 25–37.
- Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., Nagler, C. R., Ismagilov, R. F., Mazmanian, S. K., & Hsiao, E. Y. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*, *161*(2), 264–276.
- Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., & Anokhin, A. P. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227.
- Zelante, T., Iannitti, R. G., Cunha, C., De Luca, A., Giovannini, G., Pieraccini, G., Zecchi, R., D'Angelo, C., Massi-Benedetti, C., & Fallarino, F. (2013). Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity*, *39*(2), 372–385.
- Zeng, S.-L., Li, S.-Z., Xiao, P.-T., Cai, Y.-Y., Chu, C., Chen, B.-Z., Li, P., Li, J., & Liu, E.-H. (2020). Citrus polymethoxyflavones attenuate metabolic syndrome by regulating gut microbiome and amino acid metabolism. *Science Advances*, *6*(1), eaax6208.
- Zhao, Y., Chen, F., Wu, W., Sun, M., Bilotta, A. J., Yao, S., Xiao, Y., Huang, X., Eaves-Pyles, T. D., & Golovko, G. (2018). GPR43 mediates microbiota metabolite SCFA regulation of antimicrobial peptide expression in intestinal epithelial cells via activation of mTOR and STAT3. *Mucosal Immunology*, *11*(3), 752–762.

Further reading

- Britton, G. J., Contijoch, E. J., Mogno, I., Vennaro, O. H., Llewellyn, S. R., Ng, R., Li, Z., Mortha, A., Merad, M., & Das, A. (2019). Microbiotas from humans with inflammatory bowel disease alter the balance of gut Th17 and ROR γ t + regulatory T cells and exacerbate colitis in mice. *Immunity*, *50*(1), 212–224.
- Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schreter, C. E., Rocha, S., & Grdinaru, V. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell*, *167* (6), 1469–1480.
- Sharon, G., Cruz, N. J., Kang, D.-W., Gandal, M. J., Wang, B., Kim, Y.-M., Zink, E. M., Casey, C. P., Taylor, B. C., & Lane, C. J. (2019). Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell*, *177*(6), 1600–1618.
- Smith, E. A., & Macfarlane, G. T. (1996). Enumeration of human colonic bacteria producing phenolic and indolic compounds: Effects of pH, carbohydrate availability and

- retention time on dissimilatory aromatic amino acid metabolism. *Journal of Applied Bacteriology*, 81(3), 288–302.
- Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., Feldstein, A. E., Britt, E. B., Fu, X., & Chung, Y.-M. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, 472(7341), 57–63.
- Zhu, W., Gregory, J. C., Org, E., Buffa, J. A., Gupta, N., Wang, Z., Li, L., Fu, X., Wu, Y., & Mehrabian, M. (2016). Gut microbial metabolite TMAO enhances platelet hyperactivity and thrombosis risk. *Cell*, 165(1), 111–124.

MALDI–mass spectrometry imaging: the metabolomic visualization 15

Emanuela Salviati, Eduardo Sommella, and Pietro Campiglia

Department of Pharmacy, University of Salerno, Fisciano, Salerno, Italy

Introduction

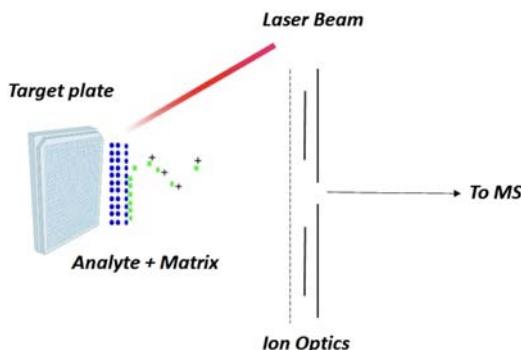
Metabolomics is often defined as a “snapshot” of living systems, capable to provide molecular information in real time, which help in the comprehension of how physiopathological events are regulated in living systems (Fiehn, 2002). The elucidation of metabolites in cells, tissues, organs or whole systems is extremely challenging and currently no single analytical method is capable to take the whole “picture” (Nicholson & Lindon, 2008). Mass spectrometry (MS) based metabolomics has proved its potential for the analysis of both polar and nonpolar metabolites, especially in combination with chromatographic methods, and MS analyzers become more and more sensitive, accurate and fast (Dettmer et al., 2007). MS coupled to different separation techniques is the gold standard for the analysis of biofluids, on the other hand, the analysis of metabolites in tissues requires their homogenization and further extraction, which causes the loss of “spatial information”. In modern, personalized medicine, there is for instance, the need of chemical characterization of tissue biopsies, or the evaluation of the drug distribution, and their effect in target organs. Imaging techniques, allowing the spatial localization of molecules, have revolutionized medicine, helping in diagnosis and understanding the pathologies. Techniques such as magnetic resonance imaging (MRI), positron emission tomography (PET), fluorescence microscopy and radiolabeling can visualize localization of target molecules, but these techniques can only monitor few molecules at time and require the use of molecular tags or labels and a priori knowledge of target compounds. MS imaging (MSI) is a powerful, label-free, analytical tool that allows mapping of the spatial distributions and the relative abundance of a wide variety of biomolecules, such as proteins, peptides, lipids, metabolites, carbohydrates and drugs in tissue samples (Caprioli et al., 1997; Norris & Caprioli, 2013). Since molecular labeling is not necessary, this technique is able to detect hundreds of molecules simultaneously in a single experiment, with subsequent *in situ* visualization of the localization of each ion in a single tissue section (Caprioli et al., 1997; Shariatgorji et al., 2016). The mass to charge ratio (m/z) of measured ions are correlated to a specific location on the

tissue, and thus the MS data are translated into images, which represent independent measurements of the detected compounds. Plotting the ion intensity in a *x* and *y* graph allows the visualization of the molecular content in a 2D map. From the first application of MSI, initially developed as a tool for intact protein imaging from the tissue surface (Norris & Caprioli, 2013), the technique has observed a rapid growth in each aspect: sample preparation, ion generation, analyzers technology and data elaboration, together with its employment in a wide range of applications in the biomedical field.

Among the several MS ionization techniques that can be used to directly analyze tissues, matrix-assisted laser desorption ionization (MALDI) has led the way in development of biological and clinical applications for MSI. This chapter describes the essential considerations for performing MALDI MSI. This chapter is not intended to be comprehensive with respect to all aspects of MSI, such as other ionization techniques (e.g., secondary ion MS, desorption electrospray ionization) or application in the proteomics field, rather it focuses on the basic concepts of MALDI-MSI, the current state and recent advances in instrumentation and method development for its application in the analysis low-molecular weight endogenous metabolites. Examples of applications are presented to highlight the potential of MSI in the fields of metabolomics and lipidomics.

Basics of MALDI mass spectrometry imaging

MSI is a multistep process involving sample preparation, analyte desorption and ionization, mass analysis and image registration. The summary of a workflow is reported in Fig. 15.1 (Schwamborn & Caprioli, 2010). The MALDI–MSI basic rely on the application of a ionization agent, that is UV-absorbing “matrix,” to the sample (in MSI a very thin tissue slice), which is usually applied as saturated solution, and the sample cocrystallizes with the matrix when the solvent evaporates. The matrix has the scope to help the desorption and ionization process after the irradiation of the cocrystallized sample with a UV laser (mostly with nitrogen lasers at 337 nm) with fast repetition rates (up to 10 kHz). The matrix absorbs the laser energy and helps in the transfer and ionization of the analytes to the gas phase, the ionized analytes are then separated by the MS analyzer. The analysis is carried out by an automatic raster of the tissue by the laser, which generates a mass spectrum at each *x* and *y* coordinates of the raster. Such as electrospray ionization, MALDI is a “soft” ionization technique and possesses the great advantage that can be applied to very different molecular weight compounds, from small metabolites, to large proteins, this allows its application in very different biological fields.

**FIGURE 15.1**

Principles and Instrumentation of MALDI-MSI. Basic MALDI-MS workflow, starting with desorption step, ionization and transfer of gas-phase analytes to MS analyzer.

Matrix choice and application

Sample preparation strategy is critical for every MS method, but it is certainly the crucial step in MSI. In fact, the choice of the matrix used and the application (coating) on the tissue has a depth impact on the overall quality of the experiment (Schwamborn & Caprioli, 2010). A wide range of matrices are now available, and their development continues. The properties of MALDI matrices are to promote ionization of the targeted analytes without producing matrix-derived peaks in the m/z region of interest that could obscure the signals of the targeted molecules. More in detail, the requirements of a MALDI matrix are:

1. A strong absorption at the typical emission wavelength of the laser used (337 or 355 nm). In this regard, almost all organic MALDI matrix contain an aromatic ring with delocalized π electrons.
2. The matrix must promote analyte ionization, and in this regard usually the matrix also contains a carboxylic acid moiety that ensures the protonation of the analytes and helps the solubilization of the matrix in aqueous solvents.
3. The matrix should be stable as long as possible under the high vacuum conditions ($1 \times 10^{-8/-9}$) of the MS instrument, failing in this leads to a change in the matrix/analyte ratio and thus compromises the repeatability of the experiment.
4. The matrix should avoid the formation of analyte clusters, such as analyte dimers, which could complicate the spectra and reduce sensitivity.
5. Matrix cocrystallization should be as homogeneous as possible to ensure repeatability from the different laser shots on the tissue.

Among the most employed matrices there are: α -cyano-4-hydroxycinnamic acid (CHCA), 2,5-dihydroxybenzoic acid (DHB), sinapinic acid (SA) and 9-aminoacridine (9-AA). Their suitability depends on the chemical nature of the

analyte and the matrix application and solubilization differs from matrix to matrix. While MALDI–MSI has proved to be a powerful tool for the analysis of proteins and peptides, it should be pointed out that the properties of small polar metabolites and lipids differ from peptides and proteins. In fact, the analysis of low molecular weight metabolites can be challenging, common matrices such as CHCA and DHB create a large background of signals under m/z 500, which add complexity to the spectrum. Among the common matrices, DHB and CHCA are popular for polar and lipid metabolites in positive mode. In negative mode, 9-aminoacridine (9-AA) and 1,5-diaminonaphthalene (DAN) have been also employed (Korte & Lee, 2014). Among other matrices, [1,8- bis(dimethylamino) naphthalene] (DMAN) (Shroff & Svatoš, 2009) and its derivative 1,8-Di(piperidinyl)-naphthalene (Weiβfloga & Svatoš, 2016) showed promising results in the analysis of lipids. Liquid Ionic Matrixes (LIMs) based on DHB were also tested with success for lipids analysis resulting in improved spatial resolution and detections sensitivity (Meriaux et al., 2010). Recently, a reactive matrix that selectively targets phenolic and primary amine groups of neurotransmitters and their metabolites was developed, based on nucleophilic aromatic substitution reaction of the 2-fluoro-1-methyl pyridinium (FMP) cation with phenolic hydroxyl and primary and secondary amine groups of neurotransmitter molecules. This was successfully applied to the imaging of neurotransmitters in rat brain with a lateral resolution of 10 μm (Shariatgorji et al., 2019). The matrix should be applied in a standardized manner allowing both desorption of the analytes and ensuring a homogenous coating while minimizing analyte delocalization, clearly the target is high signal intensity, spatial resolution and repeatability. Among different techniques for matrix deposition, there are manual spray, robotic sprayers and sublimation, each with strengths and limitations. When the biological task necessitates imaging of the whole tissue specimen, matrix can be applied uniformly to the entire surface to be imaged in one of two ways. Manual spraying of matrix is cheap and easy to implement in laboratory, this can be accomplished by glass reagent sprayers such those used for reagent deposition in thin layer chromatography. In this case the matrix is in a reagent reservoir and the tissue is sprayed usually 10–20 times. This process can lead to very high resolution images but heavily depends on the operator skill and thus can suffer of poor repeatability. To increase and automate the process, several robotic sprayers are available on the market, such as the HTX TM–Sprayer (HTX Technologies, LLC, NC, United States), SunCollect (SunChrome, Friedrichsdorf, Germany), ImagePrep (Bruker Daltonics, MA, United States), SMALDIPrep (TransMIT, Gießen, Germany) (Nishidate et al., 2019). These instruments are based on the same mechanism of manual spray but in this case the matrix solution is pumped as a fine aerosol and in an automatic and programmable manner, which can be fully adjustable depending on the tissue and spatial resolution required. Clearly, the advantage in this case is a very repeatable coating and small crystals, but as drawback all these systems are characterized by high cost. Sublimation is an approach that provide uniform coating (Hankin et al., 2007), in this process the matrix is put in a sublimation chamber,

while the tissue to be analyzed is placed inverted over the top of the matrix bed. By heat application the matrix sublimation is obtained and it condense on the surface of the cold tissue. By the variation of heating time specific thickness are obtained. This process leads to very fine crystals and this method is highly suitable when high spatial resolution imaging is required. The choice of the solvent, pH and additives for matrix application can lead to differences in the crystallization. Furthermore, the tissue nature, for example, lipid content, can influence crystal formation.

Tissue preparation for MALDI mass spectrometry imaging analysis

The MALDI MSI experiment starts with the tissue collection, experiments require animal or human tissue sections. MSI has been applied to a variety of samples such as plant or animal tissues, human biopsies, whole animal, cells and organoids. (Goodwin, 2012). Usually, immediately after the collection, samples are snap-freezed (liquid nitrogen or isopentane) and the material is stored at -80°C , different approaches are based on alcohol preserved, or formaldehyde-fixed and paraffin-embedded (FFPE) tissues. This last method has been used for many years in biobanks, tissues were kept at room temperature resulting in metabolite degradation, therefore unless for specific applications (proteomics) these fixation strategies are not recommended (Chugtai & Heeren, 2010). Another important aspect is the sample handling that must maintain the integrity and spatial organization of the molecules in biological samples, and avoid the distortion and fracturing of the tissues. The tissue preparation procedure needs still to be standardized for consistency because tissue handling and sample preparation are critical steps which can affect reproducibility and comparability of molecular end points. Postmortem changes can lead to alteration in analyte concentration, in fact the activity of many enzymes, including proteases is still extensive even 3 minutes after death (Svensson et al., 2007). In this regard, the timing is essential: Snap-freezing procedures in dissected organs from mice can take seconds, while for whole animals minutes, on the contrary formalin fixation can take on average 24 hours. It is important to remark that for snap-freezing protocols when tissues are defrosted to be used for tissue cutting and/or matrix deposition the enzymatic processes can start again, in particular proteases (Goodwin, 2012), therefore alternative stabilization protocols have been developed during the years. In this regard, heat stabilization is a powerful option. In this process, a combination of heat and pressure is applied on the tissue that acts dually preventing sample deformation and inactivating the enzymes involved in rapid degradation/turnover of biomolecules in tissues (Goodwin et al., 2010). The temperature applied on the tissue is increased quickly up to 90°C , but does not exceed 95°C in any part of the sample. This process has been now available thanks to commercial solution such as the Stabilizer T1,

Denator AB (<http://www.denator.com>). After the tissue collection, it must be sectioned in very thin (and flat) slices that will be subsequently covered with the matrix. Tissue sectioning is usually performed with a cryostat microtome, to perform this operation the embedding of the tissue is required, to allow an easy and precise handling of the tissue. Fig. 15.2 reports the synthesis of sample preparation for MALDI-MSI.

The embedding media are usually polymers such as carboxymethylcellulose or optimal cutting temperature media. It must be pointed out that this operation require some skill from the operator, besides, these polymers easily ionize in MALDI and a contamination of the tissue can lead to ion suppression and high background noise (Chughtai & Heeren, 2010), thus is imperative to avoid tissue contamination from these material. Ice or gelatine are other common strategies as embedding material that can lead to cleaner background (Chen et al., 2009). Typical tissue slices for MALDI–MSI require a thickness between 10 and 20 μm , especially when working with animal models, the temperature chamber of the cryomicrotome is typically kept between -5°C and -25°C and it depends from the tissue and fat content. The obtained tissue slices should be placed on a surface for subsequent analysis, the most common method is thaw mounting on an electrically conductive steel plate or glass slide. The thaw mounting operation requires the gently transfer, usually with the help of an artist brush, of the tissue slice on the surface of the slide, and then warming the slide and tissue together from the underside. This method is widely employed, nevertheless it can cause risk of inter and intrasample variation because of the different timing in tissue warming and attachment. Larger section (e.g., whole body) cannot prepared in this way, and in this regard tape mounting is employed. In this case adhesive tape is attached to the surface of the embedded tissue prior to section cutting. For each approach used, it is mandatory to not induce alterations on the tissue such as scratches or the formation of bubbles and edges on the tissues. The sections are finally mounted on transparent conductive glass slides, the most employed are indium-tin

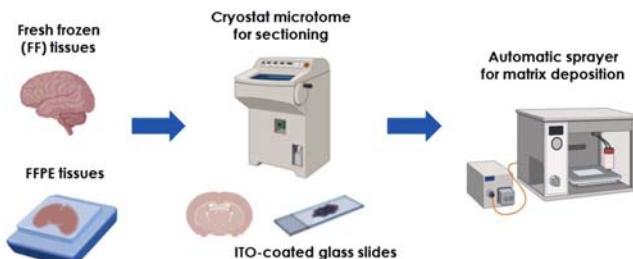


FIGURE 15.2

MALDI–MSI sample preparation workflow schematics. MALDI–MSI sample preparation workflow from tissue preparation (either fresh frozen or FFPE), tissue cutting by cryotome, and finally matrix deposition by automated sprayer.

oxide and then blocked on stainless steel MALDI targets, these slides allow also microscopic observation of MALDI samples. After mounting, the tissue slices are dried in a vacuum desiccator for at least 30 minutes to avoid moisture condensation that could cause analyte delocalization. Generally in proteomics/peptidomics applications tissues are subjected to washing step to remove abundant lipids that could suppress ionization of peptides (Kaletaş et al., 2009), but this step is not recommended for small metabolites and clearly skipped for lipidomics applications.

MALDI mass spectrometry imaging instrumentation

Given the spread of the MALDI–MSI in the pharmaceutical and biomedical field, a large number of instrumentations and solutions is now available from different vendors in all-in-one solutions, which can be summarized in two points:

1. Hardware (source and analyzer)
2. Software (data acquisition and visualization of mass spectrometry imaging data).

The parameters that are used for analyzers classification are resolution, mass accuracy, sensitivity, dynamic range, and tandem MS (MS/MS) capability. Before MS acquisition, it is essential to define the desired spatial resolution, which has a significant weight on the resulting quality and molecular detail of the overall MALDI–MSI analysis output. Clearly, an increase in spatial resolution is counterbalanced by a loss in throughput, sensitivity, and an increase in the size of the resulting datafiles. Current available MALDI instruments possess a laser focus from 200 to 20 μm , the most recent sources such as the SmartbeamTM 3D of Bruker features a laser beam of 5 μm . The choice of laser beam settings is not available on all instruments. The most employed analyzers for MALDI–MSI are Time of flight (TOF), Fourier transform ion cyclotron resonance (FT-ICR) and Orbital ion traps (Orbitrap). The largest amount of MALDI–MSI instruments are with no doubt TOF analyzers, with TOF-TOF and quadrupole-TOF (qTOF) being the most employed. TOF analyzers possess good transmission ratio (50%–100%), high sensitivity and dynamic range, especially for high molecular weight compounds ($> 5 \text{ KDa}$), furthermore MS/MS capability and the addition of ion mobility (IMS) that separates into the gas phase isobaric ions based on their cross collisional sections, further extend the capability of these analyzers with structural elucidation capability and increase resolution, respectively (McLean et al., 2007). When higher resolution is required, FT-ICR instruments possess unique capability, such as unmatched resolution capable to resolve MS signals differing of few mDa, and subppm mass accuracy that allows unambiguous molecular formula determination, furthermore, their sensitivity and dynamic range can be increased by in cell accumulation or continuous accumulation of selected ions (Amstalden van Hove et al., 2010). High resolving power coupled with MS/MS capability of Orbitrap MS devices allows the structural elucidation of isobaric analytes with

high confidence (Landgraf et al., 2009). Throughput is another important aspect, that must be taken into account when high number of samples must be analyzed. In this regard latest development in analyzers technology have reached important reduction in terms of time, for instance TOF based analyzer (TissuetyperTM from Bruker) can acquire data with a laser repetition rate of 10 kHz and an acquisition of 50 pixels/s, resulting in a drastic reduction of analysis time. Data acquisition and elaboration of MALDI–MSI files is a critical step. First software control should be designed in order to define the location of the target on the MALDI plate, usually by drawing lines around the tissue area of interest, which usually is also acquired as high resolution image in a scanner, to mark teaching or reference points on the stainless steel plate. The acquisition of mass spectra occurs in automatic way by moving the laser shots in a raster pattern. MSI data elaboration is based in the conversion of mass spectra into images, and up to now a large number of software are available which can be open access or proprietary, the last are commonly included in instrument package. A first consideration regarding MSI files is the size. Common MSI datafiles are usually in the Gigabyte range, depending on the spatial resolution and on the type of MS data (e.g., FTMS or TOF-MS spectra). This results in very demanding requirements for workstation and time consuming step in data import and export. Among opensource software there is Biomap (Novartis, Basel, Switzerland, <https://Ms-imaging.org/wp/biomap/>). The software was written in Interactive Data Language (IDL), it was previously developed for other imaging techniques and then extended to handle MSI data. It allows the selection of ROI, overlay of images and other tools such as baseline correction and spectra averaging. Datacube explorer (<https://amolf.nl/download/datacubeexplorer>) developed at FOM-institute AMOLF, similarly to Biomap allows different functions, in addition both image and spectrum-based dataview, automatic extraction of images from data and smoothing, it reads imzML file formats. Among commercial solution, the FlexImaging series from Bruker usually is present in the MALDI-TOF or FT-ICR software packages, the software allows a color-coded visualization of any ions detected, by the overlaying optical and MS images, other tools such as normalization, hierarchical clustering, principal component analysis, ROI selection and baseline correction. The main limitation is that only Bruker files can be used.

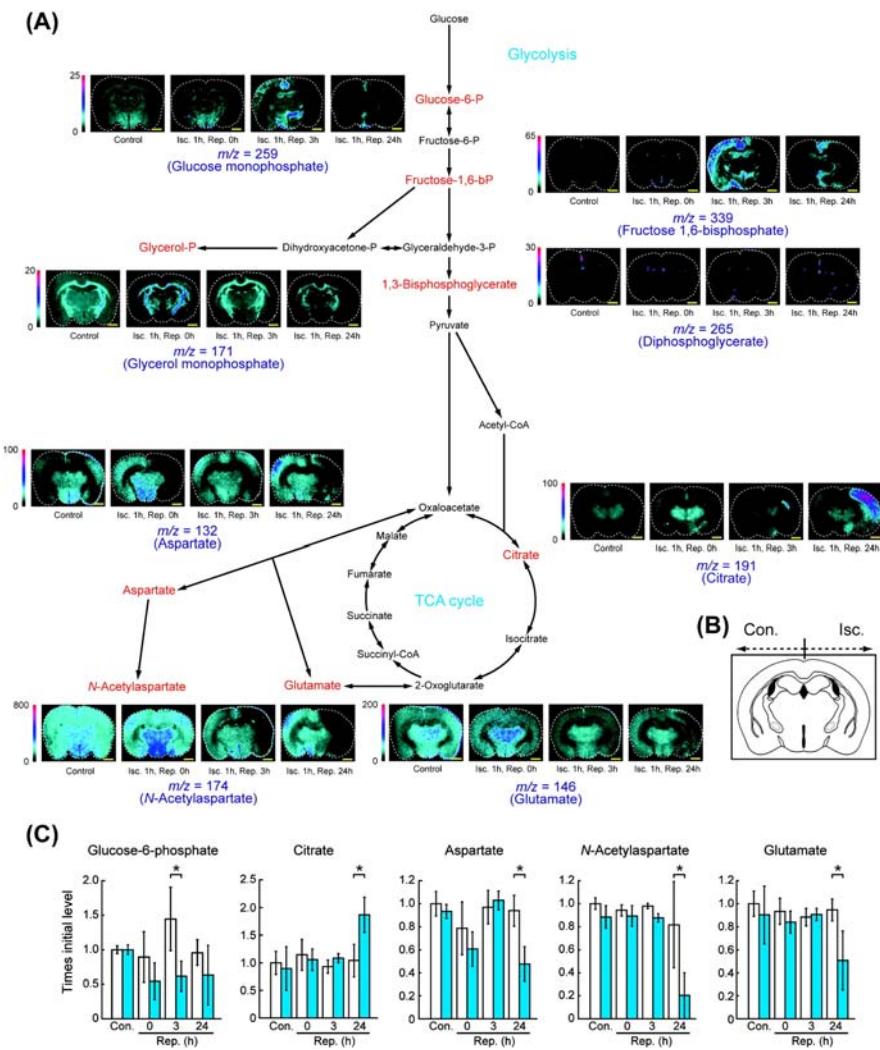
MALDI mass spectrometry imaging of endogenous metabolites

While MALDI–MSI was initially developed for proteomic analysis, pioneered by the works of Caprioli (Caprioli et al., 1997; Chaurand et al., 1999, 2006), it has been rapidly extended to the metabolomics and lipidomics field, in several biomedical applications. While the analysis of small molecular weight metabolites ($m/z < 500$ Da) can be challenging due to matrix interferences, several works

have demonstrated the effectiveness of the employment of 9-AA as matrix, resulting in very low background and very high sensitivity (Miura et al., 2012). This strategy has been successfully applied to image primary metabolites such as adenosine monophosphate, adenosine diphosphate, adenosine triphosphate, uridine diphosphate, or N-acetyl-D-glucosamine, in rat brain (Benabellah et al., 2009) as well as spatiotemporal metabolic changes following ischemia-reperfusion event in a rat model of middle cerebral occlusion, imaging simultaneously aminoacids, nucleotides, lipids, cofactors and carboxylic acids (Miura et al., 2010) (see Fig. 15.3).

Besides fresh frozen tissue, recently it has been demonstrated that also FFPE fixed tissues can be used for metabolomics analysis (Ly et al., 2016). The developed method consisted in deparaffinization of tissue slices and subsequent coating with 9-AA. The authors were able to detect a large number of features in m/z range 50–1000, noteworthy the comparison with fresh frozen tissue led to a 72% overlap, indicating a good stability in FFPE tissues. Small-molecule neurotransmitters such as dopamine (DA), γ -amino butyric acid (GABA) glutamate (Glu), and acetylcholine (ACh) are critical key messengers for central nervous system communication. Excellent results for metabolites containing primary amine groups were obtained with the employment of 2,3-diphenyl-pyranylium tetrafluoroborate (DPP-TFB) allowing the simultaneous image of 23 amino metabolites and their alteration in a model of experimental cortical spreading depression in mouse brain (Esteve et al., 2016). In a further study the combination of DPP-TFB coating and the employment of deuterated neurotransmitters analogs namely tyrosine, tryptamine, tyramine, phenethylamine, DA, 3-methoxytyramine, serotonin, GABA, glutamate, ACh, and L-alpha-glycerylphosphorylcholine spotted on tissues allowed the absolute quantitation of the selected molecules in brain tissue section at high spatial resolution (Shariatgorji et al., 2014). An improvement of neurotransmitter MALDI–MSI was recently demonstrated (Shariatgorji et al., 2019) with the employment of FMP-10 matrix that selectively targets phenolic and primary amine groups to simultaneously image the catecholaminergic and serotonergic signaling systems-including precursors and metabolites, that has been applied to the analysis of neurotransmitter alteration in rat and primate Parkinson model (Shariatgorji et al., 2019).

The lipidome is a branch of the metabolome, given the chemical differences between metabolites and lipids, analytical approaches typically used in lipidomics are different, particularly regarding sample treatment and analysis. Given their nature, almost all tissues used in MALDI–MSI have a considerable lipid content, that differs in abundance and composition among the different tissue sources, hence, given the strong signal around 700–800 m/z for other applications tissues are usually washed in organic solvent to improve s/n ratio of peptides (Lemaire et al., 2006). MALDI–MSI has been successfully applied to a large number of lipidomic studies (Zemski Berry et al., 2011), among lipids phosphatidylcholines and their different distribution in brain was investigated, resulting in a selective localization of PC(32:0), PC(34:1), PC(36:1), PC(32:0) and PC(34:1)

**FIGURE 15.3**

In situ metabolic pathway imaging visualizes changes of spatiotemporal metabolite distribution in middle cerebral occlusion (MCAO) rat brain. Wistar rat brains of control (no operation) or from rats after various periods of reperfusion following 1 h of MCAO were extirpated and immediately frozen under -80°C . Coronally sectioned brain slices ($10\ \mu\text{m}$ thickness) were then used for in situ metabolite imaging. Mass imaging data were acquired in negative ionization mode with $50\ \mu\text{m}$ spatial resolution ($14000\ \mu\text{m} \times 11000\ \mu\text{m}$, 10 shots/data point). All imaging data were normalized with the average mass spectrum for quantitative comparison of the concentration of each metabolite at different

(Continued)

(Mikawa et al., 2009). Negative ionization MALDI–MSI has been applied to image the modulation of cardiolipins in brain following a brain injury event in a Sprague Dawley rats model (Sparvero et al., 2016). Ischemic injury and betaamyloid (A β) toxicity tissue were used in MALDI–MSI to image the variation of glycosphingolipids, in particular gangliosides GD1a, GM1, GM2, and GM3 to determine alteration of their expression profiles following pathological events, similarly acylcarnitines have also been subject of MALDI–MSI analysis in models of traumatic brain injury (Caughlin et al., 2015; Mallah et al., 2019). Lipidomic applications of MALDI–MSI have found also relevant space in cardiovascular diseases (Mezger et al., 2019), as in well-known models such as ApoE knockout mouse models for the imaging of lipid alteration in atherosclerotic plaques revealing alteration of key markers of atherosclerosis progression and plaque remodeling such as Lysophosphatidylcholines and Sphingomyelins (Cao et al., 2020; Jose et al., 2014). Given the extreme complexity and the enormous amount of isomers occurring between different lipid classes, novel analytical strategies are currently used such as the employment of IMS, that thanks to the gas phase separation add a novel dimension of separation, enhancing the potential of current TOF analyzers. Several vendors now offer IMS with MALDI-TOF instruments, such as the traveling wave ion mobility from Waters SYNAPT HDMS Q-TOF systems, that have been widely applied to different class of lipids resulting in improving structural information thanks to the added selectivity of IMS (Harvey et al., 2012; Ridenour et al., 2010). Very recently a novel trapped ion mobility (TIMS)-TOF mass spectrometer has been commercialized from Bruker, and successfully applied to the separation of closely isobaric lipids such as [PC(32:0) + Na] $^+$ and [PC(34:3) + H] $^+$ (3 mDa mass difference) at very high spatial resolution and maintaining very high acquisition rates ($> 2 \text{ pixels s}^{-1}$) (Djambazova et al., 2020; Spraggins et al., 2019). Lastly novel ionization strategies such as Matrix-assisted laser desorption/ionization combined with laser-induced postionization (MALDI-2) have been combined with MALDI-TIMS-TOF currently under evaluation for numerous classes of biomolecules, resulting in a significant boost in sensitivity, with on average, a number of peaks 40% higher compared to conventional MALDI analysis in a rat brain sagittal section as well as a high number of detected features (Soltwisch et al., 2020) (Fig. 15.4).

times. (A) These data were put on the central metabolic pathway map. (B) A schematic illustration represents the structure of coronally sectioned brain. (C) Relative changes in concentrations of metabolites extracted from whole CTX. Data is shown as the mean \pm SD ($n = 5$) and represents the relative concentration of each condition to the concentration of contralateral CTX in control (Con.). Asterisk mark indicates significant differences ($P < .05$) between contralateral (open bar) and ischemic (closed bar) CTXs.

From Miura, D., Fujimura, Y., Yamato, M., Hyodo, F., Utsumi, H., Tachibana, H., & Wariishi, H. (2010).

Ultrahighly sensitive in situ metabolomic imaging for visualizing spatiotemporal metabolic behaviors.

Analytical Chemistry, 82(23), 9789–9796. <https://doi.org/10.1021/ac101998z>.

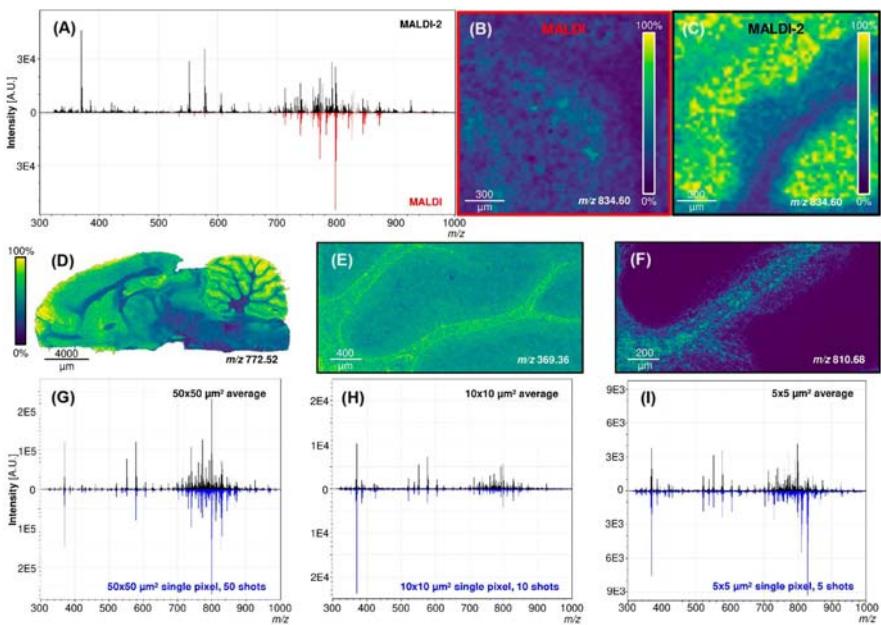


FIGURE 15.4

Overview of TIMSTOF fleX MALDI-2-MSI data of lipids registered from rat brain sections. (A) Average mass spectra of MALDI-2-MSI (top, black trace) and MALDI-MSI (bottom, red trace) analyses of rat cerebellum. The MALDI data set consists of 40249 pixels and was recorded at 22 pixels/s using a 10 kHz laser repetition rate and 20 shots/pixel. The MALDI-2 data set consists of 46818 pixels and was recorded at 16 pixels/s using a 1 kHz laser repetition rate and 20 shots/pixel. Ion distribution of m/z 834.60 ($PC(40:6) [M + H]^+$) in (B) the MALDI and (C) MALDI-2 analyses, respectively. (D–F) Additional lipid distributions revealed by the MALDI-2-MSI analyses: (D) m/z 772.52 ($PC(32:0) [M + K]^+$) at $50 \times 50 \mu\text{m}^2$, recorded using the “M5 small” laser setting, 1 kHz laser repetition rate, and 50 shots/pixel (15 pixels/s); (E) m/z 369.36 (cholesterol-H₂O [$M + H]^+$) at $10 \times 10 \mu\text{m}^2$, recorded using the “single” laser setting with “beam scan,” 1 kHz laser repetition rate, and 10 shots/pixel (33 pixels/s); (F) m/z 810.68 (HexCer (d18:1/C24:1) [$M + H]^+$) at $5 \times 5 \mu\text{m}^2$, recorded using the “single” laser setting without “beam scan,” 1 kHz laser repetition rate, and 5 shots/pixel (33 pixels/s); we note that based on the data, we cannot decide on the presence of isomeric lipoforms of this glycosphingolipid (e.g., comprising a galactose vs glucose moiety) and that of varying combinations in the ceramide part (e.g., (d20:1/C22:1). In (G)–(I), the respective average (black) and single white-matter pixel (blue) mass spectra for the datasets recorded in (D)–(F).

From Soltwisch, J., Heijls, B., Koch, A., Vens-Cappell, S., Höhndorf, J., & Dreisewerd, K. (2020). MALDI-2 on a trapped ion mobility quadrupole time-of-flight instrument for rapid mass spectrometry imaging and ion mobility separation of complex lipid profiles. *Analytical Chemistry*, 92(13), 8697–8703. <https://doi.org/10.1021/acs.analchem.0c01747>.

Metabolite annotation and quantitation in MALDI mass spectrometry imaging

One of the challenges of MALDI–MSI and other imaging techniques is the annotation of metabolites in tissues, that clearly is different from conventional metabolomics workflow based on liquid chromatography tandem MS (MS/MS) where data dependent or independent acquisition are currently employed with success for metabolite identification together with real or in silico spectral libraries. While MS/MS capability are available on almost all MALDI–MS instruments, it is important to underline that MS/MS is limited to the duty cycle of the instrument, even though latest qTOF and TOF-TOF instruments possess high acquisition rates. While proteomics identification strategies are more mature and established, automated metabolite identification remains challenging. Different vendors have in house solution, but successful other softwares have been developed, both open source or commercial, such as Metaspace (Palmer et al., 2016) or Lipostar (Tortorella et al., 2020), these software can handle data from different vendors and allow automated metabolite and lipids identification. For quantitative analysis MALDI–MSI has often been labeled as difficult because of the matrix fluctuation, and/or background in the low m/z region intense, nevertheless several group demonstrated the potential of quantitative MALDI–MSI for different metabolite classes. Different quantitative approaches available for MALDI–MSI relative or absolute quantitation, spanning from normalization strategies, to employment of analogs or isotope labeled internal standard spotted directly on the tissue. Given the vastity of the argument, excellent papers can be found in literature, but the argument will not be discussed in detail in the interest of brevity (Rzagalinski & Volmer, 2017).

Conclusion and future perspectives

MALDI–MSI and other imaging techniques for their unique ability are at the forefront of analytical methods in pharmaceutical and biomedical fields. The ability to detect at spatial level a wide range of biomolecules plays a fundamental role in several fields such as drug discovery, cancer research, and personalized medicine. While some limitations are still present such as dynamic range, lack of harmonized methods for sample treatment, matrix deposition, analyte automated identification and quantitation, the latest development in MS analyzer technology, robotic matrix deposition system, and spotters for calibration curve building are leading this powerful method further and in the next years several exciting advances will allow to reach higher sensitivity, resolving power, and repeatability.

References

- Amstalden van Hove, E. R., Smith, D. F., & Heeren, R. M. A. (2010). A concise review of mass spectrometry imaging. *Journal of Chromatography A*, 1217(25), 3946–3954. Available from <https://doi.org/10.1016/j.chroma.2010.01.033>.
- Benabdellah, F., Touboul, D., Brunelle, A., & Laprévote, O. (2009). In situ primary metabolites localization on a rat brain section by chemical mass spectrometry imaging. *Analytical Chemistry*, 81(13), 5557–5560. Available from <https://doi.org/10.1021/ac9005364>.
- Cao, J., Goossens, P., Martin-Lorenzo, M., Dewez, F., Claes, B. S. R., Biessen, E. A. L., Heeren, R. M. A., & Balluff, B. (2020). Atheroma-specific lipids in ldlr-/ and apoe-/ mice using 2D and 3D matrix-assisted laser desorption/ionization mass spectrometry imaging. *Journal of the American Society for Mass Spectrometry*, 31(9), 1825–1832. Available from <https://doi.org/10.1021/jasms.0c00070>.
- Caprioli, R. M., Farmer, T. B., & Gile, J. (1997). Molecular imaging of biological samples: Localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry*, 69 (23), 4751–4760. Available from <https://doi.org/10.1021/ac970888i>.
- Caughlin, S., Hepburn, J. D., Park, D. H., Jurcic, K., Yeung, K. K. C., Cechetto, D. F., & Whitehead, S. N. (2015). Increased expression of simple ganglioside species GM2 and GM3 detected by MALDI Imaging Mass Spectrometry in a combined rat model of A β toxicity and stroke. *PLoS ONE*, 10(6). Available from <https://doi.org/10.1371/journal.pone.0130364>.
- Chaurand, P., Norris, J. L., Cornett, D. S., Mobley, J. A., & Caprioli, R. M. (2006). New developments in profiling and imaging of proteins from tissue sections by MALDI mass spectrometry. *Journal of Proteome Research*, 5(11), 2889–2900. Available from <https://doi.org/10.1021/pr060346u>.
- Chaurand, P., Stoeckli, M., & Caprioli, R. M. (1999). Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Analytical Chemistry*, 71(23), 5263–5270. Available from <https://doi.org/10.1021/ac990781q>.
- Chen, R., Hui, L., Sturm, R. M., & Li, L. (2009). Three dimensional mapping of neuropeptides and lipids in crustacean brain by mass spectral imaging. *Journal of the American Society for Mass Spectrometry*, 20. Available from <https://doi.org/10.1021/jasms.8b03482>.
- Chughtai, K., & Heeren, R. M. A. (2010). Mass spectrometric imaging for biomedical tissue analysis. *Chemical Reviews*, 110(5), 3237–3277. Available from <https://doi.org/10.1021/cr100012c>.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78. Available from <https://doi.org/10.1002/mas.20108>.
- Djambazova, K. V., Klein, D. R., Migas, L. G., Neumann, E. K., Rivera, E. S., Van De Plas, R., Caprioli, R. M., & Spraggins, J. M. (2020). Resolving the complexity of spatial lipidomics using MALDI TIMS imaging mass spectrometry. *Analytical Chemistry*, 92(19), 13290–13297. Available from <https://doi.org/10.1021/acs.analchem.0c02520>.
- Esteve, C., Tolner, E. A., Shyti, R., van den Maagdenberg, A. M. J. M., & McDonnell, L. A. (2016). Mass spectrometry imaging of amino neurotransmitters: A comparison of derivatization methods and application in mouse brain tissue. *Metabolomics*, 12(2), 1–9. Available from <https://doi.org/10.1007/s11306-015-0926-0>.

- Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171. Available from <https://doi.org/10.1023/A:1013713905833>.
- Goodwin, R. J. A. (2012). Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences. *Journal of Proteomics*, 75(16), 4893–4911. Available from <https://doi.org/10.1016/j.jprot.2012.04.012>.
- Goodwin, R. J. A., Lang, A. M., Allingham, H., Borén, M., & Pitt, A. R. (2010). Stopping the clock on proteomic degradation by heat treatment at the point of tissue excision. *Proteomics*, 10(9), 1751–1761. Available from <https://doi.org/10.1002/pmic.200900641>.
- Hankin, J. A., Barkley, R. M., & Murphy, R. C. (2007). Sublimation as a method of matrix application for mass spectrometric imaging. *Journal of the American Society for Mass Spectrometry*, 18(9), 1646–1652. Available from <https://doi.org/10.1016/j.jasms.2007.06.010>.
- Harvey, D. J., Scarff, C. A., Crispin, M., Scanlan, C. N., Bonomelli, C., & Scrivens, J. H. (2012). MALDI-MS/MS with traveling wave ion mobility for the structural analysis of N-Linked Glycans. *Journal of the American Society for Mass Spectrometry*, 23(11), 1955–1966. Available from <https://doi.org/10.1007/s13361-012-0425-8>.
- Jose, C.-P., Nathan, H., Nana, K. K., Sheng-Ping, W., Vivienne, M., Henry, S., Alan, M., John, S., Mark, T., David, M., Vinit, S., Stephen, P., Karen, A., Michele, C., P., R. T., & G., J. D. (2014). In vivo isotopically labeled atherosclerotic aorta plaques in ApoE KO mice and molecular profiling by matrix-assisted laser desorption/ionization mass spectrometric imaging. *Rapid Communications in Mass Spectrometry*, 28, 2471–2479. Available from <https://doi.org/10.1002/rcm.7039>.
- Kaletaş, B. K., Van Der Wiel, I. M., Stauber, J., Dekker, L. J., Güzel, C., Kros, J. M., Luider, T. M., & Heeren, R. M. A. (2009). Sample preparation issues for tissue imaging by imaging MS. *Proteomics*, 9(10), 2622–2633. Available from <https://doi.org/10.1002/pmic.200800364>.
- Korte, A. R., & Lee, Y. J. (2014). MALDI-MS analysis and imaging of small molecule metabolites with 1,5-diaminonaphthalene (DAN). *Journal of Mass Spectrometry*, 49(8), 737–741. Available from <https://doi.org/10.1002/jms.3400>.
- Landgraf, R. R., Prieto Conaway, M. C., Garrett, T. J., Stacpoole, P. W., & Yost, R. A. (2009). Imaging of lipids in spinal cord using intermediate pressure matrix-assisted laser desorption-linear ion trap/orbitrap MS. *Analytical Chemistry*, 81(20), 8488–8495. Available from <https://doi.org/10.1021/ac901387u>.
- Lemaire, R., Wisztorski, M., Desmons, A., Tabet, J. C., Day, R., Salzet, M., & Fournier, I. (2006). MALDI-MS direct tissue analysis of proteins: Improving signal sensitivity using organic treatments. *Analytical Chemistry*, 78(20), 7145–7153. Available from <https://doi.org/10.1021/ac060565z>.
- Ly, A., Buck, A., Balluff, B., Sun, N., Gorzolka, K., Feuchtinger, A., Janssen, K. P., Kuppen, P. J. K., Van De Velde, C. J. H., Weirich, G., Erlmeier, F., Langer, R., Aubele, M., Zitzelsberger, H., McDonnell, L., Aichler, M., & Walch, A. (2016). High-mass-resolution MALDI mass spectrometry imaging of metabolites from formalin-fixed paraffin-embedded tissue. *Nature Protocols*, 11(8), 1428–1443. Available from <https://doi.org/10.1038/nprot.2016.081>.
- Mallah, K., Quanico, J., Raffo-Romero, A., Cardon, T., Aboulouard, S., Devos, D., Kobeissy, F., Zibara, K., Salzet, M., & Fournier, I. (2019). Matrix-assisted laser desorption/ionization-mass spectrometry imaging of lipids in experimental model of traumatic brain injury

- detecting acylcarnitines as injury related markers. *Analytical Chemistry*, 91(18), 11879–11887. Available from <https://doi.org/10.1021/acs.analchem.9b02633>.
- McLean, J. A., Ridenour, W. B., & Caprioli, R. M. (2007). Profiling and imaging of tissues by imaging ion mobility-mass spectrometry. *Journal of Mass Spectrometry*, 42(8), 1099–1105. Available from <https://doi.org/10.1002/jms.1254>.
- Meriaux, C., Franck, J., Wisztorski, M., Salzet, M., & Fournier, I. (2010). Liquid ionic matrixes for MALDI mass spectrometry imaging of lipids. *Journal of Proteomics*, 73 (6), 1204–1218. Available from <https://doi.org/10.1016/j.jprot.2010.02.010>.
- Mezger, S. T. P., Mingels, A. M. A., Bekers, O., Cillero-Pastor, B., & Heeren, R. M. A. (2019). Trends in mass spectrometry imaging for cardiovascular diseases. *Analytical and Bioanalytical Chemistry*, 411(17), 3709–3720. Available from <https://doi.org/10.1007/s00216-019-01780-8>.
- Mikawa, S., Suzuki, M., Fujimoto, C., & Sato, K. (2009). Imaging of phosphatidylcholines in the adult rat brain using MALDI-TOF MS. *Neuroscience Letters*, 451(1), 45–49. Available from <https://doi.org/10.1016/j.neulet.2008.12.035>.
- Miura, D., Fujimura, Y., & Wariishi, H. (2012). In situ metabolomic mass spectrometry imaging: Recent advances and difficulties. *Journal of Proteomics*, 75(16), 5052–5060. Available from <https://doi.org/10.1016/j.jprot.2012.02.011>.
- Miura, D., Fujimura, Y., Yamato, M., Hyodo, F., Utsumi, H., Tachibana, H., & Wariishi, H. (2010). Ultrahighly sensitive in situ metabolomic imaging for visualizing spatiotemporal metabolic behaviors. *Analytical Chemistry*, 82(23), 9789–9796. Available from <https://doi.org/10.1021/ac101998z>.
- Nicholson, J. K., & Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature*, 455 (7216), 1054–1056. Available from <https://doi.org/10.1038/4551054a>.
- Nishidate, M., Hayashi, M., Aikawa, H., Tanaka, K., Nakada, N., Miura, S. i, Ryu, S., Higashi, T., Ikarashi, Y., Fujiwara, Y., & Hamada, A. (2019). Applications of MALDI mass spectrometry imaging for pharmacokinetic studies during drug development. *Drug Metabolism and Pharmacokinetics*, 34(4), 209–216. Available from <https://doi.org/10.1016/j.dmpk.2019.04.006>.
- Norris, J. L., & Caprioli, R. M. (2013). Analysis of tissue specimens by matrix-assisted laser desorption/ionization mass spectrometry in biological and clinical research. *Chemical Reviews*, 113(4), 2309–2342. Available from <https://doi.org/10.1021/cr3004295>.
- Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., Kovalev, V., Fuchser, J., Nikolenko, S., Pineau, C., Becker, M., & Alexandrov, T. (2016). FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 14(1), 57–60. Available from <https://doi.org/10.1038/nmeth.4072>.
- Ridenour, W. B., Kliman, M., McLean, J. A., & Caprioli, R. M. (2010). Structural characterization of phospholipids and peptides directly from tissue sections by MALDI traveling-wave ion mobility-mass spectrometry. *Analytical Chemistry*, 82(5), 1881–1889. Available from <https://doi.org/10.1021/ac9026115>.
- Rzagalinski, I., & Volmer, D. A. (2017). Quantification of low molecular weight compounds by MALDI imaging mass spectrometry—A tutorial review. *Biochimica et Biophysica Acta—Proteins and Proteomics*, 1865(7), 726–739. Available from <https://doi.org/10.1016/j.bbapap.2016.12.011>.
- Schwamborn, K., & Caprioli, R. M. (2010). Molecular imaging by mass spectrometry—looking beyond classical histology. *Nature Reviews Cancer*, 10(9), 639–646. Available from <https://doi.org/10.1038/nrc2917>.

- Shariatgorji, M., Nilsson, A., Fridjonsdottir, E., Vallianatou, T., Källback, P., Katan, L., Sävmarker, J., Mantas, I., Zhang, X., Bezard, E., Svensson, P., Odell, L. R., & Andrén, P. E. (2019). Comprehensive mapping of neurotransmitter networks by MALDI–MS imaging. *Nature Methods*, 16(10), 1021–1028. Available from <https://doi.org/10.1038/s41592-019-0551-3>.
- Shariatgorji, M., Nilsson, A., Goodwin, R. J. A., Källback, P., Schintu, N., Zhang, X., Crossman, A. R., Bezard, E., Svensson, P., & Andren, P. E. (2014). Direct targeted quantitative molecular imaging of neurotransmitters in brain tissue sections. *Neuron*, 84(4), 697–707. Available from <https://doi.org/10.1016/j.neuron.2014.10.011>.
- Shariatgorji, M., Strittmatter, N., Nilsson, A., Källback, P., Alvarsson, A., Zhang, X., Vallianatou, T., Svensson, P., Goodwin, R. J. A., & Andren, P. E. (2016). Simultaneous imaging of multiple neurotransmitters and neuroactive substances in the brain by desorption electrospray ionization mass spectrometry. *NeuroImage*, 136, 129–138. Available from <https://doi.org/10.1016/j.neuroimage.2016.05.004>.
- Shroff, R., & Svatoš, A. (2009). 1,8-Bis(dimethylamino)naphthalene: A novel superbasic matrix for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis of fatty acids. *Rapid Communications in Mass Spectrometry*, 23(15), 2380–2382. Available from <https://doi.org/10.1002/rcm.4143>.
- Soltwisch, J., Heijs, B., Koch, A., Vens-Cappell, S., Höhndorf, J., & Dreisewerd, K. (2020). MALDI-2 on a trapped ion mobility quadrupole time-of-flight instrument for rapid mass spectrometry imaging and ion mobility separation of complex lipid profiles. *Analytical Chemistry*, 92(13), 8697–8703. Available from <https://doi.org/10.1021/acs.analchem.0c01747>.
- Sparvero, L. J., Amoscato, A. A., Fink, A. B., Anthonymuthu, T., New, L. A., Kochanek, P. M., Watkins, S., Kagan, V. E., & Bayir, H. (2016). Imaging mass spectrometry reveals loss of polyunsaturated cardiolipins in the cortical contusion, hippocampus, and thalamus after traumatic brain injury. *Journal of Neurochemistry*, 139(4), 659–675. Available from <https://doi.org/10.1111/jnc.13840>.
- Spraggins, J. M., Djambazova, K. V., Rivera, E. S., Migas, L. G., Neumann, E. K., Fuetterer, A., Suetering, J., Goedecke, N., Ly, A., Van De Plas, R., & Caprioli, R. M. (2019). High-performance molecular imaging with MALDI trapped ion-mobility time-of-flight (timsTOF) mass spectrometry. *Analytical Chemistry*, 91(22), 14552–14560. Available from <https://doi.org/10.1021/acs.analchem.9b03612>.
- Svensson, M., Sköld, K., Nilsson, A., Fälth, M., Nydahl, K., Svensson, P., & Andrén, P. E. (2007). Neuropeptidomics: MS applied to the discovery of novel peptides from the brain. *Analytical Chemistry*, 79(1), 14–21. Available from <https://doi.org/10.1021/ac071856q>.
- Tortorella, S., Tiberi, P., Bowman, A. P., Claes, B. S. R., Ščupáková, K., Heeren, R. M. A., Ellis, S. R., & Cruciani, G. (2020). LipoStarMSI: Comprehensive, vendor-neutral software for visualization, data analysis, and automated molecular identification in mass spectrometry imaging. *Journal of the American Society for Mass Spectrometry*, 31(1), 155–163. Available from <https://doi.org/10.1021/jasms.9b00034>.
- Weiβfloga, J., & Svatoš, A. (2016). 1,8-Di(piperidinyl)-naphthalene—Rationally designed MAILD/MALDI matrix for metabolomics and imaging mass spectrometry. *RSC Advances*, 75073–75081. Available from <https://doi.org/10.1039/C6RA17237G>.
- Zemski Berry, K. A., Hankin, J. A., Barkley, R. M., Spraggins, J. M., Caprioli, R. M., & Murphy, R. C. (2011). MALDI imaging of lipid biochemistry in tissues by mass spectrometry. *Chemical Reviews*, 111(10), 6491–6512. Available from <https://doi.org/10.1021/cr200280p>.

This page intentionally left blank

Metabolomics for oncology 16

Susan Costantini and Alfredo Budillon

*Experimental Pharmacology Unit—Istituto Nazionale Tumori-IRCCS Fondazione G. Pascale,
Naples, Italy*

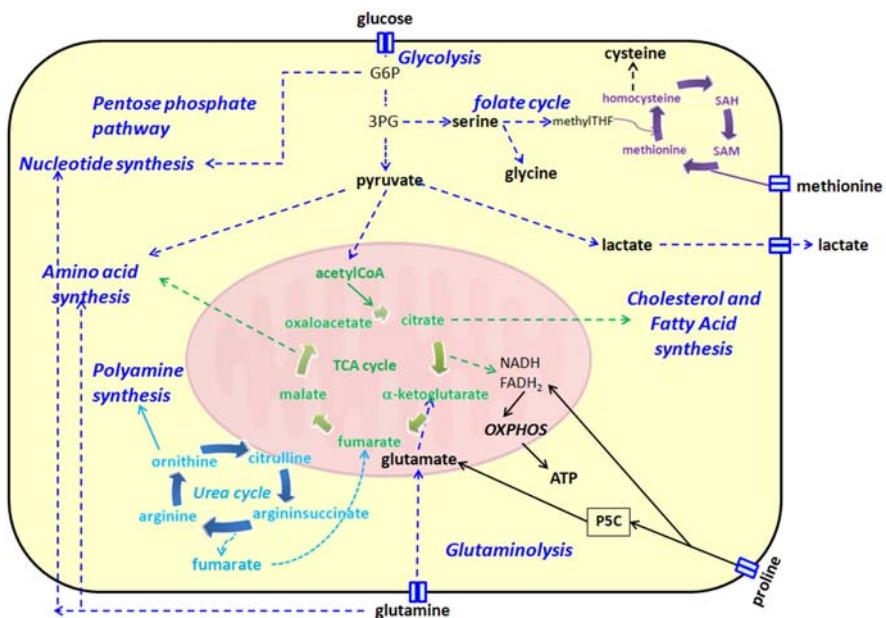
Introduction

Normal cells follow organized metabolic programs and replicate in an ordered way through the mitosis. Cancer cells proliferate in an uncontrolled way and invade the neighboring cells, present genetic abnormalities, are resistant to cell death, migrate to distant organs forming tumor metastases, and consume high levels of cellular nutrients (DeBerardinis et al., 2008; Hanahan & Weinberg, 2000; Kalyanaraman, 2017).

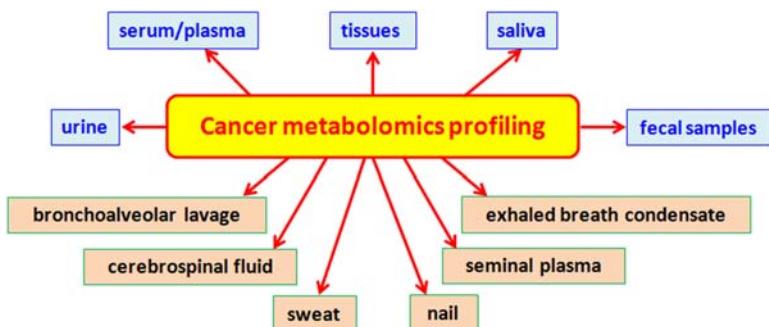
Since, during each division, cancer cells must assemble different cellular components such as DNA, proteins, phospholipids, and other organelle, they present an increased energy demand, and acquire it using mechanisms different from those of normal cells because of the metabolic reprogramming (Ward & Thompson, 2012). A common feature of cancer cell metabolism is the capability to recover nutrients from a frequently nutrient-poor environment and to use them for cellular proliferation. The alterations in intracellular and extracellular metabolites that can accompany cancer-associated metabolic reprogramming have profound effects on gene expression, cellular differentiation, and tumor microenvironment (Pavlova & Thompson, 2016). Typically, cancer cells reprogram their metabolism by consuming high levels of glucose and of some nonessential amino acids (like glutamine, serine and others), producing reactive oxygen species and potentiating the synthesis of nucleotides, fatty acids, and lipids.

In the following paragraphs, we describe the mechanisms involved in the reprogramming of cancer cell metabolism (Fig. 16.1).

Then, we reviewed current applications and examples of human cancer metabolomics based on modern techniques such as ^1H -nuclear magnetic resonance ($^1\text{H-NMR}$) spectroscopy; gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), fourier-transform infrared spectroscopy (FT-IR), vibrational and Raman spectroscopy (Fig. 16.2).

**FIGURE 16.1**

Cancer cell metabolic reprogramming. Schematic representation of the mechanisms involved in the reprogramming of cancer cell metabolism.

**FIGURE 16.2**

Cancer metabolomics profiling applications. Summary of possible cancer metabolomics profiling applications.

Reprogramming of cancer cell metabolism

Glucose and Warburg effect

Normally, most of the glucose consumed by cells is catabolized through glycolysis. This process produces two molecules of pyruvate, nicotinamide adenine dinucleotide hydride (NADH) and adenosine triphosphate (ATP) per each molecule of consumed glucose. The pyruvate is then transported to the mitochondria, and enters the tricarboxylic acid (TCA) cycle, also named as citric acid or Krebs cycle, to generate NADH, being the principal electron donor of the oxidative phosphorylation (OXPHOS) pathway and generating 30 additional molecules of ATP (Fig. 16.1). This last process provides 18-fold more ATP than glycolysis (Romero-Garcia et al., 2011). In details, through glycolysis pathway, the glucose provides metabolites such as pyruvate, 3-phosphoglycerate (3PG) and glucose 6-phosphate (G6P), that represent the principal carbon sources for biosynthesis of many macromolecules necessary for cellular proliferation. In particular, the pyruvate participates to the synthesis of:

1. acetyl-CoA, a precursor of cholesterol and fatty acid synthesis;
2. aspartate and asparagine, two nonessential amino acids.

On the other hand, 3PG is involved in the synthesis of the nonessential serine amino acid, whereas G6P is engaged in the pentose phosphate pathway (PPP) generating the ribose group for the nucleotides synthesis and participating to NADPH production (Vazquez et al., 2016). In the case of cancer cells, the majority of the pyruvate is converted to lactate (Fig. 16.1). In detail, in comparison to normal cells which rely primarily on mitochondrial oxidative phosphorylation to generate the energy needed for cellular processes, cancer cells used this catabolic process of glucose transformation into lactate (known also as glucose fermentation) that has a lower energy production (ATP) and, hence, requires a higher glucose consumption to satisfy the energy demand. This phenomenon was discovered in the 1920s by Otto Warburg that demonstrated how malignant ascites consumed a higher glucose amount and produced lactate. Therefore, this process was indicated as aerobic glycolysis or Warburg effect as a distinctive characteristic of aggressive cancer (Warburg, 1956). This high glucose demand of cancer cells is currently widely used in cancer diagnosis and to evaluate response to anticancer treatments by ^{18}F -Fluorine-deoxyglucose positron emission tomography (^{18}F -FDG-PET), that uses as tracer ^{18}F -FDG, a radioactive analog of glucose not metabolized in the glycolytic pathway and transported by glucose transporters (GLUTs) into cancer cells. High ^{18}F -FDG uptake by tumors correlates positively with both GLUT expression and poor cancer prognosis (Gatenby & Gillies, 2004; Skoura et al., 2012).

However, the reasons for increased glycolysis in cancer cells are still not clear. For examples it is controversial if Warburg effect can be only a result of defects in mitochondrial function and of impaired oxidative metabolism since many

highly proliferative cancer cell lines do not present any defect in their oxidative metabolism (Moreno-Sánchez et al., 2007). Moreover, although lactate dehydrogenase A (LDH-A), which catalyzes the conversion of pyruvate to lactate, is over-expressed in some cancers (Feng et al., 2018), its knock-out did not induce a decreased capacity to produce ATP by OXPHOS pathway (Fantin et al., 2006). Anyhow, since both glycolysis and OXPHOS pathways produce energy, the inhibition of one or both processes using selectively targeted drugs could be useful as anticancer mechanism (Scatena, 2012).

Lactate shuttle due to tumor hypoxia and Warburg effect

Cancer tissues present an oxygen gradient, containing hypoxic and aerobic regions, and, in literature, it has been demonstrated that there is a “metabolic symbiosis” between these regions (Semenza, 2011). In detail, cancer cells are well or poorly oxygenated when they are near or distant of blood vessels, respectively. Hence, when cancer cells present low oxygen levels and are hypoxic, hypoxia-inducible factor-1 α (HIF-1 α) is stabilized and, besides the vascular endothelial growth factor, the driver of cancer-induced angiogenesis, also increases the transcriptional activation of GLUTs and LDH-A, thus promoting the glucose uptake and the lactate secretion by monocarboxylated transporter 4. This circuit can be activated by lactate itself as signaling molecule for triggering HIF-1 α stabilization (Goodwin et al., 2015). On the other hand, when cancer cells are close to blood vessels and the oxygen levels are high, the lactate is uptaken by monocarboxylated transporter 1 (MCT1) and converted to pyruvate by LDH-B being thus replacing glucose energy source (Morrot et al., 2018). This process demonstrates the role of the lactate as a shuttle metabolite, used by different cancer cell subpopulations. Thus, when MCT1 is inhibited, cancer cells cannot absorb the lactate and deprive hypoxic cancer cells of correct glucose utilization. Hence, the development and use of specific MCT1 inhibitors could be useful as putative targeted chemotherapeutic agents.

Glutamine metabolism

Glutamine is the source of nitrogen for the biosynthesis of glycosylated molecules, purine and pyrimidine nucleotides, glucosamine-6-phosphate, and non-essential amino acids (alanine, aspartate, serine and proline), and serves as a substrate for fatty acid synthesis in hypoxic cells or cells with HIF-1 α activation (Metere et al., 2020). Glutamine is converted to glutamate by glutaminase (GLS), during glutaminolysis reaction; in turn the glutamate is deaminated to α -ketoglutarate by the enzyme glutamate dehydrogenase (GDH), thus fueling the TCA cycle (Fig. 16.1).

In cancer cells glutamine represents the second most consumed nutrient after glucose (Jain et al., 2012). Moreover, both GLS1, one of the two isoforms of GLS (the kidney-type GLS1 and the liver-type GLS2), and GDH resulted to be overexpressed in many cancers (Moreno-Sánchez et al., 2020; Saha et al., 2019). Notably, glutamine modulates glutaminolysis in combination with leucine that is capable to activate allosterically GDH, thus inducing α -ketoglutarate production preventing GLS inhibition by glutamate accumulation (Nguyen & Durán, 2018).

On the other hand, glutamine synthetase is the only enzyme able to synthesize glutamine through ATP-dependent condensation of glutamate and ammonia. Glutamine synthetase overexpression has been found in hepatocellular carcinoma (HCC) (Dal Bello et al., 2010), indeed its inhibition as well as glutamine depletion block the growth of liver cancer xenografts, suggesting that high glutamine is critical for liver cancer metabolism (Chiu et al., 2014).

Furthermore, it is well known that α -ketoglutarate derived from glutamine is involved in fatty acids synthesis by carboxylation reaction mediated by isocitrate dehydrogenase that catalyzes a reversible reaction between isocitrate and α -ketoglutarate. In cancer cells the inverse reductive carboxylation reaction can be able to maintain TCA cycle intermediates under mitochondria defects. Hence, several observations confirmed a role of glutamine in lipid biosynthesis in cancer models with dysfunctional mitochondria or under hypoxia (Metallo et al., 2012; Mullen et al., 2012).

Finally, considering that cancer initiation and progression has been associated with oxidative stress, and that consequently cancer cells need to increase their antioxidant capacity, it is important to underline that glutamine-derived glutamate, through cysteine and glycine condensation, is utilized for the synthesis of glutathione, that plays an important role in the cellular antioxidative mechanisms. Indeed, glutamine starvation decreases the glutathione levels in cancer cells thus potentiating oxidative-stress-inducing antitumor approach such as radiotherapy (Xiang et al., 2013). Overall, given the dependence of cancer cells on glutamine metabolism, several targeted therapy approaches have been successfully employed to target the pathways described above in cancer models (Xiang et al., 2013).

Serine metabolism

In cancer cells serine is the third most consumed metabolite after glucose and glutamine (Jain et al., 2012). Serine is a nonessential amino acid synthesized from the 3PG, a glycolytic intermediate, through a reaction catalyzed by 3PG dehydrogenase (PHGDH), an enzyme that is often genetically amplified in breast cancer (Possemato et al., 2011), melanoma (Locasale et al., 2012) and colon cancer (Jia et al., 2016), being consequently considered as a target for cancer therapy.

Serine is converted into glycine (Fig. 16.1) by releasing an one-carbon unit to the one-carbon pool through the activity of cytosolic or mitochondrial serinehydroxymethyltransferases (SHMT1 and SHMT2) (Tedeschi et al., 2015), representing the more important donor of one-carbon units used to synthesize purines and

thymidine monophosphate in cancer cells ([Labuschagne et al., 2014](#)). The one-carbon units required for purine synthesis are produced into mitochondria through SHMT2 and 5,10-methenyl-THF dehydrogenase 2 (MTHFD2) with the release of formate after the transfer to tetrahydrofolate (THF) present in folate cycle ([Lewis et al., 2014](#)). MTHFD2 is overexpressed in cancer tissues, and indicated as potential cancer therapy target ([Wei et al., 2019; Yu et al., 2020; Zhu & Leung, 2020](#)). Interestingly, considering that both serine and glycine are involved in anabolic processes linked to cancer cell growth and proliferation, some reports suggested that a restriction of dietary serine and glycine is able to block the cancer growth ([Maddock et al., 2017](#)).

Methionine metabolism

In addition to glucose, glutamine and serine, several other amino acids are involved in cancer metabolic reprogramming. In this regard, methionine is imported and recycled in the cells through the “methionine cycle” where it is converted to S-adenosyl-methionine (SAM) that serves as a methyl donor in S-adenosylhomocysteine (SAH) formation. Then, SAH is hydrolyzed to adenosine and homocysteine which is, then, converted to cysteine via the trans-sulfuration pathway. Alternatively, homocysteine acquires an one-carbon unit from 5-methyl-THF (involved in folate cycle fueled by serine and glycine), through methionine synthase, regenerating methionine ([Fig. 16.1](#)). Methionine can also be recycled from the SAM-dependent polyamine biosynthesis via the methionine salvage pathway ([Sanderson et al., 2019](#)).

Methionine has been linked to cancer development and progression, being involved in one-carbon metabolism together with serine, glycine and folate ([Newman & Maddocks, 2017](#)), and connected to different metabolic processes including redox biology, chromatin and nucleic acid methylation, and polyamine synthesis. In this context, recently it has been demonstrated that cancer-initiating cells showed an upregulation of both methionine adenosyltransferase 2A (MAT2A) expression and activity ([Wang et al., 2019](#)). Moreover, methyltransferase nicotinamide N-methyltransferase has been indicated as a driver for the oncogenic behavior of cancer-associated fibroblasts ([Eckert et al., 2019](#)).

Some studies are focusing on the search of molecules able to block methionine metabolism. For example, PF-9366, an allosteric inhibitor of MAT2A, activates MAT2A when methionine or SAM levels are low, whereas inhibits MAT2A when the levels of these metabolites are high ([Quinlan et al., 2017](#)). Acetyl-11-keto- β -boswellic acid was recently identified as a natural MAT2A inhibitor ([Bai et al., 2019](#)).

However, as reported above, although the dietary restriction of both serine and glycine is able to attenuate cancer growth in xenograft models ([Maddock et al., 2017](#)), and these two amino acids are strictly linked to the folate cycle, and, hence, involved in regulation of the methionine cycle, it remains unclear if and how methionine restriction is capable to decrease cancer occurrence risk.

Metabolism of arginine and ornithine involved in linking tricarboxylic acid and urea cycles

Arginine is another nonessential amino acid that can be converted into the nonproteinogenic amino acid ornithine through arginase (ARG1 and ARG2) activity in the last step of the urea cycle. Both arginine and ornithine participate in protein synthesis even if it is important to underline that the latter is a non proteinogenic amino acid, involved only indirectly in protein synthesis and precursor of polyamine synthesis, by the decarboxylation and the production of putrescine catalyzed by ornithine decarboxylase (ODC). ODC is overexpressed in some cancers (Somani et al., 2018) and recently it was reported that ODC inhibition by difluoromethylornithine suppresses the development of esophageal precancerous lesions by downregulating AKT/mTOR/p70S6k, ERK1/2 and p38 α signaling pathways (Xie et al., 2020).

Other data reported the role of urea cycle in cancer cell metabolism linked to TCA cycle through the fumarate (Fig. 16.1). In detail, in TCA cycle the hydration of fumarate into L-malate is catalyzed by fumarate hydratase, whereas in urea cycle the fumarate is obtained through the cleavage of argininosuccinate catalyzed by argininosuccinate lyase. Hence, it has been demonstrated that in cancer cells the fumarate hydratase deficiency induces a high rate of argininosuccinate synthesis in the urea cycle, resulting in a decrease of fumarate levels produced by the TCA cycle, and consequently, in higher arginine demand (Adam et al., 2013; Zheng et al., 2013). On the other hand, it has also been reported that the downregulation of argininosuccinate synthase, that converts aspartate and citrulline into argininosuccinate in urea cycle, supported cancer cells proliferation by increasing the aspartate availability for nucleotide biosynthesis, thus providing a metabolic link between urea cycle and pyrimidine synthesis (Rabinovich et al., 2015).

Overall, these evidences supported the role of arginine deprivation as a promising anticancer therapy (Qiu et al., 2015).

Proline metabolism

Proline, another nonessential amino acid, also plays an important role in metabolic reprogramming correlated to cancer (Phang, 2019). Proline is converted into $\Delta 1$ -pyrroline-5-carboxylate (P5C) through an enzymatic reaction catalyzed by proline dehydrogenase/proline oxidase (PRODH/POX). During this reaction, flavin adenine dinucleotide (FAD) is reduced to FADH₂, contributing to ATP generation in the OXPHOS pathway. On the other hand, P5C is converted into glutamate in a reaction catalyzed by the pyrroline-5-carboxylate dehydrogenase, and, in turn, glutamate is converted into α -ketoglutarate that enters TCA cycle producing ATP, as reported above (Fig. 16.1). Both PRODH overexpression and high levels of free proline are related with ROS production. Hence, it is clear that cancer cells might use proline to produce energy, metabolites (P5C, glutamate, α KG) and to fuel several pathways as well as ROS (D'aniello et al., 2020).

Several evidences suggested that a critical role in cancer cell proliferation and survival is played by PRODH. PRODH expression is induced under hypoxic conditions and contributes to cancer cell survival by inducing autophagy in a mouse xenograft model of human breast cancer (Liu & Phang, 2012). Indeed, PRODH depletion decreased autophagy and increased apoptosis induced by two histone deacetylase inhibitors (trichostatin A and vorinostat) in triple negative breast cancer cells (Fang et al., 2019). Moreover, it was reported that PRODH inhibition blocked breast cancer spheroids growth and lung metastases formation (Elia et al., 2017). Furthermore, PRODH depletion in nonsmall cell lung cancer cells blocked proliferation, migration and epithelial mesenchymal transition (Liu & Phang, 2012).

Lipid synthesis pathway

Acetyl-CoA generated from glucose, glutamine, or acetate, fueled the cholesterol and fatty acids synthesis, both processes enhanced during cancer cell growth. Cholesterol is synthesized through the mevalonate pathway, also known as 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) pathway. The critical role of cholesterol for cancer development and progression is confirmed by several evidences demonstrating the role of statins, the HMGCR inhibitors used as cholesterol lowering agents, in cancer prevention as well as anticancer agents also in combination with other antineoplastic drugs (Di Bello et al., 2020; Iannelli et al., 2020).

While normal cells typically acquire fatty acids by diet, cancer cells present a strong increase in “de novo fatty acid synthesis”. Indeed, fatty acid synthase (FAS) was proposed as a prognostic marker for aggressive breast cancer since 1994 (Kuhajda et al., 1994). Consequently, different FAS inhibitors have been developed as anticancer agents, however the potential toxicity of these inhibitors remains to be tested in clinical trials (Pandey et al., 2012). Another interesting target in fatty acids metabolism is stearoyl-CoA desaturase 1 (SCD1), an enzyme that converts saturated fatty acids to Δ9-monounsaturated fatty acids, whose overexpression has been found in many cancer cells correlating with aggressiveness and poor patient outcomes (Tracz-Gaszewska & Dobrzyn, 2019). Hence, it represents a promising target for anticancer therapy (Noto et al., 2017; Potze et al., 2016).

Another enzyme that should be mentioned as potentially playing a role in lipid metabolism rewiring by cancer cells, is monoacylglycerol lipase, that hydrolyzes monoglycerols and releases glycerol and free fatty acids, which resulted overexpressed in ovarian tumor tissues and is able to promote cancer aggressiveness by increasing the FFA levels (Yecies & Manning, 2010).

Finally, it should be mentioned that enhanced uptake of palmitate and fatty acid oxidation (FAO) represents one of the most used bioenergetic pathways in prostate cancer cells (Liu, Mao et al., 2020). Furthermore, FAO inhibition

demonstrated the anticancer activity in triple-negative breast cancer (Camarda et al., 2016).

Nucleotide biosynthesis pathway

Nucleotides can be obtained by “de novo synthesis pathways” or by “salvage pathways,” through the recycling of existing nucleobases and nucleosides. Cancer cells use the “de novo synthesis pathways” to obtain both purines and pyrimidines by assembling small molecules and amino acids and by generating 5-phosphoribose-1-pyrophosphate (PRPP), the activated form of ribose present in ribose 5-phosphate, obtained through the oxidative and nonoxidative branches of the PPP parallel to glycolysis (Fig. 16.1) (Lane & Fan, 2015). In the case of the pyrimidines, their ring is produced through the assembly of glutamine, bicarbonate, and aspartate, then linked to PRPP through six reactions catalyzed by carbamoyl phosphate synthetase 2, aspartate transcarbamylase, dihydroorotate, dihydroorotate dehydrogenase (DHODH) and UMP synthase. In comparison to pyrimidines, the purine ring is directly built onto PRPP using bicarbonate, glutamine, glycine and THF as substrates, through ten reactions catalyzed by enzymes including inosine monophosphate dehydrogenase (IMPDH), guanosine monophosphate synthetase, adenylosuccinate synthetase and adenylosuccinate lyase.

Notably, the expression level of DHODH has been found to be elevated in various cancers and its cancer-promoting effect was correlated to the pyrimidine synthesis function (Qian et al., 2020). On the other hand, the overexpression of the second isoform of IMPDH, involved in purine synthesis, was demonstrated in liver (He et al., 2018), kidney and bladder cancers (Zou et al., 2015). For these reasons, both these enzymes were indicated as potential cancer biomarkers and therapeutic targets. However, the use of a DHODH inhibitor (brequinar) has not shown significant clinical results (Madak et al., 2019). Conversely, an IMPDH inhibitor (mycophenolic acid) resulted capable to synergize specifically with different chemotherapies depending on the cancer cell type (Lin et al., 2011).

Applications and examples of human cancer metabolomics

Cancer-related metabolic alterations have been the object of many studies since the 1920s, with a particular attention to glucose metabolism. However, a clear understanding of cancer metabolism it is possible only through the evaluation of the so called “metabolome”, the levels of all the metabolites measured in biological samples such as cells, tissues or biological fluids (Liang et al., 2021).

This large-scale study of metabolites in biological samples, also defined metabolomics, takes advantage of several technologies, like $^1\text{H-NMR}$ spectroscopy, GC-MS, LC-MS, FT-IR, Raman and vibrational spectroscopy, that provide

semiquantitative or quantitative information simultaneously about the metabolic intermediates levels, and, hence, a complete picture on the metabolic perturbations induced by a disease. Therefore metabolomic profiling has different applications including such as the possibility to identify metabolites capable to stratify cancers into molecular subclasses, or useful as new noninvasive biomarkers that enable early diagnosis, effective disease monitoring and cancer progression, or are predictive of patients outcome (Kaushik & DeBerardinis, 2018).

In this paragraph, we will report a systematic overview of the most recent studies that used several technologies on different biological matrices (serum/plasma, urine, feces, saliva, tissues, and others) from cancer patients (Fig. 16.2).

Serum/plasma metabolomics studies

Metabolomics analysis in serum is minimally invasive and easily accessible even for disease monitoring by allowing repeated longitudinal measurements, and therefore has strong potential for the identification of novel diagnostic, prognostic and predictive biomarkers, also encouraging better patient compliance (Table 16.1). Moreover, serum metabolomics can also recapitulate alterations derived from tumor–host interaction, as well the effect of host microbiota. Indeed, many studies have been performed in the last years, mainly evaluating differences in serum metabolite concentration between cancer patients and healthy donors.

In the case of the *bladder cancer*, abnormal serum levels of dimethylamine, glutamine, histidine, lactate, malonate, and valine, measured by ¹H-NMR spectroscopy, were used to distinguish patients from healthy controls; conversely dimethylamine, glutamine and malonate discriminated low-grade versus high-grade bladder cancer patients (Bansal et al., 2013). More recently, the above signature was validated by comparing preoperative with postoperative followup levels in bladder cancer patients confirming its role as prognostic marker in this cancer type (Gupta et al., 2020). On the other hand, increased serum levels of taurine and several other amino acids (asparagine, glycine, isoleucine, phenylalanine, proline, serine, tyrosine, and valine), evaluated by LC-MS approach, were observed in bladder cancer smoker patients and correlated with overall survival (Amara et al., 2019).

In *breast cancer* patients GC-MS approach identified some significantly differentiated metabolites (seven for diagnosis, eighteen for grading and twenty-three for staging) overall evidencing that increased glycolysis, lipogenesis and production of volatile organic metabolites are critical metabolic alterations in breast cancer (Hadi et al., 2017). On the other hand, using combined LC-MS and GC-MS approaches, nine serum metabolites (α -ketoglutaric acid, ascorbic acid, creatine, phenylalanine, pyruvate, uric acid, tryptophan, tyrosine and UDP) showed progressive change from control to benign and to invasive ductal breast cancer, suggesting a role in malignant transformation (More et al., 2018). More recently, using LC-MS approach, acylcarnitines and 9,12-linoleic acid were

Table 16.1 Summary of studies related to serum metabolomics profiling on different cancers.

Cancer type	Compared groups	Techniques	References
Bladder cancer	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Bansal et al. (2013)
	Low-grade versus high-grade bladder cancer patients	$^1\text{H-NMR}$	Bansal et al. (2013)
	Preoperative versus postoperative cancer patients	$^1\text{H-NMR}$	Gupta et al. (2020)
	Smoker versus no-smoker cancer patients	LC-MS	Amara et al. (2019)
Breast cancer	Cancer patients versus healthy controls	GC-MS	Hadi et al. (2017)
	Controls vs benign versus invasive ductal cancer patients	LC-MS and GC-MS	More et al. (2018)
	Early cancer patients versus healthy controls	LC-MS	Lin, Xu et al. (2019), Lin, Ma et al. (2019)
	Pre versus postchemotherapy patients	LC-MS	Lin, Xu et al. (2019), Lin, Ma et al. (2019)
Cartilage tumor	Cancer patients versus healthy controls	$^1\text{H-NMR}$	López-Garrido et al. (2020)
Colorectal cancer	Local (stage II and III) versus liver-limited metastasis versus extrahepatic metastasis patients	GC-MS and $^1\text{H-NMR}$	Farshidfar et al. (2012)
	Metastasis patients versus healthy controls	$^1\text{H-NMR}$	Bertini et al. (2012)
	Colorectal polyps versus early cancer patients	$^1\text{H-NMR}$	Gu et al. (2019)
	Colon versus rectal cancer patients	GC-MS	Wu et al. (2020)
	Pre versus postneoadjuvant chemo-radiation therapy	LC-MS	Jia et al. (2018)
	Advanced colon cancer (stage T3) with versus without lymph node metastasis	LC-MS	Zhang et al. (2020)
	CRC patients (stage I to IV)	LC-MS	Rachieriu et al. (2021)
Endometrial cancer	Endometrial cancer patients versus healthy controls, benign endometrial disease and ovarian cancer	GC-MS	Troisi et al. (2018, 2020)
	Cancer patients versus healthy controls	LC-MS	Shi et al. (2018)
	Cancer patients versus healthy controls	LC-MS	Kozar et al. (2020)

(Continued)

Table 16.1 Summary of studies related to serum metabolomics profiling on different cancers. *Continued*

Cancer type	Compared groups	Techniques	References
Esophageal cancer	Cancer patients versus healthy controls	GC-MS	Zhu et al. (2017)
	Cancer patients versus healthy controls	LC-MS	Zhu et al. (2020)
	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Ye et al. (2021)
Gastric cancer	Chronic gastritis versus gastric cancer	LC-MS	Yu et al. (2021)
	Gastric cancer patients at initial stage versus healthy controls	LC-MS	Wang et al. (2017)
Glioma	Cancer patients versus healthy controls	LC-MS	Huang et al. (2017)
	Patients with catecholamine-producing pheochromocytomas and paragangliomas before versus after surgery	LC-MS	Erlic et al. (2019)
Head and neck cancer	Before versus after radiotherapy cancer patients	GC-MS	Wojakowska et al. (2020)
	Cancer patients treated with chemo-radiotherapy versus radiotherapy alone versus chemotherapy	LC-MS	Jelonek et al. (2020)
Hepatocellular carcinoma	Early versus advanced cancer patients	$^1\text{H-NMR}$	Casadei-Gardini et al. (2020)
	Liver cirrhosis versus cancer patients	LC-MS	Han et al. (2019)
	Cancer patients versus healthy controls	LC-MS	Stepien et al. (2021)
Kidney cancer	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Nizioł, Copié et al. (2021), Nizioł, Ossoliński et al. (2021)
Leukemia	Patients pretreated with all-trans retinoic acid and valproic acid combined with low-dose cytotoxic drugs (responders versus nonresponders)	LC-MS	Grønningssæter et al. (2019)
Lung cancer	Initial primary versus second primary cancer patients	LC-MS	Aredo et al. (2021)
	Early cancer patients versus healthy controls	LC-MS	Ros-Mazurczyk et al. (2017)

(Continued)

Table 16.1 Summary of studies related to serum metabolomics profiling on different cancers. *Continued*

Cancer type	Compared groups	Techniques	References
Melanoma	Cancer patients versus healthy controls	GC-MS	Callejón-Leblíc et al. (2019)
	Early cancer patients versus healthy controls	¹ H-NMR	Berker et al. (2019)
	Patients subjected to different immunotherapy	¹ H-NMR	Costantini et al. (2018)
Mesothelioma	Cancer patients versus healthy controls	LC-MS	Di Gregorio et al. (2021)
Ovarian cancer	Cancer patients versus healthy controls	LC-MS	Yang et al. (2018)
	Cancer patients versus healthy controls	LC-MS	Plewa et al. (2017)
	Cancer patients versus healthy controls	LC-MS	Wang et al. (2021)
Pancreatic cancer	Cancer patients versus healthy controls	¹ H-NMR	OuYang et al. (2011)
	Cancer patients versus healthy controls	¹ H-NMR	Zhang et al. (2012)
	Benign versus malignant pancreatic lesions	¹ H-NMR	Bathe et al. (2011)
	Pancreatic cancer and new onset diabetic mellitus versus new onset diabetic mellitus patients	LC-MS	He et al. (2017)
	Pancreatic cancer versus biliary tract cancer and intraductal papillary mucinous carcinoma patients	Capillary electrophoresis MS	Itoi et al. (2017)
	Pancreatic ductal adenocarcinoma patients versus healthy controls	LC-MS	Martín-Blázquez et al. (2020)
	Distal cholangiocarcinoma versus pancreatic ductal adenocarcinoma patients	LC-MS	Macias et al. (2020)
Prostate cancer	Cancer patients versus healthy controls	LC-MS	Miyagi et al. (2011)
	Benign prostatic hypertrophy versus prostate cancer	¹ H-NMR, LC-MS and GC-MS	Giskeodegard et al. (2015)
	Low grade versus high grade cancer patients	¹ H-NMR	Kumar et al. (2015)
	Gleason score 6 (3 + 3) and ≥ 7 cancer patients	LC-MS and GC-MS	Penney et al. (2021)

(Continued)

Table 16.1 Summary of studies related to serum metabolomics profiling on different cancers. *Continued*

Cancer type	Compared groups	Techniques	References
Soft tissue sarcomas	Cancer patients treated with the trabectedin regimen versus untreated	LC-MS	Miolo et al. (2020)
Thyroid	Cancer patients versus healthy controls	LC-MS	Du et al. (2021)

Table shows cancer type, the compared groups, the used techniques and the related references.

identified as metabolites useful in early breast cancer detection (Kozar et al., 2020). Moreover, nine metabolites (cysteinyl-lysine, ethyl docosahexaenoic, hulu-papeptide, lysophosphatidylethanolamine 0:0/22:4, methacholine, oleic acid amide, prostaglandin C1, ricinoleic acid and vitamin K2) were involved in the prediction of chemotherapy response in breast cancer patients (Lin et al., 2011).

In the case of *colorectal cancer* (CRC), using ^1H -NMR spectroscopy and GC-MS a metabolic signature was identified to discriminate between patients with local (stage II and III) CRC, liver-limited metastasis, or extrahepatic metastasis (Farshidfar et al., 2012). A study using ^1H -NMR approach demonstrated that metastatic CRC patients showed higher serum levels of 3-hydroxybutyrate, acetate, formate, glycerol, *N*-acetyl signal of glycoproteins, phenylalanine and proline, and lower serum levels of alanine, citrate, creatine, glutamine, lactate, leucine, pyruvate, tyrosine, and valine, compared to healthy controls and some of them were associated with short overall survival (Bertini et al., 2012). Another NMR-based study showed that the glycerolipid and pyruvate metabolisms were increased in colorectal polyps, and the lactate/citrate and acetate/glycerol rates could be used as diagnostic tool for early diagnosis of CRC (Gu et al., 2019). In summary a recent review evidenced that the most affected pathways in serum metabolomics studies of CRC patients were alanine metabolism, ammonia recycling, TCA cycle, glutathione metabolism, protein biosynthesis and urea cycle, confirming several metabolites levels (3-hydroxybutyric acid, creatinine, fumaric acid, glucose, lactic acid, malic acid, ornithine, phenylalanine, pyruvic acid, tryptophan and tyrosine) as significantly different between cancer and control samples in different studies (Amir Hashim et al., 2019). GC-MS approach was used to compare the metabolic characteristics of colon and rectal cancer patients showing that:

1. glucose and mannose altered levels were useful for colon cancer diagnosis;
2. 2-aminobutanoic acid, 3-hydroxypyridine, glucose, isoleucine, mannose, tryptophan, urea, and uric acid were useful for rectal cancer diagnosis;
3. all these metabolites improved the diagnostic accuracy of two commonly used protein markers (CEA and CA199) for both colon and rectal cancer diagnosis (Wu et al., 2020).

Using LC-MS approach, a panel of 15 metabolites was identified as predictive of response to neoadjuvant chemoradiation therapy in locally advanced rectal cancer suggesting their use for personalized treatment strategies (Jia et al., 2016). Another LC-MS serum metabolomics study selected abscisic acid, calcitroic acid and glucosylsphingosine as new good biomarkers to detect the occurrence of lymph node metastasis and to predict the survival in T3 stage colon cancer patients (Zhang et al., 2020). Furthermore, a recent high-performance LC-MS study identified several lipids, including ceramides, fatty acids, phosphatidylcholines, phosphatidylethanolamines, and sterol esters, overexpressed in CRC patients (stage I to IV) sera and associated with cancer development and progression (Răchieriu et al., 2021).

GC-MS approach metabolic profiling has been performed on sera of *esophageal cancer patients* demonstrating that four metabolites, including glucose, glutamate, lactate, and cholesterol, were strictly correlated with esophageal cancer diagnosis and prognosis (Wei et al., 2019; Yu et al., 2020; Zhu & Leung, 2020). Moreover, Zhu et al. (2020) by ultra-performance LC-MS demonstrated that indoleacrylic acid, lysophosphatidylcholine and lysophosphatidylethanolamine were associated with esophageal squamous cell carcinoma progression. Furthermore, a ¹H-NMR spectroscopy-based study evidenced lower levels of alanine, lactate and valine, and higher levels of acetate, pyruvate, glutamate, succinate, citric acid, glucose and serine in esophageal cancer patients compared to healthy donors, highlighting the alteration of metabolic pathways such as TCA cycle, glutaminolysis, pyruvate and lipidic metabolism (Ye et al., 2021).

Some ¹H-NMR-based serum metabolomics studies have been performed on *pancreatic cancer* showing:

1. higher levels of creatinine, isoleucine, leucine and triglyceride, and lower levels of 3-hydroxybutyrate, 3-hydroxyisovalerate, lactate and trimethylamine-N-oxide in cancer patients compared to healthy donors (OuYang et al., 2011);
2. higher levels of acetone, dimethylamine and N-acetyl glycoproteins, and lower levels of alanine, citrate, glutamate, glutamine, histidine, isoleucine, lysine, leucine and valine in cancer patients compared to healthy controls (Zhang et al., 2020);
3. higher levels of glutamate and glucose and lower levels of creatine and glutamine in pancreatic cancer compared with benign pancreatic lesions (Bathe et al., 2011).

Capillary electrophoresis mass spectrometry was used to analyze charged metabolites in sera of patients with pancreatic cancer and other malignant diseases such as biliary tract cancer and intraductal papillary mucinous carcinoma and showed the possibility to use isocitrate and threonine to discriminate pancreatic cancer patients from those with biliary tract cancer and intraductal papillary mucinous carcinoma (Itoi et al., 2017). Several LC-MS-based serum metabolomics studies have been reported pancreatic cancer. One study highlighted valine, leucine and isoleucine biosynthesis and degradation, primary bile acid biosynthesis, and sphingolipid metabolism, as the top three metabolic pathways associated with pancreatic cancer related

with diabetic mellitus (He et al., 2018). Another recent study demonstrated that glycerolipid, glycerophospholipid and linoleic acid metabolism, and primary bile acid biosynthesis were altered in pancreatic ductal adenocarcinoma patients versus healthy controls (Martín-Blázquez et al., 2020). A third report showed that the combination of CA19-9 plus serum levels of acylcarnitine, ceramide, phosphatidylcholines, lysophosphatidylcholines, lysophosphatidylethanolamine and sphingomyelins, distinguished between patients with distal cholangiocarcinoma and pancreatic ductal adenocarcinoma (Macias et al., 2020).

In the case of *gastric cancer*, an analysis of serum metabolomic profiles by LC-MS indicated that hexadecaspinganine, linoleamide, and *N*-Hydroxyarachidonoyl amine were good diagnostic markers for chronic gastritis and gastric cancer, and the serum levels of total apolipoprotein A1, cholesterol and high-density lipoprotein cholesterol were lower in gastric cancer patients versus chronic gastritis patients, suggesting a pathogenesis route (Wei et al., 2019; Yu et al., 2020; Zhu & Leung, 2020). Moreover, lower serum levels of 2,4-hexadienoic acid, 4-methylphenyl dodecanoate and glycerol tributanoate resulted to be independent prognostic factors of gastric cancer (Wang et al., 2019). Recently, a *meta*-analysis focused on the most current serum metabolomics studies has highlighted that the serum levels of alanine and asparagines increased, whereas those of histidine and tryptophan decreased over the course of the precancerous lesions of gastric cancer, evidencing that these four metabolites, involved in aminoacyl-tRNA biosynthesis and tryptophan catabolism, could be useful for early detection of gastric cancer (Ren et al., 2021).

Serum GC-MS-based metabolomics approach on *endometrial cancer patients* evidenced that the levels of lactate, progesterone, homocysteine, 3-hydroxybutyrate, linoleic acid, stearic acid, myristic acid, threonine, and valine were statistically different in cancer patients compared to healthy controls, benign endometrial disease or ovarian cancer (Troisi et al., 2018). This metabolic signature has been also validated on a greater cohort of patients confirming its clinical application (Troisi et al., 2020). Another study by LC-MS has been also used to search differentially expressed metabolites in early stage endometrial carcinoma, identifying a crucial role in the modulation of cancer cell behavior, of indoleacrylic acid, lyso-platelet-activating factor-16, phenylalanine, and phosphocholine (Shi et al., 2018). Very recently, using UPLC-TQ/MS, ceramides, acylcarnitines and 1-methyladenosine were identified as potential diagnostic biomarkers for endometrial cancer patients with localized (I–II) stage (Kozar et al., 2021).

Metabolomics analysis has been performed on serum samples of *glioma* by LC-MS evidencing that low levels of 2-oxoarginine, cysteine, α -ketoglutarate, chenodeoxycholate and argininate correlated with overall glioma risk (Huang et al., 2017). LC-MS also identified different serum levels of metabolites such as creatinine, glycerophospholipids, hexose, histidine, ornithine, sarcosine, sphingomyelins and tyrosine in patients with catecholamine-producing *pheochromocytomas* and *paragangliomas*, before and after surgery (Erlic et al., 2019).

¹H-NMR metabolomic profiling on sera of *HCC* patients showed statistically different levels of 1-methylhistidine, alanine, glucose, glutamine, lactate, lysine

and valine between advanced and early HCC patients, and serum tyrosine levels as good overall survival predictors for early HCC patients (Casadei-Gardini et al., 2020). Moreover, metabolomics profiling by LC-MS evidenced that four metabolites such aschenodeoxycholic acid, LPC20:5, succinyladenosine and uridine discriminated HCC from liver cirrhosis (Han et al., 2019). Stepien et al. (2021) applied high resolution LC-MS on sera of HCC patients collected prior to cancer diagnosis within the prospective European Prospective Investigation into Cancer and Nutrition cohort, and highlighted the role of metabolic pathways associated with amino acids, bile acids, steroid and phospholipids metabolism, diet, immunity and environmental exposures in HCC development (Stepien et al., 2021).

Recently, the metabolite profiles of the whole serum and serum-derived exosomes in healthy controls and *head and neck cancer* patients before and after radiotherapy were analyzed by GC-MS demonstrating that the metabolites affected by radiotherapy were associated with amino acids, lipids, nucleotides and sugars metabolism (Wojakowska et al., 2020). The combination of direct flow injection and LC-MS on sera of head and neck cancer patients treated with concurrent chemo-radiotherapy, radiotherapy alone, or induction chemotherapy evidenced that chemo-radiotherapy induced stronger effects than radiotherapy alone whereas chemotherapy alone did not induce significant changes, and the decreased levels of total phospholipids represented the most evident effect during the first step of the treatment (Jelonek et al., 2020).

Serum *lung* cancer metabolomics performed by $^1\text{H-NMR}$ showed that prolonged survival has been associated with low levels of glutamate and lipids and high levels of glycine, glutamine and valine (Berker et al., 2019). On the other hand, metabolomic profiling by GC-MS confirmed that amino acid metabolism was the most affected in lung cancer, and suggested that asparagine, glycine and valine could represent possible diagnostic biomarkers (Callejón-Leblíc et al., 2019). Moreover, lower levels of lysophosphatidylcholines detected by LC-MS showed potential usefulness in discriminating between early lung cancer patients and healthy controls (Ros-Mazurczyk et al., 2017). Recently, in a pilot case-control study by ultra-high performance LC-MS approach serum metabolomics studies on initial primary and second primary lung cancer patients showed that increased levels of 5-methylthioadenosine and phenylacetylglutamine resulted associated with second primary lung cancer (Aredo et al., 2021).

In the case of *prostate* cancer, many studies on serum metabolomic profiling were conducted (Lima et al., 2016). Through LC-MS approach, an alteration in free amino acid metabolism, including lower levels of glutamine, tryptophan and valine and higher levels of alanine, isoleucine, lysine and ornithine, were found in sera of prostate cancer patients compared to healthy controls (Miyagi et al., 2011). Moreover, the comparison between metabolomic profiling in sera of patients with benign prostatic hypertrophy and prostate cancer, performed by a combination of NMR, LC-MS and GC-MS approaches, evidenced a dysregulation of fatty acids, membrane phospholipidic and amino acid metabolisms in prostate cancer patients (Giskeødegård et al., 2015). Furthermore, $^1\text{H-NMR}$ -based

metabolomics studies showed higher levels of alanine, pyruvate and sarcosine, and lower levels of glycine, in high grade compared to low grade prostate cancer patients (Kumar et al., 2015) confirming that sarcosine can be considered as a prostate cancer aggressivity marker (Koutros et al., 2013). Finally, a comparison between serum metabolomic profiling in prostate cancer patients with Gleason score 6 (3 + 3) versus those with ≥ 7 , by LC-MS and GC-MS, evidenced higher levels of 2-hydroxyhippurate, N-methyl proline and xylitol, and lower levels of alanine, arginine, histidine, methionine, threonine, tryptophan and tyrosine in patients with Gleason score ≥ 7 (Penney et al., 2021).

An LC-MS-based serum metabolomic profiling study on *ovarian cancer* patients evidenced that the most altered pathways were fatty acid β -oxidation, phospholipid and bile acid metabolism, and that 2-piperidone and 1-heptadecanoylglycerophosphoethanolamine were potential biomarkers useful for clinical ovarian cancer diagnosis and treatment (Yang et al., 2018). Another similar study highlighted histidine and citrulline as new potential biomarkers for ovarian cancer development and progression (Plewa et al., 2017). An additional work demonstrated that serum levels of asparagine, glutamate, glutamine, glycolic acid and methionine resulted useful to distinguish epithelial ovarian cancer from healthy control, thereby as potential screening tools for this cancer (Wang et al., 2021). By ^1H -NMR approach it has been demonstrated that taurine and hypotaurine metabolism as well as synthesis and degradation of ketone bodies were statistically different in the sera of *cartilage tumor* patients compared to healthy controls (López-Garrido et al., 2020). ^1H -NMR spectroscopy showed that lower levels of choline, glycerol, glycine, lactate and leucine and higher levels of glucose can be useful to improve *kidney* cancer diagnosis and/or prognosis (Nizioł, Copié et al., 2021; Nizioł, Ossoliński et al., 2021). ^1H -NMR approach has been also applied on sera collected in metastatic *melanoma* patients subjected to different immunotherapy treatments, demonstrating that patients with different outcome grouped in separate clusters, and that a set of metabolites are able, at baseline, to discriminate patients with favorable or worst outcome (Costantini et al., 2018).

Serum metabolomic profiling by LC-MS evidenced that in patients with acute myeloid *leukemia* receiving treatment based on all-trans retinoic acid and valproic acid combined with low-dose cytotoxic drugs, the metabolites involved in fatty acid and amino acid pathways were more altered in responders versus nonresponders patients (Grønningæter et al., 2019).

Another LC-MS study showed that phenylalanine and glutamate are useful to discriminate between *thyroid* papillary carcinoma patients and healthy controls (Du et al., 2021).

Radical hemithoracic radiotherapy induced significant changes in serum metabolomics profile, by LC-MS, of malignant pleural *mesothelioma* patients mainly affecting arginine and polyamine biosynthesis pathways (Di Gregorio et al., 2021).

The serum levels of lactate and glutamate can be used as potential adjuvant diagnostic biomarkers for primary *osteosarcoma* (Lv et al., 2020). On the other

hand, in the case of patients with *soft tissue sarcomas* treated with trabectedin regimen, only citrulline and histidine correlated significantly with overall survival (Miolo et al., 2020).

Urine metabolomics studies

Urine metabolomics has been the focus of a large body of scientific studies in cancer patients because this body fluid contains many small metabolites, is collected noninvasively and can be obtained in large amounts.

Mainly, this approach has been used to find new biomarkers for urological cancers (Table 16.2). Urine levels of trans-2-dodecenoylcarnitine, serinyl-valine, feruloyl-2-hydroxyputrescine, and 3-hydroxy-nonanoylecarnitine, measured by LC-MS approach, discriminate *bladder cancer* from benign bladder lesions, whereas indolylacryloylglycine, N2-galacturonyl-L-lysine, and aspartyl-glutamate were used to distinguish high versus low grade bladder cancers (Liu et al., 2018). Same authors searched for urine metabolites to detect bladder cancer at an early stage, such as nonmuscle invasive, using LC-HRMS approach, and demonstrated that:

1. 4-sulfate, aspartyl-histidine, and tyrosyl-methionine had predictive ability to discriminate between the patients and the control group;
2. 3-hydroxy-cis-5-tetradecenoylcarnitine, 6-ketoestriol, beta-cortolone, heptylmalonic acid, and tetrahydrocorticosterone selected low versus high grade nonmuscle invasive bladder cancer;
3. tryptophan metabolism was upregulated in high grade nonmuscle invasive bladder cancer patients with the presence or absence of hematuria (Cheng et al., 2018).

¹H-NMR approach has been used to select urinary metabolomic signature of recurrence during followup of nonmuscle invasive bladder cancer patients, evidencing that the altered levels of alanine, aspartate, glutamate, phenylalanine and taurine were associated with relapse (Loras et al., 2019). Recently, GC-MS-based metabolomics studies have been applied on urine from bladder cancer patients with different tumor stages (Ta/Tis, T1and \geq T2). First of all, the analyses unveiled that all these bladder cancer patients presented lower levels of aldehydes, ketones and monoterpenes, and higher levels of several alkanes and aromatic compounds; then distinct urinary volatile profiles were found in patients with \geq T2 stage, including higher levels of 2,4 dimethylheptane, 1,2,3 trimethylbenzene, 4-methyloctane and levomenthol and lower levels of 2,5 dimethylbenzaldehyde (Pinto et al., 2021).

Many studies were performed in the last years on urine metabolomic profiling from *prostate cancer* patients. An early study by LC-MS approach demonstrated the upregulation of sarcosine, kynurenine, uracil and glycerol 3-phosphate in prostate cancer patients associated with progression (Jiang et al., 2010;

Table 16.2 Summary of studies related to urinary metabolomics profiling on different cancers.

Cancer	Comparison groups	Technique	References
Bladder cancer	Nonmuscle invasive bladder cancer versus benign bladder lesions: low grade versus high grade nonmuscle invasive bladder cancer patients	LC-MS	Cheng et al. (2018)
	Cancer versus benign bladder lesions: low grade versus high grade cancer patients	LC-MS	Liu et al. (2018)
	Followup of nonmuscle invasive bladder cancer patients	$^1\text{H-NMR}$	Loras et al. (2019)
	Ta/Tis versus T1 versus $\geq \text{T2}$ tumor stage of bladder cancer patients	GC-MS	Pinto et al. (2021)
Breast cancer	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Silva et al. (2019)
Cholangiocarcinoma	Asiatic versus united kingdom cancer patients	LC-MS	Alsaleh, Barbera et al. (2019), Alsaleh, Sithithaworn et al. (2019)
Colorectal cancer	Cancer versus healthy controls	LC-MS	Deng et al. (2019)
	Preinvasive colorectal neoplasia versus advanced CRC versus healthy controls	$^1\text{H-NMR}$	Kim et al. (2019)
	Cachectic versus precachectic versus noncachectic cancer patients	$^1\text{H-NMR}$	Ose et al. (2019)
	Healthy controls versus polyps versus CRC patients	Capillary electroforesis MS	
Endometrial cancer	Healthy controls versus endometrial hyperplasia versus cancer patients	LC-MS	Shao et al. (2016)
Esophageal cancer	Cancer versus healthy controls	$^1\text{H-NMR}$	Liang et al. (2019)
	Esophageal squamous cell carcinoma patients versus healthy controls	LC-MS	Xu et al. (2016)

(Continued)

Table 16.2 Summary of studies related to urinary metabolomics profiling on different cancers. *Continued*

Cancer	Comparison groups	Technique	References
Gastric cancer	Cancer versus benign gastric disease versus healthy controls	$^1\text{H-NMR}$	Chan et al. (2016)
	Cancer versus healthy controls	LC-MS	Chen et al. (2016)
Glioma	Pheochromocytoma and paraganglioma patients with versus without SDHx mutations	$^1\text{H-NMR}$	Martins et al. (2019)
Hepatocellular carcinoma	Cancer patients versus healthy controls	LC-MS	Liang et al. (2016)
	Healthy controls versus liver cirrhosis versus cancer patients	GC-MS	Osman et al. (2017)
Kidney cancer	Patients with benign masses and oncocytomas versus chromophobe kidney cancer patients	$^1\text{H-NMR}$	Falegan et al. (2019)
	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Niziot, Copié et al. (2021), Niziot, Ossoliński et al. (2021)
Laryngeal cancer	Cancer patients versus healthy controls	LC-MS	Chen et al. (2019)
Lung cancer	Cancer patients versus healthy controls	GC-MS	Callejón-Leblic et al. (2019)
	African american versus white patients	GC-MS	Dator et al. (2020)
Ovarian cancer	Bordeline versus malignant cancer patients	LC-MS	Liu, Liu et al. (2020), Liu, Mao et al. (2020)
Pancreatic cancer	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Davis et al. (2013)
Prostate cancer	Cancer patients versus healthy controls	LC-MS	Lee, Mahmud et al. (2020), Lee, Ang et al. (2020)
	Cancer patients versus healthy controls	GC-MS	Lima et al. (2019)
	Prostate cancer versus benign prostatic hyperplasia patients	$^1\text{H-NMR}$	Pérez-Rambla et al. (2017)
	Cancer patients versus healthy controls	LC-MS	Sreekumar et al. (2009), Jiang et al. (2010)

(Continued)

Table 16.2 Summary of studies related to urinary metabolomics profiling on different cancers. *Continued*

Cancer	Comparison groups	Technique	References
Thyroid	Gleason score 6 versus ≥ 7 in prostate cancer patients Healthy controls versus nonneoplastic nodules versus follicular adenoma versus papillary thyroid cancer	$^1\text{H-NMR}$ NMR	Yang et al. (2021) Wojtowicz et al. (2017)

Table shows cancer type, the compared groups, the used techniques and the related references.

Sreekumar et al., 2009). More recently, still by LC-MS approach, higher levels of other metabolites including intermediates of pyruvate metabolism and TCA cycle, and some amino acids [alanine, aspartate, glutamate and branched-chain amino acids (BCAAs)], were found in urine samples of prostate cancer patients and involved in cancer progression (Lee, Ang et al., 2020; Lee, Mahmud et al., 2020). $^1\text{H-NMR}$ approach evidenced in urine samples of prostate cancer compared to benign prostatic hyperplasia patients, higher levels of BCAAs, glutamate and pseudouridine, and lower levels of 4-imidazole-acetate, dimethylglycine, fumarate and glycine (Pérez-Rambla et al., 2017). Recently, still by $^1\text{H-NMR}$ -based urine metabolomics, it has been demonstrated that the combination of glycine, guanidinoacetate, and phenylacetylglycine was able to differentiate prostate cancer patients from healthy controls and to stratify patients with Gleason score 6 versus ≥ 7 (Yang et al., 2018). On the other hand, GC-MS approach identified six urinary metabolites, such as 2,5-dimethylbenzaldehyde, 3-phenylpropionaldehyde, 4-methylhexan-3-one, dihydroedulan IA, hexanal, and methylglyoxal, able to improve prostate cancer diagnosis (Lima et al., 2019).

In the case of *esophageal cancer*, lower levels of creatinine, ethanolamine, glucose, glycine, hippurate and taurine, and higher levels of acetoacetate, cis-aconitate, citrate and glutamate involved mainly in TCA cycle, glycolysis, fatty acid metabolism, and amino acid metabolism were found in patients urine samples, by $^1\text{H-NMR}$ -based metabolomics (Liang et al., 2021). Another study by LC-MS approach identified potential urinary biomarkers associated with esophageal squamous cell carcinoma that resulted to be associated with perturbations of fatty acid β -oxidation and the metabolism of amino acids as well as of purines, and pyrimidines (Xu et al., 2016).

$^1\text{H-NMR}$ -based urine metabolomics study identified 2-hydroxyisobutyrate, 3-indoxylsulfate, and alanine as the three metabolites enabling the discrimination between *gastric cancer*, benign gastric disease, and healthy controls (Chan et al., 2016). GC-MS-based urine metabolomic profiling identified ten amino acids (alanine, glycine, isoleucine, methionine, proline, serine, threonine, tryptophan, tyrosine

and valine) and four organic molecules (ethyl 2-methylacetoacetate, levulinic acid, p-cresol and benzylmalonic acid) with high diagnostic value for gastric cancer; conversely, proline, p-cresol and 4-hydroxybenzoic acid demonstrated high outcome-prediction value (Chen et al., 2016).

In the case of *CRC* patients, urinary metabolic profiles by ¹H-NMR in newly diagnosed CRC patients (stage I–IV) classified as cachectic, pre cachectic or noncachectic showed higher levels of acetone and arginine in cachectic compared to non-cachectic patients (Ose et al., 2019). Another NMR-based urine metabolomic profiling selected taurine, alanine, and 3-aminoisobutyrate as good discriminators between healthy individuals and both preinvasive colorectal neoplasia and advanced CRC, suggesting a potential for early-detection during screening (Kim et al., 2019). Similarly, LC-MS approach on urine samples from CRC patients collected in two study sites (Canada and United States), identified diacetylspermine and kynurenone as good predictors of CRC (Deng et al., 2019). More recently, capillary electrophoresis-time-of-flight MS has been used to quantify hydrophilic urinary metabolites in patients with stage 0 to IV CRC or polyps, versus healthy controls. This study highlighted higher levels of malate and citrate (TCA cycle), kynurenone (tryptophan pathway), and cystine (PPP), in CRC patients compared to healthy controls; only citrate had statistically significant different levels between polyps group and healthy controls and between CRC group and healthy controls; the combination of butyrate, 3-hydroxy-3-methylglutarate, and carnosine was able to discriminate polyps and CRC groups from healthy controls (Udo et al., 2020).

In the case of *HCC*, urinary metabolomic profiling performed by LC-Q-TOF-MS approach evidenced that five metabolites (α -N-Phenylacetyl-L-glutamine, glycocholic acid, indoleacetyl glutamine, palmitic acid, and phytosphingosine) had diagnostic value (Liang et al., 2016). Through GC-MS-based urinary metabolomics profiling it was possible to identify 13 metabolites (arabinose, citric acid, glycerol, glycine, hippuric acid, phosphate, proline, pyrimidine, serine, threonine, urea, xylitol, and xyloonic acid) able to distinguish between HCC patients and healthy controls; and eight metabolites (arabinose, citric acid, glycine, proline, serine, threonine, urea and xylitol) that discriminate between the liver cirrhosis group and healthy controls (Osman et al., 2017). The urinary metabolomics profiles of *cholangiocarcinoma* patients have been evaluated in either Thailand or United Kingdom (UK), demonstrating increased purine recycling and FAO and, hence, perturbations related to purine metabolism and lipid metabolism in Asiatic patients (Alsaleh, Barbera et al., 2019; Alsaleh, Sithithaworn et al., 2019); and altered acylcarnitine, bile acid and purine levels in UK patients (Alsaleh, Sithithaworn et al., 2019).

¹H-NMR-based metabolic profiles have been evaluated also on urine samples of *papillary thyroid* cancer patients evidencing:

1. higher levels of 2-furoylglycine and lower levels of 3-indoxylsulfate, isopropanol, acetone, citrate, 3-hydroxybutyrate, 2-phenylpropionate, compared to healthy controls;

2. higher levels of 2-phenylpropionate and 3-indoxylsulfate compared to patients with nonneoplastic nodules; and lower levels of tyrosine, 3-hydroisovalerate compared to patients with follicular adenoma (Wojtowicz et al., 2017).

In *pheochromocytoma* and *paraganglioma* ¹H-NMR-based urine metabolomics showed, in patients with succinate dehydrogenase (SDHx) mutations-associated tumors, higher levels of *N*-acetylaspartate, 3-hydroxyisovalerate and lactate (Martins et al., 2019).

Two ¹H-NMR studies evidenced in *kidney cancer* patients higher levels of myo-inositol, creatine, sucrose, *N*-dimethylglycine, 4-hydroxyphenylacetate and urea and lower levels of trimethylamine, compared to healthy controls (Nizioł, Ossoliński et al., 2021); or alterations in the levels of TCA cycle intermediates, carnitines and its derivatives, compared to patients with benign masses (Falegan et al., 2019).

In *breast cancer* patients an urinary ¹H-NMR-based metabolomics study demonstrated that creatine, glycine, trimethylamine *N*-oxide, and serine could significantly discriminate cancer patients versus healthy controls (Silva et al., 2019). In *pancreatic cancer* patients, still by ¹H-NMR-based metabolomics, higher levels of 1-methylnicotinamide, acetone, choline, dimethylamine, fucose, hypoxanthine, o-acetylcarnitine, and threonine, and lower levels of methanol and trigonelline, were demonstrated compared to healthy controls (Davis et al., 2013).

LC-MS platform was used for the following studies on urine samples, demonstrating: in *ovarian cancer*, higher levels of coniferyl alcohol, indoleacrylic acid and cerulenin, and lower levels of 16a-hydroxyestrone and (E)-casimiroedine, as markers of progression from borderline to malignant tumors (Liu et al., 2018); in *laryngeal cancer patients*, six metabolites (myristic acid, oleamide, palmitic acid, pantothenic acid, phytosphingosine and sphinganine) differentially expressed compared to healthy controls (Chen et al., 2016); in *endometrial cancer* patients, lower levels of acetylcysteine and porphobilinogen and higher levels of *N*-acetylserine, isobutyrylglycine and urocanic acid compared to healthy controls and endometrial hyperplasia (Shao et al., 2016).

GC-MS approach was applied to identify the altered metabolites in urine of *lung cancer* patients compared to healthy controls demonstrating higher levels of aconitic acid, adipic acid, malonic acid, ribitol, stearic acid, succinic acid, threonine, and uric acid and lower levels of glycine, inositol, isocitric acid, glucaric acid and phosphoric acid (Callejón-Leblíc et al., 2019). Finally, an interesting study, by LC-MS-based metabolomics, compared urine samples from whites United States citizen versus african american smokers, the latter normally at a higher risk to develop lung cancer, evidencing that in the two groups there were significant differences in the metabolic pathways including amino acids, carbohydrates, fatty acids, nicotine and nucleotides metabolism (Dator et al., 2020).

Tissue metabolomics studies

In comparison with other matrices, tumor tissues metabolomics has many advantages because it allows the analysis on the focal location, highlighting tumor

cell-intrinsic alterations as well as those metabolites that could become elusive after secretion into biofluids. However, the evaluation of tissue metabolomics is complicated by tissue heterogeneity and normally does not allow multiple longitudinal sampling during disease evolution/treatments (Vorkas & Li, 2018) (Table 16.3).

The combination of LC-MS and GC-MS approaches has been applied on invasive ductal *breast carcinoma* samples compared to benign and peripheral normal tissues identifying 36 upregulated and six downregulated metabolites including amino acids, amino sugar derivatives, fatty acids, nucleotides and organic acids (More et al., 2018). In another study, two independent analytic platforms, hydrophilic interaction chromatography (HILIC) and GC-MS, combined with different bioinformatic tools, analyzed breast cancer tissues compared to normal-adjacent mucosa, suggesting that glutamine metabolism, phosphatidylethanolamine biosynthesis, urea cycle, and ammonia recycling were significantly associated with breast cancer (Fan et al., 2020).

On *CRC* samples, LC-MS-based metabolomics analyzed tumor and normal mucosa in both formalin-fixed paraffin-embedded (FFPE) and fresh frozen tissues. Fifteen metabolites resulted altered and higher levels of hypoxanthine, iso-leucine, leucine, phenylalanine, proline and taurine, and lower levels of 1,11-undecanedicarboxylate, α -ketoglutarate, β -hydroxyisovalerate, carboxyethyl-GABA, cis-4-decanoyl carnitine, creatinine, sebacate, undecanedoate and urea distinguished tumor from normal tissues (Udo et al., 2020). In another study on CRC tissues compared with matched adjacent mucosa, high performance LC-MS approach identified higher levels of carbohydrates, choline, free fatty acids and nucleotides. Moreover, higher levels of lipid metabolites and lower levels of dipeptides were demonstrated in late-stage (III and IV) compared to early-stage (I and II) CRC; and statistically significant differential levels of forty-three metabolites, mainly lipids, discriminated adenocarcinoma versus nonadenocarcinoma tumors (Long et al., 2020).

Metabolomic profiling on *gastric cancer* tissues by capillary electrophoresis time of flight MS evidenced higher levels of lactate and lower levels of adenylate energy charge in cancer versus adjacent noncancerous tissues; moreover, aspartate, β -alanine, GDP, and glycine levels were lower in patients with recurrence compared to those nonprogressing, being β -alanine the unique independent predictor of peritoneal recurrence and an independent prognostic factor for overall survival (Kaji et al., 2020).

Different studies were conducted on *HCC* tissues. An LC-MS-based metabolomics study evidenced elevated levels of glutathione, glycolysis, gluconeogenesis, and β -oxidation with reduced levels of inflammatory-related polyunsaturated fatty acids, TCA cycle and D-12 desaturase in tumor tissues (Huang et al., 2017). Another study with the same approach analyzed the homogenate of central tumor tissue, compared with adjacent tissue and distant tissue, obtained from hepatitis B virus-related HCCs, selecting 14 metabolites differentially expressed (arachidyl carnitine, betasitosterol, oleamide, quinaldic acid, tetradecanal, phenylalanine,

Table 16.3 Summary of studies related to tissue metabolomics profiling on different cancers.

Cancer	Comparison groups	Technique	References
Breast cancer	Cancer versus normal tissues Invasive ductal breast carcinoma versus benign versus peripheral normal tissues	GC-MS GC-MS and LC-MS	Fan et al. (2020) More et al. (2018)
Colorectal cancer	Cancer versus normal tissues CRC tissues versus matched adjacent mucosa; late-stage (III and IV) versus early-stage (I and II) CRC; adenocarcinoma versus nonadenocarcinoma tumors	LC-MS LC-MS	Udo et al. (2020) Long et al. (2020)
Esophageal cancer	Cancer versus normal tissues	^1H -NMR	Ye et al. (2021)
Eyelid basal cell carcinoma	Cancer versus normal tissues	LC-MS	Huang et al. (2020)
Gastric cancer	Cancer versus normal tissue; patients with versus without recurrence	Capillary electroforesis MS	Kaji et al. (2020)
Glioma	Pheochromocytoma and paraganglioma patients with versus without SDHx mutations	LC-MS	Wallace et al. (2020)
Hepatocellular carcinoma	Hepatocellular carcinoma associated with nonalcoholic fatty liver disease versus hepatocellular carcinoma associated with cirrhosis tissues Cholangiocarcinoma versus hepatocellular carcinoma tissues	^1H -NMR Capillary electroforesis MS	Teilhet et al. (2017) Murakami et al. (2015)
	Cancer versus normal tissues Central versus distal cancer tissues	LC-MS LC-MS	Huang et al. (2013) Liu et al. (2013)
	Liver cirrhosis versus hepatocellular carcinoma	LC-MS	Han et al. (2019)
Kidney cancer	Cancer versus normal tissues	^1H -NMR	Niziot, Copié et al. (2021), Niziot, Ossoliński et al. (2021)
Lung cancer	Caspase 4 negative versus positive cancer tissues Lung cancer with or without emphysema versus lung cancer with or without chronic bronchitis	GC-MS GC-MS	Terlizzi et al. (2020) Li et al. (2020)

(Continued)

Table 16.3 Summary of studies related to tissue metabolomics profiling on different cancers. *Continued*

Cancer	Comparison groups	Technique	References
Melanoma	Primary melanoma versus extratumoral microenvironment, and metastatic melanoma versus extratumoral microenvironment tissues	LC-MS	Taylor et al. (2020)
Oral squamous cell carcinoma	Cancer versus normal tissues	$^1\text{H-NMR}$	Paul et al. (2020)
Pancreatic cancer	Benign pancreatic versus cysts with possible malignant potential versus pancreatic cancer Comparison between cancer grades or progression stages Frozen versus formalin-fixed and paraffin-embedded human pancreatic adenocarcinoma tissues	LC-MS LC-MS LC-MS	Unger et al. (2018) Hiraoka et al. (2019) Feng, Yuan et al. (2019), Feng, Zhao et al. (2019)
Prostate cancer	Cancer versus normal tissues Benign prostatic hyperplasia versus early prostate cancer versus advanced prostate cancer versus metastatic prostate cancer versus castration-resistant prostate cancer Cancer versus normal tissues; high versus low Gleason score; ERG-positive versus negative prostate cancer patients Cancer versus normal tissues <8 versus ≥ 8 Gleason score	$^1\text{H-NMR}$ $^1\text{H-NMR}$ $^1\text{H-NMR}$ and LC-MS FT-IR GC-MS and LC-MS	Lima et al. (2018) Zheng et al. (2020) Dudka et al. (2020) Felgueirasa et al. (2020) Penney et al. (2021) Metere et al. (2020)
Thyroid	Cancer versus normal tissues	$^1\text{H-NMR}$	

Table shows cancer type, the compared groups, the used techniques and the related references.

glycerophosphocholine, lysophosphatidylcholines, lysophosphatidylethanolamines and chenodeoxycholic acid glycine conjugate) (Liu et al., 2013). Through capillary electrophoresis time-of-flight MS approach, an additional study compared the tissue metabolomic profiling of intrahepatic cholangiocarcinoma and HCC patients, and their corresponding nontumor adjacent mucosa, demonstrating that hypoxanthine and taurine were the only molecules able to differentiate

intrahepatic cholangiocarcinoma from HCC groups (Murakami et al., 2015). Comparative metabolomics between nonalcoholic fatty liver disease derived HCC and cirrhosis-derived HCC, performed by ^1H -NMR spectroscopy, showed:

1. higher levels of lactate and phosphocholine, and lower level of glucose in tumor versus nontumor tissues;
2. higher levels of 3-hydroxybutyrate, tyrosine, phenylalanine and histidine in HCC associated with cirrhosis patients versus higher levels of glutamine/glutamate in HCC associated with nonalcoholic fatty liver disease patients (Teilhet et al., 2017).

Moreover, a study by ultra performance LC-MS discriminated HCC from liver cirrhosis with the combination of chenodeoxycholic acid, LPC20:5, succinyladenosine and uridine (Han et al., 2019). Furthermore, a comprehensive global metabolomics analysis identified a group of differentially expressed metabolites in liver tissues of diabetes-associated HCC patients compared with those without diabetes (Xia et al., 2019).

Relative to *kidney cancer*, ^1H -NMR-based tissue metabolomics recently highlighted lower levels of fumarate, leucine, phenylalanine, sarcosine and tryptophan in the tumor tissues compared to adjacent mucosa (Nizioł, Copié et al., 2021).

In the case of *lung cancer*, GC-MS-based tissue metabolomics evidenced in caspase-4 positive lung cancer tissues, activation of fatty acid biosynthesis (higher levels of malonic and palmitic acid) compared to noncancerous tissues or caspase-4 negative tumor (Michela et al., 2020). Moreover, a recent GC-MS-based study analyzed, for the first time, tumor tissues derived from lung cancer patients with different chronic obstructive pulmonary disease (COPD) subphenotypes like emphysema and chronic bronchitis (Cheng et al., 2018).

Regarding *pancreatic cancer*, an LC-MS-based tissue metabolomic profiling identified six metabolites significantly associated with early stage tumors compared to benign pancreatic disease group, and with a linear trend from the benign to the possible malignant potential of cysts and finally to the cancer group (Unger et al., 2018). Another LC-MS-based study, on pancreatic cancer tumors or lesions, demonstrated higher levels of proline, tryptophan and tyrosine, associated with significantly shorter survival rate, cancer grade or progression (Hiraoka et al., 2019). Another interesting ultra-performance LC-MS-based metabolomics study was performed in paired frozen and FFPE pancreatic cancer tissues, and their benign counterparts, and evidenced that the levels of ten metabolites, including N-acetylaspartate and creatinine, were significantly different in both types of tumor samples, thus representing viable candidate biomarkers of pancreatic cancer (Feng et al., 2018).

In the case of *prostate cancer*, a review published in 2018 and focused on all ^1H -NMR-based tissues metabolomics studies highlighted critical tumor-associated alterations of alanine, choline, choline-related compounds, citrate, glutamate, lactate and spermine (Lima et al., 2018). In a more recent study by ^1H -NMR approach, metabolomics profile was performed on different stages, from benign prostatic hyperplasia, to early, advanced, metastatic and castration-resistant prostate cancer, demonstrating

decreased trends of citrate, creatine and creatinine levels, and increased trends of glutamate, glutamine, and formate levels, during prostate cancer progression (Adam et al., 2013; Zheng et al., 2013). An additional study, by using ¹H-NMR and LC-MS approaches, demonstrated higher levels of phosphocholine, glutamate, hypoxanthine and arginine, and lower levels of α -glucose, as able to distinguish between cancer and benign tissues and high versus low Gleason score. Moreover, increased acylcarnitines and metabolites, involved in purine catabolism, were found in ERG-positive prostate cancer patients (Dudka et al., 2020). Another metabolomics study based on FT-IR revealed that in prostate tumor tissues there was a dysregulation in lipid metabolism, a lower polysaccharide and glycogen content, and increased nucleic acid content compared to paired normal tissues (Felgueiras et al., 2020). Very recently, a combination of GC-MS and LC-MS approaches demonstrated that higher levels of ascorbate, kynurenine, NAD and oxidized glutathione correlated with high grade (≥ 8) prostate cancer (Penney et al., 2021).

¹H-NMR-based metabolomics has been also employed on *esophageal cancer* and distant noncancerous tissues, together with patient-matched serum samples, identifying differentially expressed metabolites, including valine, alanine, glucose, acetate, citrate, succinate and glutamate, in both matrices (Ye et al., 2021).

A ¹H-NMR spectroscopy study was also performed on *thyroid cancer* and healthy thyroid tissues, showing an increase of aromatic amino acids, and a decrease of citrate, due to a shift towards glycolysis, and an oxidative stress-related decrease in activity of TCA cycle in cancer tissues (Metere et al., 2020). Another report used ¹H HRMAS NMR-based metabolomics on different specimens from *oral squamous cell carcinoma* patients (tumor, bed, margin and facial muscles), evidencing that forty-eight metabolites, including lipids, were upregulated in malignant tissues (Paul et al., 2020).

LC-MS-based metabolomics, comparing *eyelid basal cell carcinoma* with blepharoplasty-derived samples, demonstrated the presence of sixteen increased and eleven decreased metabolites in cancer tissues indicating amino acid, glutathione, nicotinamide adenine dinucleotide and polyamine metabolic rewiring (Huang et al., 2017).

A metabolomics study on freshly frozen and/or FFPE samples of *phaeochromocytomas and paragangliomas* patients showed that succinate, citrate, malate and pyruvate levels had a good predictive capacity for SDHx impairment (Wallace et al., 2020). Finally, an LC-MS-based study, comparing primary metastatic *melanoma* tissues with matched adjacent extratumoral microenvironment, highlighted higher levels of dihydroxyacetone phosphate, fructose1,6-bisphosphate, N-Acetyl-DL-glutamic acid, N-Methyl-D-aspartic acid, and uracil in tumor tissues (Taylor et al., 2020).

Fecal metabolomics studies

Fecal metabolomics studies represent a good opportunity to identify new biomarkers for clinical diagnosis and prognosis because the collection of these samples

is noninvasive and does not need qualified personnel. Moreover, fecal samples are available in large amount, have strong correlation with sick organ and also correlate with gut microbiota. Short chain fatty acids (acetate, butyrate, and propionate) are the major metabolites produced by the gut microbiota and their higher concentrations in fecal samples of solid cancer patients treated with nivolumab or pembrolizumab were recently found in the responder versus nonresponder patients and associated with longer progression-free survival (Nomura et al., 2020).

Anyhow, the majority of the studies reported are on CRC patients, some applications were demonstrated also in gastric cancer, HCC, tyroid and cervical cancers (Table 16.4). Four ¹H-NMR-based metabolomics studies were conducted on fecal samples from CRC patients and demonstrated:

1. higher levels of amino acids, glucose and lactate in CRC patients compared to healthy controls (Monleon et al., 2009);
2. altered levels of glucose, lactate, glutamate, SCFAs, and succinate in early stage (stage I/II) compared to stages III and IV CRC patients (Lin et al., 2016);
3. fecal acetate having the highest diagnostic performance for discriminating CRC from healthy subjects (Lin, Ma et al., 2019);
4. higher levels of isobutyrate, isovalerate, valerate and phenylacetate, and lower levels of amino acids (glutamine, isoleucine, ornithine and taurine), bile acids (cholate, deoxycholate and lithodeoxycholamethanol), sugars (galactose, glucose, and xylose) in cancer patients compared to normal samples (Le Gall et al., 2018).

Also ultra performance LC-MS has been used to identify the significant metabolites differentially expressed in fecal samples of CRC patients. One study showed that linoleic acid and 12-hydroxy-8,10-octadecadienoic acid can be used as biomarkers for the development of ulcerative colitis into CRC (Tang et al., 2020); another analysis demonstrated that polyunsaturated fatty acids, secondary bile acids, and sphingolipids levels could represent markers for the progression from adenoma to CRC, and, hence, are early events of carcinogenesis (Kim et al., 2019).

GC/TOF MS chromatograms of fecal samples from CRC patients showed lower levels of fructose, linoleic acid, and nicotinic acid compared to samples from healthy controls (Phua et al., 2014). Another study by GC-MS approach demonstrated that amino acids, ethanolamine, furoic acid, oxalic acid, succinic acid, and some fatty acids (hexanoic acid, oleic acid, and palmitic acid) had higher levels in stool microbial extracellular vesicles from CRC patients (Yang et al., 2020).

To evaluate the influence of gut microbiome in postoperative outcomes of *gastric cancer* patients undergoing gastrectomy, both microbiome analysis and fecal metabolomics, by capillary electrophoresis time-of-flight MS, were performed on patients' compared to healthy donors samples. This study highlighted the increased levels of deoxycholic acid and twelve amino acids, including aromatic and branched-chain amino acids (tyrosine, phenylalanine, isoleucine, leucine and valine) in the gastrectomy group (Erawijantari et al., 2020).

Table 16.4 Summary of studies related to fecal metabolomics profiling on different cancers.

Cancer	Comparison groups	Technique	References
Cervical cancer	Patients with cervical cancer before versus after pelvic radiotherapy; patients after pelvic radiotherapy with versus without radiation-induced acute intestinal symptoms	$^1\text{H-NMR}$	Chai et al. (2015)
Colorectal cancer	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Monleon et al. (2009)
	Early stage (stage I/II) versus stage III and IV CRC	$^1\text{H-NMR}$	Lin et al. (2016)
	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Lin, Xu et al. (2019), Lin, Ma et al. (2019)
	Cancer patients versus healthy controls	$^1\text{H-NMR}$	Le Gall et al. (2018)
	Cancer patients versus healthy controls	GC-MS	Phua et al. (2014)
	Extracellular vesicles cancer patients versus healthy controls	GC-MS	Kim et al. (2020)
	Ulcerative colitis versus CRC patients	LC-MS	Tang et al. (2020)
Gastric cancer	Precancerous lesions versus CRC patients	LC-MS	Kim et al. (2020)
Gastric cancer	Gastric cancer patients subjected to gastrectomy versus healthy controls	capillary electroforesis MS	Erawijantari et al. (2020)
Hepatocellular carcinoma	Patients with nonalcoholic fatty liver disease (NAFLD) related cirrhosis with versus without HCC	$^1\text{H-NMR}$ and LC-MS	Behary et al. (2021)
	Liver cirrhosis versus HCC patients	LC-MS	Cao et al. (2011)
Thyroid	Cancer patients versus healthy controls	LC-MS	Feng, Yuan et al. (2019), Feng, Zhao et al. (2019)

Table shows cancer type, the compared groups, the used techniques and the related references.

Fecal metabolomics approach by ultra-performance LC-MS was also performed on liver cirrhosis and *HCC* patients compared to healthy controls, demonstrating higher levels of lysophosphatidylcholines and lower of bile acids and bile pigments in liver cirrhosis and *HCC* patients, suggesting that these metabolites can be used to improve the early diagnosis of both cirrhosis and *HCC* ([Cao et al., 2011](#)).

More recently, a fecal metagenomic and metabolomic study, to characterize gut microbiota and its effect on peripheral immune response, was conducted on

samples from patients with nonalcoholic fatty liver disease (NAFLD) related cirrhosis, with or without HCC, compared to non-NAFLD control subjects. Gut microbiota altered composition and functional shift in HCC development was confirmed by fecal metabolomics, indicating increased production of short chain fatty acids, acetate, butyrate and formate in NAFLD-HCC samples. Notably, bacterial extracts from NAFLD-HCC samples induced immune suppression, suggesting an impact of cancer-related fecal metabolomics alterations on immune system (Behary et al., 2021).

Ultra-performance LC-MS approach was used to characterize the fecal metabolomic profile in *thyroid carcinoma* patients compared to healthy donors, demonstrating increased levels of histamine and stearic acid, and decreased of bile acid, decanoic acid and hexanoic acid. Overall choline metabolism, cell growth and death, lipid metabolism and immune system were among the more altered pathways (Feng, Zhao et al., 2019).

Considering that radiation-induced acute intestinal symptoms (RIAISs) can be a common radiotherapy complication in *cervical cancer*, ¹H-NMR-based metabolomics profiling was applied on fecal samples from patients with cervical cancer before and after pelvic radiotherapy to identify markers associated with RIAIS. The analysis evidenced that RIAIS patients after pelvic radiotherapy had lower levels of glucose, ethanol, methylamine and butyrate, and higher levels of α-ketobutyrate, alanine, aminohippurate, bile acids, creatine, creatinine, glutamine, isoleucine, lysine, phenylalanine, trimethylamine N-oxide, tyrosine, uracil and valine, compared to baseline samples treatment (Chai et al., 2015).

Saliva metabolomics studies

Saliva is enriched of secretions derived mainly from parotid, sublingual and submandibular glands and in minor amount from buccal, labial and lingual glands. It also contains bacteria, blood, cell debris, dental plaque, gingival crevicular fluid, and bronchial and nasal secretions. Notably saliva contains inorganic substances (0.2%), proteins (0.3%) and water (99.5%), but most also over 800 identified metabolites, mirroring the health status of the body. Therefore some metabolomics studies have been performed on saliva samples from different cancer patients (Table 16.5).

¹H-NMR spectroscopy has been applied to detect the salivary metabolic changes associated with *head and neck squamous cell carcinoma*, including laryngeal and oral cancers, demonstrating decreased levels of proline and increased levels of 1,2 propanediol and fucose, compared to healthy controls (Mikkonen et al., 2018). Capillary electrophoresis time-of-flight MS was used on saliva samples from *oral cancer* patients, compared to healthy controls, demonstrating forty-three metabolites significantly differentially expressed, seventeen with the same trend in oral cancer tissues, and selecting the combination of S-adenosylmethionine and pipecolate as

Table 16.5 Summary of studies related to saliva metabolomics profiling on different cancers.

Cancer	Comparison groups	Technique	References
Breast cancer	Cancer patients versus healthy controls	LC-MS	Zhong et al. (2016)
	Cancer patients versus healthy controls	LC-MS	Xavier Assad et al. (2020)
	Cancer patients versus healthy controls	SERS	Ozturk et al. (2011)
Gastric cancer	Gastric cancer patients versus healthy controls	SERS	Chen et al. (2016)
Head and neck cancer	Larynx and oral cavity cancer patients versus healthy controls	¹ H-NMR	Mikkonen et al. (2018)
	Oral cavity and oropharyngeal squamous cell carcinoma patients versus healthy controls	¹ H-NMR and LC-MS	Lohavanichbutr et al. (2018)
	Oral cancer patients versus healthy controls	Capillary electrophoresis time-of-flight MS	Ishikawa et al. (2016)
Head and neck cancer	Premalignant lesion stage versus oral squamous cell carcinoma patients versus healthy controls	Conductive polymer spray ionization MS	Song et al. (2020)
	Oral cavity and oropharyngeal cancer at baseline versus 5–6 months postchemoradiation	GC-MS	Lim et al. (2021)
	Oral squamous cell carcinoma patients versus healthy controls	LC-MS	Sridharan et al. (2019)
	Patients with oral squamous cell carcinoma versus oral submucous fibrosis versus healthy controls	Raman spectroscopy	Rekha et al. (2016)
	Nasopharyngeal carcinoma patients versus healthy controls	SERS	Lin et al. (2017)
Lung cancer	Cancer patients versus healthy controls	SERS	Qian et al. (2018)
Pancreatic cancer	Cancer patients versus healthy controls	capillary electrophoresis time-of-flight MS	Sugimoto et al. (2010)
Parotid tumor	Cancer patients versus healthy controls	¹ H-NMR	Grimaldi et al. (2018)

Table shows cancer type, the compared groups, the used techniques and the related references.

biomarkers to discriminate patients versus controls (Ishikawa et al., 2016). An additional study on the saliva of oral squamous cell carcinoma patients compared with normal controls used Q-TOF-LC-MS, evidencing higher levels of 1-methylhistidine, 2-oxoarginine, 4-nitroquinoline-1-oxide, d-glycerate-2-phosphate, inositol 1,3,4-triphosphate, norcocainenitroxide, pseudouridine and sphinganine-1-phosphate, and

lower levels of l-homocysteic acid, estradiol, valerate, neuraminic acid and ubiquinone (Sridharan et al., 2019). Using a combined approach based on different analytical platforms (LC-MS, LC-Q-TOF and NMR) salivary metabolite profiling including citrulline, glycine, ornithine and proline discriminated early stage oral cavity squamous cell carcinoma from healthy controls, representing a promising diagnostic tool (Lohavanichbutr et al., 2018). More recently, to characterize the global salivary metabolic profiles of patients with premalignant lesion and with oral squamous cell carcinoma, compared to healthy controls, Lim et al. (2021) applied conductive polymer spray ionization mass spectrometry and demonstrated that higher levels of 5-aminopentoate, adenosine, cadaverine, putrescine and thymidine, and lower levels of adrenic acid, glucose, hippuric acid, phosphocholine and serine were all markers of malignant progression from healthy controls to oral squamous cell carcinoma (Song et al., 2020). Moreover, GC–MS-based metabolomics approach, performed on saliva samples from patients with oral cavity and oropharyngeal cancer, at baseline and 5–6 months postchemoradiation, evidenced increased levels of glutamine, intermediates of pyruvate metabolism, mannitol, maltose and sorbose, and decreased levels of ethanolamine and sphingosine after treatment, associated with oral dysbiosis and oral microbiota changes that can affect the patients post-treatment quality of life (Lim et al., 2021).

¹H-NMR-based metabolomics study on saliva samples of *parotid tumor* patients enabled the identification of two dysregulated metabolic pathways (glycogenic aminoacids and ketone bodies), and of alanine and leucine as the most significant metabolites able to discriminate these patients from healthy controls, representing useful tool to improve both diagnosis and followup strategies (Grimaldi et al., 2018).

Through LC-MS-based metabolomics applied on saliva samples from healthy controls and *breast cancer* patients, Zhong et al. (2016) selected eighteen metabolites, useful to improve breast cancer diagnosis and mainly involved in phospholipid catabolism, fatty acid and sphingolipid metabolisms. Xavier Assad et al. (Xavier Assad et al., 2020) identified seven oligopeptides and six glycerophospholipids upregulated in the breast cancer group. Using capillary electrophoresis time-of-flight MS on saliva samples from pancreatic cancer patients compared to healthy controls, Sugimoto et al. (2010) evidenced that some metabolites, including aspartic acid, glutamate, glutamine, isoleucine, leucine, phenylalanine, tryptophan and valine, might represent *pancreatic cancer*-specific markers.

A recent review article described the application of Raman, specifically surface-enhanced Raman (SERS) and vibrational spectroscopy for cancer biomarker discovery by metabolic salivary fingerprint (Derrauau et al., 2020). In *oral cancer*, significant differences, in comparison with healthy controls and oral submucous fibrosis, were obtained by Raman peaks corresponding to amino acids (histidine, proline, valine), lactate, lipids and nucleic acid (Rekha et al., 2016). Specific peaks corresponding to phenylalanine, proline, valine, proteins, and collagens by SERS have been identified as differentially expressed in *nasopharyngeal carcinoma* patients (Lin et al., 2019). In the case of *lung cancer* patients

twelve peaks by SERS varied significantly and were attributed mainly to change in protein residues and in the content of nucleic acid molecules (Qian et al., 2020). More concentrated amino acids (alanine, ethanolamine, histidine, hydroxylysine, glycine, glutamate, glutamine, proline, taurine, and tyrosine) were found by SERS in saliva of *gastric cancer* patients compared with control samples (Chen et al., 2016). Finally, Ozturk et al. (2011) demonstrated by SERS that sialic acid was significantly increased in *breast cancer* patients.

Metabolomics studies on other biological matrices

Metabolomics profiling has been evaluated also on other biological matrices such as bronchoalveolar lavage, exhaled breath condensate, sweat, seminal plasma, nail, and cerebrospinal fluid from cancer patients.

For example GC-MS approach has been used to evaluate metabolomics profiling on *bronchoalveolar lavage* fluid, serum and urine, from lung cancer patients compared with healthy control and noncancerous lung lesions (Callejón-Leblie et al., 2019). Also *exhaled breath condensate* is another biological biofluid whose sampling is noninvasive and whose metabolomic composition was analyzed in lung cancer patients (Peralbo-Molina et al., 2016b). A study applying GC-MS approach on exhaled breath condensate from healthy controls, lung cancer patients, or risk factor individuals including active smokers and ex-smokers, identified two metabolites (monoacylglycerols and squalene) able to discriminate between the three groups (Peralbo-Molina et al., 2016a).

Also human *sweat* was used as clinical samples to develop a screening tool for lung cancer by demonstrating that the combination of monoglyceride, nonane-dioic acid, suberic acid and trihexose discriminates between cancer patients versus smokers (Calderón-Santiago et al., 2015).

Seminal plasma metabolites (choline, citrate, glucose and phosphocholine) and lipoproteins, identified by ¹H-NMR, resulted capable to predict low- and intermediate-risk prostate cancer versus benign samples (Roberts et al., 2017).

The analysis of *nail* metabolites in breast cancer patients demonstrated lower levels of free aromatic amino acids by LC-HRMS approach compared with healthy donors (Mitruka et al., 2020).

Finally, *cerebrospinal fluid* metabolites (diacetylspermine and fibrinogen fragments), identified by MS approach, discriminated patients with leptomeningeal carcinomatosis from those at high risk of leptomeningeal metastasis (Yoo et al., 2017). The analysis of metabolomic profile from cerebrospinal fluids of glioma patients compared to healthy controls, by MS approach, evidenced that low levels of metabolites involved in tryptophan metabolism were indicative of the absence of an inflammatory signature (Locasale et al., 2012). Metabolomics evaluation on cerebrospinal fluids by LC-MS and GC-MS approaches from patients with central nervous system lymphoma showed the presence of thirty-six significantly

upregulated metabolites involved in energy metabolism pathways, purine metabolism and amino acid metabolism, compared to noncancer patients. In particular, higher levels of lactate, 5-methylthioadenosine, and kynurenine have been found in cerebrospinal fluids collected from primary central nervous system lymphoma patients at diagnosis compared with those in controls, and in patients after lenalidomide treatment (Rubenstein et al., 2018).

Conclusion

Overall, aberrant metabolism is an emerging hallmark of cancer. Indeed, tumor metabolic changes can be used for diagnosis and to stratify cancer patients for prognosis. Notably, therapies directed at reprogramming cancer metabolism have been developed. However, although several metabolomics signatures have been proposed as novel potential biomarkers, robust and well-validated metabolic analytes are far to be implemented in clinical practice. Indeed, sensitivity and specificity vary among different technologies and the studies described above are consequently scarcely homogenous. Similarly, the enrichment of detected metabolites also varies depending on different matrices tested. Indeed, standardization of different methodologies is particularly challenging. Furthermore, compared with other omics approaches, only few meta-analyses have been performed on metabolomics studies (Lee, Mahmud et al., 2020).

References

- Adam, J., Yang, M., Bauerschmidt, C., Kitagawa, M., O'Flaherty, L., Maheswaran, P., Özkan, G., Sahgal, N., Baban, D., & Kato, K. (2013). A role for cytosolic fumarate hydratase in urea cycle metabolism and renal neoplasia. *Cell Reports*, 3(5), 1440–1448.
- Alsaleh, M., Barbera, T. A., Reeves, H. L., Cramp, M. E., Ryder, S., Gabra, H., Nash, K., Shen, Y.-L., Holmes, E., & Williams, R. (2019). Characterization of the urinary metabolic profile of cholangiocarcinoma in a United Kingdom population. *Hepatic Medicine: Evidence and Research*, 11, 47.
- Alsaleh, M., Sithithaworn, P., Khuntikeo, N., Loilome, W., Yongvanit, P., Chamadol, N., Hughes, T., O'Connor, T., Andrews, R. H., & Holmes, E. (2019). Characterisation of the urinary metabolic profile of liver fluke-associated cholangiocarcinoma. *Journal of Clinical and Experimental Hepatology*, 9(6), 657–675.
- Amara, C. S., Ambati, C. R., Vantaku, V., Piyarathna, D. W. B., Donepudi, S. R., Ravi, S. S., Arnold, J. M., Putluri, V., Chatta, G., & Guru, K. A. (2019). Serum metabolic profiling identified a distinct metabolic signature in bladder cancer smokers: A key metabolic enzyme associated with patient survival. *Cancer Epidemiology and Prevention Biomarkers*, 28(4), 770–781.

- Amir Hashim, N. A., Ab-Rahim, S., Suddin, L. S., Ahmad Saman, M. S., & Mazlan, M. (2019). Global serum metabolomics profiling of colorectal cancer. *Molecular and Clinical Oncology*, 11(1), 3–14.
- Aredo, J. V., Purington, N., Su, L., Luo, S. J., Diao, N., Christiani, D. C., Wakelee, H. A., & Han, S. S. (2021). Metabolomic profiling for second primary lung cancer: A pilot case-control study. *Lung Cancer (Amsterdam, Netherlands)*, 155, 61–67.
- Bai, J., Gao, Y., Chen, L., Yin, Q., Lou, F., Wang, Z., Xu, Z., Zhou, H., Li, Q., & Cai, W. (2019). Identification of a natural inhibitor of methionine adenosyltransferase 2A regulating one-carbon metabolism in keratinocytes. *EBioMedicine*, 39, 575–590.
- Bansal, N., Gupta, A., Mitash, N., Shakya, P. S., Mandhani, A., Mahdi, A. A., Sankhwar, S. N., & Mandal, S. K. (2013). Low- and high-grade bladder cancer determination via human serum-based metabolomics approach. *Journal of Proteome Research*, 12(12), 5839–5850.
- Bathe, O. F., Shaykhutdinov, R., Kopciuk, K., Weljie, A. M., McKay, A., Sutherland, F. R., Dixon, E., Dunse, N., Sotiropoulos, D., & Vogel, H. J. (2011). Feasibility of identifying pancreatic cancer based on serum metabolomics. *Cancer Epidemiology and Prevention Biomarkers*, 20(1), 140–147.
- Behary, J., Amorim, N., Jiang, X.-T., Raposo, A., Gong, L., McGovern, E., Ibrahim, R., Chu, F., Stephens, C., & Jebeili, H. (2021). Gut microbiota impact on the peripheral immune response in non-alcoholic fatty liver disease related hepatocellular carcinoma. *Nature Communications*, 12(1), 1–14.
- Berker, Y., Vandergrift, L. A., Wagner, I., Su, L., Kurth, J., Schuler, A., Dinges, S. S., Habbel, P., Nowak, J., & Mark, E. (2019). Magnetic resonance spectroscopy-based metabolomic biomarkers for typing, staging, and survival estimation of early-stage human lung cancer. *Scientific Reports*, 9(1), 1–12.
- Bertini, I., Cacciatore, S., Jensen, B. V., Schou, J. V., Johansen, J. S., Kruhøffer, M., Luchinat, C., Nielsen, D. L., & Turano, P. (2012). Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. *Cancer Research*, 72(1), 356–364.
- Calderón-Santiago, M., Priego-Capote, F., Turck, N., Robin, X., Jurado-Gámez, B., Sanchez, J. C., & De Castro, M. D. L. (2015). Human sweat metabolomics for lung cancer screening. *Analytical and Bioanalytical Chemistry*, 407(18), 5381–5392.
- Callejón-Leblie, B., García-Barrera, T., Pereira-Vega, A., & Gómez-Ariza, J. L. (2019). Metabolomic study of serum, urine and bronchoalveolar lavage fluid based on gas chromatography mass spectrometry to delve into the pathology of lung cancer. *Journal of Pharmaceutical and Biomedical Analysis*, 163, 122–129.
- Camarda, R., Zhou, A. Y., Kohnz, R. A., Balakrishnan, S., Mahieu, C., Anderton, B., Eyob, H., Kajimura, S., Tward, A., & Krings, G. (2016). Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. *Nature Medicine*, 22(4), 427–432.
- Cao, H., Huang, H., Xu, W., Chen, D., Yu, J., Li, J., & Li, L. (2011). Fecal metabolome profiling of liver cirrhosis and hepatocellular carcinoma patients by ultra performance liquid chromatography–mass spectrometry. *Analytica Chimica Acta*, 691(1–2), 68–75.
- Casadei-Gardini, A., Del Coco, L., Marisi, G., Conti, F., Rovesti, G., Ulivi, P., Canale, M., Frassineti, G. L., Foschi, F. G., & Longo, S. (2020). 1H-NMR based serum metabolomics highlights different specific biomarkers between early and advanced hepatocellular carcinoma stages. *Cancers*, 12(1), 241.

- Chai, Y., Wang, J., Wang, T., Yang, Y., Su, J., Shi, F., Wang, J., Zhou, X., He, B., & Ma, H. (2015). Application of ^1H NMR spectroscopy-based metabolomics to feces of cervical cancer patients with radiation-induced acute intestinal symptoms. *Radiotherapy and Oncology*, 117(2), 294–301.
- Chan, A. W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D. T., Sawyer, M. B., & Broadhurst, D. (2016). ^1H -NMR urinary metabolomic profiling for diagnosis of gastric cancer. *British Journal of Cancer*, 114(1), 59–62.
- Chen, J., Hou, H., Chen, H., Luo, Y., Zhang, L., Zhang, Y., Liu, H., Zhang, F., Liu, Y., & Wang, A. (2019). Urinary metabolomics for discovering metabolic biomarkers of laryngeal cancer using UPLC-QTOF/MS. *Journal of Pharmaceutical and Biomedical Analysis*, 167, 83–89.
- Chen, Y., Zhang, J., Guo, L., Liu, L., Wen, J., Xu, L., Yan, M., Li, Z., Zhang, X., & Nan, P. (2016). A characteristic biosignature for discrimination of gastric cancer from healthy population by high throughput GC-MS analysis. *Oncotarget*, 7(52), 87496.
- Cheng, X., Liu, X., Liu, X., Guo, Z., Sun, H., Zhang, M., Ji, Z., & Sun, W. (2018). Metabolomics of non-muscle invasive bladder cancer: Biomarkers for early detection of bladder cancer. *Frontiers in Oncology*, 8, 494.
- Chiu, M., Tardito, S., Pillozzi, S., Arcangeli, A., Armento, A., Uggeri, J., Missale, G., Bianchi, M., Barilli, A., & Dall'Asta, V. (2014). Glutamine depletion by crisantaspase hinders the growth of human hepatocellular carcinoma xenografts. *British Journal of Cancer*, 111(6), 1159–1167.
- Costantini, S., Sorice, A., Capone, F., Madonna, G., Mallardo, D., Capone, M., Ciliberto, G., Budillon, A., & Ascierto, P. (2018). Outcome prediction on melanoma patients subjected to immunotherapy treatments by ^1H -NMR metabolomic profiling approach. *Journal of Translational Medicine* 16(1), 4.
- D'aniello, C., Patriarca, E. J., Phang, J. M., & Minchietti, G. (2020). Proline metabolism in tumor growth and metastatic progression. *Frontiers in Oncology*, 10, 776.
- Dal Bello, B., Rosa, L., Campanini, N., Tinelli, C., Viera, F. T., D'Ambrosio, G., Rossi, S., & Silini, E. M. (2010). Glutamine synthetase immunostaining correlates with pathologic features of hepatocellular carcinoma and better survival after radiofrequency thermal ablation. *Clinical Cancer Research*, 16(7), 2157–2166.
- Dator, R., Villalta, P. W., Thomson, N., Jensen, J., Hatsukami, D. K., Stepanov, I., Warth, B., & Balbo, S. (2020). Metabolomics profiles of smokers from two ethnic groups with differing lung cancer risk. *Chemical Research in Toxicology*, 33(8), 2087–2098.
- Davis, V. W., Schiller, D. E., Eurich, D., Bathe, O. F., & Sawyer, M. B. (2013). Pancreatic ductal adenocarcinoma is associated with a distinct urinary metabolomic signature. *Annals of Surgical Oncology*, 20(3), 415–423.
- DeBerardinis, R. J., Sayed, N., Ditsworth, D., & Thompson, C. B. (2008). Brick by brick: Metabolism and tumor cell growth. *Current Opinion in Genetics & Development*, 18 (1), 54–61.
- Deng, L., Ismond, K., Liu, Z., Constable, J., Wang, H., Alatise, O. I., Weiser, M. R., Kingham, T. P., & Chang, D. (2019). Urinary metabolomics to identify a unique biomarker panel for detecting colorectal cancer: A multicenter study. *Cancer Epidemiology and Prevention Biomarkers*, 28(8), 1283–1291.
- Derrauau, S., Robinet, J., Untereiner, V., Piot, O., Sockalingum, G. D., & Lorimier, S. (2020). Vibrational spectroscopy saliva profiling as biometric tool for disease diagnostics: A systematic literature. *Molecules (Basel, Switzerland)*, 25(18), 4142.

- Di Bello, E., Zwergel, C., Mai, A., & Valente, S. (2020). The innovative potential of statins in cancer: New targets for new therapies. *Frontiers in Chemistry*, 39(1), 213.
- Di Gregorio, E., Miolo, G., Saorin, A., Muraro, E., Cangemi, M., Revelant, A., Minatel, E., Trovò, M., Steffan, A., & Corona, G. (2021). Radical hemithoracic radiotherapy induces systemic metabolomics changes that are associated with the clinical outcome of malignant pleural mesothelioma patients. *Cancers*, 13(3), 508.
- Du, Y., Fan, P., Zou, L., Jiang, Y., Gu, X., Yu, J., & Zhang, C. (2021). Serum metabolomics study of papillary thyroid carcinoma based on HPLC-Q-TOF-MS/MS. *Frontiers in Cell and Developmental Biology*, 9, 593510.
- Dudka, I., Thysell, E., Lundquist, K., Antti, H., Iglesias-Gato, D., Flores-Morales, A., Bergh, A., Wikström, P., & Gröbner, G. (2020). Comprehensive metabolomics analysis of prostate cancer tissue in relation to tumor aggressiveness and TMPRSS2-ERG fusion status. *BMC Cancer*, 20, 1–17.
- Eckert, M. A., Coscia, F., Chryplewicz, A., Chang, J. W., Hernandez, K. M., Pan, S., Tienda, S. M., Nahotko, D. A., Li, G., & Blaženović, I. (2019). Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts. *Nature*, 569 (7758), 723–728.
- Elia, I., Broekaert, D., Christen, S., Boon, R., Radaelli, E., Orth, M. F., Verfaillie, C., Grünewald, T. G., & Fendt, S.-M. (2017). Proline metabolism supports metastasis formation and could be inhibited to selectively target metastasizing cancer cells. *Nature Communications*, 8(1), 1–11.
- Erawijantari, P. P., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Saito, Y., Fukuda, S., Yachida, S., & Yamada, T. (2020). Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut*, 69(8), 1404–1415.
- Erlic, Z., Kurlbaum, M., Deutschbein, T., Nölting, S., Prejbisz, A., Timmers, H., Richter, S., Prehn, C., Weismann, D., & Adamski, J. (2019). Metabolic impact of pheochromocytoma/paraganglioma: Targeted metabolomics in patients before and after tumor removal. *European Journal of Endocrinology*, 181(6), 647–657.
- Falegan, O. S., Arnold Egloff, S. A., Zijlstra, A., Hyndman, M. E., & Vogel, H. J. (2019). Urinary metabolomics validates metabolic differentiation between renal cell carcinoma stages and reveals a unique metabolic profile for oncocytes. *Metabolites*, 9(8), 155.
- Fan, S., Shahid, M., Jin, P., Asher, A., & Kim, J. (2020). Identification of metabolic alterations in breast cancer using mass spectrometry-based metabolomic analysis. *Metabolites*, 10(4), 170.
- Fang, H., Du, G., Wu, Q., Liu, R., Chen, C., & Feng, J. (2019). HDAC inhibitors induce proline dehydrogenase (POX) transcription and anti-apoptotic autophagy in triple negative breast cancer. *Acta Biochimica et Biophysica Sinica*, 51(10), 1064–1070.
- Fantin, V. R., St-Pierre, J., & Leder, P. (2006). Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 9(6), 425–434.
- Farshidfar, F., Weljie, A. M., Kopciuk, K., Buie, W. D., MacLean, A., Dixon, E., Sutherland, F. R., Molckovsky, A., Vogel, H. J., & Bathe, O. F. (2012). Serum metabolomic profile as a means to distinguish stage of colorectal cancer. *Genome Medicine*, 4 (5), 1–13.
- Felgueirasa, J., Vieira Silva, J., Nunesa, A., Fernandes, I., Patrícioe, A., Maiae, N., Pelech, S., & Fardilhā, M. (2020). Investigation of spectroscopic and proteomic alterations underlying prostate. *Carcinogenesis Journal of Proteomics*, 226, 103888.

- Feng, D., Yuan, J., Liu, Q., Liu, L., Zhang, X., Wu, Y., Qian, Y., Chen, L., Shi, Y., & Gu, M. (2019). UPLC-MS/MS-based metabolomic characterization and comparison of pancreatic adenocarcinoma tissues using formalin-fixed, paraffin-embedded and optimal cutting temperature-embedded materials. *International Journal of Oncology*, 55(6), 1249–1260.
- Feng, J., Zhao, F., Sun, J., Lin, B., Zhao, L., Liu, Y., Jin, Y., Li, S., Li, A., & Wei, Y. (2019). Alterations in the gut microbiota and metabolite profiles of thyroid carcinoma patients. *International Journal of Cancer*, 144(11), 2728–2745.
- Feng, Y., Xiong, Y., Qiao, T., Li, X., Jia, L., & Han, Y. (2018). Lactate dehydrogenase A: A key player in carcinogenesis and potential target in cancer therapy. *Cancer Medicine*, 7(12), 6124–6136.
- Gatenby, R. A., & Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? *Nature Reviews. Cancer*, 4(11), 891–899.
- Giskeodegard, G. F., Hansen, A. F., Bertilsson, H., Gonzalez, S. V., Kristiansen, K. A., Bruheim, P., Mjos, S. A., Angelsen, A., Bathen, T. F., & Tessem, M. B. (2015). Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. *British Journal of Cancer*, 113, 1712–1719.
- Goodwin, M. L., Gladden, L. B., Nijsten, M. W., & Jones, K. B. (2015). Lactate and cancer: Revisiting the Warburg effect in an era of lactate shuttling. *Frontiers in Nutrition*, 1, 27.
- Grimaldi, M., Palisi, A., Rossi, G., Stillitano, I., Faiella, F., Montoro, P., Rodriguez, M., Palladino, R., D'Ursi, A. M., & Romano, R. (2018). Saliva of patients affected by salivary gland tumour: An NMR metabolomics analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 160, 436–442.
- Grønningseter, I. S., Fredly, H. K., Gjertsen, B. T., Hatfield, K. J., & Bruserud, Ø. (2019). Systemic metabolomic profiling of acute myeloid leukemia patients before and during disease-stabilizing treatment based on all-trans retinoic acid, valproic acid, and low-dose chemotherapy. *Cells*, 8(10), 1229.
- Gu, J., Xiao, Y., Shu, D., Liang, X., Hu, X., Xie, Y., Lin, D., & Li, H. (2019). Metabolomics analysis in serum from patients with colorectal polyp and colorectal cancer by 1H-NMR spectrometry. *Disease Markers*, 2019, 3491852.
- Gupta, A., Bansal, N., Mitash, N., Kumar, D., Kumar, M., Sankhwar, S. N., Mandhani, A., & Singh, U. P. (2020). NMR-derived targeted serum metabolic biomarkers appraisal of bladder cancer: A pre-and post-operative evaluation. *Journal of Pharmaceutical and Biomedical Analysis*, 183, 113134.
- Hadi, N. I., Jamal, Q., Iqbal, A., Shaikh, F., Somroo, S., & Musharraf, S. G. (2017). Serum metabolomic profiles for breast cancer diagnosis, grading and staging by gas chromatography-mass spectrometry. *Scientific Reports*, 7(1), 1–11.
- Han, J., Qin, W., Li, Z., Xu, A., Xing, H., Wu, H., Zhang, H., Li, C., Liang, L., & Quan, B. (2019). Tissue and serum metabolite profiling reveals potential biomarkers of human hepatocellular carcinoma. *Clinica Chimica Acta*, 488, 68–75.
- Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1), 57–70.
- He, X., Zhong, J., Wang, S., Zhou, Y., Wang, L., Zhang, Y., & Yuan, Y. (2017). Serum metabolomics differentiating pancreatic cancer from new-onset diabetes. *Oncotarget*, 8 (17), 29116.
- He, Y., Zheng, Z., Xu, Y., Weng, H., Gao, Y., Qin, K., Rong, J., Chen, C., Yun, M., & Zhang, J. (2018). Over-expression of IMPDH2 is associated with tumor progression

- and poor prognosis in hepatocellular carcinoma. *American Journal of Cancer Research*, 8(8), 1604.
- Hiraoka, N., Toue, S., Okamoto, C., Kikuchi, S., Ino, Y., Yamazaki-Itoh, R., Esaki, M., Nara, S., Kishi, Y., & Imaizumi, A. (2019). Tissue amino acid profiles are characteristic of tumor type, malignant phenotype, and tumor progression in pancreatic tumors. *Scientific Reports*, 9(1), 1–14.
- Huang, J., Schaefer, J., Wang, Y., Gioia, L., Pei, Y., Shi, X., Waris, S., Zhao, C., Nguyen, J., & Du, J. (2020). Metabolic signature of eyelid basal cell carcinoma. *Experimental Eye Research*, 198, 108140.
- Huang, J., Weinstein, S. J., Kitahara, C. M., Karoly, E. D., Sampson, J. N., & Albanes, D. (2017). A prospective study of serum metabolites and glioma risk. *Oncotarget*, 8(41), 70366.
- Huang, Q., Tan, Y., Yin, P., Ye, G., Gao, P., Lu, X., Wang, H., & Xu, G. (2013). Metabolic characterization of hepatocellular carcinoma using nontargeted tissue metabolomics. *Cancer Research*, 73(16), 4992–5002.
- Iannelli, F., Roca, M. S., Lombardi, R., Ciardiello, C., Grumetti, L., De Rienzo, S., Moccia, T., Vitagliano, C., Sorice, A., & Costantini, S. (2020). Synergistic antitumor interaction of valproic acid and simvastatin sensitizes prostate cancer to docetaxel by targeting CSCs compartment via YAP inhibition. *Journal of Experimental & Clinical Cancer Research*, 39(1), 1–24.
- Ishikawa, S., Sugimoto, M., Kitabatake, K., Sugano, A., Nakamura, M., Kaneko, M., Ota, S., Hiwatori, K., Enomoto, A., & Soga, T. (2016). Identification of salivary metabolomic biomarkers for oral cancer screening. *Scientific Reports*, 6(1), 1–7.
- Itoi, T., Sugimoto, M., Umeda, J., Sofuni, A., Tsuchiya, T., Tsuji, S., Tanaka, R., Tonozuka, R., Honjo, M., & Moriyasu, F. (2017). Serum metabolomic profiles for human pancreatic cancer discrimination. *International Journal of Molecular Sciences*, 18(4), 767.
- Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A. L., Kafri, R., Kirschner, M. W., Clish, C. B., & Mootha, V. K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science (New York, N.Y.)*, 336 (6084), 1040–1044.
- Jelonek, K., Krzywon, A., Jablonska, P., Slominska, E. M., Smolenski, R. T., Polanska, J., Rutkowski, T., Mrochem-Kwarcia, J., Skladowski, K., & Widlak, P. (2020). Systemic effects of radiotherapy and concurrent chemo-radiotherapy in head and neck cancer patients—Comparison of serum metabolome profiles. *Metabolites*, 10(2), 60.
- Jia, H., Shen, X., Guan, Y., Xu, M., Tu, J., Mo, M., Xie, L., Yuan, J., Zhang, Z., & Cai, S. (2018). Predicting the pathological response to neoadjuvant chemoradiation using untargeted metabolomics in locally advanced rectal cancer. *Radiotherapy and Oncology*, 128(3), 548–556.
- Jia, X., Zhang, S., Zhu, H., Wang, W., Zhu, J., Wang, X., & Qiang, J. (2016). Increased expression of PHGDH and prognostic significance in colorectal cancer. *Translational Oncology*, 9(3), 191–196.
- Jiang, Y., Cheng, X., Wang, C., & Ma, Y. (2010). Quantitative determination of sarcosine and related compounds in urinary samples by liquid chromatography with tandem mass spectrometry. *Analytical Chemistry*, 82(21), 9022–9027.
- Kaji, S., Irino, T., Kusuvara, M., Makuuchi, R., Yamakawa, Y., Tokunaga, M., Tanizawa, Y., Bando, E., Kawamura, T., & Kami, K. (2020). Metabolomic profiling of gastric

- cancer tissues identified potential biomarkers for predicting peritoneal recurrence. *Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, 23(5), 874–883.
- Kalyanaraman, B. (2017). Teaching the basics of cancer metabolism: Developing antitumor strategies by exploiting the differences between normal and cancer cell metabolism. *Redox Biology*, 12, 833–842.
- Kaushik, A. K., & DeBerardinis, R. J. (2018). Applications of metabolomics to study cancer metabolism. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1870(1), 2–14.
- Kim, E. R., Kwon, H. N., Nam, H., Kim, J. J., Park, S., & Kim, Y.-H. (2019). Urine-NMR metabolomics for screening of advanced colorectal adenoma and early stage colorectal cancer. *Scientific Reports*, 9(1), 1–10.
- Kim, M., Vogtmann, E., Ahlquist, D. A., Devens, M. E., Kisiel, J. B., Taylor, W. R., White, B. A., Hale, V. L., Sung, J., & Chia, N. (2020). Fecal metabolomic signatures in colorectal adenoma patients are associated with gut microbiota and early events of colorectal cancer pathogenesis. *MBio*, 11(1).
- Koutros, S., Meyer, T. E., Fox, S. D., Issaq, H. J., Veenstra, T. D., Huang, W.-Y., Yu, K., Albanes, D., Chu, L. W., & Andriole, G. (2013). Prospective evaluation of serum sarcosine and risk of prostate cancer in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. *Carcinogenesis*, 34(10), 2281–2285.
- Kozar, N., Kruusmaa, K., Bitenc, M., Argamasilla, R., Adsuar, A., Takač, I., & Arko, D. (2020). Identification of novel diagnostic biomarkers in breast cancer using targeted metabolomic profiling. *Clinical Breast Cancer*, Nov 11.
- Kozar, N., Kruusmaa, K., Dovnik, A., Bitenc, M., Argamasilla, R., Adsuar, A., Goswami, N., Takač, I., & Arko, D. (2021). Identification of novel diagnostic biomarkers in endometrial cancer using targeted metabolomic profiling. *Advances in Medical Sciences*, 66 (1), 46–51.
- Kuhajda, F. P., Jenner, K., Wood, F. D., Hennigar, R. A., Jacobs, L. B., Dick, J. D., & Pasternack, G. R. (1994). Fatty acid synthesis: A potential selective target for antineoplastic therapy. *Proceedings of the National Academy of Sciences*, 91(14), 6379–6383.
- Kumar, D., Gupta, A., Mandhani, A., & Sankhwar, S. N. (2015). Metabolomics-derived prostate cancer biomarkers: Fact or fiction? *Journal of Proteome Research*, 14(3), 1455–1464.
- Labuschagne, C. F., Van Den Broek, N. J., Mackay, G. M., Vousden, K. H., & Maddocks, O. D. (2014). Serine, but not glycine, supports one-carbon metabolism and proliferation of cancer cells. *Cell Reports*, 7(4), 1248–1258.
- Lane, A. N., & Fan, T. W. (2015). Regulation of mammalian nucleotide metabolism and biosynthesis. *Nucleic Acids Research*, 43(4), 2466–2485.
- Le Gall, G., Guttula, K., Kellingray, L., Tett, A. J., Ten Hoopen, R., Kemsley, K. E., Savva, G. M., Ibrahim, A., & Narbad, A. (2018). Metabolite quantification of faecal extracts from colorectal cancer patients and healthy controls. *Oncotarget*, 9(70), 33278.
- Lee, B., Mahmud, I., Marchica, J., Dereziński, P., Qi, F., Wang, F., Joshi, P., Valerio, F., Rivera, I., & Patel, V. (2020). Integrated RNA and metabolite profiling of urine liquid biopsies for prostate cancer biomarker discovery. *Scientific Reports*, 10(1), 1–17.
- Lee, K. B., Ang, L., Yau, W.-P., & Seow, W. J. (2020). Association between metabolites and the risk of lung cancer: A systematic literature review and meta-analysis of observational studies. *Metabolites*, 10(9), 362.

- Lewis, C. A., Parker, S. J., Fiske, B. P., McCloskey, D., Gui, D. Y., Green, C. R., Vokes, N. I., Feist, A. M., Vander Heiden, M. G., & Metallo, C. M. (2014). Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Molecular Cell*, 55(2), 253–263.
- Li, X., Cheng, J., Shen, Y., Chen, J., Wang, T., Wen, F., & Chen, L. (2020). Metabolomic analysis of lung cancer patients with chronic obstructive pulmonary disease using gas chromatography-mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 190, 113524, Epub 2020 Aug 2. PMID: 32795777. Available from <https://doi.org/10.1016/j.jpba.2020.113524>.
- Liang, J.-H., Lin, Y., Ouyang, T., Tang, W., Huang, Y., Ye, W., Zhao, J.-Y., Wang, Z.-N., & Ma, C.-C. (2019). Nuclear magnetic resonance-based metabolomics and metabolic pathway networks from patient-matched esophageal carcinoma, adjacent noncancerous tissues and urine. *World Journal of Gastroenterology*, 25(25), 3218.
- Liang, L., Sun, F., Wang, H., & Hu, Z. (2021). Metabolomics, metabolic flux analysis and cancer pharmacology. *Pharmacology & Therapeutics*, 224, 107827.
- Liang, Q., Liu, H., Wang, C., & Li, B. (2016). Phenotypic characterization analysis of human hepatocarcinoma by urine metabolomics approach. *Scientific Reports*, 6(1), 1–8.
- Lim, Y., Tang, K. D., Karpe, A. V., Beale, D. J., Totsika, M., Kenny, L., Morrison, M., & Punyadeera, C. (2021). Chemoradiation therapy changes oral microbiome and metabolomic profiles in patients with oral cavity cancer and oropharyngeal cancer. *Head & Neck*, 43(5), 1521–1534.
- Lima, A. R., de Lourdes Bastos, M., Carvalho, M., & de Pinho, P. G. (2016). Biomarker discovery in human prostate cancer: An update in metabolomics studies. *Translational Oncology*, 9(4), 357–370.
- Lima, A. R., Pinto, J., Azevedo, A. I., Barros-Silva, D., Jerónimo, C., Henrique, R., de Lourdes Bastos, M., de Pinho, P. G., & Carvalho, M. (2019). Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine. *British Journal of Cancer*, 121(10), 857–868.
- Lima, A. R., Pinto, J., de Lourdes Bastos, M., Carvalho, M., & de Pinho, P. G. (2018). NMR-based metabolomics studies of human prostate cancer tissue. *Metabolomics: Official Journal of the Metabolomic Society*, 14(7), 1–11.
- Lin, T., Meng, L., & Tsai, R. Y. (2011). GTP depletion synergizes the anti-proliferative activity of chemotherapeutic agents in a cell type-dependent manner. *Biochemical and Biophysical Research Communications*, 414(2), 403–408.
- Lin, X., Lin, D., Ge, X., Qiu, S., Feng, S., & Chen, R. (2017). Noninvasive detection of nasopharyngeal carcinoma based on saliva proteins using surface-enhanced Raman spectroscopy. *Journal of Biomedical Optics*, 22(10), 105004.
- Lin, X., Xu, R., Mao, S., Zhang, Y., Dai, Y., Guo, Q., Song, X., Zhang, Q., Li, L., & Chen, Q. (2019). Metabolic biomarker signature for predicting the effect of neoadjuvant chemotherapy of breast cancer. *Annals of Translational Medicine*, 7(22).
- Lin, Y., Ma, C., Bezabeh, T., Wang, Z., Liang, J., Huang, Y., Zhao, J., Liu, X., Ye, W., & Tang, W. (2019). ¹H NMR-based metabolomics reveal overlapping discriminatory metabolites and metabolic pathway disturbances between colorectal tumor tissues and fecal samples. *International Journal of Cancer*, 145(6), 1679–1689.
- Lin, Y., Ma, C., Liu, C., Wang, Z., Yang, J., Liu, X., Shen, Z., & Wu, R. (2016). NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget*, 7(20), 29454.

- Liu, S.-Y., Zhang, R.-L., Kang, H., Fan, Z.-J., & Du, Z. (2013). Human liver tissue metabolic profiling research on hepatitis B virus-related hepatocellular carcinoma. *World Journal of Gastroenterology: WJG*, 19(22), 3423.
- Liu, W., & Phang, J. M. (2012). Proline dehydrogenase (oxidase), a mitochondrial tumor suppressor, and autophagy under the hypoxia microenvironment. *Autophagy*, 8(9), 1407–1409.
- Liu, X., Cheng, X., Liu, X., He, L., Zhang, W., Wang, Y., Sun, W., & Ji, Z. (2018). Investigation of the urinary metabolic variations and the application in bladder cancer biomarker discovery. *International Journal of Cancer*, 143(2), 408–418.
- Liu, X., Liu, G., Chen, L., Liu, F., Zhang, X., Liu, D., Liu, X., Cheng, X., & Liu, L. (2020). Untargeted metabolomic characterization of ovarian tumors. *Cancers*, 12(12), 3642.
- Liu, Y., Mao, C., Wang, M., Liu, N., Ouyang, L., Liu, S., Tang, H., Cao, Y., Liu, S., & Wang, X. (2020). Cancer progression is mediated by proline catabolism in non-small cell lung cancer. *Oncogene*, 39(11), 2358–2376.
- Locasale, J. W., Melman, T., Song, S., Yang, X., Swanson, K. D., Cantley, L. C., Wong, E. T., & Asara, J. M. (2012). Metabolomics of human cerebrospinal fluid identifies signatures of malignant glioma. *Molecular & Cellular Proteomics*, 11(6), M111–014688.
- Lohavanichbutr, P., Zhang, Y., Wang, P., Gu, H., Nagana Gowda, G., Djukovic, D., Buas, M. F., Raftery, D., & Chen, C. (2018). Salivary metabolite profiling distinguishes patients with oral cavity squamous cell carcinoma from normal controls. *PLoS One*, 13(9), e0204249.
- Long, Z., Zhou, J., Xie, K., Wu, Z., Yin, H., Daria, V., Tian, J., Zhang, N., Li, L., & Zhao, Y. (2020). Metabolomic markers of colorectal tumor with different clinicopathological features. *Frontiers in Oncology*, 10, 981.
- López-Garrido, L., Bañuelos-Hernández, A. E., Pérez-Hernández, E., Tecualt-Gómez, R., Quiroz-Williams, J., Ariza-Castolo, A., Becerra-Martínez, E., & Pérez-Hernández, N. (2020). Metabolic profiling of serum in patients with cartilage tumours using ^1H -NMR spectroscopy: A pilot study. *Magnetic Resonance in Chemistry*, 58(1), 65–76.
- Loras, A., Martínez-Bisbal, M. C., Quintás, G., Gil, S., Martínez-Máñez, R., & Ruiz-Cerdá, J. L. (2019). Urinary metabolic signatures detect recurrences in non-muscle invasive bladder cancer. *Cancers*, 11(7), 914.
- Lv, D., Zou, Y., Zeng, Z., Yao, H., Ding, S., Bian, Y., Wen, L., & Xie, X. (2020). Comprehensive metabolomic profiling of osteosarcoma based on UHPLC-HRMS. *Metabolomics: Official Journal of the Metabolomic Society*, 16(12), 1–11.
- Macias, R. I., Muñoz-Bellví, L., Sánchez-Martín, A., Arretxe, E., Martínez-Arranz, I., Lapitz, A., Gutiérrez, M. L., La Casta, A., Alonso, C., & González, L. M. (2020). A novel serum metabolomic profile for the differential diagnosis of distal cholangiocarcinoma and pancreatic ductal adenocarcinoma. *Cancers*, 12(6), 1433.
- Madak, J. T., Bankhead, A., III, Cuthbertson, C. R., Showalter, H. D., & Neamati, N. (2019). Revisiting the role of dihydroorotate dehydrogenase as a therapeutic target for cancer. *Pharmacology & Therapeutics*, 195, 111–131.
- Maddocks, O. D., Athineos, D., Cheung, E. C., Lee, P., Zhang, T., van den Broek, N. J., Mackay, G. M., Labuschagne, C. F., Gay, D., & Kruiswijk, F. (2017). Modulating the therapeutic response of tumors to dietary serine and glycine starvation. *Nature*, 544(7650), 372–376.
- Martín-Blázquez, A., Jiménez-Luna, C., Díaz, C., Martínez-Galán, J., Prados, J., Vicente, F., Melguizo, C., Genilloud, O., Pérez del Palacio, J., & Caba, O. (2020). Discovery of pancreatic adenocarcinoma biomarkers by untargeted metabolomics. *Cancers*, 12(4), 1002.

- Martins, R. G., Gonçalves, L. G., Cunha, N., & Bugalho, M. J. (2019). Metabolomic urine profile: Searching for new biomarkers of SDHx-associated pheochromocytomas and paragangliomas. *The Journal of Clinical Endocrinology & Metabolism*, 104(11), 5467–5477.
- Metallo, C. M., Gameiro, P. A., Bell, E. L., Mattaini, K. R., Yang, J., Hiller, K., Jewell, C. M., Johnson, Z. R., Irvine, D. J., & Guarente, L. (2012). Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature*, 481(7381), 380–384.
- Metere, A., Graves, C. E., Chirico, M., Caramujo, M. J., Pisanu, M. E., & Iorio, E. (2020). Metabolomic reprogramming detected by 1H-NMR spectroscopy in human thyroid cancer tissues. *Biology*, 9(6), 112.
- Michela, T., Antonio, M., Chiara, C., Somma, P., De Rosa, I., Troisi, J., Scala, G., Salvi, R., Aldo, P., & Rosalinda, S. (2020). Altered lung tissue lipidomic profile in caspase-4 positive non-small cell lung cancer (NSCLC) patients. *Oncotarget*, 11(38), 3515.
- Mikkonen, J. J., Singh, S. P., Akhi, R., Salo, T., Lappalainen, R., González-Arriagada, W. A., Ajudarte Lopes, M., Kullaa, A. M., & Myllymaa, S. (2018). Potential role of nuclear magnetic resonance spectroscopy to identify salivary metabolite alterations in patients with head and neck cancer. *Oncology Letters*, 16(5), 6795–6800.
- Miolo, G., Di Gregorio, E., Saorin, A., Lombardi, D., Scalzone, S., Buonadonna, A., Steffan, A., & Corona, G. (2020). Integration of serum metabolomics into clinical assessment to improve outcome prediction of metastatic soft tissue sarcoma patients treated with trabectedin. *Cancers*, 12(7), 1983.
- Mitruka, M., Gore, C. R., Kumar, A., Sarode, S. C., & Sharma, N. K. (2020). Undetectable free aromatic amino acids in nails of breast carcinoma: Biomarker discovery by a novel metabolite purification VTGE system. *Frontiers in Oncology*, 10, 908.
- Miyagi, Y., Higashiyama, M., Gochi, A., Akaike, M., Ishikawa, T., Miura, T., Saruki, N., Bando, E., Kimura, H., & Imamura, F. (2011). Plasma free amino acid profiling of five types of cancer patients and its application for early detection. *PLoS One*, 6(9), e24143.
- Monleon, D., Morales, J. M., Barrasa, A., Lopez, J. A., Vazquez, C., & Celda, B. (2009). Metabolite profiling of fecal water extracts from human colorectal cancer. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 22(3), 342–348.
- More, T. H., RoyChoudhury, S., Christie, J., Taunk, K., Mane, A., Santra, M. K., Chaudhury, K., & Rapole, S. (2018). Metabolomic alterations in invasive ductal carcinoma of breast: A comprehensive metabolomic study using tissue and serum samples. *Oncotarget*, 9(2), 2678.
- Moreno-Sánchez, R., Marín-Hernández, Á., Gallardo-Pérez, J. C., Pacheco-Velázquez, S. C., Robledo-Cadena, D. X., Padilla-Flores, J. A., Saavedra, E., & Rodríguez-Enríquez, S. (2020). Physiological role of glutamate dehydrogenase in cancer cells. *Frontiers in Oncology*, 10, 429.
- Moreno-Sánchez, R., Rodríguez-Enríquez, S., Marín-Hernández, A., & Saavedra, E. (2007). Energy metabolism in tumor cells. *The FEBS Journal*, 274(6), 1393–1418.
- Morrot, A., Fonseca, L. M., da, Salustiano, E. J., Gentile, L. B., Conde, L., Filardy, A. A., Franklin, T. N., da Costa, K. M., Freire-de-Lima, C. G., & Freire-de-Lima, L. (2018). Metabolic symbiosis and immunomodulation: How tumor cell-derived lactate may disturb innate and adaptive immune responses. *Frontiers in Oncology*, 8, 81.
- Mullen, A. R., Wheaton, W. W., Jin, E. S., Chen, P.-H., Sullivan, L. B., Cheng, T., Yang, Y., Linehan, W. M., Chandel, N. S., & DeBerardinis, R. J. (2012). Reductive

- carboxylation supports growth in tumour cells with defective mitochondria. *Nature*, 481(7381), 385–388.
- Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., Ikeda, K., Kawada, N., Ochiya, T., & Taguchi, Y. (2015). Comprehensive analysis of transcriptome and metabolome analysis in intrahepatic cholangiocarcinoma and hepatocellular carcinoma. *Scientific Reports*, 5(1), 1–12.
- Newman, A. C., & Maddocks, O. D. (2017). One-carbon metabolism in cancer. *British Journal of Cancer*, 116(12), 1499–1504.
- Nguyen, T.-L., & Durán, R. V. (2018). Glutamine metabolism in cancer therapy. *Cancer Drug Resistance*, 1(3), 126–138.
- Nizioł, J., Copié, V., Tripet, B. P., Nogueira, L. B., Nogueira, K. O., Ossoliński, K., Arendowski, A., & Ruman, T. (2021). Metabolomic and elemental profiling of human tissue in kidney cancer. *Metabolomics: Official Journal of the Metabolomic Society*, 17(3), 1–15.
- Nizioł, J., Ossoliński, K., Tripet, B. P., Copié, V., Arendowski, A., & Ruman, T. (2021). Nuclear magnetic resonance and surface-assisted laser desorption/ionization mass spectrometry-based metabolome profiling of urine samples from kidney cancer patients. *Journal of Pharmaceutical and Biomedical Analysis*, 193, 113752.
- Nomura, M., Nagatomo, R., Doi, K., Shimizu, J., Baba, K., Saito, T., Matsumoto, S., Inoue, K., & Muto, M. (2020). Association of short-chain fatty acids in the gut microbiome with clinical response to treatment with nivolumab or pembrolizumab in patients with solid cancer tumors. *JAMA Network Open*, 3(4), e202895.
- Noto, A., De Vitis, C., Pisani, M. E., Roscilli, G., Ricci, G., Catizone, A., Sorrentino, G., Chianese, G., Taglialatela-Scafati, O., & Trisciuglio, D. (2017). Stearyl-CoA-desaturase 1 regulates lung cancer stemness via stabilization and nuclear localization of YAP/TAZ. *Oncogene*, 36(32), 4573–4584.
- Ose, J., Gigic, B., Lin, T., Liesenfeld, D. B., Böhm, J., Nattenmüller, J., Scherer, D., Zielske, L., Schrotz-King, P., & Habermann, N. (2019). Multiplatform urinary metabolomics profiling to discriminate cachectic from non-cachectic colorectal cancer patients: Pilot results from the colocare study. *Metabolites*, 9(9), 178.
- Osman, D., Ali, O., Obada, M., El-Mezayen, H., & El-Said, H. (2017). Chromatographic determination of some biomarkers of liver cirrhosis and hepatocellular carcinoma in Egyptian patients. *Biomedical Chromatography*, 31(6), e3893.
- OuYang, D., Xu, J., Huang, H., & Chen, Z. (2011). Metabolomic profiling of serum from human pancreatic cancer patients using ^1H NMR spectroscopy and principal component analysis. *Applied Biochemistry and Biotechnology*, 165(1), 148–154.
- Ozturk, L. K., Emekli-Alturfan, E., Kasikci, E., Demir, G., & Yarat, A. (2011). Salivary total sialic acid levels increase in breast cancer patients: A preliminary study. *Medicinal Chemistry*, 7(5), 443–447.
- Pandey, P. R., Liu, W., Xing, F., Fukuda, K., & Watabe, K. (2012). Anti-cancer drugs targeting fatty acid synthase (FAS). *Recent Patents on Anti-Cancer Drug Discovery*, 7(2), 185–197.
- Paul, A., Srivastava, S., Roy, R., Anand, A., Gaurav, K., Husain, N., Jain, S., & Sonkar, A. A. (2020). Malignancy prediction among tissues from Oral SCC patients including neck invasions: A ^1H HRMAS NMR based metabolomic study. *Metabolomics: Official Journal of the Metabolomic Society*, 16(3), 1–19.
- Pavlova, N. N., & Thompson, C. B. (2016). The emerging hallmarks of cancer metabolism. *Cell Metabolism*, 23(1), 27–47.

- Penney, K. L., Tyekucheva, S., Rosenthal, J., El Fandy, H., Carelli, R., Borgstein, S., Zadra, G., Fanelli, G. N., Stefanizzi, L., & Giunchi, F. (2021). Metabolomics of prostate cancer gleason score in tumor tissue and serum. *Molecular Cancer Research*, 19(3), 475–484.
- Peralbo-Molina, A., Calderón-Santiago, M., Priego-Capote, F., Jurado-Gámez, B., & De Castro, M. L. (2016a). Identification of metabolomics panels for potential lung cancer screening by analysis of exhaled breath condensate. *Journal of Breath Research*, 10(2), 026002.
- Peralbo-Molina, A., Calderón-Santiago, M., Priego-Capote, F., Jurado-Gámez, B., & De Castro, M. L. (2016b). Metabolomics analysis of exhaled breath condensate for discrimination between lung cancer patients and risk factor individuals. *Journal of Breath Research*, 10(1), 016011.
- Pérez-Rambla, C., Puchades-Carrasco, L., García-Flores, M., Rubio-Briones, J., López-Guerrero, J. A., & Pineda-Lucena, A. (2017). Non-invasive urinary metabolomic profiling discriminates prostate cancer from benign prostatic hyperplasia. *Metabolomics: Official Journal of the Metabolomic Society*, 13(5), 52.
- Phang, J. M. (2019). Proline metabolism in cell regulation and cancer biology: Recent advances and hypotheses. *Antioxidants & Redox Signaling*, 30(4), 635–649.
- Phua, L. C., Chue, X. P., Koh, P. K., Cheah, P. Y., Ho, H. K., & Chan, E. C. Y. (2014). Non-invasive fecal metabonomic detection of colorectal cancer. *Cancer Biology & Therapy*, 15(4), 389–397.
- Pinto, J., Carapito, Â., Amaro, F., Lima, A. R., Carvalho-Maia, C., Martins, M. C., Jerónimo, C., Henrique, R., Bastos, M., de, L., & Guedes de Pinho, P. (2021). Discovery of volatile biomarkers for bladder cancer detection and staging through urine metabolomics. *Metabolites*, 11(4), 199.
- Plewa, S., Horała, A., Dereziński, P., Klupczynska, A., Nowak-Markwitz, E., Matysiak, J., & Kokot, Z. J. (2017). Usefulness of amino acid profiling in ovarian cancer screening with special emphasis on their role in cancerogenesis. *International Journal of Molecular Sciences*, 18(12), 2727.
- Possemato, R., Marks, K. M., Shaul, Y. D., Pacold, M. E., Kim, D., Birsoy, K., Sethumadhavan, S., Woo, H.-K., Jang, H. G., & Jha, A. K. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature*, 476 (7360), 346–350.
- Potze, L., Di Franco, S., Grandela, C., Pras-Raves, M., Picavet, D., Van Veen, H., Van Lenthe, H., Mullauer, F., Van Der Wel, N., & Luyf, A. (2016). Betulinic acid induces a novel cell death pathway that depends on cardiolipin modification. *Oncogene*, 35(4), 427–437.
- Qian, K., Wang, Y., Hua, L., Chen, A., & Zhang, Y. (2018). New method of lung cancer detection by saliva test using surface-enhanced Raman spectroscopy. *Thoracic Cancer*, 9(11), 1556–1561.
- Qian, Y., Liang, X., Kong, P., Cheng, Y., Cui, H., Yan, T., Wang, J., Zhang, L., Liu, Y., & Guo, S. (2020). Elevated DHODH expression promotes cell proliferation via stabilizing β -catenin in esophageal squamous cell carcinoma. *Cell Death & Disease*, 11(10), 1–13.
- Qiu, F., Huang, J., & Sui, M. (2015). Targeting arginine metabolism pathway to treat arginine-dependent cancers. *Cancer Letters*, 364(1), 1–7.
- Quinlan, C. L., Kaiser, S. E., Boláños, B., Nowlin, D., Grantner, R., Karlicek-Bryant, S., Feng, J. L., Jenkinson, S., Freeman-Cook, K., & Dann, S. G. (2017). Targeting S-adenosylmethionine biosynthesis with a novel allosteric inhibitor of Mat2A. *Nature Chemical Biology*, 13(7), 785.

- Rabinovich, S., Adler, L., Yizhak, K., Sarver, A., Silberman, A., Agron, S., Stettner, N., Sun, Q., Brandis, A., & Helbling, D. (2015). Diversion of aspartate in ASS1-deficient tumours fosters de novo pyrimidine synthesis. *Nature*, 527(7578), 379–383.
- Răchieriu, C., Eniu, D. T., Moiș, E., Graur, F., Socaciu, C., Socaciu, M. A., & Hajjar, N. A. (2021). Lipidomic Signatures for Colorectal Cancer Diagnosis and Progression Using UPLC-QTOF-ESI + MS. *Biomolecules*, 11(3), 417.
- Rekha, P., Aruna, P., Brindha, E., Koteeswaran, D., Baladavid, M., & Ganesan, S. (2016). Near-infrared Raman spectroscopic characterization of salivary metabolites in the discrimination of normal from oral premalignant and malignant conditions. *Journal of Raman Spectroscopy*, 47(7), 763–772.
- Ren, Z., Rajani, C., & Jia, W. (2021). The distinctive serum metabolomes of gastric, esophageal and colorectal. *Cancers. Cancers*, 13(4), 720.
- Roberts, M. J., Richards, R. S., Chow, C. W., Buck, M., Yaxley, J., Lavin, M. F., Schirra, H. J., & Gardiner, R. A. (2017). Seminal plasma enables selection and monitoring of active surveillance candidates using nuclear magnetic resonance-based metabolomics: A preliminary investigation. *Prostate International*, 5(4), 149–157.
- Romero-Garcia, S., Lopez-Gonzalez, J. S., Bèz-Viveros, J. L., Aguilar-Cazares, D., & Prado-Garcia, H. (2011). Tumor cell metabolism: An integral view. *Cancer Biology & Therapy*, 12(11), 939–948.
- Ros-Mazurczyk, M., Jelonek, K., Marczyk, M., Binczyk, F., Pietrowska, M., Polanska, J., Dziadziuszko, R., Jassem, J., Rzyman, W., & Widlak, P. (2017). Serum lipid profile discriminates patients with early lung cancer from healthy controls. *Lung Cancer (Amsterdam, Netherlands)*, 112, 69–74.
- Rubenstein, J. L., Geng, H., Fraser, E. J., Formaker, P., Chen, L., Sharma, J., Killea, P., Choi, K., Ventura, J., & Kurhanewicz, J. (2018). Phase 1 investigation of lenalidomide/rituximab plus outcomes of lenalidomide maintenance in relapsed CNS lymphoma. *Blood Advances*, 2(13), 1595–1607.
- Saha, S. K., Islam, S., Abdullah-Al-Wadud, M., Islam, S., Ali, F., & Park, K. S. (2019). Multiomics analysis reveals that GLS and GLS2 differentially modulate the clinical outcomes of cancer. *Journal of Clinical Medicine*, 8(3), 355.
- Sanderson, S. M., Gao, X., Dai, Z., & Locasale, J. W. (2019). Methionine metabolism in health and cancer: A nexus of diet and precision medicine. *Nature Reviews. Cancer*, 19(11), 625–637.
- Scatena, R. (2012). Mitochondria and cancer: A growing role in apoptosis, cancer cell metabolism and dedifferentiation. *Advances in Mitochondrial Medicine*, 942, 287–308.
- Semenza, G. L. (2011). Oxygen sensing, homeostasis, and disease. *New England Journal of Medicine*, 365(6), 537–547.
- Shao, X., Wang, K., Liu, X., Gu, C., Zhang, P., Xie, J., Liu, W., Sun, L., Chen, T., & Li, Y. (2016). Screening and verifying endometrial carcinoma diagnostic biomarkers based on a urine metabolomic profiling study using UPLC-Q-TOF/MS. *Clinica Chimica Acta*, 463, 200–206.
- Shi, K., Wang, Q., Su, Y., Xuan, X., Liu, Y., Chen, W., Qian, Y., & Lash, G. E. (2018). Identification and functional analyses of differentially expressed metabolites in early stage endometrial carcinoma. *Cancer Science*, 109(4), 1032–1043.
- Silva, C. L., Olival, A., Perestrelo, R., Silva, P., Tomás, H., & Câmara, J. S. (2019). Untargeted urinary ^1H NMR-based metabolomic pattern as a potential platform in breast cancer detection. *Metabolites*, 9(11), 269.

- Skoura, E., Datseris, I. E., Platis, I., Oikonomopoulos, G., & Syrigos, K. N. (2012). Role of positron emission tomography in the early prediction of response to chemotherapy in patients with non–small-cell lung cancer. *Clinical Lung Cancer*, 13(3), 181–187.
- Somani, R. R., Rai, P. R., & Kandpile, P. S. (2018). Ornithine decarboxylase inhibition: A strategy to combat various diseases. *Mini Reviews in Medicinal Chemistry*, 18(12), 1008–1021.
- Song, X., Yang, X., Narayanan, R., Shankar, V., Ethiraj, S., Wang, X., Duan, N., Ni, Y.-H., Hu, Q., & Zare, R. N. (2020). Oral squamous cell carcinoma diagnosed from saliva metabolic profiling. *Proceedings of the National Academy of Sciences*, 117(28), 16167–16173.
- Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R. J., & Li, Y. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457(7231), 910–914.
- Sridharan, G., Ramani, P., Patankar, S., & Vijayaraghavan, R. (2019). Evaluation of salivary metabolomics in oral leukoplakia and oral squamous cell carcinoma. *Journal of Oral Pathology & Medicine*, 48(4), 299–306.
- Stepien, M., Keski-Rahkonen, P., Kiss, A., Robinot, N., Duarte-Salles, T., Murphy, N., Perlemuter, G., Viallon, V., Tjønneland, A., & Rostgaard-Hansen, A. L. (2021). Metabolic perturbations prior to hepatocellular carcinoma diagnosis: Findings from a prospective observational cohort study. *International Journal of Cancer*, 148(3), 609–625.
- Sugimoto, M., Wong, D. T., Hirayama, A., Soga, T., & Tomita, M. (2010). Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics: Official Journal of the Metabolomic Society*, 6(1), 78–95.
- Tang, Q., Cang, S., Jiao, J., Rong, W., Xu, H., Bi, K., Li, Q., & Liu, R. (2020). Integrated study of metabolomics and gut metabolic activity from ulcerative colitis to colorectal cancer: The combined action of disordered gut microbiota and linoleic acid metabolic pathway might fuel cancer. *Journal of Chromatography. A*, 1629, 461503.
- Taylor, N. J., Gaynanova, I., Eschrich, S. A., Welsh, E. A., Garrett, T. J., Beecher, C., Sharma, R., Koomen, J. M., Smalley, K. S., & Messina, J. L. (2020). Metabolomics of primary cutaneous melanoma and matched adjacent extratumoral microenvironment. *PLoS One*, 15(10), e0240849.
- Tedeschi, P. M., Johnson-Farley, N., Lin, H., Shelton, L. M., Ooga, T., Mackay, G., Van Den Broek, N., Bertino, J. R., & Vazquez, A. (2015). Quantification of folate metabolism using transient metabolic flux analysis. *Cancer & Metabolism*, 3(1), 1–14.
- Teilhet, C., Morvan, D., Joubert-Zakeyh, J., Biesse, A.-S., Pereira, B., Massoulier, S., Dechelotte, P., Pezet, D., Buc, E., & Lamblin, G. (2017). Specificities of human hepatocellular carcinoma developed on non-alcoholic fatty liver disease in absence of cirrhosis revealed by tissue extracts 1H-NMR spectroscopy. *Metabolites*, 7(4), 49.
- Terlizzi, M., Molino, A., Colarusso, C., Somma, P., De Rosa, I., Troisi, J., Scala, G., Salvi, R., Pinto, A., & Sorrentino, R. (2020). Altered lung tissue lipidomic profile in caspase-4 positive non-small cell lung cancer (NSCLC) patients. *Oncotarget*, 11(38), 3515–3525.
- Tracz-Gaszewska, Z., & Dobrzyn, P. (2019). Stearyl-CoA desaturase 1 as a therapeutic target for the treatment of cancer. *Cancers*, 11(7), 948.
- Troisi, J., Raffone, A., Travagliano, A., Belli, G., Belli, C., Anand, S., Giugliano, L., Cavallo, P., Scala, G., & Symes, S. (2020). Development and validation of a serum

- metabolomic signature for endometrial cancer screening in postmenopausal women. *JAMA Network Open*, 3(9), e2018327, e2018327.
- Troisi, J., Sarno, L., Landolfi, A., Scala, G., Martinelli, P., Venturella, R., Di Cello, A., Zullo, F., & Guida, M. (2018). Metabolomic signature of endometrial cancer. *Journal of Proteome Research*, 17(2), 804–812.
- Udo, R., Katsumata, K., Kuwabara, H., Enomoto, M., Ishizaki, T., Sunamura, M., Nagakawa, Y., Soya, R., Sugimoto, M., & Tsuchida, A. (2020). Urinary charged metabolite profiling of colorectal cancer using capillary electrophoresis-mass spectrometry. *Scientific Reports*, 10(1), 1–10.
- Unger, K., Mehta, K. Y., Kaur, P., Wang, Y., Menon, S. S., Jain, S. K., Moonjelly, R. A., Suman, S., Datta, K., & Singh, R. (2018). Metabolomics based predictive classifier for early detection of pancreatic ductal adenocarcinoma. *Oncotarget*, 9(33), 23078.
- Vazquez, A., Kamphorst, J. J., Markert, E. K., Schug, Z. T., Tardito, S., & Gottlieb, E. (2016). Cancer metabolism at a glance. *Journal of Cell Science*, 129(18), 3367–3373.
- Vorkas, P. A., & Li, J. V. (2018). *Tissue multiplatform-based metabolomics/metabonomics for enhanced metabolome coverage*. *Metabolic profiling* (pp. 239–260). Springer.
- Wallace, P. W., Conrad, C., Brückmann, S., Pang, Y., Caleiras, E., Murakami, M., Korpershoek, E., Zhuang, Z., Rapizzi, E., & Kroiss, M. (2020). Metabolomics, machine learning and immunohistochemistry to predict succinate dehydrogenase mutational status in phaeochromocytomas and paragangliomas. *The Journal of Pathology*, 251(4), 378–387.
- Wang, D., Li, W., Zou, Q., Yin, L., Du, Y., Gu, J., & Suo, J. (2017). Serum metabolomic profiling of human gastric cancer and its relationship with the prognosis. *Oncotarget*, 8 (66), 110000.
- Wang, X., Zhao, X., Zhao, J., Yang, T., Zhang, F., & Liu, L. (2021). Serum metabolite signatures of epithelial ovarian cancer based on targeted metabolomics. *Clinica Chimica Acta*, 518, 59–69.
- Wang, Z., Yip, L. Y., Lee, J. H. J., Wu, Z., Chew, H. Y., Chong, P. K. W., Teo, C. C., Ang, H. Y.-K., Peh, K. L. E., & Yuan, J. (2019). Methionine is a metabolic dependency of tumor-initiating cells. *Nature Medicine*, 25(5), 825–837.
- Warburg, O. (1956). On the origin of cancer cells. *Science (New York, N.Y.)*, 123(3191), 309–314.
- Ward, P. S., & Thompson, C. B. (2012). Metabolic reprogramming: A cancer hallmark even warburg did not anticipate. *Cancer Cell*, 21(3), 297–308.
- Wei, Y., Liu, P., Li, Q., Du, J., Chen, Y., Wang, Y., Shi, H., Wang, Y., Zhang, H., & Xue, W. (2019). The effect of MTHFD2 on the proliferation and migration of colorectal cancer cell lines. *OncoTargets and Therapy*, 12, 6361.
- Wojakowska, A., Zebrowska, A., Skowronek, A., Rutkowski, T., Polanski, K., Widlak, P., Marczak, L., & Pietrowska, M. (2020). Metabolic profiles of whole serum and serum-derived exosomes are different in head and neck cancer patients treated by radiotherapy. *Journal of Personalized Medicine*, 10(4), 229.
- Wojtowicz, W., Zabek, A., Deja, S., Dawiskiba, T., Pawelka, D., Glod, M., Balcerzak, W., & Mlynarz, P. (2017). Serum and urine ^1H NMR-based metabolomics in the diagnosis of selected thyroid diseases. *Scientific Reports*, 7(1), 1–13.
- Wu, J., Wu, M., & Wu, Q. (2020). Identification of potential metabolite markers for colon cancer and rectal cancer using serum metabolomics. *Journal of Clinical Laboratory Analysis*, 34(8), e23333.

- Xavier Assad, D., Acevedo, A. C., Cançado Porto Mascarenhas, E., Costa Normando, A. G., Pichon, V., Chardin, H., Neves Silva Guerra, E., & Combes, A. (2020). Using an untargeted metabolomics approach to identify salivary metabolites in women with breast cancer. *Metabolites*, 10(12), 506.
- Xia, H., Chen, J., Sekar, K., Shi, M., Xie, T., & Hui, K. M. (2019). Clinical and metabolomics analysis of hepatocellular carcinoma patients with diabetes mellitus. *Metabolomics: Official Journal of the Metabolomic Society*, 15(12), 1–10.
- Xiang, L., Xie, G., Liu, C., Zhou, J., Chen, J., Yu, S., Li, J., Pang, X., Shi, H., & Liang, H. (2013). Knock-down of glutaminase 2 expression decreases glutathione, NADH, and sensitizes cervical cancer to ionizing radiation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(12), 2996–3005.
- Xie, Y., Dong, C. D., Wu, Q., Jiang, Y., Yao, K., Zhang, J., Zhao, S., Ren, Y., Yuan, Q., & Chen, X. (2020). Ornithine decarboxylase inhibition downregulates multiple pathways involved in the formation of precancerous lesions of esophageal squamous cell cancer. *Molecular Carcinogenesis*, 59(2), 215–226.
- Xu, J., Chen, Y., Zhang, R., He, J., Song, Y., Wang, J., Wang, H., Wang, L., Zhan, Q., & Abliz, Z. (2016). Global metabolomics reveals potential urinary biomarkers of esophageal squamous cell carcinoma for diagnosis and staging. *Scientific Reports*, 6(1), 1–10.
- Yang, B., Zhang, C., Cheng, S., Li, G., Griebel, J., & Neuhaus, J. (2021). Novel metabolic signatures of prostate cancer revealed by ^1H -NMR metabolomics of urine. *Diagnostics*, 11(2), 149.
- Yang, J., Seo, H., Lee, W. H., Lee, D. H., Kym, S., Park, Y. S., Kim, J. G., Jang, I.-J., Kim, Y.-K., & Cho, J.-Y. (2020). Colorectal cancer diagnostic model utilizing metagenomic and metabolomic data of stool microbial extracellular vesicles. *Scientific Reports*, 10(1), 1–10.
- Yang, W., Mu, T., Jiang, J., Sun, Q., Hou, X., Sun, Y., Zhong, L., Wang, C., & Sun, C. (2018). Identification of potential biomarkers and metabolic profiling of serum in ovarian cancer patients using UPLC/Q-TOF MS. *Cellular Physiology and Biochemistry*, 51(3), 1134–1148.
- Ye, W., Lin, Y., Bezabeh, T., Ma, C., Liang, J., Zhao, J., Ouyang, T., Tang, W., & Wu, R. (2021). ^1H NMR-based metabolomics of paired esophageal tumor tissues and serum samples identifies specific serum biomarkers for esophageal cancer. *NMR in Biomedicine*, 34(6), e4505.
- Yecies, J. L., & Manning, B. D. (2010). Chewing the fat on tumor cell metabolism. *Cell*, 140(1), 28–30.
- Yoo, B. C., Lee, J. H., Kim, K.-H., Lin, W., Kim, J. H., Park, J. B., Park, H. J., Shin, S. H., Yoo, H., & Kwon, J. W. (2017). Cerebrospinal fluid metabolomic profiles can discriminate patients with leptomeningeal carcinomatosis from patients at high risk for leptomeningeal metastasis. *Oncotarget*, 8(60), 101203.
- Yu, C., Yang, L., Cai, M., Zhou, F., Xiao, S., Li, Y., Wan, T., Cheng, D., Wang, L., & Zhao, C. (2020). Down-regulation of MTHFD2 inhibits NSCLC progression by suppressing cycle-related genes. *Journal of Cellular and Molecular Medicine*, 24(2), 1568–1577.
- Yu, L., Lai, Q., Feng, Q., Li, Y., Feng, J., & Xu, B. (2021). Serum metabolic profiling analysis of chronic gastritis and gastric cancer by untargeted metabolomics. *Frontiers in Oncology*, 11, 443.

- Zhang, L., Jin, H., Guo, X., Yang, Z., Zhao, L., Tang, S., Mo, P., Wu, K., Nie, Y., & Pan, Y. (2012). Distinguishing pancreatic cancer from chronic pancreatitis and healthy individuals by ^1H nuclear magnetic resonance-based metabolomic profiles. *Clinical Biochemistry*, 45(13–14), 1064–1069.
- Zhang, Y., Du, Y., Song, Z., Liu, S., Li, W., Wang, D., & Suo, J. (2020). Profiling of serum metabolites in advanced colon cancer using liquid chromatography-mass spectrometry. *Oncology Letters*, 19(6), 4002–4010.
- Zheng, H., Dong, B., Ning, J., Shao, X., Zhao, L., Jiang, Q., Ji, H., Cai, A., Xue, W., & Gao, H. (2020). NMR-based metabolomics analysis identifies discriminatory metabolic disturbances in tissue and biofluid samples for progressive prostate cancer. *Clinica Chimica Acta*, 501, 241–251.
- Zheng, L., MacKenzie, E. D., Karim, S. A., Hedley, A., Blyth, K., Kalna, G., Watson, D. G., Szlosarek, P., Frezza, C., & Gottlieb, E. (2013). Reversed argininosuccinate lyase activity in fumurate hydratase-deficient cancer cells. *Cancer & Metabolism*, 1(1), 1–11.
- Zhong, L., Cheng, F., Lu, X., Duan, Y., & Wang, X. (2016). Untargeted saliva metabolomics study of breast cancer based on ultra performance liquid chromatography coupled to mass spectrometry with HILIC and RPLC separations. *Talanta*, 158, 351–360.
- Zhu, X., Wang, K., Liu, G., Wang, Y., Xu, J., Liu, L., Li, M., Shi, J., Aa, J., & Yu, L. (2017). Metabolic perturbation and potential markers in patients with esophageal cancer. *Gastroenterology Research and Practice*, 2017, 5469597.
- Zhu, Z., & Leung, G. K. K. (2020). More than a metabolic enzyme: MTHFD2 as a novel target for anticancer therapy? *Frontiers in Oncology*, 10, 658.
- Zhu, Z.-J., Qi, Z., Zhang, J., Xue, W.-H., Li, L.-F., Shen, Z.-B., Li, Z.-Y., Yuan, Y.-L., Wang, W.-B., & Zhao, J. (2020). Untargeted metabolomics analysis of esophageal squamous cell carcinoma discovers dysregulated metabolic pathways and potential diagnostic biomarkers. *Journal of Cancer*, 11(13), 3944.
- Zou, J., Han, Z., Zhou, L., Cai, C., Luo, H., Huang, Y., Liang, Y., He, H., Jiang, F., & Wang, C. (2015). Elevated expression of IMPDH2 is associated with progression of kidney and bladder cancer. *Medical Oncology*, 32(1), 1–6.

Metabolomics as a tool for precision medicine

17

Edoardo Saccenti¹ and Leonardo Tenori²

¹*Laboratory of Systems and Synthetic Biology, Wageningen University & Research,
Wageningen, The Netherlands*

²*Department of Chemistry and Magnetic Resonance Center (CERM), University of Florence,
Florence, Italy*

Systems approaches and systems medicine

Biomedical interventions and therapies have a much smaller impact on health than commonly believed; for instance, standard medical care contributes only about 10% in reducing premature deaths with respect to other contributing factors like genetic predisposition, social factors, and individual health behaviors (Schroeder, 2007).

Current medical science is largely based on using the reductionist view that all diseases can be reduced to biological causes in the body; typically, treatments of those diseases are also biological in character, such as surgery or medications (Ghaemi, 2015). This approach, which has been a predominant paradigm of science over the past two centuries, involves the notion that complex phenomena like disease onset and progression may be better understood by breaking them down into smaller, simpler components (Ahn et al., 2006).

Thus, science underlying common medical practice, from diagnosis to treatment and prevention, is based on the assumption that information about individual parts is sufficient to explain the whole (Ahn et al., 2006).

However, it is now accepted that the behavior of complex systems like the human body and its functioning cannot be predicted by the characterization of its parts alone. The difficulties in accounting for this and in including this in biomedical research and practice can be considered as the underlying factor explaining why biomedical practice is in many cases inadequate, resulting in delayed or non-timely diagnosis, suboptimal or inefficient treatments and waste of resources.

Recognizing that a reductionist approach was hampering the ability to understand how multiple variables interact with one another to create emergent biological effects, led to the introduction of systems biology. The ultimate goals of Systems Biology are to understand how the constituents of biological systems interact, potentially at different levels and at different scales, to determine the behavior of the systems and, possibly, to predict it. The idea originated with the

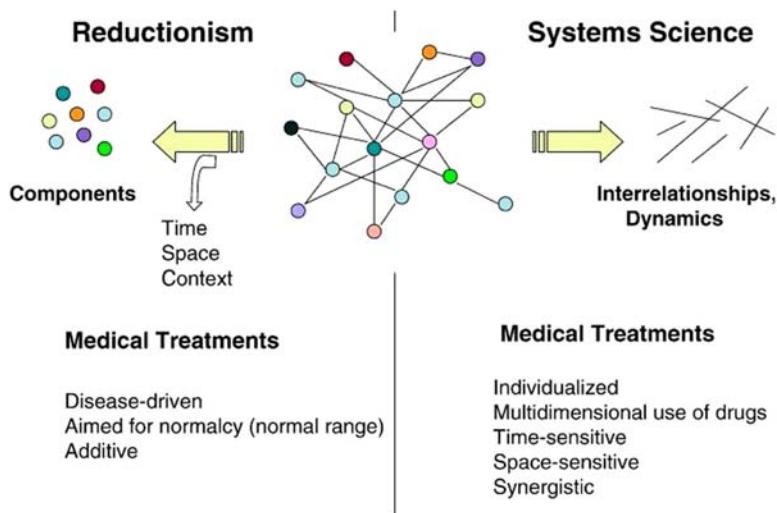
seminal works of [Wiener \(1961\)](#) and [Von Bertalanffy \(2021\)](#) and the development of new computational and theoretical physical, mathematical and statistical approaches together with high-throughput molecular techniques has led to the application of systems methods to biological problems.

Overall, systems biology is a multidisciplinary and interdisciplinary field of research, in which comprehensive profiles at the genetic, transcriptomic, proteomic, and metabolic levels are integrated and used to characterize the properties of the system and to individuate key elements playing pivotal roles in the phenomena under investigation. The knowledge gained by using such *top-down* approach ([Oltvai & Barabási, 2002](#)) is complemented by a *bottom-up* characterization of the system, which starts from a detailed molecular and biochemical knowledge of certain biological mechanisms and aims at creating mathematical models that can reproduce experimental data ([Oltvai & Barabási, 2002](#)). Ideally, systems biology works within an interactive cycle where results from *top-down* analysis are used to inform and refine mathematical predictive models, which are in turn used to generate new hypotheses and experiments able to provide new information.

Based on this principle, recent years have witnessed the development of systems medicine approaches, in which clinical practice is implemented by research carried on with computational, statistical, and mathematical multiscale analysis and modeling at both the epidemiological and individual patient levels (<http://www.easym.eu>). This new paradigm of systems science and medicine opposes, or rather complements, the traditional reductionist approach (as exemplified in [Fig. 17.1](#)) and ideally will lead to the identification of mechanisms related to disease pathophysiology, selection of novel drug targets and biomarkers, and patient risk assessment.

Systems medicine, which has also been described as 4-P (predictive, preventive, personalized, and participatory) medicine ([Hood et al., 2012](#)), is now made possible (or it is now possible to move toward) by the recent changes from data-poor (i.e., just patient data) to data-rich (i.e., molecular characterization at different levels and time resolution) applications which followed the *omics* revolution, that is the advent of high-throughput genomics, transcriptomics, metabolomics and all other *omics* disciplines ([Noble, 2008](#)). These technologies allow individuals to have relevant portions of their genomes sequenced, and multiparameter informative molecular diagnostics via blood analysis collected in a timely, standardized, and reproducible way, information that can be correlated with genetic variations. These multilevel data integration and analysis will allow the determination of a probabilistic future health history for each individual ([Oltvai & Barabási, 2002](#)), that will account not only for susceptibility to disease but also for response to drug therapies and treatments.

The overarching goal of systems medicine is to provide patients with treatments that are tailored based on their overall *omics* profiles ([Everett, 2015](#)). Since metabolites are the endpoint output of the genome, the profiles of variation of their concentrations and the patterns of metabolite–metabolite relationships offer a dynamic and actual information of the status of a system ([Rosato et al., 2018](#)).

**FIGURE 17.1**

Differences between reductionism and systems science, when analyzing the properties of a system (Tillmann et al., 2015).

From Tillmann, T., Gibson, A. R., Scott, G., Harrison, O., Dominiczak, A., & Hanlon, P. (2015). *Systems medicine 2.0: Potential benefits of combining electronic health care records with systems science models*. Journal of Medical Internet Research, 17(3), e3082. <https://doi.org/10.2196/jmir.3082>, Licensed under Creative Commons.

Recent estimations put at ~19,000 the number of metabolites expected to be found in the human body (Wishart et al., 2013), including both *endogenous* metabolites, which are gene-derived, and *exogenous* metabolites, which are environmentally derived (for instance, those derived from food intake, drugs and medicines, and microbiota), that can be used to monitor health status and diagnosed disease (Clish, 2015).

There is a long tradition for the use of metabolites to diagnose pathological conditions and impact clinical care: remarkable historical examples are blood glucose tests as screening tools for diabetes (Clarke & Foster, 2012), the use of phenylalanine to screen for phenylketonuria in newborns (Scriver, 1998) and the evaluation of the proline to citrulline ratio to diagnose ornithine- δ -aminotransferase deficiency in young children (de Sain-van der Velden et al., 2012).

Given the amount of information that is contained and that can be extracted from metabolites, it is clear that among all *omics* disciplines, metabolomics is probably the best placed to enable personalized medicine (Wishart, 2016). In addition, since metabolomics analysis allows a rapid and accurate measurement of thousands of metabolites and requires minimal sample preparation, NMR (nuclear magnetic resonance) spectroscopy on blood and tissues samples has a

direct application in the operating room, where it can provide surgeons and clinicians with real-time information (Nicholson et al., 2012a).

Individual phenotyping using nuclear magnetic resonance

Humans exhibit great phenotypic diversity, which originates from the complex interplay of genetic, epigenetic and environmental factors (Moosavi & Ardekani, 2016), and this diversity affects both disease manifestation and response to therapy. The intra-individual variability is estimated to be several folds lower than the inter-individual one, and these patterns of variability emerge at different levels, ranging from gene expression (Hughes et al., 2015) to the response of the circadian system to light (Phillips et al., 2019), from brain structure and morphology (Llera et al., 2019) to physical activity (Levin et al., 1999). This makes possible to distinguish one individual from another at different omics levels, enabling the application of personalized medicine approaches.

Small molecules of biological interest contain hydrogen atoms that can be measured through NMR spectroscopy. Thanks to the high sensitivity of the abundant ^1H nuclei, each proton (or group of equivalent protons) provides a distinctive signal, whose intensity is related to the relative concentration (and multiplicity) of that chemical species in the sample (Takis et al., 2019). NMR can measure a relatively small portion of the human metabolome, composed of all free small molecules present in concentrations $\geq 1 \mu\text{M}$. Even if the sensitivity of NMR is low compared to MS, ^1H NMR spectra of biosamples are crowded and full of chemical information (Takis et al., 2019; Vignoli et al., 2019a). When standardized pre-analytical and analytical procedures are implemented (Bernini et al., 2011a), and appropriate statistical modeling and analysis are applied, this information can be used to profile the unique metabolic fingerprints of an individual contained in different biofluids and tissues.

The first observation that humans possess an individual and recognizable metabolomic fingerprint dates back to 2005, when Assfalg et al. (2008), collecting and analyzing ≈ 40 urine samples from 22 healthy individuals of both sexes, showed the possibility of recognizing a subject within a group of individuals with almost a 100% probability (without imposing any treatment or diet) from the small molecules extracted from urines by means of ^1H NMR spectroscopy (see Fig. 17.2A).

Subsequent studies showed that this individual fingerprint is stable over a period of at least 2 (Bernini et al., 2009) to 10 years (Ghini et al., 2015).

Consistently with what observed in urine, metabolomics-based individual phenotypes have also been described in other biofluids. Wallner-Liebniann et al. (2016) showed the existence of an individual metabolic phenotype in saliva, which is able to discriminate among subjects with an accuracy ranging from 93.1% to 87.7% over 23 subjects (see Fig. 17.3A), that is only slightly lower than the one observed in the case of the individual urinary phenotype (Assfalg et al.,

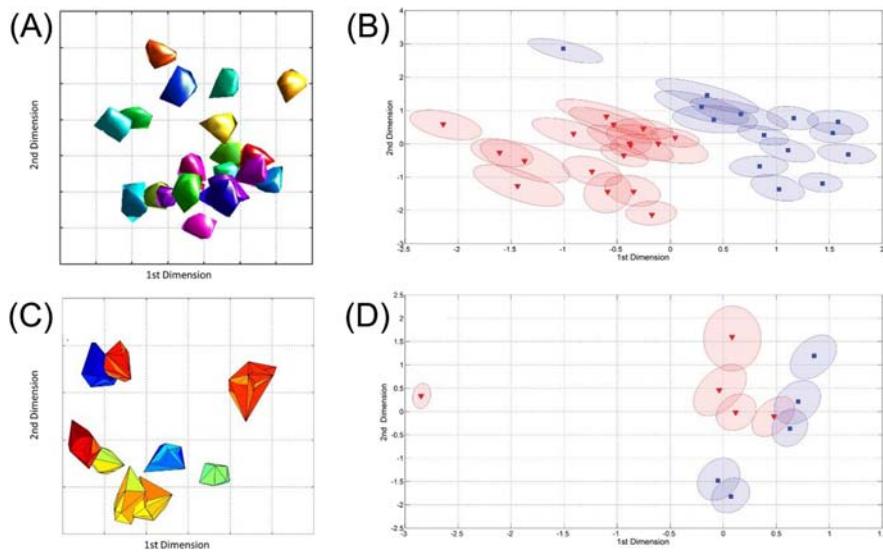


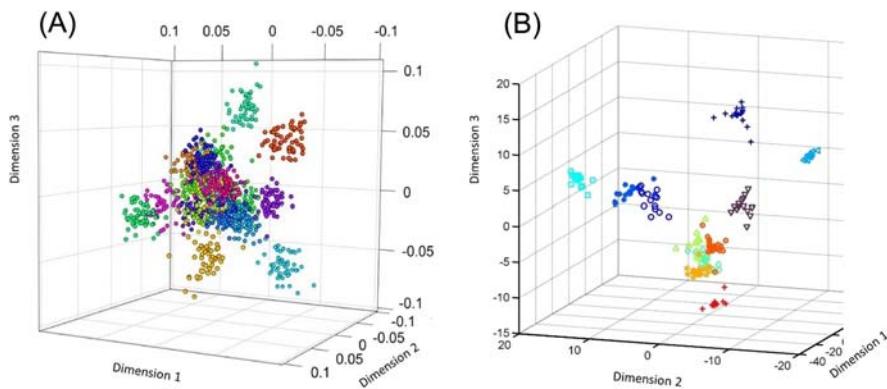
FIGURE 17.2

Modeling of the individual metabolic phenotype for two species. (A) Projection of the one-dimensional ¹H NMR spectra into PCA/CA subspace in the three most significant predictive dimensions for 22 human donors. (B) Two-dimensional plot of the multilevel model for the human static phenotype. (C) Projection of the one-dimensional ¹H NMR spectra into PCA/CA subspace in the three most significant predictive dimensions for 10 *Macaca mulatta*. (D) Two-dimensional plot of the multilevel model for the *M. mulatta* static phenotype. Each donor is color coded (Panels A, C and D). Male subjects are color coded in blue (■), female subjects is color coded in red (▼).

(A) Reproduced with permission from Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schäfer, H., Schütz, B., & Spraul, M. (2008). Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), 1420–1424. <https://doi.org/10.1073/pnas.0705685105>. Copyright (2008) National Academy of Sciences, USA; (B–D) Reproduced from Saccenti, E., Tenori, L., Verbruggen, P., Timmerman, M. E., Bouwman, J., van der Greef, J., Luchinat, C., & Smilde, A. K. (2014). Of monkeys and men: A metabolomic analysis of static and dynamic urinary metabolic phenotypes in two species. *PLoS One*, 9(9), e106077, from under Creative Commons Attribution (CC BY) license.

2008; Bernini et al., 2009), probably because saliva is more homogeneous among subjects due to homeostasis (Aure et al., 2015).

A stable and predictive individual phenotype was observed in exhaled breath of healthy donors by Martinez-Lozano Sinues et al. (2013) who reported individual classification accuracy ranging from 62% to 92%. All these studies have shown that the individual metabolic fingerprint, at least with regard to those reflected in the urine, saliva and breath metabolome is overall stable, although day by day variations can be observed.

**FIGURE 17.3**

Evidence of individual phenotype in different biofluids. (A) Projection of ^1H NMR saliva spectra from 23 donors onto the three most significant dimensions of the PCA/CA subspace for each subject. (B) Projection of breath mass spectra from 11 donors onto the first three dimensions obtained by supervised Kruskal–Wallis/PCA/CA. Each dot represents a sample. Individuals are represented by different colors and symbols.

(A) Reprinted with permission from Wallner-Liebniann, S., Tenori, L., Mazzoleni, A., Dieber-Rotheneder, M., Konrad, M., Hofmann, P., Luchinat, C., Turano, P., & Zatloukal, K. (2016). Individual human metabolic phenotype analyzed by $\text{H}-1$ NMR of saliva samples. *Journal of Proteome Research*, 15(6), 1787–1793. <https://doi.org/10.1021/acs.jproteome.5b01060>. Copyright 2016 American Chemical Society; (B) Reproduced from Martinez-Lozano Sinues, P., Kohler, M., & Zenobi, R. (2013). Human breath analysis may support the existence of individual metabolic phenotypes. *PLoS One*, 8(4), e59909. <https://doi.org/10.1371/journal.pone.0059909> under Creative Commons Attribution (CC BY) license.

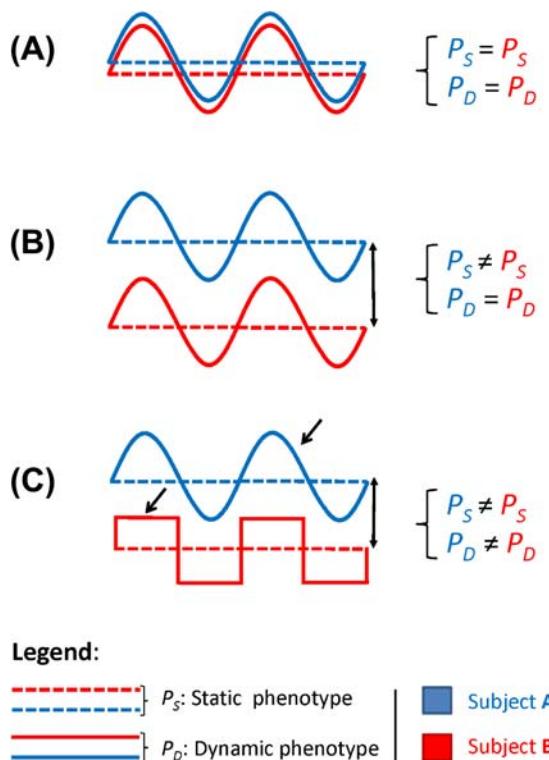
Saccenti et al. (2014) proposed the partitioning of the individual metabolic phenotype P into a static (i.e., at the level of the average metabolite concentration) part P_S and a dynamic part P_D (i.e., concerning their variation over time) as (see Fig. 17.4)

$$P = P_S + P_D \quad (17.1)$$

Further assuming that both the static and the dynamic parts are given by the contribution of intrinsic (I) (such as genetic variation) and extrinsic (E) (such as diet habits, lifestyle, and environmental conditions) factors, plus a residual part R not accounted by the model:

$$\begin{cases} P_S = I_S + E_S + I_S \times E_S + R_S \\ P_D = I_D + E_D + I_D \times E_D + R_D \end{cases} \quad (17.2)$$

Using a cohort of 31 healthy subjects, they showed that for *Homo sapiens*, 24% of the observed difference among subject-specific metabolic phenotypes is attributable to static variability, that is P_S (see Fig. 17.2B), while 76% is due to differences in the dynamic variation, that is P_D . Contextually they established the

**FIGURE 17.4**

Partitioning of the individual metabolic phenotype in static and dynamic part. The dashed lines indicate the average concentration (of a metabolite), that is taken as a proxy for the static (P_S) part of the (in this case mono-dimensional) metabolic phenotype. The solid lines indicate the time-dependent level concentration, that is the dynamic part of the metabolic phenotype (P_D). Taken together, the average concentrations of a metabolite and its modes of temporal variation make up the metabolic phenotype. Three cases are presented concerning two subjects, signified by color blue (subject 1) and red (subject 2). (A) Subjects 1 and 2 are similar with respect to both the static and dynamic phenotype. (B) Subjects 1 and 2 are similar in the dynamic phenotype but different in the static phenotype. (C) Subjects 1 and 2 are different with respect to both the static and dynamic phenotypes. The vertical double-pointed arrow (\Downarrow) indicates the difference of the average level (*dashed lines*), hence the difference of the static phenotype. The single point arrow (\downarrow) indicates the difference in the time profile shape (*solid lines*) and thus the difference of the dynamic phenotype.

Adapted from Saccetti, E., Tenori, L., Verbruggen, P., Timmerman, M. E., Bouwman, J., van der Greef, J., Luchinat, C., & Smilde, A. K. (2014). Of monkeys and men: A metabolomic analysis of static and dynamic urinary metabolic phenotypes in two species. *PLoS One*, 9(9), e106077. <https://doi.org/10.1371/journal.pone.0106077>. Reproduced under Creative Commons Attribution (CC BY) license.

existence of an individual and predictive urinary phenotype in the non-human primate rhesus macaques (*Macaca mulatta*) (see Fig. 17.2C) and showed that the 24% of the observed difference in the measured animal-specific P is attributable to P_S (see Fig. 17.2D), while 76% is due to difference in P_D , displaying a striking similarity to humans. Finally, they concluded that one quarter of the individual urinary phenotype is given by the static component and three quarters by the dynamic component (Saccenti et al., 2014):

$$\frac{P_S}{P} = \frac{1}{4} \quad \text{and} \quad \frac{P_D}{P} = \frac{3}{4}$$

The metabolic individual fingerprints originates from the complex interplay of both intrinsic and extrinsic factors. With regard to urines, it is partly determined by genetic (Bernini et al., 2009), while diet contribution is negligible (Stella et al., 2006). In addition, it seems to be significantly affected by the gut microbiome, which is involved in the regulation of multiple host metabolic pathways, signaling, and immune-inflammatory axes connecting gut, liver, muscle, and brain (Nicholson et al., 2012b) and whose dysregulation has been associated with obesity (Ley et al., 2006), diabetes (Qin et al., 2012), autoimmune diseases (Cerf-Bensussan & Gaboriau-Routhiau, 2010), and neuropsychiatric disorders (Mayer et al., 2014). An in-depth analysis of the factors shaping the individual saliva and breath phenotypes is missing, but it is not difficult to imagine an important role of the microbiome (Belstrøm, 2020).

The stability and discriminative power of the individual phenotype reflect its allostasis and resilience (Ghini et al., 2015). Allostasis indicates the adaptation process of a system to physical, physiological, and environmental challenges or stress conditions through the fluctuation of biochemical/physiological parameters (Karlamangla et al., 2002; Logan & Barksdale, 2008). The allostatic responses are the physiological changes that occur in response to external perturbations or stimuli. The resilience is the ability of the system to respond to physiological stress by reducing the risk of harm and restoring a (possibly adapted) equilibrium.

While the allostatic properties of the individual phenotype are reflected in the observed day by day variation (i.e., from its dynamic part), Ghini et al. (2015) described two remarkable examples of its resilience, which are shown in Fig. 17.5A. For Subjects A and B the urinary profiles collected in 2014 were markedly different from those collected in the previous year(s) and this resulted in a lowered discrimination accuracy of 5% and 50% instead of the >95% observed using data from previous years. The reason was found in Subject A to have developed a transient lactose intolerance possibly induced by antibiotic treatment and in Subject B to be breastfeeding at the time of collection: when these two conditions stopped, a renormalization of the metabolic profile was observed (Fig. 17.5B).

All the evidence presented indicates that a metabolomics-based representation of the individual phenotype is an accurate proxy of the healthy (or pathological) state of a person and that it can be used as a tool for metabolomics-based personalized medicine since a physio-pathological status can be directly linked to one's

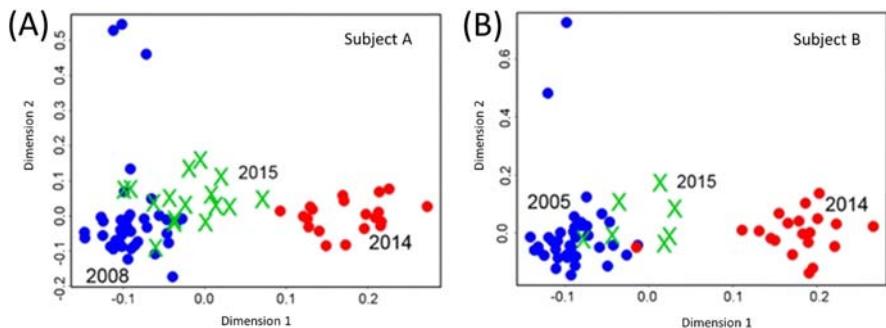


FIGURE 17.5

Resilience of the individual human urinary metabolic phenotype represented by predictive clustering of NMR urine spectra. For both subjects the individual phenotype drifted considerably between 2008 (blue dots) and 2015 (red dots) collections due to (A) antibiotic therapy and subsequent transient lactose intolerance and (B) breastfeeding. When those conditions stopped a renormalization of the metabolic profiles was observed: the 2015 (green crosses \times) phenotypes are drifting back toward the 2008 phenotype.

Adapted and reprinted with permission from Ghini, V., Saccenti, E., Tenori, L., Assfalg, M., & Luchinat, C. (2015). Allostasis and resilience of the human individual metabolic phenotype. *Journal of Proteome Research*, 14(7), 2951–2962. <https://doi.org/10.1021/acs.jproteome.5b00275>, Copyright 2015 American Chemical Society.

individual deviation(s) from his or her own reference metabolic space. This can enable the definition of metabolic fingerprints or biomarkers (or set of) that can be used for tailored prognostic, diagnostic, and treatment and applied to monitor disease progression, treatment efficacy, predisposition to drug-related side effects, and potential relapse (Koen et al., 2016).

Applications

The discovery of the existence of an individual metabolic fingerprint opens the door to the possibility of monitoring the health status of an individual during the time, by analyzing, at the systemic level, alterations in the metabolome, which are known to correlate with pathological states. By doing this, in principle, it is possible to understand if an individual is following a metabolic path of healthy ageing or if he is deviating in the direction of developing a specific disease. Of course, to make this strategy successful, it is necessary to compile a rich database of clear disease signatures. Over the last two decades, many different papers have exploited metabolomics to characterize diseases, with the intent of discovering new biomarkers and identifying biochemical pathways involved in disease pathogenesis. Just to mention very few examples, metabolomics has already increased our understanding of cellular and physiological metabolism in colorectal cancer

(Bertini et al., 2012a; Nannini et al., 2020; Turano, 2014), breast cancer (BC) (Hart et al., 2016; McCartney et al., 2017, 2018; Tenori et al., 2015; Vignoli et al., 2020a), pulmonary diseases (Bertini et al., 2013; Montuschi et al., 2018; Vignoli et al., 2020b), cardiovascular diseases (Bernini et al., 2011b; Saccenti et al., 2015; Tenori et al., 2013; Vignoli et al., 2019c, 2020c), rheumatic diseases (Vignoli et al., 2017), Down syndrome (Antonaros et al., 2020; Caracausi et al., 2018), celiac disease (Bernini et al., 2010; Bertini et al., 2008; Calabò et al., 2014; Vignoli et al., 2019b), diabetes (Dani et al., 2014; Liu et al., 2016; Mäkinen et al., 2012) obesity (Gralka et al., 2015; Ruocco et al., 2020), periodontal diseases (Aimetti et al., 2012; Romano et al., 2018, 2019).

Such sensitivity is particularly promising to monitor the individual response to illness. Often the earliest available signs of a disease are alterations of the metabolome, resulting from compensatory mechanisms which start before the disease manifestation. The detection of those early signs potentially allows an efficient prevention, in a vision of an increasingly personalized medicine. Indeed, the customization of subjects' therapeutic treatments according to their specific omic profiles/fingerprints is the basis for a future new paradigm in personalized medicine and in prevention, allowing us to really switch from the classical reactive medicine to a true predictive and preventive medicine (Bertini et al., 2012b).

For some diseases, the evidence of the workability of this approach is already established, despite further validations in larger multicentric cohorts are still needed. For example, celiac disease is characterized by several alterations in the metabolic profile of patients with respect to healthy controls, especially for what concerns energy and ketone body metabolism, as well as gut microbiota alterations (Bertini et al., 2008). Interestingly, potential celiac patients (i.e., those asymptomatic individuals with positive antibody test for CD but without any evidence of intestinal damage) present the same metabolic alterations of overt patients, suggesting that the metabolic response precedes the manifestation of the disease (Bernini et al., 2010), even when the clinical signs are not evident.

Similarly, heart failure patients and healthy controls can be discriminate by using NMR spectra of serum samples. The differences are mainly due to higher serum levels of tyrosine, phenylalanine, creatine, isoleucine, and lower serum levels of lactate, lysine, citrate, and L-dopa in patients. Strikingly, this characteristic metabolic fingerprint of heart failure is largely independent of the clinical severity of the disease, with the presence of a heart failure fingerprint in almost asymptomatic subjects. Again, metabolomics seems to be able to reveal the presence of the disease, even if the full symptoms have not yet emerged.

Acute myocardial infarction is another cardiovascular disease that continues to be challenging despite the considerable efforts made by clinicians and researchers. Metabolomic profiling approaches seem able to provide a further tool for risk stratification and management of patients. In particular, it was shown that a metabolomic-based prognostic risk model predicted death during 2 years of follow-up after acute myocardial infarction with 72.6% accuracy in the validation set (Vignoli et al., 2019c). The metabolite connectivity of patients who survived

at 2 years shows significant differences in the patterns of several low-molecular-weight molecules, implying variations in the regulation of several metabolic pathways regarding branched-chain amino acids, alanine, creatinine, mannose, ketone bodies, and energetic metabolism (Vignoli et al., 2020c).

Cancer is one of the most studied pathologies in metabolomics. The stage and the location of the tumors may differently affect the metabolome (Palmas & Vogel, 2013) of cancer patients, who exhibit significant differences in their metabolic profiles when compared to healthy controls and patients with benign diseases. For instance, metastatic colorectal cancer patients and healthy controls were clearly discriminated by multivariate statistical analysis of the serum NMR profiles: patients presented various metabolic alterations regarding energy metabolism and inflammatory response (Bertini et al., 2012a). Interestingly, it was demonstrated the possibility to stratify patients according to their overall survival, with an hazard ratio of 3.37: a far better performance with respect to traditional clinical markers (Bertini et al., 2012a).

Breast cancer (BC) is a complex and heterogeneous disease which has been extensively characterized through many platforms such as clinicopathological risk factors and various -omics techniques, including genomics and metabolomics (Asiago et al., 2010; Jobard et al., 2017; McCartney et al., 2018). In the precision medicine era, however, development of tailored oncological treatments for BC and accurate instruments to discern between patients with early-BC at high risk of disease recurrence, and those who need to be cured by locoregional therapy are missing (McCartney et al., 2017). NMR serum metabolomics could contribute significantly to this aim: in a first single center pilot study, relapse was predicted with quite good accuracy in both training set (90% sensitivity, 67% specificity, and 73% predictive accuracy, AUC (Area Under the Curve): 0.863), and validation set (82% sensitivity, 72% specificity, and 75% predictive accuracy, AUC 0.824). Tenori et al. (2015) The results have been reproduced in a multicenter study by analyzing 699 serum samples collected in the framework of an international phase III clinical trial (Hart et al., 2016). In the training set, early-BC were discriminated from metastatic BC with a discrimination accuracy of 84.9% and this model predicts disease recurrence in the validation set with an AUC of 0.747 (Hart et al., 2016).

Furthermore, the combination of the metabolomic-derived risk recurrence score with the Oncotype-DX 21-gene expression assay risk recurrence score improves the risk stratification in BC, demonstrating that metabolomics, bearing individual information, can be used to refine other kind of prognostic models (McCartney et al., 2019).

Recently, immunotherapy has presented new opportunities to fight cancer; unfortunately, not all patients respond to these therapies. Thus, in the view of more personalized treatments, a better patient selection, as well as the identification of predictive biomarkers of treatment efficacy, are of paramount importance. In this framework, metabolomics was used with the aim of selecting patients with non-small cell lung cancer who will respond to immune checkpoint inhibitors.

Interestingly, the metabolomic fingerprint of serum samples, collected before therapy, can act as a predictive tool for treatment efficacy (Ghini et al., 2020). Of course, the prospective identification of subjects that will benefit from immunotherapy could improve patient stratification, thus optimizing the treatment and avoiding unsuccessful strategies.

On the same line, metabolomics was employed to integrate genomic data in order to identify markers of unfavorable efficacy/safety profile in patients treated with phosphodiesterase inhibitors (Rocca et al., 2020).

From these few examples it emerges that ideally, in the near future of personalized medicine, individuals will be periodically screened for their individual metabolic fingerprint (Bertini et al., 2012b), permitting the monitoring of the changes of the metabolome and the deviations from the individual healthy baseline. This will allow clinicians to follow the development of diseases (even before any clinical evidence) and response to therapies. Of course, genomic, and other –omic information need to be also integrated to have a full personalized picture, and new tools for knowledge discovery and data mining need to be developed (Cacciatore et al., 2014, 2017; Saccenti & Camacho, 2020; Saccenti et al., 2014).

To practically implement these ideas, it is also important to develop standard operating procedures for each biofluid of interest, allowing experimental reproducibility and easy sample and data exchange for all the steps involved in the process of data generation, including the preanalytical phase (Bernini et al., 2011a), the analytical phase (Emwas et al., 2015), and the data analysis phase (Salek et al., 2015). The role of biobanks, that is, research infrastructures that collect, store, and redistribute biological samples and associated data (Carotenuto et al., 2015; Marcon & Nincheri, 2014), is also of key importance in personalized medicine because they can facilitate -omic-based research studies, assuring availability and quality of the biological material. With this respect, the BBMRI-ERIC European Infrastructure (<https://www.bbmri-eric.eu/>) is a precious initiative to gather all the main players from the biobanking field to boost biomedical research.

To conclude, unifying knowledge from metabolomics, other omic-sciences, medicine, information and communications technologies, and biobanking, the ultimate goal of making new personalized medicine and treatments, based on individual molecular features, seems ultimately attainable.

References

- Ahn, A. C., Tewari, M., Poon, C.-S., & Phillips, R. S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Medicine*, 3(6), e208. Available from <https://doi.org/10.1371/journal.pmed.0030208>.
- Aimetti, M., Cacciatore, S., Graziano, A., & Tenori, L. (2012). Metabonomic analysis of saliva reveals generalized chronic periodontitis signature. *Metabolomics: Official Journal of the Metabolomic Society*, 8(3), 465–474. Available from <https://doi.org/10.1007/s11306-011-0331-2>.

- Antonaros, F., Ghini, V., Pulina, F., Ramacieri, G., Cicchini, E., Mannini, E., Martelli, A., Feliciello, A., Lanfranchi, S., Onnivello, S., Vianello, R., Locatelli, C., Cocchi, G., Pelleri, M. C., Vitale, L., Strippoli, P., Luchinat, C., Turano, P., Piovesan, A., & Caracausi, M. (2020). Plasma metabolome and cognitive skills in down syndrome. *Scientific Reports*, 10(1). Available from <https://doi.org/10.1038/s41598-020-67195-z>.
- Asiago, V. M., Alvarado, L. Z., Shanaiah, N., Gowda, G. A. N., Owusu-Sarfo, K., Ballas, R. A., & Raftery, D. (2010). Early detection of recurrent breast cancer using metabolite profiling. *Cancer Research*, 70(21), 8309–8318. Available from <https://doi.org/10.1158/0008-5472.CAN-10-1319>.
- Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schäfer, H., Schütz, B., & Spraul, M. (2008). Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), 1420–1424. Available from <https://doi.org/10.1073/pnas.0705685105>.
- Aure, M. H., Konieczny, S. F., & Ovitt, C. E. (2015). Salivary gland homeostasis is maintained through acinar cell self-duplication. *Developmental Cell*, 33(2), 231–237. Available from <https://doi.org/10.1016/j.devcel.2015.02.013>.
- Belstrøm, D. (2020). The salivary microbiota in health and disease. *Journal of Oral Microbiology*, 12(1). Available from <https://doi.org/10.1080/20002297.2020.1723975>.
- Bernini, P., Bertini, I., Calabrò, A., la Marca, G., Lami, G., Luchinat, C., Renzi, D., & Tenori, L. (2010). Are patients with potential celiac disease really potential? The answer of metabolomics. *Journal of Proteome Research*, 10(2), 714–721. Available from <https://doi.org/10.1021/pr100896s>.
- Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schaefer, H., Schuetz, B., Spraul, M., & Tenori, L. (2009). Individual human phenotypes in metabolic space and time. *Journal of Proteome Research*, 8(9), 4264–4271. Available from <https://doi.org/10.1021/pr900344m>.
- Bernini, P., Bertini, I., Luchinat, C., Nincheri, P., Staderini, S., & Turano, P. (2011a). Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. *Journal of Biomolecular NMR*, 49(3–4), 231–243. Available from <https://doi.org/10.1007/s10858-011-9489-1>.
- Bernini, P., Bertini, I., Luchinat, C., Tenori, L., & Tognaccini, A. (2011b). The cardiovascular risk of healthy individuals studied by NMR metabolomics of plasma samples. *Journal of Proteome Research*, 10(11), 4983–4992. Available from <https://doi.org/10.1021/pr200452j>.
- Bertini, I., Cacciato, S., Jensen, B. V., Schou, J. V., Johansen, J. S., Kruhøffer, M., Luchinat, C., Nielsen, D. L., & Turano, P. (2012a). Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. *Cancer Research*, 72(1), 356–364. Available from <https://doi.org/10.1158/0008-5472.CAN-11-1543>.
- Bertini, I., Calabrò, A., De Carli, V., Luchinat, C., Nepi, S., Porfirio, B., Renzi, D., Saccenti, E., & Tenori, L. (2008). The metabonomic signature of celiac disease. *Journal of Proteome Research*, 8(1), 170–177. Available from <https://doi.org/10.1021/pr800548z>.
- Bertini, I., Luchinat, C., Miniati, M., Monti, S., & Tenori, L. (2013). Phenotyping COPD by 1H NMR metabolomics of exhaled breath condensate. *Metabolomics: Official Journal of the Metabolomic Society*, 10(2), 302–311. Available from <https://doi.org/10.1007/s11306-013-0572-3>.

- Bertini, I., Luchinat, C., & Tenori, L. (2012b). Metabolomics for the future of personalized medicine through information and communication technologies. *Personalized Medicine*, 9(2), 133–136. Available from <https://doi.org/10.2217/pme.11.101>.
- Cacciatore, S., Luchinat, C., & Tenori, L. (2014). Knowledge discovery by accuracy maximization. *Proceedings of the National Academy of Sciences of the United States of America*, 111(14), 5117–5122. Available from <https://doi.org/10.1073/pnas.1220873111>.
- Cacciatore, S., Tenori, L., Luchinat, C., Bennett, P. R., & MacIntyre, D. A. (2017). KODAMA: An R package for knowledge discovery and data mining. *Bioinformatics (Oxford, England)*, 33(4), 621–623. Available from <https://doi.org/10.1093/bioinformatics/btw705>.
- Calabrò, A., Antonio., Gralka, E., Luchinat, C., Saccenti, E., & Tenori, L. (2014). A metabolomic perspective on celiac disease. *Autoimmune Diseases*, 2014, e756138. Available from <https://doi.org/10.1155/2014/756138>.
- Caracausi, M., Ghini, V., Locatelli, C., Mericio, M., Piovesan, A., Antonaros, F., Pelleri, M. C., Vitale, L., Vacca, R. A., Bedetti, F., Mimmi, M. C., Luchinat, C., Turano, P., Strippoli, P., & Cocchi, G. (2018). Plasma and urinary metabolomic profiles of down syndrome correlate with alteration of mitochondrial metabolism. *Scientific Reports*, 8 (1), 2977. Available from <https://doi.org/10.1038/s41598-018-20834-y>.
- Carotenuto, D., Luchinat, C., Marcon, G., Rosato, A., & Turano, P. (2015). The Da Vinci European BioBank: A metabolomics-driven infrastructure. *Journal of Personalized Medicine*, 5(2), 107–119. Available from <https://doi.org/10.3390/jpm5020107>.
- Cerf-Bensussan, N., & Gaboriau-Routhiau, V. (2010). The immune system and the gut microbiota: Friends or foes? *Nature Reviews. Immunology*, 10(10), 735–744. Available from <https://doi.org/10.1038/nri2850>.
- Clarke, S. F., & Foster, J. R. (2012). A history of blood glucose meters and their role in self-monitoring of diabetes mellitus. *British Journal of Biomedical Science*, 69(2), 83–93.
- Clish, C. B. (2015). Metabolomics: An emerging but powerful tool for precision medicine. *Cold Spring Harbor Molecular Case Studies*, 1(1), a000588. Available from <https://doi.org/10.1101/mcs.a000588>.
- Dani, C., Bresci, C., Berti, E., Ottanelli, S., Mello, G., Mecacci, F., Breschi, R., Hu, X., Tenori, L., & Luchinat, C. (2014). Metabolomic profile of term infants of gestational diabetic mothers. *The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, 27(6), 537–542. Available from <https://doi.org/10.3109/14767058.2013.823941>.
- de Sain-van der Velden, M. G. M., Rinaldo, P., Elvers, B., Henderson, M., Walter, J. H., Prinsen, B. H. C. M. T., Verhoeven-Duif, N. M., de Koning, T. J., & van Hasselt, P. (2012). *The proline/citrulline ratio as a biomarker for OAT deficiency in early infancy*. Berlin, Heidelberg: Springer. Available from https://link.springer.com/chapter/10.1007%2F8904_2011_122, Accessed 04.01.21.
- Emwas, A.-H., Luchinat, C., Turano, P., Tenori, L., Roy, R., Salek, R. M., Ryan, D., Merzaban, J. S., Kaddurah-Daouk, R., Zeri, A. C., Nagana Gowda, G. A., Raftery, D., Wang, Y., Brennan, L., & Wishart, D. S. (2015). Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: A review. *Metabolomics*, 11(4), 872–894. Available from <https://doi.org/10.1007/s11306-014-0746-7>.

- Everett, J. R. (2015). Pharmacometabonomics in humans: A new tool for personalized medicine. *Pharmacogenomics*, 16(7), 737–754. Available from <https://doi.org/10.2217/pgs.15.20>.
- Ghaemi, S. N. (2015). Biomedical reductionist, humanist, and biopsychosocial models in medicine. In T. Schramme, & S. Edwards (Eds.), *Handbook of the philosophy of medicine* (pp. 1–19). Dordrecht: Springer. Available from https://doi.org/10.1007/978-94-017-8706-2_38-1.
- Ghini, V., Laera, L., Fantechi, B., del Monte, F., Benelli, M., McCartney, A., Tenori, L., Luchinat, C., & Pozzessere, D. (2020). Metabolomics to assess response to immune checkpoint inhibitors in patients with non-small-cell lung cancer. *Cancers*, 12(12), 3574. Available from <https://doi.org/10.3390/cancers12123574>.
- Ghini, V., Saccenti, E., Tenori, L., Assfalg, M., & Luchinat, C. (2015). Allostasis and resilience of the human individual metabolic phenotype. *Journal of Proteome Research*, 14 (7), 2951–2962. Available from <https://doi.org/10.1021/acs.jproteome.5b00275>.
- Gralka, E., Luchinat, C., Tenori, L., Ernst, B., Thurnheer, M., & Schultes, B. (2015). Metabolomic fingerprint of severe obesity is dynamically affected by bariatric surgery in a procedure-dependent manner. *American Journal of Clinical Nutrition*, 102(6), 1313–1322. Available from <https://doi.org/10.3945/ajcn.115.110536>.
- Hart, C. D., Vignoli, A., Tenori, L., Uy, G. L., To, T. V., Adebamowo, C., Hossain, S. M., Biganzoli, L., Risi, E., Love, R. R., Luchinat, C., & Leo, A. D. (2016). Serum metabolomic profiles identify ER-positive early breast cancer patients at increased risk of disease recurrence in a multicenter population. *Clinical Cancer Research*. Available from <https://doi.org/10.1158/1078-0432.CCR-16-1153>.
- Hood, L., Balling, R., & Auffray, C. (2012). Revolutionizing medicine in the 21st century through systems approaches. *Biotechnology Journal*, 7(8), 992–1001. Available from <https://doi.org/10.1002/biot.201100306>.
- Hughes, D. A., Kircher, M., He, Z., Guo, S., Fairbrother, G. L., Moreno, C. S., Khaitovich, P., & Stoneking, M. (2015). Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biology*, 16(1), 54. Available from <https://doi.org/10.1186/s13059-015-0627-z>.
- Jobard, E., Trédan, O., Bachelot, T., Vigneron, A. M., Aït-Oukhatar, C. M., Arnedos, M., Rios, M., Bonneterre, J., Diéras, V., Jimenez, M., Merlin, J.-L., Campone, M., & Elena-Herrmann, B. (2017). Longitudinal serum metabolomics evaluation of trastuzumab and everolimus combination as pre-operative treatment for HER-2 positive breast cancer patients. *Oncotarget*, 8(48), 83570–83584. Available from <https://doi.org/10.18632/oncotarget.18784>.
- Karlamangla, A. S., Singer, B. H., McEwen, B. S., Rowe, J. W., & Seeman, T. E. (2002). Allostatic load as a predictor of functional decline: Macarthur studies of successful aging. *Journal of Clinical Epidemiology*, 55(7), 696–710. Available from [https://doi.org/10.1016/S0895-4356\(02\)00399-2](https://doi.org/10.1016/S0895-4356(02)00399-2).
- Koen, N., Du Preez, I., & Loots, D. T. (2016). Chapter Three—Metabolomics and personalized medicine. In R. Donev (Ed.), *Advances in protein chemistry and structural biology* (102, pp. 53–78). Academic Press. Available from <https://doi.org/10.1016/bs.apcsb.2015.09.003>.
- Levin, S., Jacobs, D. R., Ainsworth, B. E., Richardson, M. T., & Leon, A. S. (1999). Intra-individual variation and estimates of usual physical activity. *Annals of Epidemiology*, 9 (8), 481–488. Available from [https://doi.org/10.1016/S1047-2797\(99\)00022-8](https://doi.org/10.1016/S1047-2797(99)00022-8).
- Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature*, 444(1476–4687 (Electronic)), 1022–1023.

- Liu, X., Gao, J., Chen, J., Wang, Z., Shi, Q., Man, H., Guo, S., Wang, Y., Li, Z., & Wang, W. (2016). Identification of metabolic biomarkers in patients with type 2 diabetic coronary heart diseases based on metabolomic approach. *Scientific Reports*, 6, 30785. Available from <https://doi.org/10.1038/srep30785>.
- Llera, A., Wolfers, T., Mulders, P., & Beckmann, C. F. (2019). Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *eLife*, 8, e44443. Available from <https://doi.org/10.7554/eLife.44443>.
- Logan, J. G., & Barksdale, D. J. (2008). Allostasis and allostatic load: Expanding the discourse on stress and cardiovascular disease. *Journal of Clinical Nursing*, 17(7B), 201–208. Available from <https://doi.org/10.1111/j.1365-2702.2008.02347.x>.
- Mäkinen, V.-P., Tynkkynen, T., Soininen, P., Peltola, T., Kangas, A. J., Forsblom, C., Thorn, L. M., Kaski, K., Laatikainen, R., Ala-Korpela, M., & Groop, P.-H. (2012). Metabolic diversity of progressive kidney disease in 325 patients with type 1 diabetes (the FinnDiane Study). *Journal of Proteome Research*, 11(3), 1782–1790. Available from <https://doi.org/10.1021/pr201036j>.
- Marcon, G., & Nincheri, P. (2014). The multispecialistic Da Vinci European BioBank. *Open Journal of Bioresources*, 1. Available from <https://doi.org/10.5334/ojb.af>.
- Martinez-Lozano Sinues, P., Kohler, M., & Zenobi, R. (2013). Human breath analysis may support the existence of individual metabolic phenotypes. *PLoS One*, 8(4), e59909. Available from <https://doi.org/10.1371/journal.pone.0059909>.
- Mayer, E. A., Knight, R., Mazmanian, S. K., Cryan, J. F., & Tillisch, K. (2014). Gut microbes and the brain: paradigm shift in neuroscience. *Journal of Neuroscience*, 34 (46), 15490–15496. Available from <https://doi.org/10.1523/JNEUROSCI.3299-14.2014>.
- McCartney, A., Vignoli, A., Biganzoli, L., Love, R., Tenori, L., Luchinat, C., & Di Leo, A. (2018). Metabolomics in breast cancer: A decade in review. *Cancer Treatment Reviews*, 67, 88–96. Available from <https://doi.org/10.1016/j.ctrv.2018.04.012>.
- McCartney, A., Vignoli, A., Hart, C., Tenori, L., Luchinat, C., Biganzoli, L., & Di Leo, A. (2017). De-escalating and escalating treatment beyond endocrine therapy in patients with luminal breast cancer. *Breast*, 34(Suppl. 1), S13–S18. Available from <https://doi.org/10.1016/j.breast.2017.06.021>.
- McCartney, A., Vignoli, A., Tenori, L., Fornier, M., Rossi, L., Risi, E., Luchinat, C., Biganzoli, L., & Di Leo, A. (2019). Metabolomic analysis of serum may refine 21-gene expression assay risk recurrence stratification. *NPJ Breast Cancer*, 5, 26. Available from <https://doi.org/10.1038/s41523-019-0123-9>.
- Montuschi, P., Santini, G., Mores, N., Vignoli, A., Macagno, F., Shoreh, R., Tenori, L., Zini, G., Fusco, L., Mondino, C., Di Natale, C., D'Amico, A., Luchinat, C., Barnes, P.J., Higenbottam, T. Breathomics for assessing the effects of treatment and withdrawal with inhaled beclomethasone/formoterol in patients with COPD. *Frontiers in Pharmacology* 2018, 9, 258. <https://doi.org/10.3389/fphar.2018.00258>.
- Moosavi, A., & Ardekani, A. M. (2016). Role of epigenetics in biology and human diseases. *Iranian Biomedical Journal*, 20(5), 246–258. Available from <https://doi.org/10.22045/ibj.2016.01>.
- Nannini, G., Meoni, G., Amedei, A., & Tenori, L. (2020). Metabolomics profile in gastrointestinal cancers: Update and future perspectives. *World Journal of Gastroenterology*, 26(20), 2514–2532. Available from <https://doi.org/10.3748/wjg.v26.i20.2514>.

- Nicholson, J. K., Holmes, E., Kinross, J. M., Darzi, A. W., Takats, Z., & Lindon, J. C. (2012a). Metabolic phenotyping in clinical and surgical environments. *Nature*, 491(7424), 384–392. Available from <https://doi.org/10.1038/nature11708>.
- Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012b). Host-gut microbiota metabolic interactions. *Science*, 336(6086), 1262–1267. Available from <https://doi.org/10.1126/science.1223813>.
- Noble, D. (2008). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, 93(1), 16–26. Available from <https://doi.org/10.1113/expphysiol.2007.038695>.
- Oltvai, Z. N., & Barabási, A.-L. (2002). Life's complexity pyramid. *Science*, 298(5594), 763–764. Available from <https://doi.org/10.1126/science.1078563>.
- Palmas, M. S. A., & Vogel, H. J. (2013). The future of NMR metabolomics in cancer therapy: Towards personalizing treatment and developing targeted drugs? *Metabolites*, 3(2), 373–396. Available from <https://doi.org/10.3390/metabo3020373>.
- Phillips, A. J. K., Vidafar, P., Burns, A. C., McGlashan, E. M., Anderson, C., Rajaratnam, S. M. W., Lockley, S. W., & Cain, S. W. (2019). High sensitivity and interindividual variability in the response of the human circadian system to evening light. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24), 12019–12024. Available from <https://doi.org/10.1073/pnas.1901824116>.
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., ... Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418), 55–60. Available from <https://doi.org/10.1038/nature11450>.
- Rocca, M. S., Vignoli, A., Tenori, L., Ghezzi, M., De Rocco Ponce, M., Vatsellas, G., Thanos, D., Padrini, R., Foresta, C., & De Toni, L. (2020). Evaluation of serum/urine genomic and metabolomic profiles to improve the adherence to sildenafil therapy in patients with erectile dysfunction. *Frontiers in Pharmacology*, 11. Available from <https://doi.org/10.3389/fphar.2020.602369>.
- Romano, F., Meoni, G., Manavella, V., Baima, G., Mariani, G. M., Cacciatore, S., Tenori, L., & Aimetti, M. (2019). Effect of non-surgical periodontal therapy on salivary metabolic fingerprint of generalized chronic periodontitis using nuclear magnetic resonance spectroscopy. *Archives of Oral Biology*, 97, 208–214. Available from <https://doi.org/10.1016/j.archoralbio.2018.10.023>.
- Romano, F., Meoni, G., Manavella, V., Baima, G., Tenori, L., Cacciatore, S., & Aimetti, M. (2018). Analysis of salivary phenotypes of generalized aggressive and chronic periodontitis through nuclear magnetic resonance-based metabolomics. *Journal of Periodontology*, 89(12), 1452–1460. Available from <https://doi.org/10.1002/JPERO.18-0097>.
- Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P. R., Martins dos Santos, V. A. P., & Saccenti, E. (2018). From correlation to causation: Analysis of metabolomics data using systems biology approaches. *Metabolomics*, 14(4), 37. Available from <https://doi.org/10.1007/s11306-018-1335-y>.
- Ruocco, C., Ragni, M., Rossi, F., Carullo, P., Ghini, V., Piscitelli, F., Cutignano, A., Manzo, E., Ioris, R. M., Bontems, F., Tedesco, L., Greco, C. M., Pino, A., Severi, I., Liu, D., Ceddia, R. P., Ponzoni, L., Tenori, L., Rizzetto, L., ... Nisoli, E. (2020). Manipulation of dietary amino acids prevents and reverses obesity in mice through multiple mechanisms that modulate energy homeostasis. *Diabetes*, 69(11), 2324–2339. Available from <https://doi.org/10.2337/db20-0489>.

- Saccenti, E., & Camacho, J. (2020). *Multivariate exploratory data analysis using component models. Comprehensive foodomics*. Elsevier. Available from <https://doi.org/10.1016/B978-0-08-100596-5.22902-8>.
- Sacenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. B. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3), 361–374. Available from <https://doi.org/10.1007/s11306-013-0598-6>.
- Sacenti, E., Suarez-Diez, M., Luchinat, C., Santucci, C., & Tenori, L. (2015). Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. *Journal of Proteome Research*, 14(2), 1101–1111. Available from <https://doi.org/10.1021/pr501075r>.
- Sacenti, E., Tenori, L., Verbruggen, P., Timmerman, M. E., Bouwman, J., van der Greef, J., Luchinat, C., & Smilde, A. K. (2014). Of monkeys and men: A metabolomic analysis of static and dynamic urinary metabolic phenotypes in two species. *PLoS One*, 9(9), e106077. Available from <https://doi.org/10.1371/journal.pone.0106077>.
- Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., Correa, E., Reijmers, T., Rosato, A., Tenori, L., Turano, P., Marin, S., Deborde, C., Jacob, D., Rolin, D., Dartigues, B., Conesa, P., Haug, K., Rocca-Serra, P., O'Hagan, S., Hao, J., van Vliet, M., Sysi-Aho, M., Ludwig, C., Bouwman, J., Cascante, M., Ebbels, T., Griffin, J. L., Moing, A., Nikolski, M., Oresic, M., Sansone, S.-A., Viant, M. R., Goodacre, R., Guenther, U. L., Hankemeier, T., Luchinat, C., Walther, D., & Steinbeck, C. (2015). Coordination of Standards in MetabOlonicS (COSMOS): Facilitating integrated metabolomics data access. *Metabolomics*, 11(6), 1598–1599. Available from <https://doi.org/10.1007/s11306-015-0822-7>.
- Schroeder, S. A. (2007). We can do better—Improving the health of the American people. *New England Journal of Medicine*, 357(12), 1221–1228. Available from <https://doi.org/10.1056/NEJMsa073350>.
- Scriven, C. C. (1998). A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants, by Robert Guthrie and Ada Susi. *Pediatrics*, 1963;32:318-343. *Pediatrics*, 102(1 Pt 2), 236–237.
- Stella, C., Beckwith-Hall, B., Cloarec, O., Holmes, E., Lindon, J. C., Powell, J., van der Ouderaa, F., Bingham, S., Cross, A. J., & Nicholson, J. K. (2006). Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research*, 5 (10), 2780–2788.
- Takis, P. G., Ghini, V., Tenori, L., Turano, P., & Luchinat, C. (2019). Uniqueness of the NMR approach to metabolomics. *TrAC—Trends in Analytical Chemistry*, 120, 115300. Available from <https://doi.org/10.1016/j.trac.2018.10.036>.
- Tenori, L., Hu, X., Pantaleo, P., Alterini, B., Castelli, G., Olivotto, I., Bertini, I., Luchinat, C., & Gensini, G. F. (2013). Metabolomic fingerprint of heart failure in humans: A nuclear magnetic resonance spectroscopy analysis. *International Journal of Cardiology*, 168(4), e113–e115. Available from <https://doi.org/10.1016/j.ijcard.2013.08.042>.
- Tenori, L., Oakman, C., Morris, P. G., Gralka, E., Turner, N., Cappadona, S., Fornier, M., Hudis, C., Norton, L., Luchinat, C., & Di Leo, A. (2015). Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study. *Molecular Oncology*, 9(1), 128–139. Available from <https://doi.org/10.1016/j.molonc.2014.07.012>.

- Tillmann, T., Gibson, A. R., Scott, G., Harrison, O., Dominiczak, A., & Hanlon, P. (2015). Systems medicine 2.0: Potential benefits of combining electronic health care records with systems science models. *Journal of Medical Internet Research*, 17(3), e3082. Available from <https://doi.org/10.2196/jmir.3082>.
- Turano, P. (2014). Colorectal cancer: The potential of metabolic fingerprinting. *Expert Review of Gastroenterology & Hepatology*, 8(8), 847–849. Available from <https://doi.org/10.1586/17474124.2014.945912>.
- Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., Turano, P., & Luchinat, C. (2019a). High-throughput metabolomics by 1D NMR. *Angewandte Chemie International Edition English*, 58(4), 968–994. Available from <https://doi.org/10.1002/anie.201804736>.
- Vignoli, A., Muraro, E., Miolo, G., Tenori, L., Turano, P., Di Gregorio, E., Steffan, A., Luchinat, C., & Corona, G. (2020a). Effect of estrogen receptor status on circulatory immune and metabolomics profiles of HER2-positive breast cancer patients enrolled for neoadjuvant targeted chemotherapy. *Cancers (Basel)*, 12(2). Available from <https://doi.org/10.3390/cancers12020314>.
- Vignoli, A., Orlandini, B., Tenori, L., Biagini, M. R., Milani, S., Renzi, D., Luchinat, C., & Calabro, A. S. (2019b). Metabolic signature of primary biliary cholangitis and its comparison with celiac disease. *Journal of Proteome Research*, 18(3), 1228–1236. Available from <https://doi.org/10.1021/acs.jproteome.8b00849>.
- Vignoli, A., Rodio, D. M., Bellizzi, A., Sobolev, A. P., Anzivino, E., Mischitelli, M., Tenori, L., Marini, F., Priori, R., Scrivo, R., Valesini, G., Francia, A., Morreale, M., Ciardi, M. R., Iannetta, M., Campanella, C., Capitani, D., Luchinat, C., Pietropaolo, V., & Mannina, L. (2017). NMR-based metabolomic approach to study urine samples of chronic inflammatory rheumatic disease patients. *Analytical and Bioanalytical Chemistry*, 409(5), 1405–1413. Available from <https://doi.org/10.1007/s00216-016-0074-z>.
- Vignoli, A., Santini, G., Tenori, L., Macis, G., Mores, N., Macagno, F., Pagano, F., Higenbottam, T., Luchinat, C., & Montuschi, P. (2020b). NMR-based metabolomics for the assessment of inhaled pharmacotherapy in chronic obstructive pulmonary disease patients. *Journal of Proteome Research*, 19(1), 64–74. Available from <https://doi.org/10.1021/acs.jproteome.9b00345>.
- Vignoli, A., Tenori, L., Giusti, B., Takis, P. G., Valente, S., Carrabba, N., Balzi, D., Barchielli, A., Marchionni, N., Gensini, G. F., Marcucci, R., Luchinat, C., & Gori, A. M. (2019c). NMR-based metabolomics identifies patients at high risk of death within two years after acute myocardial infarction in the AMI-Florence II Cohort. *BMC Medicine*, 17(1), 3. Available from <https://doi.org/10.1186/s12916-018-1240-2>.
- Vignoli, A., Tenori, L., Giusti, B., Valente, S., Carrabba, N., Baizi, D., Barchielli, A., Marchionni, N., Gensini, G. F., Marcucci, R., Gori, A. M., Luchinat, C., & Saccenti, E. (2020c). Differential network analysis reveals metabolic determinants associated with mortality in acute myocardial infarction patients and suggests potential mechanisms underlying different clinical scores used to predict death. *Journal of Proteome Research*, 19(2), 949–961. Available from <https://doi.org/10.1021/acs.jproteome.9b00779>.
- Von Bertalanffy, L. (2021). *General system theory: Foundations, development, applications* (revised edition). Penguin University Books. Available from <https://www.amazon.com/General-System-Theory-Foundations-Applications/dp/0807604534>, Accessed 01.04.21; 9780807604533.

- Wallner-Liebniann, S., Tenori, L., Mazzoleni, A., Dieber-Rotheneder, M., Konrad, M., Hofmann, P., Luchinat, C., Turano, P., & Zatloukal, K. (2016). Individual human metabolic phenotype analyzed by H-1 NMR of saliva samples. *Journal of Proteome Research*, 15(6), 1787–1793. Available from <https://doi.org/10.1021/acs.jproteome.5b01060>.
- Wiener, N. (2021). *Cybernetics or control and communication in the animal and the machine, reissue of the 1961* (2nd ed.). The MIT Press, <https://mitpress.mit.edu/books/cybernetics-or-control-and-communication-animal-and-machine-reissue-1961-second-edition> (accessed 01.04.21).
- Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7), 473–484. Available from <https://doi.org/10.1038/nrd.2016.32>.
- Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., & Scalbert, A. (2013). HMDB 3.0 – The human metabolome database in 2013. *Nucleic Acids Research*, 41(Database issue), D801–D807. Available from <https://doi.org/10.1093/nar/gks1065>.

Metabolomics in public health

18

Pierpaolo Cavallo^{1,2}

¹*Department of Physics, University of Salerno, Fisciano, Salerno, Italy*

²*Complex Systems Institute-National Research Council (ISC-CNR), Rome, Italy*

Introduction

Metabolomics is one of the postgenomic sciences, as well as proteomics and transcriptomics; however, unlike other methods, it has a true horizontal approach. Indeed, it studies all the universe of the small molecules, endogenous or exogenous, such as metabolic substrates and products, lipids, small peptides, vitamins and any other cofactor generated by metabolism. Its relevance is due to the fact that it unravels signals which are downstream from genes, and thus it looks at the biological patterns closer to the culmination of the disease process. In terms of Public Health, this horizontal approach allows its use for any kind of clinical and biological setting, and the exchange of information between the most disparate of those settings. To discuss the relationship between Metabolomics and Public Health, we can start from a group of eight thematic recommendations (Khoury et al., 2013) which emerged from an engagement of the National Cancer Institute to the scientific community for a cancer epidemiology vision in the 21st century. The eight themes are listed in synthesis as follows:

1. extending the reach of epidemiology research to include multilevel analysis;
2. transforming the practice of epidemiology by more access and data sharing;
3. expanding cohort studies to collect lifelong data and multiple endpoints;
4. developing and validating reliable quantitative methods to assess on a massive scale under a complex systems approach;
5. integrating “big data” science into the practice of epidemiology;
6. expanding knowledge integration to drive research, policy, and practice;
7. transforming training to address interdisciplinary and translational research;
8. optimizing the use of resources and infrastructure.

Many of these themes appear to be a guide to build a conceptual framework of Metabolomics in Public Health; Table 18.1 lists the themes and indicates, where appropriate, the possible points of interest, that will be discussed in this section. The theme 1 deals with the possibility to extend the reach of the

Table 18.1 Thematic recommendations.

Theme	Description	Metabolomics in public health point of interest
1	Extending the reach of epidemiology research to include multilevel analysis	Data integration
2	Transforming the practice of epidemiology by more access and data sharing	Systems biology in public health and metabolomics
3	Expanding cohort studies to collect lifelong data and multiple endpoints	Longitudinal and lifelong studies in metabolomics
4	Developing and validating reliable quantitative methods to assess on massive scale under complex systems approach	Quantitative methods are necessary
5	Integrating “big data” science into the practice of epidemiology	Big data and metabolomics in public health
6	Expanding knowledge integration to drive research, policy, and practice	Policies, training and resources
7	Transforming training to address interdisciplinary and translational research	
8	Optimizing the use of resources and infrastructure	

research by including multilevel analysis, and its reference point in Metabolomics can be considered the integration of multiple types of Omics data. The theme 2 aims at more access and data sharing, and this can be related to the role of Metabolomics in System Biology and Personalized Medicine, both connected to Public Health. The third theme concerns the expansion of cohort studies towards a lifelong data collection, and Metabolomics may have a role as instrument for longitudinal studies, in which the metabolomic agents play a role of exposure, mediator or outcome. The fourth theme is about the development of quantitative methods under a complex systems approach, and to discuss it we should take into account either the study design and the interaction between environment and genes.

The theme 5 regards the integration of Big Data science into the practice of Epidemiology, and Metabolomics is a structural provider of Big Data, that are no more becoming but have become a basis of Public Health. The sixth theme indicates the expansion of knowledge as a means to drive policies, and is consequence of the preceding themes, as well as the following seventh, regarding the necessity of transformation of the training towards translational research. The final eighth theme indicates the need to optimize the resources, and Metabolomics may become a relevant agent for this scope, as the molecular patterns that it includes are, in large part, composed by molecules whose assay can be obtained, once identified, with simple and straightforward methods.

Data integration

The integration of data is a powerful instrument of Public Health: the discipline in itself includes a large and widely differentiated horizon of sciences, for example Epidemiology, Biostatistics, Ecology and Environmental Toxicology, Policy and Administration, Sociology and Behavioral Sciences, and the integration of Omics sciences into these fields is a running process. There are different paradigms of integration ([Chu et al., 2019](#)) to be considered, that are at subject level, at analytic level and at biological level. The first one is the subject level integration, which can be vertical, studying multiple level of Omics in the same patient, and horizontal, integrating multiple sources of data on the same target; the second one, analytic, can involve a sequential approach, studying datatypes one-by-one in steps, or integrating multiple source populations; the third one, the biological inference, can be considered under two approaches, the data-driven and the knowledge-driven. In terms of Public Health, all the types of integration allow to use Metabolomics data—and also other Omics datatypes—for integration with the usual datasets, coming from the aforementioned disciplines. Epidemiological, clinical, environmental, administrative, social and behavioral data can be usefully integrated towards an holistic approach ([Pourbohloul & Kieny, 2011](#)) that makes possible to study health determining interactions at three different levels, which compose the holistic framework: societal, institutional and molecular. The interactions at the holistic framework level can be studied with Complex Systems Science tools, the main of which is the Network Analysis. The networks are everywhere, and, with the aid of the same instrument, it is possible to represent a set of molecular associations as a layer and information provided by Public Health data sources, such as environmental sensors network or an administrative General Practitioner data set, as additional layers. Once a network layer has been established, there are many approaches that can be used to extract relevant information, studying topology and dynamics, in order to allow an individual investigation of each single layer, either coming from Metabolomics data or from other external resources. However, studying multilayer networks is far more interesting: this approach ([Aleta & Moreno, 2019](#)) is an extension of the single layer networks, that can be completely explained by describing some entities called nodes or vertices and their interactions called links or edges. The multilayer networks are made by nodes and links arranged into different layers, and each layer includes peculiar aspects or features, so that the links can be described and measured as intralayer links, connecting nodes in the same layer, and interlayer or coupling links, connecting nodes belonging to different layers. The possible outcomes of data integration for Metabolomics and other Omics with Public Health data through Network Analysis, thus, may have a very large area of application ([Gosak et al., 2018](#)), joining the study of molecular and biochemical networks up to the individual/behavioral level and even the macroscopic level of biological, social or institutional systems. Moreover, the multilayer networks appear very

interesting to study network structure and function across time, to obtain a fourth, temporal dimension able “*...to connect... networks at the subcellular level with disease networks and epidemiology at the macroscopic level of the whole organism*” (Gosak et al., 2018), and the possibility of scaling up to the population and environment level is clear.

System biology and metabolomics in public health

Systems Biology considers the biological systems as complex systems and studies the physiology and the pathology of the living organisms under a holistic approach, in which the different levels, genome, epigenome, transcriptome, proteome, metabolome and phenotype, are connected through biological pathways in the structural sense, but also through environmental factors, including microbiome, in the functional sense. The involvement of environment, moreover, is mediated by the presence of humans, either through their individual and collective behavior, which influences human health, and through the impact of human behavior on the environment, which also influences human health. From another point of view, the complexity of the biological systems and the outcomes of the intricate relationships between individuals and environment, plus the interactions of individuals with one another, justify the wide differences that are usually found between the same disease or condition in different subjects and, as a result, the need of tailoring a treatment to each patient: this is the basis for Personalized Medicine. Personalized, or Precision, Medicine (Koenig et al., 2017), uses diagnostics based on molecular or cellular analysis of the patient, largely including Omics techniques, to obtain the most accurate possible classification for diagnostic, prognostic or therapeutic purposes. It is clear that genome, metabolome and any other Omics pattern show significant variation from the average in any given individual, but it is equally clear that the resolution power of Metabolomics can make a difference. Expanding the concept of complex systems use into the healthcare disciplines, Systems Biology, that has been defined as “an interdisciplinary effort to integrate molecular, cellular, tissue, organ, and organism levels of function into computational models that facilitate the identification of general principles” (Dammann et al., 2014) has been followed by Systems Medicine, focused on understanding the disease processes, and how to interfere with them, and by Systems Epidemiology. It adds multiple levels of analysis, as the temporal dimension could be also studied, for instance by using Public Health data bases to look for the antecedents of the disease processes in individuals or populations, thanks to the growing corpus of Electronic Healthcare Recordings, and to look for the outcomes, either on individual or population level, of the prevention or treatment of the disease. The step from Systems Epidemiology to a Systems Public Health is straightforward, especially if we consider the main properties of complex systems: adaptation, feedback and emergent behavior. For instance: let's consider the

observation that obesity distribution across population (Rutter et al., 2017) is an emergent property of the interactions between a group of factors, such as food, employment, transport, economic etc., and rarely a simple, single intervention is able to modify the outcome of these interactions.

In the same way, as shown in Fig. 18.1, we can consider that a disease is the emergent property of the interactions between a group of biological factors, namely the sequence genome—phenotype, with a group of environmental factors, and with a group of external factors, namely the health determinants. The emerging behavior of the system, that is the disease or condition, may be influenced by any and each of them, and the role of Public Health Metabolomics may probably be to capture these influences “at the last stage” before they emerge to the surface of the phenotype; for this purpose, as indicated before, it may be useful to integrate additional Omics (Chu et al., 2019), with different possible combinations. First: to study genome and metabolome, because the genomic indicators, within the limits of epigenomic variations, do not vary. The metabolite levels may vary by many factors, including the window of time and/or the sequence of interactions during which genetic factors might exert their effects on metabolites, and also by the presence of genetic influence, the so-called Genetically Influenced Metabotypes (Beger et al., 2016) that are the genetic predispositions which, by interacting with environmental factors and lifestyle habits through intermediate metabolic phenotypes, produce different outcomes in the pathogenesis of complex disorders. The second possible level is the integration of transcriptome and metabolome: the transcriptome is the representation of gene expression, and its evaluation allows the investigation of the interplay between metabolite and gene level, also, for multiple sampling, with changes in time.

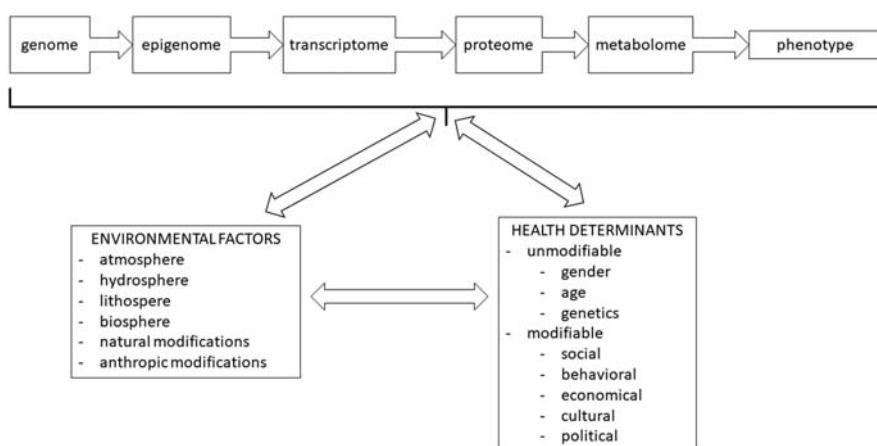


FIGURE 18.1

Interactions between molecular processes, environmental factors and health determinants.

The third level is to study the proteome and the metabolome, even if the entire set of expressed proteins tends to become very complex due to interactions between proteins, and the fourth level is the study of the microbiome and the metabolome. The latter may have a relevant interest for Public Health relationship with Metabolomics, as the microbiome, that is the microbial community that resides in an anatomical site, with its collective multi Omics features, plays a significant role in human health, and, by producing its own metabolites, it may also influence biological processes of the host organism. The interest emerges right because the microbiome—metabolome integration may take into account the processes and outcomes of the interactions of the subject with the environment, even more if we consider that gut dysbiosis has effects also on distant organs (Traversi et al., 2021). Thus, the exogenous factors, including microbes, give to Public Health scientists one more way to deepen the study of the interplay between human, environment and health determinants, mediated by the interactome and by producing the modifications depicted into the diseaseome (Barabási, 2007).

Longitudinal and life-long studies in metabolomics

Research study design represents a crucial step in Metabolomics experiments: after defining the target of inference, that is the research question, it is critical to carefully select the participants and define information about risk factors, intermediate processes, and outcomes, and then the development of the methods, planning and performing, analysis etc. will follow. Cross-sectional studies imply the collection of information about a risk factor, or exposure, at the same time with an outcome or phenotype; this approach is common (Chu et al., 2019) in Omics studies, with the collection of samples at a single time point, or even in a limited time span and with one sample collected for each subject studied. In this kind of study, usually there is a control group, which comprises subjects without the risk factor or the exposure, and the difference between cases and controls makes possible to infer the results. A different approach is the one of longitudinal studies, as the study spans over time and implies a temporal ordering of data collection: the control group may be made by the same subjects, studied before and after an exposure, or followed in time during the course of a disease or condition. In these studies, the metabolome can play different roles, as it can be considered (Chu et al., 2019) as the exposure, the mediator or the outcome of the process under observation, and this kind of approach has possible interesting implications in terms of Public Health. As the metabolome is the closest biological representation of the phenotype, it is often used as an exposure indicator, in order to predict the disease risk, and its association with environmental factors and/or health determinants can foster outcomes in Public Health studies. If the metabolome is used as the outcome, the study could be performed by addressing two or more

hypotheses, and in this case the Public Health approach could be targeted to study different clinical risk indicators, resulting from environmental factors and/or health determinants. If, finally, the metabolome is considered as a mediator of effects, the Public Health interest could be focused on its modifications following the exposure to an environmental factor and/or an health determinant. The longitudinal studies should be designed using a lifelong approach (Khoury et al., 2013), integrating the individual medical records with exposure information, and, of course, aggregating these data to obtain population-level frameworks. This kind of approach is not always possible, at least according to some normative constraints about privacy, but the possibility of building databases with existing electronic health records, linking them to Metabolomics data, gives Public Health a possible crucial role in this process. Another observation connected to the longitudinal approach is the Glocalization, that is the concept that individual, domestic characteristics of health behavior, organization and systems have interrelated impacts on global healthcare practices (Ghazvini & Shukur, 2013), and the result is an health environment in which “what happens locally has global impact, and what happens globally has local impact” (Scott et al., 2004). In terms of longitudinal studies, this means that the interaction between Metabolomics and Public Health, with regard to data sharing, will become more necessary, either in the spatial dimension and in the temporal one, for instance looking at antecedents of disease and presence of environmental, social, economic, cultural and behavioral pre-diagnostic risk factors and biomarkers.

Quantitative methods are necessary

A parallel, and unprecedented, development of a wide array of sources of quantitative data has been occurring from the end of the twentieth century. As a result, an increasing number of technologies and platforms of biomarker measurement has emerged, (Khoury et al., 2013) including all the Omics methods, but also non-coding RNA, microbiota, etc., and those for environmental, physiological, social and environmental measurements, obtained through sensor technologies, frequently incorporated into portable or wearable devices. The exposome is the environmental counterpart of the genome (Vineis et al., 2017), and it refers to the totality of environmental exposures from conception onwards. Many diseases, especially the chronic, degenerative and/or neoplastic ones, come from multifactorial interactions, in which the effects of environmental exposures interact with human genetics, and the interaction is mediated by the “Omics cascade” represented in Fig. 18.1. To complicate the framework, the outcomes of the interactions show variations, sometimes large, in reference to location and local characteristics, as climate, lifestyle, social, economic and cultural aspects, diet, occupation etc. The exploration of the exposome may be able to assess the impact of the exposures, provided that there is careful attention to the methodological

aspects, to the ability of store data and samples, and also to the main sources of exposome data (Vineis et al., 2017), the external and the internal ones. The external exposome is mainly a matter of Public Health: disease risk is assessed by measuring air and water pollutants, biosensors, satellite data, personal and wearable sensors, etc., and the integration of these data with those deriving from Electronic Health Recordings may provide quantitative and reliable data Fig. 18.2.

The internal exposome is mainly a matter of laboratory, with gathering of biological, biochemical and molecular data, through which the Metabolomics can play a significant role, because metabolome is—as already said—the last step of the cascade before the phenotype. Again, the problem will be to integrate multiple data from multiple platforms, even if a group of possibly greater challenges could be to avoid redundancies, to find an effective way of moving from proof-of-concept to wide-scale validation and translational medicine, and to defuse the possible problems resulting from ethical and legal issues. The possible interactions between environment and genes can be considered (Zeisel, 2007) under

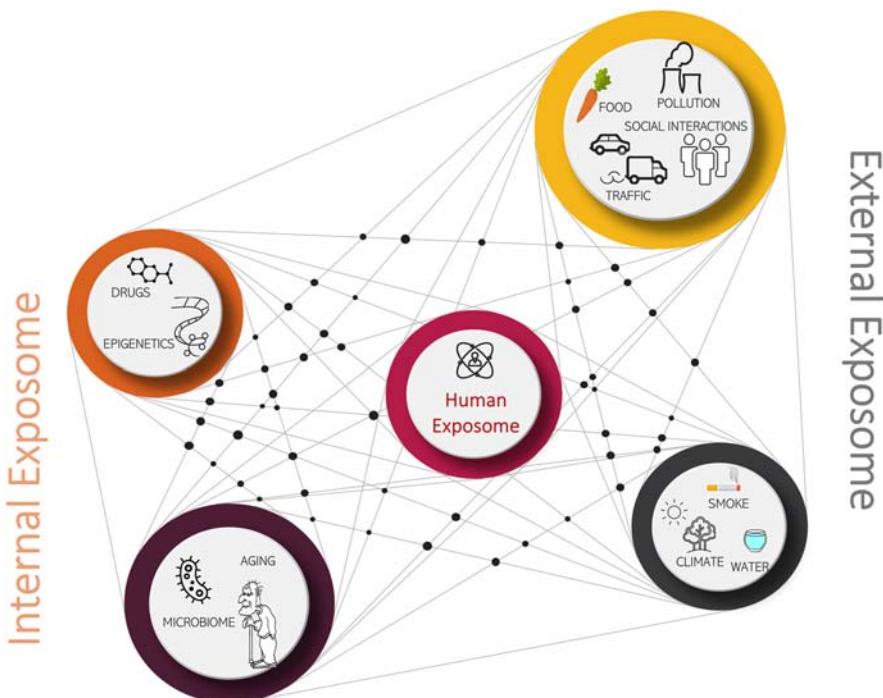


FIGURE 18.2

Internal and external exposome. Internal and external exposomes derive from different sources and establish complex connections to generate the human exposome.

three points of view: direct, epigenetic and genetic. In a direct interaction, an environmental factor interacts with a receptor or activates a biological mediator, and a gene expression is induced. In an epigenetic interaction, an environmental factor determines an alteration of the structure of a gene, and, as a consequence, its expression. In a genetic variation, the gene expression comes from a single-nucleotide polymorphism, which can become active after an environmental stimulus. The biomarkers of the possible interactions can be used for different purposes (Pepe et al., 2008), as

1. screening, a probabilistic process of detecting conditions at early stages
2. diagnosis, the process of assessment of the actual presence of the condition
3. prognosis, the prediction of outcomes for that specific case

which—in turn—may become the basis for the selection of the appropriate intervention in order to provide the best reasonable benefit. The process of identifying a biomarker is critical, as a methodology of limited quality will probably produce useless or misleading results. The quantitation must be considered as a necessity under all points of view, and a good example is represented by the PRoBE method (Pepe et al., 2008); this acronym indicates a “PRospective-specimen-collection, Retrospective-blinded-Evaluation” process. In this methodology the samples are collected making a prospective study into a certain population in which the biomarker will be applied, in absence of knowledge of the patients outcome. Then, only after that the outcome has been ascertained, it is possible to perform the selection of cases, with the outcome, and controls, without it, as well as the assay of the biomarker. The PRoBE design includes four key components (Pepe et al., 2008), namely:

1. the clinical context, that is the clinical setting in which the biomarker will be used
2. the performance criteria, that are the intended uses of the biomarker—for example, for screening, diagnosis or prediction—including the expected performance rates
3. the biomarker test, for example, a single molecule or a combination of molecules or even of biological and clinical data, including imaging
4. the study size, that must be selected according to the epidemiological relevance of the condition in study, and the number of hypotheses to be tested.

This approach can be used in any setting, and for Metabolomic studies its possible outcomes in terms of Public Health are clearly evident: in a given target population and clinical setting, the possibility of obtaining data for a large set of subjects, and then retrospectively extract the subset of interest, means that a complex analysis, such as the Metabolomics one, can be refined in a sequence of iterations, making a sort of “fine tuning” through the complex subset of biomarkers of the disease or condition in study, to obtain its metabolomic signature.

Big data and metabolomics in public health

Big data are defined as very large data bases, mainly unstructured, and in the Health Care field their definition is still under discussion ([Baro et al., 2015](#); [Murdoch & Detsky, 2013](#)). The main parameters ([Dolley, 2018](#)) that characterize them have been listed as:

1. Volume, that according to the sources, as web, social networks, search engines, messages, audio, photos and videos, internet data, etc. can be larger than petabytes and accumulate at very high speed;
2. Variety, that takes into account the disparate sources listed above, and the fact that these data are largely unstructured and may create problems for storage, mining and analyzing;
3. Velocity, that is the pace at which these data are produced and flow from the variety of sources listed before, considering also that this flow is massive and continuous in terms of accumulation, exchange, sharing and utilization;
4. Veracity, that indicates how accurate or truthful a data set is or may be, and can be influenced by the presence of bias, duplications, abnormalities or inconsistencies, and volatility, that is the rate of change and the life time duration of the data and can be considered as another “v” in itself;
5. Validity, also called Value, that is the potential value, in social or economic terms, that the data might create.
6. The types of big data for Public Health have been classified according to their source and to their nature. The classification by source ([Mooney & Pejaver, 2018](#)) is reported in [Table 18.2](#), and divides the sources into five categories.

This classification makes a differentiation in the aspect of the Big Data, by considering “tall” and “wide” data. The wide data may require to reduce the number of dimensions in the data set, for example, selecting specific variables, while the tall

Table 18.2 BDH studies classifications.

Source	Examples	Aspect	Typical uses
Omic/ biological	Metabolomics	Wide	Etiologic research, screening
Geospatial	Neighborhood characteristics	Wide	Etiologic research, surveillance
Electronic Health Records	Records of all patients with hypertension	Tall/ Often wide	Clinical research, surveillance
Personal monitoring	Daily GPS records	Tall	Etiologic research, potentially clinical decision making
Effluent data	Google search results	Tall	Surveillance, screening, identification of hidden social networks

data may require to filter observations, for example, health records unrelated to the hypothesis of interest, to obtain a more tractable data set. The first category investigates a biological aspect, while the second one is focused on a context; however, both are wide. The third one, due to the intrinsically variable nature of these recordings, may have both aspects, while the fourth, measuring a large number of data of the same type, is tall, as well as the last one, that includes data effluent from life in an electronic world. Each subtype of data offers its own unique type of challenges, for example, the Omics, as any biological data may be strongly affected by lab procedures, and it must be emphasized that in the real life a single data set may include more than one category. The classification by nature (Baro et al., 2015) has divided them into three categories. The first one is specific of big data in Omics and concerns a massive number of data collected from a limited number of individuals, so it can be defined as “small n—high p”. This category is connected also to precision medicine, (Koenig et al., 2017), in the sense that as the datasets become more voluminous and complex, they become more likely to meet the criteria for precision medicine and big data. The second category is specific of big data in public health studies, and concerns an important number of individuals but each one with a low number of variables, so “high n—small p”. The third category is referred to the medical specialties, that are characterized by an important number of individuals and variables, and it can be tagged as “high n—high p.” A natural evolution, deriving from the convergence of the first two categories, may be considered the “Precision Public Health” (Baynam et al., 2017), defined as “... a new field driven by technological advances that enable more precise descriptions and analyzes of individuals and population groups, with a view to improving the overall health of populations...”: it leverages Big Data to obtain unprecedented levels of speed in response and precision in targeting problems and optimizing care strategies (Dolley, 2018), mainly using disease surveillance and signal detection, that are obtained, by the way, through the parallel use of the unprecedented development (Khoury et al., 2013) of biomarker, social, environmental and physiological measurements previously discussed. So, Big Data in Health and Metabolomics, as the nearest to the phenotype of the Omics disciplines, offer a special and great contribution to Precision Public Health, through different approaches, such as (Dolley, 2018), disease surveillance, predicting risk, targeting treatment interventions and understanding disease. Moreover, the presence of Big Data into the Healthcare field may foster and support the advancement of the economic mission of the health care delivery (Murdoch & Detsky, 2013), in different ways:

1. may expand the capacity to generate new knowledge, through analysis of unstructured data, and the data analysis may be the similar to untargeted metabolomic studies, plus, in theory, the data base itself could also include data from any kind of Omics study;
2. may help to bring knowledge dissemination nearer to the first-line physician, General Practitioner or Specialist, giving suggestions drawn from data analysis, building a clinical decision support basis that includes the Omics;

3. may help translate personalized medicine initiatives into clinical practice by using analytical capabilities integrating systems biology, for example, Omics, with Electronic Health Recordings data to streamline genomics research towards the phenotype;
4. may allow for a transformation of health care by empowerment of the patient's ability to know, store and manage his/her own information, playing an active role, in which health-related data, for example, drug prescriptions, may be linked to other personal data, such as income, education, household environmental data, diet habits, exercise etc., to be accessed, under patient's permission, to integrate traditional medical model with determinants of health.

In this scenario, the Omics, namely the Metabolomics, may play a role or give support, and, as the BDH approach is data-driven, unbiased by “*a priori*” knowledge, as it rests on the data, and looks at the system as a whole, another common soil of Big Data in Health and Omics is the holistic approach. This approach, in terms of Public Health, considers that the health determining interactions should be studied at the same time in the three levels ([Pourbohloul & Kieny, 2011](#)) which compose the holistic framework of health: societal, institutional and molecular. At societal level there is the interplay between health determinants and environmental factors, at molecular level, the one between all the Omics networks and pathways. The level at which the holistic approach may be probably more interesting for the future is the institutional one, which can be divided into two main areas: the macro- and the micro-institutional level. The interactions at macro level can be considered the ones between institutions, such as a local government or administration with a central government or administration in terms of discussing the transfer of resources or their destination. The micro level is the one of interactions between a single citizen with the institution, that can be a medical center, under public or private governance and financing, or even the interaction of the citizen with a medical operator, that takes the form of the good, old, doctor-patient relationship. Now, this kind of approach may appear far from the Omics aspects, but we should take into account the fact that many of the behavioral aspects of the single citizen's life can be influenced by his/her relationship with the institutions at micro level, and that the environmental factors plus a large number of health determinants can be influenced by the inter-institutional relationships at macro level. Thus, we can conclude that also the institutional interactions are not, all in all, too far from the Omics cascade. In a few words: everything counts. Finally, there are some limitations ([Mooney & Pejaver, 2018](#)) to be taken into account. Indeed, since the use of machine learning techniques requires tall data sets, it may be of difficult understanding due to the “black box” nature of its processes, may leave back quality to be able to obtain quantities large enough for analysis purpose, and can be prone to bias, due to the fact that secondary data are to be used.

Policies, training, and resources

The knowledge integration (Khoury et al., 2013), is the process of combining information or data from disparate sources and from disparate fields to obtain a systematic and quick transfer of results into health benefits, also in terms of processes of decision making, policy and practice, and it implies three phases:

1. management: the identification, selection, storage and tracking of relevant information;
2. synthesis: the use of tools and methods to obtain systematic review, using either unpublished data;
3. translation: the actual use of the synthesized information to drive change or confirmation of choices in policy, practice, and research.

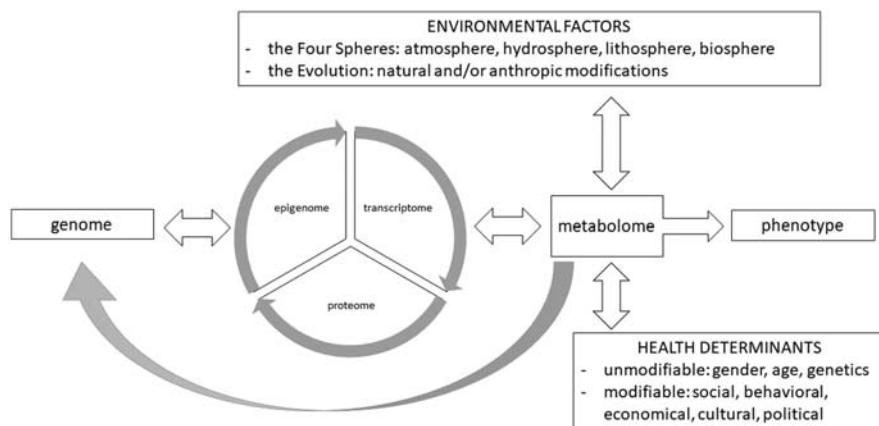
The process of integration towards translation is a fundamental component of the production function, that is the relationship between the three inputs of economic theory, capital, labor, and raw materials, into outputs as goods or services. Through progress of science and technology, this quantitative relationship changes (Murdoch & Detsky, 2013), usually requiring fewer inputs to yield the same or more output, but with a fourth input becoming necessary, that is information production, storage, exchange, management and utilization. The policies underpinning the integration processes and their evolution will need to take into account some problems and constraints, such as: the privacy concerns, for which the solutions could probably be the same used in other sensible areas, for example, the financial one; the need for secure data storage and sharing, which is connected to the former observation; the advantage, in terms of Omics correlation to Public Health data, that could be attained by using also raw data, negative and unpublished results, instead of relying only on published literature, with its subjection to selective reporting and, thus, to report only positive results. The training processes are, or better should be, always oriented by the actual needs of the processes for whom the training is performed, and by their constant evolution and transformation, that, in turn, depend on the surrounding world's evolution and transformation. The training needed for Metabolomic Scientists, thus, requires a problem-solving, action-oriented approach, with focus on innovation and translation, integrating systems biology with data science to change mental models of reasoning in the direction of considering longitudinally the processes, from the complex and multifactorial etiology to the outcomes. The training should also include implementation and dissemination of the results towards translation, including knowledge integration (Khoury et al., 2013), as discussed before. The training processes should be facing also the training syllabi of the clinicians: the innate complexity of the matter, and the intricate processes and reasoning of the Omics scientist should be shared, at least in reasonable part, with the first line practitioner, the one who deals with the patient, to make available (Zeisel, 2007) the knowledge of the weaknesses and strengths of the information given into the

metabolic context. Considering the Omics sciences and their relationship with Public Health strictly related to the Big Data in Health area, it can be considered that this integrated practice (Mooney & Pejaver, 2018) calls for new skills but does not remove the need for the traditional ones, as the training and effort required are nontrivial. Among the new skills, two (Mooney & Pejaver, 2018) appear to be specifically fostered: a change in the way of thinking, and a capacity of analyze bias. First, the “computational thinking”, that is the capacity to think like a computer when working with data, is useful to recognize which problems pose greater algorithmic challenges, and is based on two core principles, abstraction and automation. The second one is the “quantitative bias analysis”, that is the capacity to detect and understand the presence of bias, especially when working with secondary data, for which the investigator was not involved in the process of collection. The combination of rapid technology growth and limited resources (Ghazvini & Shukur, 2013) together with increasing request of healthcare quality and quantity, makes necessary to optimize the strategies, to obtain the most efficient use of every good, tangible or not, involved. Some interesting strategies could be (Khoury et al., 2013) the use of “bricolage”, that is the use of available resources in a novel way, possibly relying also on Big Data stored resources, and constant critical examination of the criteria used to evaluate the opportunity of continuing or modifying existing or replace them with new ones. Moreover, this optimization could benefit from, once again, knowledge integration, through use of existing biobanks, Electronic Health Records and, generally speaking, Big Data, for linking and mining information.

Final remarks

We are not alone. Indeed, we live and share our lives with other living organisms, both at the macro and micro levels. The macro level is made up of ourselves, human beings and communities, together with other living agents which can be seen with naked eye. The micro level consists of a much larger number of microbes, within and around us, whose knowledge and understanding of taxonomy, behavior, and interaction is still at the beginning. The metabolome is a dynamic, complex, large and constantly evolving network of molecules that arises from the biological interactions of our own metabolic pathways with the environment, composed of (Beger et al., 2016) microbes resident in our body but also of molecules that form the four Spheres, as represented in Fig. 18.3.

Fig. 18.3 shows a different representation of Fig. 18.1, in which the model takes into account the different roles of the Omics communities (Beger et al., 2016): the Central Dogma of molecular biology indicates that life arises from chromosomal DNA, which is transcribed into RNA and translated into functional proteins, but fails to consider the crucial role of the small molecules, that “do the dirty job”. In other words, the actual processes that determine the function, or

**FIGURE 18.3**

Interactions between molecular processes, environmental factors, and health determinants.

even the malfunction, of a cell, rely on the interactions of small molecules with the macromolecular components or even with other small molecules. In Fig. 18.3, we have tried to represent this scheme, considering that the genome, on the left side, interacts with a number of Omics processes, namely epigenome, transcriptome and proteome, but the main interface with the phenotype is the metabolome. The latter exchanges signals with the environmental factors and the health determinants, and can—in part—also interact with the genome itself. Unfortunately, there is still a “...tendency of most public pathway databases to ignore the polygenic nature of most vertebrate genomes ... and the polygenic, multi-organ nature of many important pathways...” (Wishart, 2019): this induces an underestimation of the importance of small molecules, not only in life, but also in modifications and adaptation of life to the mutable conditions of gender, age, behavioral, social, economic, environmental and political conditions that impact the life, the health, the disease and the death of any human being on this planet. The metabolites—let's remember it—can be indicators of something bad but also of something good, as Public Health science has clearly expressed classifying risk factors as “downside,” but also as “upside” risk factors, which someone calls “protective factors.” In this sense, Metabolomic studies in Public Health should be targeted to indicate both the types of factors, and possibly both the types of signatures, if they exist. In this way, Metabolomics should become an adult discipline, and enter with full right in clinical practice. One of the problems of the route of Metabolomics to adulthood, which is necessary to make it a standard and widespread component of the Public Health panoply of research and planning methods, is the need for standardization of methods, as we have—for instance—previously said about the PRoBE method (Pepe et al., 2008): there is a need to

reduce the bias in public pathway databases (Wishart, 2019), and probably the method of Consensus Conference should be considered as a possible target for the near future of the International Meetings of the discipline. Another problem may arise from the necessity of integrating data not only according to their belonging to the epidemiological, clinical, environmental, administrative, social and behavioral area of human species, but also in the sense of their belonging to one or more other species, different from the human one. As we are learning every day, the paradigm “we are not alone” literally includes the fact that inside our body there are many other living, and evolving, species: the microbiome. In this sense, a really holistic approach (Pourbohloul & Kieny, 2011) should take into due consideration the fact that pathogenic pathways (Hoffman et al., 2017) activated during early stages of life can explain risk disease in later times, and this happens through factors like the microbiome. Thus, the holistic approach of any omics study should look for a metabolomic signature which may include metabolites coming from, or going into, the metabolome, and, given that the interaction between biological and social factors may influence health (Hoffman et al., 2017), the interplay between Metabolomics and Public Health fields of activity becomes even more suggestive. Furthermore, the microbiome evolves as any other living entity, and its complexity changes over the time, as (Traversi et al., 2021) α -diversity, intra-individual and functional complexity, increases with age, while β -diversity, made by the interindividual variations, become less evident. This makes necessary to consider also the evolutionary aspects of the Metabolomic signatures, making much more relevant the need to develop and make effective and affordable the lifelong and longitudinal studies mentioned above. Generally speaking, the Metabolomics role in Public Health appears to be growingly important as the technologies improve in efficiency, effectiveness, availability and affordability. Public Health scientists and practitioners must become more aware of the principles and applications of Metabolomics to pursue a group of main outcomes, either methodological and operational for a biomarker identified by this technique, that are: how can be determined that the use is appropriate, the validation and stratification have been sufficient, and how could be obtained an early recognition of high risk/complexity subject. The needs for Metabolomics interaction with Public Health, into the larger framework of the Big Data in Health context, are to enhance multidisciplinary collaboration, promote strategic applications of new technologies and improve transparency. The understanding of molecular mechanisms of diseases can give better and more powerful tools of prevention, diagnosis and therapy, but also foster a better knowledge of the roots of Public Health decision making, even in terms of planning, programming and budgeting. Provided that new biomarkers mean better knowledge of what to look for, either in screening, diagnosing and monitoring pathologies, their impact on healthcare will also give a beneficial effect in reduction of the costs, as metabolites—within some limits—are less complex to assay, so a widespread use of precisely suited metabolites can make the lab testing more efficient, effective, and affordable.

References

- Aleta, A., & Moreno, Y. (2019). Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10, 45–62.
- Barabási, A.-L. (2007). Network medicine—from obesity to the “diseasome”. *The New England Journal of Medicine*, 357(4), 404–407.
- Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *Biomed Research International*. Available from <https://doi.org/10.1155/2015/639021>, Epub 2015 Jun 2.
- Baynam, G., Bauskis, A., Pachter, N., Schofield, L., Verhoef, H., Palmer, R. L., Kung, S., Helmholz, P., Ridout, M., & Walker, C. E. (2017). 3-Dimensional facial analysis—facilitating precision public health. *Frontiers in Public Health*, 5, 31.
- Beger, R. D., Dunn, W., Schmidt, M. A., Gross, S. S., Kirwan, J. A., Cascante, M., Brennan, L., Wishart, D. S., Oresic, M., & Hankemeier, T. (2016). Metabolomics enables precision medicine:”a white paper, community perspective. *Metabolomics: Official Journal of the Metabolomic Society*, 12(9), 1–15.
- Chu, S. H., Huang, M., Kelly, R. S., Benedetti, E., Siddiqui, J. K., Zeleznik, O. A., Pereira, A., Herrington, D., Wheelock, C. E., & Krumsiek, J. (2019). Integration of metabolomic and other omics data in population-based study designs: An epidemiological perspective. *Metabolites*, 9(6), 117.
- Dammann, O., Gray, P., Gressens, P., Wolkenhauer, O., & Leviton, A. (2014). Systems epidemiology: What’s in a name? *Online Journal of Public Health Informatics*, 6(3).
- Dolley, S. (2018). Big data’s role in precision public health. *Frontiers in Public Health*, 6, 68.
- Ghazvini, A., & Shukur, Z. (2013). System dynamics in e-health policy making and the “Glocal” concept. *Procedia Technology*, 11, 155–160.
- Gosak, M., Markovič, R., Dolenšek, J., Rupnik, M. S., Marhl, M., Stožer, A., & Perc, M. (2018). Network science of biological systems at different scales: A review. *Physics of Life Reviews*, 24, 118–135.
- Hoffman, D. J., Reynolds, R. M., & Hardy, D. B. (2017). Developmental origins of health and disease: Current knowledge and potential mechanisms. *Nutrition Reviews*, 75(12), 951–970.
- Khoury, M. J., Lam, T. K., Ioannidis, J. P., Hartge, P., Spitz, M. R., Buring, J. E., Chanock, S. J., Croyle, R. T., Goddard, K. A., & Ginsburg, G. S. (2013). Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology and Prevention Biomarkers*, 22(4), 508–516.
- Koenig, I. R., Fuchs, O., Hansen, G., von Mutius, E., & Kopp, M. V. (2017). What is precision medicine? *European Respiratory Journal*, 50(4).
- Mooney, S. J., & Pejaver, V. (2018). Big data in public health: Terminology, machine learning, and privacy. *Annual Review of Public Health*, 39, 95–112.
- Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA: The Journal of the American Medical Association*, 309(13), 1351–1352.
- Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M., & Potter, J. D. (2008). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: Standards for study design. *Journal of the National Cancer Institute*, 100(20), 1432–1438.
- Pourbohloul, B., & Kieny, M.-P. (2011). Complex systems analysis: Towards holistic approaches to health systems planning and policy. *Bulletin of the World Health Organization*, 89(4), 242.

- Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., Greaves, F., Harper, L., Hawe, P., & Moore, L. (2017). The need for a complex systems model of evidence for public health. *The Lancet*, 390(10112), 2602–2604.
- Scott, R. E., Jennett, P., & Yeo, M. (2004). Access and authorisation in a Glocal e-Health Policy context. *International Journal of Medical Informatics*, 73(3), 259–266.
- Traversi, D., Pulliero, A., Izzotti, A., Franchitti, E., Iacoviello, L., Gianfagna, F., Gialluisi, A., Izzi, B., Agodi, A., & Barchitta, M. (2021). Precision medicine and public health: New challenges for effective and sustainable health. *Journal of Personalized Medicine*, 11(2), 135.
- Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M., Kyrtopoulos, S., Nieuwenhuijsen, M., & Phillips, D. H. (2017). The exposome in practice: Design of the EXPOSOMICS project. *International Journal of Hygiene and Environmental Health*, 220(2), 142–151.
- Wishart, D. S. (2019). Metabolomics for investigating physiological and pathophysiological processes. *Physiological Reviews*, 99(4), 1819–1875.
- Zeisel, S. H. (2007). Nutrigenomics and metabolomics will change clinical nutrition and public health practice: Insights from studies on dietary requirements for choline. *The American Journal of Clinical Nutrition*, 86(3), 542–548.

Index

Note: Page numbers followed by “*f*” and “*t*” refer to figures and tables, respectively.

A

- Absorbance detectors, 82
Accuracy, 111–113, 348
Acetone, 46
Acetonitrile, 46
Acetylcholine (Ach), 543
6-acetylcodeine, 143
6-acetylmorphine, 143
Acid-base hydrolysis, 425
Aconiti kusnezoffii radix, 141–142
Actinobacteria, 515–516
Acute myocardial infarction, 614–615
Acylcarnitines, 223–224
Adaptive SMOTE, 353–354
Additional microbially derived metabolites, 523–524
Adenocarcinoma (AC), 162
Adenosylcobalamin, 228–229
Affinity chromatography, 14
Agglomerative hierarchical methods, 310–311
Agglomerative patterns, 307
Aging and senescence study, single-cell metabolomics in, 485–486
Alcohols, 477–480
Algorithm for Reconstruction of Accurate Cellular Networks (ARACNE), 439–441
Alignment, 268–271
Alkaloids, 144
Alloisoleucine, 233–234
Allostasis, 612
 α -cyano-4-hydroxycinnamic acid (CHCA), 537–539
 α -keto- β -methylisovaleric acid, 233–234
 α -ketoisocaproic acid, 233–234
 α -ketoisovaleric acid, 233–234
Altered biochemical pathways detection, 29–30
Alzheimer disease (AD), 17, 38–39, 247
Amino acids, 223–224
 amino acids-derived metabolites, 521–523
9-aminoacridine (9-AA), 537–539
Amniotic fluid, 40
Amphetamines, 144
Amyotrophic lateral sclerosis, 38–39
Analysis of variance (ANOVA), 291–294, 293*f*
Analytical methodologies, 42
Analytical techniques in mass spectrometry-based metabolomics, 132–136
 data analysis, 136
gas chromatography-mass spectrometry, 133
imaging mass spectrometry, 134–136
liquid chromatography-tandem mass spectrometry, 133–134
Animal cells, 481–484
Anticoagulant, 34
Aplysia californica, 481–484
ApoE
 ApoE4, 247
knockout mice models, 543–545
Apoptotic bodies, 429
Archaea, 460–477
Area Under the Curve Receiver Operating Characteristics Curve (AUCROC), 350–351
Artificial intelligence (AI), 305, 436
 for cell design, 444–446
Artificial neural networks (ANNs), 324–329, 327*f*
 training, 326–328
Aryl hydrocarbon receptor (AhR), 521–522
Aspergillus fumigatus, 160
Asproteomics, 625
Assisted reproductive technologies (ART), 10
Atmospheric pressure chemical ionization (APCI), 82
Atmospheric Pressure Ionization (API), 82
Autism spectrum disorders (ASD), 257–258
Autoimmune diseases, 612
Autoscaling, 281, 282*t*
Azidothymidine (AZT), 168

B

- B vitamins, 524
Bacteria, 460–477
Bacteroides fragilis, 17–18
Bagging, 355–356
Bayesian AuTomed Metabolite Analyzer for NMR data (BATMAN), 184
Bayesian classifier, 322
Bayesian ensembling, 358
Bayesian model averaging (BMA), 358
Bayesian model combination (BMC), 358
Bayesian optimization, 371
BAYESIL program, 184
BBMRI-ERIC European Infrastructure, 616
Benchtop NMR spectrometer (BNMR), 165
Big data, 625, 634–636
BDH studies classifications, 634*t*

- Bile acids (BAs), 516–517
 - Binning, 273–275
 - Biodiesel, 477–480
 - Bioethanol, 477–480
 - Biofluids, 37–41
 - amniotic fluid and breast milk, 40
 - CSF, 38–40
 - metabolome, 27
 - saliva, 38–40
 - sweat and tears, 41
 - Biofuels, 477–480
 - Bioinformatics, 136
 - Biological Magnetic Resonance Bank (BMRB), 181–182
 - Biological samples, 66
 - Biologics, 415–416
 - Biomarker discovery, 28–29, 138
 - Biomedical interventions and therapies, 605
 - Biomolecules, 535–536
 - Biopolymers, 101–103
 - Biplot graph, 303
 - Bipolar disorder (BD), 162
 - Birmingham Metabolite Library BML-NMR, 182
 - 1,8-bis(dimethylamino) naphthalene (DMAN), 537–539
 - Bisulfite modification, 9
 - Black box
 - deep learning models, 445–446
 - type operation, 329
 - Bligh and Dyer method, 418–419
 - Blood-based biomarkers, 18–19
 - Boosting, 356–359
 - Bootstrapping, 393–394
 - Boruta's algorithm, 361–362
 - Bottom-up characterization of system, 606
 - Branched-chain amino acids (BCAA), 228–229
 - Branched-chain fatty acids (BCFAs), 518
 - Branched-chain ketoacids (BCKA), 233–234
 - Breast cancer (BC), 613–615
 - Breast milk, 40
 - Bricolage, 637–638
- C**
- ¹³C NMR, 151
 - 1D ¹³C NMR in metabolomic studies, 164–167
 - Calibration
 - curve technique, 53, 267
 - error, 336
 - set, 335–336
 - CAMERA software program, 160
 - Cancer, 615
 - cells, 8
 - Cancer Genome Atlas, 6
 - Cannabinoids, 144
- Capillary chromatography, 70–71
 - Capillary electrophoresis (CE), 99–100, 99f, 219–223, 458–459
 - Carbohydrates, 535–536
 - fermentation, 521
 - Carcinogenesis, 9
 - Cardiovascular diseases (CVD), 244–247
 - Cartesian graph, 295
 - Case-control studies, 29
 - Cause/effects disambiguation, 256–258
 - Celiac disease, 614
 - Cell culture, 430–432
 - analysis of metabolic processes, 423–429
 - for isolation of small extracellular vesicles processes, 429–430
 - isolation of small extracellular vesicles
 - using TFF, 432–435
 - using ultracentrifugation, 430–432
 - metabolomics, 416
 - experimentation steps, 417t
 - interaction network methods in, 440t
 - pathway and network analysis tools in, 438t
 - metabolomics and lipidomics data analysis, 435–443
 - cell modeling for design and optimization of cell culture applications, 436–437
 - determination of major metabolic pathways or network data, 437
 - mechanistic modeling for cell culture optimization, design, and information gathering, 441–443, 443t
 - network analysis in cell culture metabolomics, 437–441
 - ML and hybrid models and AI for cell design, 444–446
 - sample preparation and analysis with NMR spectroscopy, 435
 - sample processing and experimentation for cell culture lipidomics and metabolomics, 417–423
 - Cell harvesting, 424
 - Cell metabolomics, 33
 - Cell modeling, 436–437
 - Cellular metabolome, 27
 - Centering, 280–281, 282t
 - Central Dogma of molecular biology, 638
 - Central nervous system (CNS), 38–39
 - Cerebrospinal fluid (CSF), 38–40, 39f, 184, 248–249
 - Chain fatty acids, 518–520
 - Chara australis*, 494–496
 - CheBi, 400, 402
 - Chebychev distance. *See* Lagrange distance

- Chemical isotope labeling liquid chromatography mass spectrometry (CIL-LCMS), 138, 253
- Chemical shift calibration, 273
- Chenodeoxycholic acid (CDCA), 520
- ChenoMX, 184–185
- Chinese hamster ovary cells (CHO cells), 436–437
- ChIP-sequencing (ChIP-seq), 9–10
- Chiral chromatography, 101–103
- Chiral stationary phase (CSP), 101–103
- Chlamydomonas*, 477–480
- C. reinhardtii*, 477–480
- Cholic acid (CA), 520
- Choline, 524
- Chromatin immunoprecipitation (ChIP), 9–10
- Chromatography, 14, 69–78, 103, 219–223
- definitions and classifications, 70–71, 71 t
 - efficiency of separation, 74–75
 - peak capacity, 76
 - qualitative and quantitative analysis in chromatography, 76–78
 - resolution, 75–76
 - retention, 72–73
 - selectivity, 74
- Circadian system, 608
- “Circle plot”, 297–299, 298 f
- Circular RNA, 427–428
- Class imbalance, 348–354
- machine learning algorithms modification, 354
 - metrics to estimate classification performances, 349–351
 - sampling strategies, 351–354
- Classification algorithms, 279–280, 355
- Classification model validation, 344–348
- k -fold cross validation, 347
 - LkOCV, 346–347
 - LOOCV, 346
 - permutation test, 347–348
- Clinical biomarker discovery, metabolomic analysis for, 138–139
- Clinical metabolism, single-cell metabolomics in, 496–498
- Clostridia, 518
- Clostridioides difficile*, 520–521, 524–525
- Clostridium*
- C. orbiscindens*, 523–524
 - Clostridium scindens*, 520–521, 524–525
 - Clostridium sporogenes*, 521–522
- Cluster analysis, 306–314, 307 f
- hierarchical clustering, 310–312
 - nonhierarchical clustering, 312–314
- Clustered, regularly interspaced, short palindromic repeats (CRISPR), 10
- CRISPR-Cas9 technology, 10
- CRISPR-Cas9-based acetyltransferase, 10
- Codeine, 143
- Codeine-6-glucuronide, 143
- Collision Induced Dissociation (CID), 123, 219–223
- Colorectal cancer (CRC), 520–521
- Complementary DNA (cDNA), 10, 12
- Confusion matrix, 349
- Consensus path database (ConsensusPathDB), 403
- Consistency, 29
- Conventional system, 491–492
- Correlation analysis, 295–296
- Correlation spectroscopy (COSY), 163, 171
- Cost matrices, 354
- Covariance, 300
- Covariance matrix, 299–301
- diagonalization of, 301
- Crohn’s disease (CD), 17–18
- Cross-validation, 345–346, 394–395
- Cumulative Fisher’s exact test. *See* Cumulative hypergeometric test
- Cumulative hypergeometric test, 404

D

- D-chiro-inositol, 18
- D-proline, 101–103
- Data analysis, 287, 435–436
- in metabolomics
 - class imbalance, 348–354
 - classification model validation, 344–348
 - data analysis deriving from metabolomics experiments, 372
 - ensemble machine learning, 354–359
 - exploratory analysis, 287–305
 - features selection, 359–368
 - hyperparameters optimization, 368–371
 - principal bioinformatic tools used in metabolomics, 373 t
 - software, 372
 - supervised machine learning, 314–343
 - unsupervised machine learning analysis, 305–314
- “Data driven” algorithms, 314–315
- Data handling, 63
- Data independent acquisition method (DIA), 131–132
- Data integration, 627–628
- Data normalization, 280
- Data preprocessing, 266–275
- mass spectrometry-based experiments, 266–271
 - nuclear magnetic resonance, 271–275
 - strategies, 282 t
- Data pretreatments, 283
- Data scaling, 283

- Data transformation, 283
 Databases, 54–55
 for NMR-based metabolomics, 181–183
 Debiased Sparse Partial Correlation algorithm (DSPC), 439
 Decision trees (DT), 317–322, 390–391, 391f
 Decision-making algorithm, 320
 Deconvolution, 267–268
 Deep sequencing, 12
 7-dehydroxylated Bas, 520–521
 7 α -dehydroxylating gut bacterium, 520–521
 Deletion, 271
 Density, 100
 gradient ultracentrifugation, 430–432
 Deoxycholic acid (DCA), 520
 1-Deoxysphingolipids, 423–424, 426
 Derivatization, 49–52, 69
 Desorption Electrospray Ionization (DESI), 458–459
 Detectors, 81–83, 90–91, 114
 Developmental biology, single-cell metabolomics in, 484–485
 Diabetes, 612
 Diabetic sensory neuropathy, 423–424
 Diagnosis, 240–241, 241f, 633
 Diagonalization of covariance matrix, 301
 1,5-diaminonaphthalene (DAN), 537–539
 1,4-dideoxy-1,4-imino-D-arabinitol (DAB), 491–492
 Diet, 140, 515–516
 Dietary polyphenols, 523–524
 Diethyl succinate (DES), 165
 Differential ultracentrifugation, 430
 Diffusion-ordered spectroscopy (DOSY), 173
 2,5-dihydroxybenzoic acid (DHB), 537–539
 Dilution, 49
 Diode array detectors (DAD), 81, 83f
 2,3-diphenyl-pyranylum tetrafluoroborate (DPP-TFB), 543
 Dipole stackings, 101–103
 Direct approach, 101–103
 Direct injection mass spectrometry, 42–43
 Direct Injection NMR (DI-NMR), 179–180
 Discriminant analysis, 324
 Disease diagnosis and therapy, 481–484
 Distance analysis, 308
 Distortions enhancement by polarization transfer (DEPT), 166
 Diversified animal applications, 481–484
 “Divide and conquer” approaches, 417–418
 Divisive hierarchical methods, 311–312
 DNA
 methylation, 7–8
 polymerase, 5
 structure, 241–242
 Dopamine (DA), 232–233, 543
 “Double” biphasic extraction methods, 421–422
 Dried blood spots (DBS), 34, 35f, 225–227
 Droplet-based microfluidics, 12–13
 DrugBank, 181–182
 Drugs, 535–536
 discovery, 459–460
 metabolomics in drug development, 139–140
DunalieLLa, 477–480
 Dynamic extractions, 66–67
 Dynamic headspace sampling (DHS), 68–69
 Dynamic headspace techniques, 68–69
 Dynamic Nuclear Polarization (DNP), 164–165
- E**
- Ectosomes, 429
 Edges. *See* Links
 Edman sequencing, 15
Eggerthella lenta, 523
 Eigenvalue matrix, 301
 Electron impact (EI), 44, 91, 114–115
 Electron ionization (EI), 115
 ion source, 115–116
 Electron Transfer Dissociation (ETD), 123
 Electron transfer flavoprotein (ETF), 229–232
 Electrospray ionization (ESI), 114–115, 118–119, 219–223, 536
 Embedded methods, 367
 for variables selection, 388–390
 latent variable methods, 389
 PLS, 390
 principal component regression, 389–390
 regularization techniques, 388–389
 Enantiomers, 101–103
 Endogenous metabolites, 607
 Energomics, 3
 Energy metabolism, 421–422
 Enrichment methods, 403, 406–407
 Ensemble machine learning, 354–359
 bagging, 355–356
 boosting, 356–359
Entamoeba histolytica, 167–168
 Environmental biology, single-cell metabolomics in, 490–491
 Enzyme-linked immunosorbent assays (ELISAs), 14–15
 Epigenetics, 7–10
 tools, 9–10
 Epinephrine, 232–233
 Error, 112–113
 Error back-propagation algorithm (EBP algorithm), 326
 ETF-dehydrogenase (ETFDH), 229–232

Euclidean distance, 308
Euglena, 477–480
 European Bioinformatics Institute (EBI), 402
 Evaporative light scattering detection (ELSD), 81
 Exogenous metabolites, 607
 Exometabolome, 32
 Exosomal proteins, 433
 Exosomes, 429
 Expanded NBS, 225–227
 Exploratory analysis, 287–305
 multivariate approach, 299–305
 univariate approach, 290–299
 Exploratory data analysis (EDA), 289
 Extracellular media, 416–417
 Extracellular vesicles (EVs), 416–417
 Extracted metabolites, 132
 Extraction
 method, 46–47
 sample extraction techniques, 66–69

F

¹⁹F in metabolomic studies, 170
Faecalibacterium prausnitzii, 17–18
 False Discovery Rate (FDR), 293
 False negatives (FN), 349
 False positives (FP), 349
 Farnesoid X receptor (FXR), 520
 Fatty acids, 144
 FdUMP, 170
 Features selection, 359–368, 392
 Boruta's algorithm, 361–362
 embedded methods, 367
 features filtering, 360–361
 features generation, 365–367
 genetic algorithm, 362–365
 Feedforward impulse propagation mechanism, 326
 Fermentable metabolites, 518–520
 Fetal bovine serum (FBS), 429–430
 Findable, Accessible, Interoperable and Reusable principles (FAIR principles), 399
 Finkel-Biskis-Jinkins osteosarcoma (FOS), 13
 Fitness, 363
 Fixed point-based approaches, 340
 Flame ionization detector (FID), 90, 91f
 Flavonoids (7,4'-dihydroxyflavone), 491–492
 Flow control, 85–87
 Flow Injection Analysis NMR (FIA-NMR), 179–180
 Flow probes, 179–180
 Fluorescence detector, 81
 Fluorescence microscopy, 535–536
 Fluorescence-based techniques, 457–458
 2-fluoro-1-methyl pyridinium (FMP), 537–539
 5-fluoro-2'-deoxyuridine (FdUrd), 170

5-fluorouracil (5-FU), 170
 5-fluorouridine (FUDr), 170
 5-fluorouridine-5'-diphosphate (FUDP), 170
 5-fluorouridine-5'-diphospho[1]-a-D-glucose (FUDPG), 170
 5-fluorouridine-5'-monophosphate (FUMP), 170
 Fluxomics, 3, 150–151
 determination of major metabolic pathways or network from, 437
 Folch extraction method, 425–426
 Fold change (FC), 294, 385
 FooDB, 181–182
 Forensic science, metabolomics in, 142–144
 Formaldehyde-fixed and paraffin-embedded tissues (FFPE tissues), 539–540
 Fourier transform algorithm, 122
 Fourier transform ion cyclotron resonance (FT-ICR), 541–542
 Fourier transform ion cyclotron resonance mass spectrometer (FTICR MS), 496–498
 Fourier transform–infrared spectroscopy, 457–458
 Fragmentation, 115–116
 Free induction decay (FID) signal, 177
 Functional genomics, single-cell metabolomics in, 488–489
 Fungi, 460–477
 Fusobacteria, 515–516
 FUTP, 170

G

G-protein-coupled BA receptor 1, 520
 G-protein-coupled receptors (GPCRs), 518
 γ -amino butyric acid (GABA), 523, 543
 Gap filling process, 271
 Gas chromatography (GC), 16–17, 43–44, 64–65, 84–91, 114, 219–223. *See also* Liquid chromatography (LC)
 column, stationary phases, and separation, 88–90
 detectors, 90–91
 mobile phase and flow control, 85–87
 sample introduction and inlets, 87–88
 temperature zones, 87
 Gas chromatography-mass spectrometry (GC-MS), 35, 49–52, 149–150, 422, 516–517
 derivatization, 49–52
 Gas-phase extractions (GPEs), 68–69
 Gastrointestinal tract (GI tract), 515–516
 Gel-based techniques, 14
 Gene Set Enrichment Analysis, 407
 Generalized log transformation, 282 t , 283
 Genes, 3
 expression technique, 12–13

- Genes (*Continued*)
 therapy, 415–416
 Genetic algorithm (GA), 315–316, 362–365,
 387–388
 operators, 364–365
 Genetic information, 4
 Genetically Influenced Metabotypes, 629
 Genome-wide association studies (GWASs), 5–6
 Genomics, 3–7, 15–16
 technologies, 6
 tools, 4–7
Giardia lamblia, 167–168
 Global metabolite analysis, 140
 Global optimization algorithms, 387–388
 Glocalization, 630–631
 Glucose metabolism impairment, 248
 Glutamate (Glu), 543
 Glutaric acidemia, 229–232
 Glutaric acidemia type 1 (GA-I), 229–232
 Glutaric acidemia type 2 (GA-II), 229–232
 Glutaryl-CoA dehydrogenase (*GCDH*), 229–232
 Glycolysis, 496–498
 Goli Metabolome Database, 54–55
 Gradient elution chromatography, 80
 Grid search system, 369–370
 Gut microbiota-derived metabolites, 515–516
 additional microbially derived metabolites,
 523–524
 amino acids-and tryptophan-derived metabolites,
 521–523
 fermentable metabolites and short chain fatty
 acids, 518–520
 metabolomics methods in host-microbiome
 studies, 516–518
 perspectives and future directions, 524–525
 secondary bile acids, 520–521
- H**
- ¹H NMR, 149–151
 1D ¹H NMR spectroscopy, 151–152
 examples, 160–163
 in metabolomic studies, 152–160
 H-bonds, 101–103
 Hard ionization techniques, 44
 Hematopoietic stem cells (HSCs), 486–487
 Hereditary sensory neuropathy type 1 (HSAN1),
 423–424
 Hereditary tyrosinemas, 233
 Heteronuclear Multiple Bond Correlations
 (HMBC), 163
 Heteronuclear Single Quantum Coherence
 (HSQC), 163
 Heuristic approach, 392–395
 bootstrap and stability selection, 393–394
 cross validation, 394–395
 Hierarchical clustering, 310–312, 311f, 312f.
 See also Nonhierarchical clustering
 agglomerative hierarchical methods, 310–311
 divisive hierarchical methods, 311–312
 Hierarchical systems, 307
 High collision dissociation (HCD), 123, 219–223
 High-level variable selection, 388–392
 decision trees, 390–391
 embedded methods for selection of variables,
 388–390
 random forests, 391–392
 support vector machine, 392
 High-performance LC (HPLC), 78, 133–134
 High-resolution magic-angle spinning NMR
 spectroscopy (HR-MAS NMR
 spectroscopy), 174–176
 High-resolution MS (HRMS), 484–485
 Histone deacetylases (HDACs), 518–519
 Histone modifications, 7–8
Homo sapiens, 610–612
 Homogenization, 46
 Homonuclear Hartmann–Hahn experiment.
 See Total COSY (TOCSY)
 Host health, 515–516
 Host physiology, 518, 520–521
 Host-microbiome studies, metabolomics methods
 in, 516–518
 Hotelling transform, 299, 340–341
 HS-SPME, 48
 Human breast adenocarcinoma cells (MCF7 cells),
 481–484
 Human Cell Atlas, 458–459
 Human embryonic kidney 293 cells (HEK293),
 436–437
 Human genome, 4, 515–516
 Human Genome Epidemiology (HuGE), 7
 Human Genome Project (HGP), 4
 Human induced pluripotent stem cells (hiPSCs),
 487
 Human Metabolome Database (HMDB), 54–55,
 131, 160, 399–401
 Human Metabolomics DataBase, 188
 Humans exhibit great phenotypic diversity, 608
 Hybrid models for cell design, 444–446
 Hybridization-based approaches, 12
 Hydrophilic interaction liquid chromatography
 (HILIC), 79–80, 422–423
 3-hydroxy-3-methylglutaryl coenzyme A reductase
 (HMGCR), 13
 3-Hydroxylkynurenine, 248–249
 Hyperparameters optimization, 368–371
 hyperparameters tuning, 369–371
 Bayesian optimization, 371

- grid search system, 369–370
- random search, 370–371
- parameters and hyperparameters in machine learning, 368–369
- Hyperpolarization techniques, 166, 168
- Hyphenated approaches, 130–131
- Hypothesis testing, 385–386
- Hypothesis-driven experimental schemes, 219–223

- I**
- Ideal analytical system, 64–65
- Ideal sample preparation approach, 65
- Illumina Infinium Human Methylation27 array, 10
- Imaging mass spectrometry (IMS), 134–136
- Immobilized pH-gradient (IPG), 15
- Immortalized cells, 33, 415–416
- Immune system, 492–494
- Immune-mediated chronic inflammatory diseases, 492–494
- Immunology, single-cell metabolomics in, 492–494
- Immunotherapy, 615–616
- IMPA LA, 408–410
- Imputation, 271
- In-cell analysis, 416–417
- Inborn error of metabolism (IEM), 223–225, 226t
 - examples of IEM diagnosed by NBS, 228–229
 - MMA, 228–229
 - markers, 227t
 - targeted metabolomics application to NBS of, 225–227
- Indirect approach, 101–103
- Individual phenotyping using nuclear magnetic resonance, 608–613
 - evidence of individual phenotype in different biofluids, 610f
 - individual human urinary metabolic phenotype represented, 613f
 - individual metabolic phenotype, 609f
 - individual phenotype in different biofluids, 610f
 - modeling of individual metabolic phenotype for two species, 609f
 - partitioning of individual metabolic phenotype, 611f
 - resilience of individual human urinary metabolic phenotype, 613f
- Indoleamine 2,3-dioxygenase 1 enzyme (IDO1), 521–522
- Inflammatory bowel diseases (IBDs), 17–18
- Information gain (IG), 320
- Inherited neuropathy, 423–424
- Inlets, 87–88
- Inorganic molecules, 27

- Instrumentation, 78–79
- “Integrative metabolomics” approach, 174–176
- Interactive Data Language (IDL), 541–542
- Interactomics, 3
- Interleukin-10 (IL-10), 518–519
- Internal exposome, 632–633
- Internal standard normalization, 275–278, 276f
- International Cancer Genome Consortium, 6
- International HapMap Project, 4
- Intestinal epithelial cells (IECs), 518–519
- Intra-class correlation coefficients (ICC), 361

- Intrauterine growth restriction (IUGR), 18
- Intrinsic value (IV), 320–321
- Ion detector, 114–115
- Ion sources, 114–119
 - EI, 115–116
 - electrospray, 118–119
 - MALDI source, 116–118
 - mass analyzers, 119–125
- Ion traps (IT), 114–115
- Ion-exchange chromatography, 14
- Ion-to-photon detectors, 114
- Ionic interactions, 101–103
- Ionization efficiency, 113
- Isocratic elution mode, 80
- Isoleucine, 228–229
- Isopropyl alcohol (IPA), 46
- Isotopes, 110–111
 - dilution, 53–54
- Isovaleric acidemia (IVA), 232
- Isovaleryl-CoA dehydrogenase (IVD), 232
- Iterative Dichotomiser 3* algorithm (ID3 algorithm), 320–321

- J**
- J-resolved (JRES), 173
- Jarvis-Patrick method, 313–314

- K**
- k*-fold cross validation, 347
- K-means method, 312–313
- KNIME, 372
- Knowledge extraction process, 287
- Knowledge integration process, 637
- Knowledge-primed neural networks (KPNN), 446
- Kolmogorov–Smirnov test (K–S test), 406–408
 - topological methods, 408
 - Wilcoxon signed rank test, 408
- Krebs cycle, 228–229
- “Kruskal–Wallis” test, 291–292
- Kyoto encyclopedia of genes and genomes (KEGG), 402

L

Lactobacillus, 521–522
 Lagrange distance, 309
 Laser ablation electrospray ionization (LAESI), 458–459
 Laser desorption ionization (LDI), 458–459
 Late-onset Alzheimer’s disease (LOAD), 18–19
 Latent variable methods, 389
 Latent variables (LV), 334, 340–341
 Le Chatelier’s principle, 240
 Least absolute shrinkage and selection operator regression (LASSO regression), 338
 Least invasive method, 34–35
 Least squares method. *See* Ordinary least squares (OLS)
 Leave-k-out cross-validation (LkOCV), 346–347
 Leave-one-out cross-validation (LOOCV), 346
 Leucine, 228–229
 Linear discriminant analysis (LDA), 324
 Linear trap quadrupole Orbitrap (LTQ), 122–123
 Links, 627–628
 Lipid extractions for cell culture analysis, 417–423, 420t
 LipidMaps, 400–401
 Lipidome, 543–545
 Lipidomics, 415–416, 543–545
 characterization, 433
 sample processing and experimentation for cell culture, 417–423
 data analysis, 435–443
 methods for optimized metabolite and lipid extractions for cell culture analysis, 417–423
 techniques, 219–223
 Lipids, 47, 535–536
 Lipopolysaccharide (LPS), 496–498
 Lipostar software. *See* Metaspace software
 Liquid based techniques, 14
 Liquid chromatography (LC), 16–17, 44, 64–65, 78–83, 114, 219–223. *See also* Gas chromatography (GC)
 detectors, 81–83
 instrumentation, 78–79
 principal separation modes, 79–81
 Liquid chromatography-mass spectrometry (LC-MS), 35, 83, 149–150, 421–422
 Liquid chromatography-tandem mass spectrometry, 133–134
 Liquid Ionic Matrixes (LIMs), 537–539
 Liquid-liquid extraction (LLE), 46
 Liquid-phase extractions (LPEs), 67
 Lithium heparin, 34
 Lithocholic acid (LCA), 520

Loading matrix, 301–302
 Loadings graph, 303
 Log transformation, 282t, 283
 Long noncoding RNAs (lncRNAs), 10–11
 Longitudinal and life-long studies in metabolomics, 630–631
 Low density lipoprotein cholesterol (LDLC), 13
 Low-level variable selection, 382–386
 supervised low-level variable selection, 384–386
 unsupervised low-level variable selection, 383–384
 Low-molecular mass chiral selectors, 101–103
 Lysophosphatidylcholines, 246, 543–545

M

Macaca mulatta. *See* Rhesus macaques (*Macaca mulatta*)
 Machine learning (ML), 278, 289–290, 305
 algorithms modification, 354
 for cell design, 444–446
 parameters and hyperparameters in, 368–369
 Macrocyclic antibiotics, 101–103
 Macrocyclic chiral selectors, 101–103
 Macromolecular chiral selectors, 101–103
 Madison–Qingdao Metabolomics Consortium Database (MMCD), 182
 Magnetic resonance imaging (MRI), 535–536
 Mahalanobis distance, 309
 MALDI mass spectrometry imaging (MALDI–MSI), 536
 basics of, 536
 principles and Instrumentation of, 537f
 of endogenous metabolites, 542–546
in situ metabolic pathway imaging visualizes changes, 544f
 instrumentation, 541–542
 matrix choice and application, 537–539
 metabolite annotation and quantitation in, 547
 tissue preparation for, 539–541
 sample preparation workflow schematics, 540f
 Manhattan distance, 309
 Mann–Whitney test, 292
 Maple syrup urine disease (MSUD), 233–234
 MarVis, 408–410
 Mass accuracy, 112–113
 Mass analyzers, 114–115, 119–125
 Orbitrap mass analyzer, 120–124
 quadrupole ion trap, 124–125
 quadrupole mass analyzer, 119–120
 time of flight mass analyzer, 120

- Mass spectrometry (MS), 14–15, 42, 109, 113–125, 219–223, 416–417, 457–458, 516–517, 535–536
 analytical techniques in mass spectrometry-based metabolomics, 132–136
 applications, 136–144
 drug development, metabolomics in, 139–140
 forensic science, metabolomics in, 142–144
 metabolomic analysis for clinical biomarker discovery, 138–139
 nutrition science, metabolomics in, 140–141
 toxicology, metabolomics in, 141–142
 untargeted metabolomics applications, 137f
 ion detector, 114–115
 ion sources, 114–119
 mass analyzer, 114
 mass spectrometry-based experiments, 266–271
 alignment, 268–271
 deconvolution, 267–268
 gap filling, 271
 peak picking and smoothing, 266–267
 mass spectrum, 109–113
 isotopes, 110–111
 resolution and accuracy, 111–113
 system for sample introduction, 114
 tandem mass spectrometry, 125–129
 untargeted metabolomics in complex samples, 129–132
 Mass spectrometry imaging (MSI), 439–441, 457–458, 535–536
 Mass-to-charge ratio (*m/z* ratio), 99–100, 109
 MassBank, 54–55
 Matrix Assisted Laser Desorption Ionization (MALDI), 15, 114–118, 536, 543–545
 Matrix choice and application, 537–539
 Matrix-assisted laser desorption/ionization MS imaging (MALDI IMS), 135
 Medical science, 605
 Mediterranean diet, 140
 Medium-level variable selection, 386–388
 global optimization algorithms, 387–388
 stepwise regression, 387
 variable selection or wrapper methods, 386–387
 Mendelian analysis, 247
 Mendelian diseases, 4–5
 MetaboAge, 251–252
 Metabolic biomarkers, 612–613
 Metabolic fingerprints, 612–613
 Metabolic flux, 423–429
 Metabolic footprint, 32
 Metabolic pathway analysis, 403–406
 overrepresentation, 404–406
 Metabolically active samples, 31–32
 Metabolically inactive samples, 31–32
 Metabolite identification, 160
 Metabolite set enrichment analysis (MSEA), 406–407
 Metabolites, 3, 27, 535–536
 identification, 54–55
 MSI, 55
 public libraries and databases, 54–55
 quantification of, 52–54
 calibration curve technique, 53
 internal standard and isotope dilution, 53–54
 Metabolites ontology, 399–403
 common metabolite databases, 401–402
 CheBI, 402
 HMDB, 401
 LipidMaps, 401
 common pathway databases, 402–403
 consensus path database, 403
 KEGG, 402
 reactome, 403
 SMPDB, 402
 Wikipathways, 403
 ontologies, 399
 for metabolites, 399–401
 Metabolome, 27, 381
 coverage, 252–253
 variability, 250–252
 Metabolome Standard Initiative, 254
 Metabolomic analysis for clinical biomarker discovery, 138–139
 Metabolomic data conversion
 data preprocessing, 266–275
 data pretreatment, 278–284
 normalization, 275–278
 Metabolomics, 3, 16–19, 27, 63–64, 237, 381, 415–416, 457–458, 535–536, 606, 615–616, 625
 analysis of metabolic processes, 423–429
 methods and protocols for isolation and metabolomics of sEVs, 427–429
 cell culture, 416
 data analysis, 435–443
 experimentation steps, 417t
 sample processing and experimentation for, 417–423
 in drug development, 139–140
 experimental design in
 analytical methodologies, 42
 applications, 28–30
 identification and quantification of metabolites, 52–55
 metabolically active *vs.* metabolically inactive, 31–32
 NMR, 42–45

- Metabolomics (Continued)**
- other biofluids, 37–41
 - quality control, 55–56
 - sample preparation, 45–52
 - sample types, 31
 - targeted metabolomics, 31
 - tissue and cells, 32–33
 - untargeted metabolomics, 30–31
 - urine, 36–37
 - whole blood, plasma, and serum, 33–35
 - workflow in design of metabolomic experiment, 28f
 - in forensic science, 142–144
 - metabolomics-based personalized medicine, 612–613
 - in nutrition science, 140–141
 - research, 63–65
 - tools, 16–19
 - in toxicology, 141–142
- Metabolomics databases and NMR software programs**, 181
- Metabolomics methods in host-microbiome studies**, 516–518
- Metabolomics profiling**, 242–244, 243f, 245f
- Metabolomics Quality Assurance & Quality Control Consortium**, 254
- Metabolomics Standards Initiative (MSI)**, 55
- MetaboMiner software program**, 160
- Metaspace software**, 543–545
- Metastatic breast cancer (MBC)**, 9
- Methane**, 477–480
- Methanol**, 46
- Methionine**, 228–229
- Methoxyamine**, 50
- Methyl tert-butyl ether (MTBE)**, 47, 418–419
- Methylcobalamin**, 228–229
- Methylmalonic acidemias (MMA)**, 228–229
- Methylmalonyl-CoA mutase (MUT)**, 228–229
- 4-methylpyridine (4MP)**, 165
- METLIN**, 54–55
- MetPA tool**, 408
- MetPer package**, 372
- Microarray techniques**, 4
- Microarray-based comparative genomic hybridization**, 4–5
- Microbes**, 477–480
- Microbial metabolism**, 477–480, 518
- Microbial metabolites**, 515–516
- Microbial technology, single-cell metabolomics in**, 460–480
- Microbiota-accessible carbohydrates (MAC)**, 518
- Microbiota-derived tryptophan metabolites modulates**, 522
- microRNA (miRNA)**, 8–9, 427–428
- Microscopic organisms**, 460–477
- Minkowski distance**, 309–310
- Missing data treatment process**, 271
- Modern mass spectrometers**, 111–112
- Modulators**, 97
- Molecular signatures**, 427–428
- Molecular techniques**, 605–606
- Molecules**, 3
- Mono-dimensional proton NMR (1D ¹H NMR)**, 170
- Monogenic traits**, 6
- Monoglycerides**, 246
- Monophasic extraction**, 46–47
- Morphine**, 143
- MPINet**, 408–410
- Multicolor fluorescence detection-based microfluidic device (MFD-MD)**, 481–484
- Multidimensional chromatography**, 92–99
- concept of multidimensionality, 92–96
 - practical and instrumental aspects, 96–99
- Multidimensional NMR experiments**, 150
- Multidimensionality**, 92–96
- Multiple linear regression (MLR)**, 386–387
- Multiple reaction monitoring (MRM)**, 127–129, 222–223
- Multiple test correction**, 404
- Multivariate approach**, 63, 223–224, 290, 299–305. *See also* Univariate approach
- loadings and scores in principal components analysis, 301–303
 - significative components, 303–304
- MxP FastQuench**, 45–46
- Myo-inositol**, 18
- N**
- N,O-bis(trimethylsilyl)-trifluoroacetamide (BSTFA)**, 50
- ¹⁵N NMR**, 151
- 1D ¹⁵N NMR in metabolomics, 167–168
- N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA)**, 50
- N-tert-butyldimethylsilyl-N-methyl-trifluoroacetamide (MTBSTFA)**, 51–52
- Naïve Bayesian (NB)**, 322–324
- Nano-scale secondary ion mass spectrometry (NanoSIMS)**, 477–480
- Nanoparticle tracking analysis (NTA)**, 433
- National Cancer Institute**, 625
- National Institute of Standards and Technology (NIST)**, 133
- National Institutes of Health**, 458–459
- Network analysis**, 627–628
- in cell culture metabolomics, 437–441
- Neurodegenerative disease**, 247–249

- Neuropsychiatric disorders, 612
 Neurotransmitters, 232–233
 Neutral loss scan (NL), 127–128
 New psychoactive substances (NPS), 142–143
 Newborn screening (NBS), 223–224
 examples of IEM diagnosed by, 228–229
 targeted metabolomics application to, 225–227
 Next-generation sequencing (NGS), 4–5
 Nodes, 627–628
 Noncoding RNAs (ncRNAs), 10, 427–428
 Nonhierarchical clustering, 312–314, 313f.
 See also Hierarchical clustering
 Jarvis-Patrick method, 313–314
 K-means method, 312–313
 Nonhierarchical systems, 307
 Nonlinear separable data, 331–332
 Nonparametric tests, 292
 Nonpolar metabolite extraction, 52
 Nontargeted analyses, 64
 Norepinephrine, 232–233
 Normal phase mode (NPLC), 79–80
 Normalization, 275–278
 centering, 280
 internal standard normalization, 275–278, 276f
 PQN, 277, 278f
 quantile normalization, 277–278
 scaling process, 281–284
 transformation, 281–284
 Noscapine, 143
 Nuclear magnetic resonance, 49
 individual phenotyping using, 608–613
 Nuclear magnetic resonance spectroscopy (NMR spectroscopy), 34, 42–45, 129–130, 149–174, 219–223, 271–275, 416–417, 457–458, 516–517, 607–608
 advantages, 185–187
 reproducibility, 186–187
 binning, 273–275
 challenges and limitations, 187–188
 chemical shift calibration, 273
 gas chromatography, 43–44
 HR-MAS NMR spectroscopy, 174–176
 improvements in NMR hardware and techniques
 and additional tools in metabolomics studies, 177–185
 databases for NMR-based metabolomics, 181–183
 flow probes, 179–180
 magnets, 178–179
 metabolomics databases and NMR software programs, 181
 probes, 179
 use of software to analyze metabolite NMR data, 183–185
 liquid chromatography, 44
 mass spectrometry, 42–44
 1D-NMR, 151–170
 pure shift nuclear magnetic resonance, 176–177
 sample preparation, 188–191
 and analysis, 435
 techniques without sampling, 44–45
 2D nuclear magnetic resonance spectroscopy, 170–174
 water signal elimination, 272–273
 Nucleic acids, 427–428
 Nutrition research, single-cell metabolomics in, 489–490
 Nutrition science, metabolomics in, 140–141
 Nutritional metabolic diseases, 140
 Nutritional metabolomics, 140–141
- O**
- Obesity, 612
 Odd-chain fatty acids, 228–229
 Omics
 cascade, 631–632
 profiles/fingerprints, 614
 revolution, 606
 sciences integration, 627–628
 technologies, 3, 187, 287, 457–458
 Oncology, 11–12
 Oncotype-DX 21-gene, 615
 1D nuclear magnetic resonance (1D-NMR), 149–170
 applications in metabolomics, 154f
¹⁹F in metabolomic studies, 170
 1D ¹³C NMR in metabolomic studies, 164–167
 1D ¹⁵N NMR in metabolomics, 167–168
 1D ¹H NMR spectroscopy, 151–152
 examples, 160–163
 in metabolomic studies, 152–160
³¹P NMR in metabolomic studies, 168–170
 strengths and weaknesses of, 153f
- Ontologies, 399
 for metabolites, 399–401
 Open-tubular column (OT column), 70–71, 88f
- Opioids, 144
 Orbitrap, 114–115
 Fusion instrument, 123, 124f
 Fusion mass spectrometers, 122–123
 mass analyzers, 43, 120–124, 121f
- Ordinary least squares (OLS), 336–337
 Organic molecules, 27
 Orthogonal partial least square discriminant analysis (OPLS-DA), 343
 Overfitting, 321, 344, 345f
 Overrepresentation algorithms, 403–406
 Oversampling, 351–352

- P**
- ³¹P NMR, 151
 - in metabolomic studies, 168–170
 - p*-value, 292
 - PaintOmics, 408–410
 - Papaverine, 143
 - Paper chromatography, 241–242
 - Para-hydrogen, 165–166
 - Parahydrogen Induced Polarization (PHIP), 164–165
 - Parallel reaction monitoring (PRM), 222–223
 - Parameters in machine learning, 368–369
 - Parametric tests, 290
 - Parent ions, 219–223
 - Pareto scaling, 281, 282_t
 - Parkinson disease (PD), 17, 247–248
 - Partial least square algorithms (PLS algorithms), 333–334
 - Partial least square discriminant analysis (PLS-DA), 223–224, 338–343, 339_f
 - geometric interpretation of, 342–343
 - latent variables, 340–341
 - OPLS-DA, 343
 - VIP, 341–342
 - Partial least square regression (PLS-R), 333–335
 - geometric interpretation of, 335
 - prediction error in partial least square, 335–336
 - Partial least squares (PLS), 389–390
 - Pasteurization, 40
 - Patho-physiological process, 460
 - Pathway analysis
 - enrichment, 406–407
 - K–S test, 407–408
 - metabolic pathway analysis, 403–406
 - metabolites ontology, 399–403
 - MSEA, 407
 - tools for metabolomic, 408–410, 409_t
 - Pathway mapping, 441
 - Peak capacity, 76
 - Peak picking, 266–267
 - Pearson’s correlation coefficient, 384–385
 - Peptides, 535–536
 - Peptostreptococcus russellii*, 521–522
 - Permutation test, 347–348
 - Personalized medicine, 137
 - Pharmacodynamics, 6
 - Pharmacogenomics, 6
 - Pharmacokinetics, 6
 - Phenotypes, 437–439
 - Phenylalanine (Phe), 232–233
 - Phenylketonuria (PKU), 232–233
 - Phosphate-buffered saline (PBS), 45–46
 - Photodiode-array (PDA), 81
 - π–π interactions, 101–103
 - Pipeolic acid, 101–103
 - Pirkle-type donor-acceptor, 101–103
 - Plant metabolomics, 138
 - Plant science and agriculture, single-cell metabolomics in, 480–481
 - Plasma, 33–35
 - metabotype, 245–246
 - monoglycerides, 246
 - Platelet activating factors (PAFs), 419–421
 - Polar metabolites, 52
 - Porous-layer OT columns (PLOT), 88
 - Positivity, 308
 - Positron emission tomography (PET), 535–536
 - “Post-hoc” test, 291–292
 - Post-pruning, 321
 - Posttranslational modifications (PTMs), 8
 - Power transformation, 282_t, 283–284
 - Pre-pruning, 321
 - Precision, 112–113
 - Precision medicine
 - applications, 613–616
 - individual phenotyping using nuclear magnetic resonance, 608–613
 - metabolomics as tool for
 - applications, 613–616
 - individual phenotyping using nuclear magnetic resonance, 608–613
 - systems approaches and systems medicine, 605–608
 - systems approaches and systems medicine, 605–608
 - Precision Medicine Initiative, 7
 - Precision oncology, 6
 - “Precision Public Health”, 634–636
 - Precursor fragmentation, 123
 - Precursor ion scan (PIS), 127, 222
 - “Precursor ions”, 219–223
 - Prediabetics, 252
 - Prediction error in partial least square, 335–336
 - 4–Predictive, preventive, personalized, and participatory medicine (4-P medicine), 606
 - Pretreatment method, 281
 - Primary cells, 33
 - cultures, 415–416
 - Principal component analysis (PCA), 223–224, 281, 290, 299, 300_f, 302_f, 389–390
 - Principal components (PCs), 299, 389–390
 - Principal components regression (PCR), 335–336, 389–390
 - Probabilistic Context Likelihood of Relatedness Algorithm (PCLR), 439–441
 - Probabilistic process, 633

Probabilistic quotient normalization (PQN), 277, 278f
 Probability density function, 340
 Probe electrospray ionization (PESI), 481–484
 Probe ESI mass spectrometry (PESI-MS), 480–481
 Product ions (PIs), 127
 scan, 127, 222
 Prognosis, 240–241, 241f, 633
 Programmed temperature vaporizer (PTV), 87–88
 Programming languages, 372
 Propionic acidemia (PA), 229
 Prostate cancer (PCa), 18
 Proteins, 3, 427–428, 535–536
 microarrays, 15
 Proteobacteria, 515–516
 Proteogenomics, 3
 Proteomics, 3, 13–16
 proteomic tools, 14–16
 Protozoa, 460–477
 Pruning operation, 321
 Public health, metabolomics in
 big data and, 634–636
 data integration, 627–628
 longitudinal and life-long studies in
 metabolomics, 630–631
 policies, training, and resources, 637–638
 quantitative methods, 631–633
 system biology and, 628–630
 thematic recommendations, 626t
 Public libraries and databases, 54–55
 Pure shift nuclear magnetic resonance, 176–177
 Purge and trap (P&T), 68–69

Q

Q Exactive, 122–123
 Quadratic discriminant analysis (QDA), 324
 Quadrupole ion trap analyzer, 124–125
 Quadrupole mass analyzer, 119–120
 Quadrupole-TOF (qTOF), 541–542
 Quadrupoles (Q), 114–115
 Qualitative analysis in chromatography, 76–78
 Qualitative response, 385–386
 fold change, 385
 hypothesis testing, 385–386
 Quality control (QC), 29, 55–56, 360–361
 Quantification, 42
 Quantile normalization (QN), 277–278
 Quantitative analysis in chromatography, 76–78
 Quantitative bias analysis, 637–638
 Quantitative methods, 631–633
 Quantitative response, 384–385
 Pearson’s correlation coefficient, 384–385

Quantitative trait locus (QTL), 6–7
 Quenching, 45–46, 66

R

R code, 408–410
 “R” approaches, 295
 Radio frequency (RF), 119
 Raman spectroscopy, 457–458
 Raman-Deuterium Isotope Probing (Raman-DIP), 477–480
 Random errors, 112–113
 Random forest (RF), 317–318, 391–392
 Random search, 370–371
 Range scaling, 281, 282t
 RapidMiner, 372
 Rational tuning, 369
 Reactome, 403
 Receiver operating characteristic curve (ROC curve), 340
 Reductionist approach, 243
 Refractive index detectors, 81
 Regressive models, 332–338
 LASSO regression, 338
 least absolute shrinkage and selection operator regression, 338
 PLS-R, 333–335
 RIDGE regression, 336–337
 Regularization techniques, 388–389
 Relative quantification, 53
 Renal Cell Carcinoma (RENCA), 165
 Reproducibility, 186–187
 Residuals sum of squares (RSS), 336–337
 Resilience, 612
 Resolution, 75–76, 111–113
 Retention, 72–73
 Retention index (RI), 89–90
 Reverse phase chromatography, 422
 Reversed-phase LC (RPLC), 79–80
 Rhesus macaques (*Macaca mulatta*), 610–612
 Rheumatoid arthritis, 492–494
 RIDGE regression, 336–337, 337f
 RNA, 8–9
 RNA-sequencing (RNA-seq), 10, 12
 Root mean square error of calibration (RMSEC), 336
 Root mean square error of cross validation (RMSECV), 336
Ruminococcus gnavus, 524–525

S

S programming language, 289–290
 SABRE in SHield Enables Alignment Transfer to Heteronuclei (SABRE-SHEATH), 167–168

- Saccharides metabolism, 477–480
Saccharomyces cerevisiae, 219–223
 Saliva, 38–40, 190
 Sample extraction techniques, 66–69
 Sample preparation, 45–52, 65–69
 and analysis with NMR spectroscopy, 435
 extraction, 46–47
 derivatization, 69
 GC-MS, 49–52
 liquid chromatography-mass spectrometry, 52
 in NMR spectroscopy, 188–191
 nuclear magnetic resonance, 49
 quenching, 45–46
 sample clean up, 47–49
 controlling metabolite concentrations, 48–49
 solvent removal, 47–48
 SPE, 48
 ultrafiltration, 48
 sample extraction techniques, 66–69
 Sample size, 254
 Sample types, 31
 Sampling strategies, 29, 351–354
 Savitzky-Golay filter, 267
 Scaling process, 281–284
 Scan modes, 127
 Schizophrenia (SCZ), 162
 Scores plot, 303
 Secondary bile acids, 520–521
 Secondary ion MS imaging (SIMS imaging), 458–459
 Selection strategies of metabolites
 heuristic approach, 392–395
 high-level variable selection, 388–392
 low-level variable selection, 382–386
 medium-level variable selection, 386–388
 variable selection levels, 383^t
 Selectivity, 74
 Separation techniques
 capillary electrophoresis, 99–100
 chiral chromatography, 101–103
 fundamentals of chromatography, 69–78
 gas chromatography, 84–91
 liquid chromatography, 78–83
 multidimensional chromatography, 92–99
 sample preparation, 65–69
 separation processes role in metabolomics research, 63–65
 SFC, 100
 steps for sample analysis in metabolomics study, 64^f
 Sequence-based approaches, 12
 Serum, 33–35
 Short chain fatty acids (SCFAs), 515–516
 Short-read sequencing, 5
- Signal Amplification by Reversible Exchange (SABRE), 164–165
 Signal-to-noise ratio (S/N), 113
 Simulated annealing (SA), 387–388
 Simulation-based kernel ML (SimKernML), 446
 Sinapinic acid (SA), 537–539
 Single cell analysis, 416–417
 Single cell metabolomics (SCM), 458–459
 in aging and senescence study, 485–486
 in clinical metabolism and disease perspective, 496–498
 in detection of metabolite dynamicity and pathway modulation, 494–496
 in developmental biology, 484–485
 diversified animal applications, 481–484
 in environmental biology, 490–491
 in functional genomics, 488–489
 in immunology, 492–494
 in microbial technology, 460–480
 in nutrition research, 489–490
 in plant science and agriculture, 480–481
 prominent SCM interventions in diverse biological applications, 461^t
 in stem cell biology, 486–487
 in system biology, 491–492
 Single cell omics, 458–459
 Single Cell Surveyor Society, 458–459
 Single human hepatocytes (HepG2/C3A), 481–484
 Single reaction monitoring (SRM), 222–223
 Single-cell metabolomics in detection of metabolite dynamicity and pathway modulation, 494–496
 Single-cell technology, 458–459
 Single-cell transcriptomics, 12
 Single-nucleotide polymorphisms (SNPs), 4
 Size exclusion chromatography, 14
 Small extracellular vesicles (sEVs), 427–428
 cell culture for isolation of, 429–430
 isolation using ultracentrifugation, 430–432
 methods and protocols for isolation and metabolomics of, 427–429
 standard strategies for isolation of, 431^t
 Small Molecules Pathway Database (SMPDB), 181–182, 402
 Small-molecule neurotransmitters, 543
 Smoothing algorithm, 266–267
 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), 14
 Soil microbes, 477–480
Solanum lycopersicum L., 480–481
 Solid-phase extraction (SPE), 35, 48, 67–68
 Solid-phase microextraction (SPME), 48, 68

- Solid-Phase-Extraction NMR (SPE-NMR), 179–180
- Solid-state NMR (ssNMR), 174
- Solvent
- removal, 47–48
 - solvent-based extraction methodologies, 67–68
 - system, 418–419
- Spectral overlap, 187–188
- Sphingomyelins, 543–545
- SPIDIA project, 190
- Squamous cell carcinoma (SCC), 162
- Stability selection, 393–394
- Standard operating procedures (SOPs), 188–189, 249, 254
- Static extractions, 66–67
- Statistical analysis, 394
- Statistical evaluation, 63
- Statistical models, 381
- Statistical TOTal Correlation SpectroscopY (STOCSY), 173, 177
- Stem cell biology, single-cell metabolomics in, 486–487
- Stepwise regression, 387
- Stimulated saliva, 37–38
- Streptococcus mutans*, 523
- Succinate (SUC), 165
- Supercritical fluid chromatography (SFC), 70–71, 100, 101^f
- Supercritical fluids, 70–71
- Supervised algorithms, 305–306
- Supervised classification algorithms, 354
- Supervised low-level variable selection, 384–386.
- See also* Unsupervised low-level variable selection
 - qualitative response, 385–386
 - quantitative response, 384–385
- Supervised machine learning, 314–343
- ANNs, 324–329
 - discriminant analysis, 324
 - DT, 317–322
 - Naïve Bayessian, 322–324
 - PLS-DA, 338–343
 - regressive models, 332–338
 - SVM, 329–332
- Support vector machines (SVM), 329–332, 392
- nonlinear separable data, 331–332
 - separating hyperplanes, 330^f
- Support-coated OT (SCOT), 88
- Sweat, 41
- Symmetry, 308
- Synechococcus sp.*, 488–489
- Synthetic Minority Oversampling Technique (SMOTE), 352, 353^f
- Synthetic polymers, 101–103
- System biology, 3, 628–630
- epigenetics, 7–10
 - genomics, 4–7
 - metabolomics, 16–19
 - proteomics, 13–16
 - single-cell metabolomics in, 491–492
 - transcriptomics, 11–13
- Systemic lupus erythematosus, 492–494
- Systems approaches, 605–608
- differences between reductionism and systems science, 607^f
- Systems approaches, 605–608, 607^f
- Systems biology, 606
- Systems Epidemiology, 628–629
- Systems medicine, 605–608, 607^f

T

- t*-tests, 291–292, 386
- T3DB, 181–182
- Tabu search (TS), 387–388
- “Tandem in space” approach, 127
- “Tandem in time” approach, 127
- Tandem mass spectrometry (MS/MS), 43, 114–115, 125–129, 219–223, 426
- instruments for, 125–127
 - tandem mass spectrometry scan modes, 127–129
 - MRM, 128–129
 - neutral loss scan, 127–128
 - precursor ion scan, 127
 - product ion scan, 127
- Tangential flow filtration (TFF), 429
- isolation of small extracellular vesicles using, 432–435
 - characterization of sEVs, 433
- Targeted analyses, 64
- Targeted metabolomics, 31, 131, 219–224, 237.
- See also* Untargeted metabolomics
 - application to NBS of IEM, 225–227
 - examples of IEM diagnosed by NBS, 228–229
 - glutaric acidemia, 229–232
 - hereditary tyrosinemas, 233
 - IEM, 224–225, 226^f
 - IVA, 232
 - MSUD, 233–234
 - PA, 229
 - PKU, 232–233
- “Task driven” algorithms, 314–315
- Tears, 41
- Temperature zones, 87
- Tert*-butyldimethylsilyl (TBDMS), 51–52
- Tetraspanins, 433
- Thebaine, 143
- Thin layer chromatography (TLC), 70–71

- Threonine, 228–229
- Time of Flight (TOF), 112–115, 541–542
analyzers, 43
mass analyzer, 120, 121f
- Tissues, 32–33
metabolome, 27
stem cells, 487
- TMA *N*-oxide (TMAO), 524
- Top-down analysis, 606
- Topological methods, 408
- Topological overlap matrix (TOM), 439
- Total COSY (TOCSY), 151
- Toxicology, metabolomics in, 141–142
- Traditional Chinese medicines (TCM), 141–142
- Training and testing sets, 345–346
- Transcriptomics, 3, 11–13, 625
tools, 12–13
- Transformation of data, 281–284
- Translation, 637–638
- Transmission efficiency, 113
- Trapped ion mobility (TIMS), 543–545
- Triangular inequality, 308
- Tricarboxylic acid (TCA), 481–484
- Trichomonas vaginalis*, 167–168
- Trimethylamine (TMA), 524
- Trimethylsilyl (TMS), 50
- Trimethylsilylation derivatization, 50
- Trimethylsilylpropanoic acid (TSP), 273
- Triple quadrupole analyzer (QqQ analyzer), 219–223
- Triple quadrupole spectrometry, 43
- tRNA, 8–9
- True negatives, 349
- True positives (TP), 349
- Trypsinization, 424
- Tryptophan, 521–522
tryptophan-derived metabolites, 521–523
- Tryptophan hydroxylase 1 (TPH1), 521–522
- Tumor genomics, 6
- Two-dimension (2D), 92
chromatography, 92
NMR techniques, 163, 170–174
- Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), 14
- Type 2 diabetes, 6–7
- Type I tyrosinemas (TYR-I), 233
- Type II diabetes (T2DM), 252
- Type II tyrosinemas (TYR-II), 233
- Tyrosine (Tyr), 232–233
- U**
- Ulcerative colitis (UC), 17–18
- Ultracentrifugation (UC), 429
- isolation of small extracellular vesicles using, 430–432
density gradient ultracentrifugation, 430–432
differential ultracentrifugation, 430
- Ultrafiltration, 48
- Ultrahigh-pressure LC (UHPLC), 78
- Undersampling, 351
- Univariate approach, 63, 290–299. *See also* Multivariate approach
tests to investigate metabolite concentration differences, 292–299
- Univariate statistics, 223–224
- Unsupervised low-level variable selection, 383–384. *See also* Supervised low-level variable selection
percentage observed, 383–384
variance based selection, 384
- Unsupervised machine learning, 305–314.
See also Supervised machine learning
cluster analysis, 306–314
- Untargeted analysis, 31
- Untargeted metabolomics, 30–31, 219–223, 237.
See also Targeted metabolomics
application, 240–242
cause/effects disambiguation, 256–258
in complex samples, 129–132
CVD, 244–247
independent cohort to validate results, 254–256
limitations, 249–250
local and nonlocal metabolomics effects, 238–240
metabolomics pipeline standardization, 254
metabolomics profiling, 242–244
moving metabolomics from laboratories to clinics, 253
neurodegenerative disease, 247–249
sample size, 254
sources of metabolome variability, 250–252
trends in, 252–253
metabolome coverage, 252–253
- Urinary metabolome, 36
- Urine metabolomics, 138
- V**
- Vaccine particles, 415–416
- Validation error, 336
- Valine, 228–229
- van Deemter equation, 70–71
- van Der Waals forces, 101–103
- Variable importance in projection (VIP), 341–342, 390
- Variable selection method, 386–387
- Variance, 300
based selection, 384

Verrucomicrobia, 515–516
Vertices. *See* Nodes
Vibrational spectroscopy, 457–458
Vitamin B3, 524–525
Volcano plot, 238–239, 294–295, 294f

W

Wall-coated OT (WCOT), 88
Water signal elimination, 272–273
Weighted gene correlation network analysis
(WCGNA), 439

Weka, 372
Western blotting, 15
“White box” metabolism modeling methods, 446
Whole blood, 33–35
Whole genome bisulfite sequencing (WGBS), 9
Whole genome sequencing (WGS), 5
Wikopathways, 403
Wilcoxon signed rank test, 408
Wrapper method, 386–387

Metabolomics Perspectives

From Theory to Practical Application

Metabolomics Perspectives: From Theory to Practical Application is an expertly written volume, which provides a thorough description of the current state-of-the-art in the metabolomics field.

The philosophy behind the book is to guide the reader in a step-by-step exploration of metabolomics experiments, ranging from sample preparation to data extraction, analysis and interpretation, and to discuss the main current applications and future perspectives of this emerging science.

Armed with critical insights, coupled with a clear writing, the book consists of three main sections. The first one introduces the pivotal theoretical fundamentals and provides a comprehensive overview of the “wet” laboratory workflow, including protocol instructions and a detailed description of experimental methods and analytical techniques. The second section covers a wide range of topics in the context of data analysis, including guidance in exploratory analysis, supervised and unsupervised machine learning approaches and validation and optimization methods. In addition to the several examples reported in the text, the book features an R package, specifically designed to perform all the described algorithms, which is hosted on a companion website (www.metabolomsperspectives.com) together with several sets of available metabolomic data. Finally, an extensive dissertation describes the latest advances and the major fields of interest for metabolomics applications, highlighting their crucial potentials for future biomedical research.

Thus, this book represents a must-read for both experienced researchers, interested in metabolomics, and newcomers to the field.

Key Features

- Provides an in-depth description of the metabolomics experimental workflow and its applications in life science and biomedical research
- Features chapter contributions from the greatest international experts in the field
- Includes an R package and several sets of metabolomics data, hosted on a companion website

About the editor

Author of more than 100 papers in international peer-reviewed journals, holder of several international patents, and speaker at dozen international conferences, Jacopo Troisi is a award-winning expert in the metabolomics field. After graduating Summa cum Laude in biotechnology, he joined Medical School in the prestigious “Scuola Medica Salernitana” at the University of Salerno (Italy). He also attended and completed studies in data analysis. Currently, he is a visiting professor of Metabolomics in both Salerno and Sannio Universities (Italy), cofounder and CEO at Theoreo Srl—Spin-off Company of the University of Salerno, and COO at the European Biomedical Research Institute of Salerno (EBRIS). His research interests are mainly focused on metabolomic profiling to provide insights in human disease onset and for the development of diagnostic tools.



ACADEMIC PRESS

An imprint of Elsevier

elsevier.com/books-and-journals

ISBN 978-0-323-85062-9

9 780323 850629