

Joseph Markowitz *Editor*

# Translational Bioinformatics for Therapeutic Development

Methods and Protocols

# METHODS IN MOLECULAR BIOLOGY

*Series Editor*

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, UK

For further volumes:  
<http://www.springer.com/series/7651>

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

# **Translational Bioinformatics for Therapeutic Development**

Edited by

**Joseph Markowitz**

*Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*



*Editor*

Joseph Markowitz

Department of Cutaneous Oncology

H. Lee Moffitt Cancer Center & Research Institute

Tampa, FL, USA

ISSN 1064-3745

ISSN 1940-6029 (electronic)

Methods in Molecular Biology

ISBN 978-1-0716-0848-7

ISBN 978-1-0716-0849-4 (eBook)

<https://doi.org/10.1007/978-1-0716-0849-4>

© Springer Science+Business Media, LLC, part of Springer Nature 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

---

## Preface

This book entitled *Translational Bioinformatics for Therapeutic Development: Methods and Protocols* is designed to introduce the reader to the realm of bioinformatics as it relates to therapeutic development. Protocols (wet lab, dry lab) are presented including the necessary introductory material needed to understand each of these specific chapters. The reader will note that therapeutic development includes drug discovery, but it is meant to be much broader to include all modalities that will eventually assist in patient care. It is the goal of this book to be useful for several groups of researchers: (1) basic/translational scientists who are interested in the methods outlined below or who have a keen interest on how their methods may be translated to the clinical setting, (2) physician scientists who require a better understanding of these topics, and (3) postdoctoral/graduate/advanced undergraduate trainees who should learn these methods to enhance their training and career development.

Translational bioinformatics, thorough learning from experiments and maximizing accessibility of data, determines how the findings of the laboratory will be beneficial in the clinical setting. The laboratory setting is grounded in basic science, which includes but is not limited to disciplines discussed in this volume such as biochemistry, biophysics, immunology, molecular biology, omics, technology-driven laboratories (i.e., nuclear magnetic resonance, mass spectrometry, genetics, flow cytometry, robotically driven drug screening laboratories), and pharmacology. In the realm of the basic laboratory, we also have dry lab disciplines, including computer science, physics, mathematical modeling, biostatistics, and computational chemistry among others. Bioinformatics is more than just the large amounts of data associated with biological experiments. Just as outlined in the first chapter in this book for clinical informatics, bioinformatics relates to the analysis of data (either experimental or simulated) that leads to meaningful interpretation. Different types of analysis are useful depending on the source of the data and the size of the data sets. Often, these data sets are large, but a bioinformatics mindset can be applied to analysis regardless of data size. We acknowledge that in many cases in this book and in our own research we are not dealing with the definition of “big data” as defined by our data science colleagues. As such, formal definitions of bioinformatics, clinical informatics, and translational bioinformatics will be provided in the first chapter.

The first section of the book begins with a series of chapters designed to introduce the reader to the clinical informatics issues that arise when embarking on translational bioinformatics studies. We begin with an overview written by Drs. Perkins and Markowitz of how to set up a clinical informatics infrastructure that will support translational bioinformatics studies using our own experiences in a US National Cancer Institute-designated Comprehensive Cancer Center (Chapter 1: “Development and Optimization of Clinical Informatics Infrastructure to Support Bioinformatics at an Oncology Center”). Chapter 2 is devoted to how to leverage pathology informatics (Dr. O’Leary) for translational and clinical studies (“Leveraging Pathology Informatics Concepts to Achieve Discrete Lab Data for Clinical Use and Translational Research”). In addition, we devote a chapter to cohort identification for translational research utilizing currently available technology (Chapter 3: “Cohort Identification for Translational Bioinformatics Studies”). Modern clinical pathway development is then described in a fashion suitable for facilitating patient treatment decisions including incorporation of decision points for entry into clinical trials (Drs. Hooda and

Fields: Chapter 4: “Transitioning Clinical Practice Guidelines into the Electronic Health Record Through Clinical Pathways”).

The next section analyzes considerations of how to measure time-to-event data that is crucial in translational experiments. Two chapters are dedicated in this area to include a general introduction chapter and a more detailed methods chapter. Chapter 5 written by Drs. Ni and Li is a general introductory article: “Variable Selection for Time-to-Event Data.” Chapter 6 authored by Drs. Pietrzak and Rempala’s group provides a method description: “Binary Classification for Failure Risk Assessment.”

The following series of chapters focus on genomic approaches to therapeutic development. This section begins with a general overview of the topic in Chapter 7 by Drs. Gray and Campbell entitled “Challenges and Opportunities of Genomic Approaches in Therapeutics Development.” The group led by Dr. Anders discusses examples of utilizing a public database such as TCGA to associate genes with cancer. Chapter 8 is entitled “Accessible Pipeline for Translational Research Using TCGA: Examples of Relating Gene Mechanism to Disease-Specific Outcomes.” In Chapter 9, Dr. Fridley’s group provides a nice overview of the statistical and bioinformatics methods that are useful for single-cell RNA sequencing experiments in the article “Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments.” Dr. Altrock’s group describes how to measure sample diversity in RNA data in Chapter 10: “Investigating Inter- and Intra-Sample Diversity of Single-Cell RNA Sequencing Datasets.”

Combining omics approaches is useful in drug discovery by providing a thorough understanding of tumor biology. A research team involving Drs. Koomen, Teer, and Eschrich describe in Chapter 11 how to combine genomics and proteomics data in the chapter “Managing a Large-Scale Multi-Omics Project: A Team Science Case Study in Proteogenomics.” Additionally, Dr. Li’s group in Chapter 12 describes “Synergistic Drug Combination Prediction by Integrating Multi-omics Data in Deep Learning Models.”

To understand disease processes and to decipher the effects of therapeutic manipulations on disease states, it is important to phenotype immune and other types of cells. Two papers are devoted to this topic. In Chapter 13, Dr. Markowitz’s group outlines the principles needed for phenotyping experiments in an article entitled “Introduction to Multi-parametric Flow Cytometry and Analysis of High-Dimensional Data.” This article is also meant as an introduction to the field such that one can be an effective collaborator and user of phenotyping data. In Chapter 14, Dr. Mailloux’s group provides a detailed example in the article entitled “High Dimensional Flow Cytometry Analysis of Regulatory Receptors on Human T Cells, NK Cells, and NKT Cells.”

Metabolomics is a key area of translational science as it is important to understand the metabolic pathways important for disease processes and the metabolites of drugs administered to patients. Two different techniques are described in this series. Dr. Patterson’s group describes a metabolomics procedure using mass spectrometry in Chapter 15: “Quantitative Analysis of Bile Acid with UHPLC-MS/MS.” Dr. Mal describes the utilization of NMR techniques in Chapter 16: “A Protocol for NMR-Based Metabolomics.” As you will see from the protocols, mass spectrometry and nuclear magnetic resonance spectroscopy techniques are complementary. Although mass spectrometry generally has greater sensitivity for individual metabolites, NMR does provide some advantages in certain situations such as the ability to measure individual nuclei simultaneously ( $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^1\text{H}$ , etc.) and to nondestructively detect metabolites not able to be measured via mass spectrometry approaches.

This book provides an introduction to the translational bioinformatics skills such that the reader can apply these techniques to their own questions in therapeutic development.

With the current state of technology, it will become clear during the course of this volume that the method of extraction of raw material for these assays influences the hypotheses that may be tested. In addition, the technologies utilized to measure the samples can influence the results. This volume addresses the wet and dry lab techniques needed to generate data sets in translational bioinformatics for therapeutic development. Bioinformaticists require robust techniques to facilitate data analysis. It is therefore critical to consider the questions being answered, the techniques for data generation, and the analysis approach in translational bioinformatics as you read each chapter in this book and determine how these strategies could be utilized in your own research or training program.

*Tampa, FL, USA*

*Joseph Markowitz*

---

# Contents

<i>Preface</i> .....	v
<i>About the Editor</i> .....	xi
<i>Contributors</i> .....	xiii
1 Development and Optimization of Clinical Informatics Infrastructure to Support Bioinformatics at an Oncology Center .....	1
<i>Randa M. Perkins and Joseph Markowitz</i>	
2 Leveraging Pathology Informatics Concepts to Achieve Discrete Lab Data for Clinical Use and Translational Research .....	21
<i>Mandy Flannery O'Leary</i>	
3 Cohort Identification for Translational Bioinformatics Studies.....	35
<i>Tiffany A. Lin, Zeynep Eroglu, Rodrigo Carvajal, and Joseph Markowitz</i>	
4 Transitioning Clinical Practice Guidelines into the Electronic Health Record through Clinical Pathways.....	45
<i>Sharjeel M. Hooda and Karen K. Fields</i>	
5 Variable Selection for Time-to-Event Data .....	61
<i>Ai Ni and Chi Song</i>	
6 Binary Classification for Failure Risk Assessment .....	77
<i>Ali Foroughi pour, Ian Loveless, Grzegorz Rempala, and Maciej Pietrzak</i>	
7 Challenges and Opportunities of Genomic Approaches in Therapeutics Development .....	107
<i>Jaimie S. Gray and Moray J. Campbell</i>	
8 Accessible Pipeline for Translational Research Using TCGA: Examples of Relating Gene Mechanism to Disease-Specific Outcomes.....	127
<i>Anders E. Berglund, Ryan M. Putney, Jordan H. Creed, Garrick Aden-Buie, Travis A. Gerke, and Robert J. Rounbehler</i>	
9 Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments .....	143
<i>Xiaoqing Yu, Farnoosh Abbas-Aghababazadeh, Y. Ann Chen, and Brooke L. Fridley</i>	
10 Investigating Inter- and Intrasample Diversity of Single-Cell RNA Sequencing Datasets .....	177
<i>Meghan C. Ferrall-Fairbanks and Philipp M. Altrock</i>	

11	Managing a Large-Scale Multiomics Project: A Team Science Case Study in Proteogenomics .....	187
	<i>Paul A. Stewart, Eric A. Welsh, Bin Fang, Victoria Izumi, Tania Mesa, Chaomei Zhang, Sean Yoder, Guolin Zhang, Ling Cen, Fredrik Pettersson, Tonghong Zhang, Zhibhua Chen, Chia-Ho Cheng, Ram Thapa, Zachary Thompson, Melissa Avedon, Marek Wloch, Michelle Fournier, Katherine M. Fellows, Jewel M. Francis, James J. Saller, Theresa A. Boyle, Y. Ann Chen, Eric B. Haura, Jamie K. Teer, Steven A. Eschrich, and John M. Koomen</i>	
12	Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models .....	223
	<i>Tianyu Zhang, Liwei Zhang, Philip R. O. Payne, and Fuhai Li</i>	
13	Introduction to Multiparametric Flow Cytometry and Analysis of High-Dimensional Data.....	239
	<i>James Sun, Jodi L. Kroeger, and Joseph Markowitz</i>	
14	High-Dimensional Flow Cytometry Analysis of Regulatory Receptors on Human T Cells, NK Cells, and NKT Cells .....	255
	<i>Ryosuke Nakagawa, Jason Brayer, Nicole Restrepo, James J. Mulé, and Adam W. Mailloux</i>	
15	Quantitative Analysis of Bile Acid with UHPLC-MS/MS.....	291
	<i>Yuan Tian, Jingwei Cai, Erik L. Allman, Philip B. Smith, and Andrew D. Patterson</i>	
16	Sample Preparation and Data Analysis for NMR-Based Metabolomics.....	301
	<i>Tapas K. Mal, Yuan Tian, and Andrew D. Patterson</i>	
	<i>Index .....</i>	315

---

## About the Editor

JOSEPH MARKOWITZ is a physician scientist and translational cancer researcher specializing in cutaneous oncology with a focus in melanoma, dissecting the immunological mechanisms of resistance to checkpoint blockade via a combination of immunology, biochemistry, and informatics approaches at a National Cancer Institute-designated Comprehensive Cancer Center: H. Lee Moffitt Cancer Center & Research Institute. His laboratory focuses on deciphering the mechanisms of resistance to checkpoint blockade. Dr. Markowitz also serves as a physician informaticist to improve clinical and research workflows using the electronic medical record.

---

## Contributors

- FARNOOSH ABBAS-AGHABABAZADEH • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- GARRICK ADEN-BUIE • *Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ERIK L. ALLMAN • *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA*
- PHILIPP M. ALTROCK • *Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- MELISSA AVEDON • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ANDERS E. BERGLUND • *Department of Bioinformatics and Biostatistics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- THERESA A. BOYLE • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JASON BRAYER • *Department of Malignant Hematology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JINGWEI CAI • *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA*
- MORAY J. CAMPBELL • *Division of Pharmaceutics and Pharmacology, College of Pharmacy, The Ohio State University, Columbus, OH, USA*
- RODRIGO CARVAJAL • *Biostatistics and Bioinformatics Shared Resource, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- LING CEN • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- CHIA-HO CHENG • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- Y. ANN CHEN • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ZHIHUA CHEN • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JORDAN H. CREED • *Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ZEYNEP EROGLU • *Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Oncologic Sciences, University of South Florida Morsani School of Medicine, Tampa, FL, USA*
- STEVEN A. ESCHRICH • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- BIN FANG • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- KATHERINE M. FELLOWS • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- MEGHAN C. FERRALL-FAIRBANKS • *Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- KAREN K. FIELDS • *Department of Breast Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- MICHELLE FOURNIER • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JEWEL M. FRANCIS • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*

- BROOKE L. FRIDLEY • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- TRAVIS A. GERKE • *Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JAIMIE S. GRAY • *Division of Pharmaceutics and Pharmacology, College of Pharmacy, The Ohio State University, Columbus, OH, USA*
- ERIC B. HAURA • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- SHARJEEL M. HOODA • *Department of Satellite and Community Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- VICTORIA IZUMI • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JOHN M. KOOMEN • *Department of Molecular Oncology/Pathology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JODI L. KROEGER • *The Flow Cytometry Core Facility, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- FUHAI LI • *Institute for Informatics (I2), Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA; Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA*
- TIFFANY A. LIN • *Collaborative Data Services Core, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- IAN LOVELESS • *College of Public Health, The Ohio State University, Columbus, OH, USA*
- ADAM W. MAILLOUX • *Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- TAPAS K. MAL • *Department of Chemistry, Pennsylvania State University, University Park, PA, USA*
- JOSEPH MARKOWITZ • *Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Oncologic Sciences, University of South Florida, Morsani College of Medicine, Tampa, FL, USA*
- TANIA MESA • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JAMES J. MULÉ • *Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- RYOSUKE NAKAGAWA • *Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- AI NI • *Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA*
- MANDY FLANNERY O'LEARY • *Pathology Informatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Clinical Pathology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Oncologic Sciences, University of South Florida, Tampa, FL, USA*
- ANDREW D. PATTERSON • *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA*
- PHILIP R. O. PAYNE • *Institute for Informatics (I2), Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA*

- RANDA M. PERKINS • *Department of Internal Medicine, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA; Department of Oncologic Sciences, University of South Florida, Morsani College of Medicine, Tampa, FL, USA*
- FREDRIK PETTERSSON • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- MACIEJ PIETRZAK • *Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA*
- ALI FOROUGHI POUR • *Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA; Department of Mathematics, The Ohio State University, Columbus, OH, USA*
- RYAN M. PUTNEY • *Department of Bioinformatics and Biostatistics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- GRZEGORZ REMPALA • *Department of Mathematics, The Ohio State University, Columbus, OH, USA; College of Public Health, The Ohio State University, Columbus, OH, USA*
- NICOLE RESTREPO • *Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ROBERT J. ROUNBEHLER • *Department of Tumor Biology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JAMES J. SALLER • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- PHILIP B. SMITH • *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA*
- CHI SONG • *Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA*
- PAUL A. STEWART • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JAMES SUN • *Department of Cutaneous Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- JAMIE K. TEER • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- RAM THAPA • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- ZACHARY THOMPSON • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- YUAN TIAN • *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA*
- ERIC A. WELSH • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- MAREK WLOCH • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- SEAN YODER • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- XIAOQING YU • *Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- CHAOMEI ZHANG • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- GUOLIN ZHANG • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*
- LIWEI ZHANG • *School of Mathematical Science, Dalian University of Technology, Dalian, Liaoning, China*
- TIANYU ZHANG • *Institute for Informatics (I2), Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA; School of Mathematical Science, Dalian University of Technology, Dalian, Liaoning, China*
- YONGHONG ZHANG • *H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*



# Chapter 1

## Development and Optimization of Clinical Informatics Infrastructure to Support Bioinformatics at an Oncology Center

Randa M. Perkins and Joseph Markowitz

### Abstract

Translational bioinformatics for therapeutic discovery requires the infrastructure of clinical informatics. In this chapter, we describe the clinical informatics components needed for successful implementation of translational research at a cancer center. This chapter is meant to be an introduction to those clinical informatics concepts that are needed for translational research. For a detailed account of clinical informatics, the authors will guide the reader to comprehensive resources. We provide examples of workflows from Moffitt Cancer Center led by Drs. Perkins and Markowitz. This perspective represents an interesting collaboration as Dr. Perkins is the Chief Medical Information Officer and Dr. Markowitz is a translational researcher in Melanoma with an active informatics component to his laboratory to study the mechanisms of resistance to checkpoint blockade and an active member of the clinical informatics team.

**Key words** Clinical informatics, Bioinformatics, Translational research, Oncology operational processes

---

### 1 Introduction

The goal of this chapter is to present a methodology based on both available literature and the experience of the authors that will inform and guide the development of clinical informatics resources to support bioinformatics research at an oncology center. The authors of this chapter have experience in creating and developing successful clinical informatics programs at both general medical and oncology-specific institutions.

The terminologies utilized in this chapter follow those commonly utilized in the field of clinical informatics as recognized by the American Board of Preventive Medicine (ABPM) for use in the medical subspecialty of clinical informatics [1–4]. Informatics is focused on the acquisition, storage, and use of information in a specific setting or domain [5]. Clinical informatics is focused on the people and bodies involved in developing, maintaining, optimizing,

and assessing the clinical application of health information systems. Whereas biomedical informatics is predominantly focused on translational laboratory data (dry or wet lab) to improve upon existing therapeutic modalities [6]. Translational bioinformatics (TBI) encompasses those efforts made to bridge clinical and biological data [7]. Biomedical informatics is “the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to improve human health [8].”

Core to the application and use of clinical informatics is the fundamental theorem of clinical informatics that the combination of humans and technology is more effective than either alone [9]. It is in line with that fundamental theorem that the implementation of various forms of clinical information systems have been occurring for over 50 years [10]. It is only recently that comprehensive electronic health records (EHRs) have become commonplace in clinical workflows [11]. These EHRs, which started as rudimentary laboratory information systems (LIS) and have matured into more complete EHRs with various capabilities directed toward physicians including the ability to integrate the clinical data with research enterprises, provided the appropriate IRB support mechanisms [12–17].

The term electronic health record (EHR) will be used in this chapter in reference to the system which serves as the patients’ health records in digital form and has the components required to qualify as a “certified EHR” as per current Centers for Medicare and Medicaid Services (CMS) criteria [18]. While in some literature the terms electronic medical record (EMR) and EHR are used interchangeably, they are now differentiated as the EMR referring to the medical records of a patient within a single clinic, and the EHR refers to a broader inclusion of data and records from multiple clinical practices and healthcare organizations [19].

The rationale for the conversion from primarily paper-based processes to electronic ones is not discussed in detail in this chapter. There have been mixed results from studies showing the benefits of general EHR conversion, which we will not define or defend here [20–22]. It is presumed that the audience is interested in understanding how to use the EHR to support biomedical research in a comprehensive cancer center. The business and regulatory components of EHR utilization and support are only reviewed here where they may impact the understanding of the use of EHRs for biomedical research. The focus of this chapter is the application of clinical informatics as it interfaces with the clinical world and that of information technology in a comprehensive cancer center. The use of clinical bioinformatics or translational biomedical informatics is dependent upon a robust and stable IT infrastructure [23–28]. With involved clinical staff and receptive information

technology resources there is an opportunity for system development and continuous optimization.

There are assumed benefits of electronic processes and documentation that benefit the growth and sustenance of biomedical research in oncology. These include, but are not limited to, improved cohort identification, expedited clinical trials screening at the individual patient level, novel association discovery, and the ability to perform true *in silico* studies [29]. It stands to reason that clinical research institutions would benefit from staff dedicated to this work who straddle both domains, the clinical and the technical. This has been demonstrated by the comparison of creating clinical information systems (CIS) under a clinician-led team design (CLDT) versus an enterprise model, with the adaptive, clinician-led design producing higher efficiency scores [30]. To integrate clinical research workflows translational bioinformaticists work well with clinical informaticists to utilize the health information infrastructure to develop research. Clinical Informatics essentially serves as the bilingual translation team between the translational research world and the clinical world, serving both.

With widespread adoption of EHRs the US healthcare system is in sudden need of highly qualified health IT and informatics staff in each organization. This sudden growth has created a job market that exceeds the current available workforce, leaving many organizations to either train their own teams or to outsource their support. Organizations that can afford to do so are paying competitively to recruit and retain the resources that have become mission critical to both clinical practice and research. If the organization being evaluated lacks these critical resources decisions must be made about internal development or outsourcing, though it may be difficult to apply cost-benefit ratios when the return on investment is measured more in clinician satisfaction than in clinical throughput and RVUs. Additionally direct comparisons of existing EHRs is difficult, further requiring internal reflection within organizations as they are left to compare and contrast both existing states and future customization with a lack of current industry standards [31].

Some organizations have chosen to hire contractors to support their growth, but this is not a long-term solution. The EHR is not going away, and utilization is only increasing as payers apply more regulatory requirements for data and the bioinformatics needs grow. Direct clinical decision support using advanced integration of data, such as the application of molecular pathology and pharmacogenomics, will require integrated clinical informatics expertise to build and maintain these processes. Organizations must commit to building the teams needed for their future state plans and this includes plans for developing translational research workflows and incorporation into the EHR.

---

## 2 Organization-Specific Issues

The development of clinical informatics for translational research requires user-centered design. User-centered design requires balancing the efficiency of mass-producing a single system for broad use versus ever-finer user-level customization, with the goal of improved efficiencies in a system defined as one that incorporates the user efficiencies as part of the greater whole [32]. The integration of the clinical and information technology worlds requires governance and structure to prevent both over- and undercustomizations of systems, or a mix of both [33]. While there is much angst over “click counts,” users may not be as aware of the cognitive burden placed by nonintuitive system design. Studies have looked at click count, overall EHR time, and error rates as proxy measures for usability [34, 35]. A great deal of the existing knowledge on clinician focus and workflow analysis is derived from prior and ongoing work done in the aerospace industry, analyzing pilot cognitive processing [36]. Additionally, there is more focus on direct patient interaction, with or without the computer as a shared resource [35, 37]. No matter how it is measured, though, utilization of the EHR has been implicated in the rising rates of provider burnout and decreased patient satisfaction [38–40]. Therefore, it is in organizations’ best interest to invest in well-structured EHR implementations and optimization plans, to avoid redundancy and duplication of efforts when customizing a system to adapt to clinicians’, and patients’, preferred workflows. These clinical systems are critical as much of the data to be used for research and stored in data warehouses are ultimately dependent upon the primary input of clinicians. Therefore, below we spend some time discussing change management in this realm.

To make system changes that are responsive to clinicians’ needs, the clinicians’ input is necessary to inform, if not drive, these changes [41]. Clinicians’ time and attention, however, especially for providers, are very limited: A 2014 paper reported that providers spend, on average, over 50 h per week in direct patient care and 5–10 h per week in pro bono work [42]. This leaves little time to advise healthcare systems on the design of the electronic health records that many providers did not ask for. This resource issue of limited physician time is a contributor to the “Knowledge Acquisition Bottleneck” problem seen in developing expert EHR systems [43]. Bioinformatics is utilizing ontology development (formal descriptions of biomedical processes and relationships between them) to standardize knowledge models for increased speed of development [44, 45]. For data to be usable and reliable, it is critical that it is entered into the EMR as discrete values, in an efficient manner, while maintaining clinical utility. To establish workflows to enter data, governance structures are needed to

ensure that all relevant stakeholders' (physicians, nurses, medical assistants, etc.) workflows are addressed.

We recommend the development of structured governance councils integrated in the organization's existing governance infrastructure, with the authority to control and prioritize the clinically facing aspects of the electronic health record and its associated workflows. Governance councils should have adequate representation from all stakeholders and no elective changes to the EHR and its associated workflows should be pursued without the explicit agreement of the impacted parties. Clinical informatics staff exist to facilitate this governance and accepted design processes, working with the clinical stakeholders to define requests and present technical options for fulfilling them. These requests are then coordinated by the clinical informatics staff to develop the changes with the information technology teams, as requested by the clinicians, with clinician validation of any nonproduction and production changes. Clinical workflow changes include those of clinical research workflows.

In any organization with limited resources there will need to be prioritization of these requests. Existing EHR change prioritization models exist, but each organization should develop a prioritization matrix consistent with its own regulatory requirements, risk management processes, revenue goals, and research needs [46]. Large enough stakeholder groups may benefit from sub-councils to deal with concerns limited to that user group, for example those clinical workflows that are predominantly within a specific discipline, with little to no impact on the workflows of the other disciplines, may be processed at discipline-specific councils and fed into the larger organizational governance. Similarly, clinical trial operational workflows, while interfacing with, and as a part of, the larger organizational clinical workflows, have unique nuances and requirements associated with the needs of trial maintenance and regulation.

## **2.1 Staffing and Workflow Issues**

Most healthcare organizations have information technology (IT) support, but it is unclear how pervasive formal clinical informatics is in the healthcare industry. As of 2017, nearly 9 in 10 (85.9%) office-based physicians had adopted any EHR [47]. While it is one thing to have a computer in the office, it is another thing entirely to have all clinical processes dependent on its function and stability. As "Meaningful Use" has swept the country so too has the need to assess and support the integration of health IT into clinical workflows. In some instances existing workflows were directly converted, like-for-like, to a digital process, but often the work itself is drastically changed by the conversion, required a deep analysis of the workflow. The workflows for complex clinical trial EHR integration is currently a major concern for cancer centers.

The in-depth work needed for these conversions is greatly facilitated by having clinicians with a strong understanding of the limitations and capabilities of health IT who are dedicated to improving clinical processes through the integration of technology. Clinical informaticists, also referred to as clinical informaticians, perform this work for healthcare organizations. Clinical informatics and reporting structure models vary widely; therefore, this chapter will only reflect on and recommend those models which the authors have had direct interaction with and organizationally benefit from in the pursuit of translational research.

Two of the disciplines needing staffing in clinical informatics are nursing and medical informatics. Nursing informatics is defined as the “science and practice (that) integrates nursing, its information, and knowledge with information and communication technologies to promote the health of people, families, and communities worldwide” (adapted from IMIA Special Interest Group on Nursing Informatics 2009). The American Nurses Association formally recognized clinical informatics in 1992 and board certification became available through the American Nurses Credentialing Center (ANCC) in 1995. Medical informatics followed this development and in 2008 the American Medical Informatics Association (AMIA) defined the core content of the sub specialty for physicians [3]. The American Board of Preventive Medicine, in collaboration with the American Board of Pathology and others, cosponsors the medical subspecialty exam which was first administered in October 2013. Additional informatics disciplines such as pharmaceutical and imaging informatics are also important for the organizational infrastructure. Basically, any field that deals with the evolution of data within the EHR will eventually require informatics expertise.

Early studies of EHR conversions (generally from paper to EHR) clearly showed that a common theme of conversions that did not “stick” is the lack of end-user involvement in the process; this is consistent with the authors’ experience as well [48]. Many clinical informatics groups or departments are started as subsets of existing clinicians that were the least resistant to, or were outright supportive of the inclusion of health IT in clinical practice. Often, they were the early “super users” or “physician champions” drafted from, or in addition to, their full-time clinical practice. Super user models of implementation have pros and cons, with a high variability of outcomes depending on the individuals chosen to act in the super user role [49]. It is not clear if, at the time, the administrations that did this anticipated that these furloughs would be short term, but it would become rapidly apparent that the conversion to EHRs is not a onetime investment and requires ongoing support and maintenance. Even if the super users’ time is financially compensated, the individuals rarely have the extended IT training or exposure that would qualify them to act as informatics experts and

to function as clinical-IT liaisons, demonstrating the need for dedicated informaticians. In addition to physicians and nurses, other staff needed are the computer scientists and programmers themselves, staff dedicated to developing specific knowledge sets for the EMR, staff for education, and dedicated informatics and IT support to answer urgent questions from the clinical staff. Auditors of the system are needed to ensure proper functioning in addition to staff with terminology expertise to facilitate transition of translational workflows into the EMR.

The evolution of clinical informatics as a subspecialty in both nursing and medicine occurred in parallel with the widespread adoption of EHRs in both hospitals and the ambulatory setting, attributed to the “Meaningful Use” components of the American Reinvestment and Recovery Act (ARRA) of 2009. Meaningful Use had multiple stages, requiring increased health IT utilization in clinical practice, and was further expanded, and later renamed, as “Promoting Interoperability,” in 2018. These regulatory requirements, combined with increased throughput demands and the reality of worsening burnout, as discussed above, have combined to create a popular focus of administrative intensity that demands continuous optimization of clinical workflows.

Regardless of the cause for the need for optimization of clinical workflows, the standardization of clinical processes is beneficial to biomedical research in oncology. Recognizing that without abundant resources to comb through paper records there is an inherent benefit to standardized electronic documentation when using clinical records for biomedical research, the required integration and utilization of electronic processes for clinical practice can feed research requirements. For example, data on prescriptions would be difficult to obtain in large volumes if it was on paper where documentation may not have included prescriptions handwritten and given to the patient. Now that electronic prescribing of medications is becoming standard practice, along with the standardized terminologies (RxNorm (standard for naming drugs), SNOMED-CT (standard utilized for pathological terms) that are required for its use, the data becomes more accessible and standardized across healthcare organizations [50].

In a paper state, or even a simplified like-for-like paper to EHR conversion where much of the data is kept in a nondiscrete format, to collect data means large scale abstraction efforts constituting various human resources across multiple departments including medical records, billing and coding, regulatory compliance, and clinical research. This is an error-prone process, with a 30% error rate per study [51–53]. The knowledge of these human abstractors can be used to develop the automated tools based on discrete data mapping where possible. Maintenance of the rules and mapping involved will be dependent on the continued involvement of these subject matter experts in each of these areas, and therefore these are

also key stakeholders that should be included in the EHR governance models and subsequently stakeholders in the governance models for translational research using the EHR.

It is worth noting, too, that the subject matter expertise needed for the processing of paper records must be maintained, as paper is the ultimate “downtime form” of the entire chart. Knowledge of how to process these paper workflows will be needed for the foreseeable future. Technically, though, there is no system that can guarantee constant uptime, not even paper. Paper may be lost, destroyed due to fire, or may not be readily available, such as in the case of workflows requiring specialized paper like the Florida controlled substances prescribing laws [54]. When accounting for all data points available, these ancillary processes must be considered. Wherever there is electronic data there must be recognition of the downtime process with the appropriate staffing for data capture related to it in order to be complete. In addition, people are needed to maintain clinical workflows and throughput, including the data of the chart. In periods of downtime processes must be utilized that do not add to the workload of the healthcare workers (clinicians, midlevel practitioners) while collecting data for translational research.

Clinicians, by training and by practice, will prioritize patient care above all else, including cumbersome EHR workflow processes [55–59]. Clinical informaticists recognize this and contribute their clinical insights to their assessments and development of health IT integrations and optimization to facilitate both the patient care and the organizational goals, often including comprehensive and accurate documentation of said care with variable outcomes [60–65]. Completely seamless integration is rare, but the more that the documentation is a product of the care provided, the more likely it is to be both complete and accurate with discrete fields available for abstraction and translational studies.

While the conversion and optimization of existing health IT into clinical workflows may have obvious benefits from clinical informaticists’ contribution, less clear may be the application of clinical informatics in an industry that appears to be on the cusp of an artificial intelligence (AI) revolution. Artificial intelligence systems have great promise in the field of developing new theories for drug development and generally fall into two classes, expert systems and machine learning systems. Expert systems are generally useful for developing clinical workflows such as in pathways, while machine learning and artificial intelligence methods are useful for hypothesis generation, including those needed to aid in translating results of murine experiments to humans, and ultimately may lead to expert systems to support pathology and other branches of medicine [66, 67]. There are new articles and reports being produced at an increasing pace on the growth and development of AI and the use of machine learning (ML). While machine learning in

the form of natural language processing (NLP) is a widely recognized and longstanding component of more advanced computational work in clinical documentation analysis, it too is seen as part of the new wave as it is applied to more and more direct clinical uses, despite having been under active study for over 30 years [68–70]. While there appears to be clear narrow use cases for AI in the near future, there is still much to be defined for the use of AI (ML, NLP, etc.) [71]. Artificial intelligence processes, such as machine learning and expert systems, have the potential to many things including transforming hypothesis generation from data warehouses, developing expert systems, creating the clinical workflows that will be needed for providing oncology-based care, and developing new hypotheses to translate findings in the laboratory to the clinic.

---

### 3 Initial Assessment of Workflows for Translational Informatics

Existing publications review proposed research IT maturity models for academic health centers [72]. While critical to an organization’s development and growth, the IT infrastructure must also be supplemented by clinician and clinical informatics expertise reviewed above for both clinical and translational informatics workflows.

#### 3.1 Governance

In assessing an organization’s current state of clinical informatics infrastructure, it helps to ask first where the decision-making occurs. Many clinicians and researchers who are focused on their own workflows may not be aware of the sometimes-Byzantine architecture of councils and committees that make the large-scale decisions that have direct impact on their practice. By looking at who and where decisions are made about what changes occur in the EHR, or new applications are added to the clinical environment, it will show where the organization may best apply high-yield interventions to improve this process.

While there have been mixed results in the application of “lean” manufacturing techniques to healthcare workflows, there are still lessons to be learned from its methodologies [73–75]. One of those is recognizing that those who know the most about a process are those who perform it. Assessing this is often called “going to the gemba,” as in where the work is done [76]. Using this principle when mapping out workflows means asking the individuals who do the work what actual steps, both intended and unintended, exist in their processes. The same may be applied to following an organization’s change governance model.

It should be determined where an organization is along the continuum of “no governance,” where changes are made without oversight and minimal review, to the extreme of “obstructive governance,” where no changes are allowed, even those that would

have beneficial impact to clinicians. Strong EHR governance should not be functioning as a barrier to EHR changes by clinicians, but a facilitator. The governance should facilitate decision-making and oversight of clinician-impacted changes by those same clinicians or their representatives. The documentation provided as an output of this governance serves those who come after in understanding why certain changes were made before. Inherently, through user-driven governance, modifications to the EHR, so deeply intertwined in clinical workflows, happens for physicians, nurses, and their colleagues, and not put upon them by well-intended administrators.

Strong, healthy change governance may not exist in an organization that lacks control over its system. In an organization without the core knowledgeable IT resources needed to build and maintain the EHR there is no opportunity to make improvements to the system. Unfortunately, many organizations' health IT departments are still too understaffed to meet the highest stages of available health IT integration [77]. If an organization is unable to sustain its existing application infrastructure it will be difficult to find the resources to optimize, let alone add, new applications to the clinical environment. Fortunately, at Moffitt we are now staffed with a core group of clinical informaticians and support staff to facilitate the implementation of these interfaces and a clinical informatics governance council currently cochaired by the authors.

### **3.2 Workflows and Staffing**

As the clinical and research staff are becoming more dependent on the IT resources that have become integrated in their workflows, it is critical that organizations invest in developing and maintaining their core IT teams. This is the first dependency for developing the clinical informatics model needed to facilitate research at any organization.

In order to assess clinical throughput, it is critical that the teams doing this work have access to the objective data needed to analyze workflows. Certain workflows will lend themselves better to data reporting than others, and some are dependent on more technical capabilities than most sites have, that is, radio frequency identification (RFID) for patient tracking in perioperative care coordination [78]. In identifying the various metrics for workflow analyses and measuring change outcomes, the clinical informatics team will need to work with their translational bioinformatics colleagues in analyzing the “big data” output of large cohort data management. In our institution, this is coordinated through a shared services model that facilitates the biomedical informatics resources used for both translational bioinformatics and the data analyses for basic science laboratory output.

As previously discussed, another staffing resource needed is that of the clinician subject matter experts, those who utilize the workflows in question. Prior to and during the initial stages of the

regulatory and legal requirements for EHR conversion as demanded by the ARRA of 2009, lack of clinician readiness for adoption of electronic health records was a barrier to adoption prior to the regulatory and legal requirements of electronic health records [79, 80]. While this may have been the case with the initial conversion and application of EHRs, many of our providers (physicians and advanced practice practitioners such as physician assistants and advanced practice registered nurses), staff nurses, and other clinical colleagues have seen the benefit to participating in the improvement of these clinical systems. In organizations that apply the information provided by the clinicians who use EHRs we have seen a positive feedback loop of involvement. Providers believe they know what needs to be fixed in the EHR and should be given the tools to do so [81]. Empowering the clinician user-base facilitates improved design of EHR workflows, thereby leading to the improved documentation previously mentioned. A vocal and involved clinical staff support the development of the clinical informatics resources needed for growing bioinformatics research. The climate at each institution needs to be assessed prior to transforming the EHRs into tools capable of capturing the discrete data points needed for translational research.

Among the key stakeholders whose voice is crucial to making any changes to the EHR clinical workflows, in a clinical research institution, such as an oncology center, research must be included in all aspects of EHR change governance. While both provider and patient are mortal, the data is not, and to be effective stewards of the data that patients and our colleagues have provided we must apply many kinds of expertise to the data's collection. When clinician-researchers, public health researchers such as epidemiologists, statisticians, and translational researchers are available and able to participate in these processes their insights as key stakeholders will facilitate decisions in a way that would not be possible with a unilateral clinical perspective. Data stewardship includes concerns of data maintenance of data from patients both past and present. Previous studies have looked at the stability of data, much depending on the physical format of data storage, but additionally the value of backward compatibility of systemic changes must be weighed against the need for expedient changes in the usability of concurrent data for active patient care. Coordination of system changes with existing data governance structures in and outside of the research domain is another key argument for the inclusion of research representation in EHR governance models. The operational use of healthcare data must complement existing research data stewardship models [82].

Even with a comprehensive and engaged change governance structure consisting of clinicians, researchers, and clinical informatics experts, there will still be ongoing optimization and application of health IT needed to improve clinical workflows to the benefit of

patients, clinicians, and researchers. This work itself, the application of clinical informatics in an operational way, can be studied and used to provide the evidence needed to inform the ongoing growth of the field. Being a relatively young subspecialty there is a need for more studies on the application and use of EHRs, a technology which was directly and broadly applied to a field with nominal real world evidence for its benefits [83]. Multiple groups have advised on the best practices for some aspects of EHRs, such as clinical decision support (CDS), but there is little literature on best practices for specific functionality such as order sets themselves [84]. Further nuanced settings and requirements, such as the best practice development of order sets at comprehensive cancer centers, is lacking in the existing literature. Development of best practice standards for use of the EHR in translational research is still in its infancy, but we believe the key factors for success are having the right governance structures, comprehensive data stewardship, and a supportive institutional culture to incorporate research into its clinical workflows without disrupting the care of patients.

At Moffitt Cancer Center, we have found an operational model for clinical informatics that is supported by experienced nurse informaticists who are able to coordinate small to moderate-scale operational EHR changes on the scale of as many as 20 working requests each, at any given time, in various stages of development, after approval and prioritization by the clinical informatics council (s). We also have specific clinical informaticists in the areas of operational management of pathology and translational research with developing expertise in radiology. Concurrent projects will inherently require clinical informatics expertise which is also facilitated by these same nurse informaticists on a scale of 2–3 organization-wide or departmental projects, again at various stages of development. Our senior informatics leadership includes Randa Perkins, MD, FAMIA, chief medical information officer (CMIO) and senior director of clinical informatics, Marc Perkins-Carrillo, MSN, RN-BC, director of clinical informatics, Tina Dieckhaus, MSN, RN-BC, manager of clinical informatics, Mandy O’Leary, MD, MPH, director of pathology informatics, and Joseph Markowitz, MD, PhD (translational informatics and co-chair of the clinical informatics council). We also benefit from strong support from our public health colleagues as well as bioinformatics and biostatistics.

---

## 4 Examples of Translational Informatics Projects

As described in detail in Chap. 3, Moffitt Cancer Center has benefited from having the Total Cancer Care Protocol that provides for patient samples to be collected when not otherwise needed for clinical care of patients and a data warehouse for translational research purposes [85]. Partnering with the company M2Gen,

the operational arm of ORIEN (Oncology Research Information Exchange Network), where multiple cancer centers collaborate and there are plans in place to help facilitate data sharing, a clinical trial was possible to facilitate a clinical trial cohort for CD30-positive tumors [86, 87]. These previous efforts represent important steps and will continue to be extremely valuable in the realm of identifying genetic alterations suitable for clinical trials, but there are other translational questions that are also useful to explore with strict protocols in place to protect patient information as permitted by institutional review boards.

Research interfaces with the data warehouse and EHR are currently in development. At the 2019 Annual Symposium of the American Medical Informatics Association (AMIA) a system tool was demonstrated, called Patient Explorer, that allows the user to have a timeline of imaging performed, clinic visit, and cancer treatment infusion [88]. This effort represents the first step toward having automated cohort identification from biobanks linked to data warehouses. It is the opinion of the authors that entry criteria to these key preclinical studies will eventually incorporate machine learning techniques and research is currently underway to facilitate this effort. With respect to the EHR, there is a field in the front end to the EHR (PowerChart) to clearly indicate whether people have enrolled in the TCC. Furthermore, the authors of this manuscript have facilitated the work to increase awareness and transparency of clinical trials enrollment by securing a visible, real-time feed from the clinical trials management system (CTMS) into the primary EHR. This data point ensures that all medical staff are aware of clinical trial participation to maximize adherence to clinical trial protocols, decreasing deviations.

So far, we have discussed implementation of biobanks and associated data warehouses for translational research. However, research is also needed in identifying appropriate practices into the EHR and clinical guidelines to facilitate both routine care and participation in translational research such as clinical trials. One of the initial reasons to formulate clinical pathways was to measure adherence to evidence-based medicine practices [89]. The rationale for clinical pathways has evolved and the ability to transform clinical guidelines into different EHRs should be model driven, but this is complex given that most EHRs are commercial entities with poor cross-vendor portability or interfacing [90]. Although much of the recent work has been done with commercially available packages there was a similarly large effort in the academic world in the early 2000s looking into the clinical pathway engines prior to the work being undertaken by the industry sector [91–98]. Clinical pathway development is more than the development of an expert system for a specific clinical algorithm as the foundation algorithm may change with new evidence, and new clinical trials may be incorporated into the pathway. However, it is clear from the literature that the work

on clinical expert systems is a prerequisite for the development of clinical pathways [99–101]. For a full description of modern pathway development utilizing one of the commercial vendors we refer you to Chap. 4.

---

## 5 Future Directions

As EHRs are being molded into tools to enhance clinical practice, best practices continue to be defined and refined, but few operational teams are taking the time to publish. Most research is coming out of large institutions with academic informaticians, whose work is invaluable and is being directly applied by operational clinical informaticians but does not always reflect the breadth of practice needs in the distributed health information technology ecosystem.

As an organization builds processes and systems to meet its clinical and research needs it will quickly find itself on the cutting edge of technology standards, or become aware of the lack thereof. An academic cancer center may be developing tests for proteins that do not have a codified value in any of the standard EHR terminologies. Even accounting for the available molecular biology databases there are not standard pivot points between the standard EHR terminologies and the molecular biology or genomics terminologies. Ontology expertise will either need to be found or developed in the organization. This is where the applied skills of colleagues within medical records, billing, and coding can be applied to the benefit of the organization. Where previously they may have been performing manual abstraction, their knowledge of medical coding and ontologies can be relied on to help design and maintain the rules and interfaces between systems, as well as conversion of nondiscrete data to mapped, discrete terms.

There are opportunities for growth in this area and there may be capabilities with platforms that act as a bridge between databases with real-time terminology feeds. For example, the Online Mendelian Inheritance in Man (OMIM) catalogue of human genes and genetic disorders, offers an application programming interface (API) providing detailed information about specific genes. APIs like this one could be referenced by an application from within a patient's chart, such as a Substitutable Medical Applications and Reusable Technologies (SMART) application built on Fast Health Interoperability Resources (FHIR) that synchronizes the patient's specific data with relevant information from APIs like OMIM [102]. The crux of applying a process concept like this is determining which data is relevant to the clinicians and patients. This will be the purview of the clinical teams and patients that make up the governance of health IT in each organization, as they decide what is clinically significant, with contributions from clinical staff, research, and all affiliated disciplines, with guidance from clinical informatics.

Similarly, exposing all of the raw genomic and proteomic data of a patient and their tumors into an EHR could easily overwhelm any of the major EHR systems available today. To avoid losing the “signal in the noise” teams will have to maintain a data ecosystem that allows for continuous cycling of clinically relevant data into the chart, and suppression of less clinically relevant data from the clinical workflow as new research develops. The regulation of this suppression and exposure process may be internal to an organization or could be developed and maintained by a third party as part of another API feed into the data aggregation process, but at the time of this publication’s writing is not currently available in the major EHR market.

These systems, with their promise and complexity, will not build and maintain themselves. There should be no false hope, either, that they will organically coalesce into something that is both benign to the clinician and beneficial to patient, clinician, and researcher. It will take knowledgeable, integrated, and empowered expertise using clinical informatics to create systems that benefit and sustain the bioinformatics research needed to “[c]ontribute to the prevention and cure of cancer [103].”

Utilization of these informatics principles in the domain of the EHR will facilitate translational research in the traditional sense of drug development, public health studies, and optimizing clinical care for patients. Ultimately, informatics resources lead to better usage of biobank repositories, data warehouses, and the EHR, as well as more efficient development of therapeutic trials. These efforts demonstrate the reason why clinical informatics exists, to better use technology to enhance the experience of patient care for the benefit of all involved.

## References

1. Detmer DE, Lumpkin JR, Williamson JJ (2009) Defining the medical subspecialty of clinical informatics. *J Am Med Inform Assoc* 16(2):167–168
2. Medicine ABoP (2017) Clinical informatics 2017 examination blueprint. Core content of the clinical informatics subspecialty. <https://www.theabpm.org/wp-content/uploads/2017/09/2017CI-Content-Outline.pdf>. Accessed 09 Jan 2019
3. Gardner RM, Overhage JM, Steen EB et al (2009) Core content for the subspecialty of clinical informatics. *J Am Med Inform Assoc* 16(2):153–157
4. Detmer DE, Shortliffe EH (2014) Clinical informatics: prospects for a new medical subspecialty. *JAMA* 311(20):2067–2068
5. Greenes RA, Shortliffe EH (1990) Medical informatics. An emerging academic discipline and institutional priority. *JAMA* 263(8):1114–1120
6. Bernstam EV, Smith JW, Johnson TR (2010) What is biomedical informatics? *J Biomed Inform* 43(1):104–110
7. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L (2011) Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc* 18(4):354–357
8. Kulikowski CA, Shortliffe EH, Currie LM et al (2012) AMIA board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *J Am Med Inform Assoc* 19(6):931–938

9. Friedman CP (2009) A “fundamental theorem” of biomedical informatics. *J Am Med Inform Assoc* 16(2):169–170
10. Weed LL (1971) The problem oriented record as a basic tool in medical education, patient care and clinical research. *Ann Clin Res* 3(3):131–134
11. Henry J, Pylypchuk Y, Searcy T, Patel V (2016) Adoption of electronic health record systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. *ONC Data Brief* 35
12. Evans RS (2016) Electronic health records: then, now, and in the future. *Yearb Med Inform (Suppl 1)*:S48–S61
13. Institute of Medicine (2003) Key capabilities of an electronic health record system: letter report. The National Academies Press, Washington, DC. <https://doi.org/10.17226/10781>
14. Schreiweis B, Trinczek B, Kopcke F et al (2014) Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. *Int J Med Inform* 83(11):860–868
15. Bruland P, McGilchrist M, Zapletal E et al (2016) Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol* 16 (1):159
16. The Innovative Medicines Initiative (2016) Electronic health records for clinical research (2011–2016). <http://www.ehr4cr.eu/>. Accessed 09 Jan 2019
17. Daniel C, Kalra D, Section Editors for the IYSO-CRI (2018) Clinical research informatics: contributions from 2017. *Yearb Med Inform* 27(1):177–183
18. CMS (2019) Certified EHR technology. <https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Certification.html>. Accessed 09 Jan 2019
19. Waegemann CP (2003) Ehr vs. cpr vs. emr. *Healthc Inform Online* 1:1–4
20. Hillestad R, Bigelow J, Bower A et al (2005) Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff (Millwood)* 24 (5):1103–1117
21. Stair TO (1998) Reduction of redundant laboratory orders by access to computerized patient records. *J Emerg Med* 16(6):895–897
22. Wilson GA, McDonald CJ, McCabe GP Jr (1982) The effect of immediate access to a computerized medical record on physician test ordering: a controlled clinical trial in the emergency room. *Am J Public Health* 72 (7):698–702
23. Chen J, Wang Y, Magrabi F (2017) Downtime in digital hospitals: an analysis of patterns and causes over 33 months. *Stud Health Technol Inform* 239:14–20
24. Harrison AM, Siwani R, Pickering BW, Herasevich V (2019) Clinical impact of intraoperative electronic health record downtime on surgical patients. *J Am Med Inform Assoc* 26(10):928–933
25. Caesar MC, McIntaggart S (2015) IT downtime - a cultural shift. *Healthc Q* 18(1):43–47
26. Becker M, Goldszal A, Detal J, Gronlund-Jacob J, Epstein R (2015) Managing a multi-site academic-private radiology practice Reading environment: impact of IT down-times on Enterprise efficiency. *J Am Coll Radiol* 12(6):630–637
27. Khatri N (2006) Building IT capability in health-care organizations. *Health Serv Manag Res* 19(2):73–79
28. Khatri N, Gupta V (2016) Effective implementation of health information technologies in U.S. hospitals. *Health Care Manag Rev* 41 (1):11–21
29. Viceconti M (2015) Biomechanics-based in silico medicine: the manifesto of a new science. *J Biomech* 48(2):193–194
30. Besiso A, Patrick JD, Dip G, Ho V, Cheng Y (2018) The impact of an enterprise electronic medical record (EEMR) model vs a clinical information system (CIS) model on usability, efficiency, and adaptability. *AMIA Annu Symp Proc* 2018:242–251
31. Ratwani RM, Hettinger AZ, Fairbanks RJ (2017) Barriers to comparing the usability of electronic health records. *J Am Med Inform Assoc* 24(e1):e191–e193
32. Tan BW, Lo TW (1991) The impact of interface customization on the effect of cognitive style on information system success. *Behav Inform Technol* 10(4):297–310
33. Moon MC, Hills R, Demiris G (2018) Understanding optimization processes of electronic health records (EHR) in select leading hospitals: a qualitative study. *J Innov Health Inform* 25(2):109–125
34. Ratwani RM, Savage E, Will A et al (2018) A usability and safety analysis of electronic health records: a multi-center study. *J Am Med Inform Assoc* 25(9):1197–1201
35. Calvitti A, Farber N, Chen Y et al (2012) Temporal analysis of physicians' EHR workflow during outpatient visits. Paper presented at: 2012 IEEE second international

- conference on healthcare informatics, imaging and systems biology, 27–28 Sept 2012
36. Estes S, Helleberg J, Long K et al (2018) Principles for minimizing cognitive assistance distraction in the cockpit. Paper presented at: 2018 IEEE/AIAA 37th digital avionics systems conference (DASC); 23–27 Sept 2018
  37. Street RL Jr, Liu L, Farber NJ et al (2014) Provider interaction with the electronic health record: the effects on patient-centered communication in medical encounters. *Patient Educ Couns* 96(3):315–319
  38. Gardner RL, Cooper E, Haskell J et al (2019) Physician stress and burnout: the impact of health information technology. *J Am Med Inform Assoc* 26(2):106–114
  39. Meyerhoefer CD, Sherer SA, Deily ME et al (2018) Provider and patient satisfaction with the integration of ambulatory and hospital EHR systems. *J Am Med Inform Assoc* 25(8):1054–1063
  40. Melnick ER, Dyrbye LN, Sinsky CA et al (2020) The association between perceived electronic health record usability and professional burnout among US physicians. *Mayo Clin Proc* 95(3):476–487
  41. Imai M (2012) Gemba Kaizen: a common-sense approach to a continuous improvement strategy, 2nd edn. McGraw Hill, New York
  42. Christopher AS, Smith CS, Tivis R, Wilper AP (2014) Trends in United States physician work hours and career satisfaction. *Am J Med* 127(7):674–680
  43. Zolhavarieh S, Parry D, Bai Q (2017) Issues associated with the use of semantic web technology in knowledge acquisition for clinical decision support systems: systematic review of the literature. *JMIR Med Inform* 5(3):e18
  44. Goble C, Stevens R (2008) State of the nation in data integration for bioinformatics. *J Biomed Inform* 41(5):687–693
  45. Gordon CL, Weng C (2015) Combining expert knowledge and knowledge automatically acquired from electronic data sources for continued ontology evaluation and improvement. *J Biomed Inform* 57:42–52
  46. Payne T (2014) Practical guide to clinical computing systems. Elsevier, Boston, MA
  47. Myrick K, Ogburn D, Ward B (2017) Percentage of office-based physicians using any electronic health record (EHR)/electronic medical record (EMR) system and physicians that have a certified EHR/EMR system, by U.S. state: National Electronic Health Records Survey, 2017. [https://www.cdc.gov/nchs/data/nehrs/2017\\_NEHRS\\_Web\\_Table\\_EHR\\_State.pdf](https://www.cdc.gov/nchs/data/nehrs/2017_NEHRS_Web_Table_EHR_State.pdf). Accessed 09 Jan 2019
  48. Sittig DF, Ash J (2011) Clinical information systems: overcoming adverse consequences. Jones and Bartlett, Sudbury, MA
  49. Yuan CT, Bradley EH, Nemphard IM (2015) A mixed methods study of how clinician 'super users' influence others during the implementation of electronic health records. *BMC Med Inform Decis Mak* 15:26–26
  50. Bodenreider O, Cornet R, Vreeman DJ (2018) Recent developments in clinical terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 27(1):129–139
  51. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP (2006) Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol* 59(7):697–703
  52. Gotzsche PC, Hrobjartsson A, Maric K, Tendal B (2007) Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 298(4):430–437
  53. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL (2005) High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol* 58(7):741–742
  54. Statutes FS (2018) 456.42 Written prescriptions for medicinal drugs. [http://www.leg.state.fl.us/Statutes/index.cfm?App\\_mode=Display\\_Statute&Search\\_String=&URL=0400-0499/0456/Sections/0456.42.html](http://www.leg.state.fl.us/Statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0400-0499/0456/Sections/0456.42.html). Accessed 3 Dec 2019
  55. Thompson WT, Cupples ME, Sibbett CH, Skan DI, Bradley T (2001) Challenge of culture, conscience, and contract to general practitioners' care of their own health: qualitative study. *BMJ* 323(7315):728–731
  56. Talty PM (1985) Time management in clinical practice. *Occup Ther Health Care* 2(4):95–104
  57. Spears BW (1981) A time management system for preventing physician impairment. *J Fam Pract* 13(1):75–80
  58. Sackett DL (2011) Clinician-trialist rounds: 2. Time-management of your clinical practice and teaching. *Clin Trials* 8(1):112–114
  59. Smeltzer CH, Hines PA, Beebe H, Keller B (1996) Streamlining documentation: an opportunity to reduce costs and increase nurse clinicians' time with patients. *J Nurs Care Qual* 10(4):66–77
  60. Jamieson T, Ailon J, Chien V, Mourad O (2017) An electronic documentation system improves the quality of admission notes: a

- randomized trial. *J Am Med Inform Assoc* 24(1):123–129
61. Jefferies D, Johnson M, Griffiths R (2010) A meta-study of the essentials of quality nursing documentation. *Int J Nurs Pract* 16(2):112–124
62. Fanucchi L, Yan D, Conigliaro RL (2016) Duly noted: lessons from a two-site intervention to assess and improve the quality of clinical documentation in the electronic health record. *Appl Clin Inform* 7(3):653–659
63. Neri PM, Volk LA, Samaha S et al (2014) Relationship between documentation method and quality of chronic disease visit notes. *Appl Clin Inform* 5(2):480–490
64. Burke HB, Hoang A, Becher D et al (2014) QNOTE: an instrument for measuring the quality of EHR clinical notes. *J Am Med Inform Assoc* 21(5):910–916
65. Slaughter SE, Hill JN, Snelgrove-Clarke E (2015) What is the extent and quality of documentation and reporting of fidelity to implementation strategies: a scoping review. *Implement Sci* 10:129
66. Brubaker DK, Proctor EA, Haigis KM, Lauffenburger DA (2019) Computational translation of genomic responses from experimental model systems to humans. *PLoS Comput Biol* 15(1):e1006286
67. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A (2019) Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 16(11):703–715
68. Kreimeyer K, Foster M, Pandey A et al (2017) Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 73:14–29
69. Yim WW, Yetisen M, Harris WP, Kwan SW (2016) Natural language processing in oncology: a review. *JAMA Oncol* 2(6):797–804
70. Hirschberg J, Manning CD (2015) Advances in natural language processing. *Science* (New York, N.Y.) 349(6245):261–266
71. Allen B Jr, Seltzer SE, Langlotz CP et al (2019) A road map for translational research on artificial intelligence in medical imaging: from the 2018 National Institutes of Health/RSNA/ACR/the academy workshop. *J Am Coll Radiol* 16(9 Pt A):1179–1189
72. Knosp BM, Barnett WK, Anderson NR, Embi PJ (2018) Research IT maturity models for academic health centers: early development and initial evaluation. *J Clin Transl Sci* 2(5):289–294
73. Moraros J, Lemstra M, Nwankwo C (2016) Lean interventions in healthcare: do they actually work? A systematic literature review. *Int J Qual Health Care* 28(2):150–165
74. Barnas K (2011) ThedaCare's business performance system: sustaining continuous daily improvement through hospital management in a lean environment. *Jt Comm J Qual Patient Saf* 37(9):387–399
75. Lot LT, Sarantopoulos A, Min LL, Perales SR, Boin I, Ataide EC (2018) Using lean tools to reduce patient waiting time. *Leadersh Health Serv (Bradf Engl)* 31(3):343–351
76. Bourgault AM, Upvall MJ, Graham A (2018) Using Gemba boards to facilitate evidence-based practice in critical care. *Crit Care Nurse* 38(3):e1–e7
77. Hersh WR, Boone KW, Totten AM (2018) Characteristics of the healthcare information technology workforce in the HITECH era: underestimated in size, still growing, and adapting to advanced uses. *JAMIA Open* 1(2):188–194
78. Southard PB, Chandra C, Kumar S (2012) RFID in healthcare: a six sigma DMAIC and simulation case study. *Int J Health Care Qual Assur* 25(4):291–321
79. Poon EG, Blumenthal D, Jaggi T, Honour MM, Bates DW, Kaushal R (2004) Overcoming barriers to adopting and implementing computerized physician order entry systems in U.S. hospitals. *Health Aff* 23(4):184–190
80. Cohen MR (2015) The challenge of EHR acceptance by physicians. *J Med Pract Manag* 31(2):117–120
81. Stanford University, Harris Poll (2018) How doctors feel about electronic health records: national physician poll by the Harris Poll. <http://med.stanford.edu/content/dam/sm/ehr/documents/EHR-Poll-Presentation.pdf>. Accessed 1 Sept 2019
82. Jansen P, van den Berg L, van Overveld P, Boiten JW (2019) Research data stewardship for healthcare professionals. In: Kubben P, Dumontier M, Dekker A (eds) *Fundamentals of clinical data science*. Springer, Cham (CH), pp 37–53
83. National Academies of Sciences, Engineering, and Medicine (2017) Real-world evidence generation and evaluation of therapeutics: proceedings of a workshop 2017. <https://doi.org/10.17226/24685>. Accessed 19 Dec 2019
84. McGreevey JD 3rd. (2013) Order sets in electronic health records: principles of good practice. *Chest* 143(1):228–235

85. Fenstermacher DA, Wenham RM, Rollison DE, Dalton WS (2011) Implementing personalized medicine in a cancer center. *Cancer J* 17(6):528–536
86. Li B, Eschrich SA, Berglund A et al (2017) Use of the total cancer care system to enrich screening for CD30-positive solid tumors for patient enrollment into a Brentuximab Vedotin clinical trial: a pilot study to evaluate feasibility. *JMIR Res Protoc* 6(3):e45
87. Wenham RM, Sullivan DM, Hulse M, Jacobsen PB, Dalton WS (2012) The creation of an integrated health-information platform: building the framework to support personalized medicine. *Pers Med* 9(6):621–632
88. Carvajal R, Gonzalez-Calderon G, Betin-Montes M et al (2019) Patient timelines for research-oriented exploration of longitudinal cancer patient data: PT explorer. Paper presented at: American Medical Informatics Association annual symposium 2019, Washington, D.C.
89. Konrad R, Tulu B, Lawley M (2013) Monitoring adherence to evidence-based practices: a method to utilize HL7 messages from hospital information systems. *Appl Clin Inform* 4 (1):126–143
90. Bockmann B, Heiden K (2013) Extracting and transforming clinical guidelines into pathway models for different hospital information systems. *Health Inf Sci Syst* 1:13
91. Boxwala AA, Peleg M, Tu S et al (2004) GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *J Biomed Inform* 37(3):147–161
92. Peleg M, Boxwala AA, Bernstam E, Tu S, Greenes RA, Shortliffe EH (2001) Sharable representation of clinical guidelines in GLIF: relationship to the Arden syntax. *J Biomed Inform* 34(3):170–181
93. Peleg M, Boxwala AA, Ogunyemi O et al (2000) GLIF3: the evolution of a guideline representation format. *Proc AMIA Symp*:645–649
94. Peleg M, Ogunyemi O, Tu S et al (2001) Using features of Arden syntax with object-oriented medical data models for guideline modeling. *Proc AMIA Symp*:523–527
95. Peleg M, Patel VL, Snow V et al (2002) Support for guideline development through error classification and constraint checking. *Proc AMIA Symp*:607–611
96. Wang D, Peleg M, Tu SW et al (2004) Design and implementation of the GLIF3 guideline execution engine. *J Biomed Inform* 37 (5):305–318
97. Wang D, Shortliffe EH (2002) GLEE—a model-driven execution system for computer-based implementation of clinical practice guidelines. *Proc AMIA Symp*:855–859
98. Peleg M, Boxwala AA, Tu S, Greenes RA, Shortliffe EH, Patel VL (2001) Handling expressiveness and comprehensibility requirements in GLIF3. *Stud Health Technol Inform* 84(Pt 1):241–245
99. Lane CD, Walton JD, Shortliffe EH (1986) Graphical access to medical expert systems: II. Design of an interface for physicians. *Methods Inf Med* 25(3):143–150
100. Tsuji S, Shortliffe EH (1986) Graphical access to medical expert systems: I. design of a knowledge engineer's interface. *Methods Inf Med* 25(2):62–70
101. Shortliffe EH (1986) Medical expert systems—knowledge tools for physicians. *West J Med* 145(6):830–839
102. Braunstein ML (2019) Health care in the age of interoperability part 6: the future of FHIR. *IEEE Pulse* 10(4):25–27
103. H. Lee Moffitt Cancer Center and Research Institute (2019) Moffitt mission statement. <https://moffitt.org/about-moffitt/>. Accessed 3 Dec 2019



# Chapter 2

## Leveraging Pathology Informatics Concepts to Achieve Discrete Lab Data for Clinical Use and Translational Research

Mandy Flannery O'Leary

### Abstract

Clinical practice is most efficient when physicians have the right information, including pathology and laboratory results, at the point of contact with the patient. In downstream workflows, subsequent groups using lab data want to have it available in a format that is easy to manipulate. With the complexity of electronic medical records, hospital information systems, and the need to accommodate data from outside systems, this is not easy to accomplish. By utilizing a group of concepts from clinical and pathology informatics, system implementations may be improved to achieve relevant laboratory data in a format that is usable by healthcare entities to improve patient care and forward endeavors in precision medicine.

**Key words** Pathology informatics, Discrete data, Lab workflows, Synoptic reporting, Laboratory information system (LIS)

---

### 1 Introduction

Physicians use clinical medical reasoning and decision-making with information garnered from histories, physicals, and diagnostic testing. Using the data gathered from this daily practice with the help of clinical trials and research endeavors, the medical world is seeing improvement in treating specific diseases. But to have the greatest impact on patient care and efficiency in healthcare, the physician needs to have the right data at the right time to make the right diagnosis, especially in regard to provision of laboratory results. This is often difficult for a myriad of reasons, such as lack of shared data resources or tools, appropriate process workflows, or system integration. Informatics processes may assist with these issues, with an end goal of expediting care by providing actionable data at the point of contact with the patient to allow the physician to make better medical decisions and improve patient safety and treatment outcomes.

When many physicians hear “informatics,” they shudder because they think of the difficulties associated with electronic health records (EHRs) or electronic medical records (EMRs) used with hospital information systems (HIS). First used in the 1960s, these computer systems were created and continue to be developed by a multitude of vendors to assist in managing medical and administrative information of hospitals. Different pieces of the HIS, including the lab systems, may be provided by different vendors, which requires data integration and data sharing between systems in order to provide data analysis downstream for medical and research purposes. While the physician may use a computerized physician order entry system (CPOE) to input orders, use the EHR to manage patients’ records and results, and use tools for clinical decision support (CDS) to guide diagnostic and treatment plans, other entities within an institution can leverage data across the HIS to perform research activities, resource planning, finance activities, or other administrative reporting. At a national level, data aggregated from the EHR allows for better epidemiology services in the public health sector and for oversight by regulatory agencies such as the Center for Medicare Services (CMS).

All of these nonclinical activities require system resources, which may divert an institution’s resources from optimizing the EMR for clinical use. Clinicians ask to see lab results in the patient’s EMR without having to click multiple times, experience multiple pop-up windows, or open extra links to other files, which creates inefficiency and causes further delays in patient care. They want to be able to trend results to view patient progress or see results presented in an integrated fashion, which requires data to be collected in a specific way. If the system is not built to be optimized to clinical use, EHRs can contribute to physician burnout and may require process redesign to make workflow improvements and decrease overall burden [1].

With the advent and growth of the informatics field, Informaticians may be consulted on best use practices for information management. Informatics is the science of information management and uses a multitude of processes and techniques to automate the collection, storage, utilization, and transmission of information. The American Medical Informatics Association (AMIA) defines biomedical and health informatics as informatics that applies principles of computer and information science to the advancement of life sciences, research, health professions education, public health, and patient care [2]. In addition, the growth of molecular and laboratory diagnostic tools and the vast multitude of pathology and laboratory information have increased the need for clinicians to have access to informatics tools and specialists. A subdiscipline, coined by Bruce Friedman, pathology informatics (PI) is described as the practice of informatics within pathology and includes the management and communication of lab

information for diagnosis, research, education, and other activities [3]. Although this subspecialty of informatics has been around for decades, it is still poorly understood by doctors within the specialty of pathology [4]; therefore, it could be surmised that clinicians in other medical disciplines are unaware of this subspecialty as well.

In fact, pathology, in general, is a “black box” to most clinicians; a specimen is sent to the lab, and reliable results are expected. Behind the scenes of the front desk or tube station by which specimens are sent, much occurs within the lab to provide and maintain tests and provide appropriate diagnostic results. For instance, considerations for on-boarding a new assay or updating an existing one include review of necessity and acquisition of new capital equipment, business analysis reviewing reimbursement rates for bringing a test in-house versus sending it out to a reference laboratory, liability and reliability of the assay including quality assurance aspects, and availability of licensed personnel with training including competency and proficiency testing. In addition to these pieces within the lab black box, there are required regulatory processes including validations, inspections, accreditation and licensing that the laboratory must meet; it has been suggested that pathology/lab is one of the most heavily regulated areas of health care [5]. Furthermore, with the large amount of lab results of different data types and required associated quality data, laboratory information systems (LIS) are employed in concert with EHRs to collate lab-generated information that may be used for clinical, quality, and/or research purposes, including personalized medicine initiatives in genomics [6]. As such, it would be useful for clinicians to be aware of what happens within the lab or PI when requesting lab data since their laboratory colleagues oversee and ensure all requirements are met in order to provide safe and correct results for patients and downstream research purposes.

There are multiple ways to accomplish implementation of builds for efficiency in the EHR and LIS to not only provide actionable diagnostic results to clinicians but to allow for discrete and searchable data for downstream research and development endeavors including precision medicine in therapeutics. This chapter will outline some suggested practices used in pathology informatics that may be used to achieve such data from laboratory processes.

---

## 2 Suggested Practices to Achieve Discrete Data from the Laboratory

### 2.1 Team Engagement

When approaching a lab data dilemma, it is important to understand that the solution may take multiple resources and team members. A first step may be engaging the appropriate team that has a stake in the successful solution of the issue at hand. Team members may include but are not limited to information

technology (IT) technical staff, laboratory staff, executive leadership, hospital administration such as quality and billing staff, and other clinical and research end users. Clinical subject matter experts (SMEs) should be involved to assist working through the issue since they may provide information and experience not readily available to all staff. Also, engagement of Clinical Informatics or PI groups may be helpful in navigating through the process and bridging expertise of team members from different areas.

## **2.2 Defining Foundational Lab Concepts**

Once the appropriate team is assembled, having everyone “on the same page and speaking the same language” is helpful. As described previously, many end users, including clinicians, who ask for new lab data or changes to how lab data is currently provided are unsure as to how the laboratory even functions. It is important to educate all the team members who will be involved to basic lab processes, since defining the issue requires some background. Foundational topics to review may include general laboratory definitions and workflows, test availability (are the assays in question performed in-house or sent elsewhere to a reference lab), and how the LIS and EHR may interact.

At a basic level, pathology departments and clinical laboratories oversee clinical, anatomic, and molecular pathology and may produce different types of data. The quality of medical data produced translates to its informational value, which heavily applies to laboratory data. Different data types include quantitative, qualitative, and descriptive textual data. Quantitative data may give a discrete numeric data value, which is frequently seen in clinical pathology, such as a potassium value. Qualitative data may have two or more exclusive or nonexclusive values, as may be seen in a drop-down choice list. Descriptive textual data is commonly seen in anatomic and molecular pathology by way of free-text reports used in interpretations. Image files are also a type of data.

All of these data types may be produced by the laboratory, depending on the lab’s test menu. The test menu for a laboratory describes the menu of assays that are performed by the lab, for example, complete blood counts and chemistry panels. There is a large number of lab assays, and as medicine improves, growth of the number of available assays is increasing. This growth is mostly seen in esoteric testing, or testing not widely available at a local or community setting that may be sent to an outside reference or specialty laboratory for analysis. To have results from esoteric testing be actionable or useful data, integration with the LIS and EHR is necessary [7]. To assist in this integration, it is useful to have a discrete orderable test for each lab test that is desired to be resulted. In other words, some assays do not have a discrete order and are ordered under a miscellaneous category. If there are a multitude of miscellaneous tests within the CPOE and EMR systems, it will be difficult to sort, view, and trend patient data. When a clinician can

order a test in a discrete fashion instead of ordering a miscellaneous test, this facilitates the eventual filing of the result in a more searchable way in the EMR under a respective category in lab results. Categories for lab results displays in the EMR may be based on the larger pathology entities of clinical, anatomic, and molecular pathology with further subcategories.

Clinical pathology (CP) employs large lab instruments that help analyze fluid specimens including blood, urine, and cerebro-spinal fluid. The simplified version of a clinical specimen workflow includes test selection and order entry in the CPOE/EMR by the ordering physician, which may include future and add-on orders. The LIS may contain dictionaries to help phlebotomy staff determine which specimen type to collect and will provide processes to print the appropriate tube labels. Once drawn, the lab specimen is received in the LIS for turnaround time tracking and status updates. If a specimen is barcoded, the LIS can direct the specimen for routing to the appropriate analyzers. Once analyzed, the LIS may assist in interpreting results through rules devised with lab oversight which have undergone validation processes. The LIS can route for further reflex testing if indicated. Lab results are typically interfaced from the LIS to the EMR. Results displays must be accurate and must be validated with the interface to match the proper patient and test order combination. These are requirements for accreditation for laboratories as dictated by the Clinical Laboratory Improvement Amendments or CLIA regulations [8].

Anatomic pathology (AP) traditionally handles solid tissue specimens which are prepared into histological slides for histologic examination by a pathologist who generates diagnostic reports. The workflow for an AP specimen is usually generated from collection of some tissue. After a tissue specimen is collected and received within the LIS, it is usually “grossed in” or described by a pathologist or pathologist assistant (PA), at which time tissue samples are taken for histology slide preparation. These samples are typically embedded in paraffin wax blocks, from which histology techs “cut” slides. After histology prepares and stains slides, Pathologists are able to histologically examine the tissue by microscopy and may order additional stains or molecular studies to assist in diagnosis. Once the pathologist is ready to report final diagnosis, he/she usually dictates a report that may be free-text or a defined selection from a checklist. This may also be coded text entry assisted by macros or templates from a data dictionary. Some reports may have an outline or “synoptic reporting” structure that may include structured data or discrete data elements that exist in a database, increasing the capability for data mining [9]. The final case sign out includes electronic signature by an attending pathologist locking and verifying the case in the LIS, which may trigger sending the result or report over to the EMR for clinician viewing.

Molecular pathology (MP) involves derivation of patient molecular data from methods such as Sanger sequencing and next-generation sequencing (NGS), which provide large data sets that require review by pathologists and bioinformatics specialists to discover mutations that may translate into pathologic findings. The complexity of datasets can be onerous, and more sophisticated software systems may be needed to handle data management [10]. In addition to the raw data files, analytic reports for this type of data may require reporting similar to AP. In this case, using free-text, as may be done in AP systems, again makes it difficult down the line to obtain data in a discrete fashion for research and reporting (business, regulatory, operational, etc.). If possible, use of discrete data for these cases may facilitate data mining.

The LIS assists in collating data; it is considered the largest analyzer in the laboratory since it essentially functions as the lab's "brain." The LIS is a conglomerate of hardware and software that assists in data processing and information management necessary to support functions within the modern lab [11]. Multiple interfaces between the LIS and other healthcare systems allow for data exchange, such as ADT or admission-discharge-transfer interfaces for patient management and order/results interfaces to transmit electronic orders through a CPOE system with results back into EMRs. Interfaces must be tested and validated to ensure proper transfer of data elements. Any new test or changes made to existing interfaces for instruments or software require testing and validation per laboratory standards.

These definitions of operations within the black box of pathology are only the very basic tools used by the lab. By providing at least this basic knowledge to them, the team will have a better understanding of the challenges the laboratory faces when implementing new services or updating existing services. When requests are submitted for lab changes by clinical, research, or administrative end users, having this knowledge will give the team a head start on devising a solution.

### **2.3 Requesting System Changes**

Issues and problems the clinicians and end users are facing regarding availability of lab data are usually raised and then presented for consideration of system changes. The ask is essentially what is being requested by the end user. The ask may be presented through multiple routes, including various ticket or project systems, and then the requests are usually reviewed and triaged by some governing structure, decided upon by each institution.

Once the ask or ticket request is approved, a version of the "waterfall model" may be used to guide the development process for the end user's ticket [12]. This model suggests six separate phases including: requirements analysis to define the requirements or why application is necessary, a specification phase where a user

defines requirements for the functionality of the system, a design phase including system architecture design, an implementation phase with development of the actual code for the application, a validation phase to install and test the system in real world and real time situations, and a maintenance phase which also involves verification that the software performs as designed and suggestions for further success. Various members of the team may be more heavily engaged at different steps in the system change process depending on their roles.

## **2.4 Understanding Current Processes**

Careful examination of the current workflow may allow the team to understand the ask and clarify the solution requirements. This review also may aid in identifying opportunities for corrections and improvements in current system processes. Keep in mind that while informatics and computerized tools may be useful in many instances, the computer system is not the answer to every issue encountered. Some workflows may be better accomplished through alternative manual processes. Furthermore, workflow design may differ due to the variety of systems, related workflows, and institutional goals [13].

## **2.5 Defining the Scope and Designing a Solution**

Once the current process has been reviewed and improvement opportunities have been identified, the team should work cooperatively to define why the change or new application is necessary and then research potential solutions. Requirements should be specified for the functionality of the system. The solution design should enable the proposed workflow to achieve the end user's ask. For example, this may include IT review of the workflow to assure the appropriate infrastructure such as hardware and networking exists for data to transmit appropriately. Working with all of the engaged teams and stakeholders will ensure the appropriate resources are in place.

Furthermore, in defining future workflow, it is important to have careful design around how to build and store laboratory data. With the growth in the number of lab tests available, institutions may not be able to keep up with test maintenance, and end users feel the effects of miscellaneous test orders and results. Usually, pathology results and reports may be stored within the LIS in a way so that they may be transmitted by a rich text format and/or a portable document format (PDF) report which can maintain the integrity of formatting and display. However, a PDF is stored as an image, and the results contained in the report are not able to be filed as discrete data elements within the EMR. One way to address this is to have discrete data elements for the report that may be sent both as discrete elements and packaged in PDF format through the interface to the EMR. Although this requires additional build and validation, it allows trending of certain fields by clinicians for patient care, preserves report format architecture, and provides

data for downstream data mining. Furthermore, the discrete data elements may be pulled to create an integrated report, which may contain AP, CP and MP studies. If orders and results are designed in computer systems to facilitate discrete information capture, then data will be more accessible and easily manipulated within the EMR, facilitating ease of use by the end user. Discrete data also enables easier data mining for long-term storage in a data warehouse or repository, and allows for optimization of data aggregation of populations for other business and clinical/research needs. All of these use cases require the ability to store and retrieve data efficiently and effectively [14, 15].

In formatting and creating reports, another tool mentioned previously used to facilitate discrete data capture is a synoptic structure or synoptic report. This report type uses a template of checklist with discrete data elements for a standardized report with clear, accurate and structured data. Synoptic reporting helps minimize error rates in reporting, promotes faster sign-out for the pathologist, and maximizes data availability with data mining and retrieval for research and other endeavors [16, 17].

To manage all of this data, applications such as knowledge bases, decision support tools, research applications, and productivity software can assist, and interfaces allow all of this information to be transferred between multiple applications. With the advent of interfaces, communication protocols and standards have been developed by some international organizations based on expert recommendations. These standards encourage compatibility between systems. Data dictionaries with standard language allow expedited data entry. Using standardized language can improve data accuracy in reporting since standardized data can be retrieved and analyzed more efficiently [18]. One such standard language which is especially useful in PI is Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT); it was developed by pathologists with the College of American Pathologists (CAP) and is now owned and licensed by the International Health Terminology Standards Development Organization [19]. SNOMED CT is a clinical ontology, a set of concepts with their properties and how they are related, with a scalable and internationally controlled vocabulary. The concepts have a unique identifier with a fully specified name and at least one synonym, and the concepts may have a hierarchical relationship or defined connections with a branch-based structure where each concept is more specific than the parent node. While SNOMED-CT may not contain all of the pathology and molecular entities, concepts may be adapted by creating extensions to enable encoding most of the morphologies [20]. Different modules including topography, morphology, function, occupation, agents, diseases, and procedures exist to assist the pathologist in describing organs, disease processes, and procedures used to prevent and treat diseases. This classification system allows

pathology reports to be reported with associated SNOMED codes so that results and report data are able to be mined for later use. For example, one group used existing guidelines in association with SNOMED codes to create clinical decision trees (CDTs) that could be used in CDS systems to help drive personalized medicine [21].

Understanding the tools available, such as standards, knowledge bases, and decision support tools, will allow the team to define a scope for the problem and design a robust solution to meet the end user's needs.

## ***2.6 Implementation and Validation***

Implementation of the solution involves development of the actual code in the various systems of the HIS including the LIS and EMR, as well as building any interfaces necessary for data integration between the systems. Once the build is complete, testing to ensure the build is correct is necessary. Once the designated teams have built the code needed for the desired workflow, the system changes must be installed and tested against real world and real time situations to ensure accurate and efficient performance. Labs are required to provide quality and highly reliable results. The LIS is built and designed to collect, process, distribute, and archive lab data. Every analyzer in the lab is validated prior to use to prove the system is performing accurately and efficiently, and the LIS is no exception. Regulatory and compliance agencies mandate that all components of an LIS system perform as expected by information system validation, which must be documented. Regulatory agencies include CMS under CLIA, as well as the FDA, The Joint Commission, and CAP. Validation is an important opportunity to identify not only failures but process improvement opportunities [22]. Laboratory end users in addition to clinical staff requesting the change and others utilizing the new workflow should be engaged in testing to verify that scope requirements are satisfied.

## ***2.7 System Maintenance***

Once in place, to assure that the solutions continue to function as intended and meet end user needs, it is advisable to perform ongoing evaluation and maintenance of the system. Suggestions for improvement may be considered at this point. Review of data for other purposes may be considered at this stage, as long as appropriate data access and consent processes have been considered. Once the required laboratory data is available for use in the HIS, it may be mined for use for future research or other business needs. For example, if molecular data is able to be provided in a discrete and minable fashion, it may have clinical relevance in the future. One group studied actionable mutations found in NGS data as it relates to cancer treatment and described the conundrum that much of the data in NGS results is nonactionable; however, having the nonactionable data available for future use and future clinical discoveries is necessary [23]. Another group in cytopathology

echoed the need for future data availability for genomics integration, and it also highlighted existing obstacles, including poor interoperability, lack of standardization between existing and legacy systems, limited commercial solutions and regulatory obstacles, that must be overcome to achieve better medical care [24]. If data is available for future use, nondiscrete data or legacy data may be used with tools such as natural language processing (NLP) to extract discrete data [25]. Unstructured text may also be studied using machine learning models, or artificial intelligence (AI) tools, to recognize medical concepts [26]. Encouraging data collection and storage, preferably in searchable formats, will allow for better documentation of medical care efforts. A systematic approach involving informatics tools allows provision of high-quality pathology results for diagnosis, prognosis and disease prediction, with future application for precision medicine [27].

---

### 3 Discussion of an Example: “A Day in the Life” Utilizing These Suggested Practices

To illustrate some of the practices discussed above, consider how they may be applied in this example, which is a common ask received in the author’s scope of work and practice at an academic, dedicated cancer center. It is not infrequent for clinicians to seek an easier way to view lab results in the EMR. For this example, a clinician asks to have faster access and discrete resulting of a single molecular test that he orders at least daily as a miscellaneous send-out test to guide cancer treatment protocols for his hematology cancer patients. To provide more detail for this example, consider this test to be for FMS-like tyrosine kinase 3 (FLT3), a receptor tyrosine kinase involved in signaling pathways controlling growth, proliferation, and survival, particularly of early blood cells called hematopoietic progenitor cells. As is required in his institution, the clinician submits a ticket request that is reviewed by governing bodies that approve the ticket and triage it to an informatics analyst to be worked. While engaging team members from the clinical area, IT, and the laboratory, the analyst begins the review of the clinician’s ask, and in doing so, maps out the current workflow for how the clinician orders the test and receives the results, as well as how it is handled within the laboratory. The current workflow requires a miscellaneous lab order in the CPOE, and results file under a miscellaneous result category. There is no way to easily search for the specific test result. Because the result data contains both quantitative and free-text information, the report is scanned into the EMR as a PDF file attachment only. In order to see the results, the clinician must search through all of the ordered miscellaneous tests and click on each one to identify the test of interest, and then must open a separate file into a result view window to see the scanned report. The team discovers that there are no means available to

compare or trend results for this assay using the current system build, making it difficult for the clinician to use the data in an actionable fashion. The laboratory does not have an interface built to the reference lab this esoteric test is sent too, and thus, it is receiving a paper result, which is attached by manually scanning the attachment into the EMR.

Upon reviewing the current workflow process and meeting with the clinician and other SMEs, the team understands that the clinical teams need to view the results in an efficient and quick manner, preferably in a discrete result field to be able to follow changes in the marker as they relate to the patient's disease status. The team lists requirements for the proposed solution, with the assistance and guidance of hematopathology experts and their clinical partners. These requirements include having a discrete orderable test for "FLT3 mutation" in the CPOE in place of the current miscellaneous order so that the result may file under "FLT3 mutation" under hematology or MP results. The result should contain discrete fields for the quantitative component and the free-text information from the report. Ideally, the free-text would also be searchable by some method, like NLP. Results should be viewable in the already existing lab flowsheet without requiring extra clicks or opening additional viewers. Research needs also require the lab result data for the FLT3 test be available for future use and easily searchable.

Since in this example a reference lab is used for performing the assay, the pathology informaticist may suggest that the lab should have methods in place to send the specimen with appropriate patient information. This may be a manual method from the ordering lab to the reference lab, but an orders interface may also be built to facilitate the data exchange. If an interface for orders or results is employed, the reference lab may work with team and be engaged in testing of the result report formats, which may include synoptic structures; the reference lab may also need to help test interfaces that may be required to enable sending discrete data electronically for the results to file back into the lab hematology tab in the EMR for the clinician. Revenue teams also might be involved to assist with billing testing. Each area affected by the proposed workflow should review the architecture to understand how their respective area will be involved, and the informatics analyst can help in navigating this.

The team decides the best solution architecture moving forward is to have a discrete order for FLT3 in the CPOE for the clinician to order with orders and results interfaces to the outside reference lab so that the results for the FLT3 assay display in the lab flowsheet in the EMR as discrete results, one result field for positive/negative result and one result field for the comments. For the implementation, the IT team builds and tests the order in the CPOE and the results back to the lab flowsheet through the LIS,

with the assistance of the reference lab and lab technologists in testing the interfaces needed to accommodate this data flow. The end users and lab technologists are engaged with final testing and validation, which is documented for regulatory inspections down the road. Before being put into production, the analysts make sure the clinician can find and order the appropriate test in the CPOE, the lab receives the correct information and specimen type, reference lab receives everything correctly to perform the assay, accurate results flow through interface to LIS and file in appropriate fields discretely, and LIS results are interfaced correctly to the EMR so that the clinician may finally view the FLT3 results in the lab flowsheet. Revenue team members are engaged to also ensure billing occurs as it should. Once all the documentation is reviewed by the lab medical director that everything is working and appropriate results are posting, the green light occurs to move the test into the live environment for the clinical team's use.

As part of ongoing maintenance, the test is routinely checked for accuracy through future updates of the software and report reviews by pathology staff for regulatory purposes. The clinician continues to report better use of time in regard to the FLT3 assay, but he continues to submit tickets for other assays needing the rework and conversion from miscellaneous status. Research is now able to more easily mine results for FLT3 for translational research endeavors since the result is now in a discrete field with a semiquantitative data result instead of a scanned image file. This concludes an example of a successful implementation for converting a miscellaneous test to a discrete orderable.

While the ask up front by the clinician may have seemed simple, the work that occurs behind the scenes to make this a reality is not menial. As described, there are many resources needed to get the data into actionable and discrete formats. That being said, the downstream ability to mine and use this data for translational purposes and precision medicine applications is vast. The ability to provide laboratory and pathology results as discrete searchable data, especially with the advent of AI tools, could potentially pave the way for faster routes to diagnosis and appropriate treatments for patients. Continuing to use informatics practices, especially in conjunction with pathology informatics, will only help in optimizing practices and lab data transformation to achieve better patient care in the future.

## References

1. Kroth PJ, Morioka-Douglas N, Veres S et al (2018) The electronic elephant in the room: physicians and the electronic health record. JAMIA Open 1(1):49–56
2. American Medical Informatics Association (2019) Why Informatics? <https://www.amia.org/why-informatics>. Accessed 18 June 2019
3. Friedman B (1990) Informatics as a separate section within a department of pathology. Am J Clin Pathol 94:S2–S6
4. Walker A, Garcia C, Baron JM et al (2016) Perceptions of pathology informatics by non-informaticist pathologists and trainees. J Pathol Inform 7:14

5. Wilson ML (2016) Regulations, standards, guidelines and benchmarks: a need for evidence-based management. *Am J Clin Pathol* 145(6):742–743
6. Beich MJ (2000) The role of the pathologist as tissue refiner and data miner: the impact of functional genomics on the modern pathology laboratory and the critical roles of pathology informatics and bioinformatics. *Mol Diagn* 5:287–299
7. Jackson BR (2019) Decision support from a reference laboratory perspective. *Clin Lab Med* 39(2):295–302
8. CLIA (2004) Clinical Laboratory Improvement Act 42CFR\$493.1291(a). [https://www.ecfr.gov/cgi-bin/text-idx?SID=69a006147ef8a38cc0f16233f35a58ab&mc=true&tpl=/ecfrbrowse/Title42/42cfr493\\_main\\_02.tpl](https://www.ecfr.gov/cgi-bin/text-idx?SID=69a006147ef8a38cc0f16233f35a58ab&mc=true&tpl=/ecfrbrowse/Title42/42cfr493_main_02.tpl). Accessed 18 June 2019
9. Kang HP, Devine LJ, Piccoli AL et al (2009) Usefulness of a synoptic data tool for reporting of head and neck neoplasms based on the College of American Pathologists cancer checklists. *Am J Clin Pathol* 132:521–530
10. Martin-Sanchez F, Maojo V, Lopez-Campos G (2002) Integrating genomics into health information systems. *Methods Inf Med* 41:25–30
11. Elevitch FR, Aller RD (1989) The ABCs of LIS: computerizing your laboratory information system. ASCP Press, Chicago
12. Royce W (1970) Managing the development of large software systems. *Proc IEEE WESCON* 26:1–9
13. Leymann F, Roller D (2000) Production workflow: concepts and techniques. Prentice-Hall, Englewood Cliffs, NJ
14. Safran C, Chute CG (1995) Exploration and exploitation of clinical databases. *Int J Biomed Comput* 39(1):151–156
15. Rassinoux AM, Miller RA, Baud RH, Scherrer JM (1998) Modeling concepts in medicine for medical language understanding. *Methods Inf Med* 37:361–372
16. Hassell LA, Parwani AV, Weiss L et al (2010) Challenges and opportunities in the adoption of College of American Pathologists checklists in electronic format: perspectives and experience of reporting pathology protocols project (RPP2) participant laboratories. *Arch Pathol Lab Med* 134:1152–1159
17. Cheung CC, Torlakovic EE, Chow H et al (2015) Modeling complexity in pathologist workload measurement: the automatable activity-based approach to complexity unit scoring (AABACUS). *Mod Pathol* 28(3):324–339
18. Houser SH, Colquitt S, Clements K, Hart-Hester S (2012) The impact of electronic health record usage on cancer registry systems in Alabama. *Perspect Health Inf Manag* 9:1f
19. SNOMED CT (2019) The global language of healthcare. <http://www.snomed.org/>. Accessed 18 June 2019
20. Sanz X, Pareja L, Rius A et al (2018) Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry. *J Biomed Inform* 78:167–176
21. Hendricks MP, Verbeek XAAM, van Veghel T et al (2019) Transformation of the National Breast Cancer Guideline into data-driven clinical decision trees. *JCO Clin Cancer Inform* 3:1–14
22. Winsten DI (1992) Taking the risk out of laboratory information systems. *Clin Lab Manage Rev* 6:39–48
23. Prawira A, Pugh TJ, Stockley TL, Siu LL (2017) Data resources for the identification and interpretation of actionable mutations by clinicians. *Ann Oncol* 28(5):946–957
24. Hanna MG, Pantanowitz L (2017) The role of informatics in patient-centered care and personalized medicine. *Cancer Cytopathol* 125 (S6):494–501
25. Baud RH, Rassinoux AM, Scherrer JR (1992) Natural language processing and semantical representation of medical texts. *Methods Inf Med* 31(2):117–125
26. Arbab A, Adams DR, Fidler S, Brudno M (2019) Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Med Inform* 7(2):e12596
27. Volynskaya Z, Chow H, Evans A et al (2018) Integrated pathology informatics enables high-quality personalized and precision medicine: digital pathology and beyond. *Arch Pathol Lab Med* 142(3):369–382



# Chapter 3

## Cohort Identification for Translational Bioinformatics Studies

Tiffany A. Lin, Zeynep Eroglu, Rodrigo Carvajal, and Joseph Markowitz

### Abstract

Translational studies for therapeutic development require cohort identification to identify appropriate biological materials from patients that can be utilized to test a specific hypothesis. Robust health information systems exist, but there are numerous challenges in accessing the information to select appropriate biological specimens needed for translational experiments. This chapter on methods describes the current standard process for cohort identification utilized by the Cutaneous Oncology Program and the Collaborative Data Services Core (CDSC) at Moffitt Cancer Center. The methods include utilization of graphical user interfaces coupled with database querying. As such, this chapter outlines the regulatory and procedural processes needed to utilize a health information management system to filter patients for cohort identification.

**Key words** Bioinformatics, Cohort identification, Translational science, Informatics, Resource paper

---

### 1 Introduction

Technological advances have vastly affected the way healthcare uses data, more specifically cancer research. Interfaces between the electronic medical record (EMR) systems and databases have grown over the years and has greatly improved health services and cohort identification for research [1, 2]. These systems are composed of health information systems and health information management systems.

Health information systems (HIS) comprise both the IT infrastructure (hardware and software) and the workflow processes required to facilitate the appropriate data capture, storage, and utilization of patient information [3].

A health information management system (HIMS), including the electronic medical record, collects, analyzes, reports, and protects medical information that is vital to the quality of patient care [4, 5]. Although the HIMS does not need to be electronic, increasing institutions are utilizing electronic medical records as their

primary source of patient medical information [6]. HIMS involves three fields: computer science, information technology, and business [7]. It starts from when patient health information is entered to the release of said information. The use of these systems should not compromise Health Insurance Portability and Accountability Act (HIPAA) privacy and security laws. In order to comply, there is a protocol that is approved by institutional and/or central Institutional Review Board (IRB) to conduct the research [8, 9]. These IRBs/protocols list the key personnel that have access to the data, the levels (roles) applied to the key personnel to determine the resources they have access to, and experimental procedures that may be conducted with patient derived biological materials [10]. The conduct of research done on clinical trials is outside the scope of this article.

There are institutional wide protocols that allow for research to be performed in a prespecified manner. For example, at Moffitt Cancer Center, there is the Total Cancer Care (TCC) protocol. TCC is an observational, prospective study with an objective of recruiting patients to participate [11, 12]. The TCC study evaluates the long-term effects of different cancers, treatments, and lifestyle. In addition, the TCC study is supported by the Health and Research Informatics (HRI) platform. HRI is the enterprise-wide patient data warehouse; it contains information from several Moffitt data source systems and from patients consented to the TCC protocol from the different consortium sites. Also, a molecular data warehouse is available to cancer investigators for future research. The molecular data warehouse contains clinically generated and research-generated molecular data. Clinical and molecular data are linked to tissue samples through data elements such as anatomic site of origin, tissue type, and histology. Patients who consent to this study will be followed throughout their lifetime; they can withdraw consent later if desired. The patients sign consent to allow release of their medical health information and the use of their tissue samples for cancer research. In addition, the protocol allows the patient to be contacted in the future if new findings could influence their care including clinical trials. TCC is currently composed of ten Florida hospitals including Moffitt Cancer Center, and eight national sites. This network of institutions recruits patients to the study, ships tissue to Moffitt Cancer Center's biobank to store and conduct molecular analysis, and provide medical data of the TCC consented patients to be integrated into a data warehouse [11].

Multiple institutions can collaborate and form networks within the scope of their individual IRB approved institutional protocols. The Oncology Research Information Exchange Network (ORIEN) founded by Moffitt Cancer Center and the Ohio State University Comprehensive Cancer Center—Arthur G. James Cancer Hospital and Richard J. Solove Research Institute, is one such type of

collaboration and now includes many more members [13, 14]. Datasets within each individual institution resides within their data warehouse with strict rules for access based upon their institutionally approved protocols. The network allows for utilization of this data as defined by the protocols. This data among other forms of data are stored in a system, a data warehouse.

A data warehouse (DW) is a system that collects, integrates, and aggregates various sources of data via Extract Transform and Load (ETL) processes within the organization for query, analyses, and reports (separate from the Electronic Health Record (EHR) but derives data from the EHR). External datasets can also be integrated into the DW. Results from a healthcare DW must be accessible to differing stakeholders, which can include healthcare regulators, physicians, hospital administrators, healthcare decision makers, and researchers. Data abstracted from data warehouses are done so using queries [15]. DWs are either set up as multidimensional hierarchical or relational databases [16]. This is different than a “flat database” that exists in an Excel-like table that only has columns, rows, and tables that makes accessing relevant information difficult. The main difference between relational and hierarchical databases is that relational databases involves joining tables that uses primary keys or foreign keys, which are common identifiers in each table to be able to join the corresponding information [17]. A hierarchical database organizes information based on its relationship to each other relative to the parent table. An example of this is a company organization chart that shows how the CEO has the VPs under their umbrella, and these VPs have managers/directors under theirs, and so on. Derivatives of hierarchical database languages such as M (Mumps) have been useful in health care [18, 19]. Many EHR commercial companies also utilize MUMPS. However, traditional hierarchical databases are difficult to utilize for cohort identification because the investigator would need to look up the parameter of interest for every single patient in the database. For cohort identification research, it is best to use a relational database as the data we deal with are not based on a relationship within a table but rather with multiple tables within the database. One can easily imagine how it can become difficult to analyze Excel tables where discrete values may be obtained from the medical record.

Optimizing the medical record to allow for discrete values to be visible for different parameters is the preuve of the HIMS. Different institutions have their own HIMS that have different variating levels of electronic integration. Part of the data-life cycle in research after it is extracted from the HIMS is create, read/retrieve, update, and delete (CRUD). The life cycle of data within the electronic medical record is beyond the scope of this article. Different users have different methods of using CRUD basic operations. In healthcare research, CRUD can be used as creating a new record in a

table, retrieving data points/health information, updating the health information, and deleting variables or rows that are no longer in use [20, 21]. Typically, CRUD operations are performed in a Structure Query Language (SQL), a programming language that can be used for creating and querying databases. When identifying a cohort of patients, accessing this health information through the electronic format can come in two ways: front-end and back-end. The front-end system is a more user-friendly graphical user interface that most of the time is a web-based application, whereas the back-end usually involves coding in a specific database management language such as SQL [22].

Multiple front-end systems with easy to use graphical interfaces exist. One such system used at Moffitt Cancer Center is a front-end system used to ascertain the data called TransMed® Systems [23]. TransMed® is used to query the Health and Research Informatics (HRI). TransMed® reports cover information related to patient demographics, clinic services, diagnosis, treatment, comorbidities, patient reported outcomes, and biospecimen samples availability for research purposes. This tool allows users to filter the data in accordance to the IRB/protocol. For those with limited coding experience, this tool provides an efficient and quick way to obtain data [24]. However, it is limited to providing simple searches and if complex queries are needed this method requires additional tools to manipulate the data. Some of these complications include potentially missing cohort subjects (dropped data points) because the interface does not support a particular type of query. As a result, separate reports must be run on the existing cohort in order to collect the data points needed.

As such, a programming language is needed to access the system at the back-end for greater user control of searchers. SQL is used to access the information within the database as it is a high-level programming language. This provides great flexibility and broad utility. SQL also allows users to build their databases vertically, where the layout is column to column rather than row to row. In addition, the language itself can handle large numbers of transactions in a single query, it is optimal for large number of table rows, and it is fast for searching, querying, and retrieving data from multiple tables. With this, there are some cons in using SQL as a way to query data and build databases. There is an in-depth complexity when interfacing a SQL database which requires more than inserting additional lines of code. Here is an example of a basic SQL query code that joins two tables (Patient and Sample) on a common field (MRN in this case): **select Patient\_ID, MRN, Gender from Patient where MRN in (select MRN from Sample where (Sample\_Type = "FFPE" and Collection\_Site\_Category = 'Thoracic' AND Current\_Quantity > 0))**. This query allows the user to pull specific medical record numbers along with the unique patient identifier from a specified table named “Patient” if there is at least one FFPE block available in the biobank collected from the

thorax. Join operations provide flexibility to relational databases, but at the same time they are computationally costly, because every join creates temporary data structures in memory containing the joined tables. There are also variations of SQL in which each variation has a subtle coding difference.

The main issue in selecting subjects for cohorts in translational studies is that the data in the data warehouse (stored as a relational database in the current example) is not in a format that can be easily read by a translational researcher. The data must be migrated to a format that is useful for the end user and abides by protocol requirements. There are two types of migration scripts: automated and manual. An automated script uses a synchronization tool and checked/ altered by a developer. A manual script is written by a database programmer (e.g., SQL script, SQL stored procedure). There are some setbacks involving the use and creating of this script. It may be impossible to preserve data in the correct place and though the script will know the “before” and “after” versions of the database, the transitional path is unclear to the user. An example of this would be migrating table A to table B. It is the same table but with a different name, however while in transition, the tool may drop all the contents of Table A and solely create Table B [22]. With this being said, manual scripts are powerful tools but safety precautions should be in place such that no data is lost.

SQL itself comes in various Relational Database Management Systems (RDBMS) such as MySQL, Oracle, and Microsoft SQL. MySQL is an open-source RDBMS. An open-sourced relational database is software whose source code is open and available for editing, studying, and redistributing. A closed-source is software whose source code is kept private to prevent copying [25]. The use of either database is important as new data emerges and comes in various forms from various sources unstructured. This new data needs to be stored in a relational database and relatively at a low-cost which is where open-sourced relational databases such as MySQL becomes useful [26]. New database platforms such as NoSQL offer increased scalability, flexible database environment with the ability to have many different document types embedded in the database instead of symbolic links and increased efficiency [27–29]. However, work in this area has not yet reached mainstream use for clinical management systems,

To summarize the requirements needed to perform searches specific for translational research projects several components are needed. The health information management system (HIMS) of an organization uses medical information to help the quality of patient care. As the medical information involves patients’ privacy, HIPAA and IRB/protocol provisions are put in place. Using the IRB/protocols, researchers are able to utilize data for their research [30]. At Moffitt, the TCC study consents patients to have access to their

medical information, tissue samples, and ability to follow up in the future if new developments occur that will affect care, such as being in a clinical trial [11]. Institutions can collaborate with respect to their individual IRB protocols and contribute with data to be aggregated and stored in data warehouses (i.e., ORIEN) [13]. Data from multiples sources are stored in a data warehouse and access to this warehouse involves complex queries [31]. Because we are looking at multiple sources and tables, having a relational database makes data accessible through querying [32]. The research-only data is separated out of the EHR of the HIMS in data warehouses such that there are no inadvertent changes to the medical record. Simple deidentified queries are performed on the user-friendly graphical tool, TransMed®. Once a cohort is identified, the honest broker in the context of an IRB approved protocol can provide the subject information to the investigator. However, searches typically require a deeper dive than the graphical tool even though this tool usually provides good insight on to where to begin the search. A database programmer who is also separate from the main research team (to maintain adherence to protocols) is needed to obtain the information. In today's environment, this usually means utilizing some variant of the relational SQL database programming language although this may change in the future with improving technologies [22]. The key is that data within the data warehouse must be migrated to an environment that is suitable for analysis by the translational investigator. Typically, this comes in the form of an Excel worksheet. Whether using front-end or back-end methods of obtaining the data, informational retrieval systems are needed to conduct searches. What come next are the software requirements for the system we use in-house and a case report.

---

## 2 Software Dependencies

- TransMed®/HRI—a commercial system used in our studies. The name of the company changed recently, but we use TransMed® in the text above as this is how the literature references the system. The new website is <https://inteliquet.com/>.
- SQL—it is a query language and there are variations of this language used in different products which include MySQL, PostgreSQL, Microsoft SQL, and Oracle. This is used to house and manage several databases. The majority of databases use SQL as a way to manage and warehouse data, however there are other databases such as Microsoft Access that do not use SQL. Microsoft Access is a great database system to use for those whose data isn't large as a nontechnical person can create a database, the cost isn't expensive as it comes with the Microsoft Suite, and there is a small learning barrier. However, if your data is big (larger than 2 gigabytes), you have a low of users and

objects, a more robust RDBMS would be the ideal data warehouse/management system to use. Microsoft Access is intended to be used as a gap between Excel and SQL server.

- R Statistical Software and Programming/R-studio [33, 34]. This is an object oriented fourth generation programming language that is used by our group for data cleanup [35, 36]. Cleaning data is detecting and correcting/removing data points that may be inaccurate/extraneous and removing the data from the final data sets.
- Microsoft Excel—the format of the deliverables is often put into Excel as that is one of the preferred visualization and analysis tool by researchers and utilized at Moffitt Cancer Center.
- PowerChart® (Cerner)—Cerner is the company that produces the electronical medical record system used by Moffitt Cancer Center and PowerChart® is the electronic medical record system that at its core is a database. It is to be noted that the data for research is stored in a data warehouse separate from the clinical database.

---

### 3 Method

1. Patients are consented to TCC protocol. Patients are approached for TCC consent in which sample collection and biobanking storage of biospecimens collected may be used for research.
2. IRB/Protocol approval for the researchers to be allowed access to data.
3. A deidentified search is conducted using TransMed® (Intelliquest®) for general query such as age, tumor, stage, and primary site.
4. For a more specific query such as tissue, and blood availability, SQL is utilized to obtain this information from the data warehouse utilizing the honest broker and the data warehouse. Information is collated utilizing SQL that will then be extracted into an Excel sheet.
5. Receive query and assess feasibility with current databases. This involves reading the protocol and checking the databases to see what fields are available for query. In some instances, there is a query where there is not a direct linkage within the database and therefore other means will have to be assessed to define such criteria. An example of this would be assessing metastatic status in body locations (sites). This is not directly measured/recorded in the databases, and therefore other means of information extraction is required. This includes pulling metastatic

site at diagnosis and tissue that have been deemed metastatic through pathology.

6. Select patients using selected inclusion and exclusion criteria.
  - (a) Through TransMed® (Intelliquest®), a front-end system, we are able to use an interface that allows us to choose criteria. However, potential drawbacks are that the system represents an overview. If a researcher is assessing a specific demographic of patients, we may need to broaden the criteria to get a general subset of patients. This general subset will allow us to manually further down the cohort through chart review.
  - (b) In SQL based systems, we program the query based on the criteria given. However, similar to HRI, if a query push is too specific in which the data is not recorded in the database, a general query will be made and a chart review will take place to finalize the cohort.
7. Once the data is abstracted, the data is cleaned based on the researchers' request. This may involve using R statistical software and programming, filtering in Excel, and additional data quality control measures are taken place within the data pull. This involves a randomization system and a manual chart review per the requirements of the IRB approved protocol to assure that the final cohort of patients reflects the eligibility criteria and appropriate data points. The data is released to the researchers.

In summary, patients who are consented to the TCC protocol are entered into a data warehouse that allows cancer investigators to query for personal health information relevant to research. A deidentified search is done to identify a cohort of patients using the research's criteria for a general query. If more specifics are needed such as biospecimens with specific characteristics, then SQL will be used to pull these results. The protocol written by the researcher will be assessed against the databases and feasibility will be determined. If there is a variable requested that is unavailable in the database, other means of pulling this variable will be discussed. Once the details have been finalized, the actual query will take place using the front-end system and/or SQL querying and data will be abstracted. The data is then cleaned using R software and sent to the researcher in Excel format once completed.

## References

1. Lau EC, Mowat FS, Kelsh MA et al (2011) Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clin Epidemiol* 3:259–272
2. Embi PJ, Payne PR (2009) Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc* 16(3):316–327

3. Berg M (2006) Health information management: integrating information technology in health care work. Routledge, New York
4. Hammond WE, Jaffe C, Cimino JJ, Huff SM (2014) Standards in biomedical informatics. In: Shortliffe EH, Cimino JJ (eds) Biomedical informatics: computer applications in health care and biomedicine. Springer London, London, pp 211–253
5. Vogel LH (2014) Management of information in health care organizations. In: Shortliffe EH, Cimino JJ (eds) Biomedical informatics: computer applications in health care and biomedicine. Springer London, London, pp 443–474
6. Yasnoff WA (2014) Health information infrastructure. In: Shortliffe EH, Cimino JJ (eds) Biomedical informatics: computer applications in health care and biomedicine. Springer London, London, pp 423–441
7. Chen ET (2013) An observation of healthcare knowledge management. Commun IIMA 13 (3):7
8. Payne TH, Graham G (2006) Managing the life cycle of electronic clinical documents. J Am Med Inform Assoc 13(4):438–445
9. Bankert EA, Amdur RJ, Amdur RJ (2006) Institutional review board: management and function. Jones and Bartlett, Sudbury, MA
10. Parker GE (2016) A framework for navigating Institutional Review Board (IRB) oversight in the complicated zone of research. Cureus 8 (10):e844
11. Li B, Eschrich SA, Berglund A et al (2017) Use of the total cancer care system to enrich screening for CD30-positive solid tumors for patient enrollment into a brentuximab vedotin clinical trial: a pilot study to evaluate feasibility. JMIR Res Protoc 6(3):e45
12. Fenstermacher DA, Wenham RM, Rollison DE, Dalton WS (2011) Implementing personalized medicine in a cancer center. Cancer J 17 (6):528–536
13. Lindsey H (2017) ORIEN uses big data to improve care for high-risk patients. Oncol Times 39(11):31, 40
14. Dalton WS, Sullivan D, Ecsedy J, Caligiuri MA (2018) Patient enrichment for precision-based cancer clinical trials: using prospective cohort surveillance as an approach to improve clinical trials. Clin Pharmacol Ther 104(1):23–26
15. Silver M, Sakata T, Su HC, Herman C, Dolins SB, O’Shea MJ (2001) Case study: how to apply data mining techniques in a healthcare data warehouse. J Healthc Inf Manag 15 (2):155–164
16. Coronel C, Morris S (2019) Database systems: design, implementation, and management. Cengage Learning, Inc. 13th edition, Boston, MA, USA
17. Murphy SN, Morgan MM, Barnett GO, Chueh HC (1999) Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp:892–896
18. Levy C, Beauchamp C, Hammond JE (1995) A managed care workstation for support of ambulatory care in Veterans Health Administration medical centers. J Med Syst 19 (5):387–396
19. O’Kane KC, McColligan EE (1997) A case study of a MUMPS intranet patient record. Healthc Inf Manage 11(3):81–95
20. Garcia-Molina H, Ullman JD, Widom J (2014) Database systems the complete book. Pearson Prentice Hall, 2nd edition Upper Saddle River, NJ, USA
21. Lianas L, Frexia F, Delussu G, Anedda P, Zanetti G (2009) pyEHR: A scalable clinical data management toolkit for biomedical research projects. Paper presented at: 2014 IEEE 16th international conference on e-health networking, applications and services (Healthcom); 15–18 Oct 2014
22. Kline K (2017) SQL in a nutshell: a desktop quick reference guide. O’Reilly Media
23. Miller ML, Ruprecht J, Wang D et al (2011) Physician assessment of disease activity in JIA subtypes. Analysis of data extracted from electronic medical records. Pediatr Rheumatol Online J 9(1):9
24. Dougoud-Chauvin V, Lee JJ, Santos E et al (2018) Using Big Data in oncology to prospectively impact clinical patient care: a proof of concept study. J Geriatric Oncol 9(6):665–672
25. Stallings W (1987) Handbook of computer-communications standards: the open systems interconnection (OSI) model and OSI-related standards. Macmillan
26. Marsan J, Pare G (2013) Antecedents of open source software adoption in health care organizations: a qualitative survey of experts in Canada. Int J Med Inform 82(8):731–741
27. Gopinath MP, Tamilzhari GS, Aarth SL, Mohanasundram (2017) An analysis and performance evaluation of NoSQL databases for efficient data management in E-health clouds. Int J Pure Appl Math 117(21):177–197
28. Ercan MZ, Lane M (2014) Evaluation of NoSQL databases for EHR systems. 25th Australasian conference on information systems, 2014, Auckland, New Zealand
29. Wang X, Williams C, Liu ZH, Croghan J (2019) Big data management challenges in health research-a literature review. Brief Bioinform 20(1):156–167

30. Dyrbye LN, Thomas MR, Mechaber AJ et al (2007) Medical education research and IRB review: an analysis and comparison of the IRB review process at six institutions. *Acad Med* 82 (7):654–660
31. Lyman JA, Scully K, Harrison JH (2008) The development of health care data warehouses to support data mining. *Clin Lab Med* 28 (1):55–71
32. Farooqui NA, Mehra R (2018) Design of a data warehouse for medical information system using data mining techniques. Paper presented at: 2018 fifth international conference on parallel, distributed and grid computing (PDGC), 20–22 Dec 2018
33. Rstudio web page. <https://rstudio.com>. Accessed 9 Jan 2020
34. The R project for statistical computing. <https://www.r-project.org/>. Accessed 9 Jan 2020
35. Chan BKC (2018) Data analysis using R programming. *Adv Exp Med Biol* 1082:47–122
36. Gabbielli M, Martini S (2010) Programming languages: principles and paradigms. Springer, London/New York



# Chapter 4

## Transitioning Clinical Practice Guidelines into the Electronic Health Record through Clinical Pathways

Sharjeel M. Hooda and Karen K. Fields

### Abstract

Clinical practice guidelines in oncology provide an evidence-based roadmap for most cancer care delivery but often lack directions for specific patient factors and disease conditions. Clinical pathways serve as a real-time clinical decision support system to translate guidelines to clinical practice. Pathways allow for the creation of a standardized, multidimensional roadmap for the continuum of care that can support clinical decision-making, maintain optimal outcomes, and limit unnecessary variation in cancer care. Here we describe the process to develop and implement clinical pathways in the electronic health record. This process includes building the appropriate foundation for a clinical pathways team with supports in the institutional ecosystem, creating visual representations of care paths, formalizing the pathway approval process, and translating clinical pathways into an electronic health record-integrated clinical decision support tool.

**Key words** Clinical pathways, Clinical decision support system, Clinical practice guidelines, EHR

---

### 1 Introduction

Clinical practice guidelines began in modern healthcare in 1990 under the direction of the Institute of Medicine's interpretation as "systematically developed statements to assist practitioner's decision about appropriate health care for specific clinical circumstances" [1]. In oncology, several professional societies have developed nationally recognized guidelines through the evaluation of available evidence and consensus meetings to identify best practices. The most prominent and extensive guidelines available for clinicians in practice are the American Society of Clinical Oncology (ASCO) clinical practice guidelines, the American Society for Radiation Oncology (ASTRO) clinical practice guidelines, the European Society for Medical Oncology (ESMO), and the National Comprehensive Cancer Network clinical practice guidelines in oncology (NCCN Guidelines®) [2–4]. While the summation of these resources cover the vast majority of cancer care

delivery, the available guidelines in oncology are often broad and identify several acceptable treatment strategies based upon the available level of evidence without accounting for specific patient factors (such as comorbidity, frailty, or prior therapy) or disease conditions (such as tumor location, histology and genetic landscape, or therapy costs). Moreover, there remains a significant degree of ambiguity and vagueness in guidelines that has led to reduced clinical practice guideline adherence [5, 6]. Previous efforts to confront these gaps include addressing the interpretation of qualitative probability terms among medical professionals and the creation of a controlled vocabulary to reduce variable interpretation, such as a lexicon of terms available for mammography results [7, 8]. In addition, the Guideline Implementability Decision Excellence Model (GUIDE-M) has been created as a tool to assist guideline developers by providing a common structure and nomenclature in guideline development [9]. Furthermore, Codish and Shiffman have previously developed a model to both address ambiguity and vagueness in the authoring and implementation process for clinical practice guidelines and foster future computerized application of such guidelines [10]. As these informatics approaches are being incorporated into the guideline development community, clinical pathways have emerged as a valuable clinical decision support system that contain several of these guiding principles to translate these evidence-based guidelines to clinical practice.

In contrast to national guidelines, clinical pathways incorporate specific consensus-derived recommendations and accompanying evidence to create a standardized, multidimensional roadmap for longitudinal care that encompasses various episodes of care and identifies key milestones and decision moments. Well-designed pathways can optimize the clinical decision-making process and limit unnecessary variation in care delivery while still allowing for adequate clinician input for personalized patient care and achieving benchmarks for both outcome and process measures. Furthermore, several pathway tools, such as Via Pathways, McKesson Value Pathways, and Eviti® Connect, have emerged that can also improve workflow with integration into the electronic health record and provide an avenue for alternative payment models between oncology practices and payers [11, 12].

The volume of oncology clinical pathways erupted in the last decade as early evidence emerged that pathways could maintain outcomes while decreasing costs of care by up to 35% [13–16]. Additional studies have shown improvements in quality and clinical outcomes, such as improved survival, appropriate biomarker testing, decreased time from diagnosis to treatment, or decreased in-hospital length of stay after surgery [17–19]. In January 2016, the ASCO Task Force on Clinical Pathways released a policy statement to identify recommendations for the development and implementation of oncology clinical pathways

[20]. ASCO furthered this initial work by developing criteria for high-quality clinical pathways in oncology and evaluating various vendors through criteria involving development, implementation, and analytics [11, 21]. By 2016, ASCO surveys revealed that more than half of all oncology practices surveyed were using clinical pathways and 46% were required to comply with multiple pathways due to relationships with payors or other outside entities [22].

In 2009, the H. Lee Moffitt Cancer Center & Research Institute (Moffitt) began the development of oncology care pathways with the goal of ensuring more consistent delivery of optimal outcomes across the care continuum. In our position as a National Cancer Institute designated Comprehensive Cancer Center, we have been able to leverage the expertise of our multidisciplinary teams consisting of medical oncologists, surgical oncologists, radiation oncologists, pharmacologists, and other related specialists to develop a systematic, evidence-based, consensus-driven, and cost-effective tool that reduces variability, creates efficiencies, controls costs, and helps physicians improve patient outcomes. Since our inception, the clinical pathways program has created 60 multidisciplinary, disease-focused clinical pathways and, in collaboration with Cerner®, has developed 20 unique clinical pathways within an Electronic Health Record (EHR)-integrated clinical decision support tool.

---

## 2 Foundational Requirements for Integrated Clinical Pathways

As described above, ASCO has developed a governance for clinical pathways development and evaluation through various policy statements that address the healthcare industry as a whole and describe both the recommended components and intended role of clinical pathways [20]. Furthermore, ASCO's developed criteria for high-quality oncology pathways programs identified 15 key criteria upon which to measure pathways in a standardized approach [21]. This then led to a subsequent evaluation of several commercialized oncology clinical pathway products with these criteria by the ASCO Task Force on Clinical Pathways [11]. While the clinical pathways program at Moffitt Cancer Center preceded the formation of these benchmarks and though none of the vendors evaluated previously have developed a completely EHR-integrated pathways program, we value these criteria and recommend all pathways programs incorporate these guiding principles as they develop and optimize their oncology clinical pathways program.

At Moffitt Cancer Center, we have developed and implemented a multidisciplinary clinical pathway solution that provides clinicians with the opportunity to provide guideline and evidence-based care through an EHR-integrated clinical decision support system. The creation of such a solution requires the availability or development

of multiple foundational structures. Furthermore, the creation is predicated on the availability of nationally developed guidelines and clinical expertise to direct each unique pathway.

The initial step is the formation of a locally developed and optimized clinical pathways team. Our team is led by a physician director and an operations director that oversee a core team consisting of a program manager and team members serving in three unique roles: clinical pathways specialist (CPS) responsible for the development and documentation of clinical content, clinical pathways informatics liaison (CPI) responsible for the translation of the content specifications into the EHR format, and systems analyst (SA) responsible for planning and building the pathways in the EHR. Our core team is further strongly supported by the clinical informatics, clinical systems, clinical trials, and data quality and business intelligence teams for integration into the EHR enterprise at our institution.

The development of a core clinical pathways team, however, is insufficient without the availability and reinforcement of a multi-disciplinary group of clinical experts. Our program is fortunate to have the clinical expertise of world-renowned faculty members across an array of disciplines including medical oncology, surgical oncology, radiation oncology, supportive care, pathology, radiology, personalized medicine, internal medicine, and various internal medicine subspecialties. The incorporation of these specialties allows for the creation of a unique product that incorporates the longitudinal journey for patients in each cancer diagnosis for which we employ a formalized pathway. In addition to our physician expertise, we have the expertise of several additional clinical disciplines, such as pharmacy and nursing, that routinely engage with our team for each pathway's development, implementation, and maintenance. Finally, the clinical pathways department must be supported by executive and cross-departmental leadership. At Moffitt Cancer Center, we have had the support of senior leadership since the program's inception in 2009 leading to the rapid growth and productivity of our pathways program. This is a requirement for any institution developing clinical pathways to ensure adequate resource availability and promotion of utilization through methods such as allocated administrative time and clinical department financial incentives.

Lastly, a robust clinical pathways program requires the availability of appropriate software to accommodate a workflow integrated solution. Diagramming software allows for elucidating the pathway nodes across all users of the pathway and to streamline core sections across various care paths. At our institution we utilize Microsoft Visio® for the creation of these diagrams, which are made available to users to review in both web format and portable document format (PDF). In addition, at Moffitt Cancer Center we have integrated our pathways with clinical trials and the availability of a

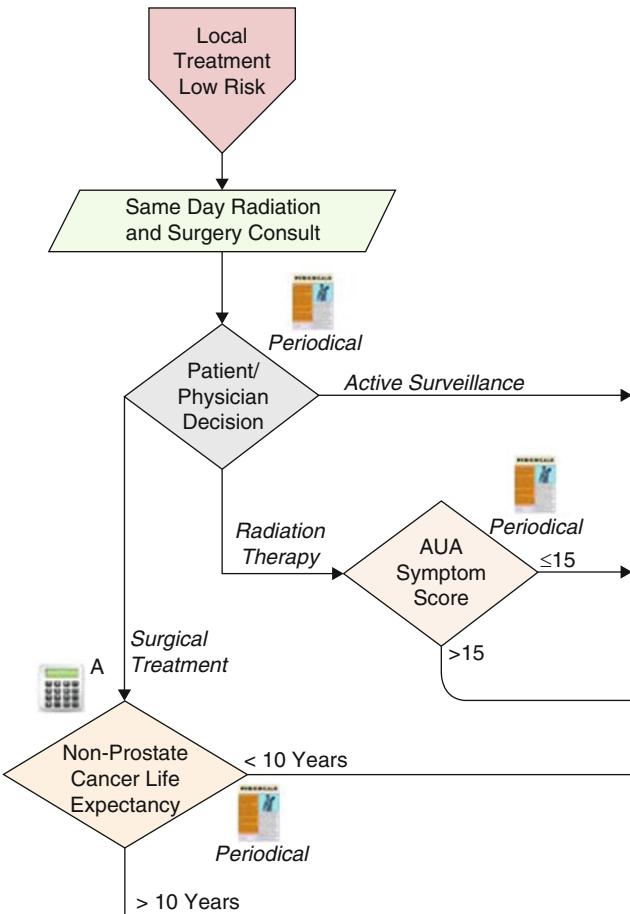
clinical trials management system (CTMS) is required for this process. We have utilized the OnCore® CTMS to create this interface and update our pathways with any changes to our clinical trials portfolio. Nonetheless, the most important and complicated technological solution requires integration within the EHR. There are multiple potential approaches, including a layered application or an embedded application. Through a unique partnership with Cerner® we have been able to develop an integrated EHR solution for PowerChart® that allows for embedded ordering within the electronic health record that incentivizes increased pathway utilization and decreased impact to a clinician's standard workflow.

---

### 3 Clinical Pathway Visual Representation

#### 3.1 Initial Pathway Diagram

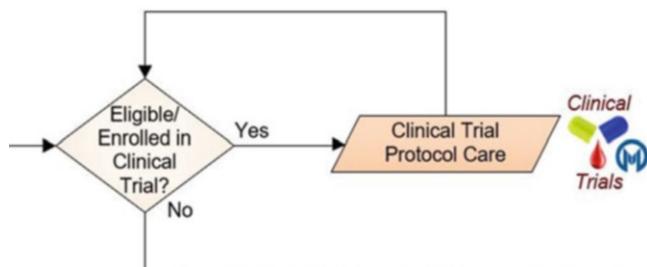
Once adequate system structures are in place, the clinical pathways team can develop individual clinical pathways. This begins with the identification of a pathway demand, which can be influenced by clinical volume, national and local prevalence, clinical expertise, and disease complexity. The initial process occurs with identification of a CPS and multiple clinical faculty members to author a newly identified pathway and create a corollary Visio® diagram. Each CPS member has a clinical background that provides them with improved capabilities to direct faculty and account for clinical workflows in pathway development. Both the faculty and CPS team members can review reporting data before the initial meeting that facilitates narrowing treatment decisions as appropriate throughout the pathway. This is followed by several consultation meetings over weeks to months to create the longitudinal care path. The presence of a clinical and informatics trained clinical pathways specialist at all these pathway development meetings allows for consistent application of pathway standards and principles across the institution. During this time, the pathways are formulated to account for the various milestones and decision points for the longitudinal course. There is a specific node in all our pathways that is identified as a "patient/physician" decision point (Fig. 1). This node occurs throughout each pathway and is a unique feature that considers both patient and disease-specific factors for clinical decision support. Patient-related factors include information such as medical comorbidities, prior therapies, and patient preference. Disease-related factors account for data related to the specific malignancy, histology, risk features, genetic landscape, disease stage, toxicities, cost, and the level of evidence for each proposed treatment. To assist clinicians at these "patient-physician" decisions, CPSs annotate the pathway with recommendations based upon the relevant patient and disease-related factors and include real-time links to references such as publications validating the evidence-based recommendation or online calculators and decision support tools.



**Fig. 1** The patient/physician decision node allows for inclusion of patient and disease-specific factors in a clinical pathway. Path annotations, periodical references, and calculators assist users in selecting a course. The Moffitt Cancer Center prostate adenocarcinoma pathway illustrates the importance of this node in the local treatment plan for low-risk prostate adenocarcinoma. (Reproduced by permission of Moffitt Cancer Center © 2011)

### 3.2 Clinical Trials Inclusion

In order to maximize the potential outcomes of cancer care delivery, clinical trials must be appropriately included as opportunities for selection in a clinical pathway. Our pathways program has developed a process to incorporate the clinical trials portfolio into each developed pathway. During the formation of the clinical pathway, faculty members will decide specific points for clinical trial evaluation. As expected, these time points can occur throughout the pathways and oftentimes these decision points will be included in multiple, if not all, sections of a care path. A close partnership between the clinical trials office utilizing the OnCore® CTMS and the clinical pathways program provides each care path with an avenue for daily updates with availability for specific clinical trials.

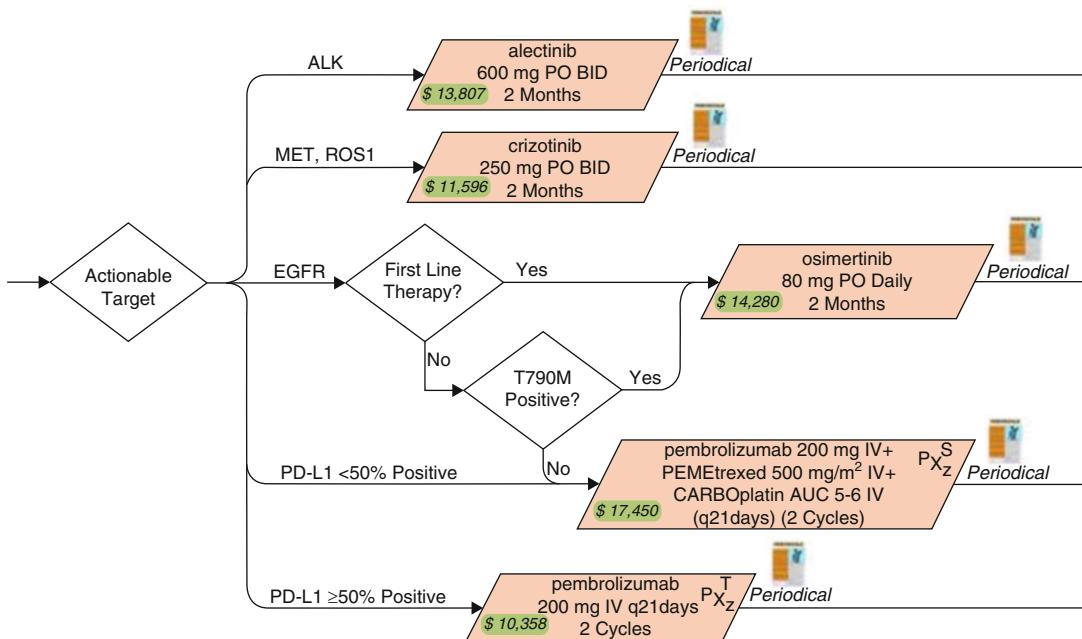


**Fig. 2** Clinical trial evaluation opportunities are embedded in the clinical pathway. All Moffitt Cancer Center clinical pathways include clinical trial consideration nodes throughout the longitudinal care path. Users can select the clinical trials icon for a curated list of potential trials, each with a hyperlink offering specific trial information. (Reproduced by permission of Moffitt Cancer Center © 2011)

Clinical trials are then linked within each pathway with users having point-of-care availability of information such as inclusion/exclusion criteria, open/closed trial status and key clinical trials personnel (Fig. 2). One of the CPSs, specifically trained in clinical trials with background in research, serves as the liaison for pathways at the clinical trials office meetings and coordinates updates for each pathway's associated clinical trials portfolio.

### 3.3 Usability Refinement

As each pathway undergoes an initial development and subsequent approval, the CPSs coordinate and hand-off the preliminary visual diagram to the SAs on the clinical pathways core team to optimize the visual representation of the care path. This includes various methods of standardization, including foci on formatting, design structure, and workflow logic (Fig. 3). Each pathway is formatted to a prescribed standard that confirms appropriate symbol usage, font type, font size, node color, appropriate icon inclusion, regimen pricing information, periodical and clinical trial links, and safety requirements such as tall-man lettering for drug names. In addition, a strict design structure is employed to ensure appropriate use of the various node types with associated text positioning standards for each object to optimize readability and clarity. Lastly, the SAs ensure that each pathway adheres to defined practices of workflow logic consistent with mechanical engineering standards, such that there is consistency across all published clinical pathways. This includes confirming that the overall flow of each page of the pathway is top-to-bottom and left-to-right, with absent or limited line cross-overs. The SAs will also confirm that each decision point adheres to a standard of “yes” answers exiting with a line on a horizontal branch and “no” answers exiting with line on a vertical branch. The SAs also review the workflow to identify unclear circuitous loops between nodes and if any objects exist without adequate entry and exit paths. Moreover, this regimented process allows our



**Fig. 3** Clinical pathway visual representations undergo a strict standardization process. All pathways include standardized formatting (such as path annotations and tall-man lettering), design structure (including appropriate use of node types, attachment of periodical references, and therapy costs represented as the annual published Average Wholesale Price for one cycle of infusional therapy or one month of oral antineoplastic therapy), and workflow logic (such as horizontal exit branches on a decision node correlating to “yes” answers). The Moffitt Cancer Center non–small cell lung cancer pathway exhibits this process for the initial targeted therapy of metastatic nonsquamous pathology with an actionable target. (Reproduced by permission of Moffitt Cancer Center © 2011)

pathways to comply to an internally defined standard set of rules and structure that creates high usability at the point of care and coherence across all published clinical pathways.

#### 4 Pathway Approval Process

Before a pathway is officially placed into the repository and available for users both internally and externally for clinical decision support, it undergoes review through a standardized three-step process: consensus, pathway approval committee, and chair approval. This process ensures that each developed clinical pathway has not only appropriate inclusion of clinical content but also incorporation of predefined pathway standards and workflow design. As a whole, this approval process serves as a governance structure for the pathways program to ensure the presence of multidisciplinary oversight that focuses on consistency and quality [23].

#### **4.1 Consensus**

Consensus allows for the review of the pathway by the clinical department members along with ancillary departments that support that diagnosis. This includes the expertise of pharmacists, advanced practice professionals (APPs) including nurse practitioners and physician assistants, inpatient and outpatient nursing team, and any additional clinical department that may be part of the care team for a diagnosis. During this time, the CPS will reach out to these various disciplines to ensure that each pathway has accounted for these various perspectives. This allows for creating pathway choices that consider details such as toxicities, periodical referenced evidence, molecular biomarkers, drug prices, clinical trials, and patient-physician shared decision points. The CPSs also routinely meet with the clinical leadership of the pathways team during this process to discuss progress and pathway creation planning.

Consensus between faculty members of the department regarding the pathway in its entirety occurs through multiple avenues. First, several meetings are organized in which multiple faculty members attend to formulate a pathway and revise it during its initial development. CPSs and clinical pathways program leadership may also attend disease specific tumor boards to discuss pathway details and garner additional guidance on formulation. This also serves to verify that pathways are created in concordance with actual practice patterns. Finally, email communication is employed to share pathway diagrams and details for any clinicians unable to attend meetings in person and final consensus is completed by reviewing and integrating the input through this asynchronous approach.

#### **4.2 Pathway Approval Committee**

The second step of the standardized approval process for each pathway is referral to the pathway approval committee (PAC). The PAC is comprised of multidisciplinary and senior members from multiple clinical sites at Moffitt Cancer Center that review each pathway for approval consideration. This review includes both a bird's-eye view and detailed workflow analysis. Furthermore, the consistency of the PAC members over time has allowed for development of their subject matter expertise in not only clinical content but also clinical decision support focus areas. For example, PAC members may identify redundant loops in pathways or decision points that lack adequate inclusion of specific patient-physician factors for decision support. This provides the opportunity to help standardize pathways with sections that have previously been employed in other care paths and limit unnecessary variation.

PAC meetings usually occur in-person and via remote conferencing to allow for multisite member input. Meetings are held weekly or biweekly, depending upon the number of pathways requiring initial approval or subsequent renewal. In-person meetings are attended by the CPS presenting the pathway and can also

include clinical faculty authors of a submitted pathway if there are significant details or specific questions to be addressed at PAC. The pathway approval committee ultimately will approve, approve with pending changes, or require resubmission for each submitted pathway. Feedback from the PAC is circulated to the clinical authors by the CPS and clinical pathways faculty and, if necessary, the pathway is resubmitted for PAC approval.

#### **4.3 Chair Approval**

The final approval for each developed clinical pathway occurs at the level of the department chair for the associated clinical program. However, this has become a formality at our institution as the chair of the department is always associated with the pathway development through either direct authorship or consensus review that occurs in prior approval steps. If there are any changes requested by PAC for approval without resubmission, the clinical pathway is updated by the primary authors and submitted for final approval directly to the chair. Once approved by the chair of the clinical department, the clinical pathway is added to the program repository and published on an interactive web page, thus making it available for use by clinicians at the bedside at our various institutional sites and any approved external partners.

---

## **5 EHR Clinical Pathway Integration**

Once a clinical pathway has been formulated to standardized specifications and final approval has been received, the pathway is available to be integrated into the EHR. The EHR integration schedule is a complicated process that is managed by the clinical pathways program manager. Each pathway requires a standardized approach for integration that includes several steps: hand-off meeting, EHR design sessions, EHR build, EHR testing, EHR training, and EHR Go-Live.

#### **5.1 Hand-Off and EHR Design**

The initial hand-off meeting occurs when the CPSs and SAs transition the pathway to the CPIs and discuss any specific or unique attributes of the pathway. Since the CPIs, CPSs, and SAs work collaboratively on several pathways and meet regularly for team status meetings, there is often already a basic understanding of the pathway in transition by all members. The next step is the EHR design sessions that occur between the CPIs, SAs, and the department clinical champions. These champions include physicians in medical oncology, surgical oncology, and radiation oncology along with APPs, pharmacists, and nurses. Oftentimes, the key faculty designing the EHR pathway include the primary authors of the online published care path. This consistency creates a strong relationship between the pathways department and the clinical departments and improves the efficiency of the design process.

These sessions are fluid and there are often situations when updates to the online published content can occur simultaneously during EHR design sessions. If such situations arise, the pathway is submitted for approval through the consensus, PAC, and chair process as designated previously. EHR design sessions are uniquely different than visual pathway design sessions in that the clinical workflow and ordering practices are focused on to develop appropriate EHR graphical user interfaces that optimize the user experience. Furthermore, the logical workflow of the EHR pages may not align exactly with the visual pathway diagram as optimizations and EHR abstracted chart information may decrease the amount of data entry or clicks required by the user. For example, previously entered staging information confirming stage IV disease could advance a user in the EHR to the metastatic disease pathways page rather than requiring initiating a pathway from the initial diagnostic page. With their informatics background, the clinical pathways informatics liaisons provide an important perspective during these EHR design and optimization sessions to incorporate pathways standards and principles as they work with the faculty to develop the EHR implementation.

### **5.2 EHR Build and Testing**

The EHR build process is the actual conversion of the visually designed pathway to an EHR pathway via creation of various EHR pages. These pages serve as the various appropriate sections of the pathway. CPIs and SAs develop order sets and data collection worksheets that incorporate all decision nodes of the pathway, such that specific choices will direct the user to either different points along the pathway or to appropriate orders and order sets available for user selection (Figs. 4 and 5). The clinical pathways team SAs then work closely with SAs within our clinical systems department to build the pages within a test environment. Once EHR clinical pathway pages are created, they undergo a round of testing internally by the SAs and CPIs followed by user acceptance testing (UAT) with the various clinician champions. Any UAT related updates, corrections, or specification requirements are subsequently completed on the EHR pages followed by subsequent EHR build in the live production environment. The clinical pathways and clinical systems teams then complete a similar testing process in the live environment to confirm pathway functionality.

### **5.3 EHR Go-Live and Training**

The final steps in the EHR clinical pathway creation process is the official Go-Live on the production system environment. At this point, end users can access the clinical pathway through the EHR and utilize the interface at the point of care for clinical encounters. Once the Go-Live occurs, CPIs will often spend several weeks training the various faculty members and APPs within the department at the point of care. These educational meetings are typically one-on-one and almost always occur during clinic to provide

Criteria		Reset
Eligible for Clinical Trial?	Yes	No
Actionable Target		
EGFR	<input checked="" type="checkbox"/>	
First Line Therapy?	Yes	No
T790M Positive?	Yes	No
Reassessment		Reassess
Response		
Complete Response (CR)		
Partial Response (PR)		
Stable Disease (SD)		
Progressive Disease (PD)		

**Fig. 4** A snapshot of the Cerner PowerChart®-integrated Moffitt Cancer Center clinical pathway page for the initial targeted therapy of metastatic nonsquamous non–small cell lung cancer with an actionable target. EHR-integrated pathway pages allow for direct user data entry into the EHR via the pathway tool to document decisions, milestones, treatments, and response. (Used with permission of Cerner)

Recommended Treatment Options	Remove All
CT Thorax w/Contrast Same Day Clinic Appointment, Non-small cell lung cancer, non-metastatic, assess treatment response, Future Order (Needs Scheduling) In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
CT Abdomen w/Contrast Same Day Clinic Appointment, Water (recommended), Liver mets suspected, Future Order (Needs Scheduling) In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
osimertinib 80 mg oral tablet = 1 tab(s), PO, DAILY, # 30 tab(s), 11 Refill(s)	
CT T/A w/Cont (Thorax/Abdomen) Same Day Clinic Appointment, Water (recommended), Follow-up disease response, Future Order (Needs Scheduling) In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
CMP Routine, Future Order In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
CBC with Diff Routine, Future Order In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
Thoracic Clinic Appointment Request Established Patient, Thoracic Onco, with MedOnc, Follow-up on initial targeted therapy for metastatic NSCLC, Standard, Future Order (Need Scheduling) In Approximately 2 month(s) Grace Period (+/-) 7 day(s)	
Supportive Care Clinic Appointment Request New Establish Pt/Consult, Supportive Care, Metastatic NSCLC, Symptom/pain management and goals of care discussion, Standard, Future Order (Need Scheduling) In Approximately 3 day(s) Grace Period (+/-) 7 day(s)	
Thoracic NSCLC Molecular Testing	

**Fig. 5** A snapshot of the Cerner PowerChart®-integrated Moffitt Cancer Center clinical pathway page for the initial targeted therapy of metastatic nonsquamous non–small cell lung cancer with an actionable target. EHR-integrated pathway pages allow for users to select from recommended orders (whether individual or ordersets) based upon the pathway node location. Users can also order alternative treatments or document off pathway as per clinician discretion. (Used with permission of Cerner)

clinicians with real-time pathways experience. Training sessions are very well received by the clinical departments and result in significantly increased pathway utilization.

#### **5.4 Version Control**

Once the clinical pathway has been formally integrated into the EHR, there are multiple methods to review a pathway to identify updates and design appropriate modifications to the visualized pathway and EHR tool. First, there are scheduled review time-points for each pathway when the primary authors and CPSs meet to review the pathway in its entirety and identify changes in best practice or recognize novel therapeutics that have been approved since the pathway's most recent evaluation. These pathway review dates occur at least annually for each pathway, but often occur more frequently for most pathways due to frequent guideline updates and novel therapeutic approvals. Contrastingly, clinical trials are updated daily through the OnCore® CTMS and clinical pathways relationship and thus do not require a scheduled review process. Second, pathways are often updated as information disseminates from the department or the pathways team regarding changes in standard of care. Minor modifications can often be completed in near real-time with approval from the primary authors and pathway leadership, but significant changes in the care path due to updated information do require the formalized approval process, including consensus, PAC, and chair approval as appropriate. Last, the pathways are reviewed routinely for “off-pathway” selections. These are reviewed for appropriateness by the clinical pathways department leadership team and the identified pathway consensus team. Importantly, these “off-pathway” selections are occasionally early indicators from expert users that the standard of care is transitioning, and the published pathway requires appropriate modification. Through these various avenues, the clinical pathways team maintains the clinical pathways as an up-to-date clinical decision support tool that can be employed at the bedside at any time.

In addition to an approach directed for individual pathways, our clinical pathways program utilizes two methods to address workflow and quality improvement across all clinical pathways that are published and developed as an EHR integration. First, the program has both a weekly status meeting and a weekly huddle that addresses any technical issues that arise either within the clinical pathways tool or the larger IT systems enterprise. The weekly status meeting includes the entire clinical pathways team along with key members within the clinical systems team and the EHR liaisons and covers topics of both interdisciplinary technical implementation and internal clinical pathways program content development and refinement. Workflow processes are evaluated at both the weekly status meeting and the weekly huddle meeting to identify any gaps in the team workflow or opportunities for workflow reengineering. Workflow reengineering typically occurs through

the clinical pathways program developed quality improvement sessions that are scheduled throughout the calendar year and focused on both team member identified areas of improvement and program leadership determined opportunities for process improvement. These methods allow for a robust program with a focus on continuous quality improvement as we continue to develop and optimize our EHR-integrated oncology clinical pathways.

## 6 Future Directions

Moffitt's clinical pathways currently cover over 90% of all cancer cases in the United States through 60 interactive web-based clinical pathways, with 75% of all cancer cases addressed with 20 EHR-integrated pathways. The clinical pathways program's continued aim is to expand our portfolio to encompass additional cancer diagnoses that clinicians will encounter at the bedside and provide a decision support tool to aid in care delivery and provide a longitudinal care roadmap. In addition, we continue to work with Cerner® to optimize the clinical pathways tool and improve the user experience and efficiencies of the EHR interaction. Furthermore, our program continues to innovate and enhance our data acquisition and analytics capabilities from pathway interactions in the EHR through our strong relationships with our internal clinical informatics, clinical systems, clinical trials, and data quality and business intelligence teams. Finally, we continue to work with payors to determine the role of shared savings through the utilization and implementation of clinical pathways. Overall, as the required application of clinical pathways continually expands in oncology, the goal across all healthcare institutions remains to enhance the capabilities and efficiencies of EHR-integrated clinical pathways for optimized cancer care delivery.

## References

1. Clinical practice guidelines: directions for a new program. National Academies Press, Washington DC, 1990
2. ASTRO clinical practice guidelines. 2019. <https://www.astro.org/Patient-Care-and-Research/Clinical-Practice-Statements/Clinical-Practice-Guidelines>. Accessed 24 Sept 2019
3. ASCO guidelines, tools, & resources. 2019. <https://www.asco.org/practice-guidelines/quality-guidelines/guidelines>. Accessed 24 Sept 2019
4. NCCN guidelines & clinical resources. 2019. <https://www.nccn.org/professionals/default.aspx>. Accessed 24 Sept 2019
5. Grol R, Dalhuijsen J, Thomas S, Veld C, Rutten G, Mokkink H (1998) Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. *BMJ* 317(7162):858–861
6. Shekelle PG, Kravitz RL, Beart J, Marger M, Wang M, Lee M (2000) Are nonspecific practice guidelines potentially harmful? A randomized comparison of the effect of nonspecific versus specific guidelines on physician decision making. *Health Serv Res* 34(7):1429–1448
7. Kong A, Barnett GO, Mosteller F, Youtz C (1986) How medical professionals evaluate expressions of probability. *N Engl J Med* 315 (12):740–744

8. Liberman L, Menell JH (2002) Breast imaging reporting and data system (BI-RADS). *Radiol Clin N Am* 40(3):409–430. v
9. Brouwers MC, Makarski J, Kastner M, Hayden L, Bhattacharyya O (2015) The guideline implementability decision excellence model (GUIDE-M): a mixed methods approach to create an international resource to advance the practice guideline field. *Implement Sci* 10:36
10. Codish S, Shiffman RN (2005) A model of ambiguity and vagueness in clinical practice guideline recommendations. *AMIA Annu Symp Proc* 2005:146–150
11. Daly B, Zon RT, Page RD et al (2018) Oncology clinical pathways: charting the landscape of pathway providers. *J Oncol Pract* 14(3):e194–e200
12. Schroeder A (2017) Clinical pathways: a current snapshot, and the journey ahead. *J Clin Pathways* 3(2):33–40
13. Hoverman JR, Cartwright TH, Patt DA et al (2011) Pathways, outcomes, and costs in colon cancer: retrospective evaluations in two distinct databases. *J Oncol Pract* 7(3S):52s–59s
14. Jackman DM, Zhang Y, Dalby C et al (2017) Cost and survival analysis before and after implementation of Dana-Farber clinical pathways for patients with stage IV non-small-cell lung cancer. *J Oncol Pract* 13(4):e346–e352
15. Neubauer MA, Hoverman JR, Kolodziej M et al (2010) Cost effectiveness of evidence-based treatment guidelines for the treatment of non-small-cell lung cancer in the community setting. *J Oncol Pract* 6(1):12–18
16. Rotter T, Kinsman L, James EL et al (2010) Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. *Cochrane Database Syst Rev* (3): CD006632
17. Bao H, Yang F, Su S et al (2016) Evaluating the effect of clinical care pathways on quality of cancer care: analysis of breast, colon and rectal cancer pathways. *J Cancer Res Clin Oncol* 142 (5):1079–1089
18. van Hoeve J, de Munck L, Otter R, de Vries J, Siesling S (2014) Quality improvement by implementing an integrated oncological care pathway for breast cancer patients. *Breast* 23 (4):364–370
19. te Marvelde L, McNair P, Whitfield K et al (2019) Alignment with indices of a care pathway is associated with improved survival: an observational population-based study in colon cancer patients. *EClinicalMedicine* 15:42–50
20. Zon RT, Frame JN, Neuss MN et al (2016) American Society of Clinical Oncology policy statement on clinical pathways in oncology. *J Oncol Pract* 12(3):261–266
21. Zon RT, Edge SB, Page RD et al (2017) American Society of Clinical Oncology criteria for high-quality clinical pathways in oncology. *J Oncol Pract* 13(3):207–210
22. American Society of Clinical Oncology (2017) The state of cancer care in America, 2017: a report by the American Society of Clinical Oncology. *J Oncol Pract* 13(4):e353–e394
23. Yu PP (2018) Oncology clinical pathways: a form of governance? *J Oncol Pract* 14 (3):144–146



# Chapter 5

## Variable Selection for Time-to-Event Data

Ai Ni and Chi Song

### Abstract

With the increasing availability of large scale biomedical and -omics data, researchers are offered with unprecedented opportunities to discover novel biomarkers for clinical outcomes. At the same time, they are also faced with great challenges to accurately identify important biomarkers from numerous candidates. Many novel statistical methodologies have been developed to tackle these challenges in the last couple of decades. When the clinical outcome is time-to-event data, special statistical methods are needed to analyze this type of data due to the presence of censoring. In this article, we review some of the most commonly used modern statistical methodologies for variable selection for time-to-event data. The reviewed methods are classified into three large categories: filter-test based method, penalized regression method, and machine learning method.

**Key words** Variable selection, Time-to-event data, Filter test, Penalized regression, Machine learning

---

### 1 Introduction

With the advancement of modern data capturing techniques such as electronic medical record and next-generation sequencing, large amounts of biomedical and -omics data are becoming increasingly available. These data offer biomedical researchers unprecedented opportunities to study the association between patients' biomarker profile and their clinical outcomes. Since the total number of available biomarkers is very large and it is likely that only a small subset of them are truly associated with the biological mechanism of the clinical outcome of interest, it is desirable to have statistical methods that accurately select a subset of biomarkers that have high probability of being associated with the clinical outcome. This biomarker selection process, also known as variable selection or feature selection, will enhance the interpretability and predictability of the statistical models [14].

---

Both the authors “Ai Ni and Chi Song” contributed equally to this work.

Time-to-event data is an important type of outcome in biomedical research. It is also known as survival data, failure time data, or lifetime data. It measures the time from a defined origin to the occurrence of an event of interest. A unique feature of time-to-event data is that the time may not be directly observed, which is called censoring in the statistical literature. For example, in a study on the number of years from the diagnosis of stage I lung cancer to death, some patients may still be alive at the end of the study and therefore their exact survival times are not observed. Special statistical methodologies are required to analyze time-to-event data taking into account the censored observations.

Variable selection is very challenging in the presence of a large number of covariates (i.e. biomarkers). The difficulties come mainly from three aspects. First, the false positive rate (i.e. the probability of selecting covariates that are not truly associated with the outcome) can be substantially inflated with a large number of statistical tests. Imaging an investigator tries to identify which of the ten thousand genes are significantly associated with a clinical outcome using a statistical test on one gene at a time and the criterion of  $p$  value  $\leq 0.05$ . Then, on average the investigator would falsely discovery 500 significant genes by chance even if none of the ten thousand genes are truly associated with the outcome. Second, the covariates under consideration are likely to be correlated with each other and the correlation structure can be very complex. This usually leads to complicated confounding relationship among the covariates, which can only be revealed by jointly modeling all candidate covariates, making the dimension of the model prohibitively high. Lastly, the strength of association of covariates with the outcome may be modified by other covariates, and the associations may be nonlinear with unknown patterns, which calls for nonlinear and highly flexible statistical models. Apparently traditional statistical methods are insufficient to handle the variable selection task with modern high-dimensional data. Over the past couple of decades, novel and powerful statistical methods have been developed to tackle this challenging. In this article, we will review some most important methods for variable selection for time-to-event data in three categories: filter-test based method, penalized regression method, and machine learning method.

---

## 2 Materials

This section summarizes the R packages and functions that can be used to carry out the variable selection methods, which will be discussed in Subheading 3.

The main component of the filter-test based variable selection methods is univariate analysis on the association of each biomarker with the time-to-event outcome. When the biomarker is a

categorical variable, a log-rank test can be performed using `survdiff` function in **survival** package. When the biomarker is a categorical or continuous variable, a Cox's proportional hazards regression can be performed using `coxph` function in **survival** package, or an accelerated failure time (AFT) regression can be performed using `survreg` function in **survival** package. Once the univariate analysis has been conducted and the  $p$  values have been obtained, one can use the `p.adjust` function available in the base R environment with the option `method=BH` to obtain the adjusted  $p$  values (also known as the  $q$ -values) using the Benjamini–Hochberg procedure [5] (see Subheading 3.1 for details).

Many of the algorithms for fitting the penalized Cox's proportional hazards regression discussed in Subheading 3.2 are implemented in R package **glmnet** [15]. The main function in that package is `glmnet`. With the option `family=cox`, it fits elastic net penalized Cox's proportional hazards regression and returns the entire solution path of the tuning parameter  $\lambda$ . To obtain the model corresponding to the optimal value of  $\lambda$ , one can use the `cv.glmnet` function, again with the option `family=cox`. This function uses  $k$ -fold cross-validation to choose the optimal value of  $\lambda$ . Both `glmnet` and `cv.glmnet` functions allow users to specify the value of the second tuning parameter  $\alpha$  in elastic net penalty. Therefore, one can fit lasso penalized Cox's proportional hazards regression by specifying  $\alpha=1$ . Moreover, both functions have `weights` option that allow users to specify weights for the penalty. Thus, by setting  $\alpha=1$  and `weight` equal the inverse of a consistent estimate of the regression coefficient (e.g. the non-penalized estimate), one can fit adaptive lasso penalized Cox's proportional hazards regression.

Group lasso and sparse group lasso penalized Cox's proportional hazards regression can be carried out using the R package **SGL** [27]. The `SGL` function in the package fits sparse group lasso penalized Cox's proportional hazards regression with the option `type=cox`. Similar to the `glmnet` function, it returns the solution path of the tuning parameter  $\lambda$ . To obtain the model corresponding to the optimal value of  $\lambda$ , one can use the `cvSGL` function, which uses  $k$ -fold cross-validation to choose the optimal value of  $\lambda$ . Both `SGL` and `cvSGL` functions allow users to specify the value of the second tuning parameter  $\alpha$  in sparse group lasso penalty. Therefore, one can fit group lasso penalized Cox's proportional hazards regression by specifying  $\alpha=0$ .

Most of the machine learning method discussed in Subheading 3.3 are implemented in R. The survival tree method can be done by using R package **rpart** and **survival**. The main function for tree construction is the `rpart` function, which is flexible enough to take any regression formula as its parameter. Suppose that we have a genomic dataset that is organized in a data frame `d`, where column `time` and `status` are the event time and censoring status, and the

other columns are gene expression data. Command `fit <- rpart(Surv(time, status) ~ ., data=d)` will construct the survival tree and save it to object `fit`.

Random survival forest is implemented in R package **randomForestSRC**. The main function in this package is `rfsrc`, whose usage is very similar to `rpart`. Command `fit <- rfsrc(Surv(time, status) ~ ., data=d)` will construct a random survival forest with default 1000 trees and save it to object `fit`. To calculate the variable importance for all the features, we can use command `importance <- vimp(fit)`.

The supervised principal component analysis is implemented in R package **superpc**. For detailed usage of these functions, please refer to the document of the package.

### 3 Methods

#### 3.1 Filter-Test Based Method

In genomic studies, tens of thousands of genes are measured at the same time, which requires fast and simple enough feature selection method. Therefore, filtering genes according to univariate analysis and testing result is the most widely used feature selection approach. This approach is simple but has decent performance in most studies. First, a univariate analysis (model fitting or statistical testing) is performed for each single gene, and a summary statistic (such as test statistic or  $p$ -value) is calculated. Then, important genes are selected based on these summary statistics.

In the univariate analysis step, the summary statistic is often selected to reflect the association between the gene feature and the outcome variable. In the time-to-event outcome, because of existence of censored subject, survival analysis methods can be used to produce this summary statistic. A natural choice is to use the  $p$ -values of the survival test. Depending on the type of gene features, different survival tests can be used. For example, if the gene features are categorical variables such as SNP genotype, a log-rank test can be used to test whether the survival function is the same for all genotypes in this SNP. However, if the genomic feature is gene expression level, we often fit a Cox regression model or an accelerated failure time (AFT) model to regress the survival outcome on the continuous gene expression feature, and use the  $p$ -value for the regression coefficient as our test  $p$ -value. In practice, to avoid the additional assumptions (e.g. linear assumption and proportional hazard assumption) made by these regression models, we can also dichotomize the gene expression levels into low and high groups by cutting at the median, then use the log-rank test to generate the  $p$ -value.

The filtering step can be achieved by either selecting features that pass a threshold or selecting an arbitrary number of features. If  $p$ -values are calculated as the statistic, one can directly select a

threshold by experience.  $p$ -Value smaller than 0.01, 0.05, or 0.1 are commonly used thresholds. Because there are multiple gene features in the genomic studies, false discovery rate (FDR) control are often achieved by calculating  $q$ -values (a.k.a adjusted  $p$ -values) using the Benjamini–Hochberg procedure [5]. Thresholding  $q$ -values at 0.05, 0.1, or 0.2 are also a common practice in genomic feature selection. We can also select features with large absolute value of the regression coefficient. However, since the overall gene expression level and variation can be different between genes, it is recommended that all genes should be standardized before the regression such that their coefficients are comparable. To ensure that no important gene is missed by filtering, more sophisticated methods such as sure independent screening are also proposed for time-to-event data feature selection [37].

The advantage of the filter-based method is that it is simple to implement and typically requires smaller sample size than other variable selection methods. It is often used in genomic studies where the number of genes are larger than the sample size by several orders. However, since it does not consider the correlation among covariates, it is usually used as a first-step feature selection method. The selected genes will usually be subject to subsequent biological experiments to confirm their functions. Alternatively, many researchers use the selected genes to further build a prediction model for the outcome. We call for caution that to evaluate the performance of the prediction model, the uncertainty of the feature selection step should also be considered. For example, if a cross-validation is used to evaluate the performance of a survival model built based on genes with  $q$ -values smaller than 0.05, the feature selection should be performed in each iteration of the cross-validation procedure. If the feature selection is performed outside the cross-validation, because the validation set data are used in the feature selection step, the evaluation will suffer from the double dipping problem, which in turn will cause an over-optimistic estimation of the performance.

### **3.2 Penalized Regression Method**

Regression is a very powerful statistical method to study the association between an outcome variable and a profile of covariates. In the statistical literature, outcome variable is also referred to as dependent variable, and covariates are sometimes called independent variables or predictors. In a regression model, each covariate is coupled with a regression coefficient that measures the strength of its association with the outcome variable. Many regression methods are available when the outcome variable is time-to-event data. A commonly used method is the Cox's proportional hazards regression [10], which assumes the hazard of event changes proportionally with covariates and the proportional change remains the same over time. This proportional hazards assumption needs to hold for each covariates in the regression model. Fortunately, this

assumption can be verified by data both graphically and statistically (Chapter 11 of [22]). If the assumption is found violated for some covariates, a penalized Cox's regression may result in incorrect selection of covariates and biased estimate of the regression coefficients. Several methods are available for remedy. If the covariate that violates the proportional hazards assumption is categorical, one can stratify the data by this covariate and fit stratified penalized Cox's proportional hazards regression. For continuous covariates, one could bin its values into a few categories and then stratify the data by it. The drawback of this method is that the effect of the stratifying variable on the survival outcome cannot be estimated. Moreover, when the number of violating covariates is large, stratification by those covariates simultaneously may be infeasible. Another strategy is to include an interaction between the violating covariate and logarithm of the survival time. By incorporating time into the covariate, the Cox's model can now accommodate non-proportional hazards. Given the versatility and popularity of the Cox's proportional hazards regression, this section focuses on this method.

Penalized regression has become an increasingly popular method for variable selection. The advantage of penalized regression over the filter-test based methods is that the importance of covariates is assessed jointly in one regression model so that the potential confounding effects among the covariates are accounted for, leading to more accurate identification of important covariates. The basic idea of penalized regression is to give a penalty to each regression coefficient when fitting the regression model so that the estimated coefficient is shrunken towards zero due to the penalty. If a regression coefficient is too small to start with, then after the shrinkage it will become exactly zero, and therefore its corresponding covariate will be excluded from the regression model, achieving the purpose of variable selection. When the number of covariates exceeds the sample size, the conventional regression model will break down due to linear dependency among the covariates. The penalized regression avoids the breakdown by automatically identifying a sparse model where the number of non-zero coefficients is less than the sample size.

Since the seminal paper that proposed the lasso penalty [30], numerous other penalties have been developed in the literature that possess various statistical properties. This section focuses on four widely used penalties: lasso, adaptive lasso, group lasso, and elastic net. These four penalties are, in the authors' opinion, the most useful and easiest to implement in biomedical research. Lasso penalty was proposed in 1996 for linear regression [30] and was extended to Cox's proportional hazards regression [31]. Lasso penalty is proportional to the absolute value of the regression coefficient. Under the Cox's proportional hazards regression model, lasso penalized regression coefficient estimate is obtained

from a sample of size  $n$  by maximizing the following objective function:

$$\sum_{i=1}^n \left[ x_i \beta - \log \left\{ \sum_{j \in R_i} \exp(x_j \beta) \right\} \right] \Delta_i - \lambda \sum_{j=1}^p |\beta_j|, \quad (1)$$

where  $x_i$  is a  $p$ -dimensional covariates of subject  $i$ ,  $\beta$  is the corresponding  $p$ -dimensional regression coefficient,  $R_i$  is the set of subjects still at risk at the failure time of subject  $i$ ,  $\Delta_i$  is the failure indicator of subject  $i$  with 1 being failure and 0 being censored,  $\lambda$  is a tuning parameter that controls the magnitude of lasso penalty. The first term in (1) is the log-partial likelihood function and the second term is the lasso penalty. Many efficient algorithms have been developed to solve the lasso penalized objective function [13, 26] and implemented in R packages. Please refer to Subheading 2 for details.

One disadvantage of lasso penalty is that the estimated non-zero coefficients are biased towards zero. Adaptive lasso penalty was proposed to overcome this disadvantage [36, 38]. In adaptive lasso penalty, a data-driven weight is multiplied to the original lasso penalty of each coefficient. The weight is typically chosen as  $1/|\tilde{\beta}|$ , where  $\tilde{\beta}$  is the maximizer of the non-penalized partial likelihood. The objective function of adaptive lasso penalized Cox's proportional hazards regression is

$$\sum_{i=1}^n \left[ x_i \beta - \log \left\{ \sum_{j \in R_i} \exp(x_j \beta) \right\} \right] \Delta_i - \lambda \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|}.$$

It is shown that the adaptive lasso penalized regression can correctly identify the zero coefficients and the estimated non-zero coefficients are asymptotically unbiased if  $\tilde{\beta}$  is a consistent estimate of the true coefficients [36]. Thus, if in addition to identifying covariates that are associated with the survival outcome the researcher is also interested in the actual magnitude of association, then adaptive lasso penalty should be preferred over the original lasso penalty. An intuitive explanation for the unbiasedness of adaptive lasso penalty is that, if the true coefficient is large, its consistent estimate tends to be large and therefore its adaptive lasso penalty is small, thereby avoiding over-penalizing large coefficients. When the number of covariates is greater than the sample size, it is not straightforward to obtain the initial consistent estimate of the coefficients. Huang et al. proposed to use the estimate from univariate regression to construct the weight and proved its asymptotic properties [19]. Since the only difference between the original lasso penalty and adaptive lasso penalty is the weight function, the algorithms developed for lasso can be easily modified to compute adaptive lasso estimator. In fact, most R packages that implement algorithms for

lasso have an option for the users to specify a weight function, where the adaptive lasso weight can be specified.

In many situations, it is desirable to select variables in groups. That is, the covariates in the same group are either all selected into or all excluded from the regression model. For example, single nucleotide polymorphism (SNP) is typically recorded as a categorical variable with three levels determined by the genotype of the two alleles, and can be represented by two dummy variables in a regression model. It is then natural to consider these two dummy variables as a group when performing variable selection. As another example, multiple genes often function in the same biological pathway. Therefore, when assessing the association between gene expression profile and the survival outcome, one may like to select all genes belonging to the same biological pathway in a group fashion. Another example where group selection is desirable is to accommodate covariates with non-proportional hazards. As mentioned earlier, one strategy for handling covariate with non-proportional hazards is to include an interaction between the violating covariate and log-survival time in the Cox regression model. To implement this strategy under penalized regression setting, one would need to select the violating covariate and its interaction with log-survival time as a group. Group lasso has been developed to address this issue [34], which was further extended and adapted to Cox's proportional hazards regression by Simon et al. [27]. To perform group lasso penalized regression, investigators first assign covariates into non-overlapping groups. Suppose there are  $L$  such groups. The objective function of group lasso penalized Cox's proportional hazards regression is

$$\sum_{i=1}^n \left[ x_i \beta - \log \left\{ \sum_{j \in R_i} \exp(x_j \beta) \right\} \right] \Delta_i - \lambda \sum_{l=1}^L \sqrt{p_l} \|\beta_{(l)}\|,$$

where  $p_l$  is the size of group  $l$ ,  $\beta$  is the vector of regression coefficients of all covariates,  $\beta_{(l)}$  is the vector of regression coefficients of covariates in group  $l$ , and  $\|\cdot\|$  denotes the Euclidean norm. With this formulation, variable selection is conducted on the group level, achieving group variable selection. Note that when all group sizes equal one, the group lasso penalty reduces to lasso penalty. Once the important groups of covariates have been identified, researchers are sometimes interested in further identifying within the selected groups which covariates are more important than others. For example, the genes in a selected biological pathway may not have the same strength of association with the outcome. Therefore, it would be useful if group lasso penalty can select covariates on the group level as well as the individual level within groups. Simon et al. developed sparse group lasso to achieve this purpose, where a linear combination of lasso penalty and group lasso penalty is used to penalize the log-partial likelihood function,

$$\sum_{i=1}^n \left[ x_i \beta - \log \left\{ \sum_{j \in R_i} \exp(x_j \beta) \right\} \right] \times \Delta_i - (1-\alpha)\lambda \sum_{l=1}^L \sqrt{p_l} \beta_{(l)} - \alpha \lambda \sum_{j=1}^p |\beta_j|,$$

where  $\alpha \in [0, 1]$ . When  $\alpha = 1$  the penalty reduces to the original lasso; when  $\alpha = 0$  it becomes group lasso. Thus,  $\alpha$  controls the relative importance of group-level and individual-level variable selection. The model identified by sparse group lasso contains the selected groups, which in turn contain only the selected individual covariates within the selected groups. An algorithm to solve the sparse group lasso problem has been developed and implemented in R package [27].

Group lasso and sparse group lasso penalties require the specification of the grouping of covariates before fitting the penalized regression model. This makes sense when the grouping of covariates is intrinsic (e.g. dummy variables representing the same categorical covariate) or can be determined by prior knowledge in the subject fields (e.g. genes in the same biological pathway). In many studies, however, the investigators do not have a definitive way to group covariates a priori but in general still would like to select covariates together that are highly correlated with each other. This is particularly true in high-dimensional variable selection. These strongly correlated covariates may jointly represent some unknown factors or mechanisms, and therefore it would make sense to include all of them in the model if any one of them is selected. Unfortunately, the original lasso penalty has been shown incapable of this task since it tends to select only one of the highly correlated covariates and ignores the others and it does not care which one to select [39]. To meet the practical need, Zou and Hastie proposed a new penalty named elastic net [39], which is a linear combination of ridge [18] and lasso penalties. Ridge penalty has been widely used in statistical analysis and other fields. It can accommodate nearly linear dependent covariates in a regression model but does not perform variable selection (i.e. it never sets a regression coefficient to zero). Elastic net penalty combines the properties of ridge and lasso penalties to achieve a balance between sparsity and group selection. It was later extended to the Cox's proportional hazards model [16, 24]. The objective function of elastic net penalized Cox's proportional hazards model is

$$\sum_{i=1}^n \left[ x_i \beta - \log \left\{ \sum_{j \in R_i} \exp(x_j \beta) \right\} \right] \times \Delta_i - \frac{1}{2}(1-\alpha)\lambda \sum_{j=1}^p \beta^2 - \alpha \lambda \sum_{j=1}^p |\beta_j|,$$

where  $\alpha \in [0, 1]$ . When  $\alpha = 1$  the penalty reduces to the original lasso; when  $\alpha = 0$  it becomes ridge penalty. Elastic net penalty is similar to sparse group lasso penalty both in format and effect. The key difference is that the latter requires pre-specification of the grouping of covariates, whereas the former automatically determines which covariates are selected as groups based on the

correlation between them. The algorithm for solving elastic net penalized Cox's regression has been implemented in R package [26].

In all the penalties reviewed above, there is one or more tuning parameters that control the magnitude or structure of the penalties. In particular, the tuning parameter  $\lambda$  controls the magnitude of the penalties and in turn controls the complexity of the selected model. Each value of  $\lambda$  gives rise to a model. When  $\lambda = 0$ , there is no penalty and all candidate covariates are selected into the model; when  $\lambda \rightarrow \infty$ , the penalty becomes so heavy that none of the candidate covariates is selected, resulting in an empty model. The collection of models corresponding to all possible values of  $\lambda$  is called the solution path of  $\lambda$ . The asymptotic properties of those penalties ensure that the solution path will contain the true model as sample size goes to infinity. However, one still has to determine which model in the solution path is the true model; in other words, which value of  $\lambda$  should be used to obtain the model. In practice, tuning parameters are often selected by a data-driven fashion. For a series of tuning parameter values, their corresponding models are evaluated using some criteria, and the value that results in the best model is chosen as the optimal tuning parameter value.

One commonly used criterion for tuning parameter selection under Cox's proportional hazards model is the log-partial likelihood, which measures the goodness of fit of the model. One problem with this measure is that it will always increase if more covariates are included in the model because larger models will always fit the data better. At some point, however, too large a model will start to explain the random noise that is unique to the data at hand, resulting in overfitting of the data. To avoid overfitting,  $k$ -fold cross-validation (CV) is often used to select the tuning parameters. In this technique, the full data is split into  $k$  non-overlapping portions. Then,  $k - 1$  portions (sometimes called training data) are used to build the penalized regression models and the  $k$ th portion (sometimes called validation data) is used to calculate the log-partial likelihood of the fitted model. In survival analysis, when the number of events is small or  $k$  is large, the partial likelihood of the validation data may be ill-defined [26]. To overcome this difficulty, van Houwelingen et al. [32] proposed the following statistic to replace the log-partial likelihood on the validation data:

$$\hat{CV}_i(\lambda) = \ell\{\hat{\beta}_{-i}(\lambda)\} - \ell_{-i}\{\hat{\beta}_{-i}(\lambda)\}, \quad i = 1, \dots, k,$$

where  $\hat{\beta}_{-i}(\lambda)$  is the penalized estimate of regression coefficients from the training data with tuning parameter  $\lambda$ ,  $\ell\{\hat{\beta}_{-i}(\lambda)\}$  is the log-partial likelihood for the full data evaluated at  $\hat{\beta}_{-i}(\lambda)$ , and  $\ell_{-i}\{\hat{\beta}_{-i}(\lambda)\}$  is the log-partial likelihood for the training data

evaluated at  $\hat{\beta}_{-i}(\lambda)$ . This procedure cycles through the  $k$  portions, resulting in  $k \widehat{CV}_i(\lambda)$ 's, which are averaged to obtain the cross-validated statistic  $\widehat{CV}(\lambda)$ . The tuning parameters that give rise to a model with the largest  $\widehat{CV}(\lambda)$  is chosen as the optimal tuning parameter. The fold number  $k$  is typically set to five or ten. Besides  $k$ -fold CV, there are other methods for tuning parameter selection including generalized cross-validation (GCV) [11], Akaike information criterion (AIC) [2], Bayesian information criterion (BIC) [25], among others. The  $k$ -fold CV, GCV, and AIC are closely connected. GCV can be thought of as CV with  $k$  equal the sample size, and AIC is approximately equal to log-GCV when the sample size is large. Even though these three methods were devised to prevent overfitting, it has been shown that they still tend to select a model that is larger than the true model [17, 23, 33]. On the other hand, BIC has been shown to select the tuning parameter that leads to the correct model as sample size goes to infinity [23, 33]. In practice with finite sample sizes,  $k$ -fold CV, GCV, and AIC often give very similar models with better prediction performance than the model chosen by BIC, which tends to be more parsimonious but may be more interpretable.

The penalized variable selection methods are generally more accurate in selecting the correct variables than the filter-based method because they take into account the correlation among the covariates. However, due to their sophistication, penalized methods are best suited for studies where the number of covariates are comparable with (e.g. in the same order of) the sample size and their performance may become suboptimal when the number of covariates are several orders higher than the sample size.

In conventional regression analysis, it is straightforward to make statistical inference on the estimated regression coefficients such as calculating  $p$  values and confidence intervals. Under penalized regression, however, these tasks turn out to be very challenging. Standard inference techniques are not valid anymore because they fail to incorporate the uncertainty of variable selection procedure into the inference. The easiest way to obtain the correct inference on a model selected by a penalized variable selection procedure is to split the original data into halves and use one half to perform penalized variable selection to identify a model and fit the selected model on the other half to obtain statistical inference on the estimated regression coefficients. However, when the size of the original data is not large enough, this strategy may suffer from low efficiency. Post-selection inference is an active research area. Many more sophisticated methods are being developed to address this challenge. An in-depth review of this topic is beyond the scope of this article. Interested readers are referred to [12] for a comprehensive review of this area.

### 3.3 Machine Learning Method

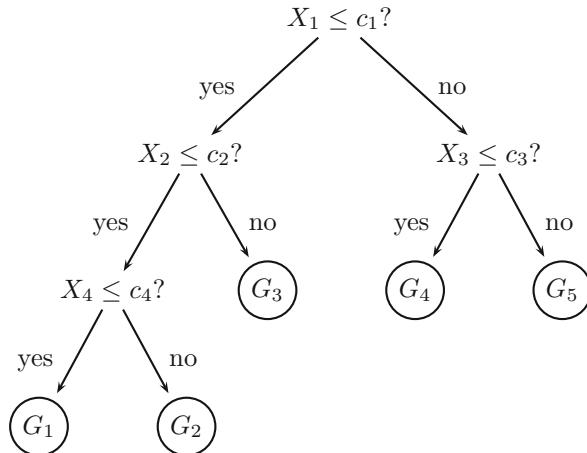
Besides traditional statistical methods, machine learning algorithms are also widely used for variable selection in time-to-event data. Machine learning methods are often categorized into supervised learning, unsupervised learning, and semisupervised learning, depending on whether the outcome variable is observed and used in the learning process. In supervised and semisupervised learning methods the outcome variables are used or partially used to build a prediction model by minimizing the difference between the predicted outcome and the observed outcome. In time-to-event data, the outcome is the survival time with or without censoring. The independent variables used to build the prediction model, such as gene expression levels, are often called predictors. Similar to penalized regression, some machine learning methods can also select the most effective predictors to build the model, which essentially is variable selection. Although penalized regression methods sometimes are also referred as machine learning methods, we will focus on other types of machine learning methods in this subsection, including survival tree, random survival forest, and supervised principle component analysis.

Machine learning methods can capture covariates with highly nonlinear relationship with the outcome such as polynomial relationship or multi-way interaction effects that are otherwise difficult to detect by other variable selection methods. However, they achieve it at the cost of interpretability of the results. Usually an estimated effect size of the selected covariate is not available from machine learning methods. In addition, machine learning methods typically require larger sample sizes than the filter-based methods and penalized variable selection methods. In the context of time-to-event data, they require a large number of events.

#### 3.3.1 Survival Tree and Random Survival Forest

Many regression-based survival analysis methods are parametric, which explicitly impose a parametric model that links the survival outcome to the predictors [6]. Assumptions such as linear assumption and proportional hazard assumption are often made to fit these models. In high-dimensional applications, such as with genomic studies, it is often infeasible to examine these assumptions for each of the thousands of genes. Moreover, because of the large number of genes, it is also intractable to explore all possible gene–gene interactions. In most applications, the interactions are only explored by post hoc analysis after feature selection. A major drawback of this approach is that the main effect of the interacting genes may not be significant enough to be selected due to a phenomenon known as Simpson’s paradox [28]. Tree-based methods, on the other hand, are less affected by this issue and can be used to explore gene–gene interactions while doing feature selection.

In the general framework, a tree-based method first divides all subjects in the data into multiple (often two) subgroups based on a selected gene feature. Then each subgroup will be further divided



**Fig. 1** An example binary decision tree with four nodes

based on a newly selected feature. The division step will continue until certain stopping rule has been met or the subgroups become “pure” enough. This strategy is also known as “recursive partitioning” or “divide and conquer.” The partitioning path will result in a classification rule look like an upside-down tree with multiple branches, which is the reason for the name “tree-based” methods or decision “trees.” Please see Fig. 1 for an example tree, where the subjects are divided into five subgroups ( $G_1, \dots, G_5$ ) based on four features ( $X_1, \dots, X_4$ ).

When constructing a tree, the features that are used to split the tree are heuristically selected. Greedy algorithms are often involved at each branch to choose the feature and cutoff that best split the subjects into subgroups. To measure the quality of the split, different criteria can be used [6]. Among them, log-rank statistic [7] and likelihood ratio statistic under Cox proportional hazard model or exponential models [8] are the most popular ones. Nonparametric criteria [9], residual-based criteria [1, 29], criteria that combine both the survival and censoring information [35] are also proposed to construct the trees. Because a decision tree uses multiple features to split the branches, the tree construction method is inherently a feature selection procedure. It is possible to select the features using the tree-based method then use other modeling strategy to further explore the relationship between the features and the survival outcome.

Because the tree construction is often a deterministic procedure based on the observed data, which is only realization of the inherently random distribution that the data could be, with limited sample size, constructing a single tree could fail to capture the random nature of the observed data and would not be able to assess the robustness of the variable selection result. To overcome this shortness of single tree construction, random forests are proposed.

A random forest is collection of trees, where each tree is constructed based on a bootstrap sample of the subjects and a random subset of the features. Then the random forest can be used to calculate the variable importance for each feature. Features with high variable importance can be selected for further analysis. The variable importance can be calculated from the out-of-bag (OOB) samples by perturbing the features [20]. For a given tree in the random forest, the OOB samples are those samples not selected by the bootstrap procedure which in turn did not contribute to the construction of this tree. After perturbing a feature, concordance index (C-index) can be calculated to measure how much the prediction performance of the random survival forest has decreased. Apparently, perturbing an important feature would significantly decrease the performance, while perturbing an unimportant feature would have little influence on the performance of the forest. Another way to calculate the variable importance for the features is based on the frequency and depth of the features being selected in the trees [21]. The depth measures how “close” a split is to the “root” of the tree. An important feature tends to be selected earlier in a tree, thus tend to has lower depth.

### 3.3.2 Supervise Principal Component

In addition to survival tree and random survival forest, supervised principle component analysis (SuperPC) [4] is another widely used machine learning method for feature selection in gene expression studies with time-to-event outcome. The feature selection in SuperPC is performed in two stages. In the first stage, a survival regression model (e.g. Cox proportional hazard model) is fitted for each of the features, and only the features with coefficients larger than a threshold will be selected into the next stage. This threshold is determined by cross-validation. In the second stage, a principal component analysis is performed using only the features selected from the first stage. Then the correlation between each feature and the first principal component is calculated as the important score. The final selected features are only those whose important score is larger than a second threshold, which again is determined by cross-validation. An alternative approach for the second stage is to treat the survival prediction problem as a binary classification problem with high-risk and low-risk groups, and use the nearest shrunken centroid method to select only those features whose mean values are different enough between the two groups [3]. The grouping of the subjects into the high-risk and low-risk is based on whether they survive beyond the median survival time. For those subjects censored before the median survival time, Kaplan–Meier estimator is employed to estimate the probability that they belong to either group. The nearest shrunken centroid method is accordingly modified to accommodate the probabilistic grouping labels. Comparing to the simple filtering methods, SuperPC will select a smaller set of features, which best predict the survival outcome in the cross-validation.

## References

1. Ahn H, Loh WY (1994) Tree-structured proportional hazards regression modeling. *Biometrics* 50:471–485
2. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov NN, Csaki F (eds) Second international symposium on information theory, pp 267–281
3. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2(4):e108
4. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. *J Am Stat Assoc* 101(473):119–137
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol)* 57(1):289–300
6. Bou-Hamad I, Larocque D, Ben-Ameur H, et al (2011) A review of survival trees. *Stat Surv* 5:44–71
7. Ciampi A, Thiffault J, Nakache JP, Asselain B (1986) Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Comput Stat Data Anal* 4(3):185–204
8. Ciampi A, Chang CH, Hogg S, McKinney S (1987) Recursive partition: a versatile method for exploratory-data analysis in biostatistics. In: *Biostatistics*. Springer, Berlin, pp 23–50
9. Ciampi A, Hogg SA, McKinney S, Thiffault J (1988) RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features. *Comput Methods Prog Biomed* 26(3):239–256
10. Cox DR (1972) Regression models and life-tables. *J R Stat Soc (Ser B)* 34(2):187–220
11. Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer Math* 31:377–403
12. Dezeure R, Bühlmann P, Meier L, Meinshausen N (2015) High-dimensional inference: confidence intervals, p-values and R-software hdi. *Stat Sci* 30:533–558
13. Efron B, Hastie T, Johnstone I, Tibshirani RJ (2004) Least angle regression. *Ann Stat* 32(2):407–451. <http://www.jstor.org/stable/3448465>
14. Fan J, Li G, Li R (2005) An overview on variable selection for survival analysis. In: *Contemporary multivariate analysis and design of experiments: in celebration of Professor Kai-Tai Fang's 65th birthday*. World Scientific, Singapore, pp 315–336
15. Friedman J, Hastie T, Tibshirani R (2009) glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1(4)
16. Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. *Biom J* 52(1):70–84
17. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference and prediction*, 2nd edn. Springer. <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
18. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67
19. Huang J, Ma S, Zhang CH (2008) Adaptive lasso for sparse high-dimensional regression models. *Stat Sin* 18:1603–1618
20. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Ann Appl Stat* 2(3):841–860
21. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010) High-dimensional variable selection for survival data. *J Am Stat Assoc* 105(489):205–217
22. Klein JP, Moeschberger ML (2006) *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, Berlin
23. Ni A, Cai J (2018) Tuning parameter selection in Cox proportional hazards model with a diverging number of parameters. *Scand J Stat* 45(3):557–570
24. Park MY, Hastie T (2007) L1-regularization path algorithm for generalized linear models. *J R Stat Soc Ser B (Stat Methodol)* 69(4):659–677
25. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
26. Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw* 39(5):1
27. Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. *J Comput Graph Stat* 22(2):231–245

28. Simpson EH (1951) The interpretation of interaction in contingency tables. *J R Stat Soc Ser B (Methodol)* 13(2):238–241
29. Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-based residuals for survival models. *Biometrika* 77(1):147–160
30. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc (Ser B)* 58:267–288
31. Tibshirani RJ (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16(4):385–395
32. van Houwelingen HC, Bruinsma T, Hart AA, van't Veer LJ, Wessels LF (2006) Cross-validated Cox regression on microarray gene expression data. *Stat Med* 25(18):3201–3216
33. Wang H, Li R, Tsai CL (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3):553–568
34. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *J R Stat Soc (Ser B)* 68(1):49–67
35. Zhang H (1995) Splitting criteria in survival trees. In: Statistical modelling. Springer, Berlin, pp 305–313
36. Zhang HH, Lu W (2007) Adaptive lasso for Cox's proportional hazards model. *Biometrika* 94(3):691–703
37. Zhao SD, Li Y (2012) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J Multivar Anal* 105(1):397–411
38. Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101(476):1418–1429
39. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc: Ser B (Stat Methodol)* 67(2):301–320



# Chapter 6

## Binary Classification for Failure Risk Assessment

Ali Foroughi Pour, Ian Loveless, Grzegorz Rempala, and Maciej Pietrzak

### Abstract

Survival analysis is tremendously powerful, and is a popular methodology for analyzing time to event models in bioinformatics. Furthermore, several of its extensions can simultaneously perform variable selection in conjunction with model estimation. While this flexibility is extremely desirable, under certain scenarios, binary class variable selection and classification methods might provide more reliable risk estimates. Synthetic simulations and real data case studies suggest that when (1) randomly censored points comprise only a small portion of data, (2) biological markers are weak, (3) it is desired to compute risk across predetermined time intervals, and (4) the assumptions of the competing time to event models are violated, binary class models tend to perform superior. In practice, it might be prudent to test both model families to guarantee adequate analysis. Here we describe the pipeline of binary class feature selection and classification for time to event risk assessment.

**Key words** Risk assessment, Survival analysis, Variable selection, Classification

---

### 1 Introduction

Time to event survival analysis (TESA) is tremendously powerful. It assigns a risk function to the failure event occurring at each future time point, i.e., assigns a distribution to the failure time of each observation based on the observed covariates rather than assigning a fixed risk statistic. Thereby, one can assess the probability of failure at any future time interval, and identify time regions where failure is more likely. Additionally, many TESA models, such as the Cox proportional hazard model, also explain the extent each covariate contributes the hazard function. TESA models can also handle fixed censoring (failure not happening until the end of study), as well as random censoring, e.g., due to loss to follow-up or patient drop out. To top it off, when used with penalty terms such as LASSO and elastic net, they also automatically perform variable selection.

---

The authors Ali Foroughi pour and Ian Loveless contributed equally to this work.

The Cox proportional hazard model is maybe the most popular method for TESA. It is simple, easy to use, and is a satisfactory model for many classical problems. However, this may be a double edged sword! In many modern applications, such as studying complex diseases, we deal with (1) small-sample high-dimensional data, (2) large number of covariates compared to sample size, (3) and complex variable interactions, which may heavily violate the proportional hazard assumption. Although more complex hazard functions can always be considered, such extensions may not be easy to work with and may be problem dependent. This may be extremely challenging for a practitioner whose main goal is to arrive at a reasonable initial statistical analysis that explains the general traits of data, not extending the survival analysis theory.

Here we will describe a methodology, based on binary classification theory and hereafter called binary classification for survival analysis (BCSA), that can be extremely useful in many practical settings where TESA fails. Note certain classification strategies, such as using a generalized linear model (GLM) with logit link inputted with binary inputs, are already studied in the survival analysis literature, e.g., in [16]; however, it seems (1) they are to a large extent under-appreciated, and (2) ignore that BCSA is not limited to GLMs. While in the statistical and signal processing literature such methods have been extensively studied, they seem to be under-appreciated in practical scenarios. To address this gap between theory and practice, here, we provide an overview of (a) the settings where BCSA can provide an insightful statistical analysis, (b) algorithms within the classification theory that can be employed, (c) how they work in conjunction with variable selection strategies in the pipeline, (d) provide numerous examples using synthetic simulations and real cancer data, and (e) provide examples that can be used as rules of thumb for determining the appropriate approach, TESA or BCSA, for the data at hand. In particular we will see that if (1) the goal is to find survivorship up to a fixed future time point or in fixed intervals rather than any potential time interval, (2) randomly censored points do not comprise a large portion of data, and (3) the number of covariates to consider is large compared with the sample size, it is wise to explore BCSA methods, particularly when the assumptions of the alternative TESA model are violated.

---

## 2 The Binary Classification Approach

Here we first describe how binary classification can be used to assign risks, i.e., probabilistic measures of failure within a time interval, to each observation. We then explain several popular classification algorithms that can be used to obtain failure risks.

## 2.1 Problem Setup

Suppose we are given a set of observations comprised of the triple  $(x, t, y)$  obtained via a trial with follow-up time  $T_f$  where  $x \in \mathbb{R}^p$  is the set of  $p$  covariates,  $t$  is the failure time, and  $y$  is the indicator of failure up to time  $T_f$ . In other words,  $y=0$  means failure did not happen by the end of the study and  $t=T_f$ ; otherwise, failure occurs before time  $T_f$ ,  $y=1$ , and  $t < T_f$ . Note for now we are assuming there is no random censoring; however, extensions that consider such scenarios will be discussed later.

In the binary classification setup the goal is to assign a risk statistic to each observation that describes the probability of failure up to time  $T_f$ . Note, without loss of generality, here we have assumed the desired time interval we wish to assign a failure risk to is the entire follow-up time. However, if a shorter time interval is of interest, say  $T_d < T_f$ , one can pretend  $T_d$  is the follow-up time and set  $y$  to be the indicator of  $t < T_d$ . More complex scenarios are discussed later. Note in BCSA we only encode if failure occurs before  $T_f$  and ignore the information encoded in the exact failure time. In particular, among two data points for which failure occurs, the one with the smaller failure time is more likely to have a larger risk; however, in many cases, any reasonable BCSA algorithm automatically and correctly infers this information from data, eliminating the need to vehemently enforce it into the statistical analysis.

To perform binary classification, train data,  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , is the collection of  $n$  ordered pairs of the form (covariate vector, failure indicator), and given a new observation, i.e., test data  $X$ , with unknown true label  $\gamma$ , we want to be able to predict  $\gamma$  based on  $X$  with a high accuracy. Therefore, we would like to compute the probability of  $\gamma=1$  given observation  $X$ , denoted by  $P(\gamma=1|X)$ . Note the probability of no failure is  $P(\gamma=0|X) = 1 - P(\gamma=1|X)$ . Finally, in classification, points with the same  $y$  are called a class, i.e., points for which  $y=0$  are called class 0, and points for which  $y=1$  are called class 1. Instead of  $P(\gamma=1|X)$  we can equivalently compute

$$\log \left( \frac{P(\gamma=1|X)}{1 - P(\gamma=1|X)} \right) = \log \left( \frac{\pi(\gamma=1)}{1 - \pi(\gamma=1)} \right) + \log \left( \frac{f(X|\gamma=1)}{f(X|\gamma=0)} \right), \quad (1)$$

where  $\pi(\gamma=1)$  is the *prior probability* of failure event ( $\gamma=1$ ) across the population before observing  $X$ . For example, the probability that a cancer patient relapses up to time  $T_f$  across the population.  $f(X|\gamma=1)$  is the *probability density function (p.d.f.)* of the distribution of class 1 at  $X$ , and similarly for  $f(X|\gamma=0)$ . Note

$$L(X) = \log \left( \frac{f(X|\gamma=1)}{f(X|\gamma=0)} \right), \quad (2)$$

is the famous *log likelihood ratio*. Note to obtain the log likelihood ratio we need the exact distributions of each class, which is never available! Instead, we would like to estimate them given our collected data,  $S$ .

Given  $S$ , the goal of BCSA is to arrive at a model to obtain  $L(X)$ , or more precisely,  $L(X|S)$ , which we then use as a *risk* function. For example, a test data  $X$  with  $L(X) > 0 / L(X) < 0$  is more likely to experience/not experience failure up to time  $T_f$ . Note given  $L(X)$  we can use Eq. 1 to compute the probability of failure. Additionally,  $L(X)$  can be used to label  $X$  as a point with high/low failure risk: given threshold  $T$ , if  $L(X) > T$ , then  $X$  is labeled as high risk, and low risk otherwise.

We now describe several methods that can serve as machineries to use the collected data,  $S$ , to compute  $L(X)$ . Note these methods have different assumptions, and hence, might be suitable for different experimental conditions. It is extremely difficult to confidently declare which one is best for each dataset, and it may be prudent to test several competing methods in practice. We believe the methods described below are suitable for a large family of practical applications; however, a deeper discussion of these algorithms can be found in [3, 26].

## 2.2 Classification Algorithms

### 2.2.1 GLM with Logit Link

Here we describe several algorithms that can be used to estimate the function that takes test point  $X$  as input and computes the log likelihood ratio of  $X$  belonging to class 1, i.e., the high risk class, which we denoted by  $L(X)$  in the previous section. Recall that  $L(X)$  is our risk function, and large  $L(X)$  is an indicator of high chance of failure.

The GLM with logit link may be the most straightforward model to consider, and is among the few algorithms that are studied in survival analysis literature, e.g., in [16]. It assumes

$$L(X) = \beta_0 + \sum_{i=1}^p \beta_i X_i, \quad (3)$$

and given train data,  $S$ , it aims to estimate coefficient vector  $\beta = [\beta_0, \beta_1, \dots, \beta_p]$ . Maximum likelihood (ML) is a popular approach for estimating  $\beta$ , and is described in detail in [20].

### 2.2.2 GLM with Probit Link

It is easier to study the probit link via its modeling of  $P(Y=1|X)$ . The probit link assumes

$$P(Y=1|X) = 1 - \Phi\left(\beta_0 + \sum_{i=1}^p \beta_i X_i\right), \quad (4)$$

where  $\Phi(\cdot)$  is the cumulative distribution function (CDF) of the standard normal distribution. Again, given data, ML estimation can be used to compute  $\beta = [\beta_0, \beta_1, \dots, \beta_p]$ . See ref. 20 for details.

### 2.2.3 GLMs with Quadratic Kernels

The GLMs discussed here assume a linear models; however, it is possible to implement more complex predictor functions, including but not limited to quadratic models. In the case of logit link, for each observation  $x_i \in S$ , we consider  $z_i = [x_1, \dots, x_p, x_1 x_2, \dots, x_{p-1} x_p]$ . We then proceed similar to the linear model substituting each  $x_i$  with  $z_i$ , and  $X$  with  $Z$ , constructed similar to  $z_i$ 's. Note in many cases, unless  $p$  is very small, the sheer number of predictors might create an ill-posed problem. In such cases one needs to use penalties, such as LASSO, ridge, or elastic net.

### 2.2.4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) assumes covariates are jointly Gaussian in each class and the covariance matrix is the same in both classes. Given data, (a) sample mean ( $\hat{\mu}_y$ ) and sample covariance ( $\hat{\Sigma}_y$ ) in each class are computed, (b) class-conditioned covariance estimates are used to compute the pooled covariance estimate  $\hat{\Sigma}_p = (n_0 - 1)/(n_0 + n_1 - 2)\hat{\Sigma}_0 + (n_1 - 1)/(n_0 + n_1 - 2)\hat{\Sigma}_1$ , where  $n_y$  is sample size in class  $y$ , and (c) pretending the sample means and pooled covariance are the true distribution parameters,  $L(X)$  is computed. It can be shown that in this case

$$L(X) = \beta_0 + \sum_{i=1}^p \beta_i X_i, \quad (5)$$

where

$$\begin{aligned} \beta_0 &= 0.5 \left( \hat{\mu}_0^T (\hat{\Sigma}_p)^{-1} \hat{\mu}_0 - \hat{\mu}_1^T (\hat{\Sigma}_p)^{-1} \hat{\mu}_1 \right), \\ \beta &= [\beta_1, \dots, \beta_p]^T = (\hat{\Sigma}_p)^{-1} (\hat{\mu}_1 - \hat{\mu}_0), \end{aligned}$$

the super-script “ $T$ ” denotes transpose, and for an invertible matrix  $M$ ,  $M^{-1}$  is its inverse. Several extensions to LDA are proposed, of which many directly aim to address problems where the number of covariates is comparable or even larger than sample size. Diagonal LDA (DLA) assumes the pooled covariance matrix is diagonal, and regularized LDA (RLDA) uses the regularized pooled covariance matrix, hereafter denoted by  $\hat{\Sigma}_r$ , instead

$$\hat{\Sigma}_r = \lambda \times s \times I + (1 - \lambda) \hat{\Sigma}_p. \quad (6)$$

where  $\hat{\Sigma}_r$  is the regularized pooled covariance matrix estimate,  $I$  is the  $p$ -by- $p$  identity matrix,  $\lambda \in (0, 1)$  is the amount of regularization, and  $s$  is mean of pooled sample variances, i.e., diagonal elements of  $\hat{\Sigma}_p$ . Note many authors fix  $s = 1$ ; however, we believe the

formulation of (6) is more appropriate as it better scales the regularization to the variability of the data at hand. Finally, note the amount of regularization to achieve “best” results depends on the underlying data parameters and sample size, and is typically determined during the internal validation process, for instance, cross validation.

### 2.2.5 Quadratic Discriminant Analysis

Similar to LDA, Quadratic Discriminant Analysis (QDA) assumes covariates in each class are Gaussian; however, it does not assume they have the same covariance matrix in both classes, providing more degrees of freedom to the designed classification rule. Again, given data we (1) find sample mean and covariances in each class, (2) pretend they are the true underlying distribution parameters, and (3) use them to compute  $L(X)$ . We have

$$L(X) = 0.5X^TAX + B^TX + C, \quad (7)$$

where

$$\begin{aligned} A &= (\hat{\Sigma}_0)^{-1} - (\hat{\Sigma}_1)^{-1}, \\ B &= (\hat{\Sigma}_0)^{-1}\hat{\mu}_0 - (\hat{\Sigma}_1)^{-1}\hat{\mu}_1, \\ C &= 0.5\left(\hat{\mu}_0^T(\hat{\Sigma}_0)^{-1}\hat{\mu}_0 - \hat{\mu}_1^T(\hat{\Sigma}_1)^{-1}\hat{\mu}_1\right) + 0.5\log\left(\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|}\right), \end{aligned}$$

and for a matrix  $M$ ,  $|M|$  is its determinant. Variants similar to LDA can also be considered for QDA. Diagonal QDA (DQDA) assumes covariance matrices are diagonal in each class, and regularized QDA (QDA) uses regularized sample covariance instead:

$$\hat{\Sigma}_{r,y} = \lambda \times s_y \times I + (1 - \lambda)\hat{\Sigma}_y, \quad (8)$$

where  $s_y$  is the mean of sample variances in each class, i.e., diagonal elements of  $\hat{\Sigma}_y$ . Again the appropriate amount of regularization is determined during validation.

### 2.2.6 $k$ Nearest Neighbors (kNN)

$k$  nearest neighbors (kNN) is another popular method for classification, which is inherently a degenerative model, i.e., does not use a mathematical model to arrive at its estimation of  $L(X)$ . For two points  $p_1, p_2 \in \mathbb{R}^p$ , let  $d(p_1, p_2)$  be the distance between  $p_1$  and  $p_2$ . Euclidean distance (also known as  $L_2$  distance) is widely popular, but is not the only distance that can be used with kNN. Manhattan distance (also known as city block and  $L_1$  distance), cosine distance, and Minkowski distance are several popular distance metrics used with kNN. kNN finds the  $k$  points with smallest distances to the test point  $X$  and assumes

$$P(Y = 1|X) = \sum_{i=1}^k w_{(i)} y_{(i)}, \quad (9)$$

where  $y_{(i)}$  is the label of  $x_{(i)}$ , the  $i^{\text{th}}$  closest point to  $X$ , and  $w_{(i)}$  is the weight associated with  $y_{(i)}$ .  $w_{(i)}$  is a decreasing function of  $d(X, x_{(i)})$ . For example, we may assume

$$w_{(i)} \propto \frac{1}{d(X, x_{(i)})}, \text{ or } w_{(i)} \propto \frac{1}{d(X, x_{(i)})^2}, \text{ or } w_{(i)} \propto e^{-d(X, x_{(i)})}.$$

Note the exact values of  $w_{(i)}$ 's can be found by enforcing  $\sum_{i=1}^k w_{(i)} = 1$ .

#### 2.2.7 Random Forests

Suppose we start with a covariate  $i$ , and use the decision rule  $x_i \leq T_i$  for some threshold  $T_i$  to partition training data to two groups. The covariate  $i$  can be chosen so that we get the best partitioning of data in terms of prediction accuracy. In other words, we use the covariate that when considered individually, gives us the best rule to assign a label to each test point. Afterwards, we focus on one of the two groups, say  $x_i > T_i$ , pretend this subset of data is our entire dataset, and find covariate  $j$  that when considered individually gives us the best rule to predict the label of a test point in this group. The process continues until a satisfactory rule is obtained. This gives us a decision tree.

A random forest (RF) is a collection of decision trees where each tree is built using  $s$  randomly selected subset of data. RF merges the results of all trees and assumes

$$P(Y = 1|X) = \sum_i w_i \hat{y}_i, \quad (10)$$

where  $\hat{y}_i$  and  $w_i$  are the predicted label and weight of tree  $i$ , respectively. RFs are described in more detail in [11, 13]. Note that RF can be used for TESA by letting the dependent variable be with time to failure value,  $t$ , instead of its indicator of being more than  $T_f$  i.e.,  $y$ .

### 3 Variable Selection

Variable selection is of paramount importance in many practical applications, and can tremendously help to avoid the curse of dimensionality. When dealing with many high-dimensional “omics” data, the large number of potential covariates can be problematic. Not only in many cases  $p > n$ , immediately creating an ill-posed problem, but also many potential covariates may not

contribute to the biological event under study, e.g., cancer relapse. Therefore, variable selection is initially performed to reduce the dimensionality of the problem. Note here we use variable and covariate interchangeably. A cornucopia of algorithms are proposed for this task; however, many can be categorized into one of the following groups: (1) performing a first phase of variable selection prior to classifier design, (2) using embedded methods, such as sparsity inducing penalties, that perform variable selection and classifier design in a single step, and (3) using wrappers that aim to select a subset of variables that yield a high accuracy estimate.

There are many issues associated with wrappers, particularly, their extensive computation cost demands search heuristics such as sequential forward search (SFS) and sequential float forward search (SFFS), and difficulties with arriving at reliable error estimates has limited their use in practice [10, 18, 23, 24]. On the other hand, first phase filtration techniques and embedded methods are typically much faster than wrappers [10], and tend to result in classifiers that enjoy higher accuracies. Here we focus on these two strategies. These variable selection methods are discussed in more detail in [6, 9, 10, 21, 26].

Embedded methods perform variable selection and classifier design in one step. They typically add penalties to the classification objective, so that classifiers which use a larger number of covariates are more penalized, i.e., less desirable. Thereby, a variable is only introduced to the model only if it improves the classification objective to a larger extent than the penalty it incurs. LASSO ( $L_1$ ), elastic net, and slab-and-spike priors are three penalties used for inducing sparsity to the classifier. LASSO and elastic net penalties typically do not tremendously increase the computation cost of the classifier design stage, while slab-and-spike priors are typically extremely computationally intensive as they require Monte Carlo Markov chain (MCMC) methods, such as Gibbs sampling, for model selection. Here we focus on LASSO and elastic net penalties. Finally, note RFs are also embedded methods as covariates which provide little information about the class labels result in low accuracies, and are not used in any of the decision trees.

Finally, one may directly perform a first phase filtration to prune the set of potential covariates. Given train data in both classes, the goal is to find the subset of variables with distributional differences across the two classes. Note the goal of many variable selection algorithms is to flag features with extreme distributional differences, which is basically the goal of differential expression analysis. Note the set of features with distributional differences does not necessarily correspond to the set of variables that result in the best classification performance for the data at hand, i.e., the actual prediction error of the trained concomitant classifier (not the theoretical Bayes error of the selected variable set) might not be minimized for the classifier using *all* variables with distributional

differences [14, 23, 24]. However, a first phase of dimensionality reduction is a very practical heuristic, resulting in classification rules with reliable predictions.

### 3.1 Feature Selection Algorithms

Here we describe several popular algorithms used for variable selection. All algorithms discussed here, except the very last one (called POFAC), are univariate filters, i.e., assess each variable individually, and do not take pairwise interactions into account. On the other hand, POFAC takes advantage of pairwise feature dependencies.

#### 3.1.1 t-test

*t*-test and its variants are one of the most popular hypothesis tests [9]. The goal of two sample *t*-test is to determine if two populations, say relapsing versus non-relapsing cancer patients, have similar means using the collected sample. It is a powerful tool to grasp differences in the means, but may fail to grasp weak markers comprised of several subpopulations [8, 10]. In particular, it is wise to evaluate higher order differences, say differences in variances, when dealing with heterogeneous data. *t*-test is a frequentist test and assigns *p*-values to each feature, where features with the smallest *p*-values are then selected. While student *t*-test assumes both populations have similar variances, Welch *t*-test does not make such assumption.

#### 3.1.2 Bhattacharyya Distance

Bhattacharyya distance (BD) evaluates how similar/different two distributions are. Assuming data in both classes is Gaussian, and pretending sample mean and variance in class  $y$ , denoted by  $\hat{\mu}_y$  and  $\hat{\sigma}_y^2$ , respectively, are the true distribution parameters, the following formula computes the distance between the two classes:

$$\text{BD} = 0.25 \log \left( 0.25 \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} + \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} + 2 \right) \right) + 0.25 \left( \frac{(\hat{\mu}_0 - \hat{\mu}_1)^2}{\hat{\sigma}_0^2 + \hat{\sigma}_1^2} \right). \quad (11)$$

The larger is the BD, the larger is the difference between the class-conditioned distributions. In other words, features with large BD are more differentially expressed, which are then selected for classification.

#### 3.1.3 Mutual Information

Mutual Information (MI) is widely used for variable selection, which aims to identify the amount of “information” each variable  $X_i$  carries about class labels  $\mathcal{Y}$ , and is denoted by  $I(X_i, \mathcal{Y})$ . Variables with highest MI are selected. To be precise, MI computes the expected number of bits we save when storing  $X_i$  by knowing  $\mathcal{Y}$ . MI is zero if and only if  $X_i$  and  $\mathcal{Y}$  are independent, and is strictly positive otherwise. It is easy and straightforward to compute MI given a training sample  $S$  for discrete variables, for instance when single nucleotide polymorphisms (SNPs) are reported in dosage. However, when working with continuous data, such as expression levels, numerical challenges arise. Several non-parametric methods

are discussed in [1], and [15] studies a parametric approach for Gaussian features. The method of [15] is rather computationally intensive and has been suggested to perform similar to BD [10]. Here, in our simulations, we focus on the non-parametric  $m$ -spacing approach of [1] with  $m=1$ .

### 3.1.4 Wilcoxon Rank Sum Test

Wilcoxon rank sum (WRS) test is a non-parametric hypothesis test that, heuristically speaking, evaluates if two variables have similar medians in both classes. More precisely, the hypothesis of WRS test is that given two randomly drawn points from the two distributions, the point from the first distribution is equally likely to be less or greater than the second point. Given the training sample, WRS (1) combines points in both classes, (2) sorts them in ascending order, (3) associates each point with their adjusted rank, being the unadjusted rank for unique values and the median of unadjusted ranks for repeated values, and (4) computes

$$U_y = R_y - \frac{n_y(n_y + 1)}{2}, \quad (12)$$

for  $y=0, 1$ , where  $R_y$  is the sum of ranked points in class  $y$ . Note  $U_0 + U_1 = n_0 n_1$ .  $U = \min \{U_0, U_1\}$  is the  $U$ -statistic, and is asymptotically normally distributed with mean  $0.5 n_0 n_1$  and variance  $n_0 n_1 (n_0 + n_1 + 1)/12$ , which can be used to compute the  $p$ -values. Finally, variables with smallest  $p$ -values are selected.

### 3.1.5 Kolmogorov–Smirnov Test

Kolmogorov–Smirnov (KS) test is another non-parametric test that can detect any mode of distributional difference between two classes. Given data, KS test (1) computes the empirical CDFs of the two classes, and (2) computes  $D^*$ , the largest difference between the two CDFs in absolute values. The asymptotic distribution of  $D^*$ , how to numerically compute p-values, and issues with current approximations in common software such as MATLAB, R, C, Fortran, and JAVA are discussed in [25]. Note that as KS is a non-parametric test, its null is rather large, resulting in larger p-values compared with parametric tests.

### 3.1.6 Optimal Bayesian Filtering

Optimal Bayesian Filtering (OBF) is a recently proposed method for variable selection and is extremely powerful. It can integrate prior biological knowledge, handle multiple classes, integrate data from different sources, and enjoys robust performance [4, 5, 7, 10]. Here we discuss a special case of OBF using Jeffreys non-informative prior. In this case, OBF ranks variables by  $s(f)$  for each variable (feature)  $f$  where variables with larger  $s(f)$  rank higher.

$$s(f) = \frac{(\hat{\sigma}_f^2)^{0.5(n_0+n_1)}}{(\hat{\sigma}_{f,0}^2)^{0.5n_0}(\hat{\sigma}_{f,1}^2)^{0.5n_1}}, \quad (13)$$

where  $\hat{\sigma}_f^2$  and  $\hat{\sigma}_{f,y}^2$  are the sample variance in both classes and in class  $y$  of variable  $f$ ; respectively. Recall,  $n_y$  is sample size in class  $y$ .

### 3.1.7 Posterior Factor-Constrained

All methods studied so far are univariate filters, meaning they assess each feature individually and rank them according to a score function. POsterior FActor-Constrained (POFAC) takes pairwise feature dependencies into account to improve its variable ranking [9]. Variables that rank high by POFAC are either individually strong, or have significant pairwise interactions with other variables. Under Jeffreys prior, POFAC ranks features by

$$\tilde{\beta}(f) = \sum_{f' \neq f} \frac{p(\{f, f'\})}{p(\{f'\})}, \quad (14)$$

where for each variable set  $A$ ,

$$p(A) = \frac{|\hat{\Sigma}_A|^{0.5(n_0+n_1)}}{|\hat{\Sigma}_{A,0}|^{0.5n_0}|\hat{\Sigma}_{A,1}|^{0.5n_1}}, \quad (15)$$

and  $\hat{\Sigma}_A$  and  $\hat{\Sigma}_{A,y}$  are sample covariance in both classes and in class  $y$ , respectively. POFAC is studied in more detail in [9].

## 3.2 Variable Selection, Biological Relevance, Prediction Accuracy, and False Discovery Rates

Indeed interpretability and reliability is crucial in analyzing biological data, and is tremendously emphasized in the literature [12, 27]. Only models that use markers with potential biological relevance and enjoy high prediction accuracies are of interest. To that end, methods that bound the false discovery rate (FDR) have become extremely popular by providing some assurance that the set of variables used are not extremely polluted with false alarms. Although it is typical to bound FDR by 5%, there is no guarantee such set results in the best prediction! How can we know whether it is the set that bounds FDR by 1% that results in highest accuracy, or the set bounding FDR by 5%, or 10%? Additionally, many methods that bound FDR only provide a universal bound, and the true ratio of false discoveries, known as the false discovery proportion (FDP), can deviate from the desired bound  $\alpha$  to a large extent. Hence, it might be wise to explore other approaches as well; in particular select the top  $K$  features and let  $K$  to be a parameter tuned during validation. Depending on the value of  $K$  and the number of variables selected by bounding FDR, which we hereafter denote by  $R$ , the following cases can be considered:

1.  $K < R$ : In this case classification requires a smaller feature set than all significant variables for reliable prediction. Since we pick the features ranking highest, this means classification imposes more conservative constraints than FDR correction, and only chooses a subset of significant variables.
2.  $K > R$ , and  $R$  is very small: In this case FDR correction may be rather conservative, outputting few features that may not be enough to explain the variance of the data. If the prediction rule is using many covariates and still suffers low accuracy, it might be the case that some important factors are overlooked and not measured in the dataset; otherwise, it may be the case that there are so many weak covariates that we cannot detect them unless we accept rather large FDPs.
3.  $K > R$ , but  $R$  is not small: In this case, many, but not all, covariates may be weak, meaning many covariates should be considered together to have a reliable prediction rule. This is further discussed below.

When FDR correction does not output a very small set of covariates, but reliable prediction is only achieved by considering much larger sets, it might be that there are many weak covariates as well as strong ones. For example, suppose we start with 10 reliable markers, but do not obtain high prediction accuracies. Suppose we are given the option to add 100 additional covariates to the model, of which only 50 are true markers. If the information of the true markers in this set is larger than the noise introduced by false alarms, the prediction accuracy can increase; however, the covariates used in the prediction model suffer large FDPs, hindering reliable biological interpretation. In these situations it might be wise to consider two modes of analysis: discovery and prediction. In the discovery mode we only report the few reliable markers, but state many important weak markers can be missed. In the prediction mode we use the large number of covariates to arrive at a reliable prediction rule; but state that although we are confident the predictions are accurate, we cannot reliably justify which patterns in the data brought us to this conclusion. Note these concerns apply to both TESA and BCSA.

## 4 How to Handle Random Censoring?

Up to this point we assumed there is no random censoring, which may almost never happen in any real-world experiment. Here we briefly discuss how these issues should be handled in BCSA.

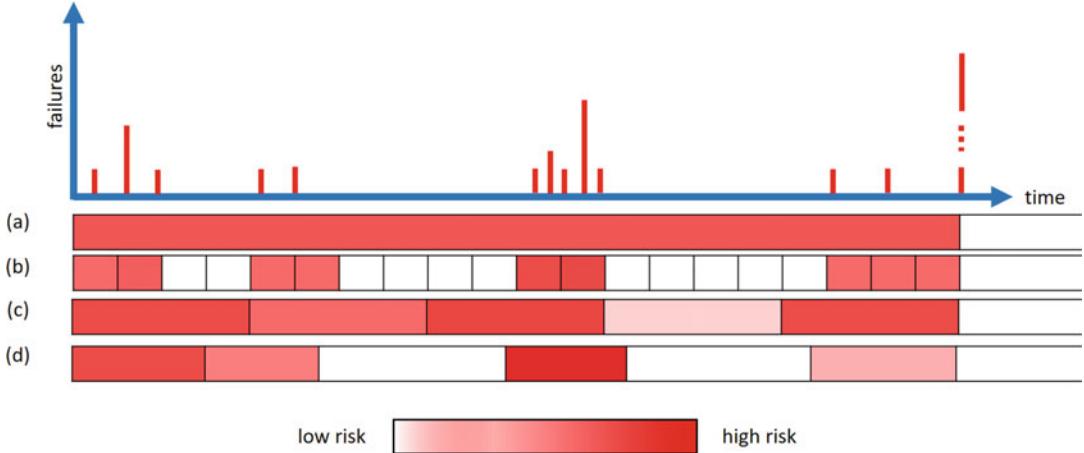
We cannot assign randomly censored points to a class, and hence they cannot be used for estimating the classifier parameters. However, the classification rule obtained using non-randomly censored points can be used to further assign a label to randomly

censored portion of the data, and further study those points in more detail. On one hand, TESA can take advantage of such points in its parameter estimation while BCSA cannot; however, the assumption violations of the TESA model might affect its performance to a large extent. Therefore, the question is if the information provided in the randomly censored points increases prediction accuracy of the TESA model more than the reduction incurred by model assumption violations. **If randomly censored points comprise a small portion of data it is extremely likely that BCSA results in more accurate predictions than TESA.** On the other hand, if a large portion of data is randomly censored, then TESA might have the advantage. However, it is not always easy to verify if we have passed the point where TESA assumption violations have a lower impact on prediction than ignoring randomly censored points for parameter estimation. Therefore, it is prudent to always check both approaches, TESA and BCSA, to be sure the right model is used.

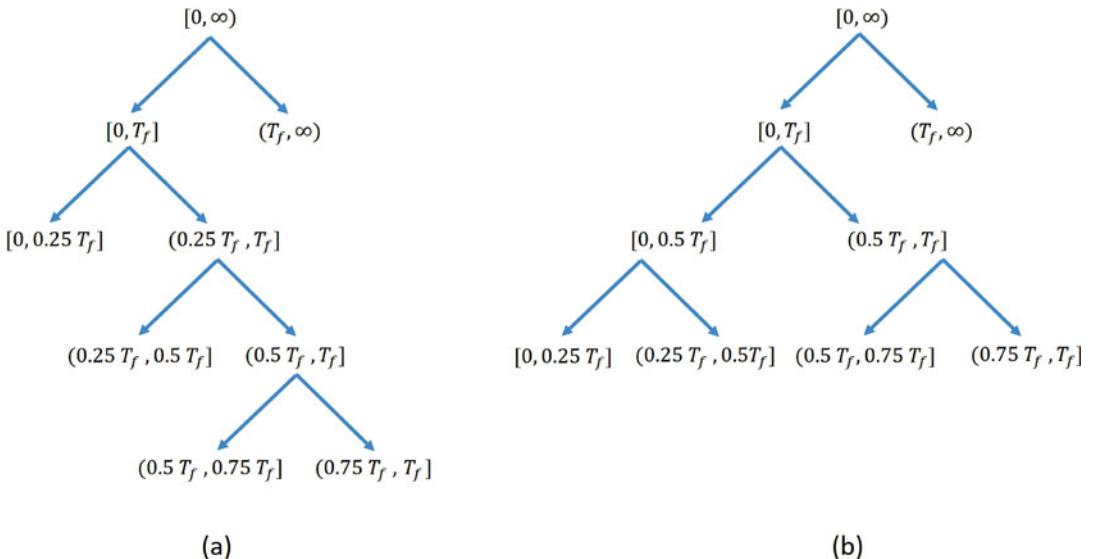
## 5 Multiple Time Intervals

TESA models can be further used to predict failure probabilities in any time interval, and do not require having the intervals fixed a priori; however, classification based methods need to be inputted with the collection of such time intervals beforehand. From this perspective TESA is much more appealing; however, if such time intervals are already predetermined, it is wise to explore *multi-class classification for survival analysis (MCSA)* methods as well, which are many times formulated as a sequence of BCSA problems. Here we provide two examples to show how sequences of BCSA models can be used to analyze risks of multiple intervals. Note that in general we may take each time unit as a potential time interval of MCSA, and hence obtain models that are almost as flexible as TESA based approaches.

First and foremost, before solving an MCSA model, irrespective of framing it as a single multiclass problem or a sequence of binary class problems, one must verify if the desired time intervals contain enough data points for reliable training. For instance, if the time intervals are too small, one may observe time intervals in which no failures occur while adjacent time intervals contain many failures. Figure 1 provides an illustration of an appropriate selection of time intervals for MCSA, and examples of inadequate intervals. Figure 1a denotes the BCSA model we have discussed here, where the goal is only to predict if failure occurs up to time  $T_f$  and Fig. 1b illustrates the case where the time intervals are too small, resulting in intervals with no failures while adjacent to high risk intervals. The time intervals in Fig. 1c, d are adequate length-wise; however, in Fig. 1c they are not appropriately chosen, as the



**Fig. 1** An example of time intervals that can be used with MCSA: (a) the BCSA time interval, (b) small time intervals resulting too much resolution for the amount of available data, particularly for the third and fourth time intervals, (c) an example of time intervals that are appropriate in length but are not well-placed, resulting in intervals that lie in both high and low risk regions, and (d) an example of appropriate time intervals in length and placement



**Fig. 2** Different time interval structures for a given set of potential time intervals: (a) a sequential approach, and (b) a bisection approach

middle intervals with many failures are broken between truly high and low risk times. Finally, Fig. 1d denotes an appropriate choice of time intervals.

Finally, note by considering different sequences of BCSA problems for the same set of given time intervals we can encode different “risk structures” in the model. Figure 2 provides two

examples how different time intervals can be concatenated. In both cases we first aim to predict if the failure is likely up to time  $T_f$ . In Fig. 2a we start with the first time interval and sequentially aim to determine if (1) the current interval is the high risk one, or (2) the high risk time interval occurs in the future. More details about such formulations can be found in [16]. Similarly, instead of going forwards in time, a backward model can be considered as well. Alternatively, in Fig. 2b we consider a bisection approach.

## 6 Synthetic Simulations

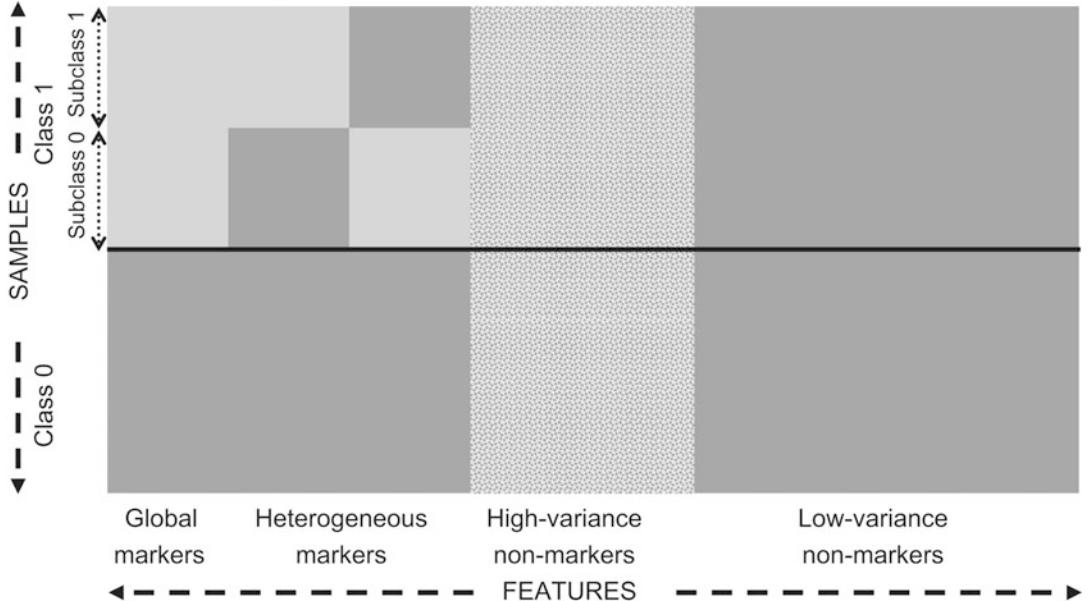
Here we perform a large family of simulations to identify (1) settings where TESA and BCSA perform best, and (2) how the different classification and variable selection algorithms stack up against each other. In particular, we consider settings where assumptions of the Cox proportional hazard model are satisfied, and settings where they are violated.

### 6.1 Data Generation

Here we use a synthetic model developed to mimic microarrays to generate data, originally introduced in [14] and extended in [6, 9]. While the model was originally proposed to mimic microarrays, it is also suitable for mimicking gene expressions as well. In this model features are either markers or non-markers. Markers construct blocks of size  $k$ , where markers in the same block are dependent, and markers in different blocks are independent of each other. Markers are either global or heterogeneous. Global markers are homogeneous within each class, while heterogeneous markers construct  $c$  subclasses in class 1, where for each block one of the subclasses follows a distribution different than class 0, and all other subclasses follow a distribution similar to class 0.

Each block of global markers is Gaussian in each class with mean  $\mu_y$  and covariance  $\sigma_y^2 \Sigma_y$ .  $\mu_0 = [0, \dots, 0]$  and  $\mu_1$  can take different values in class 1.  $\mu_1 = [1, 1/2, \dots, 1/k]$  for synergetic markers and  $\mu_1 = [1, 0, \dots, 0]$  for marginal markers. Diagonal elements of  $\Sigma_y$  are 1, and off-diagonal elements are  $\rho_y$ .  $\sigma_y^2$  and  $\rho_y$  are parameters of the model. Note the original model of [14] assumes  $\rho_0 = \rho_1$ , while they may take different values in the extensions of [6, 9]. Each block of heterogeneous markers in class 0 is similar to class 0 global markers, and the subclass in class 1 with a distribution different than class 0 is similar to class 1 markers.

Non-markers have similar distributions in both classes, and are either low variance or high variance. In the original model of [14] low variance non-markers are independent, but are similar to class 0 markers in the extensions of [6, 9]. Note we use the dependent low variance non-marker model of [9] here. High variance non-markers are independent and follow a Gaussian mixture distribution:  $wN(0, \sigma_0^2) + (1 - w)N(1, \sigma_1^2)$ , where  $w$  is drawn randomly



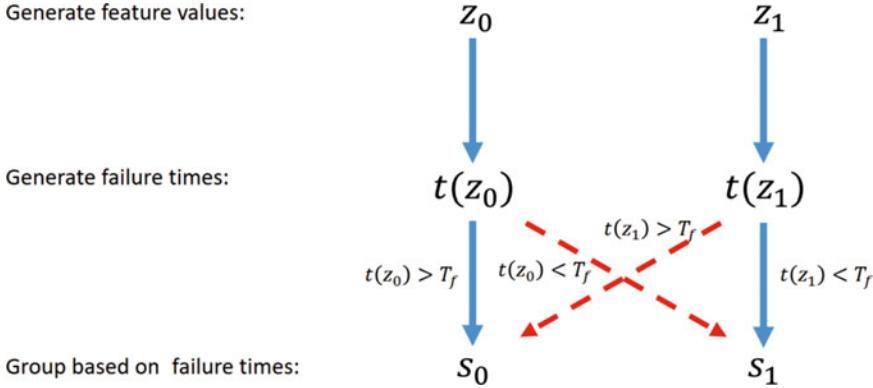
**Fig. 3** Schematic of markers and non-markers for synthetic data model adopted from [14]

from  $[0, 1]$ . Figure 3, adopted from [14], depicts how markers and non-markers look like across the two classes.

Here we use this synthetic model to generate points in two classes, denoted by  $Z_0$  and  $Z_1$ , with an equal number of points in each class, which denote the “truly” low/high risk patients, respectively. Afterwards, given expression values, we randomly generate time to failures. We assume time to failure in class  $y$ , denoted by  $t$ , follows a Weibull distribution with shape parameter  $\kappa_y$  and scale parameter  $\lambda(z)$ . Note the shape parameter only depends on the class label, while the scale parameter is a function of the observed covariates, i.e., expression levels. Note the distribution of  $z$  depends on the class label, and hence,  $\lambda$  is also affected indirectly by the class labels through  $z$ . Throughout we assume

$$\lambda(z) = e^{\beta_0 + \sum_{i \in M} \beta_i z(i)}, \quad (16)$$

where  $\beta_i$ 's are drawn independently from an exponential distribution with rate  $r$ ,  $M$  is the set of markers, and  $z(i)$  is the  $i^{\text{th}}$  marker value for observation  $z$ . Given  $t$  and follow-up time  $T_f$ , a point for which  $t > T_f$  belongs to class 0, and points for which  $t < T_f$  comprise class 1 for BCSA. In other words,  $S_y$ , train data in class  $y$  for BCSA can be written as follows:



**Fig. 4** Schematic of the data generation process. Given a truly low risk point  $z_0$  and truly high risk point  $z_1$ , we generate failure times  $t(z_0)$  and  $t(z_1)$ . We then group points with failure times less than  $T_f$  together and points with failure times larger than  $T_f$  together. Points in  $Z_0$  end up in  $S_0$  with a high probability, and a few of them are assigned to  $S_1$ . Similarly, most high risk points  $Z_1$  end up in  $S_1$ , with few points having failure times larger than  $T_f$ . Finally, failure times larger than  $T_f$  are replaced with  $T_f$ .

$$S_0 = \{z_0 \in Z_0 : t(z_0) > T_f\} \cap \{z_1 \in Z_1 : t(z_1) > T_f\},$$

$$S_1 = \{z_0 \in Z_0 : t(z_0) < T_f\} \cap \{z_1 \in Z_1 : t(z_1) < T_f\}.$$

Note  $Z_0/Z_1$  may denote points that are truly low/high risk, and  $S_0/S_1$  denote the potentially low/high risk observed points. In other words, a small portion of truly low risk points may have failure times less than  $T_f$ , and hence be associated with high risk points, and vice versa. Therefore,  $S_y$  is mostly comprised of points in  $Z_y$ , with few points which truly belong to the other class, i.e., each class is polluted with a small subpopulation of points from the other class. Note we set  $r$  in each simulation so that about 10–25% of  $Z_0/Z_1$  points belong to  $S_1/S_0$ . Figure 4 depicts the general implemented pipeline.

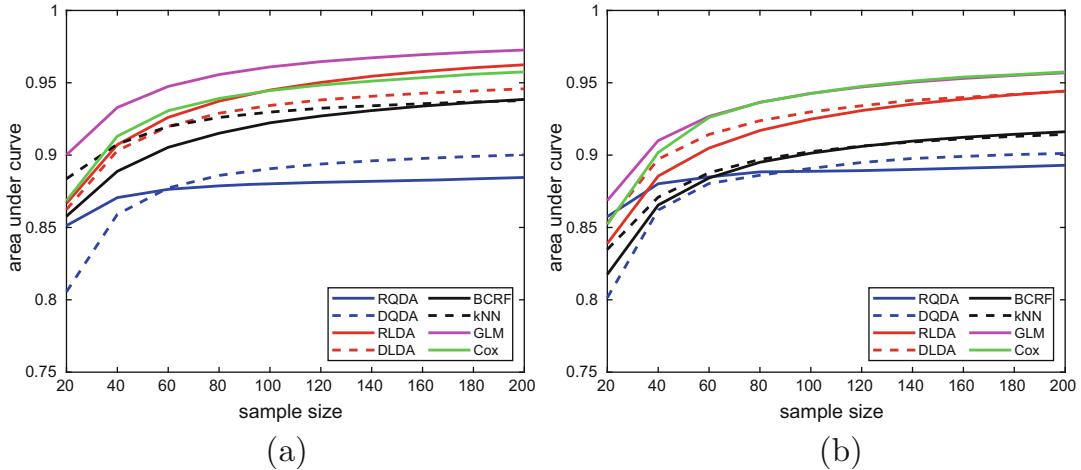
## 6.2 Prediction Evaluation

Here we perform two simulation families to compare binary classification and survival analysis methods. In the first simulation we consider a data generation model with proportional hazards that violates the assumption of Cox model that all points have similar survivorship functions. In particular, we use different values for  $\kappa_0$  and  $\kappa_1$ . In the second family of simulations we set  $\kappa_0 = \kappa_1$  so that the assumptions of the Cox model are satisfied.

Here we consider simulations where  $\kappa_0 \neq \kappa_1$ . In particular,  $\kappa_0 = 0.5$  and  $\kappa_1 = 1.5$ . We assume there are 100 markers, where 20 are global and 80 are heterogeneous,  $k=5$ , and  $c=2$ . We set  $\sigma_0 = 0.5$  and  $\sigma_1 = 0.8$ , corresponding to the “large and unequal” variance model of [14]. We consider synergetic and marginal markers, and consider two settings for  $\rho_y$ :  $\rho_0 = \rho_1 = 0.8$ , and  $\rho_0 = 0.8$  and  $\rho_1 = 0.1$ . We generate a train data, as well as a test data

**Table 1**  
**Failure time parameters of implemented data models**

Data model	Shape model	$r^{-1}$	$\beta_0$	$T_f$	$P(t(z_0) < T_f)$	$P(t(z_1) > T_f)$
Synergetic, equal correlation	$\kappa_0 \neq \kappa_1$	0.18	5	3.8	0.1471	0.1002
Synergetic, unequal correlation	$\kappa_0 \neq \kappa_1$	0.2	6	4	0.1477	0.1374
Marginal, equal correlation	$\kappa_0 \neq \kappa_1$	0.5	4	3	0.3602	0.3632
Marginal, unequal correlation	$\kappa_0 \neq \kappa_1$	1	8	2	0.297	0.3412
Synergetic, equal correlation	$\kappa_0 = \kappa_1$	0.25	6	5.6	0.1081	0.1159
Synergetic, unequal correlation	$\kappa_0 = \kappa_1$	0.2	6	4.3	0.1303	0.0665
Marginal, equal correlation	$\kappa_0 = \kappa_1$	0.35	6	10	0.3639	0.3031
Marginal, unequal correlation	$\kappa_0 = \kappa_1$	0.25	4	8	0.3152	0.3285



**Fig. 5** Average area under the curve over all 4 data generation models versus sample size for (a) unequal  $\kappa$  and (b) equal  $\kappa$  time to event models

comprised of 2000 points, comprised of 1000  $Z_0$  and 1000  $Z_1$  points. Table 1 lists the failure time parameter used for data generation.

Given data, all classifiers of Subheading 2.2 are trained, and are evaluated on the test data. In addition to the classification rules we implement several variants of the Cox model as well as a random forest for time to event analysis. To avoid an ill-posed problem due to the large number of covariates and small sample size we use LASSO, elastic net, and ridge penalties for regularization. Figure 5a plots the area under curve (AUC) of each prediction rule versus sample size for all algorithms as sample size increases from 20 to 200 in steps of 20 averaging over 200 iterations. When  $\kappa_0 \neq \kappa_1$ , BCSA methods tend to enjoy a higher AUC. In particular,

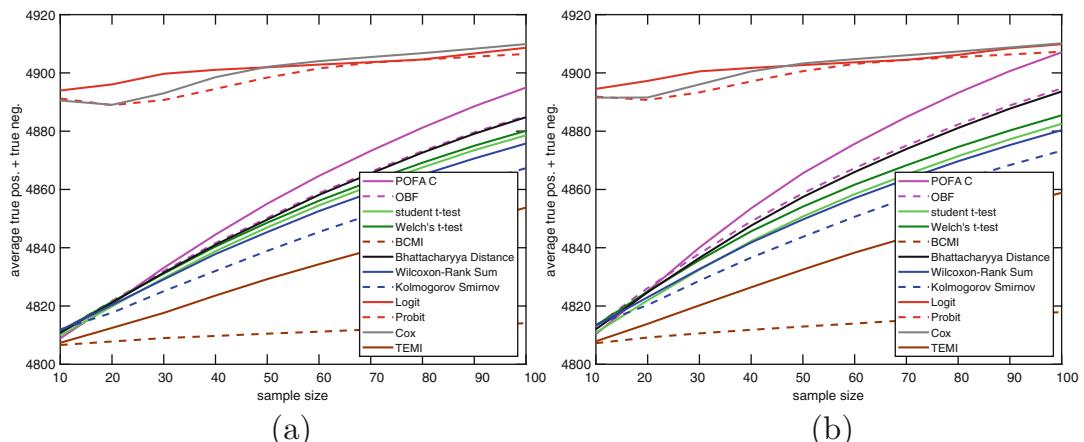
GLMs outperform other BCSA methods. We observed SARF barely performed superior than a random decision rule, and hence was omitted from the figures. GLM is followed by Cox and RLDA, which perform similarly on average. Although the assumptions of Cox are violated, it is still a contender, and enjoy superior performance compared with many other BCSA methods. Cox and RLDA are closely followed by DLDA, kNN, and BCRF. RQDA and DQDA perform inferior to all other methods; however, we observed numerical issues resulted in underestimating their AUCs for larger sample sizes. Therefore, we expect their true AUC to be higher than depicted in Fig. 5a.

We now consider simulations similar to those of Fig. 5a, except that we set  $\kappa_0 = \kappa_1 = 1$ . Figure 5b plots the AUC versus sample size averaging over all 4 data models and 200 iterations. The ranking among the different methods is to a large extent preserved compared with Fig. 5a; that is, GLM and Cox perform best, are closely followed by RLDA, DLDA, RQDA and DQDA. However, at this time DLDA slightly outperforms RLDA.

### 6.3 Variable Selection

Here we assess how feature selection algorithms of Subheading 3.1 perform. Note Cox variants and GLMS with LASSO and elastic-net penalties are embedded methods, performing variable selection in conjunction with classification. Therefore, they are included in this simulation. A mutual information based variable selection model for continuous dependent variables is proposed in [19], hereafter called (MI-C), which we include as a benchmark for variable selection using continuous time to event observations.

Assume there are 5000 features where 100 are markers, 20 global, and 80 heterogeneous. Fix  $k=5$ ,  $c=2$ , and assume



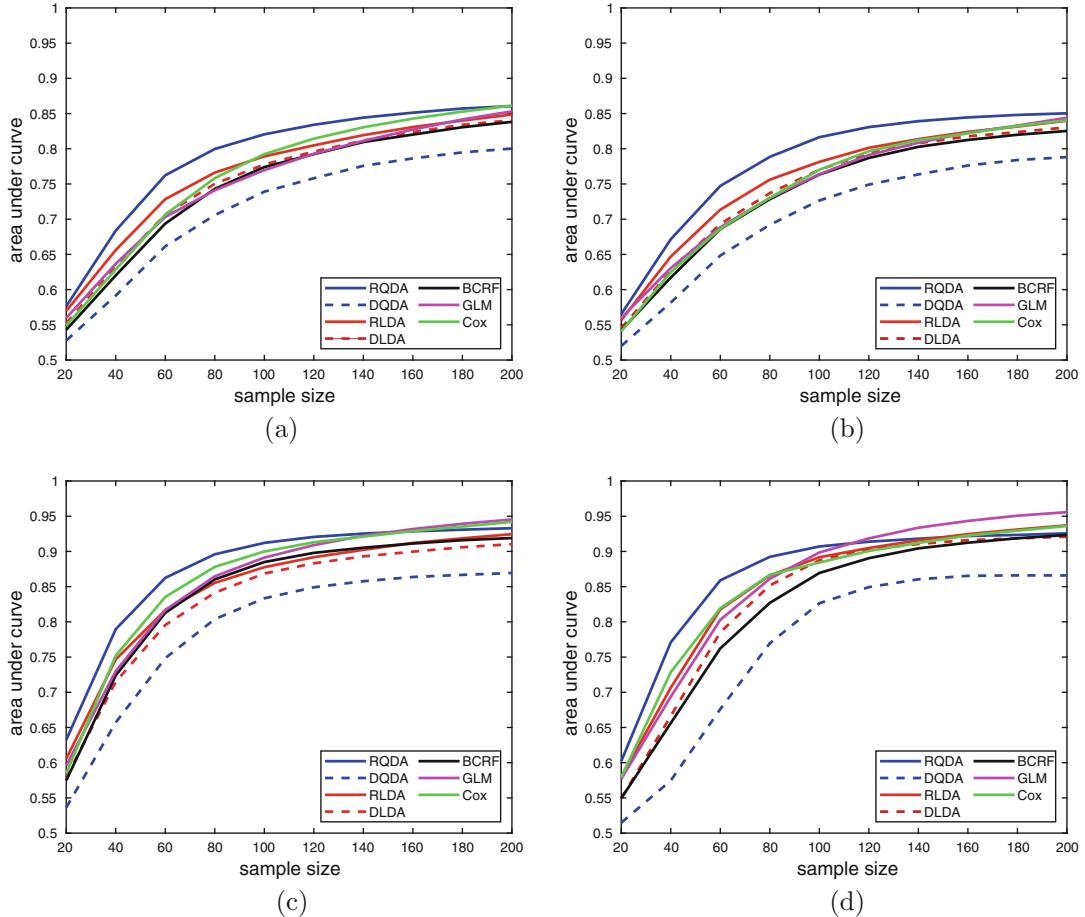
**Fig. 6** Average sum of true positives and true negatives of various variable selection algorithms versus sample size averaged over all 4 data generation models for (a) equal  $\kappa$  and (b) unequal  $\kappa$  time to event models

there are 2000 high variance non-markers. Figure 6 plots the average number of true positive and true negatives as sample size increases from 20 to 200 in steps of 20 averaging over 500 iterations. Two general patterns are observable: (1) variable selection methods tend to outperform embedded methods, (2) binary class selection methods outperform those that use continuous time to event values, and (3) mutual information arrives at better feature rankings when inputted with time to event values rather than class labels; however, it still performs inferior to other selection algorithms. Among binary class feature selection algorithms POFAC, OBF, and BD outperform others. Note that the average large number correctly labeled features (sum of true positives and true negatives) for penalized regression methods such as logit, probit, and Cox, is due to their ability to output a variable number of features, while other methods are forced to output exactly 100 features. Note not outputting any features as markers results in correctly labeling all 4900 non-markers, and the tendency of penalized regression methods to correctly label 4900 features for small samples is undesirable. On the other hand, for an algorithm that is forced to output 100 features, 50 features need to be true markers so that 4900 features are correctly labeled. Indeed a power of 50% is satisfactory for small samples.

#### **6.4 Prediction in Presence of Selection**

The prediction simulations of Subheading 6.2 are inputted with a set of markers, and in Subheading 6.3 we only focused on the variable selection performance. However, in practice, both steps are necessary. Given a large set of covariates, variable selection algorithms are used to reduce the dimensionality of the problem to improve reliability, stability, and prediction accuracy. Afterwards, prediction methods are trained on the filtered covariates. Here we consider this more realistic setting.

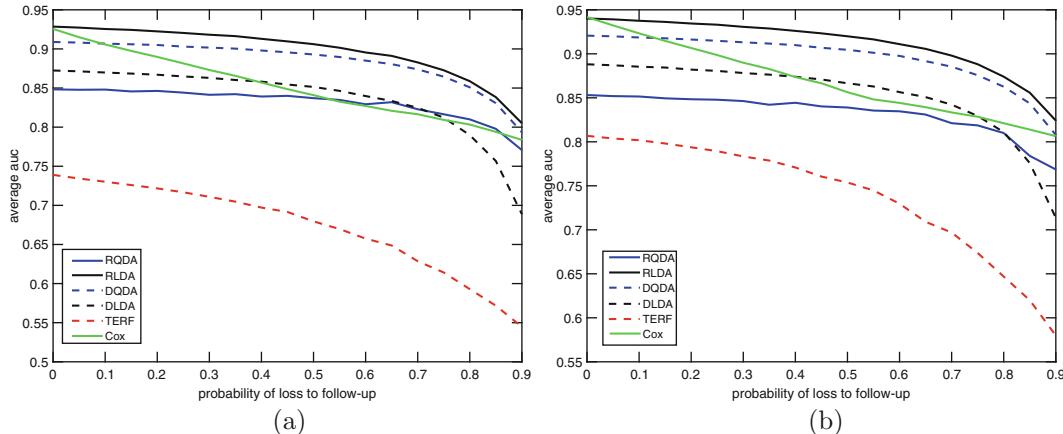
We use the data generation model of Subheading 6.3 and the survival model of Subheading 6.2 to generate the time to event values. The selected features of each variable selection algorithm of Subheading 6.3 can be used with the BCSA methods of Subheading 6.2. This creates a combinatorial issue with the total number of feature selection and classifiers to test for. Therefore, we restrict ourselves to the top performing binary class variable selection algorithms, namely BD, OBF, and POFAC. We also forgo the kNN and BCRF to reduce the computation cost. Figure 7 plots AUC of each classifier, versus sample size averaging over 200 iterations for 4 of the data generation models (two equal  $\kappa$  and two unequal  $\kappa$  models) to present the general trends observed across all simulations. For each BCSA method we report the highest AUC obtained by the tested variable selection algorithms for each sample size. This time we observe that RQDA enjoy the highest AUCs, and followed by GLM and Cox. In particular, when  $\kappa$ 's are not equal, RQDA is the clear choice.



**Fig. 7** Area under the curve of various BCSA and TESA methods versus sample size for (a) unequal  $\kappa$  and synergistic markers with equal correlations, (b) unequal  $\kappa$  and synergistic markers with unequal correlations, (c) equal  $\kappa$  and synergistic markers with equal correlations, and (d) equal  $\kappa$  and synergistic markers with unequal correlations

### 6.5 The Effect of Random Censorship

So far we have assumed all points have been followed up to time  $T_f$  and no random censorship has occurred, which is typically not the case in many real-world studies. Given our simulation observations, it is expected that when only a small portion of data has been randomly censored, not including them in the classifier design might still result in a higher accuracy than a survival model whose assumptions are violated or needs to deal with a large number of covariates in spite of being able to take advantage of the randomly censored points. On the other hand, if many points are randomly censored, it is definitely expected of survival models to outperform classification rules. An important question to answer is when this switch occurs, i.e., as the probability of random censorship increases, at which value does TESA begin to outperform BCSA. In other words, at which stage does considering randomly censored



**Fig. 8** Average area under the curve of various BCSA and TESA methods versus probability of loss to follow-up averaged over all 4 data generation models for (a) equal  $\kappa$  and (b) unequal  $\kappa$  time to event models

data improves prediction to a larger extent than TESA assumption violations reduce it. This is a very difficult question to answer, and depends on the true underlying parameters. Here we perform several simulations to get a better understanding of this pattern though.

We consider a simulation with different shape parameters ( $\kappa_0 = 0.5$  and  $\kappa_1 = 1.5$ ),  $n_0 = n_1 = 50$ , and simulate for random censorship. After the data is generated, we randomly draw from a Bernoulli random variable with probability  $p$  of being 1 for each training point, which is the indicator of random censorship. Time of random censorship for such points is drawn from a uniform distribution over  $[0, T_f]$ . Figure 8 plots the AUCs of different algorithms as probability of random censorship,  $p$ , increases, averaging over 100 iterations. Note (1) for small  $p$  classification methods perform best while survival models are appropriate for large  $p$ 's, and (2) the cross over between classification rules and survival models depends on the true distribution parameters and may happen for  $p$ 's as low as 30% or as high as 90%. Therefore, it might be wise to implement both methodologies in practical settings and pick the most competent model. Finally, note when the assumptions of the survival model, say Cox proportional hazard model, are satisfied, and there are not too many covariates in the model, survival analysis is the clear choice as it is almost guaranteed to achieve high accuracies on top of all of its appealing by-products, such as probability of failure estimates beyond  $T_f$ .

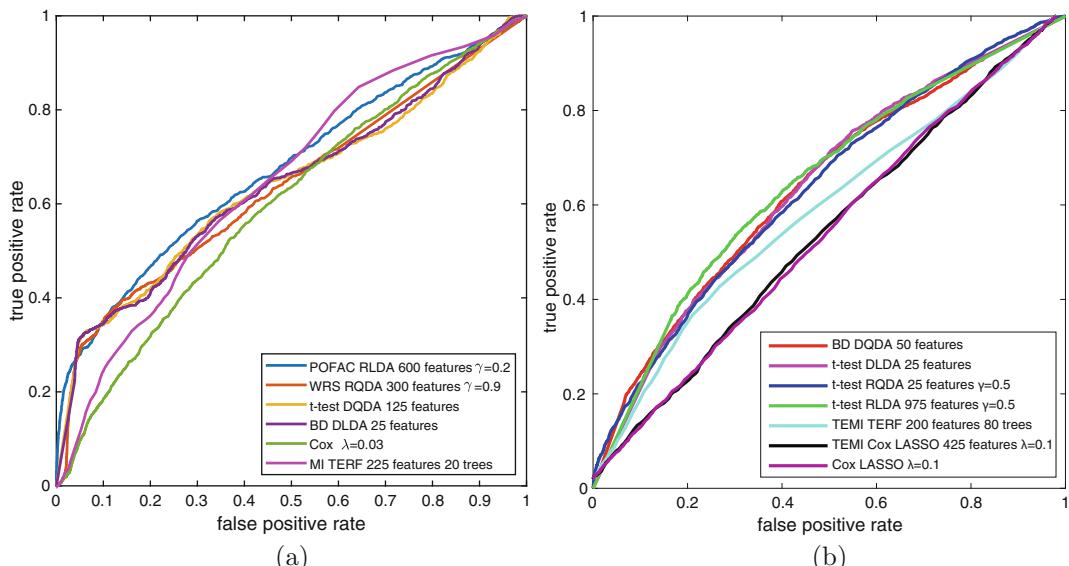
## 7 Experimental Data Analysis

Here we study two real cancer datasets to see how the different algorithms perform in more realistic settings. We analyze breast cancer RNAseq data obtained from the cancer genome atlas

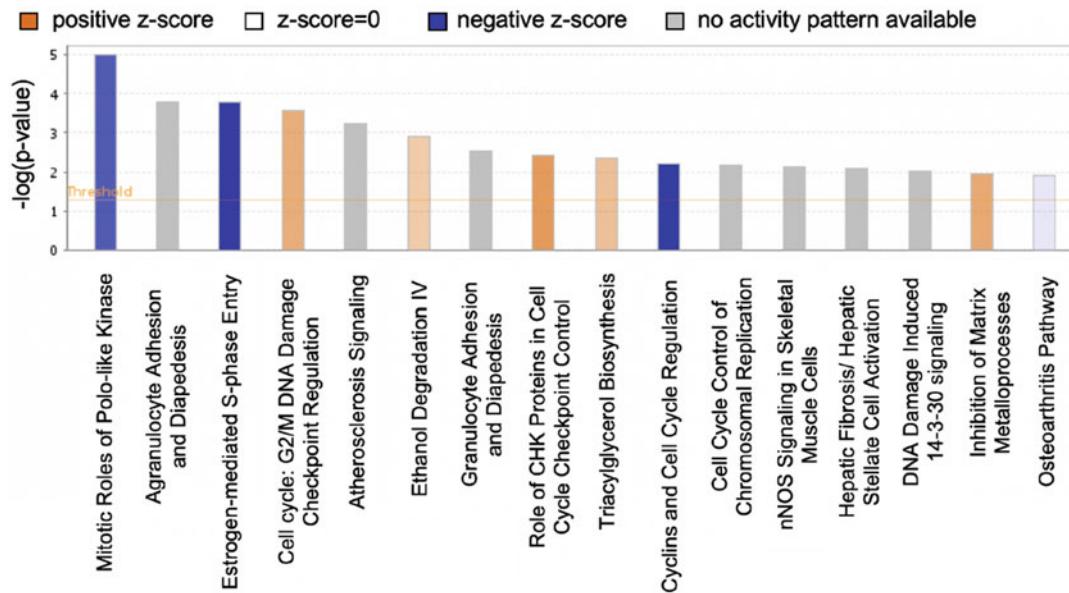
## 7.1 Breast Cancer

(TCGA), as well as a microarray lung cancer dataset, suggesting these methods are applicable to a wide range of cancer types and **Han et al.** give the details of breast cancer patients and their time to relapse values, setting  $T_f = 7$  years, we partition them to 94 relapsing and 663 non-relapsing patients. We aim to identify markers that affect the risk of cancer relapse, and develop models that assign a relapse risk within the  $[0, T_f]$  interval to each patient. Note that if a smaller interval is of interest, say  $T_d < T_f$ , patients who relapsed/did not relapse up to time  $T_d$  comprise the classes, and the rest of the BCSA pipeline remains intact. Here, as follow-up time was extensive, we chose 7 years so a reasonable number of cancer patients have relapsed up to  $T_f$  and to have extra assurance that if a patient had a high chance of relapse, then they are most probably assigned to class 1.

We randomly partition points and generate a stratified sample comprised of 78 relapsing (high risk) and 551 non-relapsing (low risk) patients, and use the remaining points for testing. Afterwards, variable selection, classification, and survival models are used to analyze the data and label test points as low risk or high risk. We perform this hold out process 100 times. Figure 9 plots the AUC of the methods that performed best. Note POFAC-RLDA using 600 genes and  $\gamma = 2$  performs superior to all other algorithms. For small false positive rates WRS-RQDA, *t*-test-DQDA, and BD-DLDA enjoyed the highest true positives rates. The Cox model using elastic net for regularization ( $\alpha = 0.5$ ) performed best with a small penalty ( $\lambda = 0.03$ ). However, it still performed inferior to BCSA models. While MI-TERF performed inferior to



**Fig. 9** Receiver operator characteristic (ROC) curve of the top performing methods for the (a) breast cancer and (b) lung cancer datasets



**Fig. 10** Top enriched pathways of the breast cancer dataset using the POFAC gene set

BCSA models, it enjoyed the highest true positive rate for larger false positive rates. However, this performance gain might not have much practical value, as large positive rates are usually not acceptable in practice.

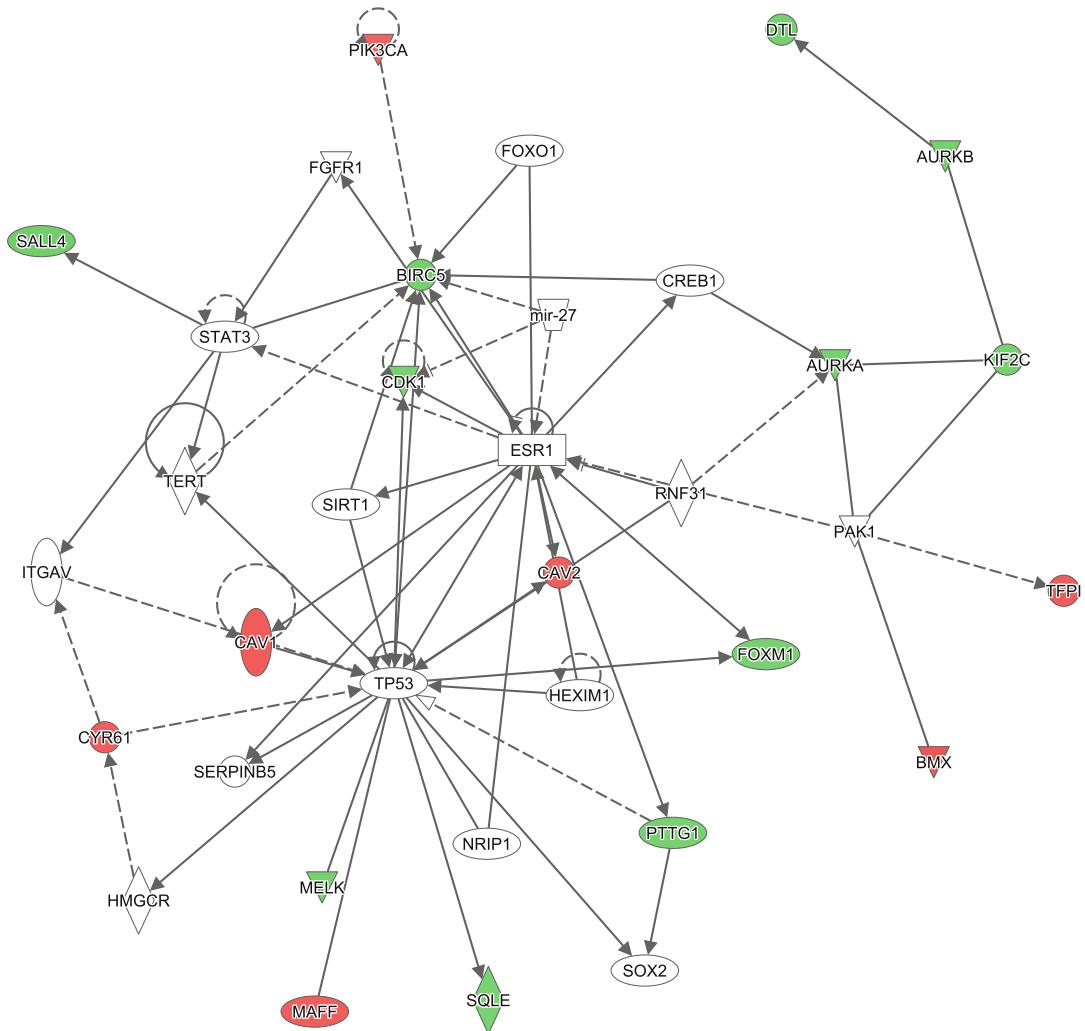
We now study if the genes selected by this are biologically relevant. Note AUC is not a desirable validation metric, and for discovery applications other validation metrics, such as false discovery rate (FDR) are more appropriate. Here we further analyze the selected genes not to argue the ability of the pipeline for biomarker discovery applications, but rather, to emphasize the selected genes are biologically relevant. Note for the variable selection methods of Subheading 6.3 can be used without the BCSA rules of Subheading 6.2. to discover differentially expressed markers.

We use Ingenuity Pathway Analysis (IPA) [17]<sup>1</sup> to find the top pathways and gene networks that are suggested to be affected given the selected variables. The top enriched pathways are provided in Fig. 10, of which many are suggested to be affected by breast cancer in the literature. The top constructed networks map to cancer, cellular development, cellular growth, cellular function, and proliferation functions. A top network is provided in Fig. 11 as an example.

## 7.2 Lung Cancer

Data obtained in [22] is deposited on gene expression omnibus (GEO) with accession number GSE68465. The data is based on the

<sup>1</sup> QIAGEN Inc., <https://www.qiagenbio-informatics.com/products/ingenuity-pathway-analysis>.

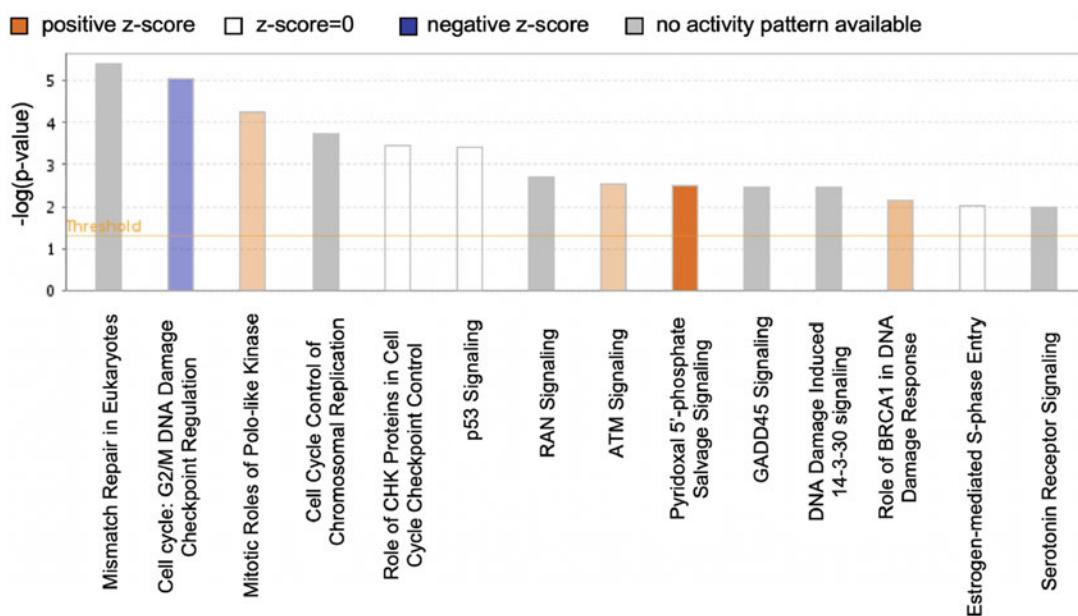


**Fig. 11** A top IPA gene network of the breast cancer dataset using POFAC genes

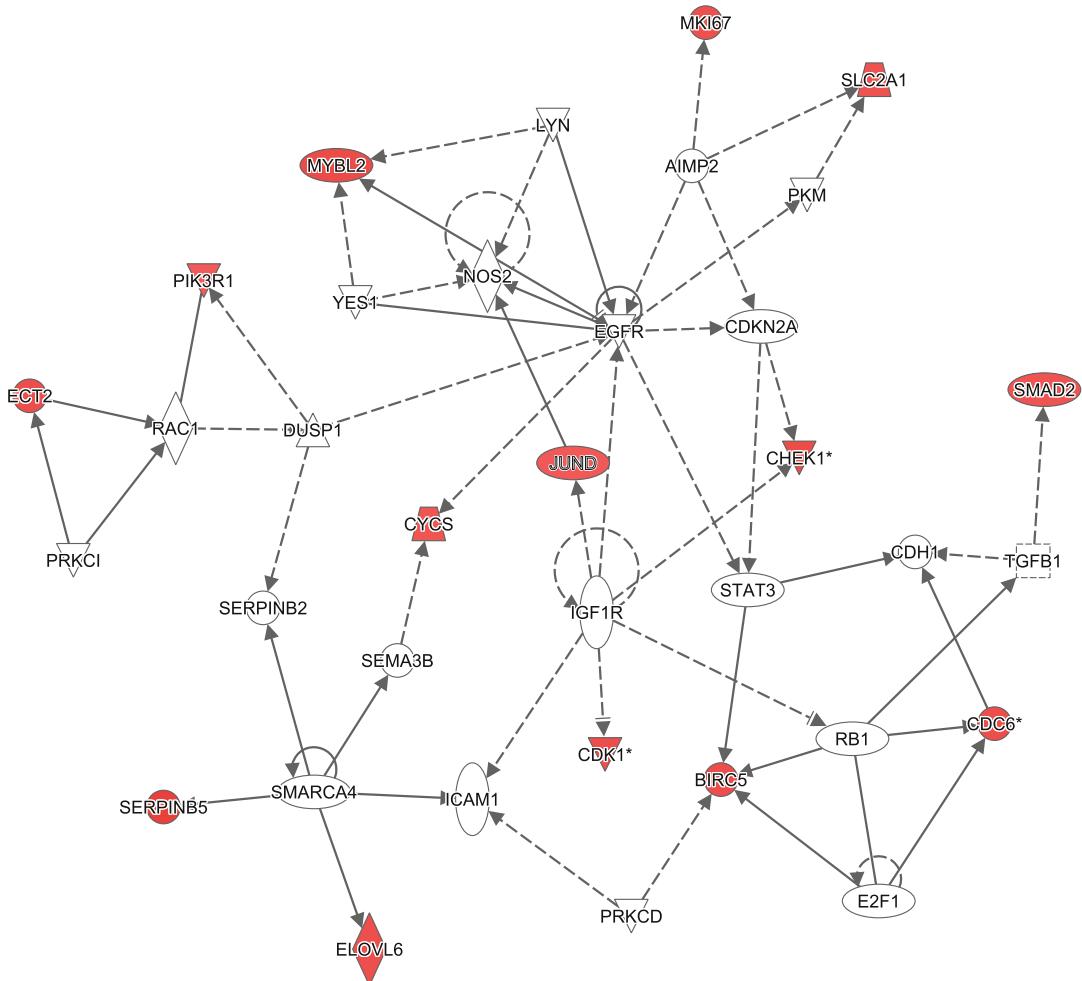
GPL96 platform, which uses the Affymetrix human genome U133A array. 224 patients who relapsed or died within  $T_f=6$  years of follow-up comprise class 1, and 238 patients who survived for this time period comprise class 0. While many patients are followed up for more than 6 years, [22] suggests the 6 year window is preferred for labeling patients as high risk (relapse or death) or low risk (no relapse or death). We input TESA with the entire follow-up results, as we believe it provides TESA more information to potentially utilize without making the model any more complex. We randomly select 204 high risk patients and 207 low risk patients for training, and use the remaining hold out points for testing. We iterate 100 times. Figure 9 plots the ROC curve of the different BCSA and TESA methods resulting in the largest AUCs. We again observe that BCSA based methods enjoy higher AUCs. This time  $t$ -

test with RLDA using 975 genes and  $\gamma = 0.5$  enjoys the highest AUC, although BD-DQDA using 50 genes and  $t$ -test RQDA using 25 genes enjoy higher true positive rates for false positive rates below 15%. TEMI TERF enjoys comparable performance with BCSA methods for true positive rates below 25%, it performs inferior for larger false positive rates. Interestingly, in this example, the Cox proportional hazard model performed much inferior than all other tested methods. Since TEMI TERF performed relatively well for small false positive rates, it might be the case that Cox's model assumption violations may be the driving force in exacerbating its performance.

We again study the top 975  $t$ -test genes obtained by considering the full dataset. Note bounding FDR by 5% using the Benjamini–Hochberg step up procedure [2], the top 712 genes are significant, suggesting RLDA is using a gene set with an FDR less than 5% for its prediction. Recall a deeper discussion on scenarios where the number of features used for prediction is smaller than the FDR feature set is provided in Subheading 3.2. The top IPA pathways and its top gene network are provided in Figs. 12 and 13, respectively. Note this top gene network maps to cancer, cell death and survival, and organismal injury and abnormalities functions.



**Fig. 12** Top enriched pathways of the lung cancer dataset using the  $t$ -test gene set



**Fig. 13** A top IPA gene network of the lung cancer dataset using *t*-test genes

## 8 Conclusion

Time to event analysis is tremendously powerful, yields satisfactory performance across a wide range of applications, and is already implemented in many software packages. While time to event analysis is a great hammer, not all problems are nails! There are certain classes of real-world scenarios where binary classification methods provide survival analysis results with higher accuracies. Therefore, it is important to check that assumptions made in the survival analysis model are not heavily violated and the model adequately fits the data. Also, it is prudent to test other analysis methods to provide assurance the appropriate model is used. In particular, in applications where (1) randomly censored points comprise a small portion of data, (2) there can potentially be many weak covariates that need

to be considered together for a reliable analysis, and (3) assumptions of the time to event model can be extremely violated, for instance when studying complex diseases, it is wise to explore binary classification for survival analysis (BCSA) methods as well. Additionally, when discovering new potential markers is of interest, binary class variable selection models seem to have an advantage. Although their prediction accuracies might not be much higher than TESA, they typically enjoy a higher probability of detection in identifying biological markers. Finally, note that in presence of few randomly censored points, one can perform BCSA on the non-censored points, treat the censored points as test data, assign them a label, and redo BCSA pretending there is no random censorship; so to incorporate the few randomly censored points in the analysis.

## References

1. Beirlant J, Dudewicz EJ, Györfi L, Van der Meulen EC (1997) Nonparametric entropy estimation: an overview. *Int J Math Stat Sci* 6 (1):17–39
2. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol)* 57:289–300
3. Bishop CM (2006) Pattern recognition and machine learning, 1st edn. Springer, New York
4. Foroughi pour A, Dalton LA (2017) Multi-class Bayesian feature selection. In: IEEE global conference on signal and information processing (GlobalSIP). IEEE, Piscataway, pp 725–729
5. Foroughi pour A, Dalton LA (2017) Integrating prior information with Bayesian feature selection. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics. ACM, New York, pp 610–610
6. Foroughi pour A, Dalton LA (2017) Robust feature selection for block covariance Bayesian models. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 2696–2700
7. Foroughi pour A, Dalton LA (2018) Bayesian feature selection with data integration. In: Proceedings of the 2018 IEEE global conference on signal and information processing (GlobalSIP), pp 504–508
8. Foroughi pour A, Dalton LA (2018) Biomarker discovery via optimal Bayesian feature filtering for structured multiclass data. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. ACM, New York, pp 331–340
9. Foroughi pour A, Dalton LA (2018) Heuristic algorithms for feature selection under Bayesian models with block-diagonal covariance structure. *BMC Bioinf* 19(3):R70
10. Foroughi pour A, Dalton LA (2018) Optimal Bayesian filtering for biomarker discovery: performance and robustness. *IEEE/ACM Trans Comput Biol Bioinf* 17(1):250–263
11. Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. *Math Intell* 27(2):83–85
12. Haury A-C, Gestraud P, Vert J-P (2011) The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS One* 6(12):e28210
13. Ho TK (2002) A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Anal Appl* 5(2):102–112
14. Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognit* 42(3):409–424
15. Huber MF, Bailey T, Durrant-Whyte H, Hanebeck UD (2008) On entropy approximation for Gaussian mixture random vectors. In: Proceedings of the 2008 IEEE international conference on multisensor fusion and integration for intelligent systems, pp 181–188
16. Kleinbaum DG, Klein M (2010) Survival analysis, 3rd edn. Springer, New York
17. Krämer A, Green J, Pollard J Jr, Tugendreich S (2013) Causal analysis approaches in ingenuity

- pathway analysis. *Bioinformatics* 30 (4):523–530
18. Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20 (15):2429–2437
19. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf* 7 (1):S7
20. McCullagh and John A. Nelder (1989) Generalized Linear Models, Second Edition, Volume 37 of Chapman & Hall/CRC Monographs on Statistics & Applied Probability, London, UK
21. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517
22. Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbins KK, Lively T, Jacobson JW, Beer DG (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 14 (8):822–827
23. Sima C, Dougherty ER (2006) What should be expected from feature selection in small-sample settings. *Bioinformatics* 22(19):2430–2436
24. Sima C, Dougherty ER (2008) The peaking phenomenon in the presence of feature-selection. *Pattern Recognit Lett* 29 (11):1667–1674
25. Simard R, L'Ecuyer P, et al (2011) Computing the two-sided Kolmogorov-Smirnov distribution. *J Stat Softw* 39(11):1–18
26. Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Academic, Boston
27. Wehenkel L, Geurts P, Huynh-Thu VA, Saeys Y (2012) Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics* 28 (13):1766–1774



# Chapter 7

## Challenges and Opportunities of Genomic Approaches in Therapeutics Development

Jaimie S. Gray and Moray J. Campbell

### Abstract

The magnitude of all therapeutic responses is significantly determined by genome structure, variation, and functional interactions. This determination occurs at many levels which are discussed in the current review. Well-established examples of structural variation between individuals are known to dictate an individual's response to numerous drugs, as clearly illustrated by warfarin. The exponential rate of genomic-based interrogation is coupled with an expanding repertoire of genomic technologies and applications. This is leading to an ever more sophisticated appreciation of how structural variation, regulation of transcription and genomic structure, both individually and collectively, define cell therapeutic responses.

**Key words** Pharmacogenomics, Genomewide association studies, Direct to consumer, Transcriptomic, Epigenetic

---

### 1 A Brief History of Genomic Approaches in Biomedical Research

Currently, biomedical research is accelerating rapidly largely as a result of an ever-expanding repertoire of next generation sequencing (NGS) technologies and applications. However, one of the earliest genomic technologies developed and applied widely was the genomewide measurement of gene expression levels using microarray approaches. These technologies have largely been superseded by NGS approaches but nonetheless illustrate many of the key opportunities and challenges of genomic approaches in biomedical research.

It is easy in the current environment of diverse and widespread NGS approaches to overlook how revolutionary the application of microarray technologies was, and how the development of this technology was vital to usher in widespread application of genomic and other high dimensional data approaches. Microarray commercialization was catalyzed by the early market leader Affymetrix in the early 1990s [1]. However, this widespread application built upon earlier technological developments including: the fluorescent

labeling of nucleic acids, the microscale printing of arrays, and the statistical approaches for the analyses of array signal [2]. This application also required the base experimental framework ideologies of community-developed standards for experimental genomic requirements [3], the ethos of publicly sharing data, and the methodological approaches comprising the development of the R framework for statistical computing [4] as a freely available community-supported platform with library packages implemented in Bioconductor [5].

Thus, the fusion of academic and commercial stakeholders was able to accelerate the widespread application of gene expression technologies across research in human and model organisms. This development was paralleled by development of similar array approaches to dissect genetic variation [2] and DNA CpG methylation [3].

Interestingly, the initial developments to enable NGS approaches were made also during the 1980s, but the technological and statistical hurdles combined with prohibitive costs in many ways allowed the halcyon days of array technologies to be sustained. Leroy Hood and colleagues first reported in 1986 the capacity to read fluorescent labeled nucleotide in an oligonucleotide [4]. Outside of large-scale genomic efforts, such as the Human Genome Project, the adoption of these early genomic approaches was severely restricted. Thus, although the Human Genome Project published a draft human genome in 2003, the widespread exploitation of this knowledge and the genomic approaches that drove it lagged behind. For example, the first publications exploiting RNA-Seq emerged in 2008 [5, 6] but then grew exponentially with over 7000 publications in PubMed in 2018 that directly list RNA-Seq in the Abstract. Coupled with the widespread application of NGS approaches has been a rapid development of the applications in terms of the measured nucleic acid composition, conformation, and biological context. This has been matched by a rapid development of sequencing technologies and scalability, which, in turn, facilitated capture of biological information hitherto considered to be unattainable [7, 8].

At the start of 2020, the number and diversity of studies exploiting NGS technologies applied across the biomedical research spectrum is arguably beyond the realistic hopes of most researchers only 20 years earlier. What has also been catalyzed by this development is the fusion of technologies to capture the interactions between different types of molecules in the cell. For example, the fusion of chromatin immunoprecipitation [9] with initial microarray approaches [10], and subsequently NGS [11] enabled the analyses of protein–DNA interactions across the genome. This has now been mirrored by a widespread number of applications, and leading to the exponential growth of such hybrid approaches and enabled high dimensional data approaches to interrogate the

highly dense and integrated experimental space of cell biology. This has then allowed biomedical researchers to study how therapeutic responses reflects the sum of genomic structure and variation, transcriptomic dynamics, diversity and interactions, protein genomic and transcriptomic interactions, and genomic and transcriptomic modifications. Combined, the application of these insights has led to the concept of pharmacogenomics in which all aspects of pharmacology are considered in the context of genomic function as a vital first step toward making precision medicine a reality. In the current review, the challenges and opportunities within each of these genomic arenas will be considered within the context of how they may impact therapeutic development. It is now reasonable to consider we are in the dawn of a new era biomedical research in which the field of genomics is beginning to transform the understanding of human health. However, it is perhaps unclear how the demands of this new era will impact the organization of research in academia and industry, and how researchers will adapt to the demands of approaching biological questions from a genomic perspective.

---

## 2 Genomic Arenas for Study in Human Health and Disease

### **2.1 Genomic Structure and Variation to Predict Therapeutic Responses**

Arguably, genetic variation was the earliest area of genomic information considered in the context of human disease and led to the application of genomewide association studies (GWAS). The current catalog of GWAS studies that pass the thresholds for design and statistical significance contains ~4200 publications that test the significance of ~150,000 genetic variation associations across a spectrum of human phenotypes and disease risks [12, 13]. Some of the earliest studies considered disease phenotypes, for example in the context of the mental illnesses including schizophrenia to understand potential disease drivers [14], as well as to consider how genetic variation in major therapeutic targets such as G protein-coupled receptors could impact therapeutic responses [15]. Allied with the application of this technology was a parallel development of the appropriate statistical and design frameworks for GWAS studies [16, 17].

From these studies, it is very clear that already there is either population and/or experimental level evidence for SNP associations with disease incidence and progression risks, as well as explaining and predicting drug responses in patients. However, while the number of potential SNP-drug relationships identified in the GWAS catalog numbers in the thousands, the actual number of FDA-supported SNP-drug relationships that are clinically actionable number in the low hundreds. That is, for example, there are clear clinical relationships for SNPs within genes in the cytochrome P450 drug-metabolizing enzymes that determine the effectiveness

of drugs such as Warfarin [18], as well as how variants in CYP2C19 are actionable in terms of metabolizing serotonin reuptake inhibitors [19]. It is likely that the number of these relationships identified will continue to grow and that many of these relationships will contribute to the understanding of how therapeutic exposures in target tissues can be modeled with pharmacokinetic/pharmacodynamic approaches.

The clinical translation of this information is impeded by a range of factors, including understanding the population specificities, and then how these relationships maybe further modified by either other aspects of a patient's gender and genetic variation, combined with other biological aspects including race, age, and obesity [20–22]. Furthermore, there are limitations in understanding exactly what these SNP phenotype associations biologically reflect. Various limiting factors have held up the development of understanding of how these genetic relationships mechanistically relate to a given phenotype. By definition, given that most of the genome does not encode for protein coding genes, then much of the germline or somatic variation occurs outside of these regions. Deciding how these variants relate to altered cellular function can be extremely challenging. For the researcher there has been a profound increase in access to well curated normal and disease cohorts in which to hypothesis test the functional impact of disease-associated variants. These resources include the large-scale pan-tissue Genotype-Tissue Expression (GTEx) project [23], where RNA-Seq has been undertaken in 50 different normal tissues, the Cancer Genome Atlas derived from multiplatform high dimensional data on 11,000 tumors representing 33 tumor types [24, 25] and the BRAIN Initiative Cell Census Consortium has undertaken large scale cell sorting combined with single cell RNA-Seq to examine specific cell-type expression patterns in the brain [26]. The potential impact of the Human Cell Atlas in this arena will no doubt be very significant [27].

Coupled with dissecting the functional relationships is the need to establish ever-larger patient cohorts for study that are combined with robust deep-phenotype data to test the significance of association outcome. One of the earliest and most powerful studies has been the Framingham Heart Study, which is a multigeneration cohort of ~15,000 individuals focused on, as the name implies, cardiac events. In recent years, samples from this cohort have undergone multiomic analyses and a remarkable series of publications and insights have emerged (reviewed in [28]). However, even these remarkable insights cannot universally be applied as the study population within the cohort does not capture all diversity. Furthermore, it is not clear that this study is actually large enough to capture rare genetic events. Therefore, there is a need to establish larger and more diverse cohorts around the world to understand how genetic variation impacts disease and therapy responses. For

example, the UK Biobank was formed to analyze the relationships between genetic variation and phenotypes and traits in 500,000 people [29]. Alternatively, other cohorts have been established focused on a single disease and ethnicity. For example, in 2018, the RESPOND cohort of 10,000 African American men was established to understand the germline and somatic genomic drivers of prostate cancer within this population with the goal to build these data into predictive polygenic risk scores of disease incidence and progression risks.

It is reasonable to anticipate that the analyses of these cohorts, combined with other commercial resources such as 23andMe [30] will profoundly impact the reporting of clinically significant germline variants and, in the case of cancer, somatic variants, that predict disease incidence and progression risks across the spectrum of human health. However, it is a much more challenging question to determine how quickly and widespread will be the adoption of this information in the clinical community. There are a number of very real limiting factors. Generally, the nuanced understanding of statistical genetics is outside the training experience of most physicians, and is even more removed from the population at large. To meet this knowledge and communication gap will require a large number of genetic counselors, of whom there is a significant shortage and ambiguities over precisely where their role should be in the clinical management spectrum [31–33]. Other issues arise from the use of direct to consumer testing (DTC) information in a clinical care setting; there have been reports of these companies having up to 40% false positives and attributing health risk to SNPs that have been designated benign by CLIA certified labs [34]. Currently, the precise approach by how the FDA approaches these technologies is an area of active examination, for example as to whether these should be termed “medical devices” which have less stricture on quality control and therefore lessened responsibility in reliability of the results [35]. Recently, these companies have been granted permission to perform BRCA mutation testing, unfortunately the mutations being tested are largely sequestered to an isolated ethnic population leaving the majority of consumers with a false sense of security on their BRCA status [36]. Given rapid changes in technology and testing of populations with low risk of BRCA mutations, there are concerns with false positives and issues of how patient’s genetic information may be utilized by commercial sources [32]. This leads back to the question of whether this information can be used to help and not hinder how medical professionals can apply this information to patients’ health care. The Genetic Information Nondiscrimination Act (GINA) of 2008 only covered the bare minimum of requirements, in both employment and health insurance, in order to make it feasible for consumers to utilize genetic testing. GINA, in fact, has only been amended once since its enactment which largely means that the

level of competency of the act falls short of potential areas for discrimination nearly 12 years later, and perhaps requires to be reviewed and updated to reflect this rapidly evolving arena.

Therefore, despite the acceleration of genomic technologies to decipher genetic variation applied to ever-larger cohorts of patients with deep phenotype data, there remains the simple but profound challenge of discerning how this information should best be assimilated and interpreted by the patient, clinical, research, and health care stakeholders.

## ***2.2 Transcriptomic Dynamics, Diversity, and Interactions in Therapeutic Responses***

Within the cell, the application of sequencing technologies has continued to reveal how signaling pathways are regulated by the genome, and how this is disrupted in various disease contexts. Cancer research has provided a clear focus for applying transcriptomic approaches given that cancer is often seen as a disease of rewired signaling events. Within cancer, one of the earliest transcriptional classifiers was established in dissecting breast cancer subtypes [37, 38]. Establishing transcriptional signatures within breast tumors that related to different phenotypes immediately established the possibility to use tumor biopsies to differential diagnosis based on gene expression patterns, and deliver more tailored therapeutic options.

One of the key findings of these analyses was to define breast tumor types by expression of the estrogen receptor alpha (ER $\alpha$ ) and relate this to the progression risks and therapeutic responses. Of course, an understanding of the relationship between sex steroids and the cancers of the reproductive system had emerged in the nineteenth century with the work of Sir George Beatson in Glasgow (UK) who began to define the relationship between the synthesis of estrogen and breast cancer risk [39]. These studies were echoed by the work of Dr. Charles Huggins and colleagues in the 1940s who established the endocrine synthesis of androgens and the relationship to prostate cancer. Subsequently, in 1972 analyses of the androgen receptor (AR) signaling system in the prostate began the field of programmed cell death [40]. However, the new genomic insights began to reveal in which tumor types these receptors could most accurately be targeted and what a significant therapeutic response may be.

Across breast, prostate and other cancers a considerable volume of preclinical and clinical studies now have examined how the transcriptome changes in response to various therapeutic approaches. Prostate and breast cancers provide clinically relevant contexts in which to dissect oncogenic transcription factor rewiring, which is heavily centered around the actions of the ER $\alpha$  and AR and their response to either activation by endogenous ligands to dissect their signaling pathways, or to synthetic antagonists to model therapeutic responses.

Prostate differentiation and homeostasis are intimately regulated by the AR. In prostate cancer the profile of AR regulated genes changes to promote growth [41–46], and dampen differentiation and apoptotic signals [47–55], but the mechanisms remain incompletely understood. Transcriptomic approaches, in the first case using microarrays and subsequently applying NGS approaches, have revealed a remarkable diversity of transcriptional responses and underscored that there are very few examples of obligate and universal transcriptional responses to these cancer-driver signals [45, 56, 57]. Of course, there are many reasons for these variations, from different experimental models and design, to an ever-increasing number of proposed mechanisms that suggest how these transcription factors function to regulate genes is extremely context-dependent and highly variable. Many investigators considered the impact of coregulators such as coactivators and corepressors which could change the transcriptional potential for a given transcription factor and indeed changes to their expression and function have altered the transcriptional responses to a range of nuclear receptors including the ER $\alpha$  and AR [58–65].

Genomewide discoveries by The Cancer Genome Atlas (TCGA) consortium [66, 67] and others have identified common changes in transcription factors and coregulators that interact with the AR [67–73]. However, establishing how these changes rewire the AR is extremely challenging given the multitude of transcription factor and coregulators (as much as 20% of the human genome encodes for transcription factor and coregulators), and the diversity of sites for their interactions across the cancer genome.

NGS studies have been applied to different aspects of the transcriptome including small noncoding RNA such as miRNA. In the past decade, roles for miRNA in prostate cancer alone has been examined in over 1000 publications [74] (reviewed in [75]). This enthusiasm derives from various attributes; their coverage [76] generates biostatistical advantages [77]; their serum stability allows reliable noninvasive [78, 79] measurements [80] using PCR technology [81] that overcomes the challenges of measuring other serum diagnostics, such as PSA in the case of prostate cancer [82]. Finally, circulating miRNAs originating from the prostate may encapsulate events within the tumor microenvironment [83–85], and in this way they can be exploited to predict disease progression risks [86–88] and possibly explain the health disparities that occur in prostate cancer [89–91]. In parallel, in breast cancer similar pathways and possibilities for the coregulation of miRNA have emerged [92–94] including to dissect the biological basis to cancer health disparities [95–97]. Other investigators have considered roles for long noncoding RNA [98, 99] and shown how these different aspects of transcriptional signals maybe coordinated [100–102] for example by acting as molecular sponges for protein coding RNA and miRNA [103–105]. Measuring, dissecting, and

modeling the coordination of all of these RNA species in the transcriptome is an ongoing challenge given that it appears there is considerable coordination between protein-coding and noncoding small and long RNA species, for example in feed forward loop structures [106–112].

The transcriptome can therefore reveal a significant prognostic signature of a disease state, as established nearly 20 years for breast cancer. It also reveals the remarkable diversity of the transcriptional events that occur within a cell, in terms of the magnitude and amplitude of transcription for a given gene in response to a drug, the specific type of transcript generated, being from a protein coding (and therefore also giving rise to splice variants) or noncoding gene, and how they are coordinated. The very dynamic range and diversity of these events that occurs, even within the same cell type receiving the same drug suggests there remains an incomplete understanding of the normal physiological process of transcription and the mechanisms whereby these processes are either disrupted or intrinsically vary from cell to cell. There are studies are being conducted that rely on the interplay between drugs and gene responses; researchers use this drug perturbed information to assess gene regulatory networks from library iterations of differential expressed genes and begin to understand which genes to target to be the most effective therapy [113]. Clearly, as studies move to single cell and small populations of purified cells answers over the spatial-temporal control of transcription will emerge [114, 115] and can be combined to help understand how disease states in a tissue arise and what are significant therapeutic responses [116].

### **2.3 Genomic and Transcriptomic Modifications**

In parallel, the distortion of transcription may arise in cancer through copy number variation and structural variations to transcription factors and coregulators [69, 117–121]. It is also clear that epigenetic reprogramming of access to enhancers and other regulatory regions occurs. Combined with these gene-specific changes can be larger alterations to three dimensional associations of chromatin the nucleus. For example, in recent years the structure of topologically associated domains (TADs) [122, 123] has been a subject of intense investigation suggesting these structures are a fundamental unit of genomic organization [124]. Alterations to TAD boundaries can profoundly change and further exacerbate distortions of transcriptome, which, in turn, can change cell sensitivity toward a given drug. For example, the understanding of mechanisms that rewire AR signaling in prostate cancer, or ER $\alpha$  in breast cancer has significant potential to distinguish indolent from aggressive early stage disease, and to help understand therapy-resistant late stage cancer. The goal of such approaches is to reduce the significant morbidity and mortality associated with these diseases.

Given these important relationships between genomic organization and coordinated gene regulation, it is not surprising that disruption to TAD structure is strongly implicated in a wide variety of diseases including in developing limb buds and leading to profound phenotypes in mice and humans [125–129]. Again in prostate cells, coordinated regulation of lncRNA and AR-regulated target genes is more likely to occur within the same TAD than within different TADs [130], supporting the idea that TADs are a fundamental organizational unit in the cell.

Enhancer reprogramming, or rewiring, has been identified as a cancer driver in a range of cancers [131–135]. One mechanism by which this occurs is through a process termed “enhancer hijacking” in which genomic disruptions to TAD boundaries allows previously insulated enhancers access to neighboring genes and changes in expression [136–141]. For example, deletion of the CTCF motif at a boundary can shift the TAD structure and change gene regulation as occurs at the HOX locus [142]. Disruption to TAD structure can lead to altered NOTCH signaling in breast cancer [143], and in the developing prostate key transcription factors have been shown to change their associations within a given TAD depending on what stage of development and differentiation the organ was undergoing [144]. In prostate cancer the frequent translocation of the ERG transcription factor with TMPRSS2 can lead to a large impact on the transcriptome which occurs in parallel to changes in chromatin organization [145–147]. Using the chromatin conformation capture technique Hi-C, which captures all genomic–genomic interactions, at least in prostate cell line models the erosion of CTCF-dependent TAD boundaries was revealed and led to an integrated understanding of how histone modifications and genomic structure may be jointly distorted as a disease driver, and in part explain the diversity of therapeutic responses.

There is considerable variation in the distribution and convergence of histone modifications across the genome which are mirrored by changes in DNA CpG methylation [148–150]. Again, these can be captured in ever-increasing detail and complexity by NGS approaches. Various workers have proposed that epigenomic events such as changes to histone modifications and DNA CpG methylation combine with reconfigured transcription factors and lead to phenotypic plasticity that ultimately drives disease, such as seen in the unrestrained growth and survival prostate cancer phenotypes. The scope for epigenomic changes to impact disease states and also to predict therapeutic responses has been intensively investigated. It is clear both in cancers, and other disease states, and even with healthy aging. Epigenetic age acceleration is measured using Horvath’s or Hannum’s clocks, and uses genomic methylation markers. These estimates accurately assess organismal stress and have been associated with cancer incidence and mortality, cardiovascular and neurologic diseases, and frailty [151–155]. One way

these physical and psychosocial stressors manifest in patients and survivors is by accelerating the patient's epigenetic age, as measured by DNA methylation sites across the genome, such that it is greater than the chronological age. Age acceleration has been associated with an increased risk of cardiovascular disease, all-cause mortality, and a variety of other health problems [154, 156–163].

However, there are divergent phenotypic consequences for many epigenetic modifications. For example, changes in epigenetic events including changes in DNA methylation at CpG islands can strongly correlate and functionally determine transcriptional responses and there is evidence that this contributes to the development of the aggressive prostate cancer phenotype arises and leads to therapy-resistant phenotype [164–168], and can be countered by treatment with epigenetic drugs [169]. It is reasoned that the mechanisms that control this epigenomic plasticity have the potential to be therapeutically significant [170, 171]. Most likely these changes determine the choice, location and impact of modified enhancers and the gene networks they control.

DNA methylation of CpG islands [172–176] is well studied in prostate cancer [177] and associated with recruitment of chromatin remodeling complexes, gene repression [59, 148, 178, 179], and worse disease outcome [180–182]. By contrast, far less is known about the cancer-related links of DNA methylation patterns in low- and moderate-density CpG regions often associated with enhancer and intergenic regions. Changes in CpG methylation levels at these regions is highly dynamic, critical in development [183, 184] and appear to promote and sustain cell differentiation [44, 185–190]. However, the precise links to the control of gene expression are only emerging.

Outside of CpG islands, in these so-called low-density CpG regions the presence of DNA methylation within transcription factor motifs can differentially impact how a given factor binds to the genome [191]. High-throughput approaches have shed light on the relationships between dynamic CpG methylation and transcription factor interactions by examining the CpG status of motifs for ~500 different TFs. This approach revealed that the DNA methylation state differentially impacted TF binding; ~40% showed decreased binding with methylation and ~60% actually increased binding [191]. These findings suggest that changes in DNA methylation in low-density regions may profoundly affect the combinations of transcription factors that bind a region and control the transcription rate of a promoter [191]. For example, RAR $\gamma$  [106] and ONECUT2 [192, 193] have opposite effects on driving prostate cell differentiation, and their genomic binding is actually antagonized by CpG methylation [191]. Therefore, the interplay between epigenomic events is highly textured depending upon the genomic location, and in the case of CpG methylation depend on the density of modifications.

Finally, the interplay of enhancers and genetic variation has been identified with integrative genomic approaches. Luca Mag-nani and coworkers [194] have mapped enhancers from primary breast cancer patient material and overlapped with GWAS identified SNPs enriched to identify those in enhancers as potential disease drivers. In prostate cancer, CTCF binding sites associated with outcome identified by GWAS have been the subject of CRISPR deletion which, in turn, leads to increased gene expression of genes within TAD loops [195].

---

### 3 Statistical and Computational Challenges in the Analyses of Genomic Approaches

Statistical and computational frameworks were essential for the capture, analyses, and interpretation of high dimensional data, which combines complex genomic analyses and there are several areas where variance can impact the data interpretation, from wet-lab experimental parameters to dry-lab analyses. An ongoing challenge is to harmonize these techniques and analytical pipelines. However, perhaps, the greatest challenge will be the data integration approaches, which will be designed, implemented, and interpreted by bioinformatically trained researchers.

It is approximately 50 years after conception of the field of bioinformatics (reviewed in [196]) and, in comparison to the 50-year widespread application of molecular biology [197–200], it is far from clear that bioinformatics is on a trajectory to also become democratized. Currently, there are significant educational and training challenges in how the biological research community can ensure that the next generation of researchers are genetically literate. This will require future workers in therapeutics to both conceive of experimental strategies in genomic terms and to be able to lead and dissect the analytical approaches. To date there have been a number of large-scale biological meta-projects such as The Human Genome Project [201], ENCODE [202, 203], and TCGA [66] which were achieved in large part through the success of the intensive data analytic initiatives within these projects. Unfortunately, at present, it is far from clear that the wider research community is able to assimilate and exploit these data, and these remarkable achievements and the organizations that drive them may prove to be the exceptions rather than the rule. This will profoundly impede how genomic approaches are implemented in therapeutic design, development, and testing.

---

### 4 Summary

The cellular and organism response to therapeutic agents is profoundly shaped by the genome through a wide variety of mechanisms. Genetic and other structural variations within genes can at a

minimum change the structure of the proteins they encode, and thereby alter the biochemical function. However, protein coding genetic variation is smaller than noncoding variation, and in these regions, many of which perform regulatory functions, the impact of variation is often more challenging to establish. There are multiple levels at which genomic approaches can dissect how cells respond to a therapy, from the regulation of coding and noncoding genes, to their interdependencies, and to the wide range of posttranslational modifications that occur on DNA, and more recently described on RNA [204]. This has led to a rapidly expanding genomic tapestry that encapsulates the potential of a cell to respond to a given therapeutic. These efforts are moving to ever-smaller numbers of more pure cell populations, and to apply genome editing approaches to functionally test genomic and epigenomic organization and structures. In turn, this understanding has the potential to predict responses in an individual and across populations. In turn, this insight offers much promise to tailor therapeutic approaches and to be a major economic catalyst. The high dimensional data approaches required to drive these discoveries require highly specialized training in bioinformatics, statistics, and genomics, supported by computational science skills. Many of these skills and insights remain in high demand and, in part, may be rate-limiting factors within academia and industry.

---

## Acknowledgments

*Funding:* This work was funded by the National Cancer Institute (NCI) grant awarded to the OSUCCC The James, CCGS P30CA016058.

*Declaration of Interest:* We declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

## References

1. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL (1993) Multiplexed biochemical assays with biological chips. *Nature* 364(6437):555–556
2. Gentalen E, Chee M (1999) A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays. *Nucleic Acids Res* 27(6):1485–1491
3. Eads CA, Danenberg KD, Kawakami K et al (2000) MethylLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res* 28(8):E32
4. Smith LM, Sanders JZ, Kaiser RJ et al (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321 (6071):674–679
5. Sultan M, Schulz MH, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321(5891):956–960
6. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18(9):1509–1517
7. Adams DR, Eng CM (2018) Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med* 379(14):1353–1362

8. Vora NL, Hui L (2018) Next-generation sequencing and prenatal 'omics: advanced diagnostics and new insights into human development. *Genet Med* 20(8):791–799
9. Orlando V, Paro R (1993) Mapping Polycomb-repressed domains in the bithorax complex using *in vivo* formaldehyde cross-linked chromatin. *Cell* 75(6):1187–1198
10. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85(1):1–15
11. Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4(8):651–657
12. MacArthur J, Bowler E, Cerezo M et al (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 45(D1): D896–D901
13. Welter D, MacArthur J, Morales J et al (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001–D1006
14. Maier W, Zobel A, Rietschel M (2003) Genetics of schizophrenia and affective disorders. *Pharmacopsychiatry* 36(Suppl 3):S195–S202
15. Sadee W, Hoeg E, Lucas J, Wang D (2001) Genetic variations in human G protein-coupled receptors: implications for drug therapy. *AAPS PharmSci* 3(3):E22
16. Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S (2005) Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet* 6(Suppl 1):S134
17. Saito A, Kamatani N (2002) Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. *J Hum Genet* 47(7):360–365
18. Cooper GM, Johnson JA, Langaele TY et al (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 112(4):1022–1027
19. Strawn JR, Powleit EA, Ramsey LB (2019) CYP2C19-guided escitalopram and sertraline dosing in pediatric patients: a pharmacokinetic modeling study. *J Child Adolesc Psychopharmacol* 29(5):340–347
20. Scherr CL, Ramesh S, Marshall-Fricker C, Perera MA (2019) A review of African Americans' beliefs and attitudes about genomic studies: opportunities for message design. *Front Genet* 10:548
21. Lauschke VM, Ingelman-Sundberg M (2019) Prediction of drug response and adverse drug reactions: from twin studies to next generation sequencing. *Eur J Pharm Sci* 130:65–77
22. Franks PW, Poveda A (2011) Gene-lifestyle and gene-pharmacotherapy interactions in obesity and its cardiovascular consequences. *Curr Vasc Pharmacol* 9(4):401–456
23. Mele M, Ferreira PG, Reverter F et al (2015) Human genomics. The human transcriptome across tissues and individuals. *Science* 348(6235):660–665
24. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216):1061–1068
25. Cerami E, Gao J, Dogrusoz U et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5):401–404
26. Ecker JR, Geschwind DH, Kriegstein AR et al (2017) The BRAIN initiative cell census consortium: lessons learned toward generating a comprehensive BRAIN cell atlas. *Neuron* 96(3):542–557
27. Regev A, Teichmann SA, Lander ES et al (2017) The Human Cell Atlas. *elife* 6
28. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS (2019) 70-year legacy of the Framingham Heart Study. *Nat Rev Cardiol* 16(11):687–698
29. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12(3):e1001779
30. Spector-Bagdady K (2016) "The Google of Healthcare": enabling the privatization of genetic bio/databanking. *Ann Epidemiol* 26(7):515–519
31. Krokosky A, Terry SF (2019) So what does that test result mean? Genetic counselors in the trenches. *Genet Test Mol Biomarkers* 23(8):507–508
32. Bamshad MJ, Magoulas PL, Dent KM (2018) Genetic counselors on the frontline of precision health. *Am J Med Genet C Semin Med Genet* 178(1):5–9
33. Haidle JL, Sternen DL, Dickerson JA et al (2017) Genetic counselors save costs across the genetic testing spectrum. *Am J Manag Care* 23(10 Spec No.):SP428–SP430
34. Tandy-Connor S, Guiltinan J, Krempely K et al (2018) False-positive results released by direct-to-consumer genetic tests highlight the

- importance of clinical confirmation testing for appropriate patient care. *Genet Med* 20 (12):1515–1521
35. Allyse MA, Robinson DH, Ferber MJ, Sharp RR (2018) Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing. *Mayo Clin Proc* 93 (1):113–120
  36. Gill J, Obley AJ, Prasad V (2018) Direct-to-consumer genetic testing: the implications of the US FDA's first marketing authorization for BRCA mutation testing. *JAMA* 319 (23):2377–2378
  37. Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747–752
  38. Perou CM, Jeffrey SS, van de Rijn M et al (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* 96 (16):9212–9217
  39. Paul J (1981) Sir George Beatson and the Royal Beatson Memorial Hospital. *Med Hist* 25(2):200–201
  40. Kerr JF, Wyllie AH, Currie AR (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer* 26(4):239–257
  41. Barfeld SJ, Urbanucci A, Itkonen HM et al (2017) c-Myc antagonises the transcriptional activity of the androgen receptor in prostate cancer affecting key gene networks. *EBioMedicine* 18:83–93
  42. Toropainen S, Niskanen EA, Malinen M, Sutinen P, Kaikkonen MU, Palvimo JJ (2016) Global analysis of transcription in castration-resistant prostate cancer cells uncovers active enhancers and direct androgen receptor targets. *Sci Rep* 6:33510
  43. Bernard D, Pourtier-Manzanedo A, Gil J, Beach DH (2003) Myc confers androgen-independent prostate cancer cell growth. *J Clin Invest* 112(11):1724–1731
  44. Thurman RE, Rynes E, Humbert R et al (2012) The accessible chromatin landscape of the human genome. *Nature* 489 (7414):75–82
  45. Wang Q, Li W, Zhang Y et al (2009) Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* 138(2):245–256
  46. Massie CE, Adryan B, Barbosa-Morais NL et al (2007) New androgen receptor genomic targets show an interaction with the ETS1 transcription factor. *EMBO Rep* 8 (9):871–878
  47. Copeland BT, Pal SK, Bolton EC, Jones JO (2018) The androgen receptor malignancy shift in prostate cancer. *Prostate* 78 (7):521–531
  48. Chattopadhyay I, Wang J, Qin M et al (2017) Src promotes castration-recurrent prostate cancer through androgen receptor-dependent canonical and non-canonical transcriptional signatures. *Oncotarget* 8(6):10324–10347
  49. Malinen M, Niskanen EA, Kaikkonen MU, Palvimo JJ (2017) Crosstalk between androgen and pro-inflammatory signaling remodels androgen receptor and NF-κappaB cistrome to reprogram the prostate cancer cell transcriptome. *Nucleic Acids Res* 45(2):619–630
  50. Olsen JR, Azeem W, Hellem MR et al (2016) Context dependent regulatory patterns of the androgen receptor and androgen receptor target genes. *BMC Cancer* 16:377
  51. Volante M, Tota D, Giorcelli J et al (2016) Androgen deprivation modulates gene expression profile along prostate cancer progression. *Hum Pathol* 56:81–88
  52. Stuchbery R, Macintyre G, Cmero M et al (2016) Reduction in expression of the benign AR transcriptome is a hallmark of localised prostate cancer progression. *Oncotarget* 7 (21):31384–31392
  53. Pomerantz MM, Li F, Takeda DY et al (2015) The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat Genet* 47(11):1346–1351
  54. Lu J, Lonergan PE, Nacusi LP et al (2015) The cistrome and gene signature of androgen receptor splice variants in castration resistant prostate cancer cells. *J Urol* 193(2):690–698
  55. Chen Y, Chi P, Rockowitz S et al (2013) ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nat Med* 19 (8):1023–1029
  56. Reid G, Metivier R, Lin CY et al (2005) Multiple mechanisms induce transcriptional silencing of a subset of genes, including oestrogen receptor alpha, in response to deacetylase inhibition by valproic acid and trichostatin A. *Oncogene* 24(31):4894–4907
  57. Waghra A, Schober M, Feroze F, Yao F, Virgin J, Chen YQ (2001) Identification of differentially expressed genes by serial analysis of gene expression in human prostate cancer. *Cancer Res* 61(10):4283–4286
  58. Lopez SM, Agoulnik AI, Zhang M et al (2016) Nuclear receptor corepressor 1 expression and output declines with prostate cancer progression. *Clin Cancer Res* 22 (15):3937–3949
  59. Doig CL, Singh PK, Dhiman VK et al (2013) Recruitment of NCOR1 to VDR target genes is enhanced in prostate cancer cells and

- associates with altered DNA methylation patterns. *Carcinogenesis* 34(2):248–256
60. Battaglia S, Maguire O, Thorne JL et al (2010) Elevated NCOR1 disrupts PPARalpha/gamma signaling in prostate cancer and forms a targetable epigenetic lesion. *Carcinogenesis* 31(9):1650–1660
61. Hodgson MC, Shen HC, Hollenberg AN, Balk SP (2008) Structural basis for nuclear receptor corepressor recruitment by antagonist-ligated androgen receptor. *Mol Cancer Ther* 7(10):3187–3194
62. Fereshteh MP, Tilli MT, Kim SE et al (2008) The nuclear receptor coactivator amplified in breast cancer-1 is required for Neu (ErbB2/HER2) activation, signaling, and mammary tumorigenesis in mice. *Cancer Res* 68(10):3697–3706
63. Banwell CM, MacCartney DP, Guy M et al (2006) Altered nuclear receptor corepressor expression attenuates vitamin D receptor signaling in breast cancer cells. *Clin Cancer Res* 12(7 Pt 1):2004–2013
64. Cheng S, Brzostek S, Lee SR, Hollenberg AN, Balk SP (2002) Inhibition of the dihydrotestosterone-activated androgen receptor by nuclear receptor corepressor. *Mol Endocrinol* 16(7):1492–1501
65. Lavinsky RM, Jepsen K, Heinzel T et al (1998) Diverse signaling pathways modulate nuclear receptor recruitment of N-CoR and SMRT complexes. *Proc Natl Acad Sci U S A* 95(6):2920–2925
66. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA et al (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45(10):1113–1120
67. Cancer Genome Atlas Research Network (2015) The molecular taxonomy of primary prostate cancer. *Cell* 163(4):1011–1025
68. Sanda MG, Feng Z, Howard DH et al (2017) Association between combined TMPRSS2:ERG and PCA3 RNA urinary testing and detection of aggressive prostate cancer. *JAMA Oncol* 3(8):1085–1093
69. Tomlins SA, Rhodes DR, Perner S et al (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310(5748):644–648
70. Armenia J, Wankowicz SAM, Liu D et al (2018) The long tail of oncogenic drivers in prostate cancer. *Nat Genet* 50(5):645–651
71. Jain P, Di Croce L (2016) Mutations and deletions of PRC2 in prostate cancer. *BioEssays* 38(5):446–454
72. Maki HE, Waltering KK, Wallen MJ et al (2006) Screening of genetic and expression alterations of SRC1 gene in prostate cancer. *Prostate* 66(13):1391–1398
73. Fraser M, Sabelnykova VY, Yamaguchi TN et al (2017) Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* 541(7637):359–364
74. Volinia S, Calin GA, Liu CG et al (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* 103(7):2257–2261
75. Jackson BL, Grabowska A, Ratan HL (2014) MicroRNA in prostate cancer: functional importance and potential as circulating biomarkers. *BMC Cancer* 14:930
76. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenines, indicates that thousands of human genes are microRNA targets. *Cell* 120(1):15–20
77. Lussier YA, Stadler WM, Chen JL (2012) Advantages of genomic complexity: bioinformatics opportunities in microRNA cancer signatures. *J Am Med Inform Assoc* 19(2):156–160
78. Loeb S, van den Heuvel S, Zhu X, Bangma CH, Schroder FH, Roobol MJ (2012) Infectious complications and hospital admissions after prostate biopsy in a European randomized trial. *Eur Urol* 61(6):1110–1114
79. Schwarzenbach H, Nishida N, Calin GA, Pantel K (2014) Clinical relevance of circulating cell-free microRNAs in cancer. *Nat Rev Clin Oncol* 11(3):145–156
80. Mihelich BL, Maranville JC, Nolley R, Peehl DM, Nonn L (2015) Elevated serum microRNA levels associate with absence of high-grade prostate cancer in a retrospective cohort. *PLoS One* 10(4):e0124245
81. Mitchell PS, Parkin RK, Kroh EM et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* 105(30):10513–10518
82. Loeb S, Chan DW, Sokoll L et al (2008) Prostate specific antigen assay standardization bias could affect clinical decision making. *J Urol* 180(5):1959–1962; discussion 1962–1953
83. Rupaimoole R, Calin GA, Lopez-Berestein G, Sood AK (2016) miRNA deregulation in cancer cells and the tumor microenvironment. *Cancer Discov* 6(3):235–246
84. Frediani JN, Fabbri M (2016) Essential role of miRNAs in orchestrating the biology of the tumor microenvironment. *Mol Cancer* 15(1):42
85. Kohlhapp FJ, Mitra AK, Lengyel E, Peter ME (2015) MicroRNAs as mediators and

- communicators between cancer cells and the tumor microenvironment. *Oncogene* 34 (48):5857–5868
86. Hudson RS, Yi M, Esposito D et al (2013) MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer. *Oncogene* 32 (35):4139–4147
  87. Goto Y, Kojima S, Nishikawa R et al (2014) The microRNA-23b/27b/24-1 cluster is a disease progression marker and tumor suppressor in prostate cancer. *Oncotarget* 5 (17):7748–7759
  88. Singh PK, Preus L, Hu Q et al (2014) Serum microRNA expression patterns that predict early treatment failure in prostate cancer patients. *Oncotarget* 5(3):824–840
  89. Cagle P, Nitre S, Srivastava A et al (2019) MicroRNA-214 targets PTK6 to inhibit tumorigenic potential and increase drug sensitivity of prostate cancer cells. *Sci Rep* 9 (1):9776
  90. Farran B, Dyson G, Craig D et al (2018) A study of circulating microRNAs identifies a new potential biomarker panel to distinguish aggressive prostate cancer. *Carcinogenesis* 39 (4):556–561
  91. Yates C, Long MD, Campbell MJ, Sucheston-Campbell L (2017) miRNAs as drivers of TMPRSS2-ERG negative prostate tumors in African American men. *Front Biosci (Landmark Ed)* 22:212–229
  92. Alunni-Fabbroni M, Majunke L, Trapp EK et al (2018) Whole blood microRNAs as potential biomarkers in post-operative early breast cancer patients. *BMC Cancer* 18 (1):141
  93. Minemura H, Takagi K, Miki Y et al (2015) Abnormal expression of miR-1 in breast carcinoma as a potent prognostic factor. *Cancer Sci* 106(11):1642–1650
  94. Niu J, Xue A, Chi Y et al (2016) Induction of miRNA-181a by genotoxic treatments promotes chemotherapeutic resistance and metastasis in breast cancer. *Oncogene* 35 (10):1302–1313
  95. Gong Z, Wang J, Wang D et al (2019) Differences in microRNA expression in breast cancer between women of African and European ancestry. *Carcinogenesis* 40(1):61–69
  96. Telonis AG, Rigoutsos I (2018) Race disparities in the contribution of miRNA isoforms and tRNA-derived fragments to triple-negative breast cancer. *Cancer Res* 78 (5):1140–1154
  97. Yao S, Graham K, Shen J et al (2013) Genetic variants in microRNAs and breast cancer risk in African American and European American women. *Breast Cancer Res Treat* 141 (3):447–459
  98. Chakravarty D, Sboner A, Nair SS et al (2014) The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat Commun* 5:5383
  99. Zhang Y, Pitchiaya S, Cieslik M et al (2018) Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARlNC1 in prostate cancer progression. *Nat Genet* 50(6):814–824
  100. Luo J, Wang K, Yeh S et al (2019) LncRNA-p21 alters the antiandrogen enzalutamide-induced prostate cancer neuroendocrine differentiation via modulating the EZH2/STAT3 signaling. *Nat Commun* 10(1):2571
  101. Shang Z, Yu J, Sun L et al (2019) LncRNA PCAT1 activates AKT and NF-kappaB signaling in castration-resistant prostate cancer by regulating the PHLPP/FKBP51/IKKalpha complex. *Nucleic Acids Res* 47 (8):4211–4225
  102. Huang P, Li F, Li L et al (2018) lncRNA profile study reveals the mRNAs and lncRNAs associated with docetaxel resistance in breast cancer cells. *Sci Rep* 8(1):17970
  103. Ma Y, Bu D, Long J, Chai W, Dong J (2019) LncRNA DSCAM-AS1 acts as a sponge of miR-137 to enhance Tamoxifen resistance in breast cancer. *J Cell Physiol* 234 (3):2880–2894
  104. Olgun G, Sahin O, Tastan O (2018) Discovering lncRNA mediated sponge interactions in breast cancer molecular subtypes. *BMC Genomics* 19(1):650
  105. Wu X, Xiao Y, Zhou Y, Zhou Z, Yan W (2019) LncRNA FOXP4-AS1 is activated by PAX5 and promotes the growth of prostate cancer by sequestering miR-3184-5p to upregulate FOXP4. *Cell Death Dis* 10(7):472
  106. Long MD, Singh PK, Russell JR et al (2019) The miR-96 and RARgamma signaling axis governs androgen signaling and prostate cancer progression. *Oncogene* 38(3):421–444
  107. Jiang L, Yu X, Ma X et al (2019) Identification of transcription factor-miRNA-lncRNA feed-forward loops in breast cancer subtypes. *Comput Biol Chem* 78:1–7
  108. Wu Q, Qin H, Zhao Q, He XX (2015) Emerging role of transcription factor-microRNA-target gene feed-forward loops in cancer. *Biomed Rep* 3(5):611–616
  109. Zhao M, Sun J, Zhao Z (2013) Synergetic regulatory networks mediated by oncogene-driven microRNAs and transcription factors in serous ovarian cancer. *Mol BioSyst* 9 (12):3187–3198

110. Yan Z, Shah PK, Amin SB et al (2012) Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. *Nucleic Acids Res* 40(17):e135
111. El Baroudi M, Cora D, Bosia C, Osella M, Caselle M (2011) A curated database of miRNA mediated feed-forward loops involving MYC as master regulator. *PLoS One* 6(3):e14742
112. Thorne JL, Maguire O, Doig CL et al (2011) Epigenetic control of a VDR-governed feed-forward loop that regulates p21(waf1/cip1) expression and function in non-malignant prostate cells. *Nucleic Acids Res* 39(6):2045–2056
113. Koido M, Tani Y, Tsukahara S, Okamoto Y, Tomida A (2018) InDePTH: detection of hub genes for developing gene expression networks under anticancer drug treatment. *Oncotarget* 9(49):29097–29111
114. Zhang XQ, Wang ZL, Poon MW, Yang JH (2017) Spatial-temporal transcriptional dynamics of long non-coding RNAs in human brain. *Hum Mol Genet* 26(16):3202–3211
115. Chou SJ, Wang C, Sintupisut N et al (2016) Analysis of spatial-temporal gene expression patterns reveals dynamics and regionalization in developing mouse brain. *Sci Rep* 6:19274
116. Berglund E, Maaskola J, Schultz N et al (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* 9(1):2419
117. Labbe DP, Brown M (2018) Transcriptional regulation in prostate cancer. *Cold Spring Harb Perspect Med* 8(11)
118. McNair C, Xu K, Mandigo AC et al (2018) Differential impact of RB status on E2F1 reprogramming in human cancer. *J Clin Invest* 128(1):341–358
119. Kron KJ, Murison A, Zhou S et al (2017) TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat Genet* 49(9):1336–1345
120. Jung SH, Shin S, Kim MS et al (2016) Genetic progression of high grade prostatic intraepithelial neoplasia to prostate cancer. *Eur Urol* 69(5):823–830
121. Castro E, Jugurnauth-Little S, Karlsson Q et al (2015) High burden of copy number alterations and c-MYC amplification in prostate cancer from BRCA2 germline mutation carriers. *Ann Oncol* 26(11):2293–2300
122. Nora EP, Lajoie BR, Schulz EG et al (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381–385
123. Dixon JR, Selvaraj S, Yue F et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380
124. Taberlay PC, Achinger-Kawecka J, Lun AT et al (2016) Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* 26(6):719–731
125. Fabre PJ, Leleu M, Mormann BH et al (2017) Large scale genomic reorganization of topological domains at the HoxD locus. *Genome Biol* 18(1):149
126. Koch L (2016) Chromatin: going a TAD out on a limb. *Nat Rev Genet* 17(12):717
127. Lupianez DG, Kraft K, Heinrich V et al (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161(5):1012–1025
128. Cutrupi AN, Brewer MH, Nicholson GA, Kennerson ML (2018) Structural variations causing inherited peripheral neuropathies: a paradigm for understanding genomic organization, chromatin interactions, and gene dysregulation. *Mol Genet Genomic Med* 6(3):422–433
129. Wu P, Li T, Li R et al (2017) 3D genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations. *Nat Commun* 8(1):1937
130. daSilva LF, Beckedorff FC, Ayupe AC et al (2018) Chromatin landscape distinguishes the genomic loci of hundreds of androgen-receptor-associated lincRNAs from the loci of non-associated lincRNAs. *Front Genet* 9:132
131. Roe JS, Hwang CI, Somerville TDD et al (2017) Enhancer reprogramming promotes pancreatic cancer metastasis. *Cell* 170(5):875–888.e20
132. Nguyen VT, Barozzi I, Faronato M et al (2015) Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nat Commun* 6:10044
133. Krum SA, Miranda-Carboni GA, Lupien M, Eeckhoute J, Carroll JS, Brown M (2008) Unique ERalpha cistromes control cell type-specific gene regulation. *Mol Endocrinol* 22(11):2393–2406
134. Jozwik KM, Chernukhin I, Serandour AA, Nagarajan S, Carroll JS (2016) FOXA1 directs H3K4 monomethylation at enhancers via recruitment of the methyltransferase MLL3. *Cell Rep* 17(10):2715–2723

135. Theodorou V, Stark R, Menon S, Carroll JS (2013) GATA3 acts upstream of FOXA1 in mediating ESRI binding by shaping enhancer accessibility. *Genome Res* 23(1):12–22
136. Haller F, Bieg M, Will R et al (2019) Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat Commun* 10 (1):368
137. Martin-Garcia D, Navarro A, Valdes-Mas R et al (2019) CCND2 and CCND3 hijack immunoglobulin light-chain enhancers in cyclin D1(–) mantle cell lymphoma. *Blood* 133(9):940–951
138. Cuartero S, Merkenschlager M (2018) Three-dimensional genome organization in normal and malignant haematopoiesis. *Curr Opin Hematol* 25(4):323–328
139. Zimmerman MW, Liu Y, He S et al (2018) MYC drives a subset of high-risk pediatric neuroblastomas and is activated through mechanisms including enhancer hijacking and focal enhancer amplification. *Cancer Discov* 8(3):320–335
140. Ryan RJ, Drier Y, Whitton H et al (2015) Detection of enhancer-associated rearrangements reveals mechanisms of oncogene dysregulation in B-cell lymphoma. *Cancer Discov* 5 (10):1058–1071
141. Northcott PA, Lee C, Zichner T et al (2014) Enhancer hijacking activates GF11 family oncogenes in medulloblastoma. *Nature* 511 (7510):428–434
142. Luo H, Wang F, Zha J et al (2018) CTCF boundary remodels chromatin domain and drives aberrant HOX gene transcription in acute myeloid leukemia. *Blood* 132 (8):837–848
143. Petrovic J, Zhou Y, Fasolino M et al (2019) Oncogenic Notch promotes long-range regulatory interactions within hyperconnected 3D cliques. *Mol Cell* 73(6):1174–1190.e12
144. Negi S, Bolt CC, Zhang H, Stubbs L (2019) An extended regulatory landscape drives Tbx18 activity in a variety of prostate-associated cell lineages. *Dev Biol* 446 (2):180–192
145. Rickman DS, Soong TD, Moss B et al (2012) Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A* 109 (23):9083–9088
146. Rickman DS, Chen YB, Banerjee S et al (2010) ERG cooperates with androgen receptor in regulating trefoil factor 3 in prostate cancer disease progression. *Neoplasia* 12 (12):1031–1040
147. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA (2014) Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res* 24(9):1421–1432
148. Long MD, Smiraglia DJ, Campbell MJ (2017) The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. *Biomol Ther* 7(1)
149. Campbell MJ, Turner BM (2013) Altered histone modifications in cancer. *Adv Exp Med Biol* 754:81–107
150. Battaglia S, Maguire O, Campbell MJ (2010) Transcription factor co-repressors in cancer biology: roles and targeting. *Int J Cancer* 126(11):2511–2519
151. Ambatipudi S, Horvath S, Perrier F et al (2017) DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer* 75:299–307
152. Horvath S, Raj K (2018) DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* 19 (6):371–384
153. Levine ME, Hosgood HD, Chen B, Absher D, Assimes T, Horvath S (2015) DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging (Albany NY)* 7 (9):690–700
154. Zheng Y, Joyce BT, Colicino E et al (2016) Blood epigenetic age may predict cancer incidence and mortality. *EBioMedicine* 5:68–73
155. Levine ME, Lu AT, Quach A et al (2018) An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* 10 (4):573–591
156. Marioni RE, Shah S, McRae AF et al (2015) DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol* 16:25
157. Keegan THM, Kushi LH, Li Q et al (2018) Cardiovascular disease incidence in adolescent and young adult cancer survivors: a retrospective cohort study. *J Cancer Surviv* 12 (3):388–397
158. Perna L, Zhang Y, Mons U, Holleczek B, Saum KU, Brenner H (2016) Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenetics* 8:64
159. Chen BH, Marioni RE, Colicino E et al (2016) DNA methylation-based measures of biological age: meta-analysis predicting time

- to death. *Aging (Albany NY)* 8(9):1844–1865
160. Illich JZ, Kelly OJ, Inglis JE (2016) Osteosarcopenic obesity syndrome: what is it and how can it be identified and diagnosed? *Curr Gerontol Geriatr Res* 2016:7325973
161. Gonnelli S, Caffarelli C, Nuti R (2014) Obesity and fracture risk. *Clin Cases Miner Bone Metab* 11(1):9–14
162. Wilson D, Jackson T, Sapey E, Lord JM (2017) Frailty and sarcopenia: the potential role of an aged immune system. *Ageing Res Rev* 36:1–10
163. Franceschi C, Garagnani P, Morsiani C et al (2018) The continuum of aging and age-related diseases: common mechanisms but different rates. *Front Med (Lausanne)* 5:61
164. Pistore C, Giannoni E, Colangelo T et al (2017) DNA methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. *Oncogene* 36(40):5551–5566
165. Camoriano M, Kinney SR, Moser MT et al (2008) Phenotype-specific CpG island methylation events in a murine model of prostate cancer. *Cancer Res* 68(11):4173–4182
166. Sinha KM, Bagheri-Yarmand R, Lahiri S et al (2019) Oncogenic and osteolytic functions of histone demethylase NO66 in castration-resistant prostate cancer. *Oncogene* 38(25):5038–5049
167. Beltran H, Prandi D, Mosquera JM et al (2016) Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* 22(3):298–305
168. Braadland PR, Urbanucci A (2019) Chromatin reprogramming as an adaptation mechanism in advanced prostate cancer. *Endocr Relat Cancer* 26(4):R211–R235
169. Labbe DP, Sweeney CJ, Brown M et al (2017) TOP2A and EZH2 provide early detection of an aggressive prostate cancer subgroup. *Clin Cancer Res* 23(22):7072–7083
170. Panja S, Hayati S, Epsi NJ, Parrott JS, Mitrofanova A (2018) Integrative (epi) genomic analysis to predict response to androgen-deprivation therapy in prostate cancer. *EBio-Medicine* 31:110–121
171. Zou M, Toivanen R, Mitrofanova A et al (2017) Transdifferentiation as a mechanism of treatment resistance in a mouse model of castration-resistant prostate cancer. *Cancer Discov* 7(7):736–749
172. Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP (1998) Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* 72:141–196
173. Jones PA (1996) DNA methylation errors and cancer. *Cancer Res* 56(11):2463–2467
174. Jones PA, Gonzalgo ML (1997) Altered DNA methylation and genome instability: a new pathway to cancer? *Proc Natl Acad Sci U S A* 94(6):2103–2105
175. Gama-Sosa MA, Slagel VA, Trewyn RW et al (1983) The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res* 11(19):6883–6894
176. Costello JF, Fruhwald MC, Smiraglia DJ et al (2000) Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat Genet* 24(2):132–138
177. Massie CE, Mills IG, Lynch AG (2016) The importance of DNA methylation in prostate cancer development. *J Steroid Biochem Mol Biol* 166:1–15
178. Bird AP, Wolffe AP (1999) Methylation-induced repression—belts, braces, and chromatin. *Cell* 99(5):451–454
179. Singh PK, Doig CL, Dhiman VK, Turner BM, Smiraglia DJ, Campbell MJ (2013) Epigenetic distortion to VDR transcriptional regulation in prostate cancer cells. *J Steroid Biochem Mol Biol* 136:258–263
180. Campbell MJ, Park S, Uskokovic MR, Dawson MI, Koefller HP (1998) Expression of retinoic acid receptor-beta sensitizes prostate cancer cells to growth inhibition mediated by combinations of retinoids and a 19-nor hexafluoride vitamin D3 analog. *Endocrinology* 139(4):1972–1980
181. Litovkin K, Van Eynde A, Joniau S et al (2015) DNA methylation-guided prediction of clinical failure in high-risk prostate cancer. *PLoS One* 10(6):e0130651
182. Haldrup C, Mundbjerg K, Vestergaard EM et al (2013) DNA methylation signatures for prediction of biochemical recurrence after radical prostatectomy of clinically localized prostate cancer. *J Clin Oncol* 31(26):3250–3258
183. Bogdanovic O, Smits AH, de la Calle Mustienes E et al (2016) Active DNA demethylation at enhancers during the vertebrate phyletypic period. *Nat Genet* 48(4):417–426
184. Charlet J, Duymich CE, Lay FD et al (2016) Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol Cell* 62(3):422–431
185. Irizarry RA, Ladd-Acosta C, Wen B et al (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at

- conserved tissue-specific CpG island shores. *Nat Genet* 41(2):178–186
186. Stadler MB, Murr R, Burger L et al (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480(7378):490–495
187. Aran D, Sabato S, Hellman A (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol* 14(3):R21
188. Aran D, Hellman A (2013) Unmasking risk loci: DNA methylation illuminates the biology of cancer predisposition: analyzing DNA methylation of transcriptional enhancers reveals missed regulatory links between cancer risk loci and genes. *BioEssays* 36 (2):184–190
189. Aran D, Hellman A (2013) DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* 154(1):11–13
190. Fleischer T, Tekpli X, Mathelier A et al (2017) DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat Commun* 8 (1):1379
191. Yin Y, Morgunova E, Jolma A et al (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356(6337)
192. Guo H, Ci X, Ahmed M et al (2019) ONE-CUT2 is a driver of neuroendocrine prostate cancer. *Nat Commun* 10(1):278
193. Rotinen M, You S, Yang J et al (2018) ONE-CUT2 is a targetable master regulator of lethal prostate cancer that suppresses the androgen axis. *Nat Med* 24(12):1887–1898
194. Patten DK, Corleone G, Gyorffy B et al (2018) Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nat Med* 24 (9):1469–1480
195. Taslim C, Chen Z, Huang K, Huang TH, Wang Q, Lin S (2012) Integrated analysis identifies a class of androgen-responsive genes regulated by short combinatorial long-range mechanism facilitated by CTCF. *Nucleic Acids Res* 40(11):4754–4764
196. Hogeweg P (2011) The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* 7(3):e1002021
197. Weaver W (1970) Molecular biology: origin of the term. *Science* 170(3958):581–582
198. Danna K, Nathans D (1971) Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. *Proc Natl Acad Sci U S A* 68(12):2913–2917
199. Saiki RK, Gelfand DH, Stoffel S et al (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239(4839):487–491
200. Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. *Science* 254 (5028):59–67
201. Roberts L, Davenport RJ, Pennisi E, Marshall E (2001) A history of the human genome project. *Science* 291(5507):1195
202. Birney E (2012) The making of ENCODE: lessons for big-data projects. *Nature* 489 (7414):49–51
203. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816
204. Soller M, Fray R (2019) RNA modifications in gene expression control. *Biochim Biophys Acta Gene Regul Mech* 1862(3):219–221



# Chapter 8

## Accessible Pipeline for Translational Research Using TCGA: Examples of Relating Gene Mechanism to Disease-Specific Outcomes

**Anders E. Berglund, Ryan M. Putney, Jordan H. Creed, Garrick Aden-Buie, Travis A. Gerke, and Robert J. Rounbehler**

### Abstract

Bioinformatic scientists are often asked to do widespread analyses of publicly available datasets in order to identify genetic alterations in cancer for genes of interest; therefore, we sought to create a set of tools to conduct common statistical analyses of The Cancer Genome Atlas (TCGA) data. These tools have been developed in response to requests from our collaborators to ask questions, validate findings, and better understand the function of their gene of interest. We describe here what data we have used, how to obtain it, and what figures we have found useful.

**Key words** TCGA, RNAseq, Methylation, Gene expression

---

### 1 Introduction

The Cancer Genome Atlas (TCGA) program, a joint project between the National Cancer Institute and the National Human Genome Research Institute, started in 2006 and molecularly characterized ~20,000 primary cancer and matched normal samples for 33 cancer types. The data were generated by 20 institutions and requires 2.5 petabytes of storage. More than just data, it has also generated multiple computational tools [1] that provide easy access to the data or different ways to visualize the data. What makes TCGA unique, compared to other resources of publicly available data sources such as GEO [2] and ArrayExpress [3], is the integration of clinical data and multiple types of molecular data available. For most tumors, there is whole exome sequencing (WES), copy-number alteration, DNA methylation, RNA sequencing gene expression (RNAseq), miRNA expression, and tissue imaging, all easily linked together. Most of the tumor types also have an accompanying TCGA publication, all available through Genome Data

Commons (GDC) (<https://gdc.cancer.gov/about-data/publications>). In addition to the link to the publication, data files are also provided with the exact version of the data that was used in the study.

One popular tool is cBioPortal (<http://www.cbioportal.org>) which provides many different ways to visualize TGCA data, as well as other publicly available datasets [4, 5]. Even though cBioPortal provides a multitude of different types of visualization, in this book chapter we describe a set of figures we have found useful in our research.

---

## 2 Materials

Most of the data presented and used in this book chapter were extracted from the NCI Genomic Data Commons (GDC) PanCancerAtlas Publications web page (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). This web page provides individual data files for the different molecular data types. The individual data types will be discussed below.

### 2.1 Sample Annotation

Initial sample annotation is based on TCGA samples code where both sample type and tumor type are derived. The build-up of the TCGA barcode is described in detail here ([https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA\\_Barcodes/](https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcodes/)). In short, the first four characters are the project name, and in this case they all are “TCGA.” Characters 6–7 encode the tissue source, referring to both the hospital at which the sample was collected and the tumor type (Table 1). Characters 9–12 are the participant identification number. Standard TCGA study abbreviations will be used in this chapter (<https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>) (Table 1). The sample types are based on the NCI GDC Sample Type Codes web page (<https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes>). Similarly, the tumor type was determined by using the NCI GDC Tissue Source Site Codes web page <https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes>. In addition to these basic sample type descriptors, additional sample information was obtained from the TCGA-Clinical Data Resource (CDR) Outcome table (TCGA-CDR-SupplementalTableS1.xlsx) available on the NCI GDC PanCancer Publications web page.

### 2.2 RNAseq Gene Expression

RNAseq gene expression data was extracted from the RNA (Final) spreadsheet (EBPlusPlusAdjustPANCAN\_IlluminaHiSeq\_RNA-SeqV2.geneExp.tsv <http://api.gdc.cancer.gov/data/3586c0da-64d0-4b74-a449-5ff4d9136611>) on the NCI GDC PanCancer Publications web page. Values less than 0 were set to 0, and the

**Table 1**  
**TCGA study abbreviations**

<b>Study abbreviation</b>	<b>Study name</b>
LAML	Acute myeloid leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
LGG	Brain lower grade glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine carcinosarcoma
UCEC	Uterine corpus endometrial carcinoma
UVM	Uveal melanoma

data were log2 transformed using the formula:  $\log_2(x + 1)$ . Samples listed in the Merged Sample Quality Annotations spreadsheet (`merged_sample_quality_annotations.tsv` <http://api.gdc.cancer.gov/data/1a7d7be8-675d-4e60-a105-19d4121bdeb>) were excluded from further analysis. Tumor type and sample type were inferred from the TCGA barcode.

### **2.3 GISTIC Copy Number Alterations Data**

The PanCancer Atlas DNA copy number data (GISTIC data) were extracted from the Copy Number file (`broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg` <http://api.gdc.cancer.gov/data/00a32f7a-c85f-4f86-850d-be53973cbc4d>). Samples listed in the Merged Sample Quality Annotations spreadsheet (`merged_sample_quality_annotations.tsv`) were excluded from further analysis. Tumor type and sample type were inferred from TCGA barcode. Based on GISTIC2.0 [6] DNA copy number analysis, the following numeric values were assigned with the cBioPortal interpretation presented in parenthesis: -2 = Deep Deletion (possibly a homozygous deletion); -1 = Shallow Deletion (possible a heterozygous deletion); 0 = Diploid; +1 = Gain (a few additional copies, often broad); +2 = Amplification (more copies, often focal).

### **2.4 Survival Data**

The publication by Liu et al. established a standardized dataset for four clinical endpoints for survival data associated with the Pan-Cancer Atlas [overall survival (OS), progression-free interval (PFI), disease-specific survival (DSS), and disease-free interval (DFI)] [7]. These clinical endpoints were extracted from the PanCan Atlas Clinical with Follow-Up file (`clinical_PANCAN_patient_with_followup.tsv` <http://api.gdc.cancer.gov/data/0fc78496-818b-4896-bd83-52db1f533c5c>).

### **2.5 Methylation Illumina 450K Data**

For data present in this chapter, only samples with Illumina's Infinium HumanMethylation450 BeadChip data were used. Raw IDAT files were downloaded from TCGA. Preprocessing the data included normalization via internal controls followed by background subtraction using the methylumi R package from Bioconductor (Davis S, Du P, Bilke S, Triche T, Bootwalla M. methylumi: Handle Illumina methylation data. R package version 2.12.02014). The calculated  $\beta$ -values were then extracted from the MethylLumi-Set object following preprocessing. The  $\beta$ -values are defined as  $\beta = M/(M + U + \alpha)$  where  $M$  is the methylated signal intensity and  $U$  unmethylated signal intensity and  $\alpha$  is an offset factor (commonly  $\alpha = 100$ ) added to  $M + U$  to stabilize  $\beta$ -values when both  $M$  and  $U$  are small. HumanMethylation450 v1.2 Manifest File was used to annotate the probes.

---

## **3 Methods**

Described below are a set of figures that can be generated using TCGA data to describe how the expression of a gene of interest is altered in cancer, as well as its association with clinical outcomes. All

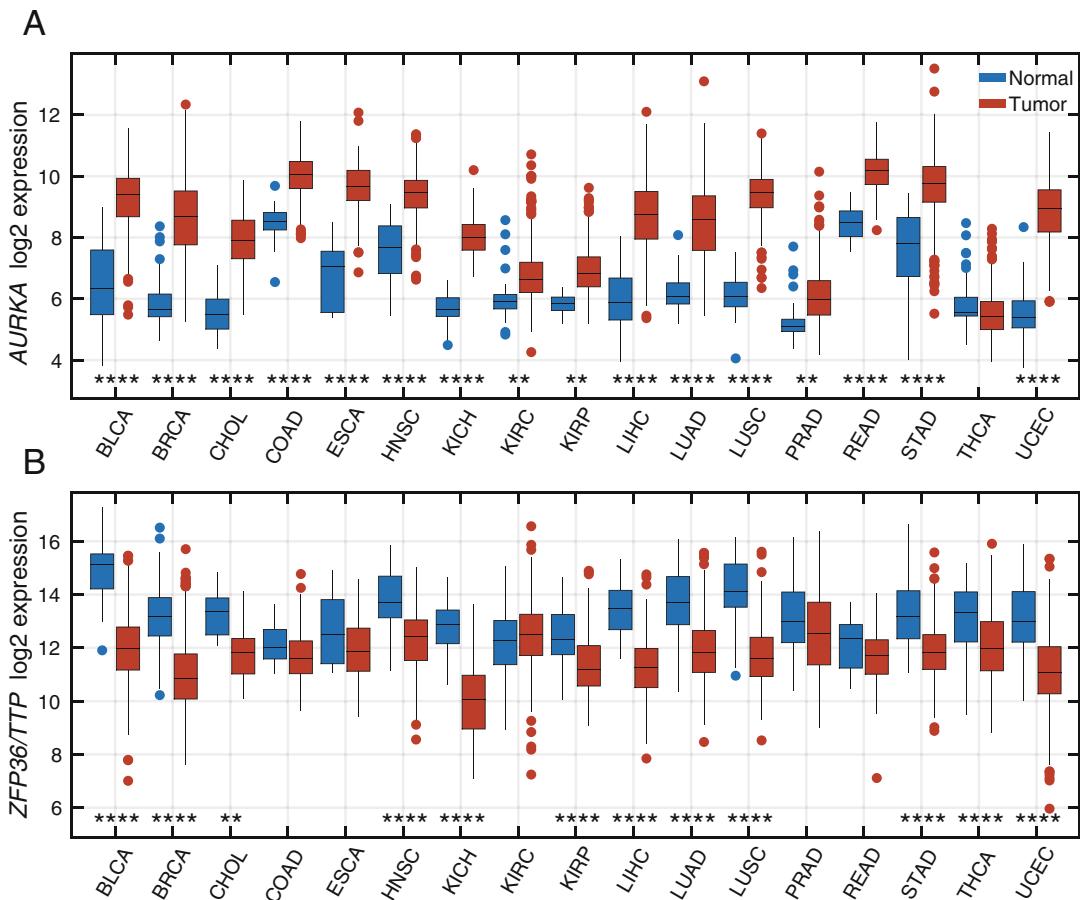
the figures below have been generated using MATLAB R2018b, Gramm [8], MatSurv (<https://www.mathworks.com/matlabcentral/fileexchange/64582-matsurv>). However, similar figures can also easily be generated using R, Python, or another similar programming language.

### **3.1 Tumor Versus Normal Box Plot**

One common question is whether a gene of interest is differently expressed between normal tissue and tumor samples. Seventeen out of the 33 tumor types in TCGA have more than 7 normal samples, which are shown in Fig. 1a, b. In these figures the gene expression in the tumor samples is presented in red, while the gene expression in the normal tissue sample is shown in blue. The box represents the gene expression between the 25th and 75th percentiles (Q1 and Q4 respectively) for each set of samples, and the black horizontal line indicates the median value (Q2). The vertical line (whiskers) indicates the range for all of the samples in each dataset, excluding any outliers (solid spheres). Outliers are defined as any data point that is more than 1.5 times the interquartile range (Tukey method), that is, points that are below  $Q1 - 1.5(Q3 - Q1)$  and points that are above  $Q3 + 1.5(Q3 - Q1)$ . The stars indicate the degree of significance in the differential expression between the tumor samples and normal tissue samples using both the *p*-value from Mann–Whitney U and the log<sub>2</sub> fold change (lfc):\* *p* < 0.05 and lfc > 0.585; \*\* *p* < 0.01 and lfc > 0.585; \*\*\* *p* < 0.001 and lfc > 1; \*\*\*\* *p* < 0.0001 and lfc > 1. The addition of fold change as a criterion for significance ensures that there is a biologically relevant change between normal and tumor tissue. The protein Aurora kinase A (*AURKA*) controls chromosome segregation during mitosis and is a marker of proliferation. Figure 1a clearly shows that the expression of *AURKA*, the gene that encodes for Aurora kinase A, is significantly increased in almost all tumors compared to adjacent normal tissue. Conversely, the RNA-binding protein tristetraprolin has been shown to be a tumor suppressor in several tumors [7]. In Fig. 1b, the gene that encodes tristetraprolin, *ZFP36/TTP*, is expressed at lower levels in tumor samples compared to normal tissue in 12 out of the 17 tumor types. It should be noted that genes may be biologically important even if there is no significant difference in expression between normal and tumor tissue. For example, *ZFP36/TTP* is shown to have no significant difference in expression between prostate tumor and normal prostate samples. However, it has been shown that decreased expression of tristetraprolin in prostate tumors is associated with increased cell proliferation and metabolism, as well as poor clinical outcomes [9–11].

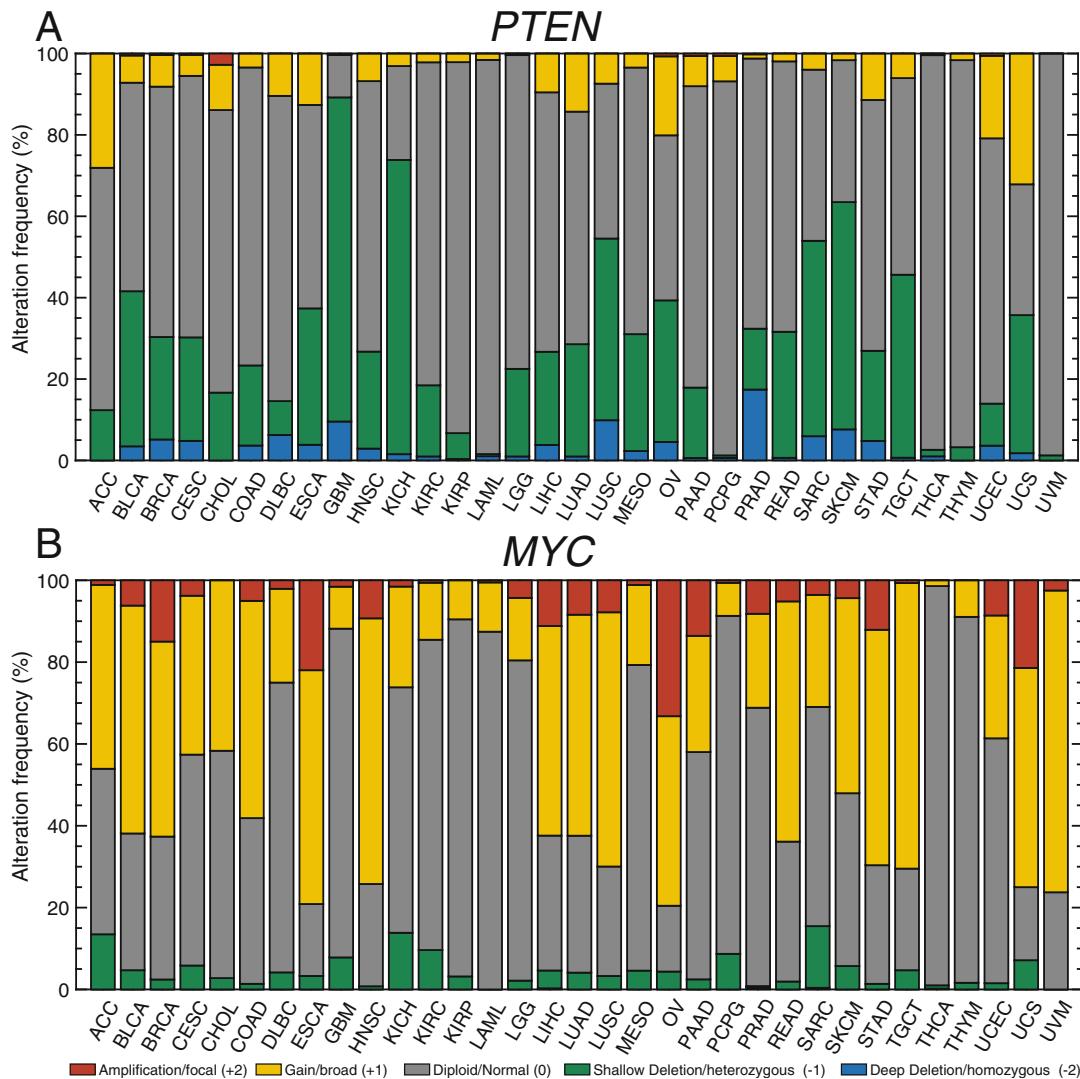
### **3.2 GISTIC Plot**

Changes in copy number, either gene deletions or gene amplifications, are common in most tumors. TCGA provides GISTIC as a measure for copy-number-alteration events across all genes and



**Fig. 1** Tumor versus normal. The expression of *AURKA* (a) and *ZFP36/TTP* (b) are compared between normal (blue) samples and tumor (red) samples across multiple tumor types. The expression is represented by box-plots representing the 25th to the 75th percentiles and the whiskers the range, excluding outliers. The stars indicate the degree of significance in the differential expression between the tumor samples and normal tissue samples using both the *p*-value from Mann–Whitney *U* and the log<sub>2</sub> fold change (lfc): \* *p* < 0.05 and lfc > 0.585; \*\* *p* < 0.01 and lfc > 0.585; \*\*\* *p* < 0.001 and lfc > 1; \*\*\*\* *p* < 0.0001 and lfc > 1

tumor types. These events can be efficiently visualized using bar plots as shown in Fig. 2a, b. For the GISTIC analysis there are five categories: (1) -2 = Deep Deletion (blue), homozygous deletion where both of the copies of the gene are deleted; (2) -1 = Shallow Deletion (green), heterozygous deletion where only one copy is deleted; (3) 0 = Diploid (gray), normal state; (4) +1 = Gain (yellow), often a few additional copies, often broad; (5) +2 = Amplification (red), more copies, often focal. The *y*-axis describes the percentage of samples within each category across the 33 tumor types. For example, in Fig. 2a, most glioblastoma multiforme (GBM) tumors show a shallow deletion for *PTEN*, which is most likely due to heterozygous deletion of the gene. There are also GBM tumors with a *PTEN* homozygous deletion, marked in



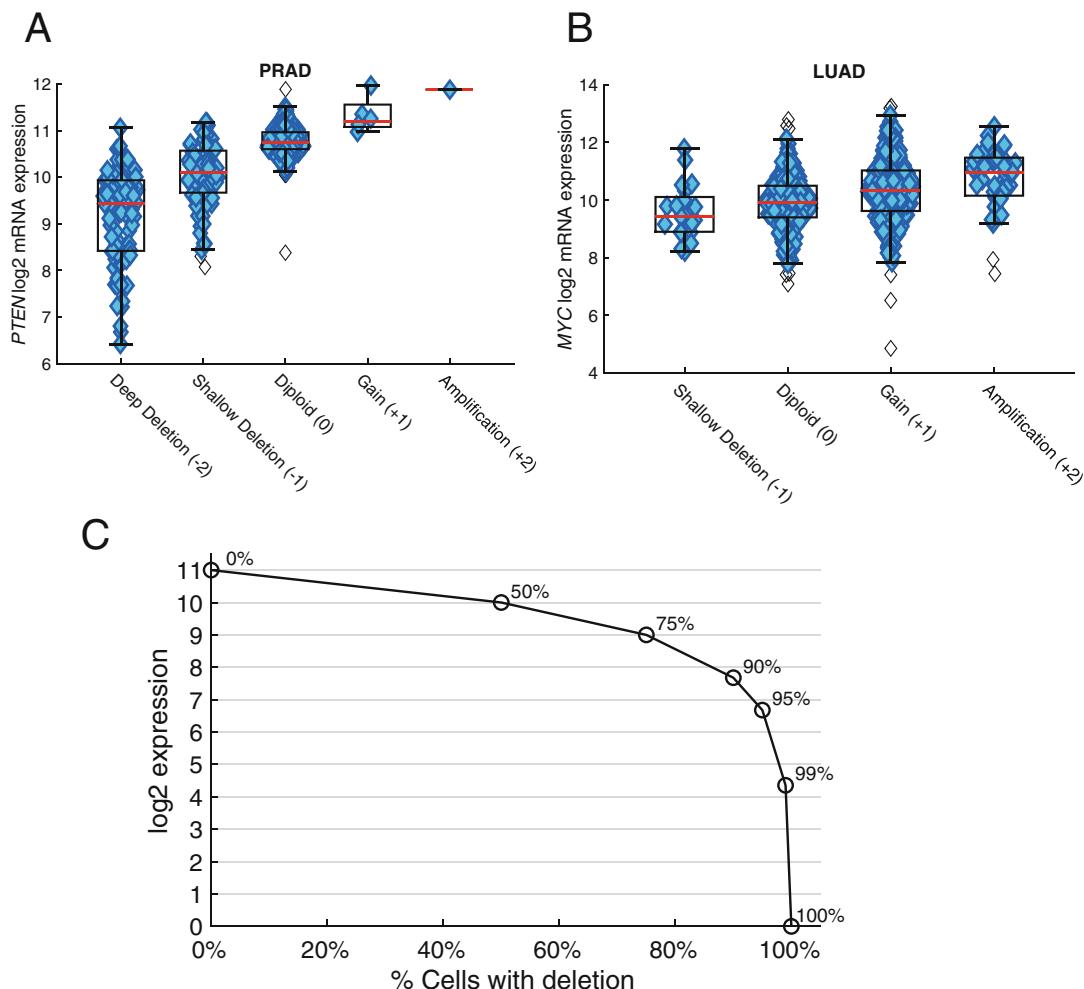
**Fig. 2** GISTIC plot. The frequency of alteration in *PTEN* (a) and *MYC* (b) across 33 TCGA tumor types: -2 = Deep Deletion (possibly a homozygous deletion); -1 = Shallow Deletion (possible a heterozygous deletion); 0 = Diploid; +1 = Gain (a few additional copies, often broad); +2 = Amplification (more copies, often focal)

blue. In addition to GBM, there is also a homozygous deletion of *PTEN* in many prostate adenocarcinoma (PRAD) tumors. On the other hand, *MYC* clearly shows an opposite trend, as it is amplified in many tumor types (Fig. 2b).

### 3.3 Gene Expression vs. GISTIC

Comparing the GISTIC score to changes in gene expression can also be very informative. The assumption is that both gene deletion and gene amplification will lead to large changes in gene expression, but this is, of course, not that simple. mRNA expression can be

regulated by a multitude of mechanisms (e.g., transcriptional regulation, mRNA degradation, or epigenetic silencing). Hence, plotting gene expression versus the GISTIC score can be informative to determine if the copy number changes also alter levels of gene expression (Fig. 3). In Fig. 3a, a clear trend is apparent in the analysis of *PTEN* gene expression versus its GISTIC score in PRAD samples. *PTEN* expression increases along the  $x$ -axis as samples with a deep (homozygous) deletion show lower expression compared samples with a shallow deletion, and both groups have lower expression than diploid samples. It is noteworthy that even in samples with a deep deletion of *PTEN*, its expression levels appear to be relatively high. This is most likely due to the fact that the deletion is only in the tumor cells, and therefore, the observed



**Fig. 3** Gene expression versus GISTIC score. The  $\log_2$  mRNA expression level for *PTEN* (a) and *MYC* (b) visualized based on GISTIC score. The expected  $\log_2$  expression level based on percentage of cells with a deletion (c)

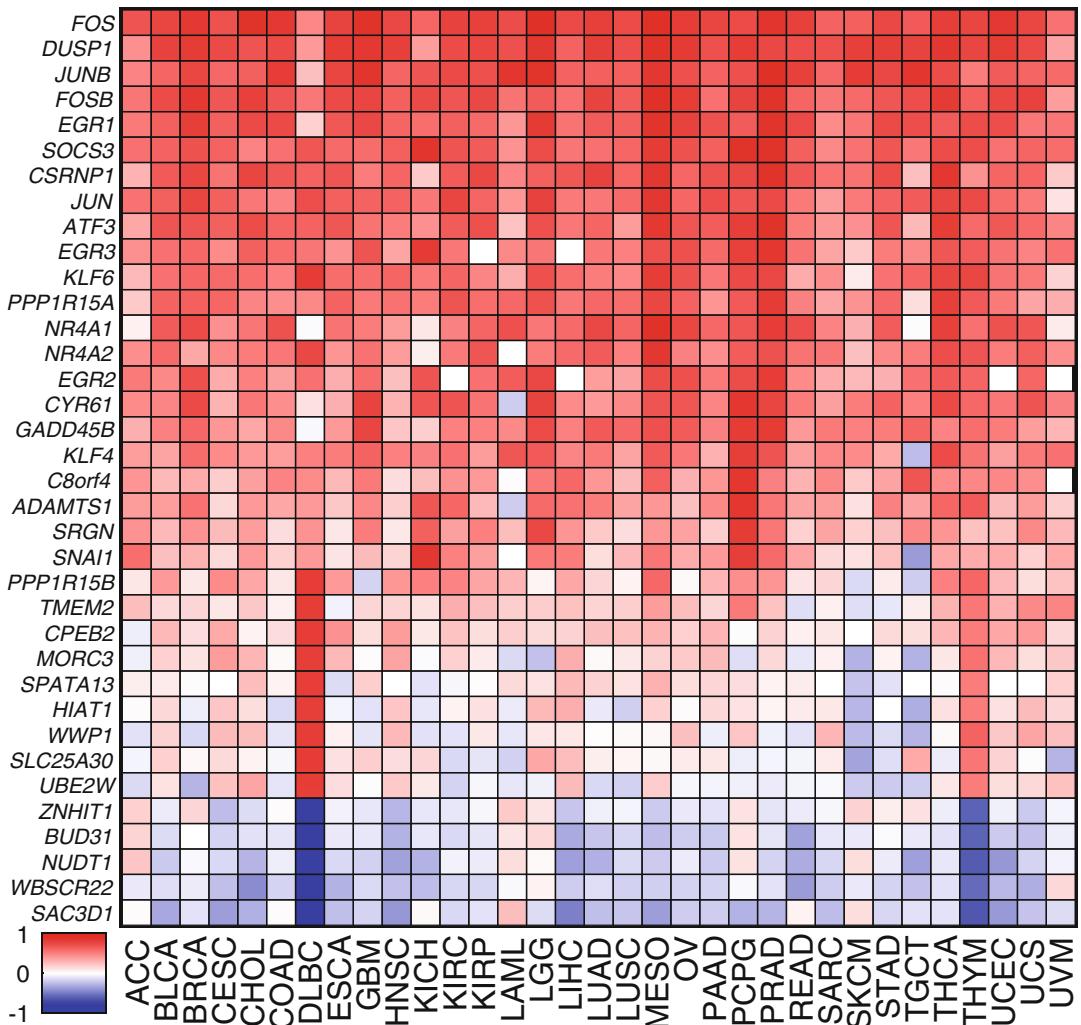
expression of *PTEN* may be coming from infiltrating immune cells or normal prostate cells. An alternative explanation is that the deletion is only present in a fraction of cancer cells. Further, it is important to remember that with RNAseq gene expression data (bulk RNAseq), the measurement represents the *average* expression level across all cells in the sample. The effect of the background expression signal based on tumor purity is illustrated in Fig. 3c. For example, if 90% of cells have *PTEN* deleted, the background expression in log2 units will be approximately 8. The TCGA project initially required 80% tumor nuclei, but was lowered to 60%. The effect of tumor purity on subsequent analysis is discussed in detail by Aran et al. [12]. Opposite to *PTEN*, *MYC* shows a much lower gene expression change based on GISTIC values (Fig. 3b). There are many potential explanations for this. For example, it is possible that upon amplification *MYC*, through its function as a transcription factor causes thousands of genes to be mis-regulated, resulting in *MYC* expression to become more normalized [13].

### 3.4 Correlation Analysis

Determining genes whose expression is correlated to a gene of interest might provide important biological information, especially by analyzing multiple tumor types. Figure 4 shows the genes that are highly correlated to *ZFP36/TTP* across all 33 TCGA tumor types and these genes are consistently correlated across all tumors. Several of these highly correlated genes are well studied, and many of them encode transcription factors, including *JUN*, *JUNB*, *FOS*, *FOSB*, *ATF3*, *KLF6*, *KLF4*, *NR4A1*, *NR4A2*, *EGR1*, and *EGR3*. Strikingly, only lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) and thymoma (THYM) show differences in some of the genes correlated to *ZFP36/TTP* compared to the other tumor types. Specifically, there is a set of genes (*SAC3D1*, *WBSCR22*, *NUDT1*, *BUD31*, and *ZNHIT1*) which are negatively correlated with *ZFP36/TTP* in DLBC and THYM, but not any of the other tumor types, suggesting that *ZFP36/TTP* expression might be regulated in a different manner in DLBC and THYM.

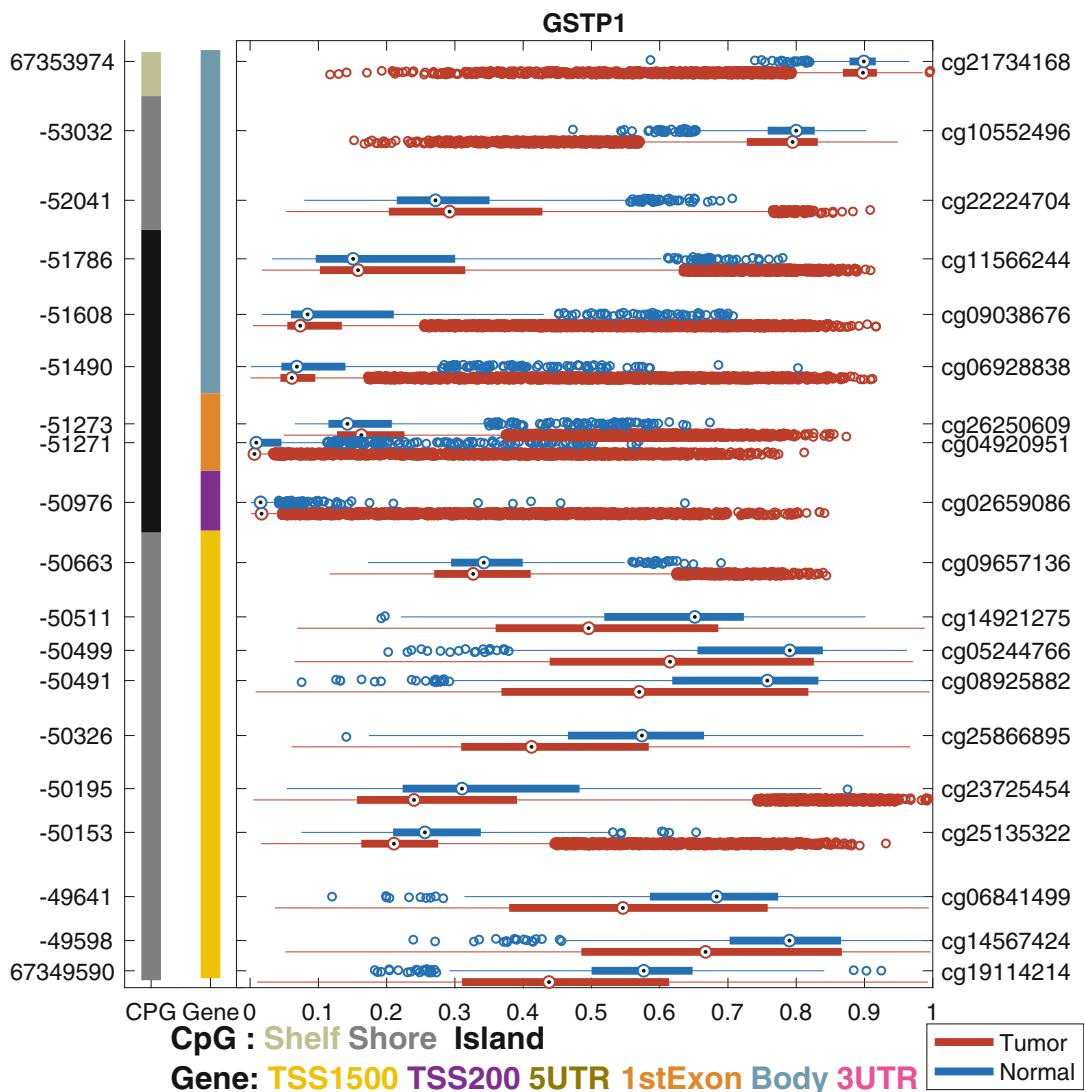
### 3.5 Epigenetic Regulation by Methylation

Abnormal methylation patterns often cause genes to be silenced or overexpressed, and this is especially important in cancer. In TCGA, almost all tumor and normal samples were assayed using the Infinium Human Methylation 450K BeadChip (450K). This chip provides ~480K methylation probes that measure the methylation level at 5'-C-phosphate-G-3' (CpG) sites. These are represented by a probe-id, a unique identifier for every probe on the Illumina chip that starts with a “cg” and followed by eight numeric values. Most genes have more than five probes. In Fig. 5, the methylation level of all 19 probes for *GSTP1* (glutathione S-transferase pi 1) across all TCGA samples is shown. The epigenetic regulation of *GSTP1* has previously been described [14]. The right y-axis shows



**Fig. 4** Genes correlated to *ZFP36/TPP* across TCGA tumor types. The most highly correlated genes to *ZFP36/TPP* expression across all TCGA tumors types are shown in a heatmap where the color indicate the Pearson correlation. Red indicate a positively correlation while blue indicate negative correlation

the probe-id, while the left *y*-axis shows the genomic position. The leftmost vertical column indicates the presence of CpG-islands in black, shores in gray, and shelves in light gray. The location of each probe within the *GSTPI* gene is presented in the second column from the left. Probes located 1500 base-pairs (bp) upstream of the transcription start site (TSS1500) are yellow, and probes 200 bp upstream (TSS200) are shown in purple. Probes located in the 5'UTR region are shown in brown (not present in *GSTPI*), the first exon is orange, the gene body is blue, and finally the 3'UTR region is pink (not present in *GSTPI*). The general assumption is that probes located in TSS1500, TSS200, and 5'UTR region will affect gene expression levels, but that is not always true. The



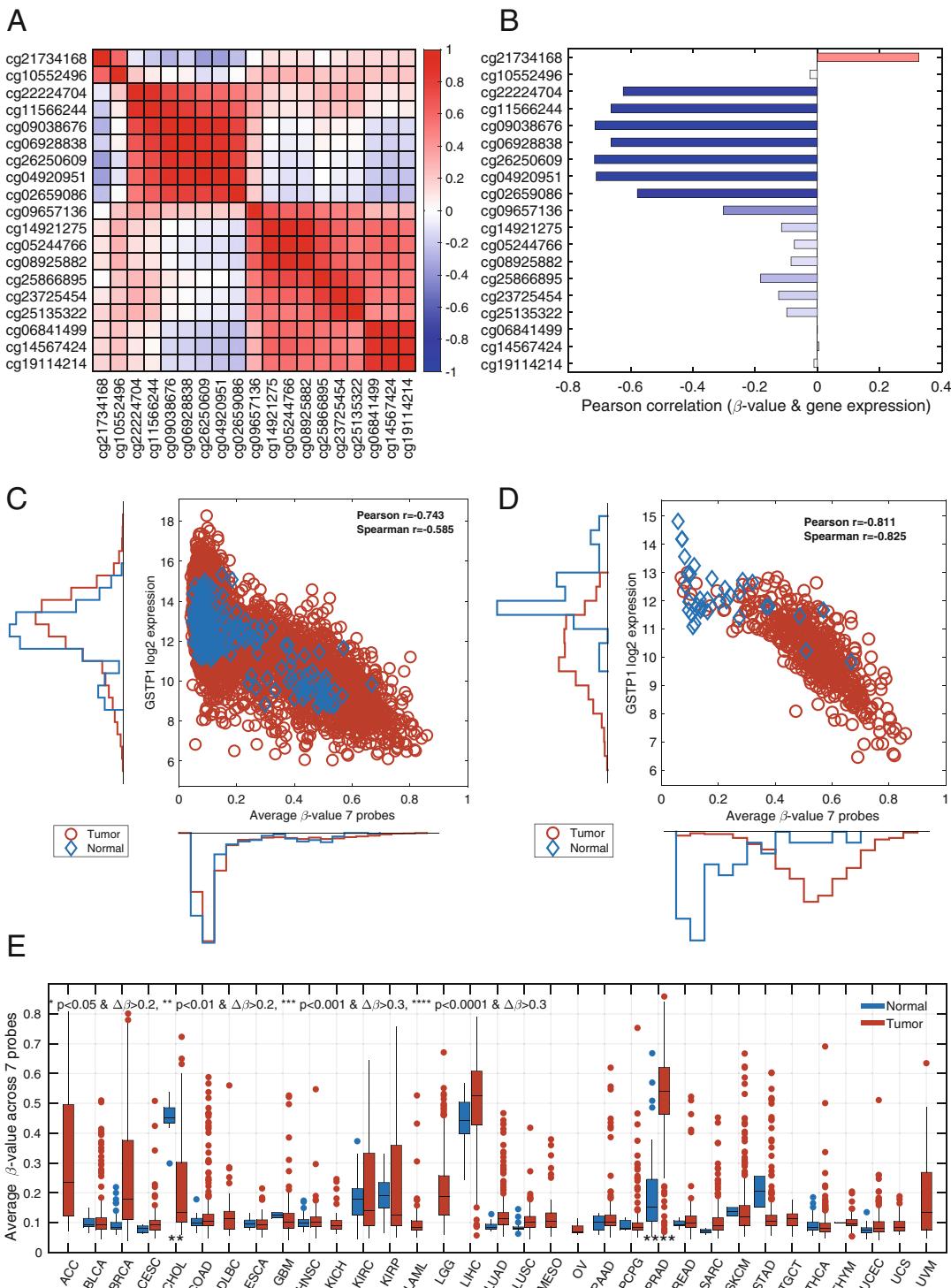
**Fig. 5** Methylation plot for *GSTM1*. The methylation level for the 16 probes of *GSTM1* across all TCGA tumor types are shown in red for tumors and in blue for normal samples. The genomic position and the probe-id are shown on the left and right y-axis, respectively. The left most column indicate CpG island and the second column where on the gene each probe is located. The  $\beta$ -value are shown on the x-axis where a zero indicate no methylation and one 100% methylation

methylation level for each probe is given a beta-value along the  $x$ -axis, where a zero means no methylation and a one means 100% methylation. The measured methylation level is an average across all DNA molecules and all the cell types in the measured sample. For each probe, the methylation level for all TCGA tumors is shown in red box plots and the normal samples in blue box plots. Starting at the bottom, there is a wide variation of methylation patterns. This is true for all probes located in the TS1500 region, ranging from

cg19114214 to cg09657136. The next section of probes cg02659086 to cg2224704 shows lower levels of methylation, which is especially true for cg02559086 and cg04920951. Both the box, corresponding to the 25th to 75th percentile, and the range are small for these two probes. The last two probes, top part of the graph, indicate a high degree of methylation. This pattern is confirmed in the probe correlation plot (Fig. 6a). Figure 6a clearly presents that three groups of probes are highly correlated. Interestingly, only the seven probes in the middle, probes cg03569086 to cg2224704, show any considerable correlation (Pearson  $r < -0.5$ ) to gene expression level (Fig. 6b). The average methylation levels for these seven probes are negatively correlated with *GSTP1* expression levels across all TCGA tumor and normal samples (Fig. 6c). It is surprising that these probes have a negative correlation since some of them are located in the gene body (Fig. 5). The average expression for these seven probes can also be visualized as tumor vs normal box-plots (Fig. 6e). For almost all tumor types, normal samples have low methylation levels, blue boxes, with the exception of cholangiocarcinoma (CHOL) and liver hepatocellular carcinoma (LIHC). Most tumor types have a low methylation level (red boxes) similar to the normal tissue samples; however, generally tumors have more outlier samples with high degree of methylation (red circles). Prostate adenocarcinoma (PRAD) is the only tumor type that has significantly higher methylation level in the tumor samples compared to the normal samples (\*\*\*\*,  $p < 0.0001$  and  $\Delta\beta > 0.3$ ). Figure 6d shows the correlation between methylation levels and gene expression levels for *GSTP1* in PRAD samples, and almost all of the prostate tumor samples show increased levels of methylation and a corresponding down regulation of *GSTP1* mRNA expression.

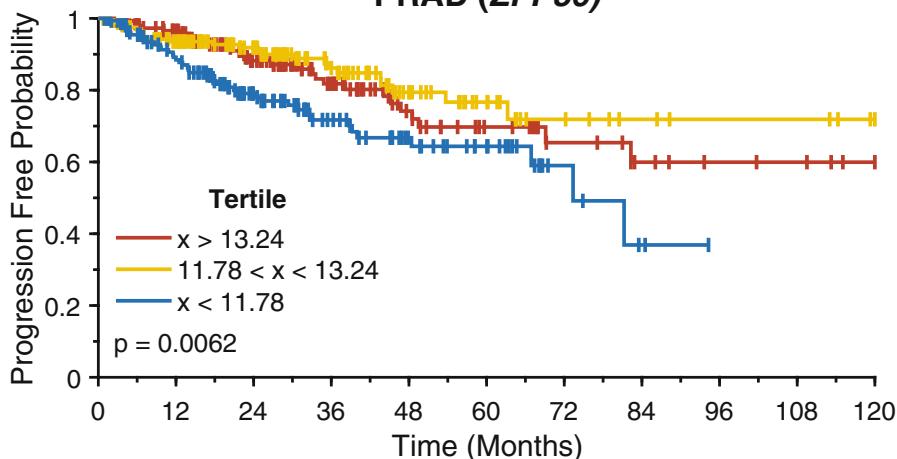
### 3.6 Survival Analysis

The correlation of gene expression levels and clinical outcomes, such as overall survival (OS) or progression free survival (PFI), is key for translation science. In a recent publication by Liu et al. all the TCGA survival data was reviewed and compiled [7]. Our previous publications [9, 11] have shown that the expression of *ZFP36/TTP* is related to time to biochemical reoccurrence (BCR) in the 333 TCGA prostate tumor samples published [15]. In that publication, only 333 samples were used since ~150 samples had a 3'-5' bias, but in this study we use all 493 prostate tumor samples with clinical data as provided in the Liu et al. publication. BCR is not explicitly available, but PFI should provide similar results. Figure 7a is a Kaplan–Meier (KM) plot for *ZFP36/TTP* gene expression using tertiles as cutoff points and PFI as clinical endpoint. It is clear that patients with low expression of *ZFP36/TTP* (blue line) have a worse clinical outcome than patients with a high *ZFP36/TTP* expression. Different cut-points can also be selected for analysis, such as median or quartile, and in the case of *ZFP36/TTP* these also



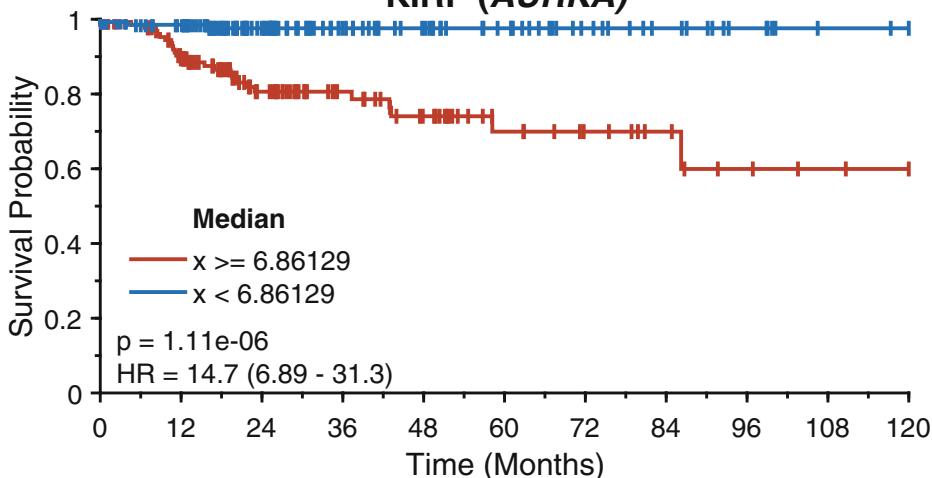
**Fig. 6** Gene expression level are dependent on methylation. The pairwise Pearson correlation between the probes for *GSTP1* shown in a heatmap (a). The correlation between the methylation level for each methylation probe and the *GSTP1* mRNA expression using all TCGA samples shown in a bar plot (b). The average methylation of the seven probes with high individual correlation compared to *GSTP1* mRNA expression across all TCGA samples (c) and in Prostate Adenocarcinoma only (d). Normal samples are indicated in blue diamonds and tumor samples in red circles. Box-plot for the average methylation level across all TCGA tumor types with normal samples in blue and tumor samples in red (e). \*,  $p < 0.05$  and  $\Delta\beta > 0.2$ , \*\*,  $p < 0.01$  and  $\Delta\beta > 0.2$ , \*\*\*,  $p < 0.001$  and  $\Delta\beta > 0.3$ , \*\*\*\*,  $p < 0.0001$  and  $\Delta\beta > 0.3$ .

A

**PRAD (*ZFP36*)**

$x > 13.24$	164	131	89	58	34	23	14	9	6	5	2
$11.78 < x < 13.24$	165	132	102	64	36	21	11	7	5	5	2
$x < 11.78$	164	125	77	47	28	19	6	2	0	0	0

B

**KIRP (*AURKA*)**

$x \geq 6.86129$	140	104	63	41	30	18	13	9	5	3	2
$x < 6.86129$	140	123	80	53	42	32	20	14	8	3	2

**Fig. 7** Survival analysis. Kaplan–Meier plot for progression free interval (PFI) in prostate adenocarcinoma (PRAD) using three groups based on their expression of *ZFP36/TTP* (a). Kaplan–Meier plot for overall survival (OS) in Kidney renal papillary cell carcinoma (KIRP) based on the expression of *AURKA* (b)

provide significant log rank  $p$ -values (median  $p = 0.037$  HR = 0.646 (0.43–0.971); quartile  $p = 0.0022$ , HR = 0.425 (0.246–0.735)). Figure 7b presents a similar analysis for the expression of *AURKA* and its association to OS. This graph shows that kidney renal papillary cell carcinoma (KIRP) patients with high expression of *AURKA*, a marker of proliferation, have a worse clinical outcome, as previously shown by others [16].

## 4 Concluding Remarks

The figures and tools presented in this chapter have been useful for translating known biological and mechanistic functions of a gene into clinically relevant insights for a specific disease of interest. The development of these tools has been driven by our collaborators and their needs. Having TCGA data compiled and ready to be used, together with these tools, figures can be quickly generated for grant applications, validation of in-house results, and in publications. We have used the concepts presented here to better understand the epigenetic regulation of cyclic GMP-AMP synthase (*cGAS*) and stimulator of interferon response cGAMP interactor 1 (*STING1*) [17], to investigate the transcriptional changes in ETS-fusion negative prostate tumors [18] and in the study of XIAP-associated factor 1 (*XAF1*) in glioblastoma [19]. Future work will include additional analyses and data types, such as mutational analysis, miRNA expression, and protein expression, which may further provide important information beyond the tools presented here. Another area not addressed in this chapter is the integrative analysis across multiple molecular data types.

## References

1. NCI. <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/tools>. Accessed Jul 2019
2. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1):207–210
3. Athar A, Füllgrabe A, George N et al (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res* 47(D1): D711–D715
4. Cerami E, Gao J, Dogrusoz U et al (2012) The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5):401–404
5. Gao J, Aksoy BA, Dogrusoz U et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6(269):pl1
6. Mermel CH, Schumacher SE, Hill B, Meyer-son ML, Beroukhim R, Getz G (2011) GIS-TIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12(4):R41
7. Liu J, Lichtenberg T, Hoadley KA et al (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173(2):400–416.e411
8. Morel P (2018) Gramm: grammar of graphics plotting in Matlab. *J Open Source Softw* 3:568
9. Berglund AE, Scott KE, Li W et al (2016) Tristetraprolin disables prostate cancer maintenance by impairing proliferation and metabolic function. *Oncotarget* 7(50):83462–83475
10. Gerke T, Beltran H, Wang X et al (2019) Low tristetraprolin expression is associated with lethal prostate cancer. *Cancer Epidemiol Biomark Prev* 28(3):584–590

11. Rounbehler RJ, Berglund AE, Gerke T et al (2018) Tristetraprolin is a prognostic biomarker for poor outcomes among patients with low-grade prostate Cancer. *Cancer Epidemiol Biomark Prev* 27:1376
12. Aran D, Sirota M, Butte AJ (2015) Systematic pan-cancer analysis of tumour purity. *Nat Commun* 6:8971
13. Lovén J, Orlando David A, Sigova Alla A et al (2012) Revisiting global gene expression analysis. *Cell* 151(3):476–482
14. Esteller M, Corn PG, Urena JM, Gabrielson E, Baylin SB, Herman JG (1998) Inactivation of glutathione S-transferase P1 gene by promoter hypermethylation in human neoplasia. *Cancer Res* 58(20):4515–4518
15. Abeshouse A, Ahn J, Akbani R et al (2015) The molecular taxonomy of primary prostate cancer. *Cell* 163(4):1011–1025
16. Ferchichi I, Kourda N, Sassi S et al (2012) Aurora A overexpression and pVHL reduced expression are correlated with a bad kidney cancer prognosis. *Dis Markers* 33(6):333–340
17. Konno H, Yamauchi S, Berglund A, Putney RM, Mulé JJ, Barber GN (2018) Suppression of STING signaling through epigenetic silencing and missense mutation impedes DNA damage mediated cytokine production. *Oncogene* 37(15):2037–2051
18. Berglund AE, Rounbehler RJ, Gerke T et al (2019) Distinct transcriptional repertoire of the androgen receptor in ETS fusion-negative prostate cancer. *Prostate Cancer Prostatic Dis* 22(2):292–302
19. Wu Q, Berglund AE, Wang D, MacAulay RJ, Mulé JJ, Etame AB (2019) Paradoxical epigenetic regulation of XAF1 mediates plasticity towards adaptive resistance evolution in MGMT-methylated glioblastoma. *Sci Rep* 9 (1):14072



# Chapter 9

## Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments

Xiaoqing Yu, Farnoosh Abbas-Aghababazadeh, Y. Ann Chen,  
and Brooke L. Fridley

### Abstract

High-throughput sequencing (HTS) has revolutionized researchers' ability to study the human transcriptome, particularly as it relates to cancer. Recently, HTS technology has advanced to the point where now one is able to sequence individual cells (i.e., "single-cell sequencing"). Prior to single-cell sequencing technology, HTS would be completed on RNA extracted from a tissue sample consisting of multiple cell types (i.e., "bulk sequencing"). In this chapter, we review the various bioinformatics and statistical methods used in the processing, quality control, and analysis of bulk and single-cell RNA sequencing methods. Additionally, we discuss how these methods are also being used to study tumor heterogeneity.

**Key words** Tumor heterogeneity, Transcriptomics, Single-cell, High-throughput sequencing, Differential expression, Normalization, Quality control

---

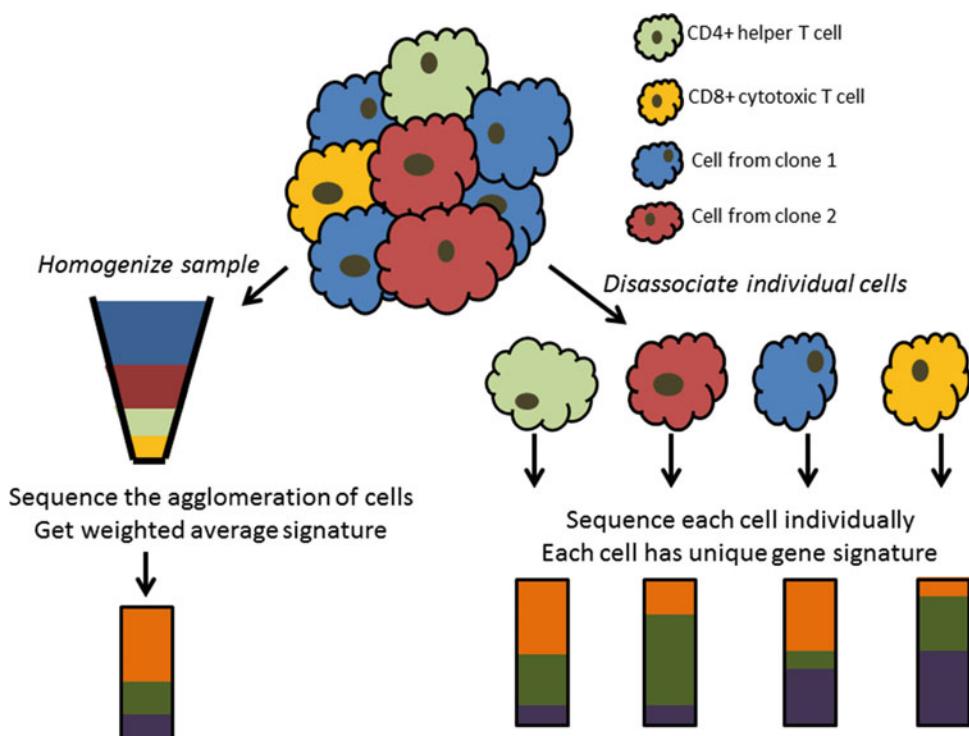
### 1 Introduction

Cancer research is becoming more computational focused in which researchers require working knowledge of data science methods to mine and extract new insights from large molecular datasets to derive novel hypotheses. Much of this need stems from the initial sequencing of the human genome to the advent of single-cell sequencing technologies, which has resulted in an explosion in the ability of cancer researchers to generate and acquire high-dimensional data. The most common type of 'omic data currently generated in the study of cancer biological systems is mRNA gene expression or transcriptomic data.

Uncontrolled cell growth and proliferation is a hallmark feature of cancer. In doing so, genes that regulate cell growth and differentiation are altered, thus leading to the uncontrollable growth and oncogenesis. Often, these changes in gene regulation are caused by mutations that disrupt the "normal" function of the gene and the downstream protein, such as in the case of p53 [1]. In other cases,

the gene regulation is changed by aberrant DNA methylation, where by silence the expression of a given gene [2]. Thus, studying the gene expression gives us clues into the biological mechanism of oncogenesis, along with tumor progression and growth. Additionally, gene expression has been used to subclassify a tumor class into smaller, more homogeneous molecular subtypes. For example, gene expression studies have been able to successfully subclassify breast cancer tumors into four to six molecular subtypes (luminal A, luminal B, HER2-enriched, triple-negative/basal-like, normal-like, claudin low) [3–7].

With the advances in technology from microarrays to high-throughput sequencing (HTS), one is able to detect other transcriptomic events, such as gene fusions or chimeras, isoform expression, and allelic expression [8]. Recently, HTS technology has advanced to the point that now one is able to sequence individual cells (i.e., “single-cell sequencing”) [9, 10]. Prior to single-cell sequencing technology, HTS would be completed on RNA extracted from a tissue sample made up of multiple cell types (i.e., “bulk sequencing”). The difference between single cell sequencing and bulk sequencing of RNA is that in the former the sequencing library represents a single cell while the latter represents a population of cells (Fig. 1). Single-cell technology allows researchers study



**Fig. 1** Illustration of differences between bulk RNA sequencing (RNA-seq) vs. single-cell RNA sequencing (scRNA-seq)

the transcriptome of different cells within the same tissue type. This technology is particularly useful in studying cancer immunology and the dissection of tumor heterogeneity, as tumors and the stromal component of tumors are a composition of (1) different cancer cells developed from different genomic events (i.e., clones, tumor heterogeneity) [11, 12] and (2) mixture of cancer cells and immune cells (i.e., tumor infiltrating lymphocytes (TILs)) [13, 14]. In the following sections we will outline the various bioinformatics and statistical methods used in the analysis of bulk and single-cell RNA sequencing.

---

## 2 Datasets Used to Illustrate the Methods

Melanoma is the fifth most common malignancy in the United States and it is estimated that 96,480 individuals will be diagnosed with melanoma in 2019 and that an estimated 7230 will die from melanoma [15]. Many genomic studies have been conducted to understand the molecular features of melanoma. Data from The Cancer Genome Atlas (TCGA) determined four major subtypes of cutaneous melanoma: *BRAF* mutant (52% of tumors), *RAS* mutant, *NFI* mutant, and Triple Wild-Type [16]. It was also found that the immune system plays a central role in the progression and treatment response in melanoma patients. To better understand the influence of immune system and tumor heterogeneity in melanoma, many studies have recently been completed to understand melanoma at the single-cell level [17–20]. To illustrate the bioinformatic and statistical methods used in the analysis of bulk and single-cell RNA-sequencing, we will use data from the TCGA melanoma study [16] and the Tirosh et al. study [20].

### 2.1 Bulk RNA-Sequencing Study: TCGA Skin Cancer Study

The RNA-seq summarized gene expression levels of skin cutaneous melanoma study (SKCM) using data obtained from TCGA project were downloaded via Genomic Data Commons (GDC) [16]. To illustrate differential expression analysis using RNA-seq data, set out to determine differentially expressed genes between primary tumors ( $N = 67$ ) and metastatic tumors ( $N = 213$ ). After filtering nonexpressed or low-expressed genes based on counts per million (CPM), 22,236 genes with CPM values above 1 in at least four libraries remain. To illustrate assessment of batch effects, technical artifacts were downloaded from *MBatch* (<https://bioinformatics.mdanderson.org/tcgabatcheffects>) for the SKCM TCGA data. This data set included the following variables: tissue source site (25 levels), plate ID (16 levels), batch ID (14 levels), and ship date (14 levels).

## **2.2 Single Cell Sequencing Study: Tirosh et al. Study**

Tirosh et al. [20] measured single-cell RNA-seq gene expression of 4645 melanoma, immune, and stromal cells from 19 melanoma tumors. These tumors included one primary acral melanoma, ten metastases to lymphoid tissues, and eight metastases to distant sites. The immune ( $CD45^+$ ) and nonimmune ( $CD45^-$ , including melanoma and stromal) cells were sorted into 96-well plates by flow cytometry (fluorescence-activated cell sorting). Single cell RNA was then isolated and sequenced with SMART-Seq2 protocol [21]. The gene expressions were quantified as  $y = \log_2(\text{TPM} + 1)$ , where TPM refers to transcripts per million. Cells with either fewer than 1700 detected genes or average housekeeping gene expression below 3 were excluded.

---

## **3 Statistical and Bioinformatics Methods for Analysis of Bulk RNA-Seq Data**

Current RNA-seq protocols still possess several essential biases and limitations, such as nucleotide composition bias, GC content, and polymerase chain reaction artifact or contaminations [22, 23]. Raw RNA-seq data must be checked and processed by quality control (QC) procedures to ensure accurate transcript measurements. Initial steps in the QC process typically involve assessing such biases of the raw reads using metrics generated by the sequencing platform or calculated directly from the raw reads (Table 1). One of the most popular tools for the generation of these quality metrics is FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The RNA-seq raw data and alignment files include various formats such as FASTA to store reference genome [24], gene transfer format (GTF) to store transcript/gene annotations, FASTQ to store raw read data [25], and the sequence alignment map (SAM/-BAM) to store read alignments [26]. RNA-seq analysis typically is consisted of major steps including raw data quality control (QC), read alignment, transcriptome reconstruction, expression quantification, and end with downstream analysis (Fig. 2).

### **3.1 Quality Control of RNA-Seq Data**

QC of raw data should be performed as the initial step which involves assessing such biases using metrics generated by the sequencing platform or calculated directly from the raw reads. In addition, depending on the RNA-seq library construction strategy and sequencing [27], trimming strategies include “adapter trimming” and “quality trimming” can be used to remove low-quality reads, trim adaptor sequences, and eliminate poor-quality bases. Adapter trimming is not necessary as most recent sequencers provide raw read in which the adapters are already trimmed, while quality trimming may be an important step depending on the analysis procedure used. Table 1 represents the widely used sequencing QC software tools.

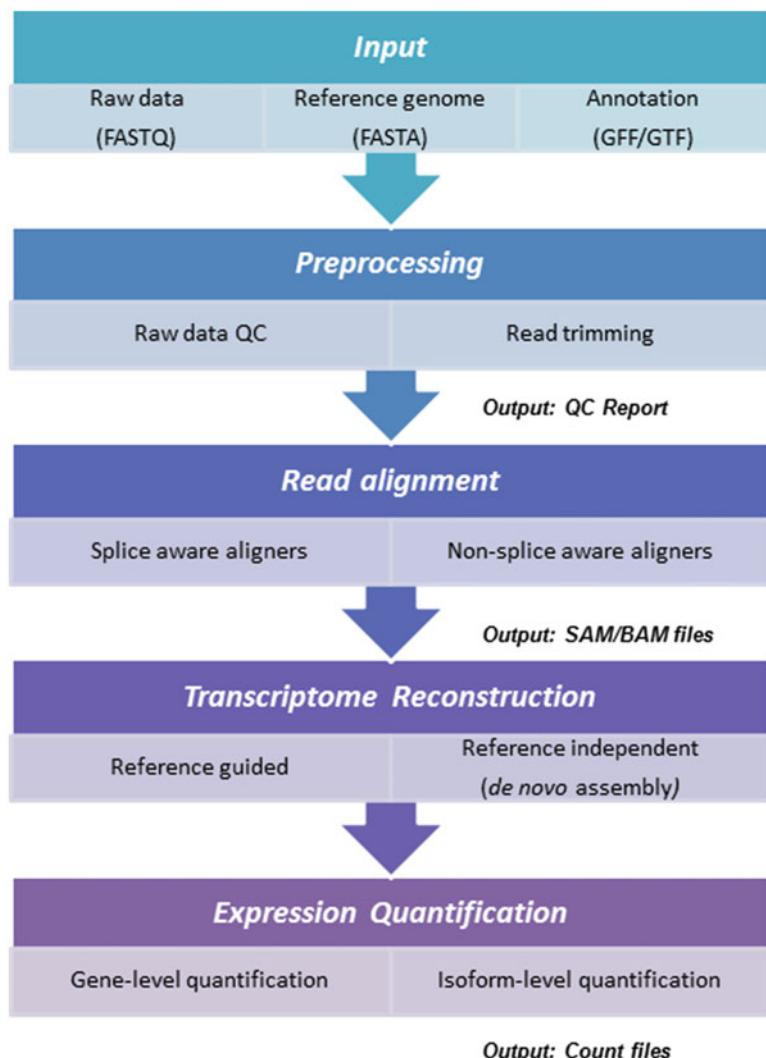
**Table 1**

**Selected list of RNA-seq analysis tools for preprocessing, read alignment, transcriptomic reconstruction, and expression quantification/abundance estimation**

Workflow	Category	Software tools	Reference
Preprocessing	RAW data QC	FastQC	Babraham Bioinformatics website
		HTQC	Yang et al. [182]
	Read trimming	NGS QC	Patel and Jain [183]
		FASTX-Toolkit	Cold Spring Harbor Laboratory website
		Trimmomatic	Bolger et al. [184]
		SolexaQA	Cox et al. [185]
Read alignment	Spliced aligner	TopHat	Trapnell et al. [186]
		STAR	Dobin et al. [30]
		MapSplice	Wang et al. [29]
		GSNAP	Wu et al. [31]
		Stampy	Lunter and Goodson [32]
	Unspliced aligner	MAQ	Li et al. [33]
		BWA	Li and Durbin [34]
		Bowtie2	Langmead and Salzberg [35]
Transcriptome reconstruction	Reference-guided	Cufflinks	Trapnell et al. [36]
		StringTie	Pertea et al. [37]
	Reference-independent	Trinity	Grabherr et al. [39]
		Oases	Schulz et al. [40]
		transABYSS	Robertson et al. [187]
Expression quantification	Gene-level quantification	featureCounts	Liao et al. [43]
		HTSeq	Anders et al. [44]
		Cufflinks	Trapnell et al. [36]
	Isoform-level quantification	StringTie	Pertea et al. [37]
		RSEM	Li and Dewey [41]
		Sailfish	Patro et al. [42]

### 3.2 Methods for Read Alignment

Reference-based alignment is the process used to determine the potential mapping locations by exact match or scoring sequencing similarity. Reads are typically aligned to either a genome or a transcriptome as a reference using two common approaches; (1) splice-aware read aligner or (2) nonsplice-aware read aligners (Table 1). The spliced read aligners use of a gapped or spliced mapper as reads may span splice junctions. Various spliced aligners have been developed including TopHat2 [28], MapSplice [29], STAR [30], and GSNAP [31]. Unspliced read aligners do not allow large gaps, such as those arising from reads spanning exon boundaries, or splice junctions including Stampy [32], mapping and assembly with quality (MAQ) [33], Burrow–Wheeler Aligner (BWA) [34], and Bowtie2 [35].



**Fig. 2** Typical bulk RNA sequencing workflow

### 3.3 Methods for Transcript Reconstruction

Transcript reconstruction includes two approaches to identify all transcripts expressed in a specimen depends on the presence or absence of a reference sequence (Table 1). When the reference annotation information is well-known, then the reference-based approaches such as Cufflinks [36] and StringTie [37] are used to reconstruct transcripts by assembly of overlapping aligned reads. When a reference genome or transcriptome is not available or is incomplete, assembled de novo algorithm directly builds transcripts from short reads using platforms such as Trinity [38], transABYSS [39], and Oases [40].

### 3.4 Methods for Gene Summarization or Abundance Estimation

One of the most widely used applications of RNA-seq is to quantify expression levels of genes and transcripts. Generally, the methods for gene quantification can be divided into two categories: “union exon”-based and “transcript”-based approaches (Table 1). Transcript-based approach fundamentally distributes reads among transcript isoforms including RSEM [41], Cufflinks [36], and StringTie [37]. However, some transcript-based quantification tools such as Sailfish [42] are alignment-free tools to estimate isoform abundances directly from a set of reference sequences. The “union exon”-based methods, such as featureCounts [43] and HTSeq [44], are widely used in RNA-seq gene quantification because of its simplicity to aggregate raw counts of mapped reads.

### 3.5 Normalization

Variability in measurement can be attributed to both the biological and technical factors. Sources of technical variation, involving, differences in library preparation across samples, sequencing error, mapping and annotation bias, sequencing composition and similarity, gene length, and sequencing depth [45–47] that can significantly reduce the accuracy of statistical inferences and also prevent researchers from properly modeling biological variation and group-specific changes in gene expression [48–51]. Some sources of between-sample technical variation are due to differences in library size or sequencing depth [47]. To correct for library size, most of methods use a common scaling factor per sample to normalize genes such as upper quartile (UQ) [45], median (Med) [45], relative log expression (RLE) [52], trimmed mean of *M*-values (TMM) [53], and quantile (Q) [54, 55]. Many of these methods are implemented in the edgeR and DESeq2 packages.

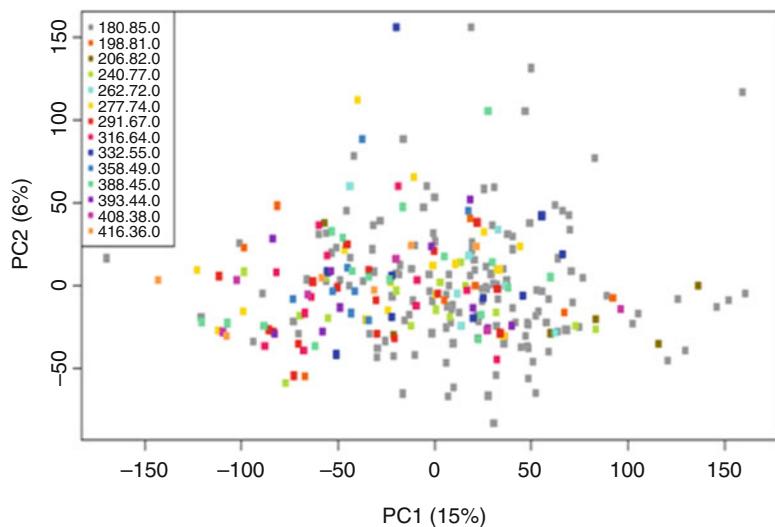
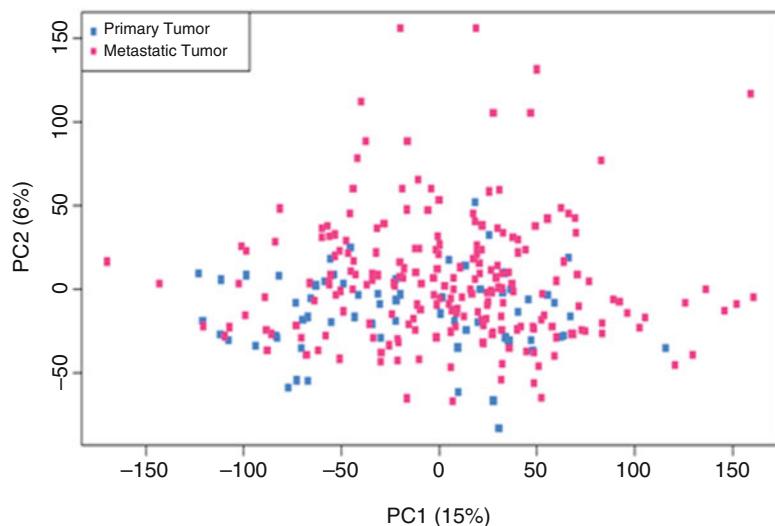
Additionally, gene length impacts the comparison of abundance estimates between genes [56], as longer genes contribute more sequenced fragments compared to shorter ones [47]. Most commonly used methods for gene length correction include TPM (transcripts per million) [57] and RPKM/FPKM (reads/fragments per kilobase per million mapped reads) [36, 47], where the former one is considered more robust to differences in RNA library size [58, 59]. However, it has been shown that scaling by gene length cannot entirely remove the positive association between gene size and read counts, and can introduce new biases to the estimates of differential expression [45, 60]. Usual normalization approaches mostly adjust the sequencing depth or/and gene length and fail to correct known or unknown technical artifacts due to the complex. To adjust known batch effect, appropriate statistical models were proposed such as linear regression model or flexible empirical Bayes method (ComBat) which is more robust and appropriate for small sample sizes [61]. ComBat can be implemented in R using the sva package and the function “ComBat.”

In addition to correcting for known artifacts, assessment and adjustment for potential latent factors is also warranted where mostly rely on singular value decomposition (SVD) or some other factor analysis approaches (e.g., Remove Unwanted Variation (RUV) in R package “ruv” [50] or surrogate variable analysis (SVA) [49] in R package “sva”). Finally, principal component analysis (PCA) is often used in which a subset of the principal components are used to normalize the data [62–65]. Note that the choice of normalization method to remove technical artifacts can affect noticeably the results of differential gene expression analyses [45, 53, 66]. For the TCGA skin cancer study, PCA was completed using filtered raw counts to assess the effects of the batch ID (Fig. 3a) and biological factor (i.e., primary and metastatic tumors groups) (Fig. 3b). Figure 3a represents the effect of batch ID with high proportion of variation for top principal component (15%) which leads to consideration of batch ID as a technical effect for normalization.

### **3.6 Methods for Differential Gene Expression Analysis**

The main methodologies for differential gene expression analysis for RNA-seq data are categorized by the distributional assumptions (Table 2). Models for read counts originated from the idea that the number of reads for each gene can be approximated by a Poisson distribution where log-linear or generalized linear model were proposed to model the mean difference between samples along with using test statistics such as likelihood ratio test, exact test, and score test for hypothesis testing are implemented in DEGseq [67], Myrna [68], and PoissonSeq [69] packages. The extended Poisson models, two-stage Poisson model [70] and generalized Poisson model [71] are also considered to adjust for overdispersion issue.

Poisson and negative binomial distributions are the two widely used models [52, 72], whereas the higher variability between biological replicates leads to incorporate negative binomial distribution to accommodate overdispersion [45]. The dispersion parameter estimation can be based on the conditional maximum likelihood, pseudo-likelihood, quasi-likelihood, local regression, and conditional inference using hypothesis testing approaches such as Wald test, likelihood ratio test, and exact test. Such methods are included in edgeR [73], DESeq [52], DESeq2 [74], and NBPSeq [75] packages. A beta-binomial model is implemented in BBSeq [76] package which accommodates the overdispersion using logistic regression where maximum likelihood approach is applied to estimate overdispersion parameter. Moreover, full or empirical Bayesian frameworks are included in ShrinkSeq [77] and baySeq [78] packages. Lastly, methods have been proposed that allow RNA-seq to be modeled using a linear model framework (i.e., Gaussian distribution) through limma package [79, 80] using voom transformation [81] or approaches that do not assume any

**a****b**

**Fig. 3** Plots from Principal Component Analysis to assess technical batch and biological factor effects globally in bulk RNA-seq experiments for the TCGA skin cancer study. **(a)** Fourteen levels of known batch ID, where batch ID was downloaded from <https://bioinformatics.mdanderson.org/BatchEffectsViewer/>; and **(b)** primary factor of interest (primary tumor and metastatic tumor) using filtered raw counts data

distribution assumption (nonparametric approaches) such as SAM-seq [82] and NOIseq [83]. It should be noted that if the modeling assumptions are valid the parametric methods will be more powerful than the nonparametric methods. However, if the modeling

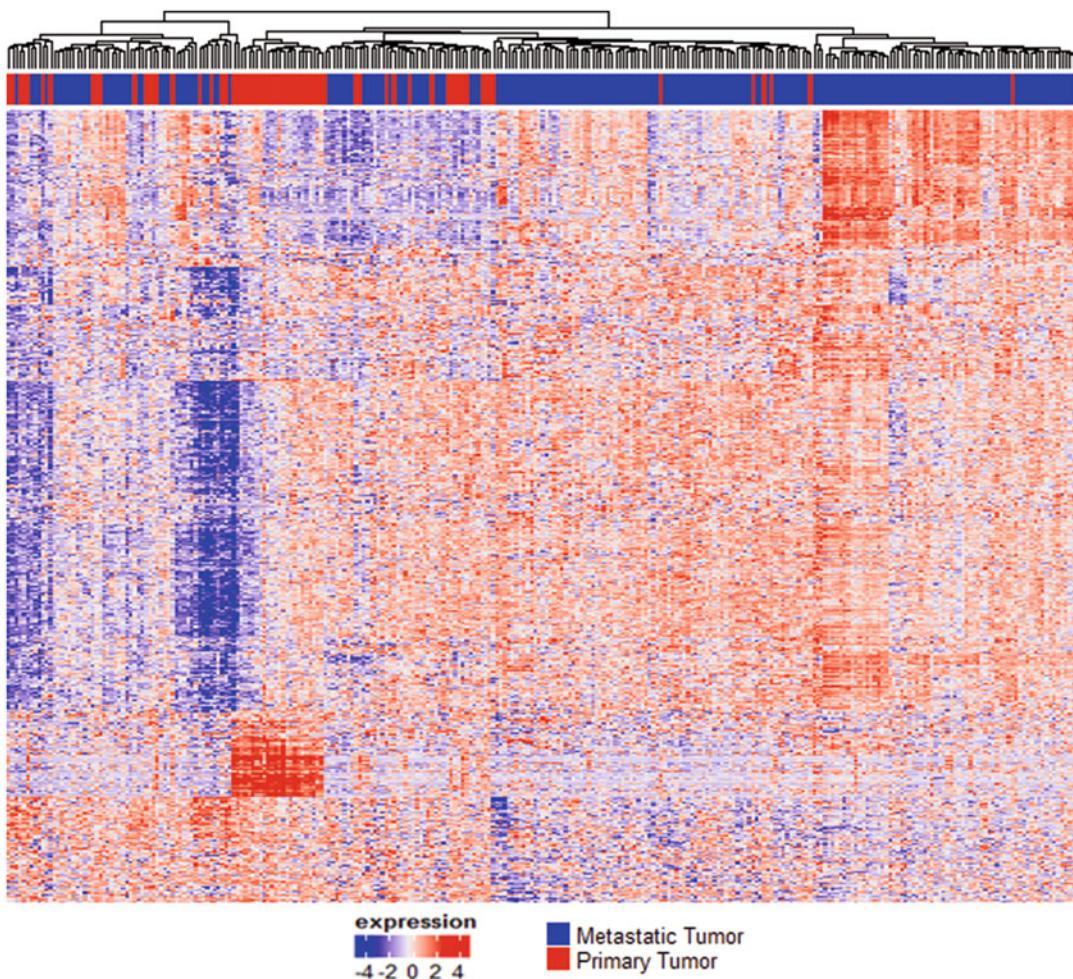
**Table 2**  
**Statistical methods to identify differential gene expressions based on RNA-seq data**

Model	Software	Reference
Poisson	DEGseq Myrna PoissonSeq	Wang et al. [67] Langmead et al. [68] Li et al. [69]
Negative binomial	edgeR DESeq DESeq2 NBSeq	Robinson et al. [73] Anders and Huber [52] Love et al. [74] Di et al. [75]
Beta-binomial	BBSeq	Zhou et al. [76]
Bayesian and empirical Bayesian	ShrinkSeq baySeq	Van de Wiel et al. [77] Hardcastle and Kelly [78]
Normal	limma+voom	Smyth [54, 79] Law et al. [81]
Nonparametric	SAMseq (samr) NOIseq	Li and Tibshirani [82] Tarazona et al. [83]

assumptions are not valid, nonparametric methods would be advisable. In practice, it is difficult to assess the modeling assumptions and thus parametric methods are mostly often used, particularly when the sample size is small. In skin-TCGA study, to identify the differentially expressed genes between primary and metastatic tumors groups (Fig. 3b), the voom-transformed UQ normalized data is considered where the design matrix contains the estimated latent artifact and the batch ID along with the tumor groups using limma package.

### 3.7 Methods for Correcting for Multiple Testing

With thousands of genes to test, controlling both the overall Type I error rate and the desired statistical power becomes important. Multiple comparison adjustment approaches can control the familywise error rate (FWER), false discovery rate (FDR) (i.e., Benjamini and Hochberg approach [84]), and Bayesian FDR ( $q$  values) [85, 86]. Various approaches control FWER and compute the adjusted  $P$  values such as Bonferroni [87], Holm [88], and Hochberg [89]. In contrast to the strong control of FWER, the FDR-based control is less conservative with the increased gain in power and has been widely used in cases where a large number of hypotheses are simultaneously tested. Figure 4 represents the heatmap of top differentially expressed genes ( $n = 8928$ ) after correcting multiple testing,  $\text{FDR}_{\text{BH}} < 0.05$ .



**Fig. 4** Heatmap of the top differentially expressed genes ( $FDR_{BH} < 0.05$ ) from the analysis of the TCGA skin cancer study

### 3.8 Studying TME Using RNA-Seq in Bulk Samples

A software tool CIBERSORT is widely used to estimate fractions of multiple cell types using gene expression data in bulk samples [90]. It is commonly used to characterize global immune landscape by estimating different proportions of different immune cells. For instance, in a recent large-scale study to characterize immune landscape by analyzing 10,000 tumors comprising 33 diverse cancer types, CIBERSORT was used to estimate immune infiltration fractions for understanding tumor-immune interaction [91]. Six immune subtypes identified are wound healing, IFN- $\gamma$  dominant, inflammatory, lymphocyte depleted, immunologically quiet, and TGF- $\beta$  dominant across cancer types and provided this as a source, iAtlas (<https://www.cri-iatlas.org/>), for researchers to understand tumor-immune interaction and potential therapeutic opportunities. In addition to cell type identification, the cell-cell interaction

from the ligand and receptor database is also incorporated. Although CIBERSORT is widely used, its performance potential is affected by statistical multicollinearity due to the inclusion of highly correlated immune cell types, and also was developed using expression on microarrays. TIMER is developed to select genes, which are negatively correlated with tumor purity for each cancer type, and then apply constrained least squares fitting to expression to predict the abundance of a subset of TILs: B cells, CD4 T cells, CD8 T cells, macrophages, neutrophils, and dendritic cells [92]. To capture the complexity in TME better, xCell attempted to infer 64 immune and stromal cell types by using gene set enrichment analyses and deconvolution method to analyze the 1822 harmonized human pure cell type transcriptomics samples [93].

---

## 4 Statistical and Bioinformatics Methods for Analysis of Single-Cell RNA-Seq Data

Single-cell RNA-sequencing (scRNA-seq) has been playing important roles in the study of tumor heterogeneity and tumor evolution. In contrast to the bulk RNA-seq where the average gene expressions are measured across a large population of cells, scRNA-seq quantifies transcriptome of individual cells. With the newly developed high-throughput cell separation technologies, thousands of cells per tumor can be profiled in parallel to capture intra-tumor heterogeneity at an unprecedented resolution. Multiple different platforms have been developed for scRNA-seq including SMART-seq [21], CEL-seq [94], Fluidigm C1 [95], Smart-seq2 [96], and more advanced droplet-based platforms including Drop-seq [97] and Chromium 10X [98]. In droplet-based platforms, cells are encapsulated in water-based droplets together with unique molecular identifiers (UMIs), a cell-specific and transcripts-specific barcoding system. These barcodes help to diminish the sequencing reads representation biases due to library amplification.

The considerable differences in cell isolation and molecule capture lead to large variations in sensitivity, specificity, and capacity of these platforms [99, 100]. However, they all rely on similar computational pipelines to reveal transcription dynamics. In the following sections, we will review the algorithms in major steps of scRNA-seq data analysis using the most commonly used droplet-based platforms, but the discussion applies to all platforms.

### 4.1 Quality Control

The droplet-based scRNA-seq platforms encapsulate thousands of cells individually into barcoded-droplets and sequence their RNA material simultaneously. All computational analyses and interpretations of results reply on the assumption of single-cell behavior, such that only one living cell exists in a single droplet. However, even for the most sensitive protocols, it is inevitable to have dead cells and doublets (multiple cells encapsulated in one droplet)

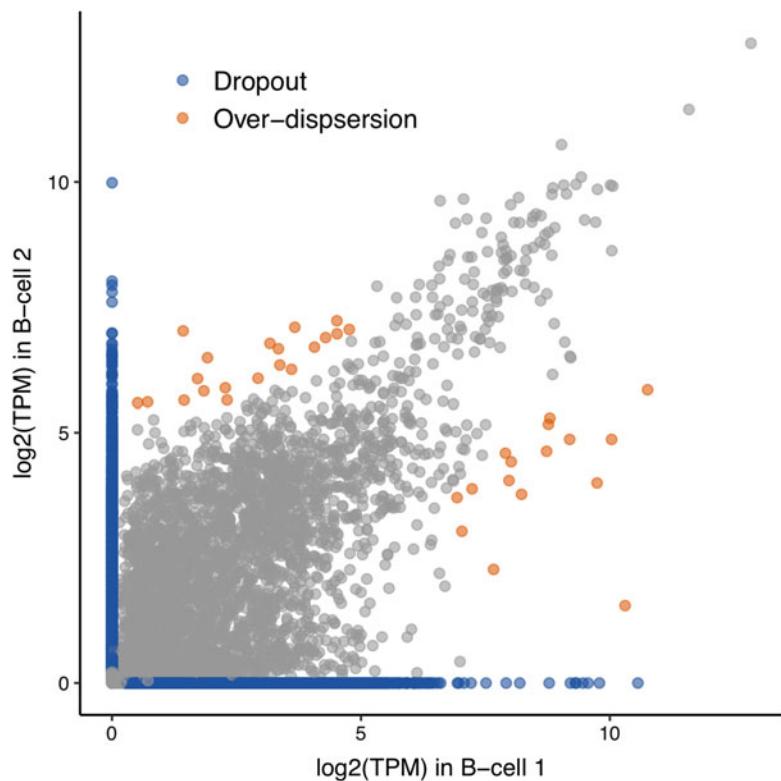
[101]. Therefore, it is essential to apply quality control (QC) to identify the low-quality droplets/barcodes/cells which ought to be excluded from downstream analyses [97, 101–105].

A common QC metric for scRNA-seq is the number of transcripts/UMIs detected per droplet. A small number of transcripts detected per barcode are often an indicator for poor droplet capture, which can be caused by cell death and/or capture of random floating RNA molecules released by dead cells. Inversely, a considerable large number of transcripts with the same barcode can suggest doublets or floating RNA encapsulated together with a living cell. Percentage of mitochondrial transcripts is another common QC metric. A high number of mitochondrial transcripts suggest the cells might be undergoing stress, for example, from cell isolation and sorting process. It is advised to remove these cells since stress level is usually not the interests of scRNA-seq analysis. In addition to the cell-level QC, gene-level QC is often performed to exclude genes expressed in only a very small proportion of cells. Removing such genes can decrease technical noise, and speed up the downstream normalization and clustering.

When deciding on the cutoffs for scRNA-seq QC, it is important to take into the consideration the cell compositions of the samples being analyzed. Different cell types actively express different number of genes, different number of mitochondrial transcripts, and different sets of transcripts, especially when comparing tumor to normal cells. A too stringent cutoff can remove a cell population of interests if this population is rare in sample of cells. Therefore, we suggest that if the researchers have no prior knowledge about the cell compositions, a possible solution is to carry out initial cell type identification and use that information to guide the QC process.

#### **4.2 Drop-Outs, Normalization, and Spike-Ins**

When analyzing scRNA-Seq data, normalization is a critical step to adjust for unwanted biological effects and technical noise collectively known as “batch effects” that mask the real signal. Similar to bulk RNA-seq, scRNA-seq batch effects can come from the variations in handling protocols, library preparation, sequencing platforms, and sequencing depth. In addition to these variations commonly seen in bulk RNA-seq, a prominent characteristic of scRNA-seq data is zero inflation, where the expression count matrix of single cells is mostly filled with zeros [106–108] (Fig. 5). There are two sources of zero inflation: (1) biological reason—the real zeros, where the cells are in transient state of transcript bursting [109] or the genes simply do not express in a subpopulation of cells; (2) technical reason—dropout, which is caused by the inefficiency of mRNA capture such that a large percentage of mRNA molecules are not captured and consequently not sequenced. Besides the dropout events, individual cells can show stronger overdispersion than typically observed in bulk RNA-seq data,



**Fig. 5** Cell-to-cell variability observed in scRNA-seq data of two single B cells

even for genes with median-to-high expression levels [108]. Furthermore, compared to bulk RNA-seq, scRNA-seq data is much more heterogeneous since the sequenced cells are usually of different populations, cell types, or statuses. Even the cells from the same population but undergoing different cell cycles can show very different expression profiles [110, 111]. To uncover the cellular heterogeneity, scRNA-seq studies often start with choosing the HVG (highly variable genes) that are most informative in distinguishing cell populations. It has been shown that the choice of HVG is highly affected by normalization [112, 113]. Therefore, it is important for scRNA-seq normalization to retain the cell-to-cell biological heterogeneity while removing the cell-specific noise at the same time.

#### 4.2.1 Normalization Methods

The global-scaling normalization methods inherited from bulk RNA-seq analysis have been widely used in scRNA-seq. However, these methods cannot accommodate the cell-specific variability in scRNA-seq data and can lead to biased estimation of scaling factors [113–115]. To optimize the modeling of the cell-to-cell variability, multiple normalization methods specifically tailored for scRNA-seq data have been developed. Table 3 provides a detailed summary of their main features, statistical models, and special considerations

**Table 3**  
**Normalization methods used in scRNA-seq data analysis**

Method	Features	Statistical models	Notes on biological variation	Notes on technical variation	Other factors
SCDE [108]	A two-component mixture model is used to capture dropout events and amplification events; Differential expression is evaluated by Bayesian approach	1. Dropouts: Poisson 2. Amplification: Negative binomial		Cell-specific technical variation is modeled by spikes-ins	Cell cycle and cell size are considered as covariates
TASC [116]	An empirical Bayes approach models the cell-specific dropout rates and amplification bias	1. Biological variation: Poisson 2. Technical variation: Logistic regression	Fraction of genes detected in each cell (CDR) is used as proxy for both biological and technical variation	Cell size is captured by CDR	
MAST [106]	Two-part generalized linear regression models dropout and amplification events	1. Probability of detection: logistic regression 2. Expression level: Gaussian			
BASICS [117]	Use integrated Bayesian hierarchical model to simultaneously quantify unexplained technical noise and cell-to-cell biological heterogeneity	1. Expression: Poisson 2. Random effect and size-specific factor: Gamma	Biological cell-specific variability is quantified by gene-specific parameters borrowing information across all cells	Technical variability is quantified based on spike-ins	Cell cycle and cell size are captured
SCALE [109]	Allele-specific transcription expression is modeled by a systematic statistical framework	1. Alleles expression status: empirical Bayes method 2. Allele-specific transcription kinetics: Poisson-Beta hierarchical model 3. Allelic difference: resampling-based test	Requires allele-specific read counts to start with	Technical variability is adjusted through spike-ins	Spike-ins/total reads serves as a proxy for cell size

(continued)

**Table 3**  
**(continued)**

<b>Method</b>	<b>Features</b>	<b>Statistical models</b>	<b>Notes on biological variation</b>	<b>Notes on technical variation</b>	<b>Other factors</b>
scran [107]	Deconvolution frameworks estimates cell-specific factor by pooling multiple cells to deal with zero inflation	Quantile regression	A ring arrangement by library size and sliding window is used to select random pool of cells; pool-based scaling factors are then deconvolved to yield cell-specific factors		
SCnorm [115]	Use quantile regression to estimate gene-specific scaling factors based on sequencing depth	Groups genes with similar sequencing depth and then estimate scale factors within each group; Spiked-ins are not required, but can be helpful	Hierarchical Dirichlet Process mixture model (DPMM)	The model simultaneously estimates the heterogeneous clusters through the DPMM and infers the technical variation parameters for imputing dropouts	
BISCUIT [118]	Employs Bayesian inference to cluster single cells considering both biological and technical variation in parallel				

when modeling cell-specific biological and technical variations. In the following sections, we will briefly discuss the key characteristics of these methods.

The first category we consider detects differential expressions among cells while adjusting technical variations that are specific to scRNA-seq data. This category includes SCDE [108], TASC [116], and MAST [106], which all employs empirical Bayesian frameworks to estimate dropout events and real amplification of transcripts. More specifically, TASC takes the cell size and cell cycle as covariates, where the former is estimated from the ratio of endogenous RNA reads to spike-ins and the latter is represented by the expression of curated genes [20]. Alternatively, MAST uses a fraction of genes detected in each cell as a proxy for technical and biological variation. Although most of these methods focus on the differential expression at the gene level, some go beyond and study the difference in allelic expression. One such example is SCALE, which models the allele-specific transcription bursting with both technical variation and cell size differences accounted for. However, SCALE requires input as allele-specific read counts at heterozygous loci, which can be challenging for many scRNA-seq platforms, especially the tag-based quantification methods including Dropseq and 10X.

Another major category is to generate normalized gene expression matrix that can be used as input for downstream analysis. Based on how the scaling factors are modeled, we further stratify these methods into two groups: cell-specific methods including scran [107] and BASiCS [117], and gene-specific methods including SCnorm [115]. BASiCS estimates the cell-specific biological variations by borrowing information across all cells and all genes while quantifies the technical variations relying on spike-ins. Alternatively, scran first clusters the cells into more homogenous groups and then deconvolutes the pooled cells to yield cell-specific factors. On the other hand, SCnorm groups genes with a similar dependence on sequencing depth and then estimates the scale factor within each gene group.

In addition to above approaches that focus on differential expression and normalized gene expression matrix, which are concepts adopted from bulk RNA-seq, several methods have been developed to specifically target the downstream heterogeneity studies in scRNA-seq data. BISCUIT [118] is a cell-type dependent normalization which uses a Bayesian probabilistic model to iteratively normalize and cluster cells. It simultaneously assigns cells to cluster and learns cell-dependent parameters within each cluster. The inferred parameters are then used to generate cell-type dependent normalization that can be fed back to improve clustering. It has been shown that BISCUIT can identify more refined subtypes of cells than global normalization methods [119].

#### 4.2.2 Drop-Out Imputation

Instead of directly normalizing endogenous genes, many chose to impute the drop-out events prior to normalization. There are mainly two strategies used in scRNA-seq data imputation:

1. To distinguish the biological zeros from technical zeros using models of expected gene expression, which is usually obtained either from borrowing information across cells or from spike-in sequences. These methods include DrImpute [120], SAVER [121], McImpute [122], scImpute [123], ALRA [124], and scRMA [125].
2. To reduce the noise by using information from neighboring data. This category includes MAGIC [126], netSmooth [127], and knn-smooth [128].

Although imputation can rescue missing information that is important to study cellular heterogeneity, concerns have been raised about their sensitivity and specificity. A positive-control based benchmarking study assessed these methods and concluded that most methods only provide small improvement [129]. Andrews and Hemberg [130] evaluated the false discovery rate of these methods using negative controls. They found that SAVER performed well on simulated data compared to others; however, all methods introduced false signals at various levels in the permuted real data. These limitations of imputing scRNA-seq data are probably due to the lack of a comprehensive and independent reference, for example, as in GWAS imputation. Until such reference is generated, caution should be used when imputing scRNA-seq data.

#### 4.2.3 Spike-Ins

As shown in Table 3, many scRNA-seq normalization methods rely on the spike-ins to estimate the cell-specific technical variations. Spikes-ins are a set of synthetic RNA sequences added to the samples in a theoretically constant and known amount, in order to calibrate the gene measurement and distinguish the biological vs. technical variations in RNA-seq experiments. The most commonly used spike-ins are the 92 External RNA Control Consortium (ERCC) molecules [46] and the Spike-in RNA variant control mixes (SIRVs, Lexogen). These extrinsic control sequences have also been used in scRNA-seq experiments [131–133], where the spike-ins with different concentrations are added with a constant amount across all cells. Vallejos et al. [113] and Lun et al. [134] have discussed the benefit of extrinsic control sequencing in scRNA-seq. However, the use of spike-ins remains challenging.

Although often neglected, calibrating the amount of spike-ins sequences is critical and should depend on the endogenous mRNA content [50]. However, due to the large and unknown heterogeneous among tumor microenvironment, it is difficult to obtain prior knowledge about the cell-type specific endogenous mRNA content before sequencing. In addition, the spike-in sequences do not reflect the gene-length and GC content in the mammalian

transcriptome, such that the technical effects may be different for the extrinsic and intrinsic genes [113]. Moreover, it has been shown that spike-ins signals can vary across technical replicates [50], and only partial spike-in sequences can be actually sequenced and aligned [113]. Furthermore, the use of spike-ins in the recent developed large-scale droplet-based platforms is not as cost-effective as in small scale platforms. For example, to reduce the doublet rate in Chromium 10X, the percentage of cell-containing droplets is deliberately designed as low as 1–10%. The spike-ins are added evenly across all droplets, not just the cell-containing ones and consequently takes up the vast majority of sequencing reads. Due to these limitations and challenges, caution should be used when employing spike-ins for technical variation estimations. Additionally, efforts should be made to design spike-ins sequences accommodating the unique characteristics of scRNA-seq experiments.

As discussed above, normalization methods that are specially tailored for scRNA-seq are theoretically and operationally superior over the global-scaling normalization inherited from bulk RNA-seq. However, these methods vary substantially in terms of their assumptions and their models, where none of them outperform others under all scenarios in the performance assessment [114, 135]. Great efforts are actively underway to develop more efficient and robust normalization for scRNA-seq.

#### **4.3 Data Integration and Batch Correction**

Due to the complexity of scRNA-seq experiments, it is often difficult for a study to process all samples at the same time and/or using the same protocols. In such situations, it is necessary to integrate samples of different batches or even of different scRNA-seq platforms. Due to the unique data structure of scRNA-seq, batch correction methods designed for bulk RNA-seq are not suitable. Several approaches have been recently proposed to deal with sample-level batches in scRNA-seq data. kBET (k-nearest neighbor correct effect) [136] quantifies the sample-level batch effects using a  $\chi^2$ -based test. MNNs (mutual nearest neighbor) [137] corrects the effect using only a subset of populations shared between batches and has been implemented in scran [102] as the mnnCorrect function. The Seurat group [138] uses cell pairwise correspondences between single cells across datasets, termed as “anchors,” to integrate gene expressions across technologies and batches. All these approaches have been shown to have better performance than bulk RNA-seq batch-correction methods.

#### **4.4 Dimension Reduction, Clustering, and Cell Type Identification**

One of the most popular uses of scRNA-seq is to identify and characterize cell types within the heterogeneous tissues or samples. The de novo identification of putative cell types has been considered as an unsupervised clustering problem. In this section, we will discuss the main classes of clustering methods having been applied

to scRNA-seq, as well as the remaining issues and challenges. We will also briefly touch on the supervised and semisupervised clustering.

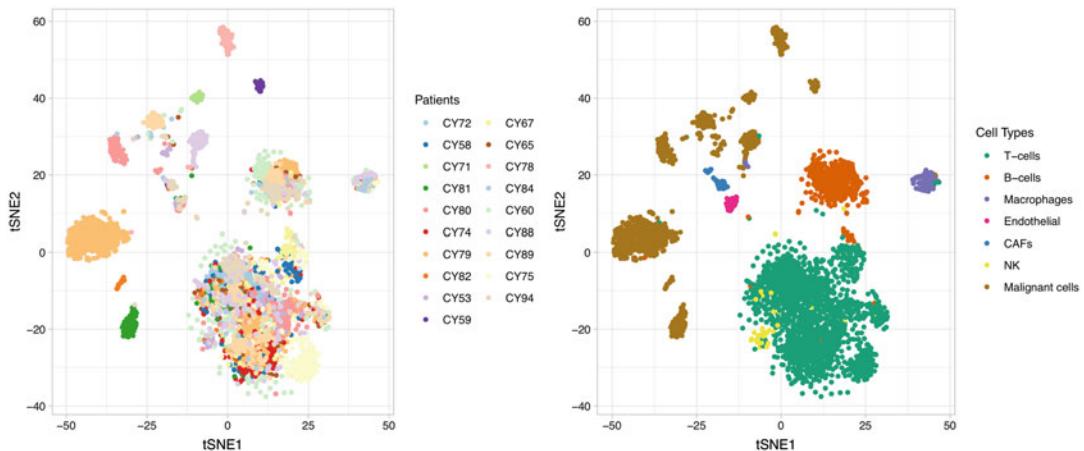
#### *4.4.1 Dimension Reduction and Feature Selection*

The high dimensional transcriptome data generated by scRNA-seq provides tremendous information for uncovering the biology of cells. However, it also introduces challenges to statistical analysis which is often referred to as the “curse of dimensionality.” With a large number of genes measured in scRNA-seq, the distance between individual cells can become small and thus make it difficult to distinguish between-population differences and within-population differences [139]. The two main approaches to deal with the issue of high dimensionality are feature selection and dimension reduction.

Feature selection removes the uninformative genes in terms of their ability to distinguish cells, in order to reduce the dimensions used in analysis and speed up calculations. The most commonly used feature selection in scRNA-seq is to select the highly variable genes (HVG), assuming that genes with high variance are more likely due to biological signals rather than technical noise [140]. The normalization and usage of spike-ins can facilitate the selection of HVG as we discussed in the previous section. Other feature selection approaches include identifying biological relevant genes based on expression correlation between cells [97, 141, 142] and using the magnitude and/or significance of the correlation to select genes.

Dimension reduction, on the other hand, completes a projection of the gene expression data onto a lower dimensional space. There are many generic dimension reduction methods that can be applied to any high dimensional data, including principal component analysis (PCA), independent components analysis (ICA) [143], Laplacian eigenmaps [144], and t-distributed stochastic neighbor embedding (t-SNE) [97, 145]. In this section we will focus on PCA and t-SNE due to their popularity in scRNA-seq data analysis. PCA uses the orthogonal transformations to project the gene count matrix onto a reduced number of linear independent dimensions called principal components. The advantage of PCA is that it is relatively fast and can preserve the distance information among cells. However, PCA is restricted to linear combinations of variables, which can be inappropriate in the context of scRNA-seq. In particular, it has been reported that the first components are often related to the number of genes detected per cell rather than the biological signals [106, 146]. To improve PCA, Risso et al. [147] proposed ZINB-WaVE which uses a zero-inflated negative binomial to deal with the dropouts in scRNA-seq data.

T-distributed stochastic neighbor embedding (t-SNE) is a stochastic method that reduces high dimensions to two or three



**Fig. 6** t-SNE projection of single cells from melanoma tumors colored by patient origin (left) and cell type (right)

embeddings while preserving the local structure among cells, such that neighbor cells stay close and distant cells remain distant. Due to the probability distribution used in embedding estimation, t-SNE follows two rules: (1) all points (cells) repel each other, and (2) each point (cell) attracts its nearest neighbors [148]. Therefore, t-SNE can specifically project cells into more distantly isolated clusters, making it almost the standard choice in visualization exploration of scRNA-seq data. Figure 6 projects 4645 single cells extracted from 19 melanoma tumors [20] onto t-SNE 2D planes. The clusters formed by t-SNE are in good agreement with the cell types identified by the original study, and show a high degree of intra-tumor heterogeneity for malignant cells but not for immune/stromal cells. However, t-SNE has its limitations. First, t-SNE is computationally expensive. This is particularly problematic in large-scale scRNA-seq studies which require analyzing hundreds of thousands of cells simultaneously. In addition, although t-SNE captures the local structure it often fails to preserve the global geometry of the data. When t-SNE places cells into distinctive clusters, the relative position of these clusters is almost arbitrary and with little biological meanings [148]. Moreover, the embedding is governed by a parameter “perplexity,” which controls the number of nearest neighbors each point is attracted to. Different perplexity choices can lead to different degrees of separation and the judgment of the appropriate perplexity is subject to the analysts. Several reviews have provided in-depth discussion and practical suggestions for using t-SNE in scRNA-seq analysis [148–150]. We would recommend readers consult with these reviews before making conclusions with t-SNE results.

More recently, a nonlinear dimension reduction method, uniform manifold approximation and projection (UMAP) [151], implemented in R package “umap,” was developed as an alternative

to t-SNE. It is claimed to preserve as much of the local structure and more of the global geometry with a shorter run time than t-SNE. Becht et al. [152] performed a systematic evaluation using well-annotated scRNA-seq data, and concluded that UMAP provided faster run times, higher reproducibility, and more meaningful cell clusters. Recognizing its advantage, several scRNA-seq analysis tools incorporate UMAP into their standard dimension reduction and visualization pipeline. Feature selection and dimension reduction are not necessarily mutually exclusive. As a matter of fact, dimension reductions are susceptible to the batch effects caused by cell-specific technical variations [106, 153]. Performing feature selection that removes genes with little biological relevant prior to dimension reduction can greatly reduce the technical noise [154]. The combination of these methods has been widely adopted by scRNA-seq analysis pipelines.

#### 4.4.2 Unsupervised Clustering

Recent single-cell sequencing technologies enable the study of tumor or TME heterogeneity at a single cell level [155, 156]. One of the ongoing challenges on scRNA-seq data analyses is that cell type recognition or subpopulation classification since the tumor–stromal interaction has been shown to be important. The diverse cell types would often be visualized in t-SNE space (Fig. 6) or other similar linear or nonlinear projections. Unsupervised clustering is a central component of scRNA-seq analysis, as it can identify cell populations and thus strongly affects any downstream analysis. Although many methods have been developed and applied to scRNA-seq, clustering and the interpretation of clusters are still facing great biological and computational challenges. Kiselev et al. [139] and Andrews and Hemberg [154] reviewed the commonly used clustering methods in the context of scRNA-seq and summarized their advantages and limitations. Duo et al. [157] systematically evaluated 15 clustering methods using simulated data and found substantial differences in their performances. In this section, we briefly review the four main classes of clustering methods.

K-means clustering iteratively assigns cells to the nearest cluster center and recomputes the new center. Although this method is fast, it requires a predetermined number of clusters and assumes the clusters are of equal sizes, which can be easily violated in TME studies. Tools that implement K-means include SC3 [158], SIMLR [159], RaceID [160], and pcaReduce [161]. The next method is hierarchical clustering which sequentially merges cells into larger clusters (bottom-up) or divides clusters into smaller communities (top-down). This method is deterministic but more computational expensive than k-means. Many scRNA-seq tools have modified hierarchical clustering either to accommodate low-depth samples by adding imputation of zeros [162] or to improve identification of small clusters by iteratively performing

dimension reduction [161, 163]. Hierarchical clustering-based tools include CIDR [162], BackSPIN [163], pcaReduce [161], SINCERA [164], mpath [165], and ascend [166].

The third type of clustering methods is density-based, which can identify dense clusters without any assumption on the shape or size on the clusters. However, it assumes equal homozygosity (density) of the clusters, requires a predetermined density cutoff, and works better with datasets with a large number of cells (e.g., droplet-based scRNA-seq assays). Such methods include DBSCAN [167], GiniClust [168] which employs DBSCAN, and monocle [169]. Lastly, there is graph-based clustering, which identifies cells densely connected by edges. Compared to k-means and hierarchical clustering, graph-based clustering does not require any predefined parameters, makes the minimal assumption on cell populations, and can scale to a large number of cells [170, 171]. The most popular application of graph-based clustering combines k-nearest-neighbor graphs and Louvain community detection [171, 172]. However, the main drawback of graph-based clustering is that it heavily relies on how well the scRNA-seq is translated into graph space. Therefore, it is often necessary to perform dimension reduction or feature selection beforehand to boost the search for nearest neighbors. Graph-based clustering has been implemented in multiple tools including Seurat [97], Phenograph [173], densityCut [174], SNN-Clip [175], SACNPY [176], and MetaCell [177].

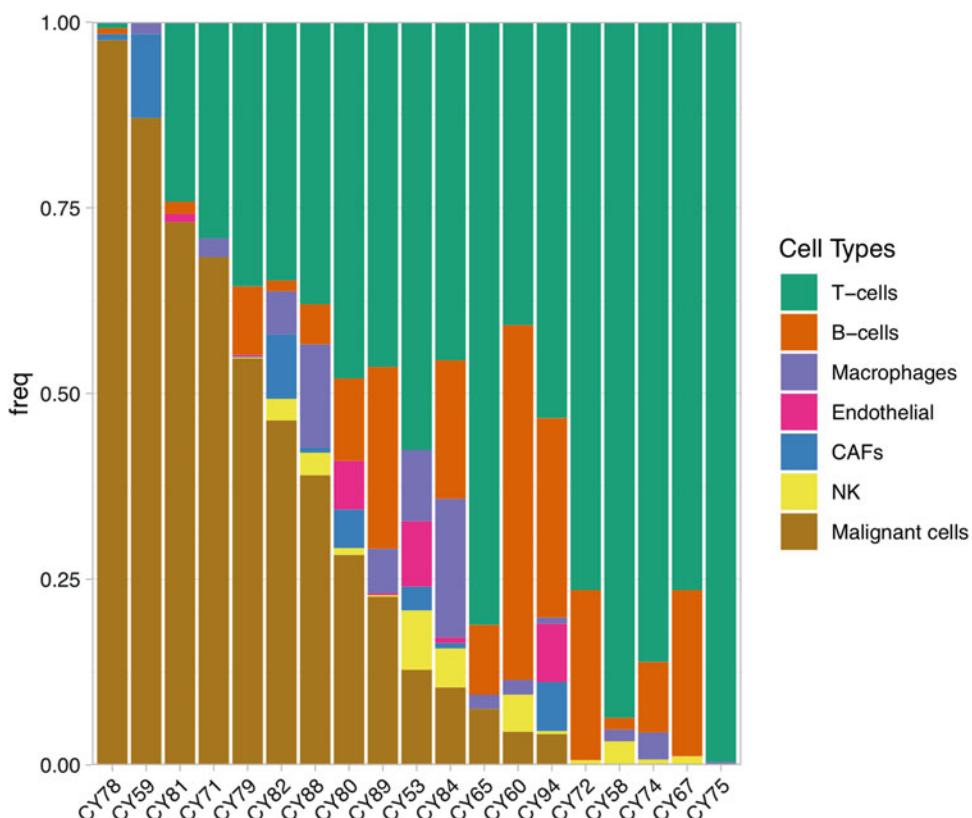
#### *4.4.3 Supervised Classifier and Cell Type Identification*

To characterize the sample heterogeneity, the groups of cells defined by unsupervised clustering are often assigned to different cell types based on enriched canonical markers. In cancer researches, the single-cell type identification has been focusing on the subpopulations of immune cells, stromal cells, and tumor-related cells present in TME and/or the circulating system. However, this canonical workflow has its limitations. First, as summarized above, each clustering method has its own drawbacks and different similarity metrics usually result in different cluster separations. Second, this process relies on the researchers' knowledge on the signature genes, and it can become arbitrary when making conclusions based on only a handful of genes. Only very recently, alternative methods have been proposed. SuperCT [178] is a supervised classifier (SC)-based machine learning method. It trains a nonlinear SC from bulk and single-cell RNA-seq data of pure cell types covering a wide range of immune and stromal cells, and then uses the trained classifiers to reveal cell types of any scRNA-seq data provided as new input. Another single-cell identifier is SingleR [179]. It constructs a reference database by collecting bulk RNA-seq data from over 1000 samples with pure cell types, and then determines the type of a single cell in scRNA-seq experiment by its Spearman correlation with each sample in the reference

database. Although these methods are still immature for application in cancer research due to the limited cancer-specific reference database, they defiantly opened up new avenues for cell type classification in scRNA-seq.

#### 4.5 Studying Heterogeneity Using scRNA-Seq

Tumor heterogeneity is commonly observed with wide range of infiltrations, as illustrated in Fig. 7 for 19 melanoma tumors from the Tirosh et al. study [20]. As the algorithms for cell type or subtype classification has been developed and improved in recent years. Some algorithms have been focused on how to quantify cellular heterogeneity. SinCHet estimates cellular heterogeneity using Shannon index over the all-possible clustering resolutions is developed to analyze cellular heterogeneity and characterize sub-population composition [180]. A recent paper proposes a general diversity index (GDI), a generalized form of ecological diversity index, to quantify heterogeneity on multiple scales and relate it to disease evolution [181]. The index takes the generalized from the low diversity order, the clonal richness, to intermediate diversity, Shannon or Simpson's indices, to higher order of diversity. The results showed that healthy individuals had lower diversity than cancer patients and little difference in diversity between pre- and post–bone marrow samples from AML patients.



**Fig. 7** Cellular composition of 19 melanoma tumors showing tumor heterogeneity

## 5 Conclusions

High-throughput sequencing (HTS) has revolutionized the study of the transcriptome and its relationship with disease. Two types of transcriptomic studies are now possible, bulk or single cell sequencing studies. With the advances in technology, many bioinformatics and statistical methods have been developed to process and analyses data from bulk sequencing (RNA-seq) and single-cell sequencing (scRNA-seq), with more methods currently being developed for scRNA-seq studies.

## References

1. Muller PA, Vousden KH (2013) p53 mutations in cancer. *Nat Cell Biol* 15(1):2–8. <https://doi.org/10.1038/ncb2641>
2. Baylin SB (2005) DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol* 2 (Suppl 1):S4–S11. <https://doi.org/10.1038/ncponc0354>
3. Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747–752. <https://doi.org/10.1038/35021093>
4. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61–70
5. Parker JS, Mullins M, Cheang MC et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27(8):1160–1167. <https://doi.org/10.1200/JCO.2008.18.1370> [pii]. *JCO.2008.18.1370* [pii]
6. Sorlie T, Perou CM, Tibshirani R et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98 (19):10869–10874. <https://doi.org/10.1073/pnas.191367098>
7. Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 100 (14):8418–8423. <https://doi.org/10.1073/pnas.0932692100>
8. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev* 10(1):57–63. <https://doi.org/10.1038/nrg2484>
9. Zhu S, Qing T, Zheng Y et al (2017) Advances in single-cell RNA sequencing and its applications in cancer research. *Oncotarget* 8(32):53763–53779. <https://doi.org/10.18632/oncotarget.17893>
10. Bian S, Hou Y, Zhou X et al (2018) Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* 362 (6418):1060–1063. <https://doi.org/10.1126/science.aao3791>
11. Navin NE (2015) Delineating cancer evolution with single-cell sequencing. *Sci Transl Med* 7(296):296fs229. <https://doi.org/10.1126/scitranslmed.aac8319>
12. Lee MC, Lopez-Diaz FJ, Khan SY et al (2014) Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc Natl Acad Sci U S A* 111(44):E4726–E4735. <https://doi.org/10.1073/pnas.1404656111>
13. Guo X, Zhang Y, Zheng L et al (2018) Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med* 24(7):978–985. <https://doi.org/10.1038/s41591-018-0045-3>
14. Zheng C, Zheng L, Yoo JK et al (2017) Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 169 (7):1342–1356.e1316. <https://doi.org/10.1016/j.cell.2017.05.035>
15. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. *CA Cancer J Clin* 69 (1):7–34. <https://doi.org/10.3322/caac.21551>
16. Cancer Genome Atlas Network (2015) Genomic classification of cutaneous melanoma. *Cell* 161(7):1681–1696. <https://doi.org/10.1016/j.cell.2015.05.044>
17. Nirschl CJ, Suarez-Farinás M, Izar B et al (2017) IFNgamma-dependent tissue-immune homeostasis is co-opted in the tumor microenvironment. *Cell* 170 (1):127–141.e115. <https://doi.org/10.1016/j.cell.2017.06.016>

18. Gerber T, Willscher E, Loeffler-Wirth H et al (2017) Mapping heterogeneity in patient-derived melanoma cultures by single-cell RNA-seq. *Oncotarget* 8(1):846–862. <https://doi.org/10.18632/oncotarget.13666>
19. Kumar MP, Du J, Lagoudas G et al (2018) Analysis of single-cell RNA-Seq identifies cell-cell communication associated with tumor characteristics. *Cell Rep* 25(6):1458–1468. e1454. <https://doi.org/10.1016/j.celrep.2018.10.047>
20. Tirosh I, Izar B, Prakadan SM et al (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352(6282):189–196. <https://doi.org/10.1126/science.aad0501>
21. Picelli S, Bjorklund AK, Faridani OR et al (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 10(11):1096–1098. <https://doi.org/10.1038/nmeth.2639>
22. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38(12):e131. <https://doi.org/10.1093/nar/gkq224>
23. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40(10):e72. <https://doi.org/10.1093/nar/gks001>
24. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85(8):2444–2448. <https://doi.org/10.1073/pnas.85.8.2444>
25. Cock PJ, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38(6):1767–1771. <https://doi.org/10.1093/nar/gkp1137>
26. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
27. Fuller CW, Middendorf LR, Benner SA et al (2009) The challenges of sequencing by synthesis. *Nat Biotechnol* 27(11):1013–1023. <https://doi.org/10.1038/nbt.1585>
28. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36. <https://doi.org/10.1186/gb-2013-14-4-r36>
29. Wang K, Singh D, Zeng Z et al (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 38(18):e178. <https://doi.org/10.1093/nar/gkq622>
30. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>
31. Wu TD, Reeder J, Lawrence M et al (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol* 1418:283–334. [https://doi.org/10.1007/978-1-4939-3578-9\\_15](https://doi.org/10.1007/978-1-4939-3578-9_15)
32. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21(6):936–939. <https://doi.org/10.1101/gr.111120.110>
33. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11):1851–1858. <https://doi.org/10.1101/gr.078212.108>
34. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> [pii]
35. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359. <https://doi.org/10.1038/nmeth.1923>
36. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515. <https://doi.org/10.1038/nbt.1621>
37. Pertea M, Pertea GM, Antonescu CM et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33(3):290–295. <https://doi.org/10.1038/nbt.3122>
38. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8):1494–1512. <https://doi.org/10.1038/nprot.2013.084>
39. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652. <https://doi.org/10.1038/nbt.1883>

40. Schulz MH, Zerbino DR, Vingron M et al (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28(8):1086–1092. <https://doi.org/10.1093/bioinformatics/bts094>
41. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. <https://doi.org/10.1186/1471-2105-12-323>
42. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* 32(5):462–464. <https://doi.org/10.1038/nbt.2862>
43. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7):923–930. <https://doi.org/10.1093/bioinformatics/btt656>
44. Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169. <https://doi.org/10.1093/bioinformatics/btu638>
45. Bullard JH, Purdom E, Hansen KD et al (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. <https://doi.org/10.1186/1471-2105-11-94>
46. Jiang L, Schlesinger F, Davis CA et al (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 21(9):1543–1551. <https://doi.org/10.1101/gr.121095.111>
47. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628. <https://doi.org/10.1038/nmeth.1226> [pii]
48. Leek JT, Scharpf RB, Bravo HC et al (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev* 11(10):733–739. <https://doi.org/10.1038/nrg2825>
49. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3(9):1724–1735. <https://doi.org/10.1371/journal.pgen.0030161>
50. Risso D, Ngai J, Speed TP et al (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 32(9):896–902. <https://doi.org/10.1038/nbt.2931>
51. Hansen KD, Wu Z, Irizarry RA et al (2011) Sequencing technology does not eliminate biological variability. *Nat Biotechnol* 29(7):572–573. <https://doi.org/10.1038/nbt.1910>
52. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106. <https://doi.org/10.1186/gb-2010-11-10-r106>
53. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11(3):R25. <https://doi.org/10.1186/gb-2010-11-3-r25>
54. Smyth GK (2005) limma: linear models for microarray data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds) *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, Berlin, pp 397–420
55. Bolstad BM, Irizarry RA, Astrand M et al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193
56. Pickrell JK, Marioni JC, Pai AA et al (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768–772
57. Li B, Ruotti V, Stewart RM et al (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26(4):493–500. <https://doi.org/10.1093/bioinformatics/btp692>
58. Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131(4):281–285. <https://doi.org/10.1007/s12064-012-0162-3>
59. Conesa A, Madrigal P, Tarazona S et al (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol* 17:13. <https://doi.org/10.1186/s13059-016-0881-8>
60. Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 4:14. <https://doi.org/10.1186/1745-6150-4-14>
61. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127. <https://doi.org/10.1093/biostatistics/kxj037>
62. Karpievitch YV, Nikolic SB, Wilson R et al (2014) Metabolomics data normalization

- with EigenMS. *PLoS One* 9(12):e116221. <https://doi.org/10.1371/journal.pone.0116221>
63. Tracy CA, Widom H (1994) Level spacing distributions and the Bessel kernel. *Commun Math Phys* 161(2):289–309
  64. Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* 29(2):295–327
  65. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190. <https://doi.org/10.1371/journal.pgen.0020190>
  66. Abbas-Aghababazadeh F, Li Q, Fridley BL (2018) Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS One* 13(10):e0206312. <https://doi.org/10.1371/journal.pone.0206312>
  67. Wang L, Feng Z, Wang X et al (2010) DEG-seq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26(1):136–138. <https://doi.org/10.1093/bioinformatics/btp612>
  68. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* 11(8):R83. <https://doi.org/10.1186/gb-2010-11-8-r83>
  69. Li J, Witten DM, Johnstone IM et al (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13(3):523–538. <https://doi.org/10.1093/biostatistics/kxr031>
  70. Auer PL, Doerge RW (2011) A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol* 10(1):Article 26
  71. Srivastava S, Chen L (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* 38(17):e170. <https://doi.org/10.1093/nar/gkq670>
  72. Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23(21):2881–2887. <https://doi.org/10.1093/bioinformatics/btm453>. btm453 [pii]
  73. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140. <https://doi.org/10.1093/bioinformatics/btp616>. btp616 [pii]
  74. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>
  75. Di Y, Schafer DW, Cumbie JS et al (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol* 10(1):24
  76. Zhou YH, Xia K, Wright FA (2011) A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 27(19):2672–2678. <https://doi.org/10.1093/bioinformatics/btr449>
  77. Van De Wiel MA, Leday GG, Pardo L et al (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* 14(1):113–128. <https://doi.org/10.1093/biostatistics/kxs031>
  78. Hardcastle TJ, Kelly KA (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422. <https://doi.org/10.1186/1471-2105-11-422>
  79. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article 3. <https://doi.org/10.2202/1544-6115.1027>
  80. Ritchie ME, Phipson B, Wu D et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47. <https://doi.org/10.1093/nar/gkv007>
  81. Law CW, Chen Y, Shi W et al (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15(2):R29. <https://doi.org/10.1186/gb-2014-15-2-r29>
  82. Li J, Tibshirani R (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* 22(5):519–536. <https://doi.org/10.1177/0962280211428386>
  83. Tarazona S, Garcia-Alcalde F, Dopazo J et al (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res* 21(12):2213–2223. <https://doi.org/10.1101/gr.124321.111>
  84. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 57(1):289–300
  85. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100(16):9440–9445
  86. Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc B Methodol* 64(Pt. 3):479–498

87. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. *BMJ* 310(6973):170. <https://doi.org/10.1136/bmj.310.6973.170>
88. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70
89. Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800–802
90. Newman AM, Liu CL, Green MR et al (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 12 (5):453–457. <https://doi.org/10.1038/nmeth.3337>
91. Thorsson V, Gibbs DL, Brown SD et al (2018) The immune landscape of cancer. *Immunity* 48(4):812–830.e814. <https://doi.org/10.1016/j.immuni.2018.03.023>
92. Li T, Fan J, Wang B et al (2017) TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res* 77(21):e108–e110. <https://doi.org/10.1158/0008-5472.CAN-17-0307>
93. Aran D, Hu Z, Butte AJ (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 18(1):220. <https://doi.org/10.1186/s13059-017-1349-1>
94. Hashimshony T, Wagner F, Sher N et al (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2 (3):666–673. <https://doi.org/10.1016/j.celrep.2012.08.003>
95. Islam S, Zeisel A, Joost S et al (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 11(2):163–166. <https://doi.org/10.1038/nmeth.2772>
96. Picelli S, Faridani OR, Bjorklund AK et al (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9 (1):171–181. <https://doi.org/10.1038/nprot.2014.006>
97. Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161(5):1202–1214. <https://doi.org/10.1016/j.cell.2015.05.002>
98. Zheng GX, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8:14049. <https://doi.org/10.1038/ncomms14049>
99. Ziegenhain C, Vieth B, Parekh S et al (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 65 (4):631–643.e634. <https://doi.org/10.1016/j.molcel.2017.01.023>
100. Svensson V, Natarajan KN, Ly LH et al (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14 (4):381–387. <https://doi.org/10.1038/nmeth.4220>
101. Ilicic T, Kim JK, Kolodziejczyk AA et al (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* 17:29. <https://doi.org/10.1186/s13059-016-0888-1>
102. Lun AT, McCarthy DJ, Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 5:2122. <https://doi.org/10.12688/f1000research.9501.2>
103. Satija R, Farrell JA, Gennert D et al (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33 (5):495–502. <https://doi.org/10.1038/nbt.3192>
104. Zhao C, Hu S, Huo X et al (2017) Dr.seq2: a quality control and analysis pipeline for parallel single cell transcriptome and epigenome data. *PLoS One* 12(7):e0180583. <https://doi.org/10.1371/journal.pone.0180583>
105. McCarthy DJ, Campbell KR, Lun AT et al (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33(8):1179–1186. <https://doi.org/10.1093/bioinformatics/btw777>
106. Finak G, McDavid A, Yajima M et al (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16:278. <https://doi.org/10.1186/s13059-015-0844-5>
107. Lun AT, Bach K, Marioni JC (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 17:75. <https://doi.org/10.1186/s13059-016-0947-7>
108. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods* 11 (7):740–742. <https://doi.org/10.1038/nmeth.2967>
109. Jiang Y, Zhang NR, Li M (2017) SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* 18(1):74. <https://doi.org/10.1186/s13059-017-1200-8>
110. Liu Z, Lou H, Xie K et al (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun* 8(1):22. <https://doi.org/10.1038/s41467-017-0039-z>

111. McDavid A, Finak G, Gottardo R (2016) The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat Biotechnol* 34(6):591–593. <https://doi.org/10.1038/nbt.3498>
112. Wang J, Huang M, Torre E et al (2018) Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc Natl Acad Sci U S A* 115(28):E6437–E6446. <https://doi.org/10.1073/pnas.1721085115>
113. Vallejos CA, Risso D, Scialdone A et al (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 14(6):565–571. <https://doi.org/10.1038/nmeth.4292>
114. Cole MB, Risso D, Wagner A et al (2019) Performance assessment and selection of normalization procedures for single-cell RNA-Seq. *Cell Syst* 8(4):315–328.e318. <https://doi.org/10.1016/j.cels.2019.03.010>
115. Bacher R, Chu LF, Leng N et al (2017) SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* 14(6):584–586. <https://doi.org/10.1038/nmeth.4263>
116. Jia C, Hu Y, Kelly D et al (2017) Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res* 45(19):10978–10988. <https://doi.org/10.1093/nar/gkx754>
117. Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 11(6):e1004333. <https://doi.org/10.1371/journal.pcbi.1004333>
118. Prabhakaran S, Azizi E, Carr A et al (2016) Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *JMLR Workshop Conf Proc* 48:1070–1079
119. Azizi E, Prabhakaran S, Carr A et al (2017) Bayesian inference for single-cell clustering and imputing. *Genomics Comput Biol* 3(1):e46. <https://doi.org/10.18547/gcb.2017.vol3.iss1.e46>
120. Gong W, Kwak IY, Pota P et al (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 19(1):220. <https://doi.org/10.1186/s12859-018-2226-y>
121. Huang M, Wang J, Torre E et al (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 15(7):539–542. <https://doi.org/10.1038/s41592-018-0033-z>
122. Mongia A, Sengupta D, Majumdar A (2019) McImpute: matrix completion based imputation for single cell RNA-seq data. *Front Genet* 10:9. <https://doi.org/10.3389/fgene.2019.00009>
123. Li WV, Li JJ (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 9(1):997. <https://doi.org/10.1038/s41467-018-03405-7>
124. Linderman GC, Zhao J, Kluger Y (2018) Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*:397588. <https://doi.org/10.1101/397588>
125. Chen C, Wu C, Wu L et al (2018) scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. *bioRxiv*:459404. <https://doi.org/10.1101/459404>
126. van Dijk D, Sharma R, Nainys J et al (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell* 174(3):716–729.e727. <https://doi.org/10.1016/j.cell.2018.05.061>
127. Ronen J, Akalin A (2018) netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res* 7:8. <https://doi.org/10.12688/f1000research.13511.3>
128. Wagner F, Yan Y, Yanai I (2017) K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*:217737. <https://doi.org/10.1101/217737>
129. Zhang L, Zhang S (2018) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans Comput Biol Bioinform*. <https://doi.org/10.1109/TCBB.2018.2848633>
130. Andrews TS, Hemberg M (2018) False signals induced by single-cell imputation. *F1000Res* 7:1740. <https://doi.org/10.12688/f1000research.16613.2>
131. Buettner F, Natarajan KN, Casale FP et al (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33(2):155–160. <https://doi.org/10.1038/nbt.3102>
132. Katayama S, Tohonen V, Linnarsson S et al (2013) SAMstr: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29(22):2943–2945. <https://doi.org/10.1093/bioinformatics/btt511>
133. Ding B, Zheng L, Zhu Y et al (2015) Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31(13):2225–2227. <https://doi.org/10.1093/bioinformatics/btv122>

134. Lun ATL, Calero-Nieto FJ, Haim-Vilmovsky L et al (2017) Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res* 27(11):1795–1806. <https://doi.org/10.1101/gr.222877.117>
135. Vieth B, Parekh S, Ziegenhain C et al (2019) A systematic evaluation of single cell RNA-Seq analysis pipelines: library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies. *bioRxiv*:583013. <https://doi.org/10.1101/583013>
136. Buttner M, Miao Z, Wolf FA et al (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 16(1):43–49. <https://doi.org/10.1038/s41592-018-0254-1>
137. Haghverdi L, Lun ATL, Morgan MD et al (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36(5):421–427. <https://doi.org/10.1038/nbt.4091>
138. Stuart T, Butler A, Hoffman P et al (2018) Comprehensive integration of single cell data. *bioRxiv*:460147. <https://doi.org/10.1101/460147>
139. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev* 20(5):273–282. <https://doi.org/10.1038/s41576-018-0088-9>
140. Brennecke P, Anders S, Kim JK et al (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 10(11):1093–1095. <https://doi.org/10.1038/nmeth.2645>
141. Fan J, Salathia N, Liu R et al (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 13(3):241–244. <https://doi.org/10.1038/nmeth.3734>
142. Usoskin D, Furlan A, Islam S et al (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat Neurosci* 18(1):145–153. <https://doi.org/10.1038/nn.3881>
143. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13(4–5):411–430
144. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15(6):1373–1396. <https://doi.org/10.1162/089976603321780317>
145. Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605
146. Hicks SC, Townes FW, Teng M et al (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19(4):562–578. <https://doi.org/10.1093/biostatistics/kxx053>
147. Risso D, Perraudeau F, Gribkova S et al (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 9(1):284. <https://doi.org/10.1038/s41467-017-02554-5>
148. Kobak D, Berens P (2018) The art of using t-SNE for single-cell transcriptomics. *bioRxiv*:453449. <https://doi.org/10.1101/453449>
149. Wattenberg M, Viegas F, Johnson I (2016) How to use t-SNE effectively. *Distill.pub*. <https://doi.org/10.23915/distill.00002>
150. Linderman GC, Rachh M, Hoskins JG et al (2019) Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* 16(3):243–245. <https://doi.org/10.1038/s41592-018-0308-4>
151. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv e-prints*
152. Becht E, McInnes L, Healy J et al (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37:38. <https://doi.org/10.1038/nbt.4314>. <https://www.nature.com/articles/nbt.4314#supplementary-information>
153. Tung PY, Blischak JD, Hsiao CJ et al (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* 7:39921. <https://doi.org/10.1038/srep39921>
154. Andrews TS, Hemberg M (2018) Identifying cell populations with scRNASeq. *Mol Asp Med* 59:114–122. <https://doi.org/10.1016/j.mam.2017.07.002>
155. Navin NE (2014) Cancer genomics: one cell at a time. *Genome Biol* 15(8):452. <https://doi.org/10.1186/s13059-014-0452-9>
156. Wang Y, Navin NE (2015) Advances and applications of single-cell sequencing technologies. *Mol Cell* 58(4):598–609. <https://doi.org/10.1016/j.molcel.2015.05.005>
157. Duo A, Robinson MD, Soneson C (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 7:1141. <https://doi.org/10.12688/f1000research.15666.2>

158. Kiselev VY, Kirschner K, Schaub MT et al (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 14(5):483–486. <https://doi.org/10.1038/nmeth.4236>
159. Wang B, Zhu J, Pierson E et al (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 14(4):414–416. <https://doi.org/10.1038/nmeth.4207>
160. Grun D, Lyubimova A, Kester L et al (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525(7568):251–255. <https://doi.org/10.1038/nature14966>
161. Zurauskienė J, Yau C (2016) pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17:140. <https://doi.org/10.1186/s12859-016-0984-y>
162. Lin P, Troup M, Ho JW (2017) CIDR: ultra-fast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 18(1):59. <https://doi.org/10.1186/s13059-017-1188-0>
163. Zeisel A, Munoz-Manchado AB, Codeluppi S et al (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–1142. <https://doi.org/10.1126/science.aaa1934>
164. Guo M, Wang H, Potter SS et al (2015) SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput Biol* 11(11):e1004575. <https://doi.org/10.1371/journal.pcbi.1004575>
165. Chen J, Schlitzer A, Chakarov S et al (2016) Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun* 7:11988. <https://doi.org/10.1038/ncomms11988>
166. Senabouth A, Lukowski SW, Alquicira Hernandez J et al (2017) *ascend*: R package for analysis of single cell RNA-seq data. *bioRxiv*:207704. <https://doi.org/10.1101/207704>
167. Ester M, Kriegel H-P, et al (1996) A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the Proceedings of the Second International Conference on Knowledge discovery and data mining, Portland, Oregon
168. Jiang L, Chen H, Pinello L et al (2016) Gini-Clust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* 17(1):144. <https://doi.org/10.1186/s13059-016-1010-4>
169. Trapnell C, Cacchiarelli D, Grimsby J et al (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32(4):381–386. <https://doi.org/10.1038/nbt.2859>
170. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* 105(4):1118–1123. <https://doi.org/10.1073/pnas.0706851105>
171. Blondel VD, Guillaume J-L, Lambiotte R et al (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008:10008
172. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. *Phys Rev E* 80(5):056117. <https://doi.org/10.1103/PhysRevE.80.056117>
173. Levine JH, Simonds EF, Bendall SC et al (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162(1):184–197. <https://doi.org/10.1016/j.cell.2015.05.047>
174. Ding J, Shah S, Condon A (2016) density-Cut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics* 32(17):2567–2576. <https://doi.org/10.1093/bioinformatics/btw227>
175. Xu C, Su Z (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31(12):1974–1980. <https://doi.org/10.1093/bioinformatics/btv088>
176. Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19(1):15. <https://doi.org/10.1186/s13059-017-1382-0>
177. Baran Y, Sebe-Pedros A, Lubling Y et al (2018) MetaCell: analysis of single cell RNA-seq data using k-NN graph partitions. *bioRxiv*:437665. <https://doi.org/10.1101/437665>
178. Xie P, Gao M, Wang C et al (2019) SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res*. <https://doi.org/10.1093/nar/gkz116>
179. Aran D, Looney AP, Liu L et al (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 20(2):163–172.

- <https://doi.org/10.1038/s41590-018-0276-y>
180. Li J, Smalley I, Schell MJ et al (2017) SinCHet: a MATLAB toolbox for single cell heterogeneity analysis in cancer. *Bioinformatics* 33(18):2951–2953. <https://doi.org/10.1093/bioinformatics/btx297>
181. Ferrall-Fairbanks MC, Ball M, Padron E et al (2019) Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity. *JCO Clin Cancer Informatics* 3:1–10. <https://doi.org/10.1200/cci.18.00074>
182. Yang X, Liu D, Liu F et al (2013) HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* 14:33. <https://doi.org/10.1186/1471-2105-14-33>
183. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7(2):e30619. <https://doi.org/10.1371/journal.pone.0030619>
184. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
185. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485. <https://doi.org/10.1186/1471-2105-11-485>
186. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111
187. Robertson G, Schein J, Chiu R et al (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* 7(11):909–912. <https://doi.org/10.1038/nmeth.1517>



# Chapter 10

## Investigating Inter- and Intrasample Diversity of Single-Cell RNA Sequencing Datasets

Meghan C. Ferrall-Fairbanks and Philipp M. Altrock

### Abstract

Tumor heterogeneity can arise from a variety of extrinsic and intrinsic sources and drives unfavorable outcomes. With recent technological advances, single-cell RNA sequencing has become a way for researchers to easily assay tumor heterogeneity at the transcriptomic level with high resolution. However, ongoing research focuses on different ways to analyze this big data and how to compare across multiple different samples. In this chapter, we provide a practical guide to calculate inter- and intrasample diversity metrics from single-cell RNA sequencing datasets. These measures of diversity are adapted from commonly used metrics in statistics and ecology to quantify and compare sample heterogeneity at single-cell resolution.

**Key words** Tumor heterogeneity, Single-cell RNA sequencing, Diversity index, Tumor progression, Cancer evolution

---

### 1 Introduction

Intratumor heterogeneity (ITH) is a major determinant of tumor progression, the evolution of resistance to therapy, and can fuel tumor evolution and the development of metastasis. ITH is present on multiple different levels, ranging from genetic [1] to epigenetic/cell phenotypic [2, 3] and metabolic [4] to microenvironmental heterogeneity [5]. Single-cell DNA and RNA sequencing have made it possible to identify ITH in a way that cannot be captured by bulk sample profiling [6, 7], because they can, in principle, characterize important differences or common features on the level of the individual cell. Estimating cellular heterogeneity by way of diversity and uncertainty about the identity of an individual in the context of others in a sample is thus an important task. One important quantitative method to assess heterogeneity is by calculating the degree of variation between individual entities, which can be achieved using the concept of a diversity index [8]. Here, we present a method to use single-cell RNA sequencing data and clustering algorithms to calculate a general diversity index

in order to estimate intratumor heterogeneity, and use it as a starting point for clinical correlations [9], or mathematical modeling [10, 11].

ITH is of clinical interest because it serves as a reservoir for therapeutic resistance and is likely a driver of clinical progression with single and combination therapies, when targeted therapies. The clinical implications of ITH have not been explored in all types of cancer, on all scales of heterogeneity. Further, it is unknown whether certain therapeutics could directly decrease ITH and thus serve to mitigate this critical resistance mechanism. The primary objective of this manuscript is to introduce a multiscale approach to measure ITH using single-cell RNA sequencing. This method can be applied downstream of a number of computational and statistical approaches that integrates scRNA-seq data, and will become an important step in the quest to generate foundational evidence that ITH as a relevant clinical factor in those cancer types that have been lacking behind in terms of describing and clinically assessing tumor heterogeneity. Eventually, it would be the goal to describe ITH such that it can be modified by, for example, epigenetic therapeutics that either increase or reduce it to avert rapid resistance evolution.

Single-cell RNA sequencing (scRNA-seq) can be used to estimate cellular diversity, especially in the context of intratumor heterogeneity. Novel scRNA-seq technologies have become a cost-effective method to identify transcriptomic changes at high resolution. Intratumor heterogeneity can be identified for many disease at various stages [12], and have the potential to bring about novel ways to understand tumor evolution [13]. We build our methodology on the fact that single cell transcriptome profiling of leukemias can directly measure intraleukemic heterogeneity (ILH). A scRNA-seq study in chronic myeloid leukemia (CML) has demonstrated that scRNA-seq was capable of segregating patients with discordant responses to targeted tyrosine kinase inhibitor therapy [14], and it was recently shown that scRNA-seq data-based cellular diversity quantification can segregate various other healthy and cancer states [15].

As a summary statistic for ILH we show how to calculate and use a diversity index often applied in ecology [16, 17], using the general nonspatial diversity index [18] called  ${}^qD$ . This approach considers diversity on all possible orders  $q$ , and allows to compare states according to specific diversity indices, which emerge as special cases (e.g.,  ${}^0D$ , or  ${}^2D$ ). The species (clonal) richness of a sample is given by  ${}^0D$ . The Simpson index, that is, the probability that any two cells are identical, emerges from  ${}^2D$ . Most notably, the Shannon index [19]—a measure of uncertainty about the state of the heterogeneous cell population estimated from a subsample of it—can be derived from the limit of  $q$  approaching the value of 1. These indices have been used previously to quantify cancer heterogeneity

[10, 11, 20]. This general representation of diversity allows flexibility in the choice of the optimal  $q$ , potentially tailored for its biological or clinical application.

### 1.1 Chapter Outline

This chapter describes one established framework for quantifying inter- and intrasample heterogeneity of single-cell RNA sequencing experiments, followed by an example previously described comparing these diversity metrics for acute myeloid leukemia (AML) patients to healthy donors and discussion of interpretation of these diversity metrics. The chapter concludes with notes about how robustness and error of these types of metrics. The outline for this chapter is as follows. Subheading 2 contains the computational materials, including R libraries, and example single-cell RNA sequencing datasets that can be used to gain an intuition of the method. In Subheading 3, we present a single-cell RNA sequencing quality control analysis, methods of data clustering, and the diversity score calculation pipeline. Subheading 4 contains a worked example of calculating a universal diversity metric, applied to an AML dataset of four patient samples. Finally, in Subheading 5 we present notes/discussion from the example R code.

## 2 Materials

The methods presented in this chapter are one way to calculate diversity metrics applied to FASTQ files generated from single-cell RNA sequencing experiments. The materials required include:

1. FASTQ dataset.
2. Cell Ranger Pipelines.
3. Statistical programming package R.
4. R libraries.
5. UMAP.

### 2.1 Count Matrix Generation

#### 2.1.1 Downloading and Installing Tools

Single-cell RNA sequencing (scRNA-seq) data is often exported or postprocessed into a FASTQ file. FASTQ files are the most common way scRNA-seq data is stored in publicly available datasets.

1. Download and install the Cell Ranger tar file on a Linux distribution from the 10X Genomics website (<https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation>).
2. Install *Seurat* (version 3.0.0) package available on CRAN in R on version 3.4 or greater [21, 22]. Additional packages that are useful to speed up computation, especially if running these analyses on a cluster include *future* to access multiple process for parallelizing the Seurat commands and *bigmemory*.

3. Install UMAP (and required dependencies) with python as described on GitHub (<https://github.com/lmcinnes/umap/blob/master/README.rst>) [23]. UMAP visualization for clustered scRNA-seq data can be performed in R using the Seurat package but requires that UMAP first be installed via python.

## **2.2 Preparing Data for Diversity Analysis**

1. Run *cellranger count* on each sample fastq file. This creates a number of output files, including a folder with the filtered feature-barcode matrix containing a MEX formatted counts matrix. *cellranger count* should be run on each sample's FASTQ file individually.
2. Once FASTQ files have been run through *cellranger count*, open R Studio and initiate libraries using *library(packagename)* for the following libraries: Matrix, dplyr, readr, rdetools, data.table, ggplots2, iterators, Seurat.
3. Import data into R using the *Read10X* and then the *CreateSeuratObject* commands. Use these two commands for each additional sample. To create a multisample dataset, use the *merge* command to merge the individual Seurat Objects.

## **3 Methods**

This method was adapted and expanded from the tutorials for analysis of single-cell RNA sequencing data from the developers of the Seurat package [21, 22] in R available at: <https://satijalab.org/seurat/>. These tools are not the only way to analyze single-cell RNA sequencing data, but are best practices that we have found useful in quantifying differences in diversity measures across different polyclonal and malignant populations.

### **3.1 Quality Control and Normalization of Count Matrix**

Raw datasets need to be corrected to remove batch effects. This can be done by assessing the distribution of genes captured per cell in the dataset and the individual dataset's distribution of mitochondrial gene.

Typical gene distribution cutoff to determine which cells to include in further analysis involved a lower limit of cells with at least 200 genes detected and an upper limit of genes detected as:

$$\text{Mean number of genes detected} \pm 2 \times (\text{standard deviation of genes detected})$$

These gene distribution cutoffs are an attempt to avoid counting doublet cells as distinct single-cells [24].

Typical mitochondrial DNA content upper cutoff is very problem specific. A general rule of thumb is to exclude cells with over 5% mitochondrial content. This cutoff is an attempt to exclude

cells that are dying and may not be capturing the biology researchers are interested in exploring [24]. This cutoff needs to be increased for cell- and disease-specific cases, for example, cardiac cells are known to have increased mitochondrial content and so the maximum mitochondrial content cutoff would need to be increased.

These cutoffs to correct for doublet and high mitochondrial content concerns are implemented with by using the *subset* function on the full Seurat object dataset.

Then researchers can normalize and scale the expression data using *NormalizeData* and *ScaleData* functions. One common normalization technique is to normalize the feature expression for each cell by the total expression. Scaling the data allows researchers to removed unwanted sources of variation, including RNA count information and mitochondrial content by shifting the expression level to have a mean around 0 and variance across cells of 1.

## **3.2 Cluster Detection**

### **3.2.1 Dimension Reduction**

Principal component analysis (PCA) is often performed on single-cell RNA sequencing datasets to identify the largest sources of variation in the dataset (i.e., the principal components) as well as that some clustering algorithms require dimension reduced datasets to be used with those algorithms. PCA is implemented on a Seurat object using the *RunPCA* function. Dimension reduction can be performed on the entire scaled dataset or on the subset of variable feature gene set.

### **3.2.2 Community-Based Detection Methods**

Clustering the single cells based on similar expression profiles offers an axis on which diversity can be quantified by across individual samples. One approach used here is a graph-based clustering approach, where the cells are embedded in a graph structures with edges drawn between similar cells [21, 22]. The graph was then partitioned into highly connected communities and the Louvain algorithm is applied to optimize based on modularity. The modularity scores the quality of the optimized clusters. High modularity reflects the presence of community structure in the graph [25]. For our analysis, modularity less than 0.6 were further refined for lack of community structure present in the network analyzed. High confidences in community network structure are present in networks with modularity greater than 0.8. Using Seurat, cluster determination is implemented by *FindNeighbors* and *FindClusters* functions, respectively. In the *FindClusters* function, there is a resolution parameter (defaulted at 0.6) that allows researchers to adjust the granularity of downstream clustering. Increasing the resolution parameter in *FindClusters*, increases the number of distinct clusters identified and should be optimized for large datasets. Visualization of these clusters can be implemented using the *RunUMAP* function that utilizes the principal component

dimension reduction to create a 2D visualization of the clustered data, which can be displayed using the *DimPlot* function to plot the clustered dataset either by cluster identity or sample identity (or any other meta data groupings added to the dataset).

### **3.3 Diversity Score Calculation**

#### *3.3.1 Generalized Diversity Index*

The generalized diversity index takes the frequency of each sample identity in each of the clusters identified in the dataset and quantifies the diversity score that can be calculated over a range of resolution scales. The mathematical formulation is:

$${}^q D = \left( \sum_{i=1}^n p_i^q \right)^{\frac{1}{1-q}}$$

where  $n$  is the number of clusters identified,  $p_i$  is the frequency of each cluster, and  $q$  is the resolution or “order of diversity.” The most common diversity metrics, Shannon Entropy and Simpson Index are permutations of this generalized index [26]. Shannon entropy [27] is calculated by  $q = 1$  with  $\log({}^1 D)$  and the Simpson index [28] is calculated by  $q = 2$  corresponding to  $1/{}^2 D$ .

This generalized diversity index,  ${}^q D$ , can be calculated from the clustered data set by first counting the number of unique barcodes per cluster, then grouping cells by cluster and then by sample type. From the raw per-cluster-per-type grouping, the cell counts can be converted to frequencies, which can be directly input the equation for  ${}^q D$ , which can be solved over a range of  $q$  (one range capturing most dynamics if from  $10^{-2}$  to  $10^2$ ).

#### *3.3.2 Kolmogorov Smirnov Distance*

Another metric that can be used to quantify the differences between samples is using the Kolmogorov Smirnov (KS) distance between two discrete distributions. The KS distances is a nonparametric test, where similar distributions have smaller KS distances. The KS distance can be calculated by taking the probability mass function for a given sample across all the clusters identified by the aggregate dataset, converting that to a cumulative probability distribution. Then the KS distance is calculated using the supremum, or least upper bound (practically, the maximum value of a finite set of numbers)

$$d_{\text{KS}} = \sup(\text{abs}(c_{1,i} - c_{2,i}))$$

where  $c_1$  and  $c_2$  are the cumulative probability function of two different samples. The maximum value of the absolute differences of the cumulative distributions is what is known as the KS distance. Previously, we have shown the KS distances is smaller for like samples (two healthy or two AML) and larger between different samples (for example, healthy versus AML) [15].

---

## 4 Example

### 4.1 Background

In this section, we demonstrate how inter- and intrasample diversity may be used to quantify the bone marrow mononuclear cell (BMMC) heterogeneity of two acute myeloid leukemia (AML) patients compared to BMCCs of two healthy donors previously described in Ferrall-Fairbanks et al. [15] and code publicly available. This draws upon existing, publicly available data from Zheng et al. [6].

### 4.2 Method

Download the AML dataset (<https://github.com/MathOnco/scRNA-seqITH/blob/master/Pipelines/data/AML-Data.zip>) and R code (<https://github.com/MathOnco/scRNAsqITH/blob/master/Pipelines/MiMB-Diversity-Pipeline.R>) from GitHub to follow along with the worked-example of quantifying heterogeneity between two healthy donor and two AML BMCCs (see Notes 1 and 2).

---

## 5 Notes

### 1. Metric interpretation

Quantifying a generalized diversity metric enabled us to distinguish between leukemic states based on the high-dimensional single-cell patient samples. From an ecological perspective, diversity can be measured across a number of different spatial scales and by solving for a continuum of diversity indices we can examine a sample's diversity across these scales. For low  $q$ , generalized diversity index (GDI) represents the clonal richness, assuming that clusters of similar gene expression represent a ‘clone’ and as  $q$  approaches 0,  ${}^0D$  becomes the number of clusters identified. On the other side of the spatial scale, for high  $q$ , the contribution of the major clone(s) is weighted more, attempting to quantify species evenness. In a clinical setting, this would likely represent the dominant one or two phenotypes of tumor, that are easily detected by clinicians and may drive therapy selection. Intermediate values of  $q$  correspond to classical measures of sample diversity, such as Shannon index ( $H$ ,  $q = 1$ ) [27] that has been used in oncology to analyze tumor evolution and single-cell tumor imaging data [29]. We have seen that diversity scores can be very similar around  $q = 1$  and as a result the Shannon index can therefore be a problematic diversity indicator. However, GDI at a range of  $q$ , point to differences in the number of major drivers of tumor evolution, possibly prior to detection/sampling.

## 2. Robustness

One limitation of scRNA-seq is that missing data does not necessarily reflect that those transcripts are not expressed in the sample. In scRNaseq, any given cell only captures at most about 10% of the whole transcriptome, but in aggregation with other single cells, the entire genome is covered. As a result, in order to test the robustness this diversity metric, one can down-sampled their dataset and cluster to determine how the diversity index may change with the subset dataset. In Ferrall-Fairbanks et al., we down-sampled the dataset by randomly removing as much as 50% of the cells from each of the healthy and AML samples (this was repeated 1000 times) and found that the clustering did not change more than 1–2 clusters in either direction. With these new clustering, if the AML diversity curve was shifted down two units and healthy diversity curve was shifted up two units, you still would have separation between the conditions, suggesting that this metric is pretty robust. Furthermore, during this down-sampling exercise, we found that if we down-sampled to roughly 1500 cells, we would often capture the same number of clusters as the full dataset, suggesting that at least for this AML versus healthy BMMC comparisons, 1500 cells is a lower limit of cells needed to capture these diversity dynamics.

## 6 Summary

Translational bioinformatics is an emerging field at the intersection of molecular bioinformatics, statistics, and clinical applications. In the age of ever refined molecular insights in large data sets, it is important to develop pipelines that allow effective integration and biological interpretation with a focus on cancer evolution. Importantly, these pipelines have to scale when applied to large data sets, as well as deliver biological interpretation. With the pipeline described here, we have provided the theoretical and computational basis for tools that allow quantification and comparing diversity sample diversity at single cell resolution. The generalized diversity score applied to single cell sequencing samples allows comparing heterogeneity across different normal and malignant samples, and is based on clustering of single cell data. Hereby, particular choices of imputation and clustering procedures [30, 31] and batch correction [32, 33]—likely to undergo further development in the near future [34]—can be integrated into this concept—our method does not rely on a particular clustering algorithm.

Further therapeutic target development can be gleaned from the diversity analysis by exploring differentially expressed gene signatures between normal and malignant patients, or between different patients of different stages. How the measure of diversity,

and the associated gene signatures, change over time in a given patient may also offer important insights into therapeutically relevant targets, which needs to be explored further.

## References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW (2013) Cancer genome landscapes. *Science* 339 (6127):1546–1558
2. Marusyk A, Almendro V, Polyak K (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 12(5):323–334
3. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. *Nature* 501(7467):328–337
4. Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson AR (2015) Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer Res* 75 (8):1567–1579
5. Tabassum DP, Polyak K (2015) Tumorigenesis: it takes a village. *Nat Rev Cancer* 15 (8):473–483
6. Zheng GX, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8:14049
7. Paguirigan AL, Smith J, Meshinchi S, Carroll M, Maley C, Radich JP (2015) Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. *Sci Transl Med* 7(281):281re282
8. Lou J (2006) Entropy and diversity. *Oikos* 113 (2):363–375
9. Dagozo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 15(2):81–94
10. Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* 514(7520):54–58
11. Altrock PM, Liu LL, Michor F (2015) The mathematics of cancer: integrating quantitative models. *Nat Rev Cancer* 15(12):730–745
12. Park Y, Lim S, Nam JW, Kim S (2016) Measuring intratumor heterogeneity by network entropy using RNA-seq data. *Sci Rep* 6:37767
13. Hu Z, Sun R, Curtis C (2017) A population genetics perspective on the determinants of intra-tumor heterogeneity. *Biochim Biophys Acta* 1867(2):109–126
14. Giustacchini A, Thongjuea S, Barkas N et al (2017) Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med* 23 (6):692–702
15. Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM (2019) Leveraging single-cell RNA sequencing experiments to model intra-tumor heterogeneity. *JCO Clin Cancer Informatics* 3:1–10
16. MacArthur RH (1965) Patterns of species diversity. *Biol Rev* 40:510–533
17. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432
18. Tuomisto H (2010) A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* 164(4):853–860
19. Shannon CE (1997) The mathematical theory of communication. 1963. *MD Comput* 14 (4):306–317
20. Almendro V, Cheng YK, Randles A et al (2014) Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep* 6(3):514–527
21. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36 (5):411–420
22. Stuart T, Butler A, Hoffman P et al (2019) Comprehensive integration of single-cell data. *Cell* 177(7):1888–1902.e1821
23. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. *J Open Source Softw* 3 (29)
24. AlJanahi AA, Danielsen M, Dunbar CE (2018) An introduction to the analysis of single-cell RNA-sequencing data. *Mol Ther Methods Clin Dev* 10:189–196
25. Newman ME (2006) Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 103(23):8577–8582
26. Morris EK, Caruso T, Buscot F et al (2014) Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecol Evol* 4 (18):3514–3524

27. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423, 623–656
28. Simpson EH (1949) Measurement of diversity. *Nature* 163:688
29. Almendro V, Kim HJ, Cheng YK et al (2014) Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res* 74 (5):1338–1348
30. Qi R, Ma A, Ma Q, Zou Q (2019) Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform.* <https://doi.org/10.1093/bib/bbz062>
31. Petegrosso R, Li Z, Kuang R (2019) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief Bioinform.* <https://doi.org/10.1093/bib/bbz063>
32. Hicks SC, Townes FW, Teng M, Irizarry RA (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19(4):562–578
33. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 16(1):43–49
34. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 20 (5):273–282



# Chapter 11

## Managing a Large-Scale Multiomics Project: A Team Science Case Study in Proteogenomics

**Paul A. Stewart, Eric A. Welsh, Bin Fang, Victoria Izumi, Tania Mesa, Chaomei Zhang, Sean Yoder, Guolin Zhang, Ling Cen, Fredrik Pettersson, Yonghong Zhang, Zhihua Chen, Chia-Ho Cheng, Ram Thapa, Zachary Thompson, Melissa Avedon, Marek Wloch, Michelle Fournier, Katherine M. Fellows, Jewel M. Francis, James J. Saller, Theresa A. Boyle, Y. Ann Chen, Eric B. Haura, Jamie K. Teer, Steven A. Eschrich, and John M. Koomen**

### Abstract

Highly collaborative scientists are often called on to extend their expertise to different types of projects and to expand the scope and scale of projects well beyond their previous experience. For a large-scale project involving “big data” to be successful, several different aspects of the research plan need to be developed and tested, which include but are not limited to the experimental design, sample collection, sample preparation, metadata recording, technical capability, data acquisition, approaches for data analysis, methods for integration of different data types, recruitment of additional expertise as needed to guide the project, and strategies for clear communication throughout the project. To capture this process, we describe an example project in proteogenomics that built on our collective expertise and experience. Key steps included definition of hypotheses, identification of an appropriate clinical cohort, pilot projects to assess feasibility, refinement of experimental designs, and extensive discussions involving the research team throughout the process. The goal of this chapter is to provide the reader with a set of guidelines to support development of other large-scale multiomics projects.

**Key words** Experimental design, Planning, Big data, Proteogenomics, Informatics, Biostatistics, Cancer, Landscape paper

---

### 1 Introduction

In this current era of team science, large-scale projects often utilize different omics approaches to address complex questions in biology and human disease. As an example, our research team was interested in undertaking a proteogenomics project to assess frozen tumor tissues from a cohort of ~100 lung squamous cell carcinoma

patients to support improved molecular classification [1]. As lung cancer definitions have expanded from organ site to histology and subsequently to genomic classifiers like mutation status or chromosomal translocations producing fusion proteins, we expected that this approach could yield even more understanding of lung squamous cell carcinomas by linking genotypes to phenotypes. This effort provided novel challenges to a research team that had experience in each relevant category of research (including lung cancer, genomics, proteomics, bioinformatics, and statistics), because we had not undertaken a project of this scale that required integration of all of our contributions and built on previous publications from other researchers.

By 2014, cancer genomics had produced two landscape papers for lung squamous cell carcinoma (LSCC) from The Cancer Genome Atlas (TCGA) [2] and Wilkerson et al. [3]. Both publications could classify LSCC tumors using gene expression data, providing genomic signatures that could serve as a benchmark for comparison of additional genomic datasets. Different proteomics datasets generated with liquid chromatography–tandem mass spectrometry (LC-MS/MS) had also been published for lung tumor tissues [4–11]. In addition, integration of different omics approaches had also been published, with the first examples being comprehensive genomics with integrated analysis of deoxyribonucleic acid (DNA) copy number and mutation status, ribonucleic acid (RNA) gene expression arrays, and microRNA expression arrays [12]. This research was quickly followed by proteogenomics datasets [13, 14]. From this background, our research team was interested in applying these techniques to the molecular classification of LSCC tumors, which are poorly understood, leaving patients with high medical need due to the lack of treatment options.

In order to undertake this project, a strategy was developed to examine each step, determine the available resources and required knowledge, and undertake the necessary pilot projects as practice for the full scale project. This process will be outlined, from hypothesis formulation, cohort identification, data collection, data release, tissue processing, RNA processing, DNA processing, protein processing, analytical variable collection, data integration, and bioinformatics analysis. Each step of this process will be described; salient points will be discussed both in specific terms for this project and in lessons learned for future projects. This chapter is focused on the people, process, and project management associated with a complex multidisciplinary multiomics project.

---

## 2 Materials

### 2.1 Team

1. Invest in the people that can contribute to the project. The biggest asset to the project is early, frequent and sustained involvement by key faculty and staff members of the shared resources required for the study.
2. Engage laboratory leadership early and establish strong working relationships. In our study, the tissue core, proteomics core, genomics core and bioinformatics core leadership were significantly engaged throughout the project, and consistently interacted with the principal investigators and other members of the laboratory team.
3. Include domain experts from every relevant field. Broad engagement is important. The collaborative investigators included expertise in oncology, bioinformatics and proteomics. We included the shared resources project manager, Collaborative Data Services, Tissue Shared Resource, Genomics Shared Resource, Proteomics Shared Resource, Biostatistics and Bioinformatics Shared Resource throughout the process.
4. Recruit an expert in logistics. Oversight of the workflow, timelines, and coordination of a complex project requires a person with training and experience dedicated to project management.

### 2.2 Large Biobank

1. Accumulate a large pool of available patient tissues, so that the various tissue requirements for each assay can be accommodated.
2. Expect attrition of tissues, because some samples from the initial list will not meet the specific quality or quantity requirements or clinical characteristics defined in the inclusion and exclusion criteria for the study.
3. Maintain institutional support for and long term commitment to the development of the resources to provide the best chance for success. These efforts were supported by Moffitt's institutional commitment to the Total Cancer Care protocol, which consents patients to tissue donation and lifetime follow up.

### 2.3 Data Resources for Querying Clinical Information and Biobanked Samples

1. Set up database access to key clinical and research systems to enable effective querying. This database includes elements from the cancer registry, tissue bank, pathology reports, self-reported questionnaires, and deidentified medical records.
2. Make tissue quality data elements available for initial queries. This step focuses on fields such as the amount of tissue available, the histology of the tumor, the smoking status of the patient (or other disease-relevant behavioral variables), and the pathology staging.

3. Use multiple data sources to resolve ambiguities in clinical data elements, particularly in diagnosis and treatment records.

#### **2.4 Data Concierge with Honest Broker Capability for Managing Data Queries**

1. Recruit or develop a team member with knowledge of key data resources and their characteristics, because this background knowledge is required for effective querying of existing samples for cohort selection.
2. Invest in training and educating the Data Concierge/Honest Broker. While digital or electronic data may be available for research use, there are many subtle aspects that require significant training and experience to fully utilize the available data. For instance, some data resources provide more detailed information, but are frequently missing for a patient; other resources may have poor data quality for key elements. Without this expertise, navigation of complex data systems can involve significant error rates.
3. Filter the sample set for final consideration iteratively to maximize use of the combination of tissue bank information, clinical medical records, cancer registry follow-up information, and self-reported data.
4. Benefit from the learning process of the Data Concierge/Honest Broker. The domain knowledge gained by an honest broker focused on a single disease can be invaluable in cohort selection and refinement. As an example, certain criteria can significantly impact the number of available samples that qualify for the study. It is important to know the steps when “the funnel closes” and to adjust queries from requirements to preferences to maintain cohort size.

#### **2.5 Experimental Resources**

1. Invest in appropriate instrumentation, software, and staff expertise.
2. Complete thorough literature searches for existing methods. For instance, review journals, including *Nature Methods*, *Nature Protocols*, *Bioinformatics*, *Molecular and Cellular Proteomics*, *Journal of Proteome Research*, *Journal of Visualized Experiments*, and the book series, *Methods in Molecular Biology*, which all publish methods-oriented content.
3. Develop detailed protocols describing each part of the process informed by the use of pilot projects.
4. Record dates, run order, and sample quality (where feasible) for every sample preparation and assay step to enable downstream identification of batch or process related technical artifacts. Include any additional metadata that could be informative or could be used to study confounding variables.

---

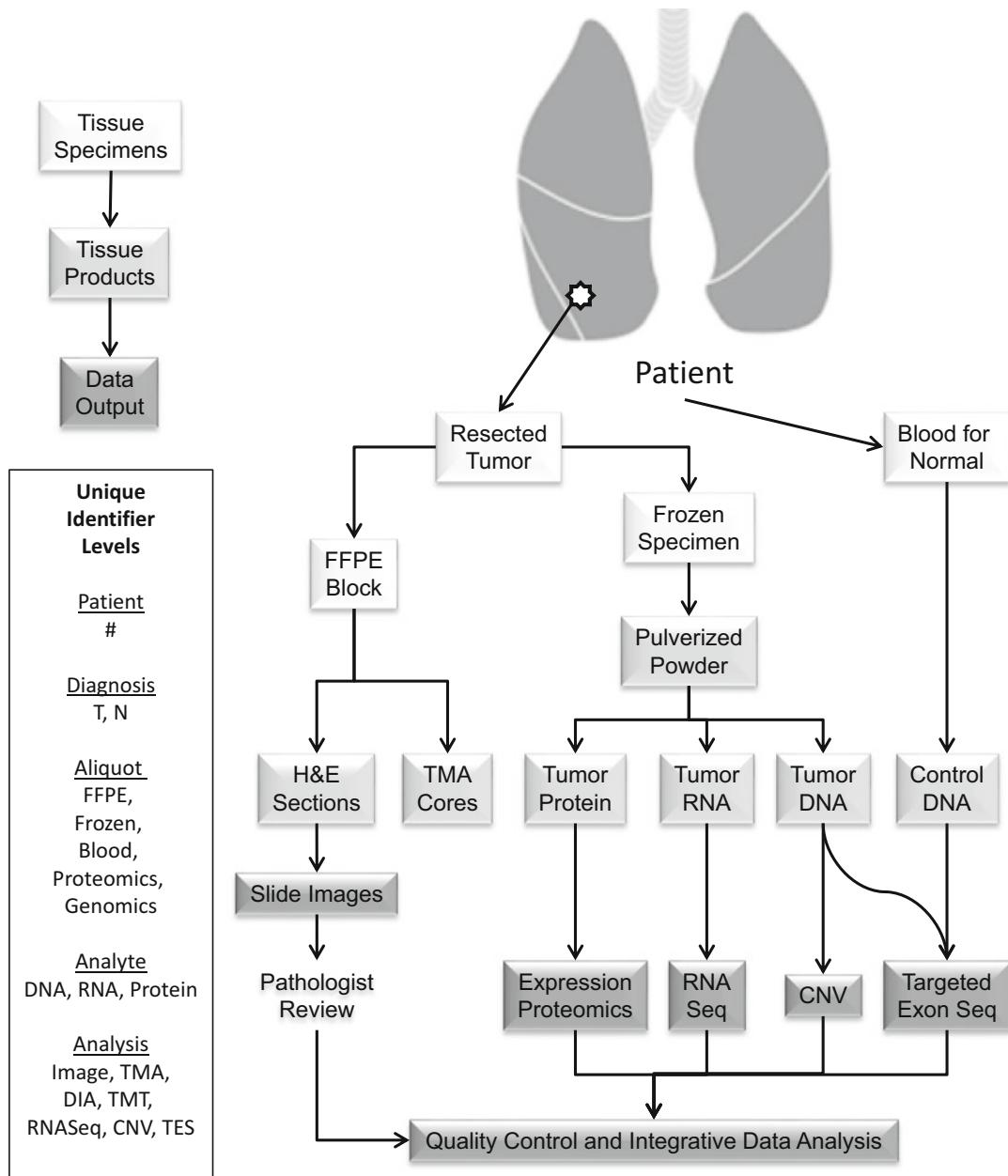
### 3 Methods

#### 3.1 Experimental Design

1. Develop the hypothesis of the study. The question(s) being asked will drive many downstream decisions and must be stated clearly at the outset. This step includes essential starting characteristics, such as the disease type. In our study, one hypothesis was that protein expression adds more insight into lung squamous cell carcinoma tumor biology that will impact clinical classification and can guide the development of novel therapeutic strategies.
2. Select the types of omics used in the study. This decision is made partly on the basis of scientific insight, availability of technical expertise, funds, and other considerations. The omics types used in our study included targeted DNA sequencing (since DNA mutation frequency has been previously cataloged by other studies), DNA copy number variation, RNA sequencing for gene expression, and expression proteomics (*see* Fig. 1). For the purposes of this discussion, we will focus mainly on DNA and RNA sequencing in combination with expression proteomics.
3. Determine the general experimental workflow. As the focus of our study was to characterize LSCC tumors in a retrospective study, the workflow was to collect all necessary information, validate processing techniques, and then produce omics data in parallel through multiple molecular assays from multiple research groups (*see* Fig. 1).

#### 3.2 Cohort Design

1. Perform initial cohort identification from existing sources to determine how many total samples are available for consideration. This process includes collecting information on the number, type, and quality of data elements that are available for later analysis [15]. For instance, a challenge such as collecting lung tumors requires having access to reliable smoking status and patient survival information. The availability of these tissue and data resources should be known at the outset of the project (*see* Fig. 2).
2. Develop criteria for the minimum tissue amount by carefully reviewing tissue processing requirements for procurement, RNA/DNA/protein extraction, and assay processing with the respective research groups. The goal is to use a single frozen specimen divided into multiple aliquots for the diverse molecular assays.
3. Overestimate the minimum tissue criteria to plan for the future and enable follow-up experiments, if needed. Once a unique cohort with extensive clinical and molecular characterization has been created, many new assays will suddenly become available for further study. Having enough tissue to prioritize new

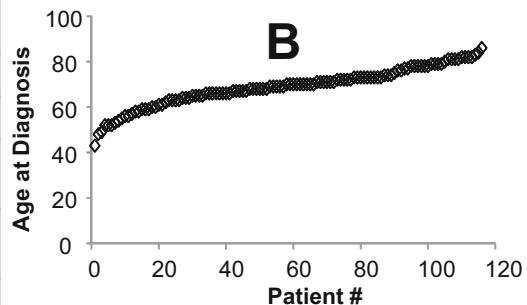
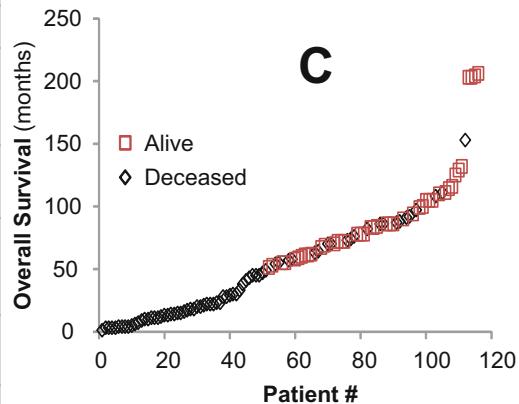


**Fig. 1** Experimental workflow for lung squamous cell carcinoma proteogenomics. The overall process begins with tissue specimens and proceeds to generate datasets (upper left). To organize a study of this type, multiple identifiers are required to link the samples and data back to the patient (lower left). The overall workflow diagram is presented to show the overall process and emphasize the integrative nature and illustrate the high need for communication and consensus in terms of strategy

molecular types will further enhance the value of the work. In our study, B cell receptor sequencing and phosphoproteomics have arisen from our initial studies and require additional banked material in different processed forms.

**A**

Criterion	Data
Dates of Diagnosis	1994-2011
Age Range	43-86
Gender	77 Male, 39 Female
Race/Ethnicity	113 Caucasian (1 Hispanic), 3 African American
Vital Status	43 Alive, 73 Deceased
Overall Survival	1-206 months
Lung	54 Left, 62 Right
1 <sup>st</sup> Course of Treatment	S (83), SC (22), SR (6), SCR (5) <u>Surgery, Chemotherapy, Radiation</u>
TNM Stage (#)	1 (2), 1A (22), 1B (39), 2 (1), 2A (10), 2B (24), 3A (9), 3B (9)
Avg. Overall Survival by Stage (months)	Stage 1: 65 Stage 2: 54 Stage 3: 47
Differentiation	57 poorly, 56 moderately, 2 well, 1 N/A
Node Positive Disease	33
Recurrence	29

**B****C**

**Fig. 2** Summary of selected clinical data for lung squamous cell carcinoma patients. Demographics and tumor characterization data (**a**) are presented with plots of age at diagnosis (**b**) and overall survival (**c**)

4. Once a full cohort has been identified, proceed to perform cohort selection. Questions to ask during this process include: Which samples from the initial cohort will be best to use? How do we group or randomize these samples? Answers can be dependent on tissue amount, tissue preservation, tissue age, clinical (or pathologic) staging, existing molecular data from clinical assays, presence of prior treatments, and age of patient. Other considerations include any funding limitations, which should focus the project on the pre-specified hypothesis. The cohort selection step involves many tradeoffs, in terms of availability and desirability of specimens and data.
5. Recheck the identified cohort. Data for the cohort selection may come from different sources, or may be recorded in multiple locations. Check all sources of available data to verify the

disease status, tissue availability, and other parameters. For instance, if the patient has no history of lung cancer despite having an annotated tissue, further investigation is required. Site of origin should also reflect whether the tumor is a primary or a metastatic lesion.

6. Perform pathology review of hematoxylin and eosin stained (H&E) slides. If available, utilize previously stained H&E slides to allow a pathologist to visually verify that the samples under consideration are the disease being studied. Annotations can be confusing or inaccurate; visual verification ensures that the extensive effort planned for each tissue is justified.
  7. Determine a batch order for samples. Inevitably, processing differences can arise over the course of the experiment. Use a predetermined batch order for all processing steps, including tissue processing, genomics/proteomics core processing, plate organization for experiment runs, and data acquisition order. Ideally, batches can be created that represent the largest unit of work that can be performed at once. In our case, the numbers of samples included in two proteomics experiments aligned with a single genomics experiment; therefore, the batches were created to support both. This batching requires agreement and participation among research groups to maintain the ordering throughout processing steps. This process enables the batch information to be recorded and examined as a confounding variable (*see Note 1*).
  8. Randomize sample order. Batch effects are inevitable; therefore, ensure that samples representing key phenotypic variables are randomly and evenly distributed across batches. In our project, the biostatisticians designed a full run order for samples and verified there were no statistically significant differences between batches in the stage, recurrence, gender, vital status, stratified stage/recurrence combinations, and age of samples.
1. Open broad communication channels for all members of the team and use them frequently. This communication should seem excessive at first, to ensure that all members of the team have opportunities to provide input into the planning and execution process. In our project, a series of 4–5 breakout meetings were held with each group prior to initiation of the study, followed by monthly coordination meetings with the whole team (*see Note 2*). In our experience, once members of a research lab see the level of commitment to quality and process, they will engage in careful and thoughtful design. Rely on the expertise and experience of the project manager or logistics expert to assist with the steps in this section.

### 3.3 Operational Design

2. Develop the Institutional Review Board (IRB) protocol for the study. Err on the side of providing extensive process detail in the protocol and listing all possible personnel for the project. Even if the details are not required for IRB approval, the documentation of the process, the thought required, and the development of good teamwork/communication are essential. This process includes recording the standard operating procedures (SOPs) for each shared resource as appendices to the protocol. In particular, add enough detail and options in the protocol to allow for flexibility once the project begins, in case of any changes of plan.
3. Publish all protocols as part of a methods supplement. Once the IRB protocol is fully developed, it becomes a reference source for the publications that follow.
4. Establish sample naming conventions prior to initiation of the project. Unique identifiers should be associated with each aliquot of tissue. Should each shared resource use the same identifier? Will the proteomics aliquot be named “identifier\_-protein”? Note this decision requires consultation with shared resources, and includes logistics such as ensuring the identifier can fit on a printed label for tubes/boxes used to store tissues.
5. Plan to profile more samples than needed to accommodate sample losses. Despite best efforts, some tissues will not be of sufficient quality or quantity to proceed. Attrition occurs throughout the project, so planning for failures at the outset leads to sufficient sample size at completion. Despite our best efforts in this project, the original 116 samples resulted in 108 samples in the final dataset (*see Note 3*).
6. Plan to pilot unfamiliar processes or experiments. If particular processes will be new or different from standard operating procedures, implement pilot experiments using specimens not used in the main project. For instance, our project needed to determine the feasibility of large-scale Tandem Mass Tag (TMT) chemical labeling for expression proteomics compared to label-free proteomics. We designed and executed a pilot study to ensure the quality and content of data was equivalent. Equally important, we were able to work out the process prior to initiation of the project and publish the results for a smaller scale case study [16] (*see Note 4*).
7. The full experimental schedule from beginning to end should be considered before each experiment begins. Ideally, tissue samples should be processed and experiments performed in a continuous sequence. This rule can help reduce time-based batch effects. Holidays and routine instrument downtime should be avoided if possible.

### **3.4 Clinical Data Collection**

1. Extract clinical data. As part of the initial experimental design, available clinical variables should be evaluated and selected. Once the cohort is selected, extract the clinical variables from source systems into a single dataset.
2. Engage board-certified pathologists to review hematoxylin and eosin stained slides to verify tumor content, and record as much detail as possible. Verification of malignancy and evaluation of the tissue percentages of tumor, normal, necrosis, stroma, immune infiltration, and other features provides confirmation that the sample is worth analyzing. Cutoff criteria can be established for each tumor type to determine whether a sample has sufficient histologic quality for analysis.
3. Involve a biostatistician to review and summarize the clinical variables as well as confounding variables. Biostatisticians spend large amounts of time cleaning up datasets and summarizing cohorts. As a result, they have a highly trained sense for anomalous data or trends. Engage them early to verify data follow expected distributions. Include omics scientists in these discussions to explore pre-analytical and analytical variables that may need to be considered as part of the analysis plan.
4. Expect clinical data surprises, even late in the project. Not all clinical data are visible and these data represent the needs of clinical care when originally collected. This subset of the medical history almost certainly does not represent the total information needed with regard to the patient. For instance, in a tertiary cancer center (such as our institution), prior clinical treatments for the disease may not occur locally. Thus, exclusion of tumors from patients with prior treatment may be difficult to ascertain due to lack of available data.
5. Be strategic on identifying key data and prioritize data collection. Our Total Cancer Care (TCC) protocol [17] allows for investigators to contact patients again to update follow-up data, including survival outcomes. Based on the need for the recurrence variable, patient contact was prioritized for those patients who did not have a documented recurrence. Other clinical variables can be used to suggest high risk of recurrence and thus allow for further prioritization.

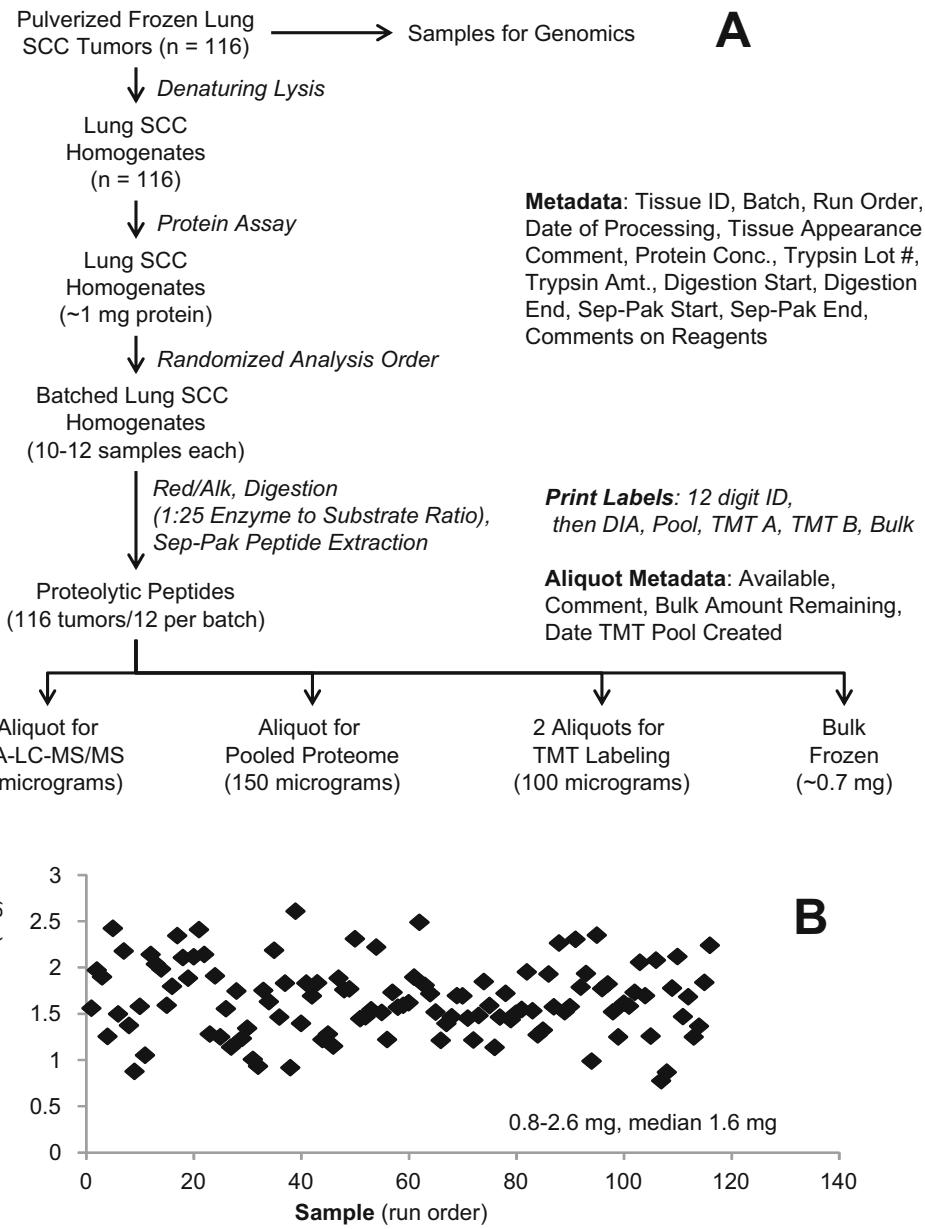
### **3.5 Tissue Processing**

1. Create a spreadsheet of samples to record metadata and tissue handling details. Include the time and date of tissue handling, time removed from freezer, and number of strikes required for pulverization.
2. Allocate sufficient team member time for tissue processing. Plan around employee time off and other large projects to ensure the same individual performs the same role for the entire project.

3. Decide on the appropriate batch size for tissue processing. For our project, batches of 12 tumors were processed together, because the biopulverization device had 12 wells, which needed to be cleaned to eliminate sample cross-contamination and re-equilibrated to  $-80^{\circ}\text{C}$  between batches.
4. Determine the number of people who will process the samples. Two person sample processing was selected for this project. One person carried out the majority of the physical tasks of biopulverization (“masher”), while the other person was responsible for data recording (“reporter”). This system sped up the process and significantly reduced the opportunity for errors (*see Note 5*).
5. Define the aliquots and relative amounts needed for downstream analysis. Because equal samples were needed for proteomics and genomics, we had to decide whether to cut individual aliquots from the bulk tumor or to divide the sample after homogenization. The latter option was chosen in an effort to minimize the impact of tumor heterogeneity between datasets.
6. Process samples in bulk and divide the pulverized powder into aliquots for downstream analysis to maximize consistency of sample preparation, timing, and freeze/thaw effects (*see Note 6*).

### 3.6 Proteomics

1. Determine the approach that will be used; *see Fig. 3* for an example workflow. At the time this project began, bottom-up proteomics included several different strategies for analysis of protein expression. Analysis of individual samples using peptide fractionation, LC-MS/MS, and peptide counting had been published for the Clinical Proteomics Tumor Assessment Consortium (CPTAC) colorectal cancer proteogenomics project [13]. However, in a setting with more limited funding and less available instrument time, chemical labeling (a barcoding strategy for sample multiplexing) was more appealing. Two types of chemical labeling were available from commercial vendors at the time this project was designed: isobaric tags for relative and absolute quantification, (iTRAQ, Sciex), and tandem mass tags (TMT, Proteome Sciences/Thermo). iTRAQ experiments typically used four reporter ion tags, while TMT used six (*see Note 7*). This choice has further support in the fact that subsequent CPTAC breast and ovarian cancer datasets [18, 19] were acquired using chemical labeling strategies using newer 10- or 11-plex TMT reagents.
2. Develop a pilot project to characterize the output of each experiment. We had internally already compared 4-plex iTRAQ and 6-plex TMT, and found that they provided similar

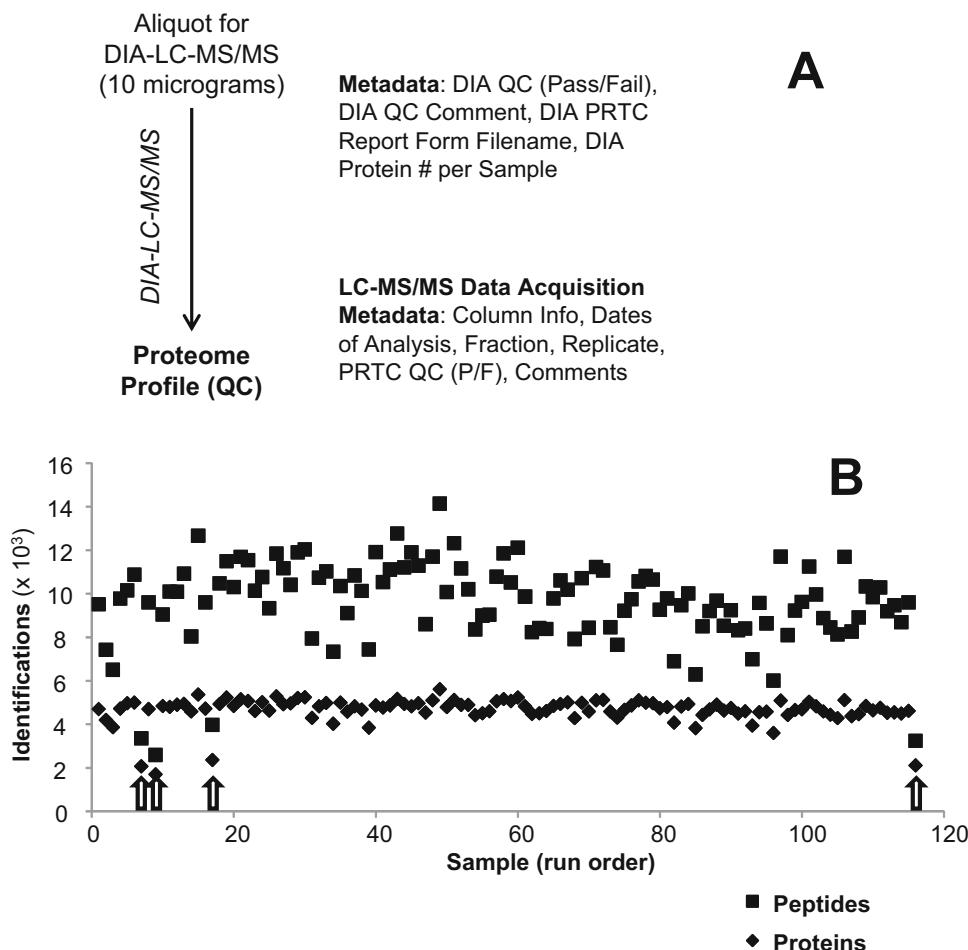


**Fig. 3** Experimental workflow diagram for proteomics sample preparation and data for protein recovery from frozen tumors. The workflow (a) describes the steps in sample processing and digestion. The pooled proteome serves as a control for batch-to-batch comparison of the 29 TMT experiments. Bulk digest is saved for future posttranslational modification (PTM) experiments or quantitative follow-up. Protein assay results are included for each tumor homogenate in chronological order of processing (b)

amounts of protein identifications and concordant quantitative results [16]. Therefore, the higher multiplexing of 6-plex TMT was favored. However, we did pursue and publish a pilot study comparing LC-MS/MS of individual samples with TMT

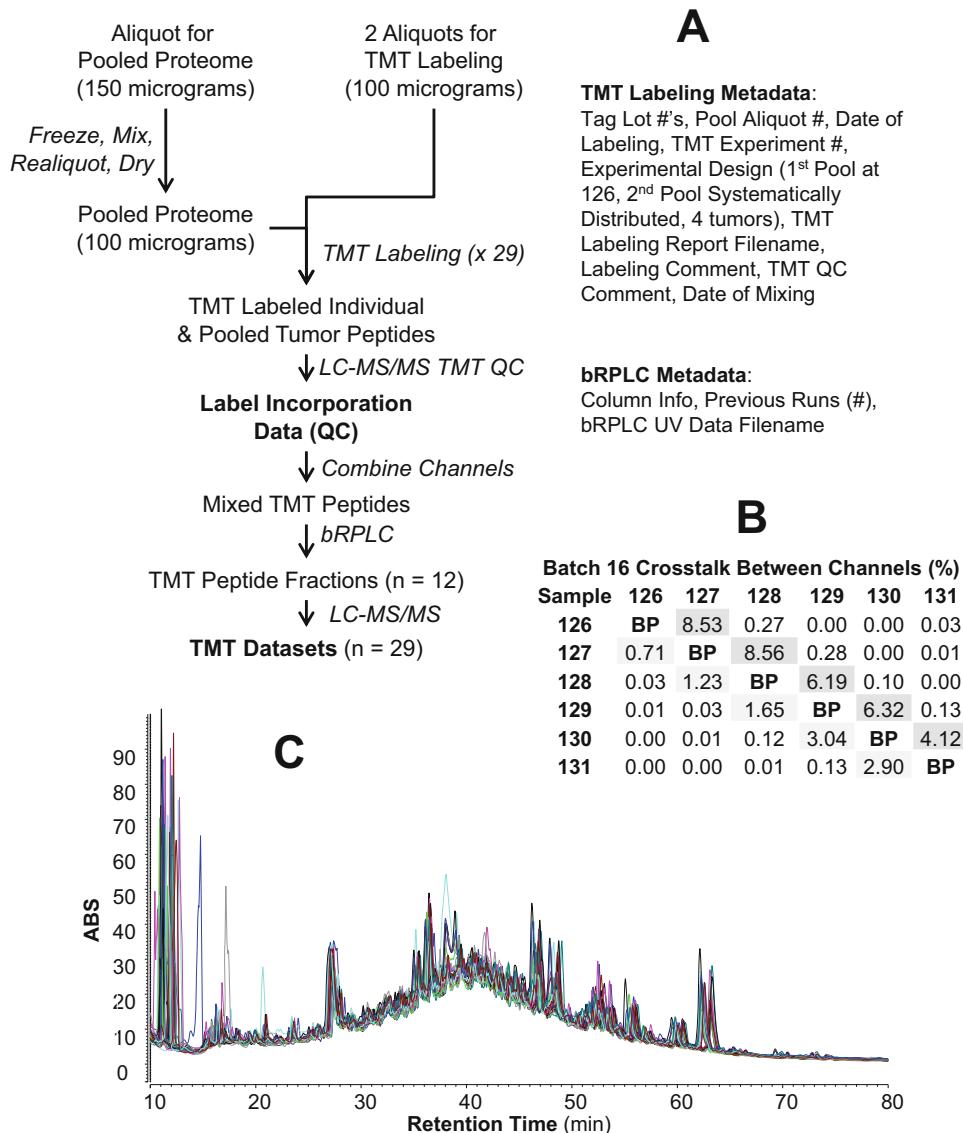
analysis of three lung adenocarcinomas and three squamous cell carcinomas, which indicated that TMT would be an optimal strategy, particularly in its cost-effectiveness [20]. One limitation for sample multiplexing is the decrease in likelihood that a tumor-specific peptide is detected.

3. Having chosen a strategy, define the different stages at which quality control experiments are needed. Overemphasize these quality control metrics in the first pass experiment and then decide whether to retain them in future projects or reduce their stringency.
4. Rapidly assess your sample preparation and proteolytic digestion to equalize samples. LC-MS/MS with data independent acquisition (DIA) [21] was used for quality control as shown in Fig. 4 to verify that each tumor homogenate had been processed effectively using the readout of identified peptides and proteins (*see Note 8*).



**Fig. 4** Digestion quality control using data independent acquisition LC-MS/MS. The workflow (a) describes data acquisition and associated metadata. The numbers of identified peptides and proteins indicated four samples (arrows) that failed quality control and needed to be digested again (b)

5. Determine the number of aliquots of digest and the corresponding amounts that are needed to support current and future experiments. In this project, we stored two aliquots with 100 µg of digest for TMT labeling experiments (providing a backup sample), one aliquot of 150 µg of digest for a pooled control sample (see next step), and retained the rest of the bulk digest in another tube (~0.7 mg), which could be used for experiments like phosphoproteomics that benefit from higher input amounts.
6. Define a strategy for evaluation of batch-to-batch variability and normalization using a control sample. Several choices are available, including cell lines or tumor pools. In this case, a pooled tumor proteome was selected, so that the content of all of the tumors could be represented in the pooled sample. This choice is important in TMT experiments, since the tumor pool can provide the full proteome that can be encountered in every sample, as opposed to a cell line pool that lacks stroma and other histological components of tissue. An aliquot of each digest was saved to create a tumor pool in this experiment, as shown in the workflow in Fig. 5.
7. Utilize predetermined sample order and batch assignment for TMT labeling. For each 6-plex TMT experiment, four individual tumors were analyzed with two pooled samples. The rationale for this design was that the protein expression levels in the individual tumors are usually measured as ratios to the pooled sample, whether directly or as an intermediate normalization step. Therefore, to assess consistency, the second pool was included to generate a pool-to-pool ratio, which should have a value close to 1 and would serve as a quality control for each individual measurement.
8. Rotate the placement of the second pooled sample, as shown in Fig. 5, to minimize any systematic interference that may occur between channels and impact the utility of the controls.
9. Define appropriate quality controls for chemical labeling. In this experiment, every labeled channel was individually analyzed with LC-MS/MS; *see* workflow diagram in Fig. 5. MaxQuant was used to identify and quantify both labeled and unlabeled peptides; the number of labeled peptides divided by the total number of identified peptides was used as a surrogate for quantification of labeled peptide signal to unlabeled peptide signal. Samples with >98% labeled peptides were accepted. This rapid method proved to be suitable, in part because the response factors of labeled and unlabeled peptides are not the same in LC-MS/MS, and careful quantification of the peptide peak intensities or areas is not expected to be directly quantitative (*see Note 9*).

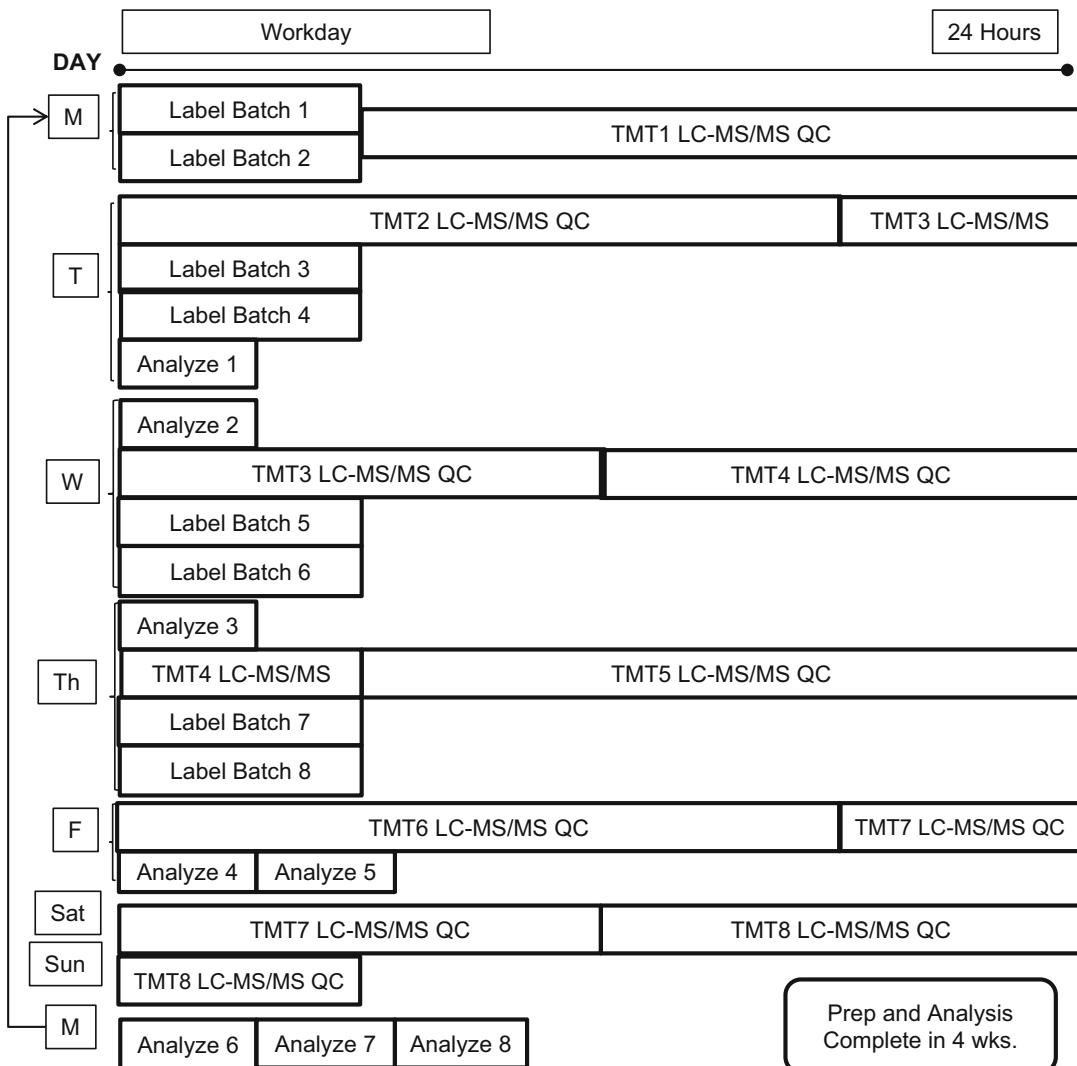


**Fig. 5** Tandem mass tag labeling and peptide fractionation for expression proteomics. The workflow (a) describes sample processing, data acquisition, and associated metadata. The amount of cross talk between TMT channels for Batch 16 are shown as an example of the quality control data (b). Overlaid UV chromatograms show the reproducibility of peptide separation across the 29 TMT batches (c)

- Define a strategy for peptide fractionation as well as associated metadata and quality control metrics. Basic pH reversed phase liquid chromatography with fraction concatenation was selected for this project, and each 6-plex TMT was separated using this technique. Ultra-violet (UV) chromatograms were recorded and overlaid to verify performance (see Fig. 5). Based on previous experiments, LC-MS/MS of 12 fractions

produced the optimal balance between data and cost/time; in addition, this amount of peptide fractionation produced minimal amounts of chimeric tandem mass spectra, which negatively impact TMT quantification. Use of higher numbers of fractions will produce more data, but the main impact is the addition of more peptides per protein rather than detection of higher numbers of proteins.

11. Define an injection control to evaluate LC-MS/MS instrument performance. The peptide retention time calibrator (PRTC) mix or other stable isotope-labeled standard peptides can be used to assess signal intensity, reproducibility of retention time, and mass measurement accuracy among other parameters for certification of instrument performance.
  12. Block dedicated Instrument time with no interruptions, as shown in Fig. 6 for the TMT labeling quality control. This point is critical to minimize variability. This project had three major blocks scheduled for LC-MS/MS: 116 data independent acquisition (DIA) LC-MS/MS analyses for digestion quality control (QC), or ~1 month of instrument time; 174 LC-MS/MS runs to check TMT labeling efficiency, or ~1.5 months of instrument time; and 696 LC-MS/MS experiments for duplicate analysis of each of 12 fractions for the 29 6-plex experiments, or 3 months of instrument time. Future experiments should try to minimize the amount of instrument time for the TMT label verification experiment.
  13. Minimize the number of LC column changes within a scheduled block. We determined that column changes were the largest source of process-related variability between samples.
  14. Perform QC assessment at the end of each TMT 6-plex experiment to verify equal instrument performance. Tune, calibrate, and exchange columns as needed to maintain high quality data.
  15. Graph QC metrics routinely. Examine the performance of the injection standards (e.g., PRTC mix) and outputs for each experiment, including numbers of identified and quantified peptides and proteins. Intervene as necessary to maintain data quality (*see Note 10*).
  16. Note samples that fail QC. To maximize available time and capability for intervention, samples that fail QC and should be re-analyzed at the end of the experiment. Selected samples that pass QC are included in repeated analyses to verify consistency with earlier data.
- 3.7 Analysis of Expression Proteomics Data**
1. Define a strategy for database searching to identify peptides and quantify them based on reporter ion intensities. While our group had experience with a number of approaches, it was challenging to analyze a dataset with 696 LC-MS/MS



**Fig. 6** Example operational diagram for proteomics weekly lab activities and instrument usage for TMT labeling and quality control. Lab members label peptides with the TMT reagents, acquire LC-MS/MS data for quality control, and perform database searches and quantitative analysis for labeling. One LC-MS/MS instrument is dedicated to perform the experiments during this time

experiments. For smaller scale experiments, we had experience generating catalogs with Sequest (Thermo) and Mascot ([www.matrixscience.com](http://www.matrixscience.com)) database searches that could be displayed in Scaffold ([www.proteomesoftware.com](http://www.proteomesoftware.com)) as well as Andromeda searches prior to MaxQuant relative quantification [22, 23]. In order to be able to assemble all of the data in a single analysis, IDPicker [24, 25] was selected and three search engines, Myrimatch [26], MS-GF+ [27], and Comet [28], were used to maximize the capability to identify peptides, similar to the strategy for the colorectal CPTAC project [13].

2. Create a database that includes real protein sequences as well as decoys. A commonly used option creates a decoy database by reversing the existing sequences. UniProt ([www.uniprot.org](http://www.uniprot.org)) and RefSeq are often used for database searches.
3. Choose appropriate parameters for database searching, as well as filtering criteria for inclusion and exclusion of results. First, the *m/z* tolerances or allowed mass measurement error should be specified. For high resolution data, 10 parts per million (ppm) is a recommended starting point; for low resolution data or to accommodate errors in calling the monoisotopic peak, then 0.5–0.6 Th can be used. Enzyme specificity (i.e., fully tryptic with no cleavages prior to proline) and posttranslational modifications (e.g., TMT label incorporation, carbamidomethylation of cysteine, methionine oxidation, etc.) should be specified. In particular, the false discovery rate (FDR) for peptide sequence identification should be a point of focus, and different consequences can be observed for selection of peptide or protein FDR. In addition, be certain that any contaminant filters are not eliminating proteins that could be biologically relevant (e.g., some keratins can be used to distinguish between cell types). Filtering based on the amount of missing values for a particular protein or protein group should also be considered.
4. Determine whether quantitative comparisons will be made using ratios or peak height/area/intensity. While many TMT experiments rely on ratios, we decided to use the pooled samples (channel 126) to create scaling factors, so that the expression levels could be compared using relative intensity.
5. Normalize the data. Iterative rank order normalization [29] was used to reduce or eliminate batch-to-batch variation in the samples and across the 29 TMT 6plex experiments.
6. Define the method for peptide assembly into protein groups and choose a strategy based on maximal data reporting or parsimony.
7. Choose the strategy for combining the quantification of the individual peptides into a single value for protein or protein group quantification (e.g., calculating the sum or geometric mean of all quantified peptides). In proteomics, a challenge here is to overcome missing data. Not all peptides will be consistently quantified across TMT experiments due to the stochastic nature of sampling in data dependent tandem mass spectrometry and variations in instrument performance. While other data acquisition strategies (e.g., data independent acquisition) may have lower amounts of missingness, these methods are not compatible with the sample multiplexing required for increased throughput in this project.

### **3.8 Nucleic Acid Processing**

1. Determine how samples are used for nucleic acid extraction. Options include splitting the sample into two aliquots for separate DNA/RNA extraction, or performing DNA/RNA isolation from the same aliquot. In our project, frozen pulverized tissues would have required additional thaw time to further separate into two aliquots. Therefore, the isolation approach was used to serially extract DNA and RNA from the same sample.
2. Pilot the sequencing experiments to be used in the project. Nucleic acid processing approaches do not produce universally superior purified products, but instead depend on the specific assays used downstream. In our project, we determined that the sequencing experiments needed different levels of RNA/DNA purity than hybridization-based approaches. DNase treatment of RNA requires particular attention when performing RNA-seq.

### **3.9 Molecular Genomics**

1. Determine sequencing approach to be used for DNA. If targeted DNA sequencing is to be run, determine capture kit to use (whole exome, cancer gene, custom) and confirm genes and genomic regions of interest are included. A comprehensive cancer panel plus three custom genes was used to characterize the samples based on mutations in known cancer genes relevant to LSCC.
2. Determine sequencing approach to be used for RNA. A total RNA stranded approach was used based on prior experience with these data and a desire to comprehensively assess transcribed sequence.
3. Determine the appropriate batch size for processing. This step should be done in collaboration with the other teams to ensure a consistent batch size (or multiples of a larger, common batch size). The batch size will be related to the size of plates for library preparation, the number of runs that can be performed on a sequencer, and other factors.
4. Determine when top-off processing will occur. For sequencing, the individual runs may not produce a sufficient number of reads to meet the predetermined threshold. An additional run is usually performed typically after the initial batch, but before moving to the next batch.
5. Determine when sample failures will be rerun. Assume that some sequencing samples or runs will fail. The timing of reruns should be predetermined. These samples could be run immediately following the batch in which they fail. However, this process will require coordination and discussion, slowing down the execution of the process. In our project, a final batch of samples was planned, which consisted of the failed samples as

well as reanalysis of a selected sample that passed QC to assess consistency with the prior data.

6. Select the specific assays to perform as a pilot study. Based on preliminary results from the TCGA paper detailing somatic mutations in lung squamous cell carcinoma, a targeted exon panel was compared to whole exome sequencing to determine whether each effectively detected the key mutations in lung SCC.
7. Perform an integration pilot study. RNA/DNA extraction from tissue can have significant impacts on the RNA/DNA sequencing steps that follow. In our project, we uncovered incomplete DNase activity once RNA isolation was completed. The resulting DNA contamination in RNA causes specificity issues with sequencing that were previously not identified in other contexts. It required the coordination of the tissue extraction, genomics, and bioinformatics groups to solve this issue.
8. Perform QC: Assess read counts, base qualities, and other vendor provided sequencing metrics to ensure high quality of generated sequence data.

### **3.10 DNA Alignment**

1. Select a reference. DNA is typically aligned against the entire reference genome. Even when using targeted sequencing approaches (including whole exome) sequences are aligned to the entire genome. Many regions of the human genome are repetitive, and such sequences are often present in targeted results as off-target bycatch. In order to not force these off-target sequences to align to targeted regions incorrectly (which often yields false-positive variant identification), the entire genome is used as an alignment target. Often, the genome is pre-processed or indexed by the alignment program to facilitate fast execution. The hs37d5 human genome version was used in our analysis.
2. Prepare input FASTQ files. Adapter removal and low quality base trimming were performed as part of the Illumina sequence processing.
3. Run alignment using the Burrows-Wheeler Aligner or BWA [30] and sort resulting sequence alignment map (SAM)/binary alignment map (BAM) using samtools [31], possibly run samtools fixmate and fillmd. This step results in a properly formatted BAM output that is compatible with tools for downstream analysis.
4. Mark duplicate reads with Picard (<https://broadinstitute.github.io/picard/>). This process identifies reads that come from the same original molecule based on position information. If molecular barcodes are used at the DNA molecule level,

this information can be used to precisely identify duplicate sequences.

5. Realign indels using Genome Analysis Tool Kit (GATK) [32]. Sequences neighboring known and newly detected insertions and deletions can be misaligned due to penalties associated with accounting for an insertion or deletion. Realignment sequence reads to possible insertions or deletions often removes such errors by allowing for an easier alignment.
6. Perform Base Quality Score Recalibration using GATK. This step adjusts base quality scores based on features of the bases and sequence reads.
7. Complete final cleaning with samtools calmd and index the BAM file.
8. Perform QC: Evaluate alignment rates, duplication rates, and nonhuman alignment rates. Deviations from expected values would indicate a problem with the experiment at some stage in the process.

### **3.11 RNA Alignment**

1. Select a reference genome or transcriptome. In the case of a whole genome reference, a gene definition file can be provided to facilitate faster or more accurate alignment based on known exons. If the RNA-seq experiment is strand-specific, the orientation may need to be provided to improve alignment. The hs37d5 human reference genome with RefSeq gene definitions was used in our analysis.
2. Run alignment with Tophat2 [33].
3. Calculate QC metrics using Picard. The MEDIAN\_CV\_COVERAGE metric can be especially useful as an additional RNA quality metric of the amplified sample, independent from any recorded DV200/300 or RNA integrity number (RIN) from TapeStation or Bioanalyzer assays.
4. Optional: If RNA-seq data will be used for variant detection, alignment refinement is recommended as in Subheading 3.10 above. Indel realignment may not handle intron sequence as well as RNA-seq aligners and may make alignment worse with little benefit, so this step may be excluded. A special GATK step (SplitNCigarReads) is needed to account for N's in the alignment output to avoid problems with downstream tools (*see Note 11*).
5. Complete final indexing of the output BAM file.
6. Perform QC: Evaluate alignment rates, nonhuman alignment rates, coverage variation. Deviations from expected values would indicate a problem with the experiment at some stage in the process.

### **3.12 Genomic Variation Detection**

1. Identify somatic mutations. Several approaches can be taken (*see Note 12*). A tumor-only mutation calling approach was applied to a focused set of known cancer genes. Tumor samples include an individual's inherited variants. To enrich for somatic mutations, population variation databases like the 1000 Genomes Project were used to remove common inherited variants. A pool of nontumor samples was also used as a control to remove false variants arising from technical artifacts.
2. Run variant detection using GATK. All samples were analyzed simultaneously using GATK UnifiedGenotyper on all BAM files at once, thereby using observation across all samples to help improve sensitivity. This step also results in output of reference genotypes, allowing for the ability to distinguish between high-quality reference and missing genotypes. The data are output as standard VCF (Variant Call Format), see <https://samtools.github.io/hts-specs/VCFv4.2.pdf>.
3. Filter putative mutations. Mutations are first restricted to the target region to include on-target mutations [16]. GATK variant quality score recalibration (VQSR) was used to improve accuracy of highly sensitive (but less specific) GATK variant identification. Mutations are further filtered by removing variants found commonly in the 1000 Genomes Project or in a pool of nontumor samples. This selection has previously been shown to improve somatic mutation specificity, although inherited variants will still remain [34].
4. Annotate variants using Annotate Variation (ANNOVAR) [35]. Annotation adds contextual information, including overlap with a gene, prediction of a resulting amino acid change, and overlap with a variety of population databases (1000 Genomes [36], Exome Aggregation Consortium [37]), and genotype/phenotype databases (such as ClinVar found at <http://www.ncbi.nlm.nih.gov/clinvar/>).
5. Perform QC: Assess mutation counts, mutation types, and depth of coverage across targeted region.

### **3.13 RNA Gene Expression**

1. Define the approach for gene expression analysis. RNA expression has long been a valuable tool in understanding the intricate biological interactions within living cells. Massively parallel sequencing generates a sequence read representing information from an individual molecule, and can therefore be used to count instances of a given molecule that represents its prevalence in the original sample. Measuring gene expression from RNA sequencing data has been a popular application and has prompted development of novel methods to address new challenges inherent in RNA sequencing. A gene-level expression approach based on read counting was used for this project.

2. Count reads aligning to defined genes using HTseq [38].
3. Normalize gene-level expression using DESeq2 [39].
4. Perform QC: Assess number of genes detected and distribution of expression values. Utilize dimension reduction techniques (like Principal Component Analysis or t-Distributed Stochastic Neighbor Embedding (t-SNE)) to observe patterns of difference in the data. Such patterns may indicate experimental influences on expression due to batch or others effects. These artifacts would not represent biological differences of interest.

### 3.14 Individual Omics Analysis

1. Follow a predetermined analysis plan. Use similar work in other studies to guide the development of specific questions and the use of specific techniques. For example, multiomics studies from The Cancer Genome Atlas and the Clinical Proteomics Tumor Assessment Consortium. [13, 18, 19].
2. Handle missing data. Different types of molecular data (RNA-Seq and Proteomics) may have values that are missing in samples due to the nature of the assays that generate the data. The type of missingness may differ in molecular types [40], including missing due to random sampling or missing due to signal below the limit of detection. Imputation is frequently employed to address this effect. In our project we removed protein/RNA observations with high levels of missingness (>10%). In cases where the techniques could not use missing data, we imputed using minimum value imputation (typically with a value of 0).
3. Consider appropriate statistical tests for the molecular type. Statistical tests involve basic assumptions on the data involved; verify these assumptions in each molecular data type considered before applying the tests. For instance, a Student's *t* test may be appropriate if the input data is expected to be normally distributed, but a Mann–Whitney test does not require this assumption.
4. Control for multiple testing. In high-dimensional omics analyses, performing many different hypothesis tests leads to high false positives using an uncorrected *p*-value threshold for significance [41]. Use false discovery rates [41, 42] or other multiple testing corrections to accommodate the large number of tests [43, 44].
5. Perform enrichment analysis. Individual gene/protein findings can be difficult to interpret. Use enrichment techniques to identify pathways or gene ontology terms that appear more frequently than expected in the significance analysis. Examples include Enrichr [45, 46] and Gene Set Enrichment Analysis (GSEA) [47].

### 3.15 RNA, DNA, and Protein Integration

1. Set the rationale for integrative analysis of omics data. DNA sequencing, RNA sequencing, and proteomics information each give distinct but related snapshots of a biological system. By incorporating more than one of these perspectives, we can gain a greater understanding of biological systems and how they are perturbed in human disease.
2. Consider appropriate processing techniques for each omics data type. Consolidation of the types requires matching gene/protein identifiers, which can be complex. In the case of transcripts and proteins, many different isoforms exist with highly similar sequences, making these annotations very difficult to correctly assign. In addition, protein groups identified by the same set of peptides will correspond to multiple genes, so that element of the gene–protein linkage strategy also needs to be defined.
3. Define mutations altering expression. Studies of oncogenes in cancer have shown that a single mutation can result in a dramatic change in downstream gene and protein levels. A single mutation can alter the activity of a protein product, resulting in significant changes in downstream signaling cascades and expression programs. Statistical approaches can be used to identify such changes using molecular measurements of mutations (DNA-seq), gene expression (RNAseq) and/or protein expression using mass spectrometry-based proteomics (*see Note 13*). Sample groups can be defined by mutation status, and then differential gene or protein expression can be determined between samples with the presence or absence of the mutation. Differences in frequency of specific mutations across groups (defined by clinical features, gene, or protein expression) can also be calculated. Mutations are not always common across samples, which impacts the ability to detect differences. This limitation can be addressed by “collapsing” mutations with similar function to the gene level or even up to pathways. However, each collapsing step results in a loss of specific biological information.
4. Examine mutations modulating protein stability. Protein levels can be determined for genes harboring mutations suspected of affecting protein stability within a sample. In particular, a marked decrease in protein amount may suggest nonsense-mediated decay in the case of a truncating mutation.
5. Examine mutations that have an impact on posttranslational modifications, like phosphorylation. Phosphoproteomic experiments can identify phosphorylation differences that may be directly associated with mutations. This process includes mutations that would alter the ability of a kinase to phosphorylate a given protein site, as well as mutations at or near

phosphorylation sites that affect the ability of a protein to be phosphorylated.

6. Confirm expression of mutations at the protein level. Protein expression can be used to confirm that a given mutation is expressed at the protein level. This goal may require a separate set of experiments, particularly if the expression proteomics data was acquired using multiplexed chemical labeling. Targeted proteomics with a focus on the mutant peptides can be a very effective approach for detection. Individual tumor analysis with in depth label free expression proteomics may also provide the necessary data.
7. Examine RNA/protein concordance [20]. A variety of studies have demonstrated a surprisingly low correlation (usually ranging from 0.25 to 0.35) between protein expression and transcript expression. Calculate the paired transcript/protein correlation across samples to identify proteins that are well correlated with transcript expression and those that are not. Use Gene Ontology enrichment analysis (Enrichr [45, 46]) to determine classes of molecules that do and do not correlate. In our project, this process identified transcripts involved in nonsense-mediated decay and other categories that do not correlate well with protein expression. Extract information about processes controlled primarily by transcription (that show strong correlation of gene expression and protein expression) and those that are modulated by other means (e.g., protein posttranslational modification).
8. Use clustering for molecular categorization of samples. Perform ConsensusClustering (using ConsensusClusteringPlus [48] in R/Bioconductor) to find stable patient subgroups in RNA and protein expression data. Compare these subgroups using visualization techniques including ComplexHeatmap [49].
9. Jointly cluster multiple molecular types using iCluster [50] or similar tool to identify similar structures in proteomics and RNA expression data.

### **3.16 Performing an Integrated Proteogenomics Study in Lung Squamous Cell Carcinoma**

The approaches above were used by our team to profile 116 tumor tissues from patients with lung squamous cell carcinoma [1]. Beginning in 2015, the investigators developed the hypothesis that proteomics data can provide additional insights into lung squamous cell tumor biology and provide value for translational research. To address this hypothesis, we determined to combine proteomics characterization with other omics types more commonly used in cancer research. The TCGA project provided information that DNA sequencing and RNA expression were important for LSCC. Copy number variation plays a role in lung cancer so it was considered as well, since this is difficult to extrapolate from other data

sources. At this point in the project, the general workflow was to perform a parallel analysis of tumor tissue by multiple shared resources. Our team had already worked together in an interdisciplinary way to integrate medical oncology, proteomics, genomics and bioinformatics expertise.

Since not everything would be understood prior to implementation, we designed a pilot study to exercise the strategy on a set of five initial samples. As expected, this step provided significant insights into the process, data flow, timing and group communication. It also demonstrated the seriousness of the group to doing the study well and with completeness. Group meetings became regularly scheduled at this point to ensure issues were addressed and operational details could be worked out in a timely, focused fashion. Technology limitations, instrument time, and overall logistical limits were identified as part of the pilot study. Importantly, the handoff between tissue procurement, tissue processing and downstream genomics processing during the pilot allowed us to identify a needed process change to ensure quality sequencing results. In fact, the sequencing pilot was rerun to ensure the changes produced the expected results. The proteomics characterization pilot involved the use of several competing technologies, which were then evaluated to determine the approach we would take with respect to the larger project [16].

Before and during the pilot phase of the project, several key members of the research group met with all the individual teams to discuss logistical issues and process design. Through the course of the project, we realized that discussing key data elements to capture during processing provoked further discussion about logistics and operations. Building a table of experimental conditions that needed to be captured for each technology allowed us to think through the process that would ultimately generate these experimental details. It also involved everyone in the process and demonstrated the commitment to undertaking a quality study.

Once the pilot projects had been completed and changes implemented, the execution of the full project began. Tissues were processed in a pre-specified, randomized order followed by parallel profiling by multiple research groups. Monthly meetings allowed everyone to provide updates on the tissue processing, genomics, proteomics, data provisioning and bioinformatics progress. Importantly, as unexpected issues arose, the group could address them, knowing the goals of the project and having worked together on similar problems from the pilot. While the data generation process required multiple steps and incurred inevitable delays, we did not need to deviate significantly from the originally developed plan. As data was produced from each shared resource, the bioinformatics team worked with each group for QC. Additionally, the study team designed the detailed analysis plan, while the data generation was occurring. The analysis and interpretation phase of

the project required significant time and effort, as expected, however the entire group was involved in the interpretation.

The project yielded insights into lung squamous cell carcinoma tumor biology [1]. The group identified three major subgroups of LSCC, relating to immune and redox activity as well a mixed subtype. These groupings partially overlapped existing classifications, but the proteomics data clearly indicated immune-related biology that was not evident initially from the genomics. Interestingly, these subgroups did not correspond to differences in overall survival, likely because specific interventions for these molecular subtypes are not yet available. However, the presence of B-cell-rich tertiary lymph node structures was observed and found to be related to outcomes. These structures did not completely align with the immune subgroup, suggested more complex biology of the tumor-immune microenvironment remains to be uncovered in this disease.

This LSCC project was able to characterize 116 tumors using proteomics, genomics, RNA expression and clinical data. Ultimately, we analyzed and presented the final 108 tumor cohort in *Nature Communications* [1], providing a multiomic view of the disease and categorizing its major drivers and phenotypes. The complexity of the project, particularly at a single institution, required extensive coordination and communication. Most importantly, it required a commitment to quality and thoroughness from many people with expertise that spanned medical oncology, genomics, proteomics, tissue processing, bioinformatics, and biostatistics working effectively as a team.

### 3.17 Conclusions

In order to tackle a large-scale project for the first time, the approach must be systematic and thorough. The first step is to define the hypothesis or hypotheses to be tested to set the direction of the research. An extensive inventory of personnel, resources, expertise, and instrumentation will help define capabilities and expose any missing components needed for the research. The next step involves cohort selection: flexibility in study design and inclusion/exclusion criteria can be critical to success in this phase. Constant communication is required through the data acquisition and data analysis phases to overcome potential stumbling blocks. As new work is published on these complex proteogenomics projects [51–57], additional literature review and comparison of methods is useful to refine your strategy. The steps outlined in this chapter describe the efforts that were used for a large-scale proteogenomics project including ~100 tumors from different patients, and they will serve as a basis for future projects of this scale and scope at our institution.

---

## 4 Notes

1. Subheading **3.2, step 7:** Batch order is an essential step of the design process. No matter how careful a molecular laboratory prepares for the project, variation will occur. It is not uncommon for experiments to assay the controls first, followed by treatment conditions, thereby making it difficult to perform comparisons of treatment effects without being confounded by time-related variables.
2. Subheading **3.3, step 1:** Meetings are a crucial part of adequate communication and coordination. A number of recurring team meetings should be used to organize, plan and make experimental design decisions. As the project matures, the content shifts more to operational issues, timelines and challenges. Finally, once the data is generated, the focus further shifts into analytical strategies and interpretation of results. These changes should lead to the inclusion of additional collaborators as needed to fully understand the data and subsequent biological knowledge that is created. Each new member of the research team will need to be brought up to speed on the process in order to contribute effectively.
3. Subheading **3.3, step 5:** Tissue management is essential for maximizing the benefit of the project. This process includes ensuring that sufficient tissue is available at the outset of the project, and that the derived biospecimens are adequately stored, organized, and labeled so that future studies can use the samples. Nomenclature should be a point of consensus, so it is clear to all collaborators.
4. Subheading **3.3, step 6:** Piloting various processes is invaluable. The team ran at least three pilots to ensure that the DNA/RNA extraction, processing, and sequencing processes were completely determined. Although the processes in use were appropriate for specific applications, the project required a specific protocol to ensure minimal DNA contamination in RNA aliquots prior to sequencing. This quality control checkpoint would not have been determined, understood, and resolved without appropriate piloting.
5. Subheading **3.5, step 1:** Thoroughly document tissue handling. In our project, we recorded information including batch number, run order, pulverization date, operator, pulverizer, pulverizer well ID, number of hammer strikes, batch start time, batch end time, sample processing start time, sample processing end time, and free-text notes. Process information cannot be recovered based on memory or post hoc analysis, so it must be captured during the activity. It is important to record as many potential sources of batch effect and technical variation

as is practical, in case they explain observed variation in the final assays.

6. Subheading **3.5, step 6:** Continuity by homogenizing or minimization of heterogeneity between the tissue used from proteomics and genomics. The team chose to homogenize the specimen prior to distributing components to individual core facilities for RNA/DNA/protein processing. The applied technologies operate on bulk tumor, not single cells; therefore, they represent a composite signal of many cells (e.g. tumor, stroma, and immune infiltrates). Since this mixture is inherent to each technology, the team ensured a representative mixture of cell types was available for each. Thus, artifacts due to sampling will be consistent and analytical approaches can rely on more evenly distributed cell types.
7. Subheading **3.6, step 1:** While higher multiplexing is possible with 8-plex iTRAQ and 10-plex or 11-plex TMT, we did not want to further dilute the signal from each individual tumor or reduce the speed of LC-MS/MS data acquisition. The analysis of individual samples provides the best chance to detect mutations and other rare features that are unique to an individual tumor. The use of chemical labeling favors detection and quantification of proteins that are expressed across all samples. For initial molecular classification, the robust signals that are detected in comparison of chemical labeling should provide the largest differences between tumor subtypes; for finer detail in an iterative experiment, individual sample analysis will likely provide even deeper insight.
8. Subheading **3.6, step 4:** In this dataset, analysis of four samples produced fewer peptide and protein identifications than the other samples. Re-digestion of these samples produced the same results, ruling out a technical problem with digestion of these samples. Therefore, the samples were included for further analysis. Due to the sample multiplexing, the impact of having fewer protein identifications is lessened by the contributions of the other samples and the pools. These samples also were not outliers in principal component analysis (PCA) plots and other quality control assessments. Another benefit provided by this single sample LC-MS/MS analysis is that the data can be used to define accessible biology and candidate biomarkers that are easily detectable in tumor tissues, defining the low-hanging fruit for translation to clinical measurements.
9. Subheading **3.6, step 9:** This expenditure of instrument time may appear high (analysis of six samples from each of 29 TMT experiments), but we did discover one sample that was not labeled and two instances when reagent vials for different tags had been incorrectly labeled; for example, reagents for two

channels, reporter ion 128 and reporter ion 129, were incorrectly labeled. The unlabeled sample was labeled again; the notes were corrected for proper assignment of the reporter ions in the other two cases. Therefore, it is important to retain this step and highly worthwhile to evaluate faster ways of acquiring the data. Ion mapping (MS/MS of every *m/z* value) during an infusion experiment should verify the correct reporter ion and provide limited information about tagging efficiency (positive or negative results), but it will not clearly estimate the percentage of tagged peptides.

10. Subheading 3.6, step 14: The data analysis strategy for the quality control steps does not need to be same as the full process used to create the dataset that will be used for molecular classification of tumors. This process can be fast and easy, as long as it is consistent. Rapid searches of every 24 LC-MS/MS data files with Sequest (Thermo) or Mascot ([www.matrixscience.com](http://www.matrixscience.com)) with visualization in Scaffold ([proteomesoftware.com](http://proteomesoftware.com)) or quantitative analysis with MaxQuant [22] would work well for this step.
11. Subheading 3.11, step 5: Genotyping approaches are now being used on RNA sequence data as well. Although additional challenges exist (including but not limited to more heterogeneous depth of coverage, reverse transcriptase errors, posttranscriptional gene editing, and allele specific expression), genetic information is represented in RNA sequencing data and is accessible using RNA sequencing.
12. Subheading 3.12, step 1: Somatic mutations can be identified in many ways. A tumor sample is always sequenced because it will contain the cancer specific mutations of interest. The main challenge is removing inherited variants. This goal can be achieved by excluding genetic variants observed in population variation databases and by excluding variants observed in local nontumor samples run with the same capture kits. A more specific approach is to use a matched nontumor sample from the same patient, which serves as a precise control to remove inherited variants.
13. Subheading 3.15, step 4: Proteomics can be used to confirm that a given mutation is expressed at the protein level. There are many challenges to this type of experiment, including peptide size based on protease digestion, detectability of the appropriate peptide fragment via mass spectrometry-based proteomics, and sample purity. Evidence of expression of a mutation can be important, especially in experiments seeking to identify immune epitopes presented by the cell, but low sensitivity for specific mutated peptides (particularly in sample-multiplexed chemical labeling experiments) limits utility.

## Acknowledgments

The authors would like to thank all of our past and present collaborators for providing projects and previous experience that could be integrated into the development of this new capability. We would also like to acknowledge the contributions of the coauthors to the original proteogenomics paper. J.K. would like to thank Daniel Liebler, PhD, Steven Carr, PhD, and Henry Rodriguez, PhD, for their mentoring and thank Proteome Sciences for a sponsored research agreement that introduced his lab to Tandem Mass Tag chemical labeling. This work was supported by the Collaborative Data Services Core, Tissue Core, Proteomics and Metabolomics Core, Molecular Genomics Core, and Biostatistics and Bioinformatics Shared Resource at H. Lee Moffitt Cancer Center (Moffitt Cancer Center), which are partially funded by the NCI Cancer Center Support Grant (P30-CA076292). Project funding was provided through the Moffitt Lung Cancer Center of Excellence; the academic time of the faculty was supported by the Moffitt Cancer Center.

## References

- Stewart PA, Welsh EA, Slebos RJC, Fang B, Izumi V, Chambers M, Zhang G, Cen L, Pettersson F, Zhang Y, Chen Z, Cheng CH, Thapa R, Thompson Z, Fellows KM, Francis JM, Saller JJ, Mesa T, Zhang C, Yoder S, DeNicola GM, Beg AA, Boyle TA, Teer JK, Ann Chen Y, Koomen JM, Eschrich SA, Haura EB (2019) Proteogenomic landscape of squamous cell lung cancer. *Nat Commun* 10(1):3578. <https://doi.org/10.1038/s41467-019-11452-x>
- Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489 (7417):519–525. <https://doi.org/10.1038/nature11404>
- Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, Muldrew K, Miller CR, Randell SH, Socinski MA, Parsons AM, Funkhouser WK, Lee CB, Roberts PJ, Thorne L, Bernard PS, Perou CM, Hayes DN (2010) Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res* 16(19):4864–4875. <https://doi.org/10.1158/1078-0432.CCR-10-0199>
- Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, Beer DG (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteom* 1(4):304–313. <https://doi.org/10.1074/mcp.m200008mcp200>
- Ha ES, Choi S, In KH, Lee SH, Lee EJ, Lee SY, Kim JH, Shin C, Shim JJ, Kang KH, Phark S, Sul D (2013) Identification of proteins expressed differently among surgically resected stage I lung adenocarcinomas. *Clin Biochem* 46(4–5):369–377. <https://doi.org/10.1016/j.clinbiochem.2012.11.014>
- Kawamura T, Nomura M, Tojo H, Fujii K, Hamasaki H, Mikami S, Bando Y, Kato H, Nishimura T (2010) Proteomic analysis of laser-microdissected paraffin-embedded tissues: (1) Stage-related protein candidates upon non-metastatic lung adenocarcinoma. *J Proteome* 73(6):1089–1099. <https://doi.org/10.1016/j.jprot.2009.11.011>
- Kikuchi T, Hassanein M, Amann JM, Liu Q, Slebos RJ, Rahman SM, Kaufman JM, Zhang X, Hoeksema MD, Harris BK, Li M, Shyr Y, Gonzalez AL, Zimmerman LJ, Liebler DC, Massion PP, Carbone DP (2012) In-depth proteomic analysis of nonsmall cell lung cancer to discover molecular targets and candidate biomarkers. *Mol Cell Proteom* 11 (10):916–932. <https://doi.org/10.1074/mcp.M111.015370>
- Pernemalm M, De Petris L, Branca RM, Forshed J, Kanter L, Soria JC, Girard P, Validire P, Pawitan Y, van den Oord J,

- Lazar V, Pahlman S, Lewensohn R, Lehtio J (2013) Quantitative proteomics profiling of primary lung adenocarcinoma tumors reveals functional perturbations in tumor metabolism. *J Proteome Res* 12(9):3934–3943. <https://doi.org/10.1021/pr4002096>
9. Wei Y, Tong J, Taylor P, Strumpf D, Ignatchenko V, Pham NA, Yanagawa N, Liu G, Jurisica I, Shepherd FA, Tsao MS, Kislinger T, Moran MF (2011) Primary tumor xenografts of human lung adeno and squamous cell carcinoma express distinct proteomic signatures. *J Proteome Res* 10 (1):161–174. <https://doi.org/10.1021/pr100491e>
10. Zeng GQ, Zhang PF, Deng X, Yu FL, Li C, Xu Y, Yi H, Li MY, Hu R, Zuo JH, Li XH, Wan XX, Qu JQ, He QY, Li JH, Ye X, Chen Y, Li JY, Xiao ZQ (2012) Identification of candidate biomarkers for early detection of human lung squamous cell cancer by quantitative proteomics. *Mol Cell Proteom* 11(6): M111.013946. <https://doi.org/10.1074/mcp.M111.013946>
11. Zhang W, Wei Y, Ignatchenko V, Li L, Sakashita S, Pham NA, Taylor P, Tsao MS, Kislinger T, Moran MF (2014) Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches. *Proteomics* 14 (6):795–803. <https://doi.org/10.1002/pmic.201300382>
12. Lazar V, Suo C, Orear C, van den Oord J, Balogh Z, Guegan J, Job B, Meurice G, Riponche H, Calza S, Hasmats J, Lundeberg J, Lacroix L, Vielh P, Dufour F, Lehtio J, Napieralski R, Eggermont A, Schmitt M, Cadrelan J, Besse B, Girard P, Blackhall F, Validire P, Soria JC, Dessen P, Hansson J, Pawitan Y (2013) Integrated molecular portrait of non-small cell lung cancers. *BMC Med Genet* 6:53. <https://doi.org/10.1186/1755-8794-6-53>
13. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shad-dox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Town-send RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slepko RJ, Liebler DC, Nci C (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* 513 (7518):382–387. <https://doi.org/10.1038/nature13438>
14. Li L, Wei Y, To C, Zhu CQ, Tong J, Pham NA, Taylor P, Ignatchenko V, Ignatchenko A, Zhang W, Wang D, Yanagawa N, Li M, Pintilie M, Liu G, Muthuswamy L, Shepherd FA, Tsao MS, Kislinger T, Moran MF (2014) Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. *Nat Commun* 5:5469. <https://doi.org/10.1038/ncomms6469>
15. Michener WK (2015) Ten simple rules for creating a good data management plan. *PLoS Comput Biol* 11(10):e1004525. <https://doi.org/10.1371/journal.pcbi.1004525>
16. Stewart PA, Fang B, Slepko RJ, Zhang G, Borne AL, Fellows K, Teer JK, Chen YA, Welsh E, Eschrich SA, Haura EB, Koomen JM (2017) Relative protein quantification and accessible biology in lung tumor proteomes from four LC-MS/MS discovery platforms. *Proteomics* 17(6). <https://doi.org/10.1002/pmic.201600300>
17. Fenstermacher DA, Wenham RM, Rollison DE, Dalton WS (2011) Implementing personalized medicine in a cancer center. *Cancer J* 17 (6):528–536. <https://doi.org/10.1097/PPO.0b013e318238216e>
18. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clouser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, Kawaler E, Mundt F, Krug K, Tu Z, Lei JT, Gatza ML, Wilkerson M, Perou CM, Yellapantula V, Huang KL, Lin C, McLellan MD, Yan P, Davies SR, Townsend RR, Skates SJ, Wang J, Zhang B, Kinsinger CR, Mesri M, Rodriguez H, Ding L, Paulovich AG, Fenyö D, Ellis MJ, Carr SA, Nci C (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534 (7605):55–62. <https://doi.org/10.1038/nature18003>
19. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D, Sun S, Yang F, Chen L, Wang J, Shah P, Cha SW, Aiyetan P, Woo S, Tian Y, Gritsenko MA, Clauss TR, Choi C, Monroe ME, Thomas S, Nie S, Wu C, Moore RJ, Yu KH, Tabb DL, Fenyö D, Bafna V, Wang Y, Rodriguez H, Boja ES, Hiltke T, Rivers RC, Sokoll L, Zhu H, Shih IM, Cope L, Pandey A, Zhang B, Snyder MP, Levine DA, Smith RD, Chan DW, Rodland KD, Investigators C (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* 166(3):755–765. <https://doi.org/10.1016/j.cell.2016.05.069>
20. Stewart PA, Parapatics K, Welsh EA, Muller AC, Cao H, Fang B, Koomen JM, Eschrich SA, Bennett KL, Haura EB (2015) A pilot proteogenomic study with data integration identifies MCT1 and GLUT1 as prognostic markers in lung adenocarcinoma. *PLoS One* 10(11):e0142162. <https://doi.org/10.1371/journal.pone.0142162>

21. Prakash A, Peterman S, Ahmad S, Sarracino D, Frewen B, Vogelsang M, Byram G, Krastins B, Vadali G, Lopez M (2014) Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis. *J Proteome Res* 13(12):5415–5430. <https://doi.org/10.1021/pr5003017>
22. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367–1372. <https://doi.org/10.1038/nbt.1511>
23. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10(4):1794–1805. <https://doi.org/10.1021/pr101065j>
24. Holman JD, Ma ZQ, Tabb DL (2012) Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker. *Current Protoc Bioinformatics*. Chapter 13:Unit13 17. <https://doi.org/10.1002/0471250953.bi1317s37>
25. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL (2009) IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* 8(8):3872–3881. <https://doi.org/10.1021/pr900360j>
26. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6(2):654–661. <https://doi.org/10.1021/pr0604054>
27. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5:5277. <https://doi.org/10.1038/ncomms6277>
28. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13(1):22–24. <https://doi.org/10.1002/pmic.201200439>
29. Welsh EA, Eschrich SA, Berglund AE, Fenstermacher DA (2013) Iterative rank-order normalization of gene expression microarray data. *BMC Bioinformatics* 14:153. <https://doi.org/10.1186/1471-2105-14-153>
30. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
32. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498. <https://doi.org/10.1038/ng.806>
33. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36. <https://doi.org/10.1186/gb-2013-14-4-r36>
34. Teer JK, Zhang Y, Chen L, Welsh EA, Cress WD, Eschrich SA, Berglund AE (2017) Evaluating somatic tumor mutation detection without matched normal samples. *Hum Genomics* 11(1):22. <https://doi.org/10.1186/s40246-017-0118-2>
35. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164. <https://doi.org/10.1093/nar/gkq603>
36. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, Tasse AM, Flieck P (2017) The international Genome sample resource (IGSR): a worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res* 45(D1):D854–D859. <https://doi.org/10.1093/nar/gkw829>
37. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissono D, Boehnke M, Danesh J,

- Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–291. <https://doi.org/10.1038/nature19057>
38. Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169. <https://doi.org/10.1093/bioinformatics/btu638>
39. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>
40. Li Q, Fisher K, Meng W, Fang B, Welsh E, Haura EB, Koomen JM, Eschrich SA, Fridley BL, Chen YA (2019) GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz488>
41. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100(16):9440–9445. <https://doi.org/10.1073/pnas.1530509100>
42. Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19(3):368–375. <https://doi.org/10.1093/bioinformatics/btf877>
43. Dudoit S, van der Laan MJ, Pollard KS (2004) Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol* 3:Article 13. <https://doi.org/10.2202/1544-6115.1040>
44. O'Brien DA, Gabel CA, Welch JE, Eddy EM (1991) Mannose 6-phosphate receptors: potential mediators of germ cell-Sertoli cell interactions. *Ann N Y Acad Sci* 637:327–339. <https://doi.org/10.1111/j.1749-6632.1991.tb27320.x>
45. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. <https://doi.org/10.1186/1471-2105-14-128>
46. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 44(W1):W90–W97. <https://doi.org/10.1093/nar/gkw377>
47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545–15550. <https://doi.org/10.1073/pnas.0506580102>
48. Wilkerson MD, Hayes DN (2010) ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26(12):1572–1573. <https://doi.org/10.1093/bioinformatics/btq170>
49. Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32(18):2847–2849. <https://doi.org/10.1093/bioinformatics/btw313>
50. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25(22):2906–2912. <https://doi.org/10.1093/bioinformatics/btp543>
51. Chambers MC, Jagtap PD, Johnson JE, McGowan T, Kumar P, Onsongo G, Guerrero CR, Barsnes H, Vaudel M, Martens L, Gruning B, Cooke IR, Heydarian M, Reddy KL, Griffin TJ (2017) An accessible proteogenomics informatics resource for cancer researchers. *Cancer Res* 77(21):e43–e46. <https://doi.org/10.1158/0008-5472.CAN-17-0331>
52. Kwon OK, Ha YS, Lee JN, Kim S, Lee H, Chun SY, Kwon TG, Lee S (2019) Comparative proteome profiling and mutant protein identification in metastatic prostate cancer cells by quantitative mass spectrometry-based proteogenomics. *Cancer Genom Proteom* 16(4):273–286. <https://doi.org/10.21873/cgp.20132>
53. Zhan X, Cheng J, Huang Z, Han Z, Helm B, Liu X, Zhang J, Wang TF, Ni D, Huang K (2019) Correlation analysis of histopathology and proteogenomics data for breast cancer. *Mol Cell Proteom*. <https://doi.org/10.1074/mcp.RA118.001232>
54. Mertins P, Tang LC, Krug K, Clark DJ, Gritsenko MA, Chen L, Clauser KR, Clauss TR, Shah P, Gillette MA, Petyuk VA, Thomas SN, Mani DR, Mundt F, Moore RJ, Hu Y, Zhao R,

- Schnaubelt M, Keshishian H, Monroe ME, Zhang Z, Udeshi ND, Mani D, Davies SR, Townsend RR, Chan DW, Smith RD, Zhang H, Liu T, Carr SA (2018) Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nat Protoc* 13(7):1632–1661. <https://doi.org/10.1038/s41596-018-0006-9>
55. Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi DV, Edwards NJ, Thanagudu RR, Ketchum KA, Kinsinger CR, Mesri M, Rodriguez H, Stein SE (2016) A description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) common data analysis pipeline. *J Proteome Res* 15 (3):1023–1032. <https://doi.org/10.1021/acs.jproteome.5b01091>
56. Tabb DL, Wang X, Carr SA, Clauser KR, Mertins P, Chambers MC, Holman JD, Wang J, Zhang B, Zimmerman LJ, Chen X, Gunawardena HP, Davies SR, Ellis MJ, Li S, Townsend RR, Boja ES, Ketchum KA, Kinsinger CR, Mesri M, Rodriguez H, Liu T, Kim S, McDermott JE, Payne SH, Petyuk VA, Rodland KD, Smith RD, Yang F, Chan DW, Zhang B, Zhang H, Zhang Z, Zhou JY, Liebler DC (2016) Reproducibility of differential proteomic technologies in CPTAC fractionated xenografts. *J Proteome Res* 15(3):691–706. <https://doi.org/10.1021/acs.jproteome.5b00859>
57. Wu P, Heins ZJ, Muller J, Katsnelson L, de Brujin I, Abeshouse AA, Schultz N, Fenyo D, Gao J (2019) Integration and analysis of CPTAC proteomics data in the context of cancer genomics in the cBioPortal. *Mol Cell Proteom.* <https://doi.org/10.1074/mcp.TIR119.001673>



# Chapter 12

## Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models

Tianyu Zhang, Liwei Zhang, Philip R. O. Payne, and Fuhai Li

### Abstract

Intrinsic and acquired drug resistance is a major challenge in cancer therapy. Synergistic drug combinations could help to overcome drug resistance. However, the number of possible drug combinations is enormous, and it is infeasible to experimentally screen all drug combinations with limited resources. Therefore, computational models to predict and prioritize effective drug combinations are important for combination therapy discovery. Compared with existing models, we propose a novel deep learning model, AuDNNsynergy, to predict the synergy of pairwise drug combinations by integrating multiomics data. Specifically, three autoencoders are trained using the gene expression, copy number, and genetic mutation data of tumor samples from The Cancer Genome Atlas (TCGA). Then the gene expression, copy number, and mutation of individual cancer cell lines are coded using the three trained autoencoders. The physicochemical features of individual drugs and the encoded omics data of individual cancer cell lines are used as the input features of a deep neural network that predicts the synergy score of given pairwise drug combinations against the specific cancer cell lines. The comparison results showed the proposed AuDNNsynergy model outperforms, specifically in terms of rank correlation metric, four state-of-the-art approaches, namely, DeepSynergy, Gradient Boosting Machines, Random Forests, and Elastic Nets.

**Key words** Prediction methods, Multiomics, Deep learning models

---

### 1 Introduction

Synergistic combination therapies are important for overcoming intrinsic or acquired drug resistance in cancer therapy [1–4]. Many drug combinations have been reported to be synergy to inhibit tumor cell growth in preclinical models. For example, PI3K and HDAC inhibitors are synergistic for Group 3 Medulloblastoma [5]. The LKB1 inactive, non-small cell lung cancer model is sensitive to mTOR and PI3K inhibitors [6]. Vemurafenib and tretinoin are potentially synergistic, predicted by a computational model, for drug resistant melanoma [7]. Sorafenib and bevacizumab are potentially synergistic for ovarian cancer [8]. Drug combinations to inhibit RAS signaling and autophagy are reported to be synergistic in pancreatic cancer models [9, 10]. Moreover, the targeted

therapies (e.g., BRAF, ERK inhibitors) and immune checkpoint blockade can increase the response rates and durability in melanoma models [11]. In addition, the targeted therapies (e.g., gefitinib and erlotinib) can increase the sensitivity of chemotherapy in non-small cell lung cancer models [12]. The mechanisms of synergy of drug combinations are diverse and heterogeneous [3, 13, 14]. High-throughput screening (HTS) can be used to identify individual drugs [5], but it is limited to experimentally identify effective drug combinations from the enormous number of possible combinations. Therefore, computational models to predict effective drug combinations for specific cancer subtypes or individual patients are important and urgently needed. To facilitate the development of computational models of drug combination prediction, increasing numbers of datasets of drug combinations have been generated. For example, 178 drug combinations were collected from US Food and Drug Administration (FDA) Orange Book [15], and 239 drug combinations were derived from literature reports [16]. The Drug Combination DataBase (DCDB) database [17] collected 1363 drug combinations from the [ClinicalTrials.gov](#) database, FDA Orange Book, and PubMed database [17]. Moreover, experimental screening of combinations of 39 drugs against 38 cell lines were reported by Merck Research Laboratories [18]; and drug combination screening of 104 drugs against 59 cell lines was conducted by Dr. Doroshow's group in Division of Cancer Treatment and Diagnosis, National Cancer Institute [2]. The data of these two drug combination screening projects are publicly available.

Despite a few reported computational approaches reported, drug combination prediction remains an open problem. Connectivity Map (CMap) data [19, 20] is the most widely used resource for drug and drug combination prediction. Specifically, about 70 cancer cell lines are treated by over 3000 drugs and compounds. The gene expression data of these cancer cell lines before and after drug treatment are collected. Then the upregulated genes and downregulated genes after drug treatment are used as gene expression "signatures" of given drugs to characterize the mechanism of action of drugs [21]. The gene set enrichment analysis (GSEA) [22] approach was introduced to associate drugs with diseases based on the gene expression signatures. In other words, drugs that can downregulate the upregulated genes in disease (compared with normal tissue) are top-ranked to be effective for treatment of the disease. In some studies, drug combinations are prioritized based on drug clustering, using Connectivity Map data [19, 20], and target distribution on the disease signaling networks that are derived based on the disease genomics data [23–26]. Based on the CMAP gene expression signatures of drugs, two drugs with distinct signatures that are both related to the same disease genes are hypothesized to be synergistic [27]. These models define general

rules based on the biological knowledge for drug combination prediction but do not utilize conventional experimental screening experiments. In another example, a supervised model (using the experimental screening data to train the parameters of the model), taking the gene expression and copy number as the input features, was proposed to predict drug combination synergy [28]. As compared to the supervised model, a semisupervised learning-based model was developed to predict synergistic drug combinations [29]. The difference between supervised and semisupervised prediction models are as follows. In the supervised model, only the training data are used to train the model (i.e., learning the model parameters), and the trained model is then used to predict the label of testing data (samples in testing dataset will not affect the prediction model). In the semisupervised model, both the training dataset and testing dataset are used to predict the labels of samples in the testing dataset.

Recently, deep learning models, which are formulated as a multilayer artificial neural network, have generated the best prediction results in image analysis [30], natural language processing (NLP) [31], and medical informatics [32]. Deep learning models are one-of-a-kind machine learning models. However, compared with the traditional machine learning models, deep learning models are formatted in the artificial neural network and have enormous number of parameters. Thus, the large number of parameters can make the model complex enough to learn the complex prediction tasks. Deep learning models also have been developed for drug combination prediction. For example, the DeepSynergy model was proposed in [33] using a deep neural network (DNN), and the comparison results indicated the deep learning model outperforms other traditional machine learning models.

This methods paper will describe a novel deep learning model, *AuDNNsynergy* (*Deep Neural Network Synergy model with Autoencoders*), for drug combination prediction by integrating multiomics data of a large set of tumor samples from The Cancer Genome Atlas (TCGA), a dataset including genomics data, including gene expression, copy number variation, methylation, mutation, and proteomics data, of over 30,000 cancer samples. In this model, multiomics datasets (i.e., gene expression, mutation, copy number variation), are encoded by using three autoencoders that can potentially remove the tissue specific bias. Then the encoded multiomics data of individual cell lines and the physicochemical features of individual drugs are used as the input of a deep neural network to predict the synergy scores of given drug pairs. The comparison results indicated that the AuDNNsynergy model outperformed four state-of-the-art approaches. In this paper, we introduce the AuDNNsynergy deep learning model, and use it as an example to predict drug combinations using multiomics data.

---

## 2 Materials

### 2.1 Drug Combination Screening Dataset

The high-throughput drug combination screening dataset is downloaded as the drug screening dataset [18]. Specifically, 583 pairwise drug combinations of 38 individual drugs were screened against 39 human cancer cell lines. The 39 cell lines include seven types of cancers: lung, breast, skin, large intestine, pleura, prostate, and ovary. Tumor cell viability after 48 h of drug treatments (four doses for each drug) was measured to estimate the synergy of drug combinations. In total, there are 23,062 data points.

### 2.2 Omics Data of Cancer Cell Lines and TCGA Samples

RNA-seq data of the 1156 cancer cell lines and 10,535 tumor samples from The Cancer Genome Atlas (TCGA) are downloaded from Cancer Cell Line Encyclopedia (CCLE) and UCSC Xena Server [34] (URL: <https://xenabrowser.net/datapages/?dataset>). The somatic mutation data of the screened cell lines and 9104 TCGA tumor samples are downloaded from Catalogue of Somatic Mutations in Cancer (COSMIC) [35] and UCSC Xena web server (<https://xenabrowser.net/>) respectively. The copy number of 10,845 TCGA pan-cancer tumors and the 39 cancer cell lines can be accessed from UCSC Xena and CCLE respectively. Some cancer cell lines data cannot be found in CCLE and COSMIC. We used the genomics data of EFM-129A and COLO320 for the EFM-129B and COLO320DM cell lines, respectively. Briefly, the EFM-129B was established from the same person of EFM-129A after 14 days, and the COLO320DM sample is a derivative cell line of COLO320.

### 2.3 Physicochemical Features of Drugs

To characterize the chemical structures of individual drugs, we used the counts of extended connectivity fingerprints with a radius of 6 (ECFP\_6) [36]. ECFP is the most widely used topological fingerprints for molecular characterization, and substructure and similarity searching [36]. In this method we utilize the predefined physicochemical properties (e.g., solubility and passive permeability), and presence or absence of toxicophore substructures of compounds, which are known to be toxic. These features can be calculated using jCompoundMapper [37] and Chemopy [38], respectively.

---

## 3 Methods

In this section, we will introduce the methods of using a deep learning model, AuDNNsynergy, to predict the synergy of drug combinations. We will (1) preprocess the genomics data, (2) design the structure of AuDNNsynergy model, and (3) evaluate and compare the model performance.

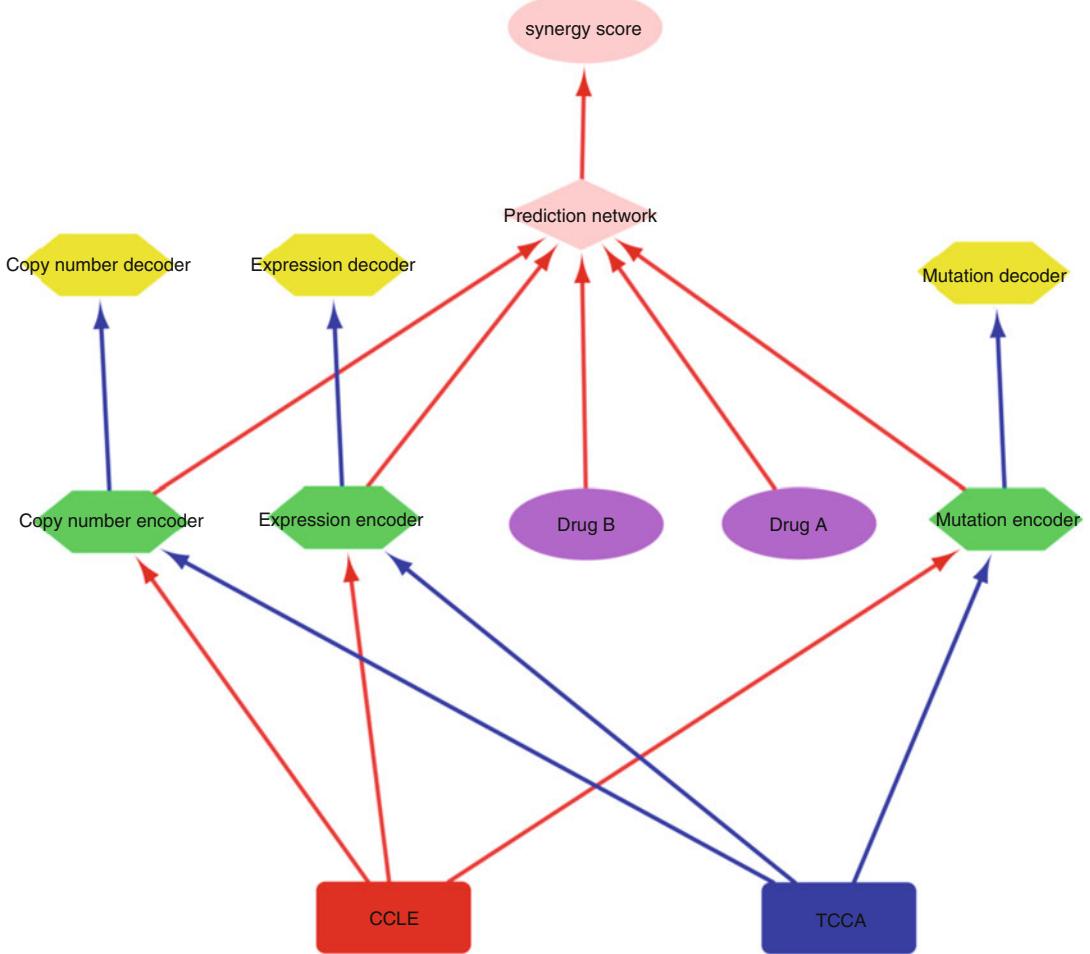
### 3.1 Data Preprocessing

The gene expression data are normalized as  $\log_2(\text{tpm} + 0.001)$ , where tpm denotes the number of transcripts per million. The mutation data (only non-silent mutations are used) is formalized as a binary variable: 1 (with mutation) and 0 (no mutation). Copy number variation (CNV) are estimated using GISTIC2 method [39]. For synergy score of drug combinations, the Loewe Additivity model is used [40]. All features are filtered by removing elements with zero variance and then normalized to have zero-mean and unit variance. We denote  $E^T, E^C, M^T, M^C, C^T$  as processed data, where  $E$ ,  $M$ , and  $C$  represent gene expression, mutation, copy number and drug, and the T and C indicate tumor and cell lines. For example,  $E^T$  represents TCGA expression data after preprocessing. After preprocessing, each cell line contained 50,196, 17,643, and 20,318 features of expression, mutation, and copy number, and each drug represented using the 2431 physicochemical features, denoted by  $D$ . In total, each drug combination against a given cancer cell line has 93,019 features.

### 3.2 Model Architecture of AuDNNsynergy

The AuDNNsynergy model consists of two major components: three autoencoders and a deep neural network, as shown in Fig. 1.

*Three autoencoders.* The three autoencoders, expression autoencoder ( $\text{En}_{\text{ex}}$ ), mutation autoencoder ( $\text{En}_{\text{mu}}$ ), and copy number autoencoder ( $\text{En}_{\text{co}}$ ), are trained using TCGA multiomics data to transfer the knowledge embedded in the large-scale genomics data of TCGA samples. The cell lines' gene expression, copy number, and mutation data are then encoded using the trained three autoencoders, and the outputs are used as the input of the prediction network. Specifically, we applied six densely connected layers neural network for each autoencoder. Moreover, to choose the optimal autoencoder structures and parameters, the numbers of neurons and batch sizes are set as hyperparameters, resulting in 20 alternatives of autoencoders. The architecture of autoencoders is symmetric with input, output layers and five hidden layers having the number of neurons selected from [8192, 4096, 2048, 1024, 512], and the batch size selected from [512, 256]. The linear activations are used in the output layer and bottleneck layer, and the rectified linear activations are used for other layers. Wikipedia lists the detailed definitions of a list of activation functions (URL: [https://en.wikipedia.org/wiki/Activation\\_function](https://en.wikipedia.org/wiki/Activation_function)). The mean square error (MSE) is used as the cost function minimized by Adam optimizer with default learning rate. The uniform distribution is used for initialization, and the variation of moving average over five epochs on the validation set is used as early-stopping criterion for training iterations. Each epoch means one training iteration (to update model parameters multiple times) using all training samples. The early-stopping



**Fig. 1** Overview of the AuDNNsynergy architecture

process is used to avoid the overfitting problem, which means that the model has good performance on training data, but has poor performance on the testing data.

*Deep neural network.* The number of layers, the number of neurons, and the learning rate are set as hyperparameters. The number of layers is selected from [2, 3]; the number of neurons is selected from [8192, 4096, 2048, 1024, 512]; and learning rate (to update the model parameters) is selected from [ $10^{-1}$ ,  $10^{-3}$ ,  $10^{-5}$ ]. The linear activations are used in the output layer, and rectified linear activations are used in the hidden layers. The batch size of 64, a dropout rate of 0.2 and 0.5 for the input and hidden layers, and stochastic gradient descent optimizer with momentum 0.5 (which considers the previous gradient information (momentum) and current gradient information) to minimize MSE cost function are used. The

uniform distribution is used for initialization, and a moving average over 15 epochs on a validation set early-stopping is used for training iterations. Mathematically, the output  $\alpha^i$  of layer  $i$  is calculated by  $\alpha^i = A^i(W^i\alpha^{i-1} + b^i)$ , where  $\alpha^i$  denotes the production of layer  $i$  ( $\alpha^0$  is the input),  $A^i$ ,  $W^i$  and  $b^i$  are the activation function, weight matrix and bias vector of layer  $i$  respectively. The weights and biases are learned to minimizes the MSE cost function.

### **3.3 Model Evaluation and Comparison of Drug Combination Prediction**

#### **3.3.1 Model Computational Environment**

#### **3.3.2 Performance Evaluation of AuDNNsynergy**

#### **3.3.3 Synergistic Drug Combination Prediction for TCGA Samples**

We deployed the model on Google Cloud Platform with system Ubuntu 16.04, four vcpu, 32G memory, and NVIDIA Tesla V100 GPU. The AuDNNsynergy is run under CUDA 9.0, cuDNN 9.0 using Keras 2.1 with GPU supporting Tensorflow 1.4 backend.

The optimal hyperparameters of the three autoencoders are as follows. For the gene expression autoencoder,  $En_{ex}$ , the number of neurons in three hidden layers are: 8192, 2048, and 512, respectively, with the batch size 256. For the mutation and copy number autoencoders,  $En_{mu}$  and  $En_{co}$ , the number of neurons in the three hidden layers are the same: 8192, 2048, and 1024 neurons respectively, with a batch size 256. The two-layer deep neuron network with 8192 and 2048 neurons, with the  $10^{-5}$  learning rate, has the best performance. Fivefold cross-validation is used to evaluate the performance of different models. The performance of the models is measured using the rank correlation coefficient and mean squared error (MSE) metrics between the experimental synergy scores and the predicted synergy scores. The rank correlation coefficient is practically more important than the MSE in terms of prioritizing the most effective drug combinations for the design of further experimental evaluations.

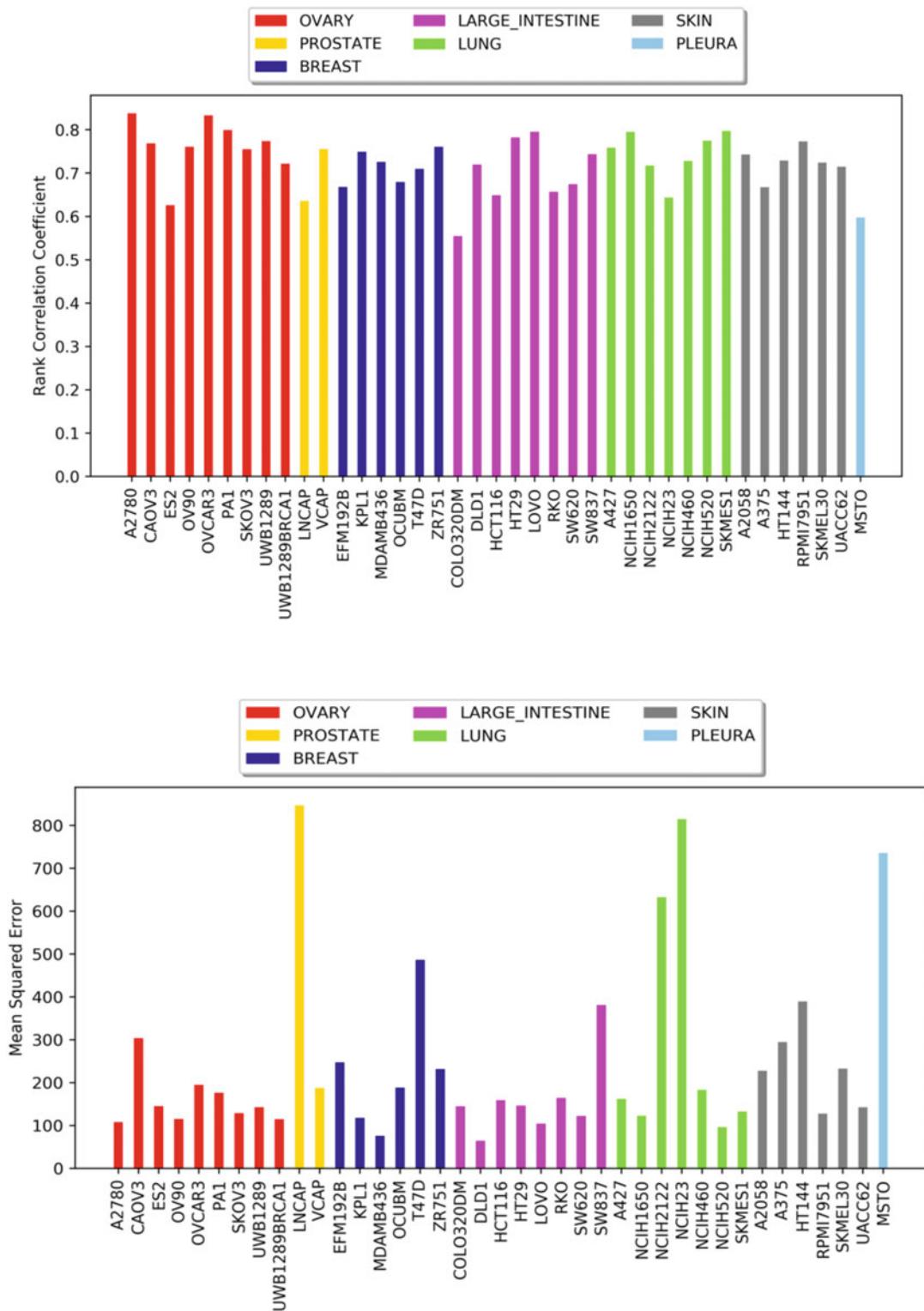
We also applied AuDNNsynergy to predict synergy scores of the 583 drug combinations on TCGA samples. Although the omics profiles of cell lines and tumors might be different, some prediction results are still meaningful. For example, drug combination dasatinib–erlotinib has a high synergy score (rank 7/583) for NSCLC harboring an EGFR mutation. Dasatinib is an inhibitor of SRC while erlotinib is an EGFR inhibitor. The dasatinib–erlotinib combinations has been reported to be synergistic, safe for NSCLC with activating EGFR mutations [41]. The other examples are the ABT-888/BEZ-235, MK-4827/BEZ-235, and MK4827/MK8669 combinations, which are top ranked in the prediction (top 4–7%). ABT-888 and MK-4827 target PARP while BEZ-235

and MK-8669 target MTOR in breast cancer samples. In [42], everolimus and olaparib combination showed synergy for the treatment of a BRCA2 mutated luminal breast cancer patient-derived xenograft. Interestingly, olaparib is a PARP inhibitor while everolimus is a mTOR inhibitor, and their combination could lead to unrepaired DNA damage and tumor regression *in vivo*, through a cross talk between DNA repair and mTOR pathways [42]. These examples indicated that the proposed model can be used to predict drug combinations for new tumor samples not included in the 39 cancer cell lines in the screening data. The training of the three autoencoders using the TCGA samples might be useful to transfer and identify important omics features to generalize the model on new tumor samples.

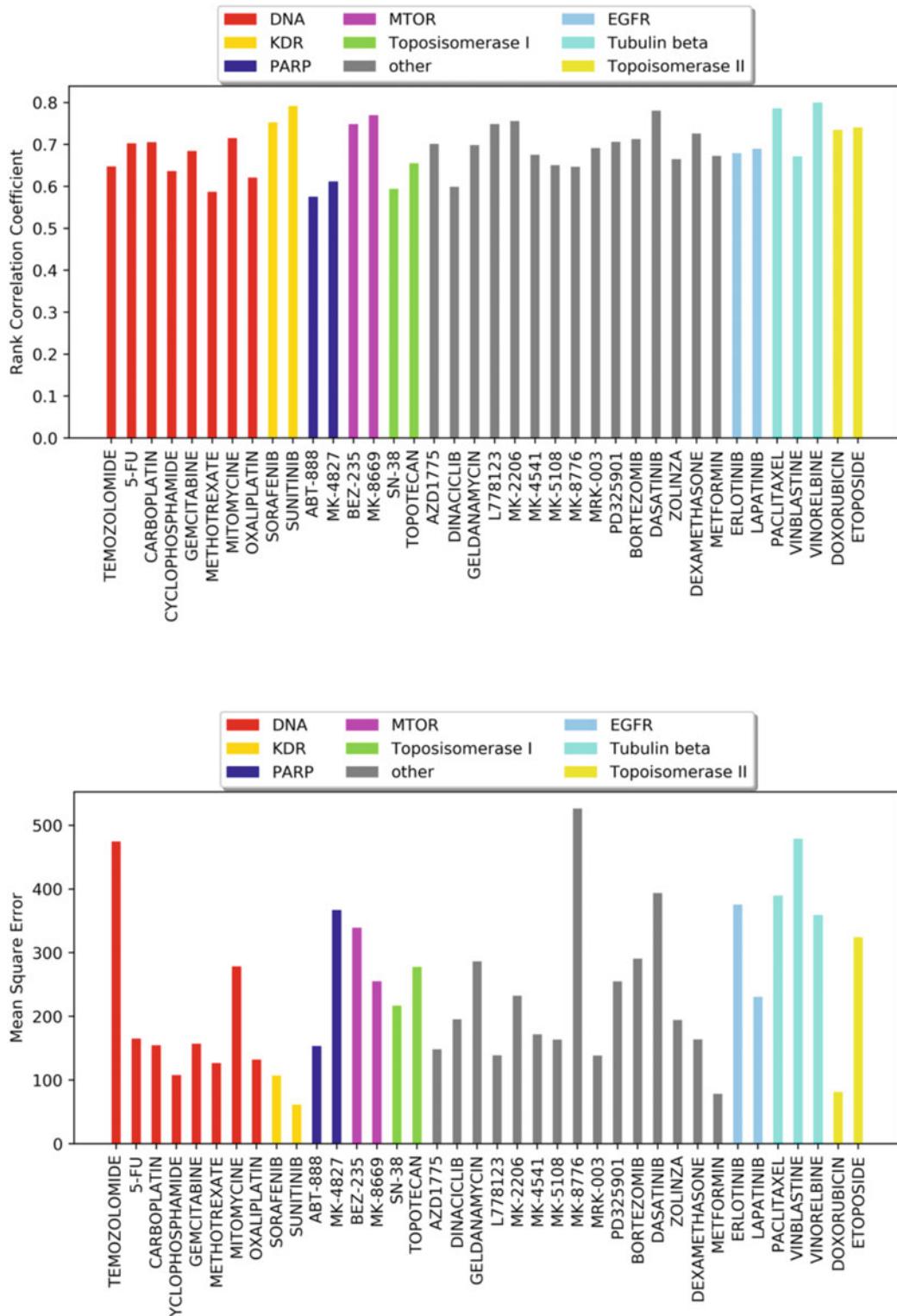
The prediction results, in terms of rank correlation and MSE, of the AuDNNsynergy model are shown in Figs. 1 and 2. Figure 3 shows the rank correlation coefficient and MSE of the prediction per cancer cell line. As shown, there are a few cell lines with very high MSE values. We investigated the cancer cell lines with great MSE values, and found that these cell lines: LNCAP, MSTO, NCIH2122 HT144, T47D are all metastatic, SW837 is high grade (grade IV) rectal cancer, which might be more drug resistant. The NCIH23 (not metastatic) lung cancer cell line also has a very high MSE. Also, the prediction accuracy in terms of MSE varies more than the rank correlation coefficient metric among cell lines, and the MSE and rank correlation coefficient are not always consistent, like the prediction results of T47D and SW837 cell lines. It indicates the challenge of predicting the complex drug response of individual cancer cell line.

For individual drugs, the rank correlation coefficients ranged from 0.55 to 0.86. Only two drugs (methotrexate and ABT-888) have rank correlation coefficients below 0.6, and ~67% of the drugs had rank correlation coefficients higher than 0.7. The MSEs for drugs varies from 64 to 846. Four drugs (temozolomide, MK-8776, dasatinib, and vinblastine) have MSEs larger than 400. The large MSE of drugs might be caused by the “block” effects of these drugs (effective or resistant for most cell lines) [43]. About 67% of the drugs have MSEs lower than 200. Specifically, sunitinib and doxorubicin are relatively easy to predict, with rank correlation above 0.7 and MSEs <100.

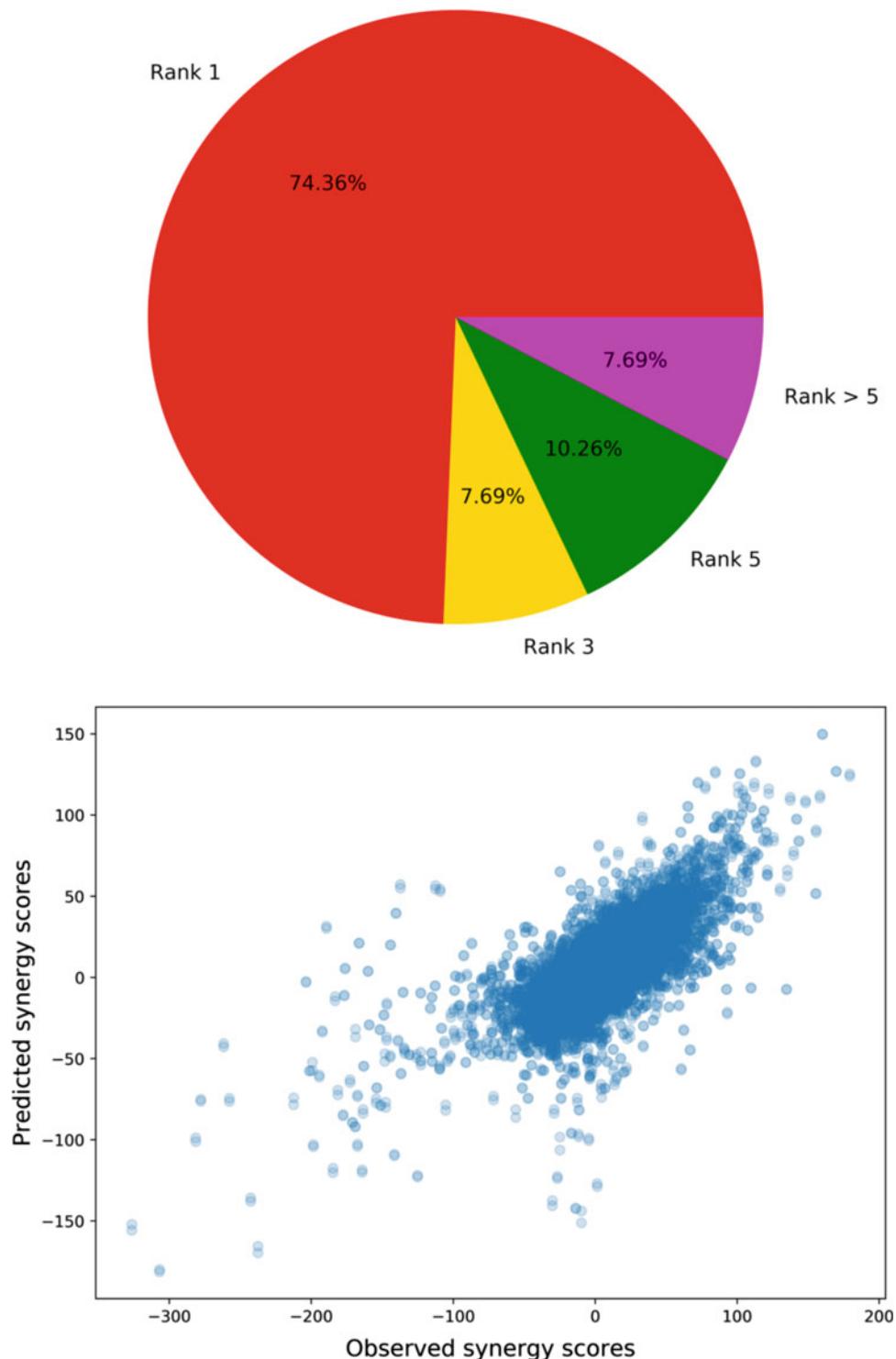
Moreover, we ranked observed synergy scores and compared the drug pairs with the highest predicted scores. Interestingly, there are about 74.36% of synergistic drug combinations in the drug combination screening dataset are top-ranked in the prediction of AuDNNsynergy (*see* Fig. 4-top), which indicates that the prediction model is reliable to rank and prioritize synergistic drug combinations. We also investigated the correlation between the prediction and experimental synergy scores (*see* Fig. 4-bottom),



**Fig. 2** Rank correlation coefficient (top) and MSE (bottom) of the AuDNNsynergy prediction per cancer cell line



**Fig. 3** Rank correlation coefficient (top) and MSE (bottom) of the AuDNNsynergy prediction per drug



**Fig. 4** Ranking correlation between prediction and experimental validations (top); and the overall correlation of predicted and the experimental validated synergy scores

which indicates the high correlation between the prediction and experiment synergy scores. Therefore, the rank correlation should be a better metrics for predicting and ranking effective drug combinations.

### 3.3.4 Model Comparison

The software package in this methods paper can be compared with four existing models, including DeepSynergy [33], Gradient Boosting Machines [44], Random Forests [45], and Elastic Nets [46]. Fivefold cross-validation is used to evaluate the performance of different models. Specifically, all these models, except DeepSynergy, are evaluated using the 93,019 features generated in this study. The DeepSynergy model is evaluated on the 8846 features reported in [33] because the model taking 93,019 features as input is too complex (too many parameters) to be trained using NVIDIA Tesla V100 GPU. Table 1 shows the tested hyperparameters for these models, and the best performance are used for the comparison.

## 4 Summary of the AuDNNsynergy Model

Within this paper, we illustrated the usage of a deep learning model, AuDNNsynergy, to predict drug combinations based on multiomics data and chemical structure features. Taking the genomics data of patients, the deep learning model can predict a list of drug combinations that might be effective for the treatment of individual patients. Compared with existing models, multiomics data, that is, gene expression (RNA-seq, copy number, and mutation), are integrated using three autoencoders (that reduce the large number of genomics data (high dimension) into a relatively small number of variables (low dimension)). Also, the three autoencoders are trained using the multiomics data of about 10,000 tumor samples from TCGA data, which might be able to identify the important metafactors across over 30 types of cancer. We compared the

**Table 1**  
**Hyperparameter settings of four drug combination prediction methods**

Method	Hyperparameter
DeepSynergy	Number of hidden layers neurons $\in \{[8192, 4096], [8192, 8192], [8192, 2048]\}$
Gradient boosting machines	Number of estimators $\in [128, 256, 512, 1024]$ features considered $\in [\log_2(\# \text{ of features}), \text{square}(\# \text{ of features})]$
Random forests	Number of estimators $\in [128, 256, 512, 1024]$ features considered $\in [\log_2(\# \text{ of features}), \text{square}(\# \text{ of features})]$
Elastic nets	$\alpha \in [0.1, 1, 10, 100]$ $L_1$ ratio $\in [0.25, 0.5, 0.75]$

AuDNNsynergy model with other four existing methods. The comparison results indicated that the AuDNNsynergy model is better because it integrates more genomics features, similar gene expression profiles, copy number variation, and genetic mutations. AuDNNsynergy also outperforms the other existing models specifically in terms of rank correlation coefficient.

Precision medicine aims to identify effective drugs or drug combinations for individual patients. However, only a few genetic mutations have been identified as druggable targets, and there is no actionable target for a significant number of patients. Therefore, sophisticated computational models are urgently needed to integrate the genomics data of patients with existing big data of pharmacogenomics, and to predict the potential effective drugs and drug combinations for cancer patients who are resistant to given treatments.

Table 2 shows the detailed model comparison results in terms of the following metrics: rank correlation coefficient, MSE, root mean square error (RMSE), and Pearson correlation. The models are evaluated on the 93,019 features (multiomics data) and the 8846 features (\*marked) respectively. As shown in Table 2, the proposed AuDNNsynergy model outperforms the other four models specifically in terms of the rank correlation (0.73), which is important for drug combination prediction (to rank the drug combinations for further experimental validation). For the other performance metrics, the AuDNNsynergy model is slightly better than the other models. In addition, Table 3 shows performance evaluation using the area under the receiver operator characteristics curve (ROC AUC), area under the precision–recall curve (PR AUC),

**Table 2**  
**Rank correlation, MSE, confidence interval, RMSE and Pearson correlation for five methods on the 93,019 features dataset, and the 8846 features dataset**

Method	Rank correlation	MSE	RMSE	Pearson correlation
AuDNNsynergy	$0.73 \pm 0.02$	$241.12 \pm 43.52$	$15.46 \pm 1.44$	$0.74 \pm 0.03$
Deep synergy	$0.66 \pm 0.02$	$255.49 \pm 49.54$	$15.91 \pm 1.56$	$0.73 \pm 0.04$
Gradient boosting machines	$0.64 \pm 0.02$	$310.73 \pm 50.40$	$17.57 \pm 1.42$	$0.64 \pm 0.02$
Gradient boosting machines	$0.64 \pm 0.01$	$308.06 \pm 49.34$	$17.50 \pm 1.39$	$0.64 \pm 0.02$
Random forests	$0.64 \pm 0.02$	$326.31 \pm 53.54$	$18.00 \pm 1.45$	$0.63 \pm 0.02$
Random forests	$0.63 \pm 0.02$	$330.18 \pm 51.45$	$18.12 \pm 1.37$	$0.61 \pm 0.03$
Elastic nets	$0.49 \pm 0.02$	$422.72 \pm 54.35$	$20.52 \pm 1.30$	$0.44 \pm 0.02$
Elastic nets	$0.49 \pm 0.02$	$424.38 \pm 54.39$	$20.56 \pm 1.30$	$0.44 \pm 0.02$

**Table 3**  
**ROC AUC, PR AUC, ACC, PREC, and Kappa for five methods**

Performance metrics	ROC AUC	PR AUC	ACC	PREC	Kappa
AuDNNsynergy	0.91 ± 0.02	0.63 ± 0.06	0.93 ± 0.01	0.72 ± 0.06	0.51 ± 0.04
Gradient boosting machines	0.87 ± 0.02	0.54 ± 0.04	0.93 ± 0.01	0.72 ± 0.03	0.39 ± 0.05
Random forests	0.86 ± 0.03	0.51 ± 0.04	0.92 ± 0.02	0.57 ± 0.04	0.21 ± 0.04
Elastic nets	0.77 ± 0.04	0.33 ± 0.09	0.92 ± 0.01	0.23 ± 0.29	0.15 ± 0.09
DeepSynergy*	0.90 ± 0.03	0.59 ± 0.06	0.92 ± 0.03	0.56 ± 0.11	0.51 ± 0.04

accuracy (ACC), precision (PREC), and Cohen's Kappa. As can been seen, the proposed model has slightly better performance in terms of all the metrics. Also, compared with DeepSynergy, the proposed model has much better precision.

## Acknowledgments

The authors would like to thank Institute for Informatics (I2) colleagues for the helpful discussions. This study is partially supported by the start-up funding, Institute for Informatics (I2), and department of Pediatrics, and Children's Discovery Institute II (MI-II-2019-802).

## References

- Humphrey RW, Brockway-Lunardi LM, Bonk DT et al (2011) Opportunities and challenges in the development of experimental drug combinations for cancer, *Journal of the National Cancer Institute*, vol 103. Oxford University Press, Oxford, pp 1222–1226
- Holbeck SL, Camalier R, Crowell JA et al (2017) The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res* 77:3564–3576
- Ahronian LG, Corcoran RB (2017) Strategies for monitoring and combating resistance to combination kinase inhibitors for cancer therapy. *Genome Med* 9(1):37
- Nelander S, Wang W, Nilsson B et al (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol* 4:216
- Pei Y, Liu KW, Wang J et al (2016) HDAC and PI3K antagonists cooperate to inhibit growth of MYC-driven medulloblastoma. *Cancer Cell* 29:311–323
- Shukuya T, Yamada T, Koenig MJ et al (2019) The effect of LKB1 activity on the sensitivity to PI3K/mTOR inhibition in non-small cell lung cancer. *J Thorac Oncol* 14:1061–1076
- Regan-Fendt KE, Xu J, DiVincenzo M et al (2019) Synergy from gene expression and network mining (SynGeNet) method predicts synergistic drug combinations for diverse melanoma genomic subtypes. *NPJ Syst Biol Appl* 5:6
- Azad NS, Posadas EM, Kwitkowski VE et al (2008) Combination targeted therapy with sorafenib and bevacizumab results in enhanced toxicity and antitumor activity. *J Clin Oncol* 26:3709–3714
- Kinsey CG, Camolotto SA, Boespflug AM et al (2019) Protective autophagy elicited by RAF→MEK→ERK inhibition suggests a treatment strategy for RAS-driven cancers. *Nat Med* 25(4):620–627

10. Bryant KL, Stalnecker CA, Zeitouni D et al (2019) Combination of ERK and autophagy inhibition as a treatment approach for pancreatic cancer. *Nat Med* 25(4):628–640
11. Prieto PA, Reuben A, Cooper ZA, Wargo JA (2016) Targeted therapies combined with immune checkpoint therapy. *Cancer J* 22 (2):138–146
12. Maiione P, Gridelli C, Troiani T, Ciardiello F (2006) Combining targeted therapies and drugs with multiple targets in the treatment of NSCLC. *Oncologist* 11:274–284
13. Zhang T, Xu J, Deng S et al (2018) Core signaling pathways in ovarian cancer stem cell revealed by integrative analysis of multi-marker genomics data. *PLoS One* 13(5):e0196351
14. Massarelli E, Varella-Garcia M, Tang X et al (2007) KRAS mutation is an important predictor of resistance to therapy with epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. *Clin Cancer Res* 13(10):2890–2896
15. Zhao XM, Iskar M, Zeller G, Kuhn M, van Noort V, Bork P (2011) Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Comput Biol* 7 (12):e1002323
16. Huang H, Zhang P, Qu XA, Sanseau P, Yang L (2014) Systematic prediction of drug combinations based on clinical side-effects. *Sci Rep* 4:7160
17. Liu Y, Wei Q, Yu G, Gai W, Li Y, Chen X (2014) DCDB 2.0: a major update of the drug combination database. *Database* 2014: bau124
18. O’Neil J, Benita Y, Feldman I et al (2016) An unbiased oncology compound screen to identify novel combination strategies. *Mol Cancer Therap* 15:1155–1162
19. Lamb J, Crawford ED, Peck D et al (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313:1929–1935
20. Subramanian A, Narayan R, Corsello SM et al (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171:1437–1452
21. Li F, Wang L, Kong R, Sheng J, Cao H, Mancuso J, Xia X et al (2016) DrugMoaMiner: a computational tool for mechanism of action discovery and personalized drug sensitivity prediction. *IEEE*
22. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545–15550
23. Huang L, Li F, Sheng J et al (2014) DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics* 30: i228–i236
24. Xu J, Regan-Fendt K, Deng S, Carson WE, Payne PRO, Li F (2018) Diffusion mapping of drug targets on disease signaling network elements reveals drug combination strategies. *Pac Symp Biocomput* 23:92–103
25. Wu H, Miller E, Wijegunawardana D, Regan K, Payne PRO, Li F (2017) MD-Miner: a network-based approach for personalized drug repositioning. *BMC Syst Biol* 11:86
26. Regan KE, Payne PRO, Li F (2017) Integrative network and transcriptomics-based approach predicts genotype-specific drug combinations for melanoma. *AMIA Joint Summits on Translational Science Proceedings*. AMIA Joint Summits on Translational Science, pp 247–256
27. Lee JH, Kim DG, Bae TJ et al (2012) CDA: combinatorial drug discovery using transcriptional response modules. *PLoS One* 7:e42573
28. Li H, Li T, Quang D, Guan Y (2018) Network propagation predicts drug synergy in cancers. *Cancer Res*. <https://doi.org/10.1158/0008-5472.CAN-18-0740>
29. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G (2016) NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput Biol* 2(7):e1004975
30. Krizhevsky A, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Neural Inf Process Syst*. <https://doi.org/10.1145/3065386>
31. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *J ArXiv abs/1810.04805*
32. Rajkomar A, Oren E, Chen K et al (2018) Scalable and accurate deep learning with electronic health records. *NPJ Dig Med* 1:18
33. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G (2018) DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* 34:1538–1546
34. Mary Goldman BC, Brooks AN, Zhu J, Haussler D (2018) The UCSC Xena Platform for cancer genomics data visualization and interpretation. *biorxiv*
35. Forbes SA, Beare D, Boutsikakis H et al (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 45(D1): D777–D783

36. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754
37. Hinselmann G, Rosenbaum L, Jahn A, Fechner N, Zell A (2011) JCompoundMapper: an open source Java library and command-line tool for chemical fingerprints. *J Cheminf* 3:3
38. Cao DS, Xu QS, Hu QN, Liang YZ (2013) ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29(8):1092–1094
39. Mermel CH, Schumacher SE, Hill B, Meyer-son ML, Beroukhim R, Getz G (2011) GIS-TIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12(4):R41
40. Ianevski A, He L, Aittokallio T, Tang J (2017) SynergyFinder: a web application for analyzing drug combination dose-response matrix data. *Bioinformatics* 33(15):2413–2415
41. Gold KA, Lee JJ, Harun N et al (2014) A phase I/II study combining erlotinib and dasatinib for non-small cell lung cancer. *Oncologist* 19:1040–1041
42. Botty RE, Coussy F, Hatem R et al (2018) Inhibition of mTOR downregulates expression of DNA repair proteins and is highly efficient against BRCA2-mutated breast cancer in combination to PARP inhibition. *Oncotarget* 9:29587–29600
43. Sheng J, Li F, Wong STC (2015) Optimal drug prediction from personal genomics profiles. *IEEE J Biomed Health Inf* 19:1264–1270
44. Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38:367–378
45. Ho TK (1995) Random decision forests, Proceedings of 3rd International Conference on Document Analysis and Recognition (vol. 1, pp. 278–282), Montreal, Quebec, Canada. <https://doi.org/10.1109/ICDAR.1995.598994>
46. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B Stat Methodol* 67:301–320



# Chapter 13

## Introduction to Multiparametric Flow Cytometry and Analysis of High-Dimensional Data

James Sun, Jodi L. Kroeger, and Joseph Markowitz

### Abstract

Multiparametric flow cytometry is a technique utilized in translational experiments that utilizes fluorescently tagged antibodies and functional fluorescent dyes to measure proteins on the surface or in the cytoplasm of cells and to measure processes occurring within cells themselves. These fluorescent molecules, or fluorophores, can be tagged to antibodies to measure specific biological molecules such as proteins inside or on the surface of cells. Small organic compounds such as the nucleic acid binding dye propidium iodide (PI) can permeate compromised cell membranes when cells are no longer viable or used to measure DNA content of cycling cells. Successful completion of flow cytometry experiments requires expertise in both the preparation of the samples, acquisition of the samples on instruments, and analyses of the results. This chapter describes the principles needed to conduct a successful multiparameter flow cytometry experiment needed for drug development with references to well established internet resources that are useful to those less experienced in the field. In addition, we provide a brief introduction to data analysis including complex analysis of 10+ parameters simultaneously. These high-dimensional datasets require novel methods for analysis due to the volume of data collected, which are also introduced in this chapter.

**Key words** Flow cytometry, Phenotyping, Informatics

---

### 1 Introduction

The recent advances in cancer immunotherapy for melanoma treatment are attributed to the development of checkpoint blockade including anti-PD1 and anti-CTLA4 antibodies [1, 2]. Identification of the role of *BRAF* mutations in melanoma has also led to the development of a new class of targeted therapies inhibiting *BRAF* and MEK [3]. *BRAF* targeted therapy also has some immunological consequences, evidence in a study demonstrating increased T-cell infiltration into tumors after *BRAF* therapy, suggesting combination type approaches [4]. Many trials are underway exploring the effect of combination on the immune microenvironment [5]. These discoveries can be attributed to an improved understanding of the tumor microenvironment. Therefore,

physician scientists require a way to understand the immune system using many parameters simultaneously. One such method is the use of flow cytometry for analyzing cell samples from the peripheral blood of patients undergoing checkpoint blockade. This focused informatics paper will provide an overview of the principles of multiparametric flow cytometry needed for translational researchers. In order to understand the analysis of the flow cytometry and other high-dimensional data (mass cytometry), it is necessary to understand the basic experimental method. Therefore, we begin our discussion of flow cytometric methods followed by how to utilize available methods to analyze the data in a way that is meaningful in translational research for therapeutic development. We describe available tools to analyze these high dimensional datasets and specific instances where they may be useful. Finally, we describe the extension of these techniques to other platforms such as imaging mass cytometry [6, 7]. Applications of this technology to multidimensional immunofluorescence utilizing FFPE tissues can analyze up to 9 and 50 parameters depending on the throughput of the instrumentation [8, 9]. There are multiple other platforms that are available, but we limit ourselves to the ones that we are most familiar. The application of flow cytometry to cancer research has led to better understanding of the tumor microenvironment and identification of treatment targets [10].

---

## 2 Materials

### 2.1 Flow Cytometry Analysis

A flow cytometry apparatus is a microfluidic laser-based system for high throughput analysis of cell populations. Samples are suspended in fluid and passed through a chamber that is exposed to light of different wavelengths [11, 12]. Specimens are analyzed based on light scattering from structural properties of cells and fluorescent emission from fluorescent probes (fluorochromes) bound to cells or cell components [11]. Cells need to pass through the laser source in single-file for accurate single-cell analysis. Emitted light is deflected off particles in the fluid stream based on their structure. In a general model, forward scatter (FSC) is collected along the axis of the light source and represents particle size. Side scatter (SSC) is collected from reflected light collected at 90° to the light source and represents particle granularity or complexity [11, 13]. In the early days of flow cytometry, FSC and SSC measurements were used to differentiate cell types. With the development of new light sources capable of emitting light and filters capable of collecting the emission over narrow bandwidths (range of wavelengths), the scattered signal from the biological sample can be directed to a wavelength-specific detector [13]. Three types of filters are utilized; longpass filters, which only permit wavelengths above a specified level, shortpass filters, which only permit

wavelengths below a specified level, and bandpass filters, which permit wavelengths in a defined range. Using a series of filters and mirrors in geometric arrangements (e.g., octagon or triangle), fluorescence from one light source can be directed into as many as eight unique detectors for different fluorescent molecules on certain types of commercial flow cytometers.

The number of fluorescent particles in a cell (autofluorescence) is fortunately limited, therefore fluorochromes are used to stain samples and identify features of interest, including proteins (receptors, cytosolic proteins, etc.) and nucleic acids. Use of these markers can also give the researcher an idea of cellular activity or viability [11]. Fluorochromes are limited by overlapping emission spectra, limiting the number of fluorochromes that can be used simultaneously. Tandem dyes can increase the number of fluorochromes used by combining two different dyes such that the emission spectrum of one excites the other (fluorescence resonance energy transfer, FRET) but the user should be careful in selecting tandem dyes to minimize spillover into different “channels” of the flow cytometer and attention should be paid to the stability of the tandem dye [14, 15].

## 2.2 Controls in Flow Cytometry

### 2.2.1 Compensation in Flow Cytometry

Spectral overlap, defined as signal spillover from one fluorochrome (fluor) measured by a detector for another fluorochrome, is a limitation of traditional flow cytometry techniques. Spectral overlap requires the mathematical process of compensation prior to data analysis to eliminate signal from fluors not associated with the target of interest for a set voltage of the photomultiplier tube [16, 17]. If there are “ $n$ ” fluorescent parameters in an experiment, “ $n$ ” compensation controls are needed with an unstained sample for autofluorescence. These controls must emit light equal to or greater than the fluorescence emitted by that fluor in the biological sample. As such, there are two possibilities to use as a single-color control, the cellular sample itself or a bead that has the ability for the fluorescent antibody to bind [18, 19]. It is recommended to record a sufficient number of events to calculate a robust compensation matrix. The process of correcting spectral spillover is relatively simple with a limited number of samples (e.g., 1–4 parameters) but becomes more complex with multidimensional data generated in high-parameter flow cytometry experiments. In the pioneer days of flow cytometry, compensation was typically performed manually using hardware compensation on older generation analog flow cytometers. With these types of cytometers it is recommended to collect the data without compensation and perform the compensation on external flow cytometry program applications (e.g., FCS Express [De Novo Software, Glendale, CA, USA] and FlowJo [BD Biosciences, Ashland, OR, USA], flowCore in R [20]). Otherwise, if the manual compensation is not accurate, the data can be saved with errors that cannot be corrected. In the

modern era, most commercial flow cytometers have digital electronics with software that is capable of calculating the compensation matrix post-acquisition, and the raw data file contains the matrix recognizable by other third-party compatible analysis software [18]. Mathematically, the vector representing the amount of each fluorochrome associated with the cells is a matrix product of the vector representing the signal from each of the “channels” on the instrument multiplied by the inverse of the compensation matrix (sometimes called the spectral matrix) [16]. With fewer parameters, data can be obtained with minimal compensation by utilizing dyes with limited fluorescent overlap. High-parameter experiments require compensation matrices where the number of elements contained within the matrix is the square of the number of parameters. To facilitate compensation, either cells or polymer beads may be utilized. Compensation beads are nonbiologic materials that are either fluorochrome-impregnated or are anti-immunoglobulin-coated capture beads [21]. Compensation beads provide several advantages. Firstly, beads stain more intensely and uniformly than cells, facilitating compensation of even rare events. They can be used with the same antibodies to be used in an experiment, to ensure compensation of the same fluorochromes. Lastly, using compensation beads preserves cells for the experiment rather than controls. Premade fluorochrome-coated beads should be used with caution as the fluorescent signal from such a bead may not match the signal of a tested fluor due to variations among manufacturers [16]. Anti-immunoglobulin capture beads function by binding to antibodies. As long as the beads bind to the fluorochrome-conjugated antibody of interest and has signal equal to or brighter than the intensity of the expression on the biological specimen, even small populations of cells in a test sample can be accurately compensated. Furthermore, beads are more uniform than cells and allow for precise calculation of signal spill-over. Despite the accuracy of capture beads, there are several shortcomings. Fluorescent-conjugated antibodies may be brighter on cells than on the beads and cannot be used to compensate in these situations [16]. Capture beads cannot be used with non-antibody reagents, such as live-dead discrimination with nucleic acid binding dyes, which require amine-reactive beads for compensation. They also cannot be used if the beads do not bind to the fluorochrome-conjugated antibody that requires compensation [16]. One other limitation to anti-immunoglobulin capture beads is that they must be matched to the animal source of the antibody.

### 2.2.2 Live–Dead Controls

Cell viability can be determined by morphological changes or by changes in membrane permeability [22]. The benefit of assessing cell viability using flow cytometry is the simultaneous assessment of viability parameters on the basis of dye exclusion or retention [21]. Several classes of dyes can be used but generally fall into

three classes (DNA binding, amine binding, or vital dyes) [23]. DNA-binding dyes such as 4',6-diamidino-2-phenylindole (DAPI), propidium iodide (PI), or 7-aminoactinomycin D (7-AAD) can be used as dead cell discriminators [23]. Live cells are normally impermeable to many in this family of dyes but compromised membranes of nonviable cells are permeable. These dyes are easy to use but cannot be used when intracellular markers are measured with fixation and permeabilization methods. Another class of dyes is synthetic amine dyes, which bind to amine groups of proteins after penetrating the membrane of dead cells and are available in a wide variety of excitation and emission profiles. However, unlike DNA binding dyes, these can be used in fixation and permeabilization experiments [24]. Vital dyes are used less often in phenotypic flow cytometry experiments but this may change as more choices for excitation/emission in vital dyes become available. A key point is that vital dyes require an intact cell and functional esterases and this dye should be utilized prior to any fixation/permeabilization steps due to the need for an esterase to cleave the dye into its fluorescence form [23].

### 2.2.3 Doublet Discrimination

Doublet/aggregate events can occur in data acquisition, due to the aggregation of cells. Aggregates will yield erroneous results as the user would have assigned a higher fluorescence intensity to single-cell measurements. Furthermore, larger aggregates can clog the instrument and cause problems in data acquisition. The first step in addressing sample aggregation is in sample preparation to minimize adhesion between cells in the sample. Several sample protocols and considerations are presented in Table 1. Doublet discrimination is necessary for the accuracy of single cell analysis. There are a couple of options for gating out aggregates in the downstream analysis. It is performed by plotting FSC-H vs. FSC-W and plotting SSC-H and SSC-W. In this case, the doublets can easily be distinguished by FSC-W and SSC-W. When using flow cytometers that do not measure width (FSC-W) accurately, an area measurement (FSC-A) can be used in conjunction with FSC-H to identify aggregates. When an aggregate passes through the cytometer, the FSC-H measurement will be the same, however the FSC-A measurement will be increased compared to single cells, facilitating the identification of doublets/aggregates [25, 26]. Therefore, in practice one should analyze height, area, and width to determine which measurements work best for your samples.

### 2.2.4 Fluorescence-Minus-One Control

Fluorescence-minus-one controls, commonly known as FMO controls, are used to properly analyze flow cytometry data and interpret the results. A FMO control sample is stained with all fluorescent reagents *except* the one of interest to determine the threshold that best discriminates the positive and negative signals for the fluor of

**Table 1**  
**Selected resources**

<i>Tutorials<sup>a</sup></i>
<ul style="list-style-type: none"> <li>– BD Biosciences (<a href="http://www.bdbiosciences.com/us/support/s/itf_launch">http://www.bdbiosciences.com/us/support/s/itf_launch</a>)</li> <li>– Thermo Fisher (<a href="http://www.thermofisher.com/us/en/home/support/tutorials.html">http://www.thermofisher.com/us/en/home/support/tutorials.html</a>)</li> <li>– Beckman Coulter (<a href="http://beckman.com/resources">http://beckman.com/resources</a>)</li> </ul>
<i>Sample preparation<sup>a</sup></i>
<ul style="list-style-type: none"> <li>– University of Wisconsin Carbone Cancer Center (UWCCC) Flow Lab (<a href="https://cancer.wisc.edu/research/wp-content/uploads/2017/03/Flow_TechNotes_Minimizing-Aggregates-in-Samples_20170918.pdf">https://cancer.wisc.edu/research/wp-content/uploads/2017/03/Flow_TechNotes_Minimizing-Aggregates-in-Samples_20170918.pdf</a>)</li> <li>– University of Iowa Carver College of Medicine (<a href="https://medicine.uiowa.edu/flowcytometry/protocolssample-prep/sample-preparation-analysis/">https://medicine.uiowa.edu/flowcytometry/protocolssample-prep/sample-preparation-analysis/</a>)</li> <li>– The Rockefeller University (<a href="https://www.rockefeller.edu/fcrc/sampleprep/">https://www.rockefeller.edu/fcrc/sampleprep/</a>)</li> </ul>
<i>Blogs and other online resources<sup>a</sup></i>
<ul style="list-style-type: none"> <li>– Excyte Expert Cytometry (<a href="http://expertcytometry.com">http://expertcytometry.com</a>)</li> <li>– Dr. Mario Roederer's Home Page (<a href="http://drmr.com">http://drmr.com</a>)</li> <li>– Purdue Cytometry E-mail Archive (<a href="http://lists.purue.edu/pipermail/cytometry">http://lists.purue.edu/pipermail/cytometry</a>)</li> <li>– University of Chicago UCFlow Blog (<a href="http://ucflow.blogspot.com">http://ucflow.blogspot.com</a>)</li> <li>– Boston University Introduction to Flow Cytometry: A Learning Guide (<a href="http://www.bu.edu/flow-cytometry/files/2010/10/BD-Flow-Cytom-Learning-Guide.pdf">http://www.bu.edu/flow-cytometry/files/2010/10/BD-Flow-Cytom-Learning-Guide.pdf</a>)</li> </ul>

<sup>a</sup>The authors report no affiliation or commercial interests in any of these resources or entities. These are recommended for scientific content only

interest [16]. When comparing the test sample to the FMO control, only one variable has changed and any differences in signal are attributed to the biology and not the background contributed from the other fluors. Typically, for experiments that are run longitudinally, all FMO controls are run at the beginning of the study, and the important FMOs are run for subsequent individual experiments to distinguish between positive and negative populations when there is no clear definition [27].

## 2.2.5 Isotype Controls

The FMO control is the experimental control to determine positive versus negative based on the properties of the fluors in the sample. Another type of negative control that is controversial in the literature is an isotype control, which was originally meant to identify nonspecific background cell staining. The isotype control is based on the theory that an antibody with the same isotype will have the same fluorescent signal. Thus, if a sample is stained with an isotype of a tested antibody, nonspecific binding from the antibody can be identified and a gate defined to exclude the background signal [28, 29]. Isotype controls make several assumptions that cannot be confirmed. One assumption is that the control isotype binds to off-targets with the same affinity as the experimental antibody. Furthermore, the assumption must be made that there are no primary targets for the isotype control. Lastly, the fluorochrome-

to-protein ( $F/P$ ) ratio, which is the amount of fluorochrome bound to the antibody or isotype, is assumed to be the same between the isotype control and experimental antibody. In summary, in order for an isotype control to be a valid proxy for an experimental antibody, it must only bind to background proteins and not a target of interest, and must have the same fluorescent properties as the antibody. Since these properties cannot be known or verified, the isotype control is not an appropriate method of standardization of positive vs. negative gating for a particular antibody but may help in explaining patient to patient variability and inherent “cell stickiness” for flow cytometry experiments [29, 30]. The confusion regarding this topic is that in the not so distant past, isotype controls were considered the gold standard negative control. If the isotype has robust properties it should be equivalent to the fluorescence-minus-one experiments. In many situations, isotype controls may be abandoned or may be experimentally proven to be the same as FMO controls, but we urge the reader to discuss this with an expert in flow cytometry. One area where isotype controls are still useful is in measuring phenotypes from patient derived peripheral blood mononuclear cell samples where the above conditions are met.

### 2.2.6 Nonspecific Binding

If isotype controls cannot accurately account for nonspecific binding in many cases, another solution is needed. Nonspecific binding is most commonly caused by an excess of antibody relative to the intended target and the first step is appropriate sample preparation by optimizing antibody concentrations with a titration assay [31]. Another reason for nonspecific binding is due to Fc receptors on the surface of most immune cells, which may bind to antibodies with varying affinities. Blocking reagents can be used to block these receptors and minimize this effect (Table 1) with one such reagent being nonspecific IgG [32]. From the above discussion, it is apparent that this is still a field with polarized views but the most important thing to realize is that there needs to be a way to account for nonspecific binding. There are also practical considerations of the experiment to consider and for this reason, we highly encourage that the novice in flow cytometry consult an expert prior to embarking on this journey.

## 2.3 Running the Experiment

- Familiarize yourself with many flow cytometry protocols and discuss the project with an experienced flow cytometrist if possible (Table 1).
- Sample preparation for phenotyping (single cell suspension [33]).
  - Cells should be harvested, washed and filtered to obtain single cell suspension with target concentration of  $5 \times 10^5$  to  $5 \times 10^6$  cells/ml in ice cold buffer (typically a combination of PBS, and 1–5% fetal bovine serum). Cell fixative buffer

contains a fixative such as 0.5–4% paraformaldehyde in a PBS-based buffer.

Obtaining a single cell suspension is critical (Table 1).

Live cell assays should not have fixative.

Paraformaldehyde should be kept in the refrigerator in a container wrapped in foil.

- Add cell suspension to test tube and add blocking solution as needed to minimize nonspecific binding and background fluorescence.
- Centrifuge sample at 4 °C and discard supernatant. Temperature is a variable that will need to be optimized.
- Add primary labeled antibody ( $\pm$  isotype to a different tube) and incubate for at least 30 min at 4 °C in the dark.

In-house experiments are typically conducted in a total of 100  $\mu$ l of cell suspension buffer.

- Wash the cells by centrifugation with the buffer utilized for the staining and resuspend in ice cold flow buffer to be utilized on the cytometer.
- Store on ice (or refrigerator) in the dark until time for analysis. Cells should be filtered prior to analysis.
- Certain types of analysis require fresh samples without fixation, but this procedure is for the vast majority of phenotyping experiments where fixation of cells is appropriate.
- For intracellular markers, a single cell suspension is created as above. The key difference is making the cell membrane permeable to staining.

Stain cells with viability dye.

Then fix cells with fixation buffer and incubate at experimentally determined temperature. Cells should be fixed away from light.

Permeabilization buffer is then added and the sample is centrifuged.

The type of permeabilization buffer and the procedures associated with staining should be optimized. We recommend conducting a literature search prior to conducting these types of experiments.

Cells should be filtered prior to analysis.

- Check the controls.
  - Compensation controls.
  - FMO controls.
  - Biological controls.
  - QC controls (beads for the experiment).

This paper assumes that you have staff running the flow cytometer capable of properly performing daily quality control on the instrument.

- Live–Dead control.
- For the first run of the experiment, all FMO controls will be prepared and acquired.
- On subsequent runs, the FMOs that do not have clear positive/negative cutoffs along with compensation controls and samples will be measured.
  - Use standardized reference beads to control for flow cytometer settings across experiments.
- Include viability dye (exclude dead cells), doublet discrimination (exclude aggregates). Target population frequencies and data set coefficient of variation (CVs) will be the determining factors in how many total events to acquire.

## **2.4 Principles of Multidimensional Data Analysis**

### **2.4.1 The “Curse of Dimensionality”**

The information generated from a flow cytometry experiment with few variables can be easily plotted on histograms and dot plots and manually interpreted. Several commercial programs exist (i.e., FCS Express, FlowJo among others), and freely available software for those with more limited budgets are available in R but requires knowledge of the R programming environment. As discussed above, cells can be identified using simple measurements like size (e.g., FSC), or more commonly, using markers identified by fluorochromes. Flow cytometry data is most commonly displayed on one- or two-parameter plots. As the complexity of flow cytometry increases, traditional visualization methods quickly become inadequate to display high-dimensional data. Dot plots are unable to convey meaningful data if individual points cannot be visualized and are impractical as the number of  $2 \times 2$  plots required to visualize multiparametric data are cumbersome and it becomes impossible to transmit meaningful or representative data. As one increases the number of parameters in a flow experiment, the data becomes increasingly difficult to visualize.

### **2.4.2 Manual Data Display Utilizing Available Software Packages**

With increasing number of parameters and data measured, new solutions are required to adequately convey meaningful data. The logarithmic scale is typically used to include a large range of values, however negative and small near-zero values are not effectively visualized. Displaying flow cytometry data presents a challenge because oftentimes, the fluorescent signal of a sample can be nearly zero or negative after compensation, precluding visualization on a logarithmic scale. Rare cell populations can thus be missed as signals “pile up” around zero [34]. Several groups have proposed solutions to overcome the limitation of the logarithmic scale by transforming the data linearly at low values and logarithmically at

higher values [35]. Others utilize a biexponential scale that more accurately displays populations with low fluorescence intensity and visually centers data around the true median of a sample [36]. Dot plots are limited by resolution with increasing data points, limiting the amount of data that can be visualized. Other recommendations include quantile contour plots to accurately represent data [34]. The key point to remember is that the transformation of the data makes it easier to visualize the population of interest but it does not alter the collected data. However, one must be vigilant when transforming data and detecting rare populations that this rare population is not indeed an artifact.

## **2.5 Informatics Analysis of High-Dimensional Flow Cytometry Data**

As we continue to improve the capabilities of flow cytometry techniques, challenges in analyzing these datasets have arisen. Conventionally, when a limited number of parameters were measured (e.g., 2–3), data could be manually analyzed with sequential gating strategies [37]. For example, in the case of Tregs, a user may gate on FSC-A vs. SSC-A to obtain a lymphocyte gate, then gate on CD3, to have their final gate be a combination of CD4 and FOXP3. However, manual analysis of flow cytometry data can introduce variability and bias due to manual gating of positive and negative populations [38–40]. Most importantly, manual analysis of multi-dimensional flow data is time-consuming. Each flow plot can only compare two parameters, which quadratically increases with additional markers [41]. Modern flow cytometers are able to measure 30+ parameters simultaneously and it is not practical to manually compare each permutation of flow plots [42, 43].

The introduction of computational flow cytometry techniques allows for rapid analysis of large, complex datasets and to efficiently present the data in a meaningful way to identify distinct cell populations within a sample. In an experiment with multiple markers, one can consider each marker to be a dimension. We provide several examples of techniques that can provide automated gating techniques utilizing dimensionality reduction and clustering techniques to rapidly sort populations of cells.

One technique to visualize this data is dimensionality reduction where combinations of markers are utilized to visualize the data, representing high-dimensional data in a lower-dimensional (e.g., 2–3 dimensions) projection, while also accounting for the nonparametric nature of cellular data. This allows for identification of rare populations and also simplifies the dataset for visualization. t-SNE (t-distributed stochastic neighbor embedding) is a nonparametric technique for dimensionality reduction [44]. This technique maps the high-dimensional data to a lower dimension suitable for visualization [43]. Given that t-SNE represents combinations of the markers, large trends in the data can be visualized on a 2D map [44–46]. Because each cell is plotted as an individual data point, t-SNE can be time consuming for very large datasets, but for most

flow cytometry applications the run times are reasonable. As flow cytometry machines and memory storage capacities have increased, variations of t-SNE may be needed in the future to facilitate practical analysis. Approximated-tSNE (A-tSNE) is an adaptation of t-SNE that minimizes computation time by approximating distances between data points [47].

Clustering techniques tend to reduce the dimensionality of the data to create a visual representation in a reduced number of dimensions. Traditional clustering methods (e.g., k-means clustering) belong to a class of data mining techniques called unsupervised learning techniques that automatically processes data instead of responding to user feedback [43]. The dataset is automatically sorted by assigning cells with similar features to similar clusters. SPADE (spanning-tree progression analysis of density-normalized events) [48] and FlowSOM (flow cytometry data analysis using self-organizing maps) [49] are examples of algorithms that automatically cluster datasets then visualizes data on a minimum spanning tree. SPADE first applies a density-dependent downsampling step to simplify the data. This is needed as the computational requirements to process the entire dataset (even 100,000 events) would be prohibitively long. SPADE then groups cells into cell-type clusters. Abundant cells and rare population subsets are both described as a single node and therefore it is easier to visualize less abundant cells. The data is visualized using a minimal spanning tree that highlights the differences between phenotypes (nodes of different phenotypes would be further apart on the tree). The downside of phenotypic trees is that the nodes at the far ends of the branches may in actuality be similar, but artificially separated due to limitations of this visualization method. In addition, there is a lot of manual curation to analyze the phenotypes described on the minimal spanning tree. The information regarding the numbers of cells in each phenotype (node) is then superimposed on the tree by increasing the size of the node. Work in our laboratory has focused on reducing the manual labor needed for analysis [50]. We also have in-house software that allows for new samples to be superimposed on existing phenotypic trees. FlowSOM clusters cells using a self-organized map, which is an algorithm that automatically sorts cells based on similarities in marker expression. Data is displayed on a phenotypic tree similar to SPADE, with the goal of connecting similar nodes and including some information regarding the different markers that define the phenotype for that node [49]. The aforementioned limitations also apply to FlowSOM visualizations as one would visualize the phenotypic tree for each sample separately.

ACCENSE (Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding) aims to resolve the loss of single-cell resolution, which is a problem inherent to clustering approaches [51]. ACCENSE performs dimensionality reduction

(t-SNE) to maintain single-cell resolution then clusters based on the t-SNE data. The number of clusters is defined by the user. ACCENSE is limited by the input size for its clustering algorithm, thus limiting its scalability to larger data sets. Another limitation occurs at the clustering step, which, after downsampling, does not cluster all of the cells in the sample [52]. This limits its application for single-cell resolution because the downsampling step may exclude rare populations of cells. With this being said, it is still a very powerful algorithm.

Cytobank is a commercial, web-based platform for flow cytometry analysis that utilizes a t-SNE-based algorithm (viSNE [53]) for downsampling and flowSOM or SPADE for clustering. While Cytobank does not introduce a novel workflow, it incorporates all analyses of a flow cytometry experiment into a single platform with graphical user interface, including data analysis [52, 54]. Flow cytometry data can be obtained from public databases but once data is clustered, no new data points can be superimposed and samples must be visualized individually.

Cytosplore is a program that was designed specifically for analyzing CyTOF data to identify cell populations with attention to details such as lineage and subset definitions [52, 55]. A modified SPADE algorithm is used to cluster data into similar groups for the user to select major lineages of interest to be further evaluated. A modified tSNE is then utilized to start exploring the immune cell subsets. Gaussian Mean Shift (GMS) clustering is performed on this output to identify cell subsets and then the cells can be phenotyped based on the clusters. Multiple analysis steps allow for the determination of cell phenotypes via use of the phenotypic tree and a heat map illustrating the density of different parameters. The Cytosplore method claims to increase the efficiency of the phenotyping as illustrated in software comparisons to SPADE and tSNE [52]. The Cytosplore method is listed in this chapter as one should pay attention to the platform for which the software was developed as one may see spurious results if software packages are applied to different experimental platforms without some minor alterations.

ImmPort is a free, publicly accessible database of immunology datasets. It was initially created for internal data sharing by the Division of Allergy, Immunology and Transplantation in the National Institute of Allergy and Infectious Diseases (NIAID) [56]. It has since been opened to the public, retaining data storage and sharing functions with the addition of flow cytometry analysis tools (ImmportGalaxy) [57]. The clustering algorithms have some differences in Galaxy, but the principles are generally the same. A nice feature of the Galaxy platform is that there are public data sets for which the analyses have already been published prior to analyzing your own data.

Cell phenotyping in flow cytometry is an evolving field that has outgrown manual interpretation of results. Marker based

phenotyping of cells in suspension is critical to study immune cell phenotypes in this era of immune-based therapies. There is increasing demand for efficient and accurate analysis techniques to extract meaningful information from this data. We provided an introduction to modern flow cytometry (measurement and analysis) to provide the reader an overview of the preparation and analysis of multiparameter phenotyping datasets. In terms of experimentation, it is advised to consult with an expert in flow cytometry. Two general themes that have emerged are the concepts of dimension reduction and downsampling. In general, high-dimensional datasets require simplification prior to analysis, which is usually accomplished by dimensional reduction so that the data can be visually interpreted. Another approach is downsampling, only analyzing a representative subset of sample containing tens of thousands of cells. A subsequent clustering step sorts the sample by feature similarity to identify overall populations in the sample. There are methods that combine these approaches for efficient identification of cell phenotypes. The field of single-cell analysis using computer aided clustering techniques is expanding due to the complexity of flow cytometry data and continues to lead to new discoveries and applications in medicine. Multiple software packages have been created to answer specific questions that may be posed by immunological researchers. The methods described above are currently in the process of being altered for the analysis of cells and their geometric relationships within paraffin embedded sections by multiple groups including our own. Attention should always be paid to the experimental platform for which the software was developed.

## References

1. O'Day SJ, Hamid O, Urba WJ (2007) Targeting cytotoxic T-lymphocyte antigen-4 (CTLA-4): a novel strategy for the treatment of melanoma and other malignancies. *Cancer* 110 (12):2614–2627
2. Tumeh PC, Harview CL, Yearley JH et al (2014) PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* 515(7528):568–571
3. Davies H, Bignell GR, Cox C et al (2002) Mutations of the BRAF gene in human cancer. *Nature* 417(6892):949–954
4. Liu C, Peng W, Xu C et al (2013) BRAF inhibition increases tumor infiltration by T cells and enhances the antitumor activity of adoptive immunotherapy in mice. *Clin Cancer Res* 19 (2):393–403
5. Pelster MS, Amaria RN (2019) Combined targeted therapy and immunotherapy in melanoma: a review of the impact on the tumor microenvironment and outcomes of early clinical trials. *Ther Adv Med Oncol* 11:1758835919830826
6. Angelo M, Bendall SC, Finck R et al (2014) Multiplexed ion beam imaging of human breast tumors. *Nat Med* 20(4):436–442
7. Giesen C, Wang HA, Schapiro D et al (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 11(4):417–422
8. Coltharp C, Zheng Y, Schaefer R et al (2018) Advances in multiplex fluorescence immuno-histochemistry: 9 color imaging; whole slide multispectral. *J Immunother Cancer* 6(Suppl 1):P433
9. Goltsev Y, Samusik N, Kennedy-Darling J et al (2018) Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* 174(4):968–981.e915

10. Gorris MAJ, Halilovic A, Rabold K et al (2018) Eight-color multiplex immunohistochemistry for simultaneous detection of multiple immune checkpoint molecules within the tumor microenvironment. *J Immunol* 200(1):347–354
11. Adan A, Alizada G, Kiraz Y, Baran Y, Nalbant A (2017) Flow cytometry: basic principles and applications. *Crit Rev Biotechnol* 37(2):163–176
12. Pockley AG, Foulds GA, Oughton JA, Kerkvliet NI, Multhoff G (2015) Immune cell phenotyping using flow cytometry. *Curr Protoc Toxicol* 66:18.18.11–18.18.34
13. Virgo PF, Gibbs GJ (2012) Flow cytometry in clinical pathology. *Ann Clin Biochem* 49(Pt 1):17–28
14. Leavesley SJ, Britain AL, Cichon LK, Nikolaev VO, Rich TC (2013) Assessing FRET using spectral techniques. *Cytometry A* 83(10):898–912
15. Bushnell T (2017) Why understanding fluorochromes is important in flow cytometry. <https://expertcytometry.com/why-understanding-fluorochromes-is-important-in-flow-cytometry/>. Accessed 26 Aug 2019
16. Roederer M (2002) Compensation in flow cytometry. *Curr Protoc Cytom. Chapter 1: Unit 1.14*
17. De Rosa SC, Roederer M (2001) Eleven-color flow cytometry. A powerful tool for elucidation of the complex immune system. *Clin Lab Med* 21(4):697–712, vii
18. Roederer M (2001) Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry* 45(3):194–205
19. Baumgarth N, Roederer M (2000) A practical approach to multicolor flow cytometry for immunophenotyping. *J Immunol Methods* 243(1–2):77–97
20. Hahne F, LeMeur N, Brinkman RR et al (2009) flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10:106
21. Cossarizza A, Chang HD, Radbruch A et al (2017) Guidelines for the use of flow cytometry and cell sorting in immunological studies. *Eur J Immunol* 47(10):1584–1797
22. Johnson S, Nguyen V, Coder D (2013) Assessment of cell viability. *Curr Protoc Cytom. Chapter 9:Unit 9.2*
23. McCracken J (2015) 3 Reagents for identifying live, dead, and apoptotic cells by flow cytometry. <https://expertcytometry.com/3-reagents-for-identifying-live-dead-and-apoptotic-cells-by-flow-cytometry/>. Accessed 26 Aug 2019
24. Perfetto SP, Chattopadhyay PK, Lamoreaux L et al (2010) Amine-reactive dyes for dead cell discrimination in fixed samples. *Curr Protoc Cytom. Chapter 9:Unit 9.34*
25. Bauer KD (1993) Quality control issues in DNA content flow cytometry. *Ann N Y Acad Sci* 677:59–77
26. Wersto RP, Chrest FJ, Leary JF, Morris C, Stetler-Stevenson MA, Gabrielson E (2001) Doublet discrimination in DNA cell-cycle analysis. *Cytometry* 46(5):296–306
27. Wang L, Hoffman RA (2017) Standardization, calibration, and control in flow cytometry. *Curr Protoc Cytom* 79:1.3.1–1.3.27
28. Maecker HT, Trotter J (2006) Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry A* 69(9):1037–1042
29. Bushnell T (2017) Strengths and weaknesses of isotype controls in flow cytometry. <https://expertcytometry.com/strengths-and-weaknesses-of-isotype-controls-in-flow-cytometry/>. Accessed 26 Aug 2019
30. McCracken J (2015) When to use (and not use) flow cytometry isotype controls. <https://expertcytometry.com/when-to-use-and-not-use-flow-cytometry-isotype-controls/>. Accessed 26 Aug 2019
31. Hulspas R, O’Gorman MR, Wood BL, Gratama JW, Sutherland DR (2009) Considerations for the control of background fluorescence in clinical flow cytometry. *Cytometry B Clin Cytom* 76(6):355–364
32. Andersen MN, Al-Karradi SN, Kragstrup TW, Hokland M (2016) Elimination of erroneous results in flow cytometry caused by antibody binding to Fc receptors on human monocytes and macrophages. *Cytometry A* 89(11):1001–1009
33. Reichard A, Asosingh K (2019) Best practices for preparing a single cell suspension from solid tissues for flow cytometry. *Cytometry A* 95(2):219–226
34. Herzenberg LA, Tung J, Moore WA, Herzenberg LA, Parks DR (2006) Interpreting flow cytometry data: a guide for the perplexed. *Nat Immunol* 7(7):681–685
35. Novo D, Wood J (2008) Flow cytometry histograms: transformations, resolution, and display. *Cytometry A* 73(8):685–692
36. Parks DR, Roederer M, Moore WA (2006) A new “Logicle” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry A* 69(6):541–551
37. Lugli E, Roederer M, Cossarizza A (2010) Data analysis in flow cytometry: the future just started. *Cytometry A* 77(7):705–713

38. Nomura L, Maino VC, Maecker HT (2008) Standardization and optimization of multiparameter intracellular cytokine staining. *Cytometry A* 73(11):984–991
39. Pachon G, Caragol I, Petriz J (2012) Subjectivity and flow cytometric variability. *Nat Rev Immunol* 12(5):396; author reply 396
40. Gouttefangeas C, Chan C, Attig S et al (2015) Data analysis as a source of variability of the HLA-peptide multimer assay: from manual gating to automated recognition of cell clusters. *Cancer Immunol Immunother* 64 (5):585–598
41. Palit S, Theis FJ, Zielinski CE (2018) Meeting the challenges of high-dimensional data analysis in immunology. *bioRxiv*:473215
42. Perfetto SP, Chattopadhyay PK, Roederer M (2004) Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol* 4(8):648–655
43. Saeys Y, Van Gassen S, Lambrecht BN (2016) Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol* 16(7):449–462
44. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605
45. Wattenberg M, Viégas F, Johnson I (2016) How to use t-SNE effectively. Distill
46. Team AVC (2017) Comprehensive guide on t-SNE algorithm with implementation in R & Python. <https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/>. Accessed 11 Sept 2019
47. Pezzotti N, Lelieveldt BPF, Van Der Maaten L, Hollt T, Eisemann E, Vilanova A (2017) Approximated and user steerable tSNE for progressive visual analytics. *IEEE Trans Vis Comput Graph* 23(7):1739–1752
48. Qiu P, Simonds EF, Bendall SC et al (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 29(10):886–891
49. Van Gassen S, Callebaut B, Van Helden MJ et al (2015) FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* 87(7):636–645
50. Garg SK, Ott MJ, Mostofa AGM et al (2019) Multi-dimensional flow cytometry analyses reveal a dichotomous role for nitric oxide in melanoma patients receiving immunotherapy. *Front Immunol* 11:164
51. Shekhar K, Brodin P, Davis MM, Chakraborty AK (2014) Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci U S A* 111 (1):202–207
52. Höllt T, Pezzotti N, van Unen V et al (2016) Cytopllore: interactive immune cell phenotyping for large single-cell datasets. *Comput Graph Forum* 35(3):171–180
53. el Amir AD, Davis KL, Tadmor MD et al (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 31(6):545–552
54. Chen TJ, Kotecha N (2014) Cytobank: providing an analytics platform for community cytometry data analysis and collaboration. *Curr Top Microbiol Immunol* 377:127–157
55. van Unen V, Hollt T, Pezzotti N et al (2017) Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat Commun* 8(1):1740
56. Bhattacharya S, Andorf S, Gomes L et al (2014) ImmPort: disseminating data to the public for the future of immunology. *Immunol Res* 58(2–3):234–239
57. Bhattacharya S, Dunn P, Thomas CG et al (2018) ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci data* 5:180015



# Chapter 14

## High-Dimensional Flow Cytometry Analysis of Regulatory Receptors on Human T Cells, NK Cells, and NKT Cells

Ryosuke Nakagawa, Jason Brayer, Nicole Restrepo, James J. Mulé, and Adam W. Mailloux

### Abstract

The field of flow cytometry has witnessed rapid technological advancements in the last few decades. While the founding principles of fluorescent detection on cells (or particles) within a uniform fluid stream remains largely unchanged, the availability more sensitive cytometers with the ability to multiplex more and more fluorescent signals has resulted in very complex high-order assays. This results in the co-use of fluorophores with increased levels of emission overlap and/or spillover spreading than in years past and thus requires careful and well thought out planning for flow cytometry assay development. As an example, we present the development of a large 18-color (20 parameter) flow cytometry assay designed to take an in depth analysis of effector lymphocyte phenotypes, with careful attention to assay controls and panel design.

**Key words** T cell, NK cell, NKT cell, Immune checkpoint, NK receptor, Memory markers, Innate immune cells, Adaptive immune cells

---

### 1 Introduction

Cytotoxic lymphocytes can be found among effector populations in both adaptive and innate arms of immunity. This includes the classically defined cytotoxic T cells ( $T_C$ ) which express the CD8 coreceptor and recognize foreign antigens presented in the context of major histocompatibility complex class I (MHC-I; human leukocyte antigen, HLA, in humans) [1], and natural killer (NK) cells which are regulated by a balance of positive and negative signals through activating and inhibitory NK receptors [2, 3]. In addition to these archetypes, another major population called natural killer T (NKT) cells has been described that shares numerous phenotypic markers with both T cells and NK cells [4], but which is derived from its own unique thymic selection [5] and recognizes lipid antigens presented by the molecule CD1d [6, 7]. Investigating regulatory receptors that govern these populations has led to the development of potent immunotherapy drugs, including those that

block checkpoint receptors [8]. However, the study of checkpoint receptors is often limited to conventional T cell populations and the study of NK cell receptors is often limited to conventional NK cells, despite the expression of both groups of receptors on T, NK, and NKT cells. This 18-color, 20-parameter flow cytometry panel not only allows for the phenotypic characterization of major T cell, NK cell, and NKT cell subsets, but combines several well-known activating NK cell receptors with four well-studied checkpoint receptors to be analyzed therein. It was developed using human peripheral blood mononuclear cells (PBMCs), but could in theory be applied to any human cell source containing effector lymphocyte populations.

Among T cells, cytotoxic activity is largely restricted to the CD8<sup>+</sup> compartment in which T<sub>C</sub> begin as CD45RA<sup>+</sup> naïve T cells and circulate between secondary lymphoid tissues after emigration from the thymus after TCR gene rearrangement. These naïve T<sub>C</sub> can then become activated through engagement with MHC-I if the presented peptide matches the rearranged TCR specificity. This engagement induces the downregulation of central markers such as CCR7, CD62L, and CD27 and the acquisition of effector functionality, including cytokine and cytotoxic granule production. Following activation, T<sub>C</sub> downregulate CD45RA expression and become memory T cells, of which two classically defined subtypes exist, longer-lived central memory (CM) T cells which reacquire the expression of central markers, or shorter lived, but more responsive effector memory (EM) T cells which lack central markers [9]. Because T cell clones undergo a process of negative selection in the thymus during development, highly self-reactive clones are deleted as a part of central tolerance [10]. However, subsequent regulatory mechanisms that protect against autoimmune reactivity are also needed, not the least of which are immune checkpoint receptors, such as lymphocyte-activation gene 3 (LAG-3), T-cell immunoglobulin and mucin-domain containing-3 (TIM-3), cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), and programmed cell death protein 1 (PD-1) [11–16]. Importantly, immune checkpoint receptors are transiently upregulated following activation, and are not limited to conventional T cells, but can also be expressed on NK or NKT populations [17–22].

NK cells, which can be identified in humans as CD56<sup>+</sup>, CD16<sup>high</sup> or CD16<sup>low</sup>, and CD3<sup>−</sup> [23], are large granular lymphocytes that do not express T cell receptor. Instead, NK activation and cytotoxic function is regulated through a net sum of positive and negative signals given by the engagement of activating and inhibitory NK receptors [2, 3]. A common ligand for inhibitory NK receptors are MHC-I molecules, which can be downregulated on malignant or infected cells, whereas common ligands for activating NK receptors are often upregulated on infected or malignant cells [24–26]. Hence NK cells serve a complementary, nonredundant

role with T<sub>C</sub> cells. In addition to NK receptors, NK cells can also express immune checkpoint receptors under certain conditions [17–22], although by comparison the role of immune checkpoint receptors in NK cell function is far less studied than in conventional T cells.

NKT cells recognize lipid antigen presented in the context of the CD1d molecule [6, 7], and in this way reflect the adaptive nature of conventional T cells but also display innate-like properties regarding their regulation through NK cells receptors under specific conditions [27, 28]. Despite overlapping cell surface marker expression, NKT cells are derived from common lymphoid progenitors via thymic selection similar to conventional T cells, but through a TCR restricted to CD1 isoforms [4]. While lipid antigen recognition makes NKT cells important for recognizing organisms such as *mycobacterium* [29], it can also lead NKT cells to play a complex and important role in regulating autoimmune disease in which pathologic or commensal bacteria result in the presentation of lipid antigens [30]. There are two major subsets of NKT cells that have been defined: type I or invariant NKT cells (iNKT) that express a single TCR specific (Va24Ja18) for alpha-galactosylceramide (aGalCer) presented by CD1d, and type II or variant NKT cells (vNKT) with diverse TCR clonality specific for a range of glycolipids or phospholipids presented by CD1 molecules [31].

---

## 2 Materials

1. Flow Buffer: 1% FBS, 1% BSA, 200 nM EDTA in PBS.
2. 12 × 75 mm test tubes.
3. Antibody: anti-human CD19 BV510 (Table 1).
4. Antibody: anti-human CD3 BV786 (Table 1).
5. Antibody: anti-human CD4 BUV805 (Table 1).
6. Antibody: anti-human CD8 AF700 (Table 1).
7. Antibody: anti-human CD16 BUV496 (Table 1).
8. Antibody: anti-human CD56 APC-Fire750 (Table 1).
9. Antibody: anti-human  $\alpha$ 24 $\beta$ 18 PE-Cy7 (Table 1).
10. Antibody: anti-human CD45RA BV6350 (Table 1).
11. Antibody: anti-human CD27 BUV737 (Table 1).
12. Antibody: anti-human CD62L BUV395 (Table 1).
13. Antibody: anti-human CCR7 BB515 (Table 1).
14. Antibody: anti-human NKG2D PerCP-Cy5.5 (Table 1).
15. Antibody: anti-human NKp30 BV421 (Table 1).

**Table 1**  
**Antibodies and reagents**

Specificity	Clone	Fluorochrome	Vendor	Purpose	Optimal titer <sup>a</sup>	SI <sup>b</sup>
CD19	HIB19	BV510	BioLegend	Utility	2.5 µl	22
Live/Dead Aqua			ThermoFisher	Utility		
CD3	UCHT1	BV786	BD Biosciences	Lymph. Lin.	2.5 µl	15
CD4	L3T4	BUV805	BD Biosciences	Lymph. Lin.	2.5 µl	69
CD8	HIT8a	AF700	BioLegend	Lymph. Lin.	0.1 µl	42
CD16	3G8	BUV496	BD Biosciences	Lymph. Lin.	5 µl	26
CD56	5.1H11	APC-Fire750	BioLegend	Lymph. Lin.	0.6 µl	10
va24ja18	6B11	PE-Cy7	BioLegend	Lymph. Lin.	0.3 µl	5
CD45RA	HI100	BV650	BioLegend	Memory Markers	5 µl	53
CD27	L128	BUV737	BD Biosciences	Memory Markers	5 µl	21
CD62L	SK11	BUV395	BD Biosciences	Memory Markers	0.3 µl	71
CCR7	3D12	BB515	BD Biosciences	Memory Markers	5 µl	17
NKG2D	1D11	PerCP-Cy5.5	BD Biosciences	Reg. Receptors	5 µl	2.6
NKp30	P30-15	BV421	BD Biosciences	Reg. Receptors	10 µl	3.5
NKp46	9E2	PE-Dazzle594	BioLegend	Reg. Receptors	1.2 µl	11
PD-1	EH12.2H	BV605	BioLegend	Reg. Receptors	5 µl	3.6
CTLA-4	L3D10	APC	BioLegend	Reg. Receptors	5 µl	3.2
TIM-3	7D3	BV711	BD Biosciences	Reg. Receptors	0.3 µl	4.5
LAG-3	T47-530	PE	BD Biosciences	Reg. Receptors	5 µl	3.3

*AF* Alexa Fluor, *BUV* Brilliant Ultra Violet™, *PE* R-phycerythrin, *BB* Brilliant Blue™, *BV* Brilliant Violet™, *Cy* cyanine, *PerCP-Cy5.5* Peridinin-chlorophyll Cy5.5, *APC* allophycocyanin, *MFI* median fluorescent intensity, *Lymph. Lin.* lymphocyte lineage, *Reg.* regulatory

<sup>a</sup>Volume of commercially available antibody used per  $1.0 \times 10^6$  cells in 100 µl staining volume

<sup>b</sup>SI, stain index as calculated using the listed optimal titer volume using the formula: SI = [(MFI of positive cells) – (MFI of negative cells)]/(2 × SD of negative cells)

16. Antibody: anti-human NKp46 PE-Dazzle594 (Table 1).
17. Antibody: anti-human PD-1 BV605 (Table 1).
18. Antibody: anti-human CTLA4 APC (Table 1).
19. Antibody: anti-human TIM-3 BV711 (Table 1).
20. Antibody: anti-human LAG-3 PE (Table 1).
21. Fixable Live/Dead™ Aqua.
22. FcR Blocking reagent.
23. Ficoll-Paque PLUS.
24. RBC lysis buffer: 0.15 M NH<sub>4</sub>Cl, 10.0 mM KHCO<sub>3</sub>, 0.1 mM EDTA in ddH<sub>2</sub>O pH 7.2–7.4, 0.2-µm filtered.

25. Brilliant Stain Buffer (BD Biosciences).
26. Cell Sample (here human peripheral blood mononuclear cells; PBMC).

---

### 3 Methods

#### 3.1 Staining Protocol

1. PBMC were prepared from fresh blood using density gradient centrifugation with Ficoll-Paque PLUS or RBC lysis buffer as per the manufacturer's instructions.
2. Some aliquots of PBMC were treated with 3000 IU/ml recombinant human IL-2 and/or 10 µg/ml of alpha-galactosylceramide (aGalCer) for 2 days.
3. Single-color compensation and fluorescence-minus-one (FMO) controls are performed using PBMC.
4. Wash cells in flow buffer by adding 2–3× volume, spinning down in the centrifuge at 300–400 × g for 5 min and then carefully pour off supernatant.
5. Resuspend the cells at  $1.0 \times 10^7$  cells/ml in flow buffer.
6. Add 100 µl into each 12 × 75 mm test tube ( $1.0 \times 10^6$  cells).
7. Spin cells in a centrifuge at 1300–1400 RPM for 5 min and then carefully pour off supernatant.
8. Vortex briefly to break up cell pellet, and add 1 ml of PBS.
9. Add 1 µl of Fixable Live/Dead™ Aqua to each appropriate tube, vortex briefly, and incubate at RT for 20 min in the dark.
10. Cells were washed as in **step 3** twice.
11. Brilliant Stain Buffer (optional; see **Note 1**) can be added at this point to the flow buffer to minimize interaction between brilliant violet and brilliant ultraviolet conjugates.
12. Vortex briefly to break up cell pellet, and add the appropriate amount of anti-CD16 BUV496.
13. Incubate for 30 min at 4 °C in the dark.
14. 2 µl FcR blocking reagent was added to each tube for 15 min.
15. Vortex briefly (do not wash), and add the appropriate volume of each remaining antibody to the appropriate tubes in the residual volume (approximately 100 µl) of buffer left in the tube (Table 1).
16. Incubate for 30 min at 4 °C in the dark.
17. Cells were washed as in **step 3** twice.
18. Resuspend in 250 µl of flow buffer.
19. Analyze the cells using a BD LSRII with the optical configurations listed in Table 2, or on a similarly configured cytometer.

**Table 2**  
**BD LSR II SORP configuration**

Laser		Detector		Optics		Channel	
Wavelength (nm)	Power (mW)	Array	Position	Mirror	Filter	Name	OMIP Use
488	50	Trigon	A	685 LP	710/50 BP	Blue A	PerCP-Cy5.5
			B	505 LP	525/50 BP	Blue B	BB515
			C	-	488/10 BP	Blue C	SSC
405	50	Octagon	A	760 LP	785/50 BP	Violet A	BV 786
			B	685 LP	710/50 BP	Violet B	BV 711
			C	635 LP	670/30 BP	Violet C	BV 650
			D	595 LP	610/20 BP	Violet D	BV 605
			E	505 LP	525/50 BP	Violet E	BV 510 and Fixable L/D™ Aqua
			F	-	450/50 BP	Violet F	BV 421
640	40	Trigon	A	755 LP	780/60 BP	Red A	APC-Cy7
			B	685 LP	730/45 BP	Red B	Alexa Fluor 700
			C	-	670/30 BP	Red C	APC
561	50	Octagon	A	755 LP	780/60 BP	YG A	PE-Cy7
			B	635 LP	-	-	-
			C	600 LP	610/20 BP	YG C	PE-Dazzle
			D	-	582/15 BP	YG D	PE
355	60	Octagon	A	770 LP	820/60 BP	UV A	BUV 805
			B	690 LP	740/35 BP	UV B	BUV 737

(continued)

**Table 2**  
(continued)

Laser		Detector		Optics		Channel	
Wavelength (nm)	Power (mW)	Array	Position	Mirror	Filter	Name	OMIP Use
			C	410	515/30	UV C	BUV 496
				LP	BP		
			D	—	379/28	UV D	BUV 395
					BP		

SSC side scatter, *AF* Alexa Fluor, *BUV* Brilliant Ultra Violet™, *PE* R-phycocerythrin, *BB* Brilliant Blue™, *BV* Brilliant Violet™, *Cy* cyanine, *PerCP-Cy5.5* Peridinin-chlorophyll Cy5.5, *APC* allophycocyanin, *LP* long pass, *BP* band pass

### 3.2 Development Strategy

This panel can differentiate a broad spectrum of T, NK, and NKT cell subsets, and analyze the expression of activating NK cell receptors and checkpoint receptors therein. It was optimized using human peripheral blood mononuclear cells (PBMC), but could be performed on any source of human cells containing T, NK, or NKT cells. To build the panel, dyes and markers of interest were first assigned categories based on their usefulness to exclude dead or unwanted cell populations (utility), identify cellular subsets (lymphocyte lineage markers), identify memory phenotypes (memory markers), or measure activating NK or checkpoint receptors (regulatory receptors) (Table 1). Fluorescent conjugates were then selected for each marker considering fluorochrome brightness, expected density of each antigen, commercial availability, and spillover spreading (Fig. 1) that occurs on the BD LSRII detailed in Table 2. Regarding the latter, spillover spreading was largely limited to spillover into the UV C channel (in which the BUV496 fluor is assayed). Therefore, we chose to analyze CD16 using BUV496 as this marker typically stains brightly with good signal to noise ratios.

All antibody-fluorochrome conjugates were titrated individually using PBMC freshly prepared from human blood as a function of stock volume used in residual buffer (approximately 100 µl) on  $1.0 \times 10^6$  cells. Blood was prepared using density gradient centrifugation (Ficoll-Paque PLUS; GE Healthcare Life Sciences). Staining index was then calculated for each dilution using the formula:  $SI = [(MFI \text{ of positive cells}) - (MFI \text{ of negative cells})]/(2 \times SD \text{ of negative cells})$ . Optimal antibody volumes were selected based on calculated SI (Fig. 2). The frequency of type I NKT cells in human PBMC is generally low and highly variable ranging from undetectable to 1%. Therefore, three blood samples were used to titrate the antibody detecting the iNKT va24ja18 TCR (clone 6B11), and the sample displaying the highest frequency was used to calculate the SI. Resting lymphocytes express little or no checkpoint receptors, but can be induced under certain activating conditions. To titrate antibodies against these antigens, we stimulated PBMC using

# Signal

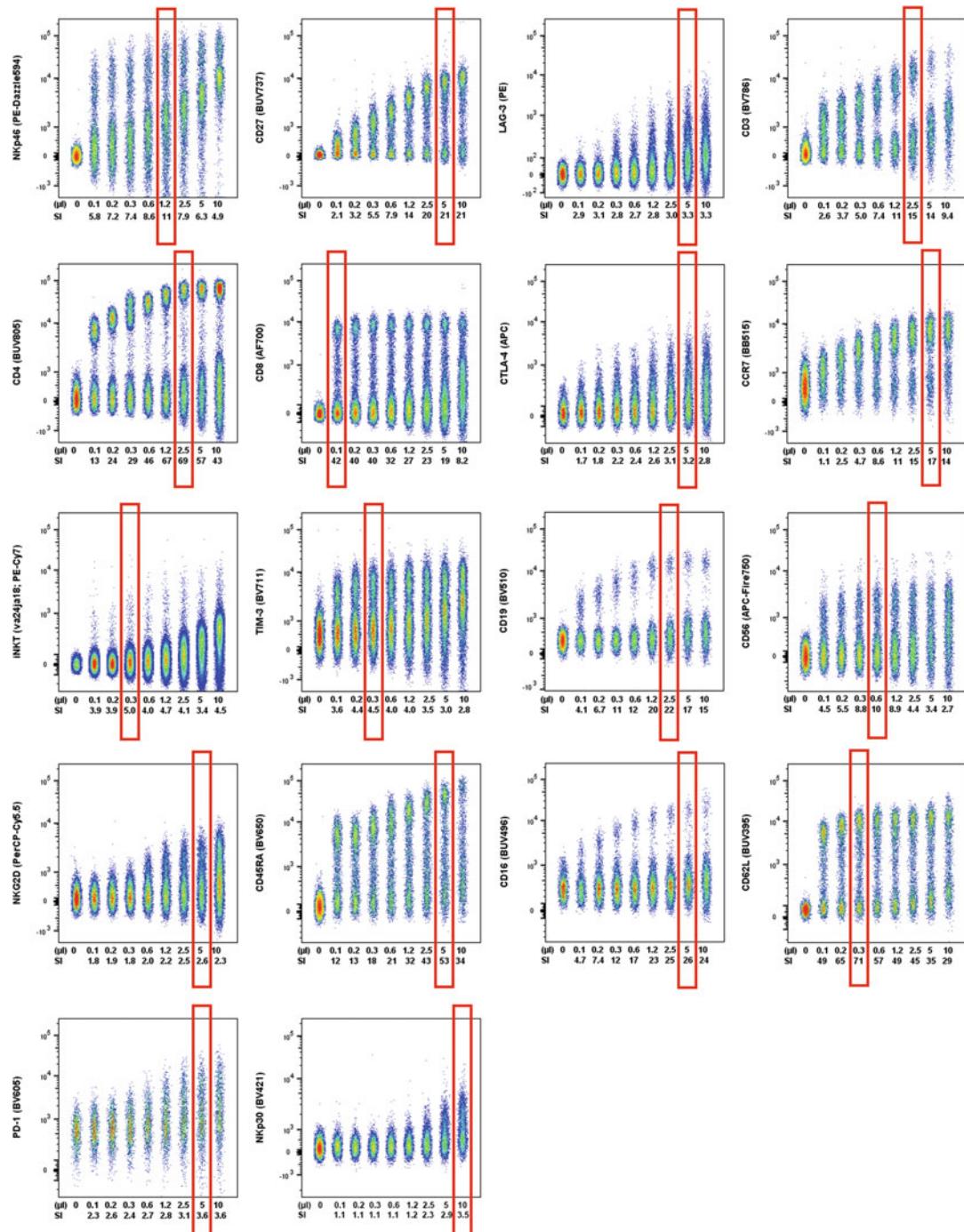
## PMT

SSM	CCR7 (BB515)	CD16 (BUV496)	CD27 BUV737)	CD3 (BV786)	CD4 (BUV805)	CD45RA (BV650)	CD56 (APC-Fire750)	CD62L (BUV395)	CD8 (AF700)	CTLA-4	LAG-3 (PE)	L_D Aqua CD19 (BV510) dump	NKG2D (PerCP-Cy5.5)	NKp30 (BV421)	NKp46 (PE-Dazzle594)	PD-1 (BV605)	TIM-3 (BV711)	iNKT (va24ja18 PE-Cy7)
CCR7 (BB515)	0.00	0.00	0.01	0.37	0.27	0.52	0.40	0.13	0.00	0.00	0.00	0.63	0.00	0.86	0.66	0.00	0.70	0.00
CD16 (BUV496)	4.54	0.00	0.95	0.00	0.20	0.00	0.44	0.34	0.55	0.28	3.23	1.69	0.00	0.00	1.06	0.00	0.00	0.00
CD27 BUV737)	0.00	0.00	0.00	0.45	1.58	0.36	1.46	0.00	2.13	0.22	0.20	0.00	2.52	1.16	0.84	0.00	0.71	0.53
CD3 (BV786)	0.00	0.00	1.48	0.00	1.72	0.92	1.67	0.00	0.76	0.18	0.00	1.35	0.81	3.80	0.77	0.60	0.84	0.52
CD4 (BUV805)	0.00	0.00	0.41	0.33	0.00	0.42	1.42	0.18	0.33	0.13	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.59
CD45RA (BV650)	0.00	0.00	0.88	0.67	0.46	0.00	0.68	0.00	1.16	1.75	0.12	0.67	0.95	0.94	0.78	1.00	1.34	0.29
CD56 (APC-Fire750)	2.19	6.94	0.25	0.73	1.17	1.17	0.00	0.17	1.59	1.03	0.25	0.00	0.57	0.00	1.08	0.77	0.00	1.79
CD62L (BUV395)	2.39	14.01	0.86	0.00	0.21	0.77	0.51	0.00	0.59	0.47	0.03	0.00	1.10	2.17	0.00	0.95	0.73	0.00
CD8 (AF700)	1.35	5.26	1.00	0.27	0.37	0.00	2.18	0.12	0.00	0.68	0.14	0.52	1.71	0.00	0.00	0.39	0.00	0.81
CTLA-4	0.13	0.17	0.45	0.14	0.20	0.53	1.17	0.02	2.43	0.00	0.07	0.00	0.56	0.06	0.13	0.14	0.17	0.46
LAG-3 (PE)	0.36	0.29	0.11	0.06	0.05	0.24	0.00	0.02	0.10	0.13	0.00	0.10	1.56	0.08	1.34	0.37	0.11	0.24
qua CD19 (BV510) dump	1.57	0.00	0.16	0.30	0.31	0.67	0.23	0.19	0.01	0.04	0.16	0.00	0.73	0.90	0.43	0.99	0.00	0.16
NKG2D (PerCP-Cy5.5)	3.38	12.91	0.74	0.00	0.00	0.00	0.00	0.31	1.79	1.27	0.00	2.52	0.00	0.00	0.00	0.99	0.00	0.33
NKp30 (BV421)	1.02	0.00	1.26	1.62	0.55	1.24	1.76	0.00	0.87	0.00	0.44	3.54	0.28	0.00	2.04	0.66	5.31	0.66
NKp46 (PE-Dazzle594)	3.12	6.79	0.85	0.00	0.00	0.00	0.00	0.36	0.50	0.92	15.09	1.39	3.99	0.00	0.00	0.99	0.00	0.56
PD-1 (BV605)	0.16	0.36	0.77	0.42	0.30	1.47	0.08	0.02	0.27	0.33	1.05	0.16	1.10	0.32	1.47	0.00	0.70	0.45
TIM-3 (BV711)	0.51	0.41	2.60	1.60	1.18	0.90	1.64	0.06	2.91	0.77	0.00	0.76	3.80	1.30	0.00	0.29	0.00	0.47
iNKT (va24ja18 PE-Cy7)	3.21	8.58	0.88	0.00	0.25	0.00	0.62	0.35	0.03	0.69	0.56	1.10	1.84	0.00	0.00	0.89	0.00	0.00

**Fig. 1** Spillover spreading matrix (SSM) calculated from single color controls. The matrix is color-coded to reflect SS values with red being higher SS values

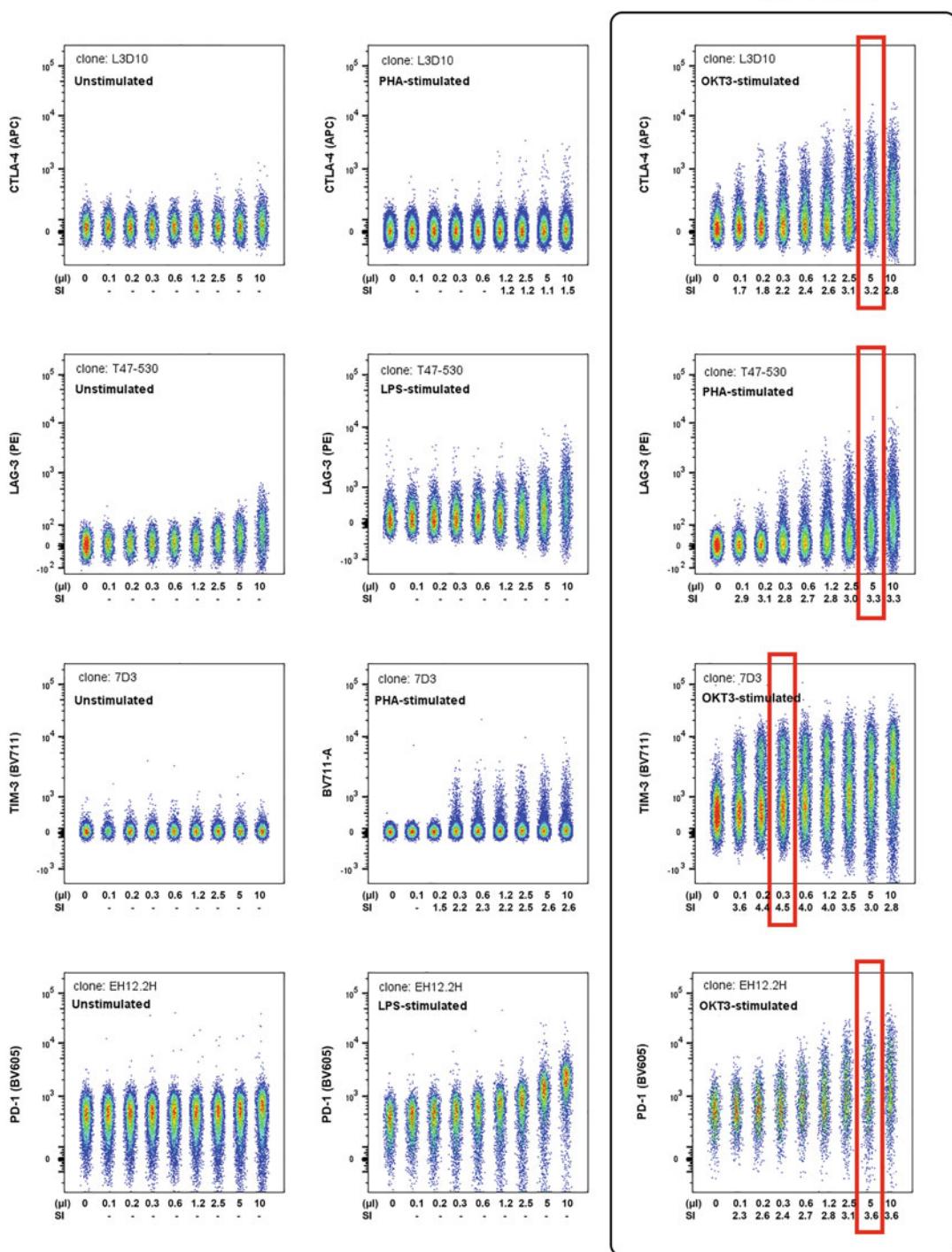
PHA, LPS, or anti-TCR (OKT3) overnight, and selected the best induction condition to calculate SI (Fig. 3).

To build the panel, we progressively added each category set of dyes/antibodies listed in Table 1, beginning with the utility markers using Fixable Live/Dead™ Aqua as per the manufacturer's instructions and with the optimal dilution of staining antibodies following the staining protocol below. We then performed the flow staining using the utility markers and markers in the lymphocyte lineage category. The staining performance of utility markers was then compared to those stained with utility markers alone, and no significant interference was observed. Flow staining was then performed a third and fourth time adding in cumulative succession the markers in the "memory markers" and "regulatory receptors" categories, each time comparing the staining performance of the previously added groups with no significant interferences observed. Compensation was applied using single color controls and a compensation matrix calculated using FlowJo™ software (version X; BD Biosciences). Gates were drawn at fluorescent intensities based on fluorescence minus one (FMO) controls, or along natural breaks in staining patterns that are at least as bright as the FMO(s). Pairwise dot plots of FMO controls are displayed in Fig. 4a–r.

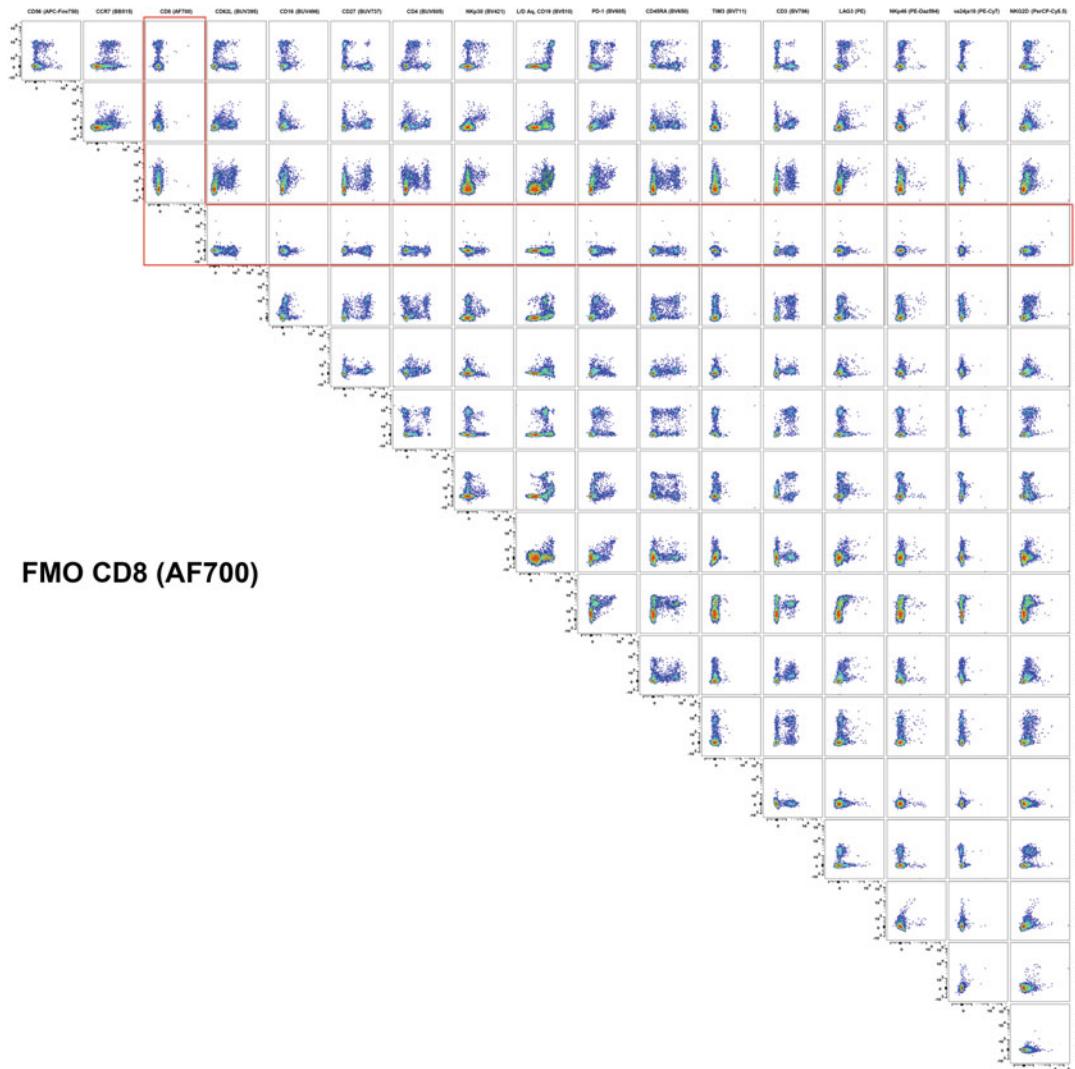


**Fig. 2** Titrations of antibodies used in this OMIP performed on PBMC. Dilutions are  $\mu\text{l}$  of stock antibody concentration provided from the vendor. Staining index was calculated for each dilution using the formula:  $\text{SI} = [(\text{MFI of positive cells}) - (\text{MFI of negative cells})]/(2 \times \text{SD of negative cells})$ . Selected dilutions are demarcated by a red box

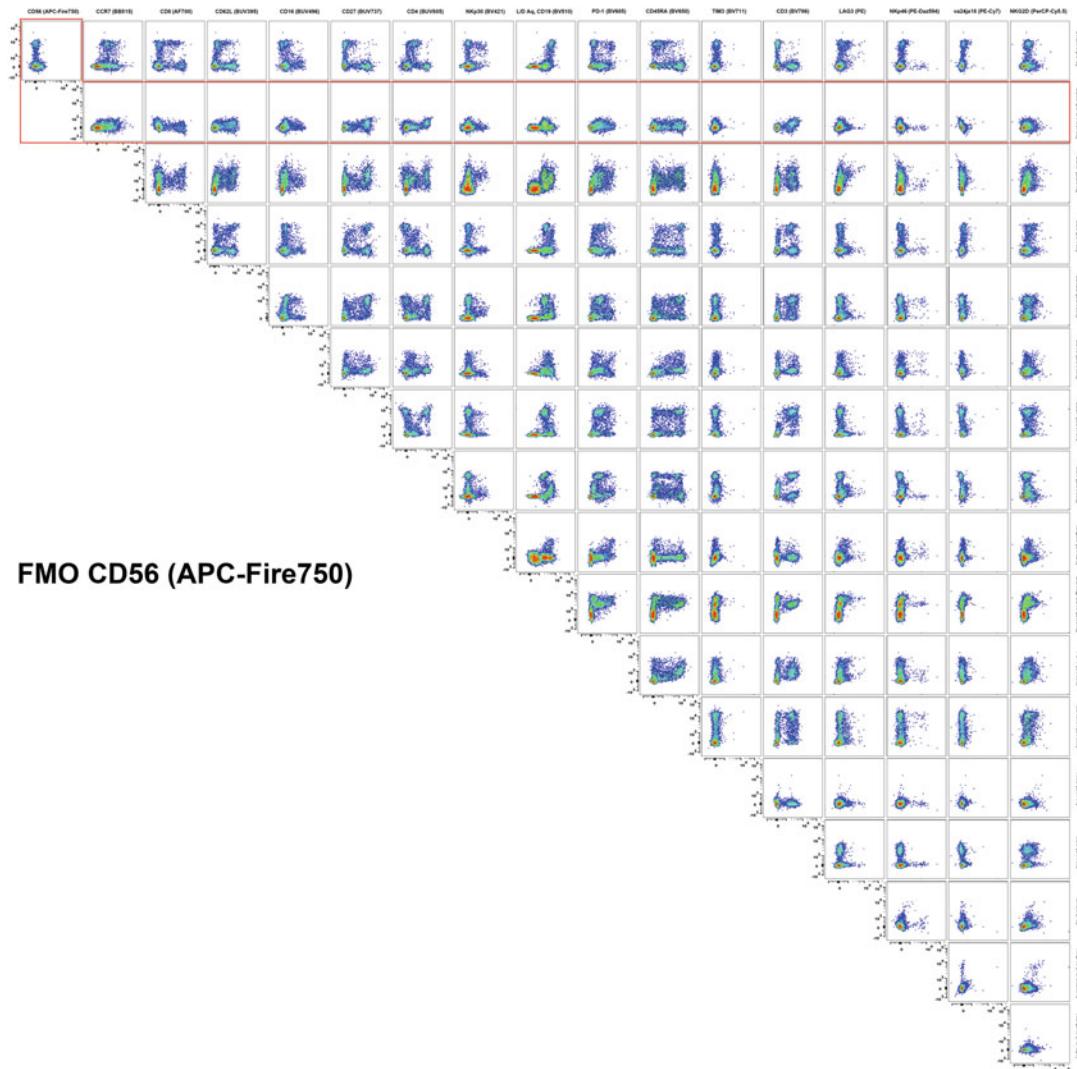
**Stimulation Conditions  
Used for Titration**

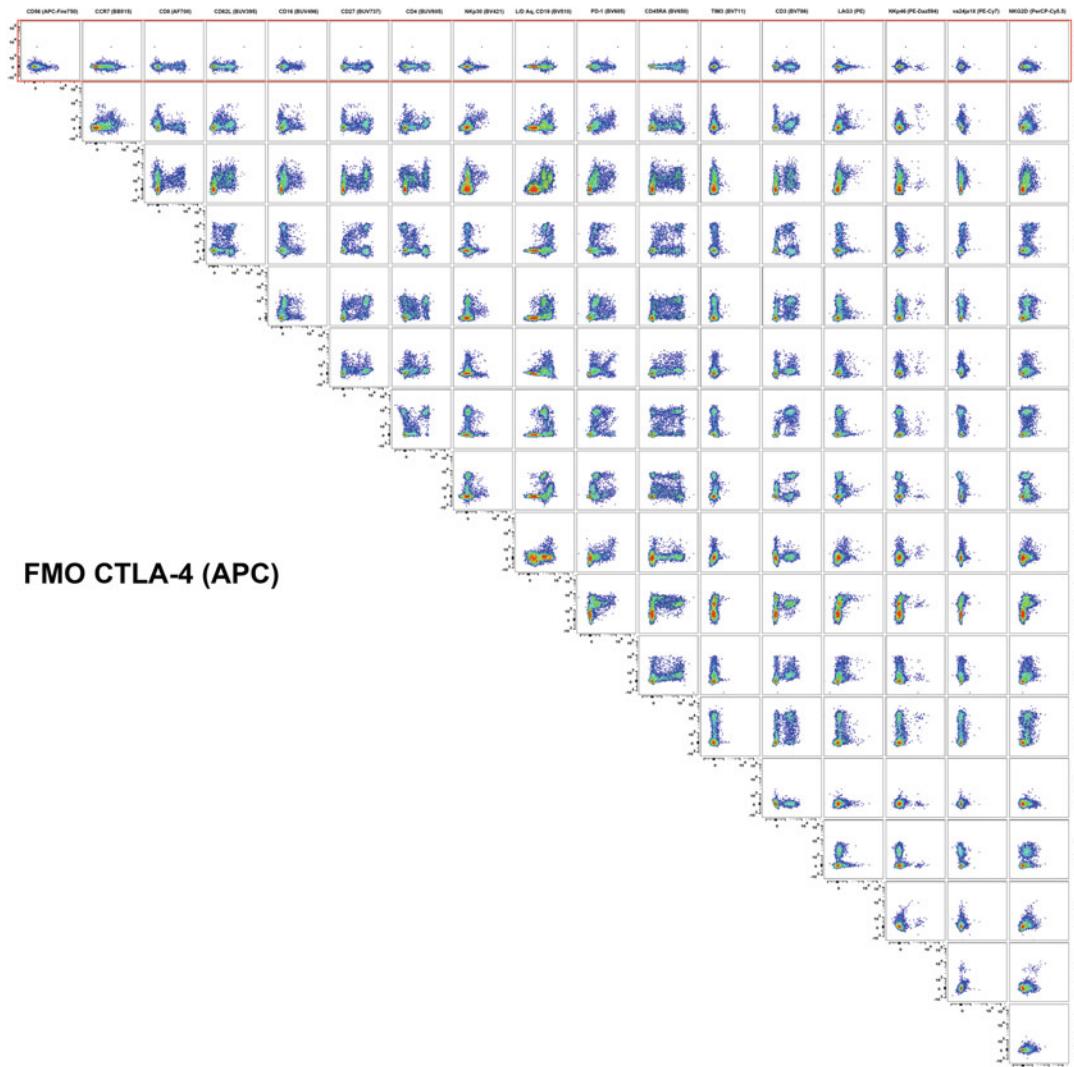


**Fig. 3** Activation conditions needed to titrate checkpoint receptor antibodies. Titrations of antibodies used in this OMIP performed on PBMC. Dilutions are  $\mu\text{l}$  of stock antibody concentration provided from the vendor. Staining index was calculated for each dilution using the formula:  $\text{SI} = [(\text{MFI of positive cells}) - (\text{MFI of negative cells})]/(2 \times \text{SD of negative cells})$ . Selected dilutions are demarcated by a red box

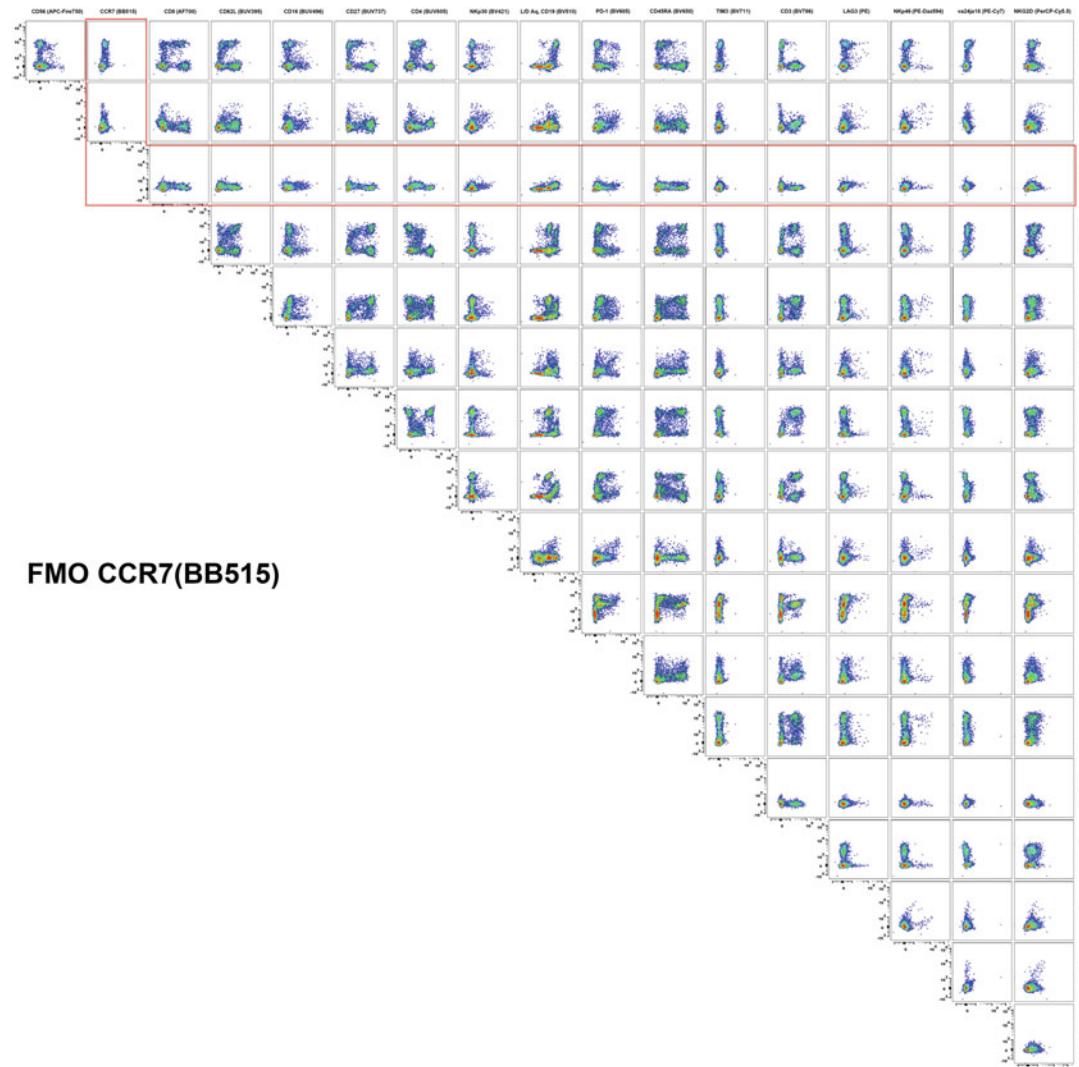


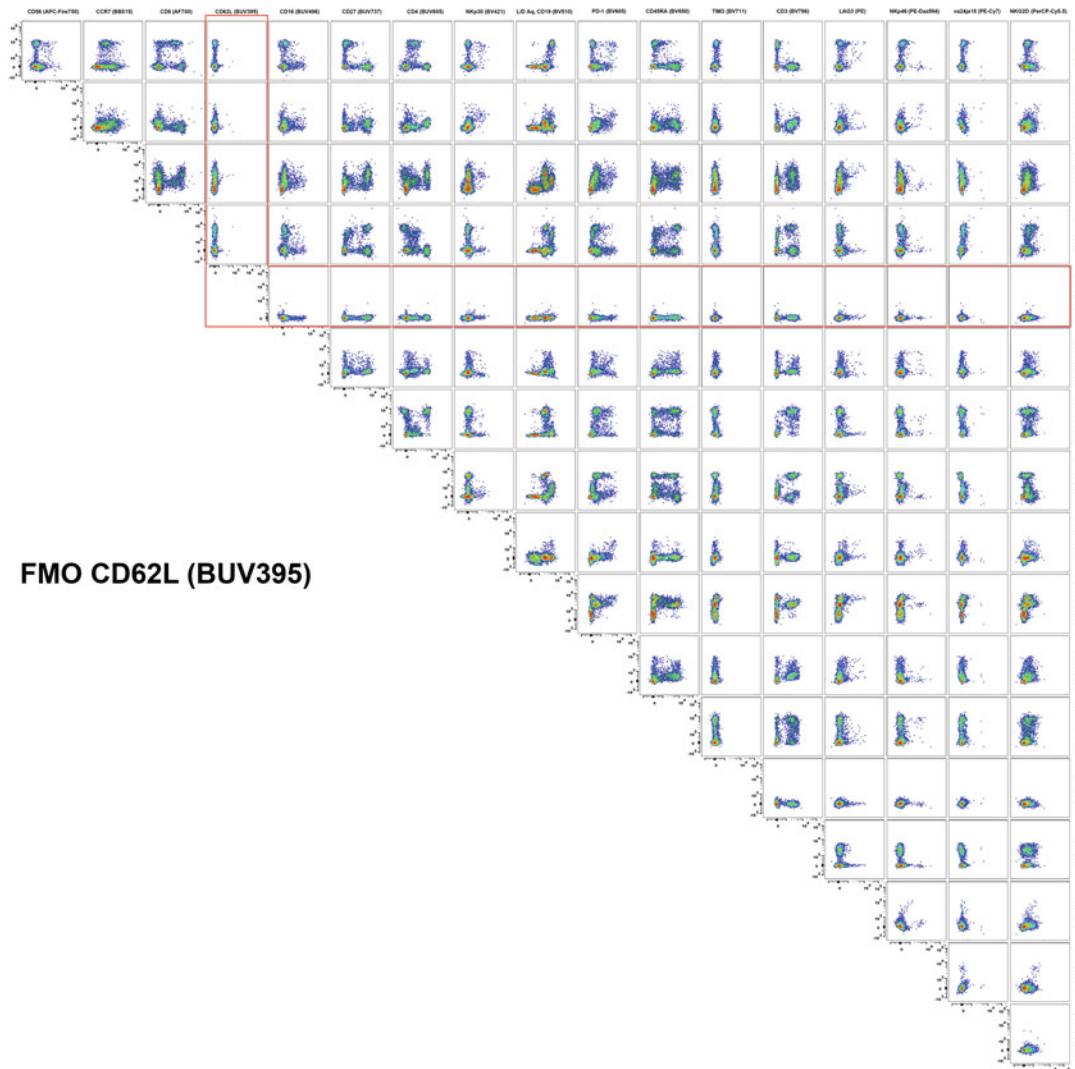
**Fig. 4 (a–r)** Fluorescence minus one (FMO) controls for the staining parameters in this OMIP. Displayed are  $N \times N$  plots showing every possible dot plot combination, and plots relevant to each FMO are outlined in red. Prior to displaying these plots, doublet events were gated out using forward and side scatter height and width characteristics and then cells were selected for based on FSC versus SSC characteristics similar to the main figure

**Fig. 4** (continued)

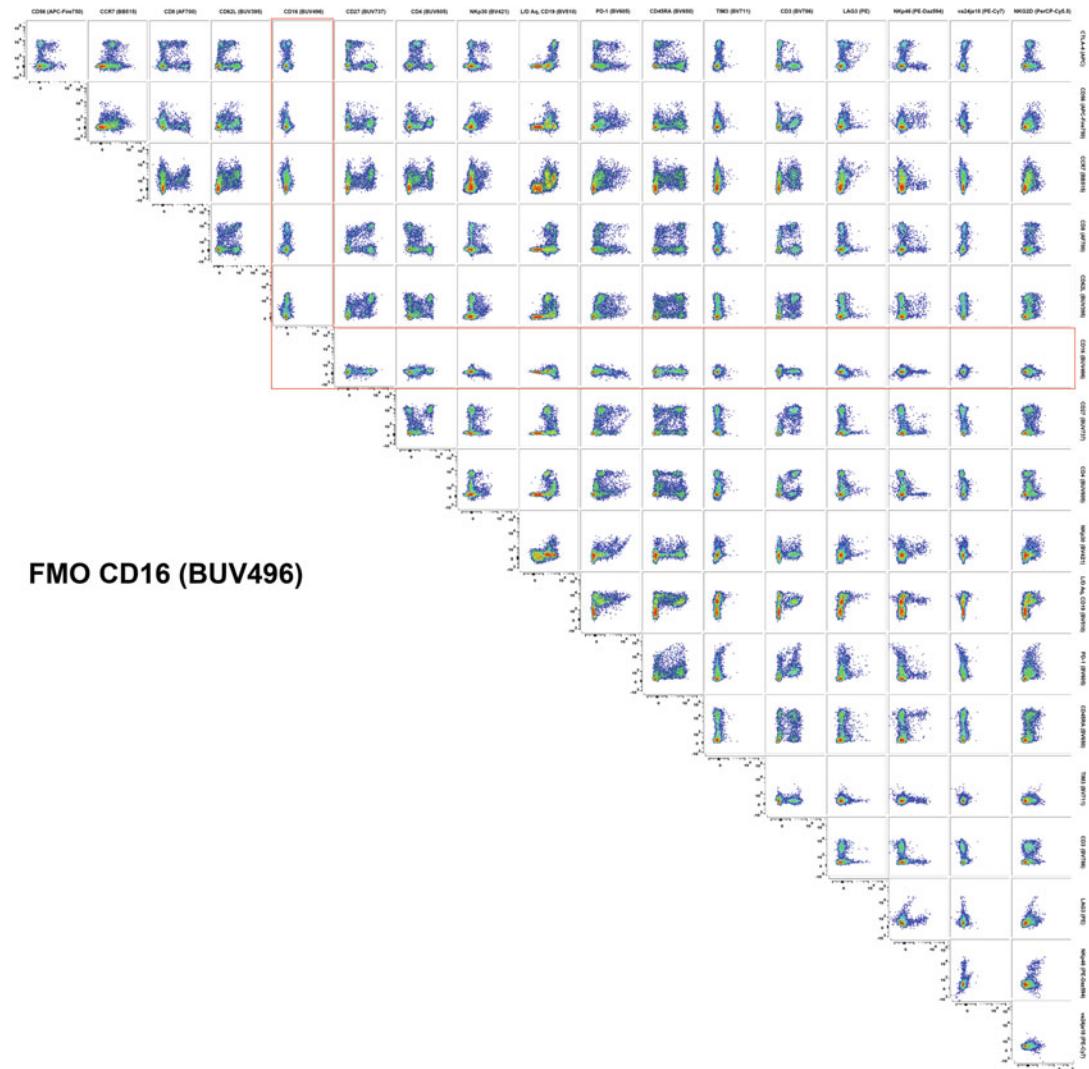


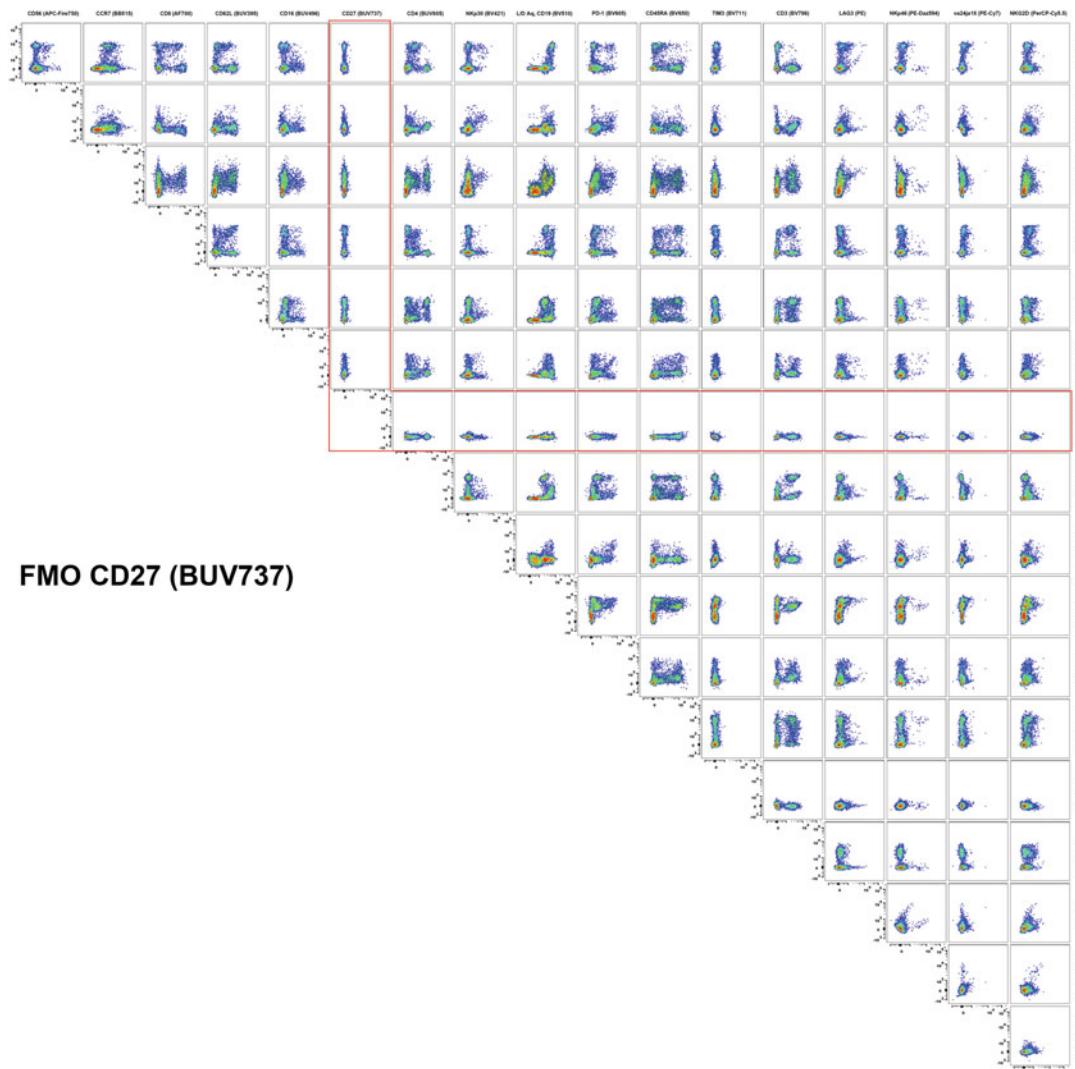
**Fig. 4** (continued)

**Fig. 4** (continued)

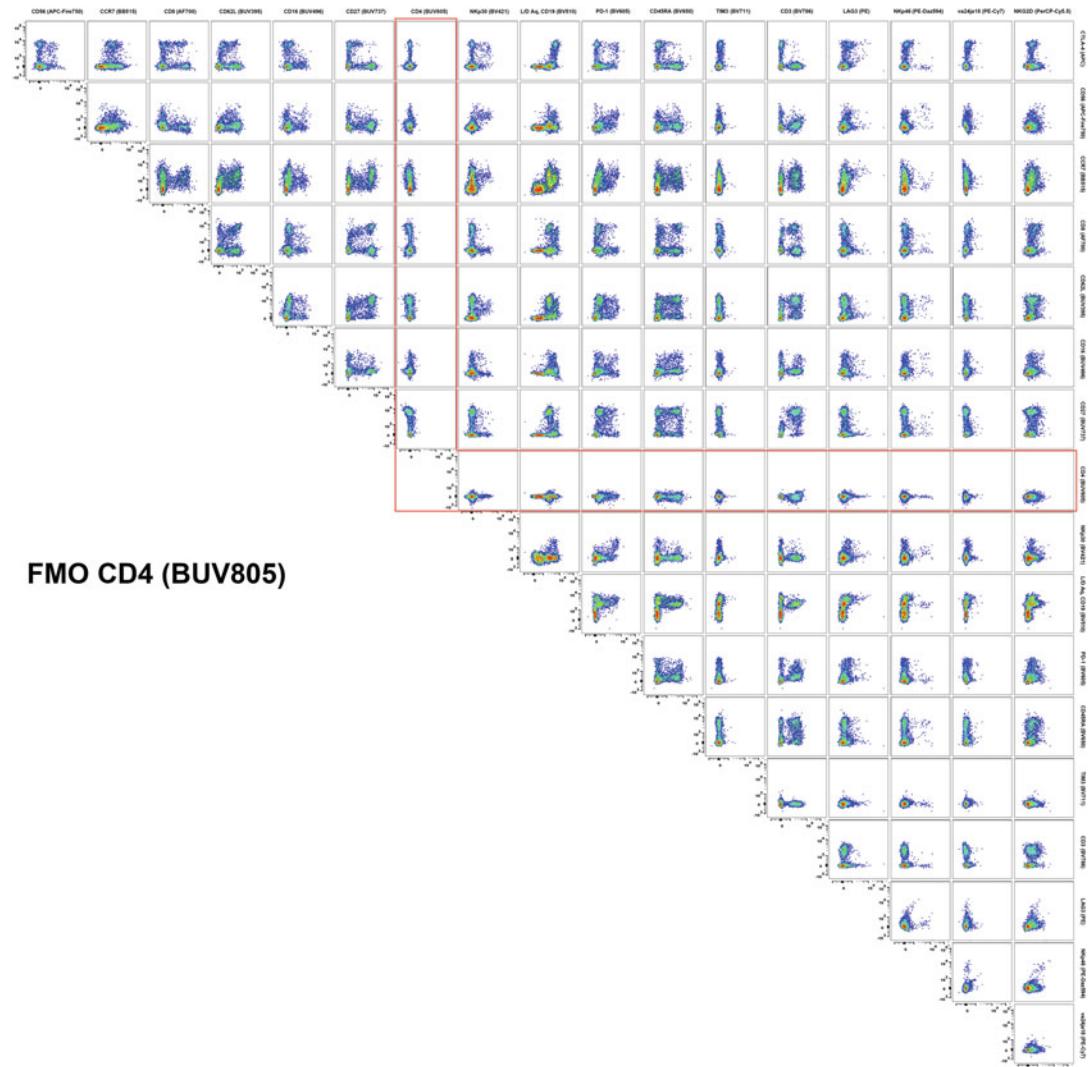


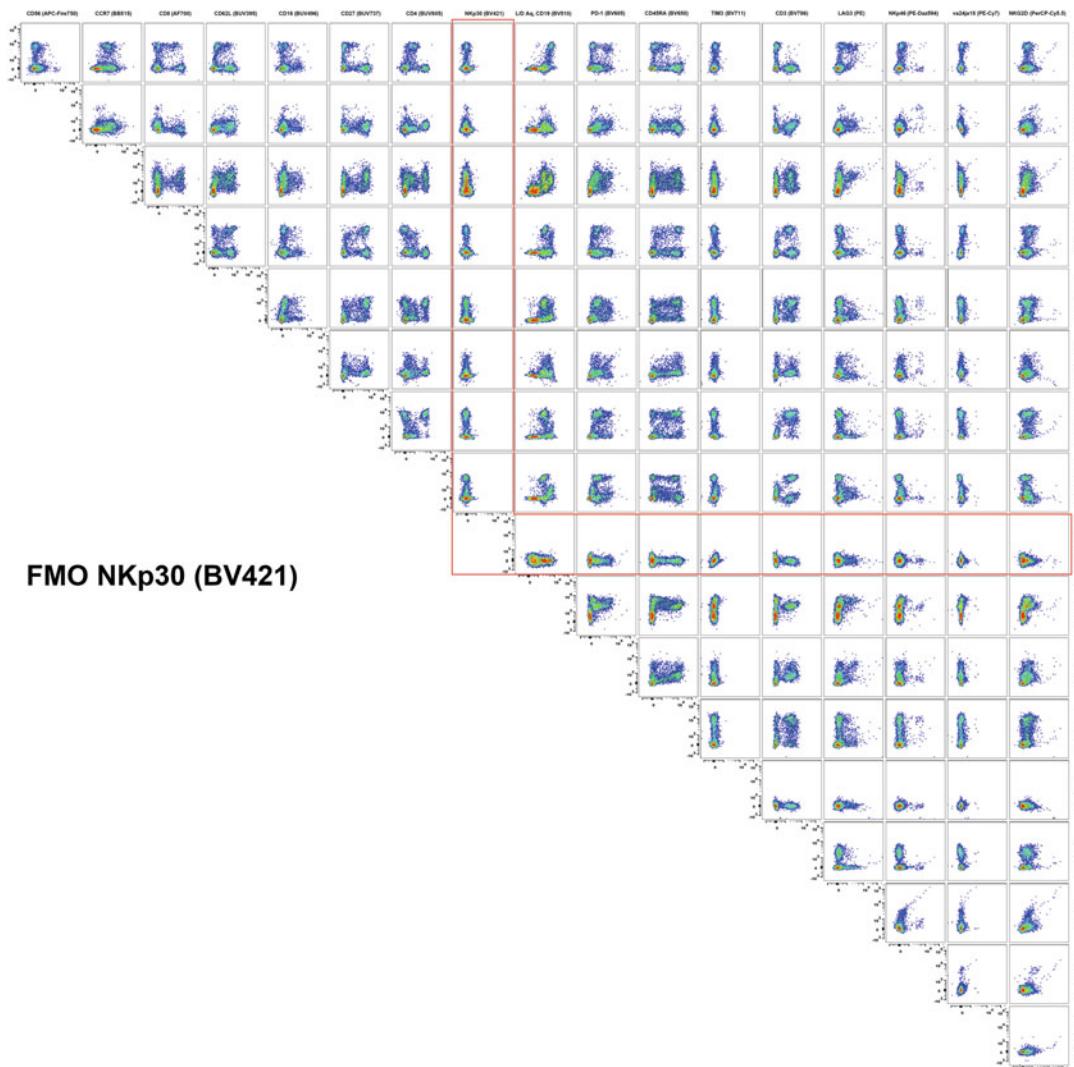
**Fig. 4** (continued)

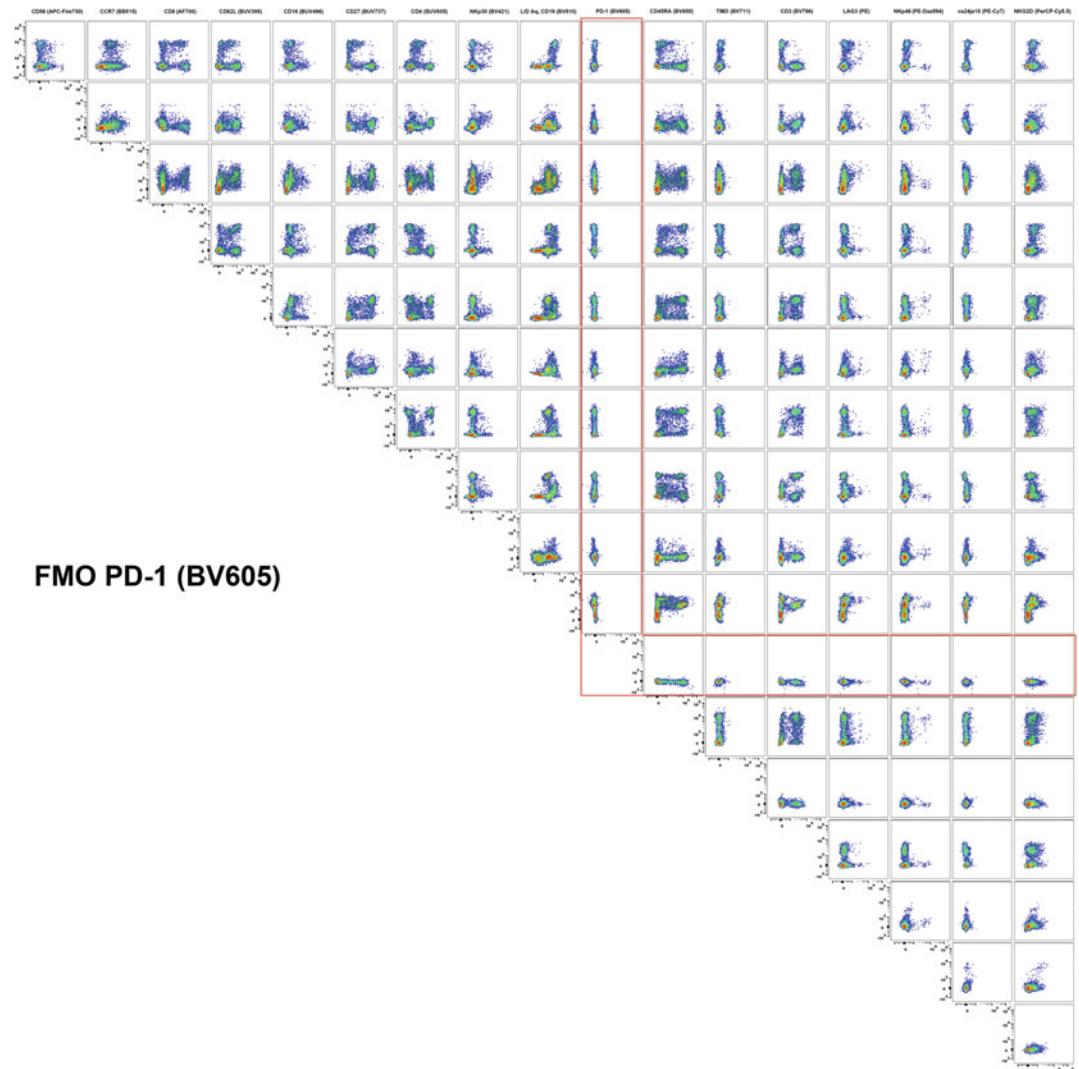
**Fig. 4** (continued)

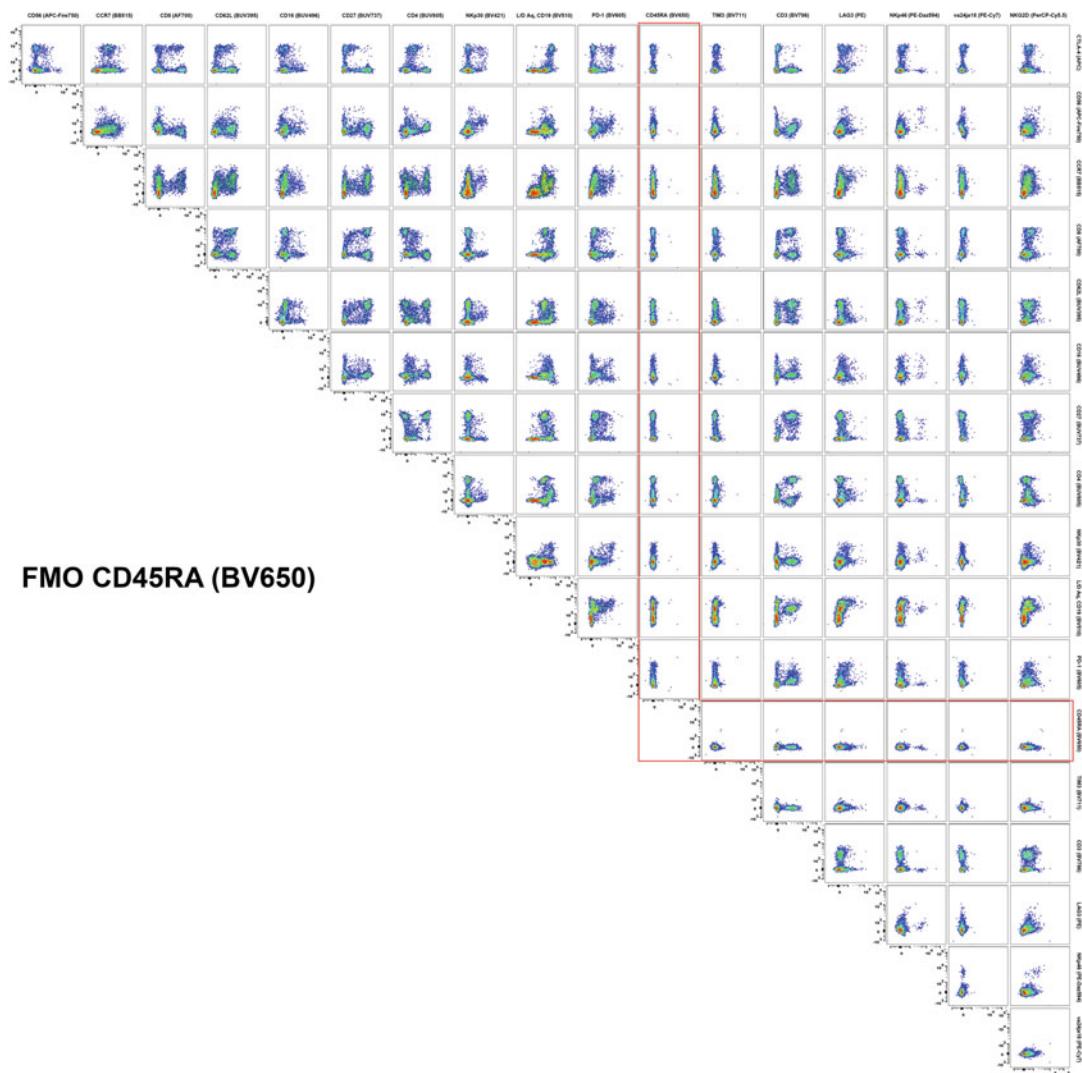


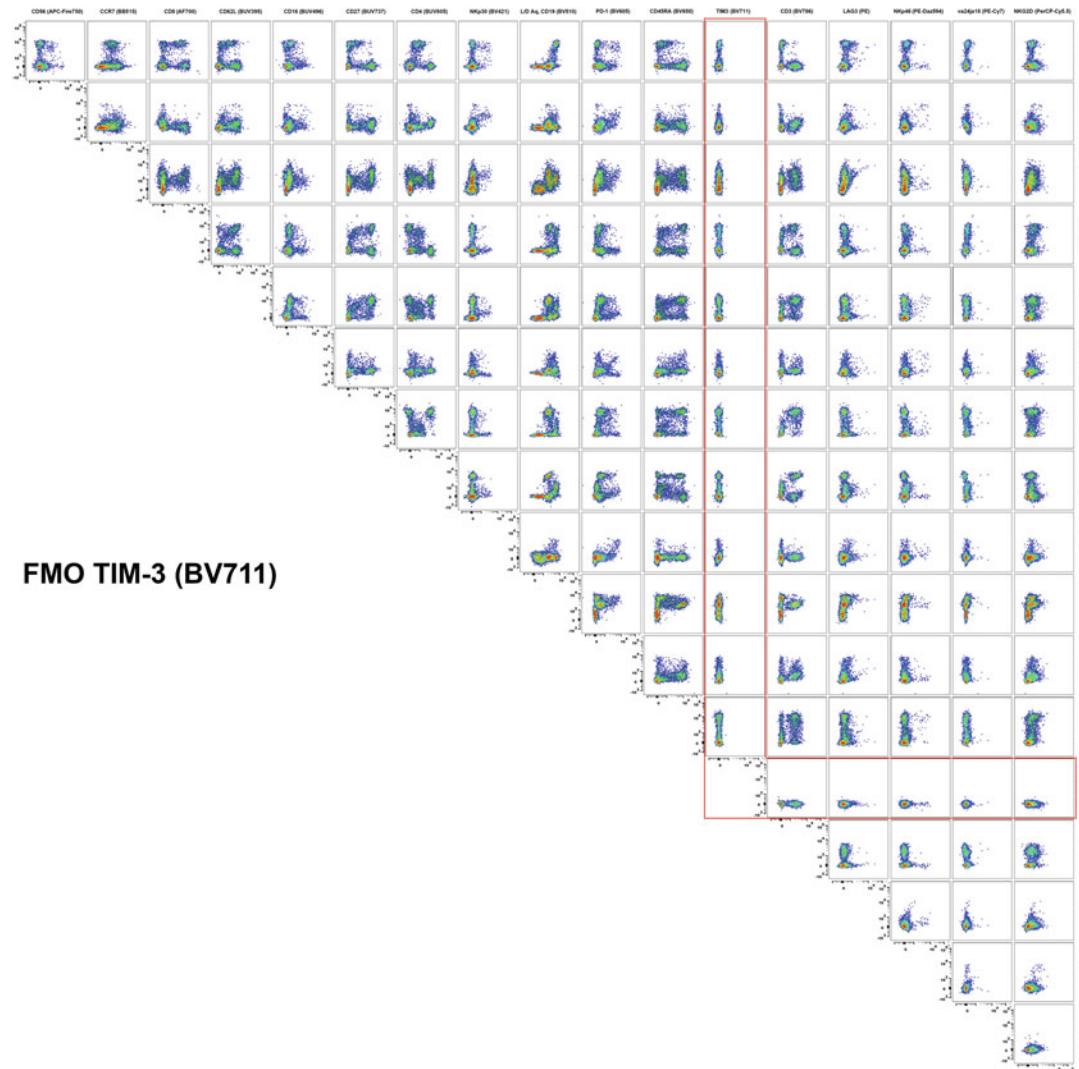
**Fig. 4** (continued)

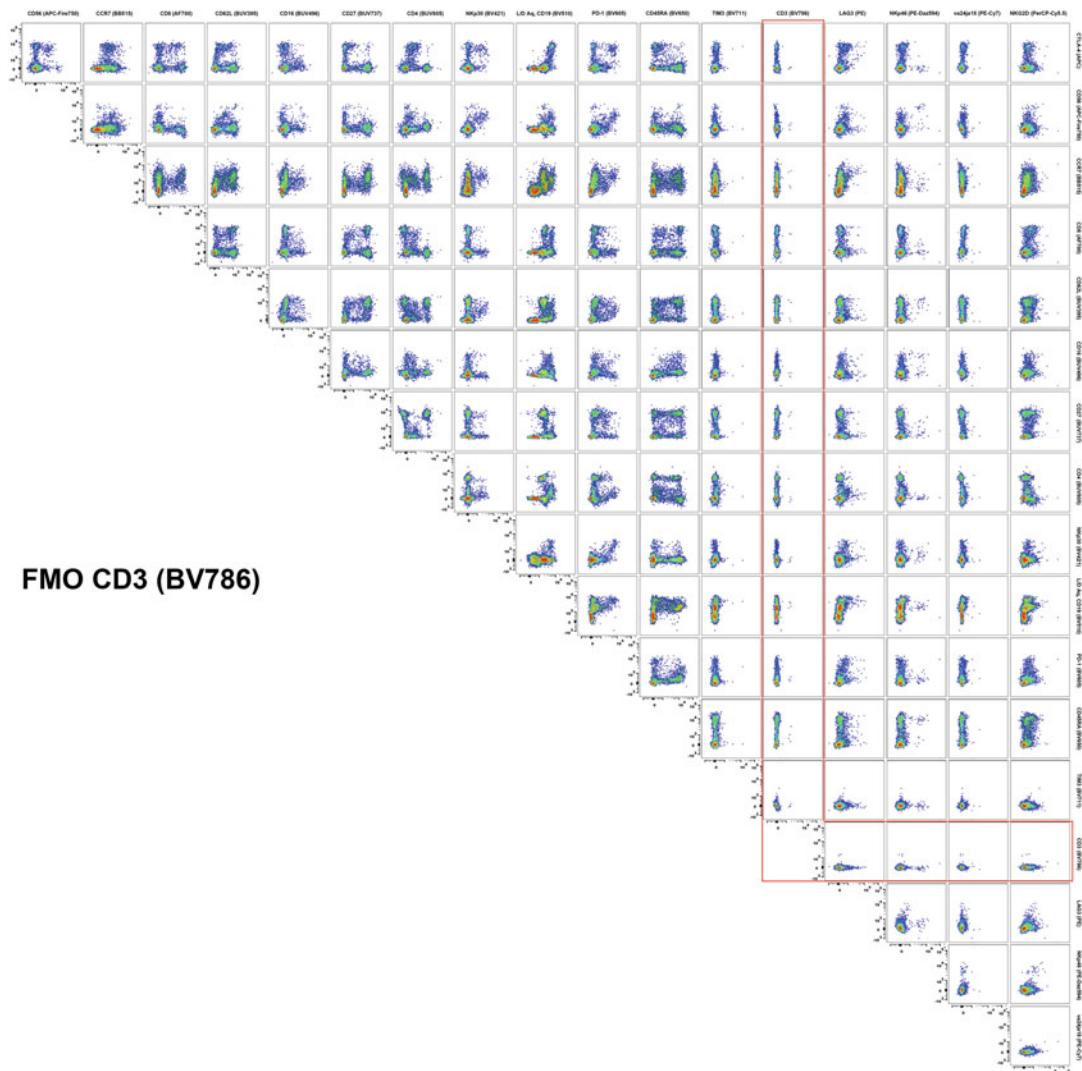
**Fig. 4** (continued)

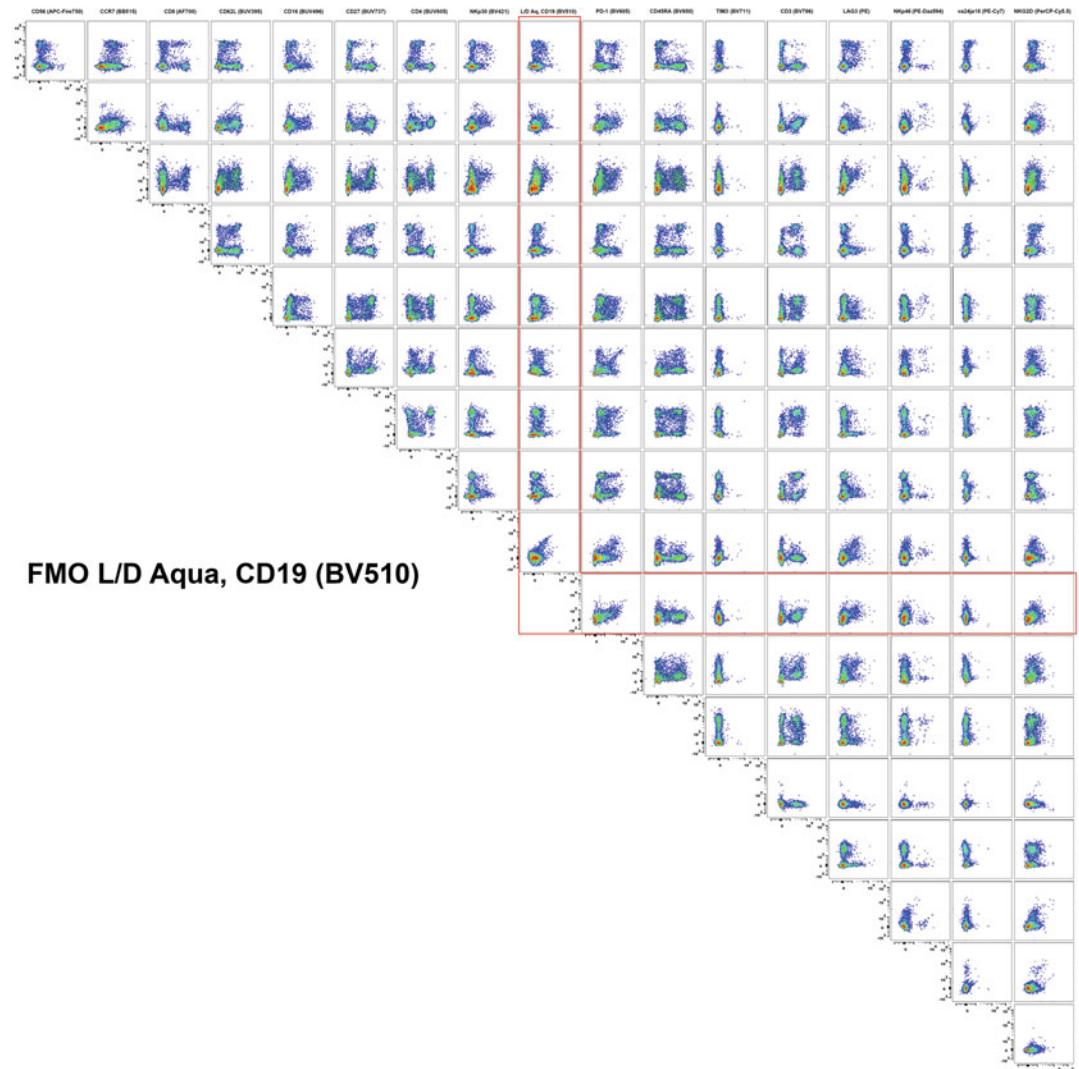
**Fig. 4** (continued)

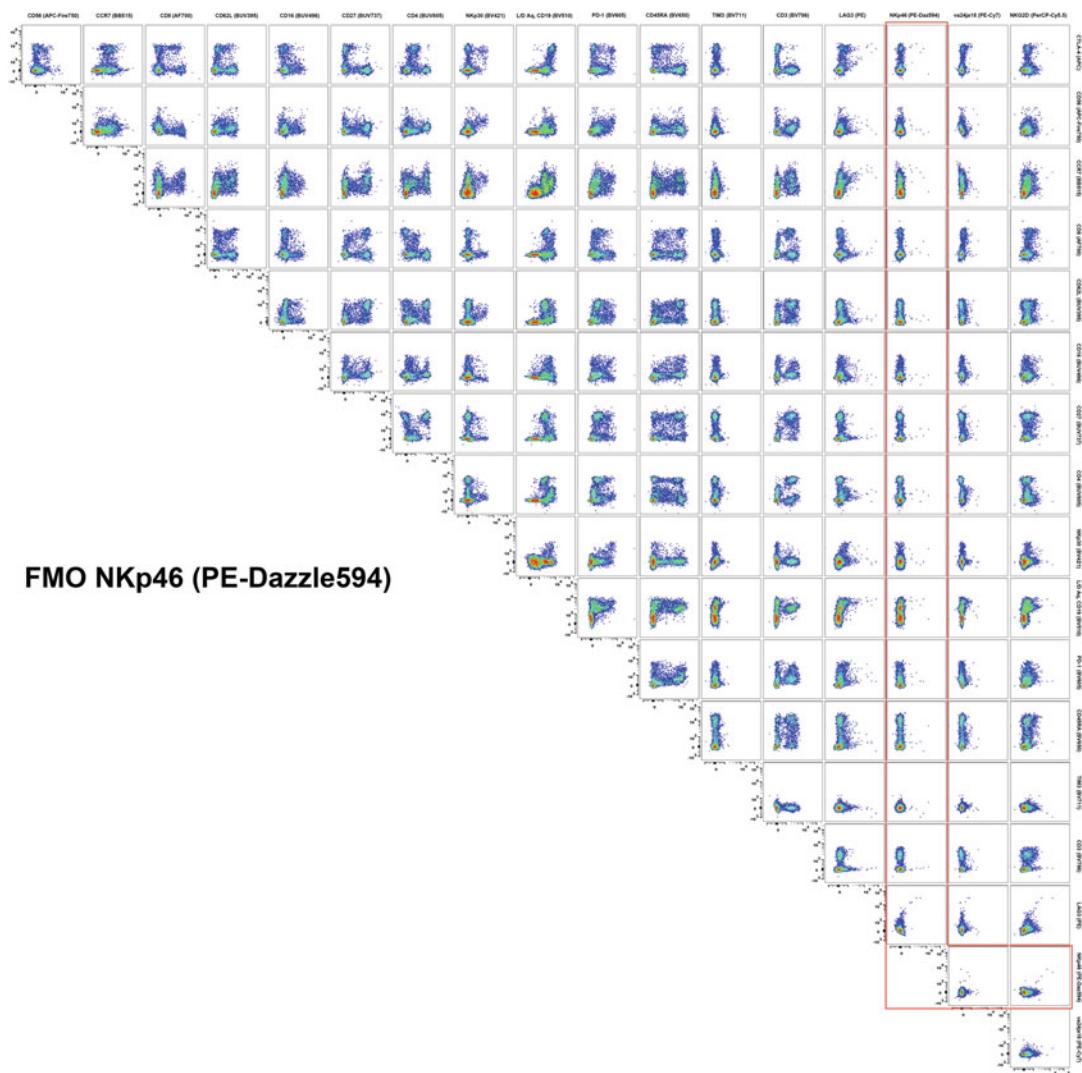
**Fig. 4** (continued)

**Fig. 4** (continued)

**Fig. 4** (continued)

**Fig. 4** (continued)

**Fig. 4** (continued)

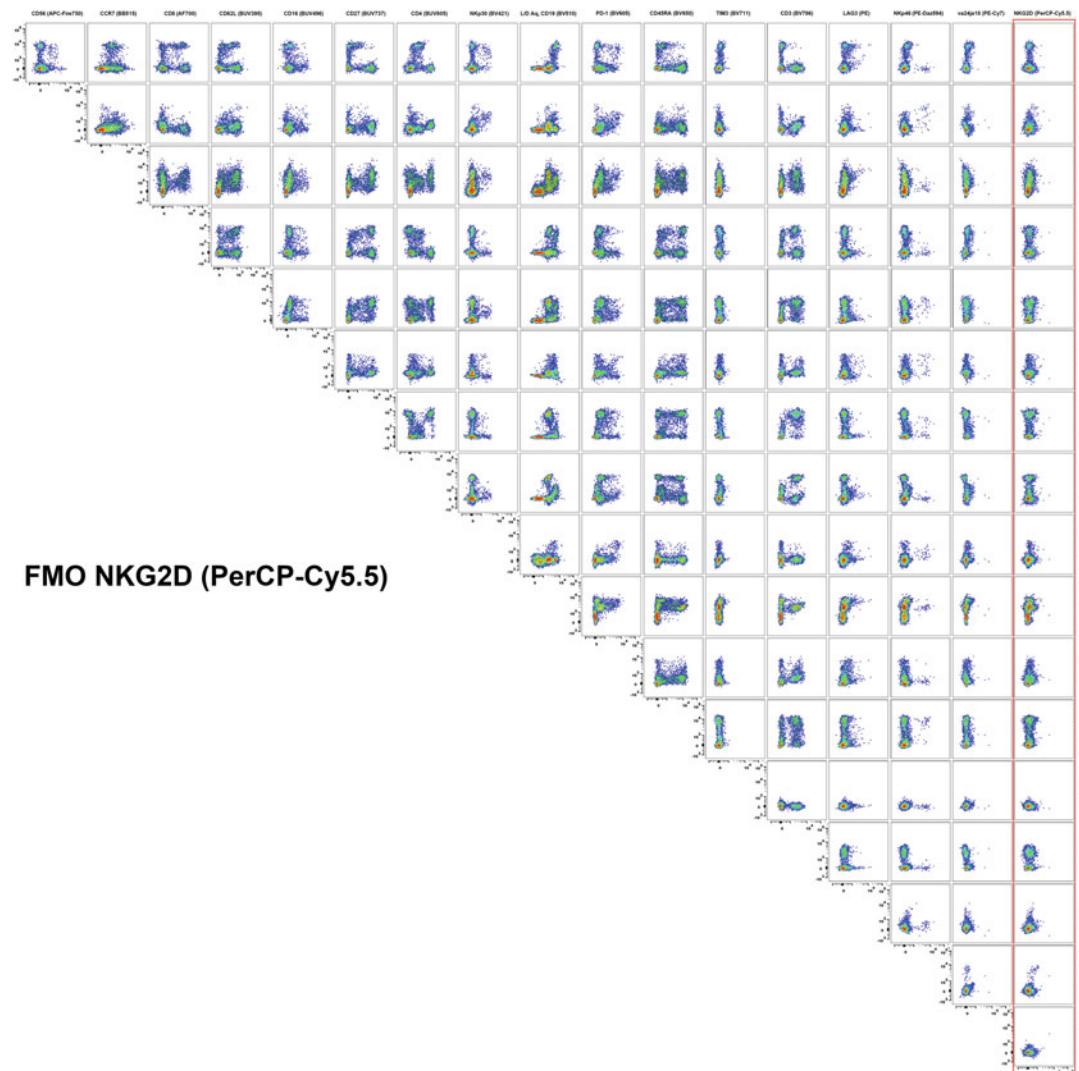


**Fig. 4** (continued)

**Fig. 4** (continued)



**Fig. 4** (continued)

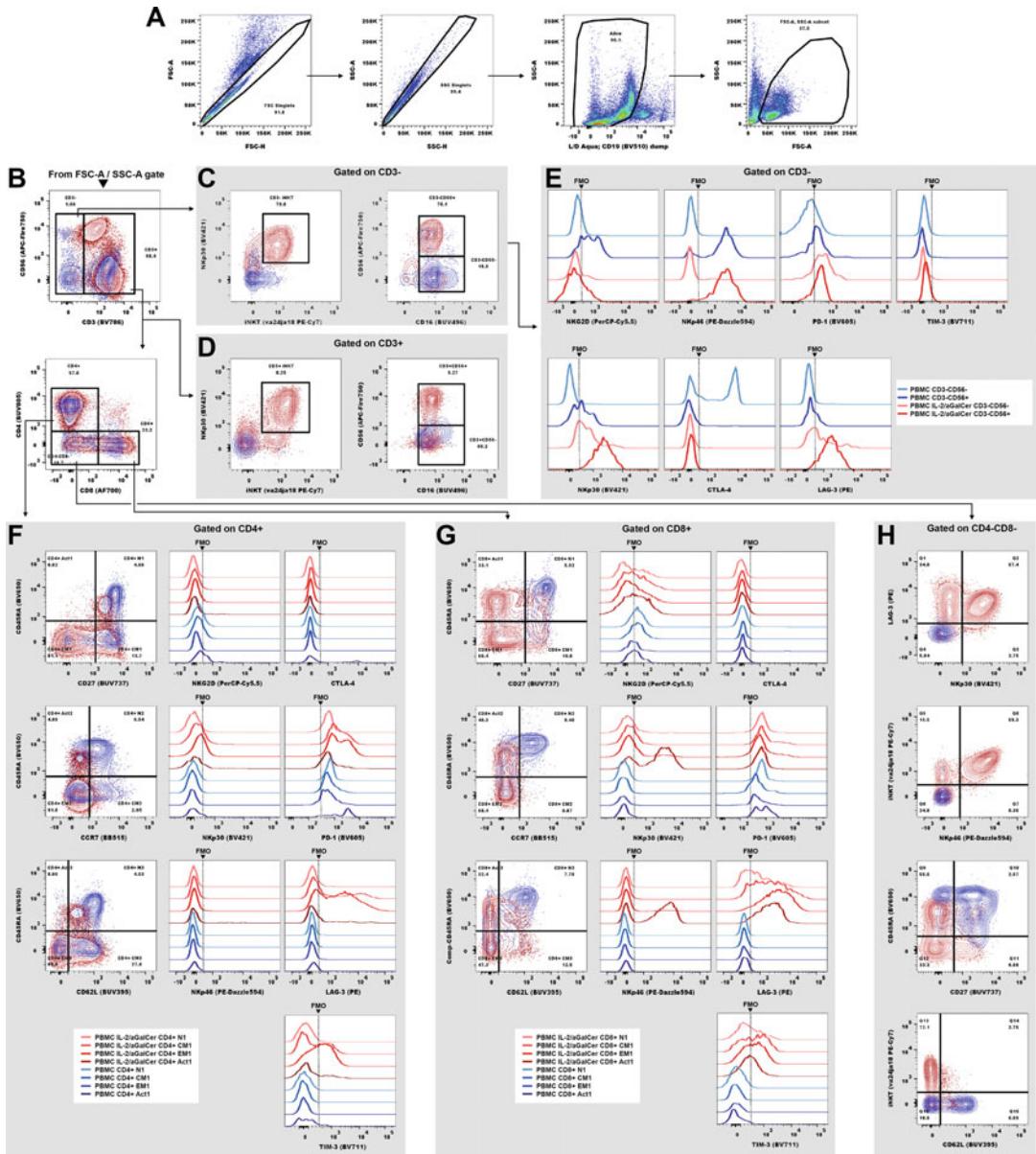
**Fig. 4** (continued)

### 3.3 Gate-Based Analysis

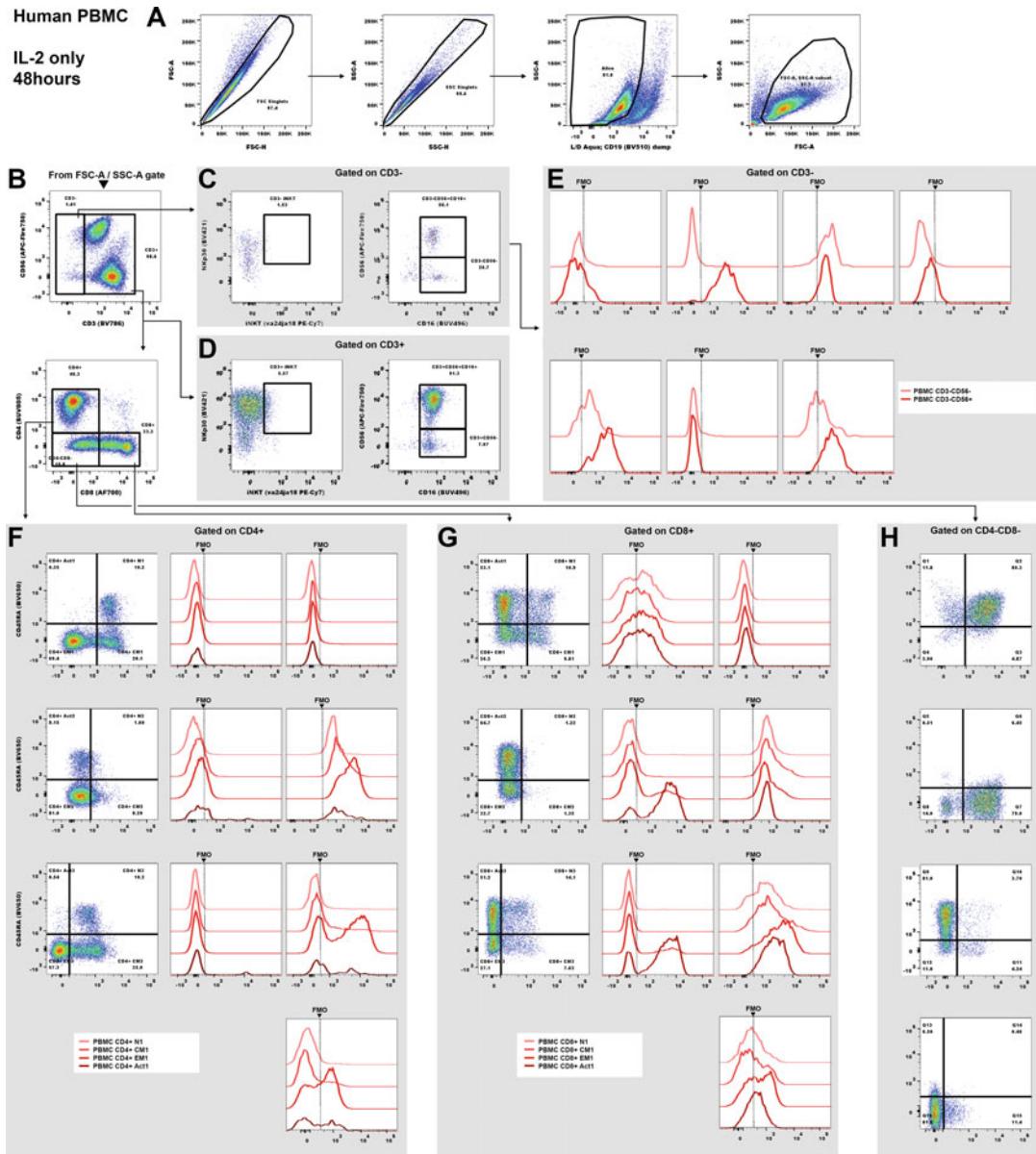
This flow cytometry panel includes a basic framework that allows for the differentiation between conventional NK and T cell subsets and for the identification of invariant and variant NKT cells. This panel was made by first titrating the staining dilution of each antibody and calculating the optimal staining index (Figs. 2 and 3), and then in a step-wise manner building the panel based on category sets of parameters (Table 1), adding each category in turn. One possible gate-based analysis exemplified here, first excludes doublet events and aggregates using forward-scatter (FSC) height and width properties, and then by side-scatter (SSC) height and width properties. This panel includes as a utility anti-CD19 conjugated to brilliant violet® 510 (BV510), which can be excluded from analysis in the same detector as the viability marker Live/Dead™ Aqua. Viable CD19<sup>-</sup> cells are then selected based on low or negative median fluorescent intensity (MFI) using this “dump” channel, and then leukocytes are gated on based on FSC and SSC area parameters (Fig. 5a).

One possible gating schema to analyze this panel, as demonstrated in Figs. 5, 6, and 7, begins by gating out doublet events using forward scatter height and width properties followed by side scatter height and width properties. Viable, CD19<sup>-</sup> events are then selected based on negative or low median fluorescent intensity (MFI) on the Violet E channel before lymphocytes are gated on using side scatter and forward scatter characteristics. T cells and NKT cells are then separated from CD3<sup>-</sup> cells using CD3 staining. Conventional NK cell subsets can then be analyzed in the CD3<sup>-</sup> fraction using CD56, and CD16, whereas NKT populations can be analyzed using CD56, CD16 and the iNKT TCR. Importantly, some iNKT staining is observed in the CD3<sup>-</sup> fraction in cells treated with IL-2 and aGalCer, although this is a minor number of cells given that the CD3<sup>-</sup> fraction represents less than 2% of expanded lymphocytes (Fig. 5) and was not observed in IL-2 or aGalCer-only controls (Figs. 6 and 7). Conventional T cells can then be divided into T<sub>H</sub> and T<sub>C</sub> populations using CD4 and CD8 staining, and naïve, activated, central memory and effector memory can then be defined using the naïve marker CD45RA along with CD27, CCR7, or CD62L. A smaller population of CD3<sup>+</sup>CD4<sup>-</sup>CD8<sup>-</sup> cells can be seen, and may have some overlap with NKT populations. Once subdivided, the expression of activating NK receptors and checkpoint receptors can be assessed on each subpopulation.

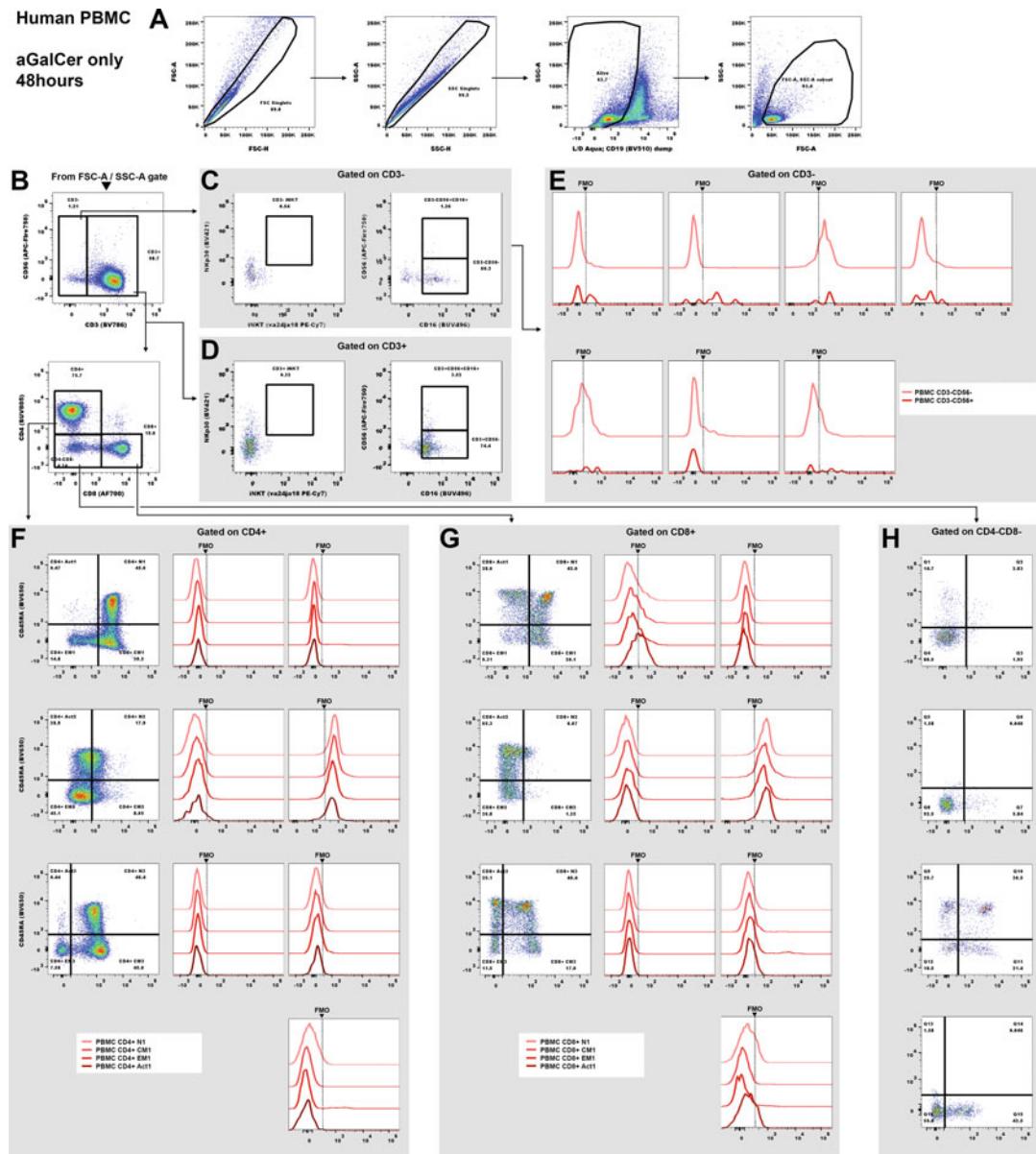
In healthy human blood, the frequency of NKT cells, and in particular iNKT cells, is low and highly variable, ranging from 0% to 1% but averaging less than 0.1% of peripheral blood leukocytes [32]. In addition, many of the regulatory receptors examined by this panel are not expressed on resting lymphocytes, but are inducible in response to activation or other stimuli. To best exemplify the potential of this flow cytometry panel, resting PBMC were analyzed



**Fig. 5** Gating schema for the analysis of regulatory receptors on T cell, NK cell, and NKT cell subsets. **(a)** Single cells are gated on using FSC and then SSC height and width parameters, and then CD19<sup>-</sup> viable cells are gated on based on low MFI on the BV510 channel. Lymphocytes are then selected for by gating on FSC versus SSC parameters. **(b)** CD3<sup>+</sup> are then further gated into CD4<sup>+</sup>, CD8<sup>+</sup>, and CD4<sup>-</sup>CD8<sup>-</sup> populations for further analysis. **(c-e)** Subsets of CD3<sup>+</sup> (**c**) or CD3<sup>-</sup> cells (**d**) can then be subdivided based on CD56 expression or by the expression of va24ja18, before the expression of NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 is analyzed and compared between fresh and IL-2/aGalCer-treated PBMC and defined subpopulations (**e**). **(f-g)** Naïve and memory subsets can be defined on T<sub>H</sub> (**f**) and T<sub>C</sub> cells (**g**) using CD45RA expression versus CD27, CCR7, or CD62L expression. NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 were then analyzed and compared on fresh and IL-2/aGalCer-treated PBMCs and between N, CM, EM, or Act T<sub>H</sub> (**f**) and T<sub>C</sub> cells (**g**). Lastly, CD3<sup>+</sup>CD4<sup>-</sup>CD8<sup>-</sup> cells were analyzed for the expression of LAG-3, NKp30, NKp46, va24ja18, CD45RA, CD27, and CD62L (**h**). Unstimulated PBMC are depicted or overlaid in blue density plots or histograms, while PBMC stimulated with IL-2 and  $\alpha$ -GalCer for 48 h are depicted in red density plots or histograms.



**Fig. 6** IL-2 treatment control. (a) Single cells are gated on using FSC and then SSC height and width parameters, and then CD19<sup>-</sup> viable cells are gated on based on low MFI on the BV510 channel. Lymphocytes are then selected for by gating on FSC versus SSC parameters. (b) CD3<sup>+</sup> are then further analyzed into CD4<sup>+</sup>, CD8<sup>+</sup>, and CD4<sup>-</sup>CD8<sup>-</sup> populations for further analysis. (c-e) Subsets of CD3<sup>+</sup> (c) or CD3<sup>-</sup> cells (d) can then be subdivided based on CD56 expression or by the expression of va24ja18, before the expression of NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 is analyzed on defined subpopulations (e). (f-g) Naïve and memory subsets can be defined on T<sub>H</sub> (f) and T<sub>C</sub> cells (g) using CD45RA expression versus CD27, CCR7, or CD62L expression. NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 were then analyzed on N, CM, EM, or Act T<sub>H</sub> (f) and T<sub>C</sub> cells (g). Lastly, CD3<sup>+</sup>CD4<sup>-</sup>CD8<sup>-</sup> cells were analyzed for the expression of LAG-3, NKp30, NKp46, va24ja18, CD45RA, CD27, and CD62L (h)



**Fig. 7** aGalCer treatment control. **(a)** Single cells are gated on using FSC and then SSC height and width parameters, and then CD19<sup>-</sup> viable cells are gated on based on low MFI on the BV510 channel. Lymphocytes are then selected for by gating on FSC versus SSC parameters. **(b)** CD3<sup>+</sup> are then further gated into CD4<sup>+</sup>, CD8<sup>+</sup>, and CD4<sup>-</sup>CD8<sup>-</sup> populations for further analysis. **(c-e)** Subsets of CD3<sup>+</sup> (**c**) or CD3<sup>-</sup> cells (**d**) can then be subdivided based on CD56 expression or by the expression of va24ja18, before the expression of NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 is analyzed on defined subpopulations (**e**). **(f-g)** Naïve and memory subsets can be defined on T<sub>H</sub> (**f**) and T<sub>C</sub> cells (**g**) using CD45RA expression versus CD27, CCR7, or CD62L expression. NKG2D, NKp30, NKp46, CTLA-4, LAG-3, TIM-3, and PD-1 were then analyzed on N, CM, EM, or Act T<sub>H</sub> (**f**) and T<sub>C</sub> cells (**g**). Lastly, CD3<sup>+</sup>CD4<sup>-</sup>CD8<sup>-</sup> cells were analyzed for the expression of LAG-3, NKp30, NKp46, va24ja18, CD45RA, CD27, and CD62L (**h**).

alongside PBMC cultured in recombinant human interleukin 2 (IL-2) and aGalCer for 2 weeks to expand iNKT cells and to induce the expression of checkpoint receptors on different lymphocyte populations (Fig. 5b–h). After gating on single viable CD19<sup>−</sup> leukocytes, conventional T<sub>H</sub> and T<sub>C</sub> cells can be gated based on CD3<sup>+</sup> expression, and then CD4<sup>+</sup> and CD8<sup>+</sup> expression respectively (Fig. 5b), while NK, iNKT, and vNKT cells can be gated using CD56 and CD16 expression, or va24ja18 expression respectively (Fig. 5c). While the majority of expanded iNKT cells are CD3<sup>+</sup>, some expanded va24ja18<sup>+</sup> cells can be found in the CD3<sup>−</sup> gate as well (Fig. 5d). Differential expression of checkpoint receptors or activating NK receptors were then compared between CD3<sup>+</sup>CD56<sup>+</sup> and CD3<sup>+</sup>CD56<sup>−</sup> cells in fresh or IL-2/aGalCer-treated PBMC (Fig. 5e).

For conventional T cell populations, naïve (N), central memory (CM), effector memory (EM), and activated (Act) phenotypes can be defined using CD45RA staining versus a central marker such as CD27, CCR7, or CD62L. For conventional T<sub>H</sub> and T<sub>C</sub> cells, phenotypes defined by these central markers are largely overlapping. Here, we compared the expression of checkpoint receptors or activating NK receptors on N, CM, EM, and Act T cells in both T<sub>H</sub> (Fig. 5f) and T<sub>C</sub> (Fig. 5g) subsets as defined by CD45RA and CD27 expression, although a similar comparison could be made using CD45RA staining along with CCR7 or CD62L expression.

Lastly, the gating schema here explores the phenotypes of CD3<sup>+</sup>CD4<sup>−</sup> and CD3<sup>+</sup>CD8<sup>−</sup> cells, which are more frequent in IL-2/aGalCer-treated PBMCs (Fig. 5b). A large portion of these treated cells express the activating NK receptors NKp30 and NKp46, are va24ja18<sup>+</sup>, have up-regulated Lag-3 expression, and downregulated CD45RA, CD62L, and CD27 expression (Fig. 5h) suggesting that they are expanded iNKT cells and overlap with those populations defined in Fig. 5d. In contrast, these CD3<sup>+</sup>CD4<sup>−</sup>CD8<sup>−</sup> cells in fresh PBMC lack expression of NKp30, NKp46, va24ja18, LAG-3, and maintain expression of central markers such as CD62L and CD27 and are largely positive for the naïve marker CD45RA, suggesting that these cells are largely naïve CD4<sup>−</sup>CD8<sup>−</sup> T cells.

Beyond the gating schema exemplified here, it is now appreciated that NK, and NKT cells exist along a spectrum of phenotypes, which now encompass six distinct subsets of NKT cells [4], and two distinct populations of NK cells [33] which can be further subdivided based on the expression of CD27, CCR7, CD62L, and NK receptors [34, 35]. This phenotypic spectrum is further complexed by the description of three types of innate lymphoid cells (ILCs) called ILC-1, ILC-2, and ILC-3 [36] which express no antigen receptors, but instead react to environmental queues through activating and inhibitory receptors including checkpoint and NK receptors [37, 38]. ILC 1–3 may serve a supportive or

regulatory role analogous to helper T cells ( $T_H$ ) subsets  $TH_1$ ,  $TH_2$ , and  $TH_{17}$  respectively, which greatly influence the abundance and effectiveness of other cytotoxic populations [39]. In healthy PBMC, ILC make up less than 0.01% of circulating lymphocytes [40], and thus the PBMC samples used here suboptimal for ILC analysis. However, this panel in theory could also be useful to investigators wishing to study ILC in other context. In such case, the addition of CD127 would aid in the differentiation of ILC from other lymphocytes, and the addition of CRTH2 and c-kit would aid in the subdifferentiation of ILC-1, ILC-2, and IL-C3 cell types within the ILC compartment [41].

## 4 Notes

- Brilliant Stain Buffer is a proprietary additive available from BD Biosciences that minimizes the interactions of Sirigen polymer dyes (such as Brilliant Violet or Brilliant Ultraviolet fluorophores). If only one or less Brilliant Violet or Brilliant Ultraviolet fluorophore is used, this additive has no benefit.

## Acknowledgments

This work was supported by the Moffitt Cancer Center—Innovative Core Projects (Project number 16060201), NCI–NIH (1 R01 CA148995-01; P30CA076292; P50CA168536), the V Foundation, the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, and the Chris Sullivan Foundation.

## References

- Klebanoff CA, Gattinoni L, Restifo NP (2012) Sorting through subsets: which T-cell populations mediate highly effective adoptive immunotherapy? *J Immunother* 35(9):651–660
- Colonna M, Nakajima H, Navarro F, Lopez-Botet M (1999) A novel family of Ig-like receptors for HLA class I molecules that modulate function of lymphoid and myeloid cells. *J Leukoc Biol* 66(3):375–381
- Raulet DH, Vance RE (2006) Self-tolerance of natural killer cells. *Nat Rev Immunol* 6(7):520–531
- Godfrey DI, Stankovic S, Baxter AG (2010) Raising the NKT cell family. *Nat Immunol* 11(3):197–206
- Bendelac A (1995) Positive selection of mouse NK1+ T cells by CD1-expressing cortical thymocytes. *J Exp Med* 182(6):2091–2096
- Bendelac A, Lantz O, Quimby ME, Yewdell JW, Bennink JR, Brutkiewicz RR (1995) CD1 recognition by mouse NK1+ T lymphocytes. *Science* 268(5212):863–865
- Exley M, Garcia J, Balk SP, Porcelli S (1997) Requirements for CD1d recognition by human invariant Valpha24+ CD4–CD8– T cells. *J Exp Med* 186(1):109–120
- Patel SA, Minn AJ (2018) Combination cancer therapy with immune checkpoint blockade: mechanisms and strategies. *Immunity* 48(3):417–433
- van den Broek T, Borghans JAM, van Wijk F (2018) The full spectrum of human naive T cells. *Nat Rev Immunol* 18(6):363–373
- Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the T

- cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol* 14(6):377–391
11. Carter L, Fouser LA, Jussif J et al (2002) PD-1: PD-L inhibitory pathway affects both CD4(+) and CD8(+) T cells and is overcome by IL-2. *Eur J Immunol* 32(3):634–643
  12. Grosso JF, Kelleher CC, Harris TJ et al (2007) LAG-3 regulates CD8+ T cell accumulation and effector function in murine self- and tumor-tolerance systems. *J Clin Invest* 117(11):3383–3392
  13. Latchman Y, Wood CR, Chernova T et al (2001) PD-L2 is a second ligand for PD-1 and inhibits T cell activation. *Nat Immunol* 2(3):261–268
  14. Ngiow SF, von Scheidt B, Akiba H, Yagita H, Teng MW, Smyth MJ (2011) Anti-TIM3 antibody promotes T cell IFN-gamma-mediated antitumor immunity and suppresses established tumors. *Cancer Res* 71(10):3540–3551
  15. Yang YF, Zou JP, Mu J et al (1997) Enhanced induction of antitumor T-cell responses by cytotoxic T lymphocyte-associated molecule-4 blockade: the effect is manifested only at the restricted tumor-bearing stages. *Cancer Res* 57(18):4036–4041
  16. Zhu C, Anderson AC, Schubart A et al (2005) The Tim-3 ligand galectin-9 negatively regulates T helper type 1 immunity. *Nat Immunol* 6(12):1245–1252
  17. Alvarez IB, Pasquinelli V, Jurado JO et al (2010) Role played by the programmed death-1-programmed death ligand pathway during innate immunity against Mycobacterium tuberculosis. *J Infect Dis* 202(4):524–532
  18. Beldi-Ferchiou A, Lambert M, Dogniaux S et al (2016) PD-1 mediates functional exhaustion of activated NK cells in patients with Kaposi's sarcoma. *Oncotarget* 7(45):72961–72977
  19. Gleason MK, Lenvik TR, McCullar V et al (2012) Tim-3 is an inducible human natural killer cell receptor that enhances interferon gamma production in response to galectin-9. *Blood* 119(13):3064–3072
  20. Ndhllovu LC, Lopez-Verges S, Barbour JD et al (2012) Tim-3 marks human natural killer cell maturation and suppresses cell-mediated cytotoxicity. *Blood* 119(16):3734–3743
  21. Pesce S, Greppi M, Tabellini G et al (2017) Identification of a subset of human natural killer cells expressing high levels of programmed death 1: A phenotypic and functional characterization. *J Allergy Clin Immunol* 139(1):335–346.e333
  22. Zhang Q, Bi J, Zheng X et al (2018) Blockade of the checkpoint receptor TIGIT prevents NK cell exhaustion and elicits potent anti-tumor immunity. *Nat Immunol*. <https://doi.org/10.1038/s41590-018-0132-0>
  23. Cooper MA, Fehniger TA, Caligiuri MA (2001) The biology of human natural killer-cell subsets. *Trends Immunol* 22(11):633–640
  24. Bauer S, Groh V, Wu J et al (1999) Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science* 285(5428):727–729
  25. Marcus A, Gowen BG, Thompson TW et al (2014) Recognition of tumors by the innate immune system and natural killer cells. *Adv Immunol* 122:91–128
  26. Pende D, Cantoni C, Rivera P et al (2001) Role of NKG2D in tumor cell lysis mediated by human NK cells: cooperation with natural cytotoxicity receptors and capability of recognizing tumors of nonepithelial origin. *Eur J Immunol* 31(4):1076–1086
  27. Kuylenstierna C, Bjorkstrom NK, Andersson SK et al (2011) NKG2D performs two functions in invariant NKT cells: direct TCR-independent activation of NK-like cytotoxicity and co-stimulation of activation by CD1d. *Eur J Immunol* 41(7):1913–1923
  28. Yu J, Mitsui T, Wei M et al (2011) NKp46 identifies an NKT cell subset susceptible to leukemic transformation in mouse and human. *J Clin Invest* 121(4):1456–1470
  29. Behar SM, Dascher CC, Grusby MJ, Wang CR, Brenner MB (1999) Susceptibility of mice deficient in CD1D or TAPI to infection with *Mycobacterium tuberculosis*. *J Exp Med* 189(12):1973–1980
  30. Liao CM, Zimmer MI, Wang CR (2013) The functions of type I and type II natural killer T cells in inflammatory bowel diseases. *Inflamm Bowel Dis* 19(6):1330–1338
  31. Arrenberg P, Halder R, Dai Y, Maricic I, Kumar V (2010) Oligoclonality and innate-like features in the TCR repertoire of type II NKT cells reactive to a beta-linked self-glycolipid. *Proc Natl Acad Sci U S A* 107(24):10984–10989
  32. Berzins SP, Cochrane AD, Pellicci DG, Smyth MJ, Godfrey DI (2005) Limited correlation between human thymus and blood NKT cell content revealed by an ontogeny study of paired tissue samples. *Eur J Immunol* 35(5):1399–1407
  33. Stabile H, Fionda C, Gismondi A, Santoni A (2017) Role of distinct natural killer cell subsets in anticancer response. *Front Immunol* 8:293
  34. Carrega P, Bonaccorsi I, Di Carlo E et al (2014) CD56(bright)perforin(low) noncytotoxic human NK cells are abundant in both

- healthy and neoplastic solid tissues and recirculate to secondary lymphoid organs via afferent lymph. *J Immunol* 192(8):3805–3815
- 35. Mamessier E, Pradel LC, Thibult ML et al (2013) Peripheral blood NK cells from breast cancer patients are tumor-induced composite subsets. *J Immunol* 190(5):2424–2436
  - 36. Spits H, Artis D, Colonna M et al (2013) Innate lymphoid cells—a proposal for uniform nomenclature. *Nat Rev Immunol* 13(2):145–149
  - 37. Montaldo E, Vacca P, Vitale C et al (2016) Human innate lymphoid cells. *Immunol Lett* 179:2–8
  - 38. Morita H, Moro K, Koyasu S (2016) Innate lymphoid cells in allergic and nonallergic inflammation. *J Allergy Clin Immunol* 138(5):1253–1264
  - 39. Spits H, Di Santo JP (2011) The expanding family of innate lymphoid cells: regulators and effectors of immunity and tissue remodeling. *Nat Immunol* 12(1):21–27
  - 40. Hazenberg MD, Spits H (2014) Human innate lymphoid cells. *Blood* 124(5):700–709
  - 41. Artis D, Spits H (2015) The biology of innate lymphoid cells. *Nature* 517(7534):293–301



# Chapter 15

## Quantitative Analysis of Bile Acid with UHPLC-MS/MS

**Yuan Tian, Jingwei Cai, Erik L. Allman, Philip B. Smith, and Andrew D. Patterson**

### Abstract

Bile acids are important end products of cholesterol metabolism, having been shown to serve as signaling molecules and intermediates between the host and the gut microbiota. Here we describe a robust and accurate method using ultrahigh-pressure liquid chromatography coupled with tandem mass spectrometry (UHPLC-MS/MS) for the quantification of bile acids in stool/cecal and tissue samples.

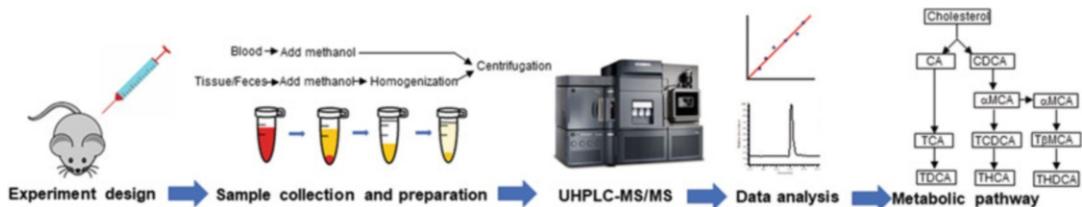
**Key words** Bile acid, UHPLC-MS/MS, Quantitation, Metabolomics

---

### 1 Introduction

Bile acids are hydroxylated steroids primarily synthesized from cholesterol in the liver. Bile acids are typically conjugated with amino acids glycine (Gly) and/or taurine (Tau) by two enzymes in the liver, bile acid CoA ligase (BAL) and bile acid CoA:amino acid *N*-acyltransferase (BAAT) [1, 2]. However, bacterial enzymes known as bile salt hydrolases (BSH) are capable of deconjugating conjugated bile acids, which are expressed by numerous commensal inhabitants of the human gastrointestinal tract including *Lactobacillus*, *Clostridium*, *Bifidobacterium*, and *Enterococcus* [3–5]. Similarly, primary bile acids, such as cholic acid (CA) and chenodeoxycholic acid (CDCA), may be deconjugated and dehydroxylated by intestinal bacteria to generate secondary bile acids, such as deoxycholic acid (DCA) and lithocholic acid (LCA) [6]. A recent study also demonstrated the conversion of DCA to isoDCA by *Ruminococcus gnavus*, which favors the growth of the abundant genus *Bacteroides* [7]. Therefore, the composition of the intestinal bile acid pool can be dramatically shaped by bacterial metabolism and vice versa [8–10].

In addition to their role in dietary lipid absorption, bile acids have gained increasing consideration as signaling molecules with potential effects on endocrine function [11–13]. Recent research



**Fig. 1** UHPLC-MS/MS workflow for bile acid quantitation

reported that bile acids can function as gut hormones capable of influencing metabolic processes via various ligand-activated nuclear receptors, such as the farnesoid X receptor (FXR), pregnane X receptor (PXR), vitamin D receptor (VDR), and cell surface G protein-coupled receptors (GPCRs), such as GPBAR-1 (TGR5) [11]. Therefore, accurate, sensitive, and reproducible analytical methods for the quantitation of primary and secondary bile acids in biological samples are desirable to better understand the physiological and metabolic function of bile acid in healthy and pathological settings.

The analysis of bile acids in biological samples is complicated, largely due to their chemical diversity, wide concentration range, and matrix effects [14]. Multiple analytical platforms including liquid chromatography coupled with mass spectrometry (LC-MS) [8, 15], gas chromatography coupled with mass spectrometry (GC-MS) [16, 17], and nuclear magnetic resonance (NMR) [17, 18] have been used for bile acid detection, identification, and quantitation. Among these techniques, LC-MS is most widely used for its high selectivity and sensitivity with selected ion monitoring (SIM) [16] and multiple reaction monitoring (MRM) modes [19]. In this chapter we describe a validated ultra high-pressure liquid chromatography coupled with tandem mass spectrometry (UHPLC-MS/MS) method based on a published protocol [20] with minor modifications for the quantification of bile acids in stool and tissue samples (Fig. 1).

## 2 Materials

All the solvents and chemicals should be high performance liquid chromatography (HPLC) grade or of the highest analytical purity.

### 2.1 Standards

Dissolve bile acid standards (Table 1) to a stock concentration of 5 mg/mL in methanol. Store standards in 1.8 mL glass tubes at -20 °C. Table 1 and Fig. 2 show the 70 bile acid standards including 36 nonconjugated, 12 Tau-conjugated, 8 Gly-conjugated, and 14 deuterated internal standards obtained from Sigma-Aldrich (St. Louis, MO), Steraloids (Newport, RI), and Medical Isotopes (Pelham, NH).

**Table 1**  
**UHPLC-MS/MS settings for the 70 bile acid standards**

No.	Bile acid	Formula	MRM or SIM (m/z)	CE (eV)	RT (min)	IS
1	Ursocholanic acid	C <sub>24</sub> H <sub>40</sub> O <sub>2</sub>	359.3	10	12.09	LCA-d4
2	Lithocholenic acid	C <sub>24</sub> H <sub>38</sub> O <sub>3</sub>	373.3	10	11.27	LCA-d4
3	5-Cholenic acid-3 $\beta$ -ol	C <sub>24</sub> H <sub>38</sub> O <sub>3</sub>	373.3	10	11.08	LCA-d4
4	3-Ketocholanic acid	C <sub>24</sub> H <sub>38</sub> O <sub>3</sub>	373.3	10	11.39	LCA-d4
5	Isolithocholic acid	C <sub>24</sub> H <sub>38</sub> O <sub>3</sub>	375.3	10	11.17	LCA-d4
6	Allolithocholic acid	C <sub>24</sub> H <sub>40</sub> O <sub>3</sub>	375.3	10	11.10	LCA-d4
7	3 $\alpha$ ,12 $\alpha$ ,23-Nordeoxycholic acid	C <sub>23</sub> H <sub>38</sub> O <sub>4</sub>	377.3	10	10.93	LCA-d4
8	9(11),(5 $\beta$ )-Cholenic acid-3 $\alpha$ -ol-12-one	C <sub>24</sub> H <sub>36</sub> O <sub>4</sub>	387.3	10	10.56	LCA-d4
9	5 $\alpha$ -Cholanic acid-3,6-dione	C <sub>24</sub> H <sub>36</sub> O <sub>4</sub>	387.3	10	10.41	LCA-d4
10	3,7-Diketocholanic acid	C <sub>24</sub> H <sub>36</sub> O <sub>4</sub>	387.3	10	10.41	LCA-d4
11	3,6-Diketocholanic acid	C <sub>24</sub> H <sub>36</sub> O <sub>4</sub>	387.3	10	10.47	LCA-d4
12	3,12-Diketocholanic acid	C <sub>24</sub> H <sub>36</sub> O <sub>4</sub>	387.3	10	10.47	LCA-d4
13	8(14),(5 $\beta$ )-Cholenic acid-3 $\alpha$ ,12 $\alpha$ -diol	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	389.3	10	10.77	LCA-d4
14	5 $\beta$ -Cholenic acid-7 $\alpha$ -ol-3-one	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	389.3	10	10.96	LCA-d4
15	5 $\alpha$ -Cholanic acid-3 $\alpha$ -ol-6-one	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	389.3	10	10.44	LCA-d4
16	3 $\alpha$ -Hydroxy-7 ketolithocholic acid	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	389.3	10	10.63	LCA-d4
17	3 $\alpha$ -Hydroxy-12 ketolithocholic acid	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	389.3	10	10.68	LCA-d4
18	Lithocholic acid (LCA)	C <sub>24</sub> H <sub>40</sub> O <sub>3</sub>	375.3	10	11.37	LCA-d4
19	5 $\beta$ -Cholanic acid-3 $\beta$ ,12 $\alpha$ -diol	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	10.69	LCA-d4
20	Chenodeoxycholic acid (CDCA)	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	11.08	CDCA-d4
21	Deoxycholic acid (DCA)	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	11.14	DCA-d4
22	Hyodeoxycholic acid (HDCA)	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	10.78	HDCA-d4
23	Isodeoxycholic acid	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	11.27	DCA-d4
24	Murocholic acid (MuroCA)	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	10.17	DCA-d4
25	Ursodeoxycholic acid (UDCA)	C <sub>24</sub> H <sub>40</sub> O <sub>4</sub>	391.3	10	10.55	UDCA-d4
26	3,7,12-Dehydrocholic acid	C <sub>24</sub> H <sub>34</sub> O <sub>5</sub>	401.2	10	6.25	UDCA-d4

(continued)

**Table 1**  
**(continued)**

No.	Bile acid	Formula	MRM or SIM (m/z)	CE (eV)	RT (min)	IS
27	3 $\alpha$ -Hydroxy-7,12-diketocholanic acid	C <sub>24</sub> H <sub>36</sub> O <sub>5</sub>	403.2	10	6.96	UDCA-d4
28	3 $\alpha$ -Hydroxy-6,7-diketocholanic acid	C <sub>24</sub> H <sub>36</sub> O <sub>5</sub>	403.2	10	10.58	UDCA-d4
29	5 $\beta$ -Cholanic acid-3 $\alpha$ ,6 $\alpha$ -diol-7-one	C <sub>24</sub> H <sub>38</sub> O <sub>5</sub>	405.3	10	9.42	CA-d4
30	3-Dehydrocholic acid	C <sub>24</sub> H <sub>38</sub> O <sub>5</sub>	405.3	10	10.14	CA-d4
31	12-Dehydrocholic acid	C <sub>24</sub> H <sub>38</sub> O <sub>5</sub>	405.3	10	9.52	CA-d4
32	$\alpha$ -Muricholic acid	C <sub>24</sub> H <sub>40</sub> O <sub>5</sub>	407.3	10	9.52	CA-d4
33	$\beta$ -Muricholic acid	C <sub>24</sub> H <sub>40</sub> O <sub>5</sub>	407.3	10	9.65	CA-d4
34	$\omega$ -Muricholic acid	C <sub>24</sub> H <sub>40</sub> O <sub>5</sub>	407.3	10	9.43	CA-d4
35	Cholic acid (CA)	C <sub>24</sub> H <sub>40</sub> O <sub>5</sub>	407.3	10	10.65	CA-d4
36	Hyocholic acid (HCA)	C <sub>24</sub> H <sub>40</sub> O <sub>5</sub>	407.3	10	10.44	CA-d4
37	Glycoursocholanic acid	C <sub>26</sub> H <sub>43</sub> NO <sub>3</sub>	416.3 → 74	40	11.39	GLCA-d4
38	Glycolithocholic acid (GLCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>4</sub>	432.3 → 74	40	10.75	GLCA-d4
39	Glycoursodeoxycholic acid (GUDCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>5</sub>	448.3 → 74	40	7.37	GUDCA-d4
40	Glycohyodeoxycholic acid (GHDCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>5</sub>	448.3 → 74	40	7.76	GUDCA-d4
41	Glycochenodeoxycholic acid (GCDCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>5</sub>	448.3 → 74	40	9.87	GCDCA-d4
42	Glycodeoxycholic acid (GDCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>5</sub>	448.3 → 74	40	10.23	GDCA-d4
43	3,7,12-Glycodehydrocholic acid	C <sub>26</sub> H <sub>37</sub> NO <sub>6</sub>	458.3 → 74	40	3.16	GDCA-d4
44	Glycocholic acid (GCA)	C <sub>26</sub> H <sub>43</sub> NO <sub>6</sub>	464.3 → 74	40	6.99	GCA-d4
45	Taurolithocholic acid (TLCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>5</sub> S	466.3 → 80	40	11.16	TLCA-d4
46	Tauroursocholanic acid	C <sub>26</sub> H <sub>45</sub> NO <sub>4</sub> S	482.3 → 80	60	10.38	TLCA-d4
47	Tauroursodeoxycholic acid (TUDCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>6</sub> S	498.3 → 80	60	6.35	TUDCA-d5
48	Taurohyodeoxycholic acid	C <sub>26</sub> H <sub>45</sub> NO <sub>6</sub> S	498.3 → 80	60	6.69	TUDCA-d5
49	Taurochenodeoxycholic acid (TCDCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>6</sub> S	498.3 → 80	60	8.74	TUDCA-d5
50	Taurodeoxycholic acid (TDCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>6</sub> S	498.3 → 80	60	9.18	TUDCA-d5

(continued)

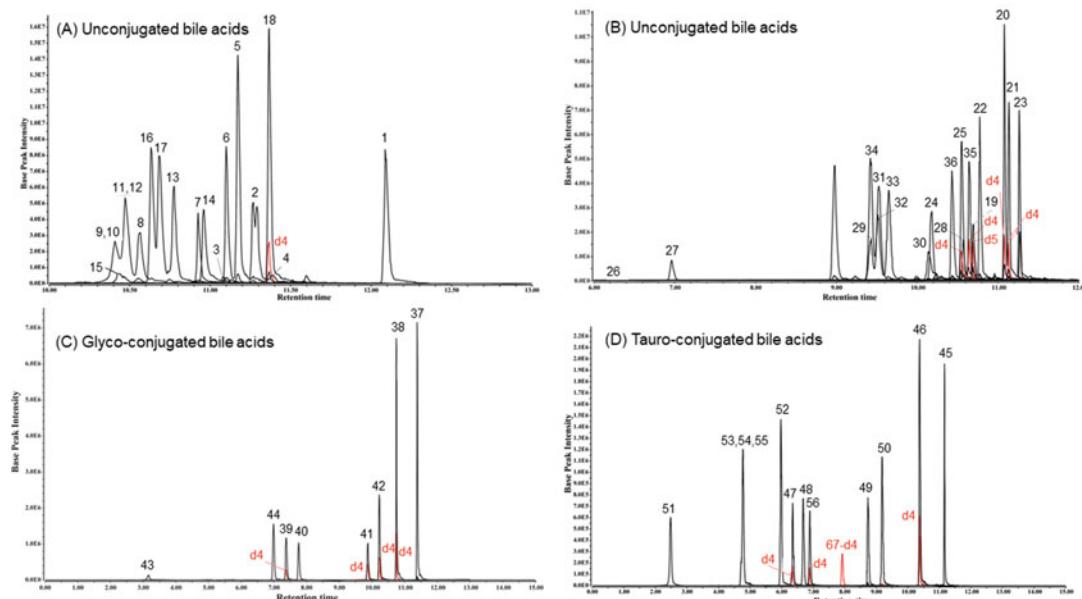
**Table 1**  
(continued)

No.	Bile acid	Formula	MRM or SIM (m/z)	CE (eV)	RT (min)	IS
51	3,7,12-Taurodehydrocholic acid	C <sub>26</sub> H <sub>39</sub> NO <sub>7</sub> S	508.2 → 80	60	2.48	TUDCA-d5
52	Taurohyocholic acid (THCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>7</sub> S	514.3 → 80	60	5.98	TCA-d4
53	Tauro- $\beta$ muricholic acid (T $\beta$ MCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>7</sub> S	514.3 → 80	60	4.78	TCA-d4
54	Tauro- $\alpha$ muricholic acid (TaMCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>7</sub> S	514.3 → 80	60	4.78	TCA-d4
55	Tauro- $\omega$ muricholic acid (ToMCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>7</sub> S	514.3 → 80	60	4.76	TCA-d4
56	Taurocholic acid (TCA)	C <sub>26</sub> H <sub>45</sub> NO <sub>7</sub> S	514.3 → 80	60	6.90	TCA-d4
57	Lithocholic acid-d4	C <sub>24</sub> H <sub>36</sub> D <sub>4</sub> O <sub>3</sub>	379.6	10	11.37	
58	Deoxycholic acid-d4	C <sub>24</sub> H <sub>36</sub> D <sub>4</sub> O <sub>4</sub>	395.6	10	11.14	
59	Chenodeoxycholic acid-d4	C <sub>24</sub> H <sub>36</sub> D <sub>4</sub> O <sub>4</sub>	395.6	10	11.08	
60	Ursodeoxycholic acid-d4	C <sub>24</sub> H <sub>36</sub> D <sub>4</sub> O <sub>4</sub>	395.6	10	10.55	
61	Hyodeoxycholic acid-d5	C <sub>24</sub> H <sub>36</sub> D <sub>5</sub> O <sub>4</sub>	396.6	10	10.69	
62	Cholic acid-d4	C <sub>24</sub> H <sub>36</sub> D <sub>4</sub> O <sub>5</sub>	411.6	10	10.64	
63	Glycolithocholic acid-d4	C <sub>26</sub> H <sub>39</sub> D <sub>4</sub> NO <sub>4</sub>	436.6 → 74	40	10.75	
64	Glycochenodeoxycholic acid-d4	C <sub>26</sub> H <sub>39</sub> D <sub>4</sub> NO <sub>5</sub>	452.5 → 74	40	9.87	
65	Glycodeoxycholic acid-d4	C <sub>26</sub> H <sub>39</sub> D <sub>4</sub> NO <sub>5</sub>	452.6 → 74	40	10.22	
66	Glycoursodeoxycholic acid-d4	C <sub>26</sub> H <sub>39</sub> D <sub>4</sub> NO <sub>5</sub>	452.6 → 74	40	7.36	
67	Glycocholic acid-d4	C <sub>26</sub> H <sub>39</sub> D <sub>4</sub> NO <sub>6</sub>	468.6 → 74	40	7.93	
68	Taurolithocholic acid-d4	C <sub>26</sub> H <sub>41</sub> D <sub>4</sub> NO <sub>5</sub> S	486.6 → 80	60	10.37	
69	Tauroursodeoxycholic acid-d5	C <sub>26</sub> H <sub>40</sub> D <sub>5</sub> NO <sub>6</sub> S	503.7 → 80	60	6.79	
70	Taurocholic acid-d4	C <sub>26</sub> H <sub>41</sub> D <sub>4</sub> NO <sub>7</sub> S	518.7 → 80	60	6.58	

SIM selected ion monitoring, MRM multiple reaction monitoring transitions, CE collision energy, RT retention time, IS internal standard

## 2.2 Reagents

1. Mobile phase A (*see Note 1*): 100 mL of acetonitrile added to 1 L of ultrapure water, containing 1 mM ammonium acetate adjusted to pH = 4.15 with acetic acid.
2. Mobile phase B (*see Note 1*): a mixture of acetonitrile and 2-propanol (1:1).
3. Wash solvent: 2-propanol.
4. Purge solvent: a mixture of water and 2-propanol (9:1).



**Fig. 2** Extracted ion chromatograms of 70 bile acid species: unconjugated (**a**, **b**), glyco-conjugated (**c**), and tauro-conjugated (**d**). Deuterated internal standards are denoted in red. For more details, see Table 1

5. Solution of internal standards for calibration curve: Prepare 56 bile acid standards at concentrations of 0, 0.16, 0.3, 0.63, 1.25, 2.5, 5, 10  $\mu$ M containing deuterated internal standards (16 total at 0.5  $\mu$ M each) (*see Notes 2 and 3*). Store standards in 1.8 mL glass tubes at  $-20^{\circ}\text{C}$ .

## **2.3 UHPLC-MS/MS**

1. UHPLC: ACQUITY UPLC system with an ACQUITY BEH C8 (2.1 × 100 mm, 1.7 µm) UPLC column (Waters, Milford, MA).
  2. MS/MS: Waters Xevo TQ-S Triple Quadrupole mass spectrometer (Waters, Milford, MA).

### 3 Methods

### **3.1 Sample Collection (See Note 4)**

1. Tissue (Liver and Intestine), Cecal Content, Feces/Stool.
    - (a) Collect the tissue sample according to approved protocols.
      - Wash liver and intestine tissue in a petri dish with ice-cold 0.1 M PBS.
    - (b) Transfer sample to a screw cap tube, freeze in liquid nitrogen, and store at -80 °C for analysis.
  2. Serum.
    - (a) Collect whole blood from the portal vein, cardiac puncture, or orbital sinus according to a standard protocol [21], in a BD Microtainer® blood collection tube.

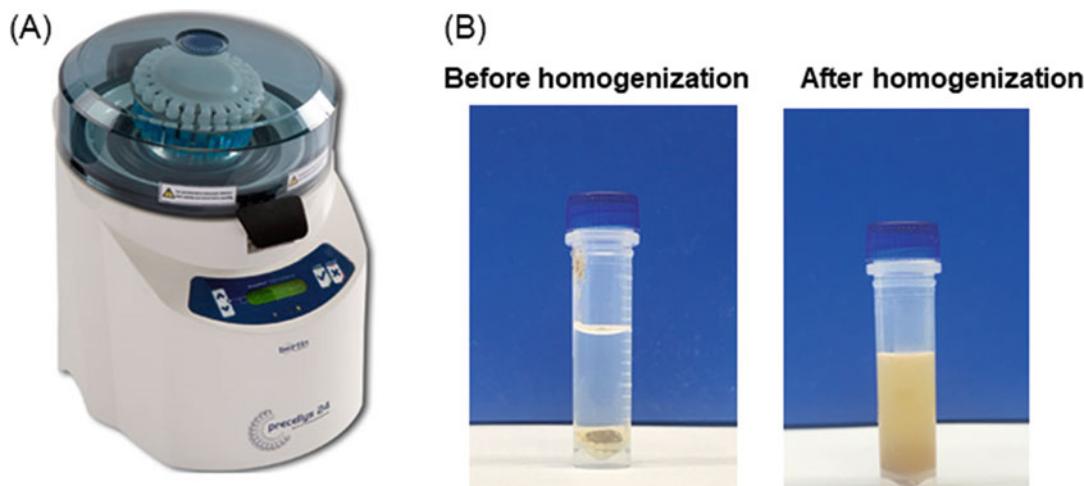
- (b) Incubate samples on ice for 20 min.
- (c) Centrifuge samples at 4 °C, 5 min,  $5000 \times g$ .
- (d) Transfer supernatant to a screw cap tube, freeze in liquid nitrogen, and store at –80 °C for analysis.

3. Plasma.

- (a) Collect whole blood from the portal vein, cardiac puncture, or orbital sinus according to a standard protocol [21], in a BD Microtainer® blood collection tube containing ethylenediaminetetraacetic acid (EDTA) (1.5 mg per 1 mL blood).
- (b) Incubate samples on ice for 20 min.
- (c) Centrifuge samples at 4 °C, 5 min,  $5000 \times g$ .
- (d) Transfer supernatant to a screw cap tube, freeze in liquid nitrogen, and store at –80 °C for analysis.

**3.2 Sample Preparation (See Note 5)**

1. Tissue (Liver and Intestine), Cecal Content, Feces/Stool.
  - (a) Thaw samples on ice.
  - (b) Weigh 50 mg tissue or 25 mg feces/stool/cecal content in screw cap tubes (*see Note 6*) and add ~50 µL Zirconia/Silica beads.
    - Use 0.1 mm beads for cecal/feces/stool and 1 mm beads for tissue samples.
  - (c) Add 1 mL of ice-cold methanol containing 0.5 µM deuterated internal standards.
  - (d) Homogenize the samples thoroughly with the Precellys 24 tissue homogenizer (Fig. 3a), adjust homogenization settings as necessary (*see Notes 7 and 8*). An ideal setting is 20 s × 2 at 6500 rpm.



**Fig. 3** (a) Precellys 24 tissue homogenizer and (b) mouse feces sample before and after homogenization

- Additionally, freeze-thaw cecal/feces/stool samples three times (this step is not needed for tissue samples).

- Centrifuge samples at 4 °C, 15 min, 14,000 × *g*.
- Transfer 200 µL supernatant to a 11 mm polypropylene snap closure autosampler vial.

2. Serum/Plasma.

- Thaw samples on ice.
- Centrifuge samples at 4 °C, 15 min, 14,000 × *g*.
- Transfer 50 µL of supernatant to a new screw cap tube.
- Add 150 µL of ice-cold methanol (*see Note 5*) containing 0.5 µM deuterated internal standards.
- Vortex for 10 s and incubate samples at –20 °C, 20 min.
- Centrifuge samples at 4 °C, 15 min, 14,000 × *g*.
- Transfer 150 µL supernatant to a low volume 11 mm polypropylene snap closure autosampler vial.

**3.3 UHPLC-MS/MS  
(See Notes 9–11)**

1. Inlet method: the gradient separation is described in Table 2. LC parameters were as follows: run time of 15 min, preinject wash time of 3.0 s, postinject wash time of 6.0 s, column temperature of 60 °C, and injection volume of 1 µL (*see Note 12*).

**Table 2**  
**Chromatographic gradient separation table for bile acid profiling and targeted analysis**

Time (min)	Flow rate (mL/min)	% A	% B	Curve
0	0.300	80	20	6
0.10	0.300	80	20	6
9.25	0.300	65	35	6
11.50	0.325	15	85	6
11.80	0.400	0	100	6
12.00	0.475	0	100	6
12.10	0.500	0	100	6
12.40	0.500	0	100	6
12.45	0.425	45	55	6
12.50	0.425	90	10	6
12.60	0.400	90	10	6
12.70	0.350	90	10	6
12.80	0.300	90	10	6
15.00	0.300	90	10	6

2. Mass spectrometry parameters were as follows: electrospray ionization, negative ion mode, capillary voltage of 1.5 kV, cone voltage of 60 V, source temperature of 150 °C, desolvation temperature of 600 °C, desolvation gas flow of 1000 L/h, and cone gas flow of 150 L/h.

### 3.4 Data Analysis

After injection of sample, peak integration was performed with TargetLysnx 4.1 (Waters, Milford, MA). Peaks were smoothed using Smoothing Iterations = 1 and Smoothing Width = 2. The data were exported to Excel to calculate the analyte/internal standard peak area ratios. Sample concentration is calculated from the calibration curve for each bile acid.

---

## 4 Notes

1. Sonicate the mobile phase for 10 min to remove dissolved gases.
2. The concentration range covered by the standards should cover the expected concentration range in the experimental samples.
3. The standard samples should be run from the lowest concentration to the highest to minimize any carry-over.
4. The rapid quenching of samples, especially tissues, should be considered before harvesting.
5. All preparation steps should be performed on ice or with cold solvents.
6. Use screw cap tubes for preparation and homogenization to prevent loss of sample and contamination.
7. After homogenization, check that the samples are thoroughly homogenized (Fig. 3b). If necessary, perform an additional cycle until all samples have been homogenized.
8. Incubate samples on ice for several minutes between each of the two homogenization steps to prevent overheating of the sample.
9. Method blanks should be prepared to check for contamination and interferences. Follow all preparation steps for blanks along with samples. The blank samples should be run (a) at the beginning of each UHPLC session, (b) between all the samples, (c) and after high concentration standards.
10. Prepare a quality control (QC) sample that is a pool of all the samples, in equal ratios, for monitoring method performance.
11. Each analytical run should include a set of standards for standard curve, blank, and QC samples.
12. Typically, 1–10 µL is injected for UHPLC-MS analysis (will vary based on metabolites of interest).

## References

1. Falany CN, Xie XW, Wheeler JB et al (2002) Molecular cloning and expression of rat liver bile acid CoA ligase. *J Lipid Res* 43(12):2062–2071
2. Falany CN, Johnson MR, Barnes S, Diasio RB (1994) Glycine and taurine conjugation of bile-acids by a single enzyme - molecular-cloning and expression of human liver bile-acid CoA-amino acid N-acyltransferase. *J Biol Chem* 269(30):19375–19379
3. Hofmann AF, Hagey LR (2008) Bile acids: chemistry, pathochemistry, biology, pathobiology, and therapeutics. *Cell Mol Life Sci* 65(16):2461–2483
4. Urdaneta V, Casadesus J (2017) Interactions between bacteria and bile salts in the gastrointestinal and hepatobiliary tracts. *Front Med* 4:163
5. Begley M, Hill C, Gahan CGM (2006) Bile salt hydrolase activity in probiotics. *Appl Environ Microbiol* 72(3):1729–1738
6. Begley M, Gahan CGM, Hill C (2005) The interaction between bacteria and bile. *FEMS Microbiol Rev* 29(4):625–651
7. Devlin AS, Fischbach MA (2015) A biosynthetic pathway for a prominent class of microbiota-derived bile acids. *Nat Chem Biol* 11(9):685
8. Li F, Jiang CT, Krausz KW et al (2013) Microbiome remodelling leads to inhibition of intestinal farnesoid X receptor signalling and decreased obesity. *Nat Commun* 4:2384
9. Jiang CT, Xie C, Lv Y et al (2015) Intestine-selective farnesoid X receptor inhibition improves obesity-related metabolic dysfunction. *Nat Commun* 6:1016
10. Ridlon JM, Kang DJ, Hylemon PB, Bajaj JS (2014) Bile acids and the gut microbiome. *Curr Opin Gastroenterol* 30(3):332–338
11. Li TG, Chiang JYL (2014) Bile acid signaling in metabolic disease and drug therapy. *Pharmacol Rev* 66(4):948–983
12. Chiang JYL (2013) Bile acid metabolism and signaling. *Compr Physiol* 3(3):1191–1212
13. Mertens KL, Kalsbeek A, Soeters MR, Eggink HM (2017) Bile acid signaling pathways from the enterohepatic circulation to the central nervous system. *Front Neurosci* 11:617
14. Dutta M, Cai J, Gui W, Patterson AD (2019) A review of analytical platforms for accurate bile acid measurement. *Anal Bioanal Chem* 411(19):4541–4549
15. Tian Y, Cai JW, Gui W et al (2019) Berberine directly affects the gut microbiota to promote intestinal farnesoid X receptor activation. *Drug Metab Dispos* 47(2):86–93
16. Perwaiz S, Tuchweber B, Mignault D, Gilat T, Yousef IM (2001) Determination of bile acids in biological fluids by liquid chromatography-electrospray tandem mass spectrometry. *J Lipid Res* 42(1):114–119
17. Kakiyama G, Muto A, Takei H et al (2014) A simple and accurate HPLC method for fecal bile acid profile in healthy and cirrhotic subjects: validation by GC-MS and LC-MS. *J Lipid Res* 55(5):978–990
18. Tian Y, Zhang LM, Wang YL, Tang HR (2012) Age-related topographical metabolic signatures for the rat gastrointestinal contents. *J Proteome Res* 11(2):1397–1411
19. Han J, Liu Y, Wang RX, Yang JC, Ling V, Borchers CH (2015) Metabolic profiling of bile acids in human and mouse blood by LC-MS/MS in combination with phospholipid-depletion solid-phase extraction. *Anal Chem* 87(2):1127–1136
20. Sarafian MH, Lewis MR, Pechlivanis A et al (2015) Bile acid profiling and quantification in biofluids using ultra-performance liquid chromatography tandem mass spectrometry. *Anal Chem* 87(19):9662–9670
21. Parasuraman S, Raveendran R, Kesavan R (2010) Blood sample collection in small laboratory animals. *J Pharmacol Pharmacother* 1(2):87–93



# Chapter 16

## Sample Preparation and Data Analysis for NMR-Based Metabolomics

Tapas K. Mal, Yuan Tian, and Andrew D. Patterson

### Abstract

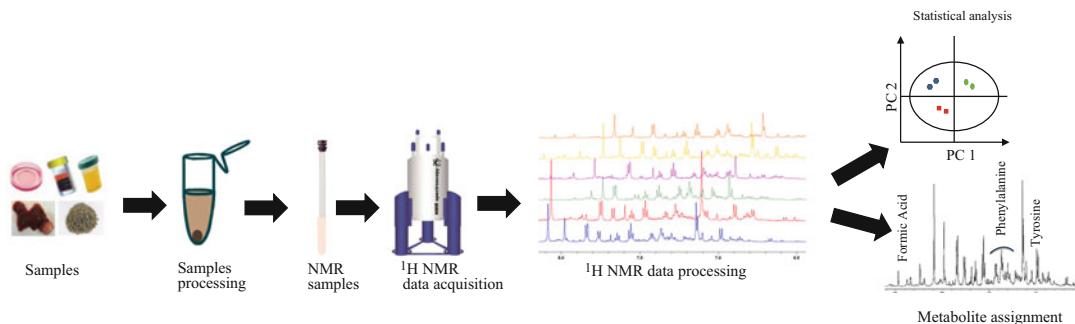
NMR spectroscopy has become one of the preferred analytical techniques for metabolomics studies due to its inherent nondestructive nature, ability to identify and quantify metabolites simultaneously in a complex mixture, minimal sample preparation requirement, and high degree of experimental reproducibility. NMR-based metabolomics studies involve the measurement and multivariate statistical analysis of metabolites present in biological samples such as biofluids, stool/feces, intestinal content, tissue, and cell extracts by high-resolution NMR spectroscopy—the goal then is to identify and quantify metabolites and evaluate changes of metabolite concentrations in response to some perturbation. Here we describe methodologies for NMR sample preparation of biofluids (serum, saliva, and urine) and stool/feces, intestinal content, and tissues for NMR experiments including extraction of polar metabolites and application of NMR in metabolomics studies. One dimensional (1D)  $^1\text{H}$  NMR experiments with different variations such as pre-saturation, relaxation-edited, and diffusion-edited are routinely acquired for profiling and metabolite identification and quantification. 2D homonuclear  $^1\text{H}$ - $^1\text{H}$  TOCSY and COSY, 2D J-resolved, and heteronuclear  $^1\text{H}$ - $^{13}\text{C}$  HSQC and HMBC are also performed to assist with metabolite identification and quantification. The NMR data are then subjected to targeted and/or untargeted multivariate statistical analysis for biomarker discovery, clinical diagnosis, toxicological studies, molecular phenotyping, and functional genomics.

**Key words** Metabolomics, NMR, TOCSY, COSY, HSQC, HMBC, HRMAS, CPMG

---

### 1 Introduction

NMR spectroscopy has made a tremendous impact in the metabolomics field [1, 2]. NMR is a quantitative, noninvasive, and nondestructive technique that provides information on molecular structures of metabolites in solution and semisolid state [3]. It is a reliable and robust technique for metabolomics application where reproducibility is imperative [4, 5]. NMR allows for the detection of a wide range of metabolites with variable concentrations simultaneously in a sample and provides a snapshot of metabolites at any given time point. With the advancement of NMR technology, metabolites with as low as 1  $\mu\text{M}$  concentration can be detected



**Fig. 1** A schematic workflow for NMR-based metabolomics

with high confidence in a few minutes by employing a high field NMR instrument equipped with a cryogenic probe.

Metabolomics studies are usually performed on biofluids and cell or tissue extracts by solution NMR. Biofluids are relatively easy to obtain, particularly saliva, urine, and serum [3, 6–8]. A wide range of fluids including seminal fluids, amniotic fluid, cerebrospinal fluid, synovial fluid, blister and cyst fluids, bronchoalveolar lavage and dialysis fluids have been studied [5, 8, 9]. In addition, tissue, stool/feces, and cell extracts are also studied [2, 10–13]. Solid-state and high-resolution magic angle spinning (HRMAS) NMR have been employed for metabolomics studies on intact tissues and organs [14–18]. Metabolite imaging and metabolic analysis are also been performed through magnetic resonance spectroscopy (MRS) and magnetic resonance imaging (MRI) [19–21]. These studies allow for measuring changes of metabolites in a biological sample or a given system in response to a challenge in real-time (physical, chemical, environmental, or other external stressors) to normal cellular homeostasis. Therefore, NMR-based metabolomics has tremendous promise in the application of human health. Metabolomics has been applied for disease diagnosis [15, 22–27], to understand the effect of nutrition on health and relation to gut microbiota [8, 28], epidemiological studies [4], and biomarker discovery [6, 29–35]. In this chapter, we describe sample preparation that is a key step for producing reliable results independent of operators and analytical instruments, NMR data acquisition, and analysis (Fig. 1).

## 2 Materials

All solvents and chemicals must be HPLC or high purity analytical grade reagents. Reagents used such as methanol (VWR), anhydrous sodium/potassium phosphate dibasic and monobasic molecular biology grade (Sigma), sodium 3-(trimethylsilyl) [2,2,3,3-<sup>2</sup>H<sub>4</sub>] propionate (TSP) from Sigma and D<sub>2</sub>O from Cambridge Isotope Laboratories.

<b>2.1 NMR Stock Buffers</b>	0.1 M $\text{Na}^+/\text{K}^+$ $\text{PO}_4$ Buffer (100 ml, pH = 7.4). $\text{K}_2\text{HPO}_4$ : 1.408 g.
<b>2.1.1 Cell Extract (100 ml, 0.1 M)</b>	$\text{NaH}_2\text{PO}_4$ : 0.242 g. $\text{H}_2\text{O}$ : 50 ml. $\text{D}_2\text{O}$ : 50 ml. TSP: 0.005 g.
<b>2.1.2 Cell Media (25 ml, 0.4 M)</b>	0.4 M $\text{Na}^+/\text{K}^+$ $\text{PO}_4$ Buffer (25 ml, pH = 7.4). $\text{K}_2\text{HPO}_4$ : 1.394 g. $\text{NaH}_2\text{PO}_4$ : 0.240 g. $\text{D}_2\text{O}$ : 25 ml. TSP: 0.00125 g.
<b>2.1.3 Serum/Plasma (100 ml, 0.045 M)</b>	0.045 M $\text{Na}^+/\text{K}^+$ $\text{PO}_4$ Buffer (100 ml, pH = 7.4). $\text{K}_2\text{HPO}_4$ : 0.633 g. $\text{NaH}_2\text{PO}_4$ : 0.110 g. 0.9% saline: 0.9 g in 50 ml water (+ 50 ml $\text{D}_2\text{O}$ ). (no TSP, which will associate with proteins in serum or plasma).
<b>2.1.4 Urine (100 ml, 1.5 M)</b>	1.5 M $\text{Na}^+/\text{K}^+$ $\text{PO}_4$ Buffer (100 ml, pH = 7.4). $\text{K}_2\text{HPO}_4$ : 21.113 g. $\text{NaH}_2\text{PO}_4$ : 3.636 g. $\text{D}_2\text{O}$ : 100 ml. TSP: 0.005 g. $\text{NaN}_3$ : 0.01 g (preservative).
KF Stock Solution (25 ml, 5 M)	$\text{KF}\cdot 2\text{H}_2\text{O}$ : 11.885 g. $\text{H}_2\text{O}$ : 25 ml.
EDTA-d12 Stock Solution (25 ml, 120 mM)	EDTA-d12: 0.913 g. $\text{KOH}$ : 0.593 g. $\text{H}_2\text{O}$ : 25 ml.
<b>2.1.5 Tissue Extract, Cecum Content, and Feces (100 ml, 0.1 M)</b>	0.1 M $\text{Na}^+/\text{K}^+$ $\text{PO}_4$ Buffer (100 ml, pH = 7.4). $\text{K}_2\text{HPO}_4$ : 1.408 g. $\text{NaH}_2\text{PO}_4$ : 0.242 g. $\text{H}_2\text{O}$ : 50 ml. $\text{D}_2\text{O}$ : 50 ml. TSP: 0.005 g. $\text{NaN}_3$ : 0.01 g (preservative).

*2.1.6 Saliva (100 ml, 0.1 M)* 0.1 M Na<sup>+</sup>/K<sup>+</sup> PO<sub>4</sub> Buffer (100 ml, pH = 7.4):  
 K<sub>2</sub>HPO<sub>4</sub>: 1.408 g.  
 NaH<sub>2</sub>PO<sub>4</sub>: 0.242 g.  
 D<sub>2</sub>O: 100 ml.

(no TSP, which will associate with proteins present in saliva)

## 2.2 NMR Sample Preparation

While there are a number of methods available for the extraction of metabolites from biofluids, tissues, and cells, we follow the following for metabolomics sample preparation. We found these methods work well for us in terms of reproducibility and reliability of results independent of operators and instruments. In addition, these methods require minimal processing and efforts to prepare NMR samples.

### 2.2.1 Cell Extraction

1. Collect cell samples from 150 mm dishes (about  $2 \times 10^7$  per dish).
2. Add 1 ml of precooled methanol–H<sub>2</sub>O (2:1).
3. Homogenize the samples thoroughly with a Precellys 24 tissue homogenizer and an ideal setting is 20 s × 2 at 6500 rpm.
4. Freeze-thaw three times with liquid nitrogen and 37 °C water bath.
5. Centrifuge at  $3200 \times g$  for 10 min, 4 °C.
6. Transfer the supernatants to 2 ml new microcentrifuge tubes.
7. Add 1 ml solution of methanol–H<sub>2</sub>O (2:1) to the pellets and repeat the above procedure (**items 4 and 5**).
8. Combine the supernatants.
9. Dry down and then store them at –80 °C until further use.
10. Resuspend in 0.6 ml 0.1 M PBS, centrifuge at 4 °C, 12,000 ×  $g$  for 10 min, then transfer 0.55 ml supernatants to 5 mm NMR tubes.

### 2.2.2 Cell Media

1. Add 6 ml methanol to 3 ml medium, keep in ice for 30 min.
2. Centrifuge at  $1600 \times g$  for 10 min, 4 °C.
3. Transfer 3 ml to new 5 ml microcentrifuge tubes.
4. Dry down.
5. Add 0.5 ml H<sub>2</sub>O, vortex samples for 10 s.
6. Transfer 0.48 ml to new 1.5 ml EP tubes.
7. Add 0.12 ml PBS, vortex.
8. Centrifuge at 12,000 ×  $g$  for 10 min, 4 °C.
9. Transfer 0.55 ml supernatants to 5 mm NMR tubes.

**2.2.3 Serum/Plasma**

**When the Volume of Serum/Plasma Enough**

1. 200  $\mu$ l samples mixed with 400  $\mu$ l PBS (0.045 M).
2. Vortex samples for 10 s.
3. Centrifuge at 18,100  $\times \mathcal{g}$  for 10 min, 4 °C.
4. Transfer 550  $\mu$ l supernatants into 5 mm NMR tubes.

**When the Volume Serum/Plasma, Not Enough ( $\leq 50 \mu$ l)**

1. 30  $\mu$ l samples mixed 30  $\mu$ l PBS (0.045 M).
2. Vortex samples for 10 s.
3. Centrifuge at 18,100  $\times \mathcal{g}$  for 10 min, 4 °C.
4. Transfer 60  $\mu$ l supernatants into 1.7 mm NMR tubes with microsyringe.

**2.2.4 Urine**

1. 500  $\mu$ l urine mixed with 14  $\mu$ l KF (5 M).
2. Vortex samples for 10 s.
3. Centrifuge at 12,000  $\times \mathcal{g}$  for 10 min, 4 °C.
4. Add 8.3  $\mu$ l EDTA-d12 (0.12 M) into 5 mm NMR tubes.
5. Transfer 450  $\mu$ l supernatants into NMR tubes and mix.
6. Add 45  $\mu$ l PBS (1.5 M) into NMR tubes and mix.

**2.2.5 Tissue Extract**

1. Weigh ~50 mg tissues in homogenization tubes and make a record, mark and save on ice.
2. Add 8–10 Zirconia/Silica beads and 1 ml of precooled methanol–H<sub>2</sub>O (2:1) to homogenization tubes and vortex for 10 s.
3. Homogenize the samples thoroughly with the Precellys 24 tissue homogenizer and an ideal setting is 20 s  $\times$  2 at 6500 rpm.
4. Centrifuge at 4 °C, 12,000  $\times \mathcal{g}$  for 10 min.
5. Transfer the supernatants to 2 ml microcentrifuge tubes.
6. Add 0.6 ml solution of methanol–H<sub>2</sub>O (2:1) to the pellets, repeat the above procedure (**items 4 and 5**).
7. Combine the supernatants.
8. Dry down and store them at –80 °C until further use.
9. Resuspend in 0.6 ml 0.1 M PBS, centrifuge at 4 °C, 12,000  $\times \mathcal{g}$  for 10 min, then transfer 0.55 ml supernatants to 5 mm NMR tubes.

**2.2.6 Cecum Content and Feces Extract**

1. Weigh 50–60 mg samples in homogenization tubes and make a record, mark and save it on ice.
2. Add 8–10 Zirconia/Silica beads and 1.0 ml PBS (0.1 M) solution containing 50% D<sub>2</sub>O to homogenization tubes and vortex for 30 s.
3. Homogenize the samples thoroughly with the Precellys 24 tissue homogenizer and an ideal setting is 20 s  $\times$  2 at 6500 rpm.
4. Freeze-thaw two times with liquid nitrogen.

5. Centrifuge at 4 °C, 12,000 ×  $\text{g}$  for 10 min.
6. Transfer the supernatants to 2 ml new microcentrifuge tubes.
7. Add 0.6 ml PBS solution to the pellets followed by vortexing for 30 s and centrifuging at 4 °C, 12,000 ×  $\text{g}$  for 10 min.
8. Combine the supernatants, centrifuge at 4 °C and 16,000 ×  $\text{g}$  for 10 min and transfer the supernatants (0.55 ml) to 5 mm NMR tubes.

#### 2.2.7 Saliva

1. 200  $\mu\text{l}$  samples mixed with 400  $\mu\text{l}$   $\text{Na}^+/\text{K}^+$   $\text{PO}_4$  Buffer (0.1 M).
2. Vortex samples for 10 s.
3. Centrifuge at 18,100 ×  $\text{g}$  for 15 min, 4 °C.
4. Transfer 550  $\mu\text{l}$  supernatants into 5 mm NMR tubes.

### 2.3 NMR Data Acquisition and Processing

1D  $^1\text{H}$  NMR experiments are performed for metabolomics studies because metabolites are  $^1\text{H}$  rich and  $^1\text{H}$  NMR experiments are automatable, reliable, fast, and easily performed. The most common experiments performed for metabolomics studies on the Bruker (Bruker, Billerica, MA) AVANCE spectrometer are 1D nuclear Overhauser enhancement spectroscopy (NOESY)-presat (noesypr1d) [36–38], Carr-Purcell-Maiboom-Gill (CPMG)-presat (cpmgpr1d) [39–41], J-resolved [42], and diffusion edited [43] experiments. In our facility, we perform the noesypr1d pulse sequence for samples such as cell extracts and culture media, urine, and tissue, stool/feces, and cecum extract samples, and cpmgpr1d for serum, saliva, and plasma samples on a Bruker AVANCE NEO 600 MHz instrument equipped with a temperature-controlled SampleJet™ that can hold over 450 samples. The following steps are carefully performed to set up NMR experiments:

1. Set sample temperature to 298 K.
2. Insert NMR sample into the magnet (wait for a few minutes to allow the sample to equilibrate).
3. Lock, tune and match, shimming (automated).
4. 90° pulse width measured using Bruker “pulsecal” command in topspin.
5. Determine the offset of water signal for water suppression.

Note: normally water presaturation is effective; however, the power level must be carefully calibrated for optimal water suppression without suppressing metabolites resonance close to it.

#### 6. NMR experiments:

- (a) noesypr1d experiment is run with 16 ppm spectral width, 100 ms mixing time, 64k time domain data, 10 s relaxation delay (to make sure acquiring fully relaxed data),

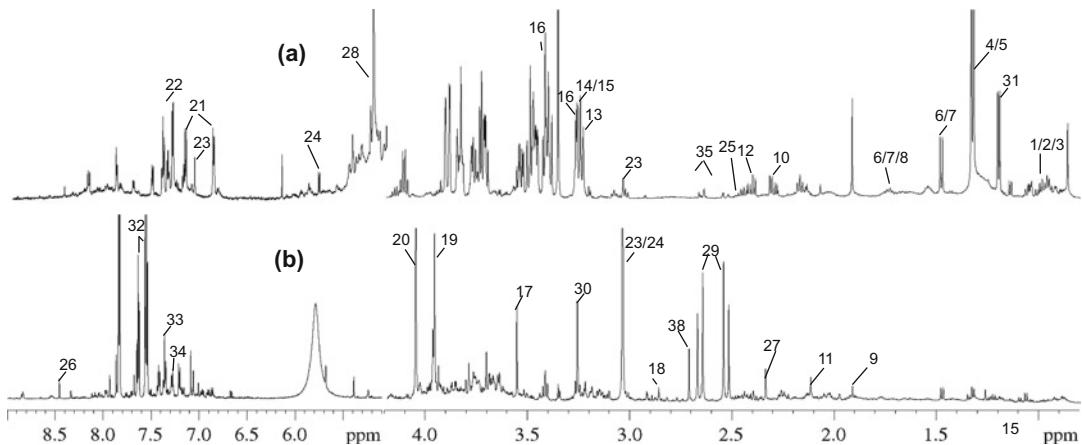
64 number of scans. The total acquisition time with these parameters is ~15 min/sample.

- (b) noesygppr1d experiment is very similar to noesypr1d except it uses gradients to improve water suppression.
  - (c) cpmgpr1d experiment is performed for serum and plasma samples to attenuate the NMR signals of larger molecules, such as proteins, lipids. It is run with a total echo time of 70 ms.
7. The NMR data are processed with 1.0 Hz exponential multiplication line broadening, zero-fill to double the number of Fourier domain points to 128k before Fourier transformation.
  8. Fourier transformed data are carefully phased before baseline correction is performed.
  9. The spectral data are calibrated against the TSP internal standard ( $\alpha$ -glucose for serum/plasma/saliva).

While we use Bruker TopSpin (<https://www.bruker.com/products/mr/nmr/software.html>, free for academic use) and Mnova (Mestrelab Research, <https://mestrelab.com/software/mnova/>) for NMR data processing there are freeware available for processing including ACD/NMR Processor Academic Edition (<https://www.metabolomicsworkbench.org/tools/externaltools.php>), SpinWorks (<http://home.cc.umanitoba.ca/~wolowiec/spinworks/index.html>), matNMR 3 (<http://matnmr.sourceforge.net/>), and iNMR (<http://www.inmr.net/free.html>).

## 2.4 Data Analysis

Figure 2 is a representative spectrum of serum and urine originating from a cancer patient. While we use Chenomx software ([www.chenomx.com](http://www.chenomx.com)) in combination with publically available metabolite databases such as the human metabolite database (HMDB, [www.hmdb.ca](http://www.hmdb.ca)), biological magnetic resonance data bank (BMRB, [www.wise.bmrb.edu/metabolomics](http://wise.bmrb.edu/metabolomics)) for metabolomics assignments and quantification. There are software available to automatically identify [44–46] and quantify [46, 47] metabolites. While Chenomx does a good job in data analysis, due to the smaller  $^1\text{H}$  chemical shift window there are severe overlaps of peaks that lead to ambiguity in metabolite identification and quantification in 1D NMR spectra. Often two dimensional (2D) homonuclear  $^1\text{H}$ - $^1\text{H}$  correlation spectroscopy (COSY) [48, 49] and total correlation spectroscopy (TOCSY) [50] along with heteronuclear  $^1\text{H}$ ,  $^{13}\text{C}$  single quantum coherence ( $^1\text{H}$ - $^{13}\text{C}$ -HSQC) [51], heteronuclear multiple bond correlation (HMBC) [52] and J-resolved NMR spectroscopy (J-Res) [53, 54] are performed to aid in the metabolite assignment and quantification. Figure 3 shows a representative  $^1\text{H}$ - $^{13}\text{C}$ -HSQC data of cell extracts [51].

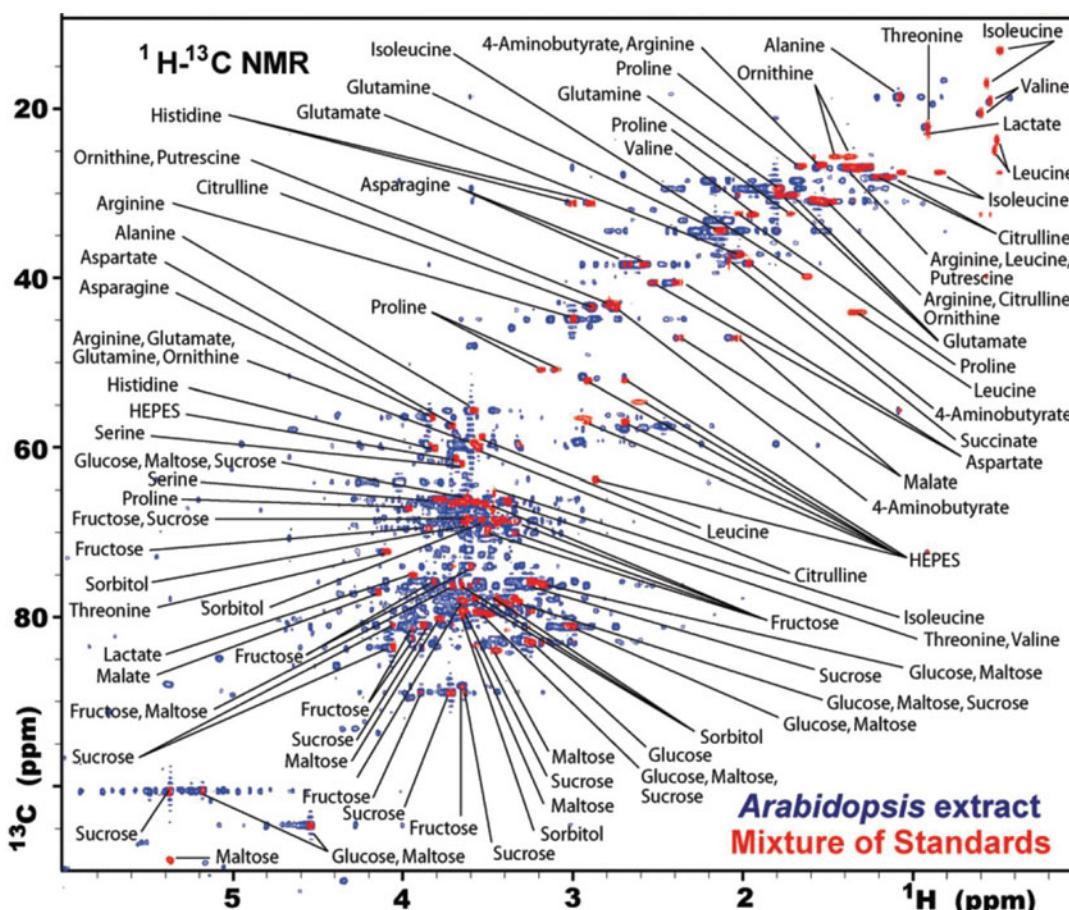


**Fig. 2** Representative 600 MHz  $^1\text{H}$  NMR spectra of (a) serum and (b) urine from a colorectal cancer patient with assigned metabolites: (1) isoleucine; (2) leucine; (3) valine; (4) lactate; (5) threonine; (6) alanine; (7) lysine; (8) arginine; (9) acetate; (10) glutamate; (11) methionine; (12) glutamine; (13) choline; (14) phosphocholine (PC); (15) glycerophosphocholine (GPC); (16) taurine; (17) glycine; (18) trimethylamine (TMA); (19) creatine; (20) creatinine; (21) tyrosine; (22) phenylalanine; (23) histidine; (24) uracil; (25) succinate; (26) formate; (27) pyruvate; (28) glucose; (29) citrate; (30) dimethylamine (DMA); (31) 3-hydroxybutyrate; (32) hippurate; (33) indoxyl sulfate; (34) phenylacetylglycine

### 3 Metabolomics Profiling and Multivariate Analysis

The post-processing 1D  $^1\text{H}$  NMR spectra are used for metabolomics profiling since each spectrum contains information about a wide range of structurally diverse metabolites simultaneously present at different concentrations providing a “snapshot” of metabolites in a given time point in a biological sample. Metabolomics profiling is carried out either supervised or unsupervised fashion. We follow the following steps for metabolomics profiling:

1. The spectral region of the processed 1D  $^1\text{H}$  data between 0.5 and 10.0 ppm is integrated into bins of 0.004 ppm using AMIX (Bruker BioSpin) or Chenomx.
2. The regions between 4.2 and 5.1 ppm are discarded to eliminate the effects of imperfect water suppression. Urea resonance between 5.53 and 6.25 ppm for urine samples and residual solvent signals such as methanol are removed.
3. The areas of all bins are then normalized to the total intensity or dry sample weight.
4. Multivariate data analysis is performed using SIMCA 15 (<https://umetrics.com/products/simca>). There are other freeware available including MetaboAnalyst 4.0 (<https://www.metaboanalyst.ca/>) [55], muma (Metabolomics Univariate and Multivariate Analysis, <https://cran.r-project.org/web/>

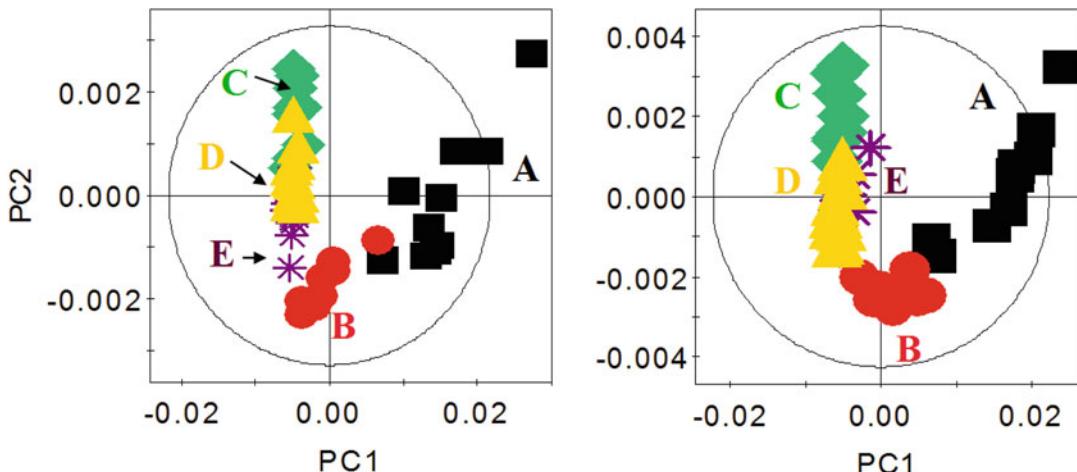


**Fig. 3** 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC NMR spectra of a synthetic mixture (red) overlaid onto a spectrum of aqueous whole-plant extract from *Arabidopsis thaliana* (blue). (Reproduced by permission of ACS Publications [51]. Copyright 2007 ACS Publications)

packages/muma/index.html) and R package (<http://cran.cnr.berkeley.edu/>).

5. First unsupervised principal component analysis (PCA) is carried out on the mean-centered (Ctr) binned data.
  6. Supervised latent structure-discrimination analysis (PLS-DA) or orthogonal projection to latent structure-discrimination analysis (OPLS-DA) are performed using unit variance scaled (UV) data.

Multivariate statistical analysis allows for distinguishing the change in metabolic profiling within biological samples. Figure 4 shows a PCA analysis discriminating metabolites between adjacent intestinal contents [56].



**Fig. 4** PCA score plots derived from 1D  $^1\text{H}$  NMR data of the intestinal contents from 12-week old rats (left) and 15-week old rats (right): (A) jejunal content (black filled box), (B) ileal content (red filled circle), (C) cecal content (green filled diamond), (D) feces (yellow filled triangle), and (E) colonic content (purple asterisk). (Reproduced by permission of ACS Publications [56]. Copyright 2012 ACS Publications)

## References

- Bell JD, Sadler PJ, Morris VC, Levande OA (1991) Effect of aging and diet on proton NMR spectra of rat urine. *Magn Reson Med* 17:414–422
- Fan TWM, Lane AN, Higashi RM (2012) The handbook of metabolomics. Humana, New York
- Beckonert O, Keun HC, Ebbels TMD et al (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* 2:2692–2703. <https://doi.org/10.1038/nprot.2007.376>
- Dumas ME, Maibaum EC, Teague C et al (2006) Assessment of analytical reproducibility of  $^1\text{H}$  NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Anal Chem* 78:2199–2208. <https://doi.org/10.1021/ac0517085>
- Emwas A-H, Roy R, McKay RT et al (2019) NMR spectroscopy for metabolomics research. *Metabolites* 9:1–39. <https://doi.org/10.3390/metabo9070123>
- Nicholson JK, Lindon JC, Holmes E (1999) “Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29:1181–1189. <https://doi.org/10.1080/004982599238047>
- Dame ZT, Aziz F, Mandal R et al (2015) The human saliva metabolome. *Metabolomics* 11:1864–1883. <https://doi.org/10.1007/s11306-015-0840-5>
- Nicholls AW, Mortishire-Smith RJ, Nicholson JK (2003) NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem Res Toxicol* 16:1395–1404. <https://doi.org/10.1021/tx0340293>
- Bolland ME, Holmes E, Lindon JC et al (2001) Investigations into biochemical changes due to diurnal variation and estrus cycle in female rats using high-resolution  $^1\text{H}$  NMR spectroscopy of urine and pattern recognition. *Anal Biochem* 295:194–202. <https://doi.org/10.1006/abio.2001.5211>
- Le Belle JE, Harris NG, Williams SR, Bhakoo KK (2002) A comparison of cell and tissue extraction techniques using high-resolution  $^1\text{H}$ -NMR spectroscopy. *NMR Biomed* 15:37–44. <https://doi.org/10.1002/nbm.740>
- Hauser A, Eisenmann P, Muhle-goll C et al (2019) Efficient extraction from mice feces for NMR metabolomics measurements with

- special emphasis on SCFAs. *Metabolites* 9:1–19. <https://doi.org/10.3390/metabo9030055>
12. Lin Y, Ma C, Liu C et al (2016) NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget* 7:29454–29464. <https://doi.org/10.18632/oncotarget.8762>
  13. Tian Y, Cai J, Gui W et al (2019) Berberine directly affects the gut microbiota to promote intestinal farnesoid X receptor activation. *Drug Metab Dispos* 47:86–93. <https://doi.org/10.1124/dmd.118.083691>
  14. Beckonert O, Coen M, Keun HC et al (2010) High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nat Protoc* 5:1019–1032. <https://doi.org/10.1038/nprot.2010.45>
  15. Swanson MG, Vigneron DB, Tabatabai ZL et al (2003) Proton HR-MAS spectroscopy and quantitative pathologic analysis of MRI/3D-MRSI-targeted postsurgical prostate tissues. *Magn Reson Med* 50:944–954. <https://doi.org/10.1002/mrm.10614>
  16. Tate AR, Foxall PJD, Holmes E et al (2000) Distinction between normal and renal cell carcinoma kidney cortical biopsy samples using pattern recognition of  $^1\text{H}$  magic angle spinning (MAS) NMR spectra. *NMR Biomed* 13:64–71. [https://doi.org/10.1002/\(SICI\)1099-1492\(200004\)13:2<64::AID-NBM612>3.0.CO;2-X](https://doi.org/10.1002/(SICI)1099-1492(200004)13:2<64::AID-NBM612>3.0.CO;2-X)
  17. Martínez-Bisbal MC, Martí-Bonmatí L, Piquer J et al (2004)  $^1\text{H}$  and  $^{13}\text{C}$  HR-MAS spectroscopy of intact biopsy samples ex vivo and in vivo  $^1\text{H}$  MRS study of human high grade gliomas. *NMR Biomed* 17:191–205. <https://doi.org/10.1002/nbm.888>
  18. Somashekar BS, Amin AG, Rithner CD et al (2011) Metabolic profiling of lung granuloma in *Mycobacterium tuberculosis* infected guinea pigs: ex vivo  $^1\text{H}$  magic angle spinning NMR studies. *J Proteome Res* 10:4186–4195. <https://doi.org/10.1021/pr2003352>
  19. Calvo N, Beltrán-Debón R, Rodríguez-Gallego E et al (2015) Liver fat deposition and mitochondrial dysfunction in morbid obesity: an approach combining metabolomics with liver imaging and histology. *World J Gastroenterol* 21:7529–7544. <https://doi.org/10.3748/wjg.v21.i24.7529>
  20. Lin AQ, Shou JX, Li XY et al (2014) Metabolic changes in acute cerebral infarction: findings from proton magnetic resonance spectroscopic imaging. *Exp Ther Med* 7:451–455. <https://doi.org/10.3892/etm.2013.1418>
  21. Simões RV, Martinez-Aranda A, Martín B et al (2008) Preliminary characterization of an experimental breast cancer cells brain metastasis mouse model by MRI/MRS. *Magn Reson Mater Phys Biol Med* 21:237–249. <https://doi.org/10.1007/s10334-008-0114-6>
  22. Moka D, Vorreuther R, Schicha H et al (1998) Biochemical classification of kidney carcinoma biopsy samples using magic-angle-spinning  $^1\text{H}$  nuclear magnetic resonance spectroscopy. *J Pharm Biomed Anal* 17:125–132
  23. Cheng LL, Pohl U (2006) The role of NMR-based metabolomics in cancer. In: Lindon JC, Nicholson JK, Holmes E (eds) *Handbook of metabonomic and metabolomics*. Elsevier, Amsterdam, pp 345–374
  24. Lindon JC, Holmes E (2006) A survey of metabolomics approaches for disease characterization. In: Lindon JC, Nicholson JK, Holmes E (eds) *Handbook of metabonomic and metabolomics*. Elsevier, Amsterdam, pp 413–442
  25. Moolenaar SH, Engelke UFH, Wevers RA (2003) Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. *Ann Clin Biochem* 40:16–24. <https://doi.org/10.1258/000456303321016132>
  26. Brindle JT, Antti H, Holmes E et al (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using  $^1\text{H}$ -NMR-based metabolomics. *Nat Med* 8:1439–1444. <https://doi.org/10.1038/nm802>
  27. Mickiewicz B, Duggan GE, Winston BW et al (2014) Metabolic profiling of serum samples by  $^1\text{H}$  nuclear magnetic resonance spectroscopy as a potential diagnostic approach for septic shock. *Crit Care Med* 42:1140–1149. <https://doi.org/10.1097/CCM.0000000000000142>
  28. Nicholson JK, Holmes E, Wilson ID (2005) Gut microorganisms, mammalian metabolism and personalized health care. *Nat Rev Microbiol* 3:431–438. <https://doi.org/10.1038/nrmicro1152>
  29. Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* 1:153–161. <https://doi.org/10.1038/nrd728>
  30. Gartland KPR, Sanins SM, Nicholson JK et al (1990) Pattern recognition analysis of high resolution  $^1\text{H}$  NMR spectra of urine. A non-linear mapping approach to the classification of toxicological data. *NMR Biomed* 3:166–172. <https://doi.org/10.1002/nbm.1940030404>

31. Gartland KPR, Beddell CR, Lindon JC, Nicholson JK (1991) Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine. *Mol Pharmacol* 39:629–642
32. Ebbels T, Keun H, Beckonert O et al (2003) Toxicity classification from metabolomic data using a density superposition approach: “CLOUDS”. *Anal Chim Acta* 490:109–122. [https://doi.org/10.1016/S0003-2670\(03\)00121-1](https://doi.org/10.1016/S0003-2670(03)00121-1)
33. Antti H, Ebbels TMD, Keun HC et al (2004) Statistical experimental design and partial least squares regression analysis of biofluid metabolomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemom Intell Lab Syst* 73:139–149. <https://doi.org/10.1016/j.chemolab.2003.11.013>
34. Cloarec O, Dumas ME, Craig A et al (2005) Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic  $^1\text{H}$  NMR data sets. *Anal Chem* 77:1282–1289. <https://doi.org/10.1021/ac048630x>
35. Crockford DJ, Holmes E, Lindon JC et al (2006) Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabolomic toxicology studies. *Anal Chem* 78:363–371. <https://doi.org/10.1021/ac051444m>
36. Emwas AH, Saccenti E, Gao X et al (2018) Recommended strategies for spectral processing and post-processing of 1D  $^1\text{H}$ -NMR data of biofluids with a particular focus on urine. *Metabolomics* 14:1–23. <https://doi.org/10.1007/s11306-018-1321-4>
37. Lacy P, McKay RT, Finkel M et al (2014) Signal intensities derived from different NMR probes and parameters contribute to variations in quantification of metabolites. *PLoS One* 9:1–10. <https://doi.org/10.1371/journal.pone.0085732>
38. Mo H, Raftery D (2008) Pre-SAT180, a simple and effective method for residual water suppression. *J Magn Reson* 190:1–6. <https://doi.org/10.1016/j.jmr.2007.09.016>
39. Meiboom S, Gill D (1958) Modified spin-echo method for measuring nuclear relaxation times. *Rev Sci Instrum* 29:688–691. <https://doi.org/10.1063/1.1716296>
40. Carr HY, Purcell EM (1954) Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Phys Rev* 94:630–638
41. Wishart DS (2008) Quantitative metabolomics using NMR. *Trends Anal Chem* 27:228–237. <https://doi.org/10.1016/j.trac.2007.12.001>
42. Aue WP, Karhan J, Ernst RR (1976) Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *J Chem Phys* 64:4226–4227. <https://doi.org/10.1063/1.431994>
43. Wu DH, Chen A, Johnson CS (1995) An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *J Magn Reson Ser A* 115:260–264
44. Cañuelo D, Gómez J, Salek RM et al (2018) rDolphin: a GUI R package for proficient automatic profiling of 1D  $^1\text{H}$ -NMR spectra of study datasets. *Metabolomics* 14:1–5. <https://doi.org/10.1007/s11306-018-1319-y>
45. Cui Q, Lewis IA, Hegeman AD et al (2008) Metabolite identification via the Madison Metabolomics Consortium Database [3]. *Nat Biotechnol* 26:162–164. <https://doi.org/10.1038/nbt0208-162>
46. Tardivel PJC, Canlet C, Lefort G et al (2017) ASICS: an automatic method for identification and quantification of metabolites in complex 1D  $^1\text{H}$  NMR spectra. *Metabolomics* 13:1–9. <https://doi.org/10.1007/s11306-017-1244-5>
47. Röhnisch HE, Eriksson J, Müllner E et al (2018) AQuA: an automated quantification algorithm for high-throughput NMR-based metabolomics and its application in human plasma. *Anal Chem* 90:2095–2102. <https://doi.org/10.1021/acs.analchem.7b04324>
48. Le Guennec A, Giraudeau P, Caldarelli S (2014) Evaluation of fast 2D NMR for metabolomics. *Anal Chem* 86:5946–5954. <https://doi.org/10.1021/ac500966e>
49. Féraud B, Govaerts B, Verleysen M, de Tullio P (2015) Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolic Informative Content and comparison with  $^1\text{H}$ -NMR. *Metabolomics* 11:1756–1768. <https://doi.org/10.1007/s11306-015-0830-7>
50. Sandusky P, Raftery D (2005) Use of selective TOCSY NMR experiments for quantifying minor components in complex mixtures: application to the metabolomics of amino acids in honey. *Anal Chem* 77:2455–2463. <https://doi.org/10.1021/ac0484979>
51. Lewis IA, Schommer SC, Hodis B et al (2007) Method for determining molar concentrations of metabolites in complex solutions from two-dimensional  $^1\text{H}$ - $^{13}\text{C}$  NMR spectra. *Anal*

- Chem 79:9385–9390. <https://doi.org/10.1021/ac071583z>
52. Bernini P, Bertini I, Luchinat C et al (2009) Individual human phenotypes in metabolic space and time. J Proteome Res 8:4264–4271. <https://doi.org/10.1021/pr900344m>
53. Fonville JM, Maheir AD, Coen M et al (2010) Evaluation of full-resolution J-resolved  $^1\text{H}$  NMR projections of biofluids for metabolomics information retrieval and biomarker identification. Anal Chem 82:1811–1821. <https://doi.org/10.1021/ac902443k>
54. Ludwig C, Viant MR (2010) Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. Phytochem Anal 21:22–32. <https://doi.org/10.1002/pca.1186>
55. Chong J, Wishart DS, Xia J (2019) Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. Curr Protoc Bioinformatics 68:1–128
56. Tian Y, Zhang L, Wang Y, Tang H (2012) Age-related topographical metabolic signatures for the rat gastrointestinal contents. J Proteome Res 11:1397–1411. <https://doi.org/10.1021/pr2011507>

# INDEX

## A

- ACCENSE ..... 249, 250  
ArrayExpress ..... 127  
AuDNNsynergy model ..... 225–232, 234–236

## B

- Bile Acid measurement, vi ..... 291–299  
Binary classification approach ..... 78–83

## C

- Carr-Purcell-Maiboom-Gill (CPMG) ..... 306  
cBioPortal ..... 128, 130  
Change Control ..... 10, 131  
Chimera ..... 144  
Chromosomal integration ..... 131  
Clinical informatics ..... 1–15, 24, 189  
Clinical informatics infrastructure  
    examples of translational informatics  
        projects ..... 12–15  
        governance ..... 4, 5, 8, 11, 14  
        organizational specific issues ..... 5, 8  
        staffing ..... 2, 5, 7, 11, 13  
        workflows ..... 3–5, 7–12, 15  
Clinical pathways ..... 13, 14, 45–58  
Cohort identification ..... 3, 13,  
    35–42, 188, 191  
Correlation analysis (TCGA) ..... 135  
Correlation spectroscopy (COSY) ..... 307  
Cytobank® ..... 250  
Cytosplore ..... 250

## D

- Danio rerio*, *see Zebrafish*  
Data warehouse (DW) ..... 4, 12, 13,  
    15, 36, 37, 39–42  
Deep learning models, vi ..... 223–236  
DeepSynergy ..... 225, 234, 236  
Deletion vectors ..... 68, 79, 80, 229, 242  
Discrete data from laboratory ..... 7, 11, 23–31  
DNA Alignment ..... 206, 207  
Drying chamber ..... 2, 117, 201, 304, 305

## E

- EHR clinical pathway integration ..... 54–58  
Elastic nets ..... 63, 69, 70,  
    77, 81, 84, 94, 95, 99, 234–236  
Electronic health records (EHRs) ..... 2, 3, 5–7,  
    11–14, 22, 23, 37, 46, 47  
Epigenetic regulation by methylation ..... 135–138  
Expression profiling ..... 68, 156, 181, 235

## F

- Failure risk assessment, vi ..... 77–104  
Feeder cells, *see* Mouse embryonic fibroblasts (MEF)  
Flow cytometry  
    compensation ..... 241, 242,  
        246, 247, 259, 262  
    doublet discrimination ..... 243, 247  
    fluorescence minus one (FMO) ..... 243, 245  
    live-dead controls ..... 242, 243, 247  
FlowSOM ..... 249, 250  
Fluorescence minus one (FMO) ..... 61, 243–247,  
    259, 262, 265

## G

- Gene expression omnibus (GEO) ..... 100  
Genome data commons (GDC) ..... 127, 128, 145  
Genomewide association studies (GWAS) ..... 109,  
    117, 160  
Genomic approaches in biomedical research ..... 107–109  
Genomic modifications ..... 109, 114–117  
Genotyping ..... 216  
Germline transmission ..... 110, 111  
GISTIC plot ..... 131–133  
Gradient boosting machines ..... 234–236

## H

- Health information management system  
    (HIMS) ..... 35–37, 39, 40  
Heteronuclear multiple bond correlation  
    (HMBC) ..... 307  
High-resolution magic angle spinning  
    (HRMAS) ..... 302

**I**

- ImmPort ..... 250  
 Inter-sample diversity ..... 177–185  
 Intra-sample diversity  
     cluster detection ..... 181, 182  
     diversity score calculation ..... 182  
     quality control and normalization of count  
         matrix ..... 180, 181  
 Isotypes ..... 244–246

**L**

- Laboratory Information System (LIS) ..... 2,  
     23–27, 29, 31, 32  
 LacZ, *see* β-galactosidase  
 Laser-supported injection, *see* Microinjection

**M**

- Mass spectrometry, v, vi ..... 210, 216, 292, 299  
     UHPLC-MS/MS, vi ..... 291–299  
 Metabolomics ..... 301–310  
 MicroRNAs ..... 188  
 miRNAs, *see* MicroRNAs  
 Molecular pathology (MP) ..... 3, 24–26, 28, 31  
 Multi-omics ..... 187–216, 223–236

**N**

- Nematode, *see* *C. Elegans*  
 NoSQL ..... 39  
 Nuclear magnetic resonance (NMR) ..... v, vi,  
     292, 301–310  
     HMBC ..... 307  
     HRMAS ..... 302  
     HSQC ..... 307, 309  
 Nuclear magnetic resonance-based  
     metabolomics ..... 301–310

**O**

- Oncology Research Information Exchange Network  
     (ORIEN) ..... 13, 36, 40

**P**

- PanCancer Atlas  
     copy Number Alterations data ..... 130  
     RNASeq ..... 128  
     survival Analysis ..... 130  
     TCGA ..... 226  
 Pathology informatics ..... 12, 21–32  
 Pathway development ..... 47–49, 54  
 Phenotype ..... 109–112, 115,  
     116, 183, 188, 208, 213, 245, 249–251, 261, 287  
 Phenotyping NK and NKT cells ..... 255, 256,  
     261, 287

- Polymerase chain reaction (PCR) ..... 113, 146  
 Proteogenomics ..... 187–216  
 Proteomics ..... 15, 188,  
     189, 191, 192, 194, 195, 197, 198, 200–203,  
     209–213, 216, 217, 225

**R**

- Random forests (RF) ..... 73, 74, 83, 234–236  
 Relational database ..... 37, 39, 40  
 Risk Assessment, vi ..... 77–104  
 RNA Alignment ..... 207

**S**

- scRNASeq  
     correcting for multiple testing ..... 152  
     data integration and batch correction ..... 161  
     datasets ..... 143, 145,  
         161, 165, 177–185  
     differential gene expression ..... 150, 152  
     dimension reduction, clustering, and cell type  
         identification ..... 161  
     drop-outs, normalization, and spike-ins ..... 155  
     gene summarization/abundance estimation ..... 149  
     normalization ..... 50, 149,  
         155–157, 159–161  
     normalization methods ..... 150, 156, 157,  
         159–161  
     quality control (QC) ..... 146, 154,  
         155, 179, 180  
     read alignment ..... 146, 147  
     spike-ins ..... 155, 157, 160–162  
     statistics and Bioinformatics ..... 143–167  
     studying heterogeneity using scRNA-seq ..... 166  
     TME using RNA-seq in bulk samples ..... 153  
     transcript reconstruction ..... 148  
 Spanning-tree progression analysis of density-normalized  
     events (SPADE) ..... 249, 250  
 Structure Query Language (SQL) ..... 38–42  
 Synergistic Drug Combination, vi ..... 223–236

**T**

- Tandem Mass Spectrometry ..... 188, 292  
 The Cancer Genome Atlas (TCGA)  
     survival analysis ..... 138  
 Time to event data  
     machine learning ..... 62, 63, 72, 74  
     materials (available scripts) ..... 62  
     penalized regression method ..... 62, 65, 72  
     supervised principal component ..... 64  
     survival tree and random survival forest ..... 72, 74  
     variable selection ..... 61–73  
 Total Cancer Care (TCC) ..... 12, 13, 36,  
     39, 41, 42, 189, 196

- Total correlation spectroscopy (TOCSY) ..... 307  
Transcriptomic dynamics ..... 109, 112  
Transcriptomic modifications ..... 114  
Transgene  
    silencing ..... 134  
TransMed® ..... 38, 40–42  
Transposon ..... 81
- UHPLC-MS/MS ..... vi, 291–299
- V  
Variable selection ..... 61–74, 78,  
    83–88, 91, 95, 96, 99, 104