

Computational Epigenetics and Diseases

Translational Epigenetics Series

Volume 9

Edited by
Loo Keat Wei



Computational Epigenetics and Diseases

Translational Epigenetics Series

Trygve O. Tollefsbol, Series Editor

Transgenerational Epigenetics

Edited by Trygve O. Tollefsbol, 2014

Personalized Epigenetics

Edited by Trygve O. Tollefsbol, 2015

Epigenetic Technological Applications

Edited by Y. George Zheng, 2015

Epigenetic Cancer Therapy

Edited by Steven G. Gray, 2015

DNA Methylation and Complex Human Disease

By Michel Neidhart, 2015

Epigenomics in Health and Disease

Edited by Mario F. Fraga and Agustin F. F. Fernández, 2015

Epigenetic Gene Expression and Regulation

Edited by Suming Huang, Michael Litt, and C. Ann Blakey, 2015

Epigenetic Biomarkers and Diagnostics

Edited by Jose Luis García-Giménez, 2015

Drug Discovery in Cancer Epigenetics

Edited by Gerda Egger and Paola Barbara Arimondo, 2015

Medical Epigenetics

Edited by Trygve O. Tollefsbol, 2016

Chromatin Signaling and Diseases

Edited by Olivier Binda and Martin Fernandez-Zapico, 2016

Genome Stability

Edited by Igor Kovalchuk and Olga Kovalchuk, 2016

Chromatin Regulation and Dynamics

Edited by Anita Göndör, 2016

Neuropsychiatric Disorders and Epigenetics

Edited by Dag H. Yasui, Jacob Peedicayil and Dennis R. Grayson, 2016

Polycomb Group Proteins

Edited by Vincenzo Pirrotta, 2016

Epigenetics and Systems Biology

Edited by Leonie Ringrose, 2017

Cancer and Noncoding RNAs

Edited by Jayprokas Chakrabarti and Sanga Mitra, 2017

Nuclear Architecture and Dynamics

Edited by Christophe Lavelle and Jean-Marc Victor, 2017

Epigenetic Mechanisms in Cancer

Edited by Sabita Saldanha, 2017

Epigenetics of Aging and Longevity

Edited by Alexey Moskalev and Alexander M. Vaiserman, 2017

The Epigenetics of Autoimmunity

Edited by Rongxin Zhang, 2018

Epigenetics in Human Disease, Second Edition

Edited by Trygve O. Tollefsbol, 2018

Epigenetics of Chronic Pain

Edited by Guang Bai and Ke Ren, 2019

Epigenetics of Cancer Prevention

Edited by Anupam Bishayee and Deepak Bhatia, 2019

Translational Epigenetics
Volume 9

Computational Epigenetics and Diseases

Edited by

Loo Keat Wei

Universiti Tunku Abdul Rahman, Kampar, Malaysia



ACADEMIC PRESS

An imprint of Elsevier

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

Copyright © 2019 Elsevier Inc. All rights reserved.

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

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

Notices

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

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

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

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-814513-5

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



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Publisher: Andre Wolff

Acquisition Editor: Rafael Teixeira

Editorial Project Manager: Megan Ashdown

Production Project Manager: Punithavathy Govindaradjane

Cover Designer: Greg Harris

Typeset by TNQ Technologies

Contents

Contributors	xvii
--------------------	------

CHAPTER 1 Computational Epigenetics and Disease.....1

Loo Keat Wei

Introduction.....	1
Computational Approaches in DNA Methylation	1
Computational Approaches in Histone Modifications.....	3
Computational Approaches in miRNAs.....	4
Computational Epigenetics in Metabolic and Cardiac Disorders	4
Computational Epigenetics in Neurological Disorders	5
Computational Epigenetics and Cancer	6
Conclusions.....	7
Acknowledgment	7
References.....	7

CHAPTER 2 Computational Methods for Epigenomic Analysis.....11

Ho-Ryun Chung

Introduction.....	11
Unbiased Detection of ChIP-Enrichment.....	12
Segmentation of the Epigenome Into Chromatin States.....	16
The Differential Epigenome	19
References.....	21

CHAPTER 3 Statistical Approaches for Epigenetic Data Analysis.....23

Thorsten Dickhaus

Introduction.....	23
Statistical Modeling	24
Statistical Methodology.....	25
Formulation of Multiple Test Problems	25
Test Statistics and Their Limiting Null Distributions.....	26
Multiple Test Procedures: Closure Principle	27
Finite Sample Modification: Studentized Permutation Approach.....	28
Real Data Analysis	29
Discussion	30
Acknowledgments.....	31
References.....	31

CHAPTER 4 Bioinformatics Methodology Development for the Whole Genome Bisulfite Sequencing	33
<i>Deqiang Sun</i>	
Introduction.....	33
Results.....	35
Beta-Binomial Hierarchical Model for Both Sampling and Biological Variations.....	35
Credible Methylation Difference (CDIF) Is a Single Metric for Both Statistical and Biological Significance of Differential Methylation.....	35
Functions and Performance of the MOABS Pipeline.....	37
Simulated BS-seq Data Reveal the Superior Performance of MOABS.....	39
MOABS Improves the Detection of Allele-Specific DNA Methylation	44
MOABS Reliably Reveals Differential Methylation Underlying TFBSS	47
MOABS Detects Differential 5hmC in ES Cells Using RRBS and oxBS-seq	52
Discussion	57
Methods.....	57
Distribution for Difference of Two Binomial Proportions	57
Distribution for Difference of Difference	58
Distribution for Measurements With Replicates.....	58
Acknowledgments.....	59
References.....	59
Supplementary Methods	61
Methylation Ratio of One Locus Follows a Beta Distribution.....	61
CI for Single Binomial Proportion.....	62
CI for Difference of Two Binomial Proportions in Detail	63
Identification of DMCs for Two or More Samples	64
Identification of DMRs for Two Samples by Simply Grouping DMCs.....	65
Identification of DMRs for Two Samples by Hidden Markov Model	65
Identification Hypomethylated Regions from One Sample.....	66
Supplementary References	66
CHAPTER 5 Data Analysis of ChIP-Seq Experiments: Common Practice and Recent Developments	67
<i>Qi Zhang</i>	
The Design of ChIP-Seq	68
The Quality of ChIP-Seq Data	69
Mapping ChIP-Seq Reads	69
Peak Calling.....	70
Differential Enrichment Detection	71

All-in-One Data Analysis Pipelines for ChIP-Seq	72
Beyond the Standard Pipeline: Allelic-Imbalance Detection From ChIP-Seq	73
Summary	75
References.....	75
CHAPTER 6 Computational Tools for microRNA Target Prediction	79
<i>Nurul-Syakima Ab Mutalib, Siti Aishah Sulaiman and Rahman Jamal</i>	
Introduction.....	79
Principles of microRNA Target Prediction	80
Seed Sequence Complementarity.....	81
Free Energy.....	82
G–U Wobble	82
Evolutionary Conservation Status	82
3' UTR Compensatory Binding.....	82
Target-Site Accessibility.....	83
Target-Site Abundance	83
Local AU Flanking Content	83
Machine Learning	84
Pattern-Based Approach.....	84
microRNA Target Prediction Tools	84
Conclusion and Future Direction	86
References.....	98
Further Reading	105
CHAPTER 7 Integrative Analysis of Epigenomics Data	107
<i>Cenny Taslim, Stephen L. Lessnick and Simon Lin</i>	
Introduction.....	107
Quality Control and Data Preprocessing.....	109
Relationship Between Histone Modification Pattern, Transcription Factor Binding, and mRNA Expression Level	110
Regression Analysis.....	111
Mixture Model	112
Identification of Functional Regulatory Regions	114
Association Between Multiple Transcription Factors Using Self-Organizing Map (SOM)	115
Prediction of Chromatin and Transcription Binding Sites Directly From DNA Sequences Using Deep Learning	116
Discussion	117
Acknowledgments.....	118
References.....	118

CHAPTER 8 Differential DNA Methylation and Network Analysis in Schizophrenia	121
<i>Huang Kuo Chuan</i>	
Introduction.....	121
Methodology for DNA Methylation.....	121
Methylation Schizophrenia Network.....	123
Novel Prediction Applications	123
Candidate Genes in Schizophrenia.....	123
SDMGs and Disease Mechanism of Schizophrenia	123
Corresponding Pathways and Schizophrenia	125
Schizophrenia and Epigenetic Review	126
Findings Highlight the Significance of Antipsychotic Drugs on DNA	
Methylation in Schizophrenia Patients.....	127
References.....	128
CHAPTER 9 Epigenome-Wide DNA Methylation and Histone Modification of Alzheimer's Disease	131
<i>Ankush Bansal and Tiratha Raj Singh</i>	
Background	131
Epigenetics Association With the Nervous System.....	131
Epigenetic Mechanisms in AD.....	132
Epigenetic Changes in AD	132
Epigenetic Modifications.....	133
DNA Methylation	133
Hypomethylation in AD	134
Hydroxymethylation in AD	134
Gene-Wise DNA Methylation Changes in AD.....	134
Genome-Wide DNA Methylation Alternations in AD	135
DNA Repair and Methylation in AD	135
Histone Modifications.....	136
Histone Acetylation Changes in AD.....	136
Gene-Wise Histone Alterations in AD.....	136
Epigenomics.....	137
Molecular Mechanisms Linking Genomic Risk Factors to AD	137
Polymorphisms and AD	137
Systems Level Modules for AD	138
Future Directions	141
References.....	142
Further Reading	148

CHAPTER 10	Epigenomic Reprogramming in Cardiovascular Disease	149
<i>Yang Zhou, Jiandong Liu and Li Qian</i>		
Introduction.....	149	
Decipher Histone Codes of CM Transcription	150	
Identify Chromatin Modification Landscapes and Dynamics During Heart Development	151	
Dynamics of Regulatory cis-Elements in Heart Disease.....	152	
DNA Methylation During Heart Development and in Disease.....	153	
DNA Methylation Is Orchestrated in Normal Heart.....	153	
DNA Methylation Is Potential Therapeutic Target in Heart Disease.....	154	
DNA Hydroxymethylation Regulates Gene Expression in Cardiac Development and Hypertrophy.....	155	
Chromatin Conformation in Cardiomyocytes	155	
Rapid Chromatin Switch During Somatic Reprogramming.....	156	
Conclusion	157	
References.....	157	
CHAPTER 11	Bioinformatic and Biostatistic Methods for DNA Methylome Analysis of Obesity	165
<i>Sarah Amandine Caroline Voisin</i>		
Which DNA Methylation Assessment Technique Should I Use?	165	
Which Software and Data Sets Should I Use to Analyze DNA Methylation Data in the Context of Obesity?	167	
How Do I Annotate My DMRs to Specific Genes?	169	
What Does a Difference of 5% in Methylation Mean?.....	170	
How Do I Know Whether My DMRs Are a Cause or a Consequence of Obesity?	171	
How Can I Be Sure That My DMRs Are Not Due to Differences in Cell Type Proportions?	172	
References.....	173	
CHAPTER 12	Epigenomics of Diabetes Mellitus	181
<i>Ivanka Dimova</i>		
Basics of Epigenetics.....	182	
Epigenetic Regulation in Type 2 Diabetes Mellitus	185	
Epigenetics in Vascular Complications of Type 2 Diabetes Mellitus	187	
Epigenetics and Cancer Development in Type 2 Diabetes Mellitus	188	
Role of microRNAs (miRNAs) in Type 2 Diabetes Mellitus.....	191	
Future Perspectives and Epigenetic Drugs.....	192	
Conclusion	193	
References.....	194	

CHAPTER 13 Epigenetic Profiling in Head and Neck Cancer 201

*Javed Hussain Choudhury, Sharbadeb Kundu, Fazlur Rahaman Talukdar,
Ruhina S. Laskar, Raima Das, Shaheen Laskar, Bishal Dhar, Manish Kumar,
Sharad Ghosh, Rosy Mondal, Yashmin Choudhury and Sankar Kumar Ghosh*

Introduction.....	201
Epigenetic Alterations in Cancer	202
DNA Methylation Profiling in Head and Neck Cancer.....	204
Techniques Available for Epigenetic Profiling of HNC	206
Methylation Specific PCR	206
Combined Bisulfite Restriction Analysis Assay	206
Bisulfite Sequencing	206
Pyrosequencing.....	208
Whole Genome Bisulfite Sequencing	208
Array or Bead Hybridization Techniques for Epigenetic Profiling.....	208
Enrichment-Based Methods	209
Methylated DNA Immunoprecipitation	209
Computational Epigenetics Analysis	209
Bioinformatics Tools for Computational Epigenomics	210
Methods for Analyzing and Interpreting the DNA Methylation Data	210
Conclusion and Future Perspectives	214
References.....	215

CHAPTER 14 Epigenome-Wide DNA Methylation Profiles in Oral Cancer 219

Raghunath Chatterjee, Shantanab Das, Aditi Chandra and Baidehi Basu

Introduction.....	219
Epigenetic Regulation in Oral Cancer	220
Need for Computational Tools in Epigenetics Study	221
Available Methods and Computational Tools for Oral Cancer Methylomics	221
Tools for Methylomics by Bisulfite-Sequencing Method.....	221
Tools for Methylomics by Bisulfite-Microarray Method	223
Tools for Methylomics by Enrichment-Based Method.....	223
DNA Methylation Data Visualization	224
DNA Methylomics in Oral Cancer	224
DNA Methylation Biomarker for OSCC	224
Advancement in DNA Methylation Study in OSCC	227
Conclusion	228
References.....	228

CHAPTER 15 Computational Epigenetics for Breast Cancer	233
<i>Juan Xu, Yongsheng Li, Tingting Shao and Xia Li</i>	
Introduction.....	233
DNA Methylation in Breast Cancer.....	233
Histone Modification in Breast Cancer.....	235
Noncoding RNA Regulation in Breast Cancer	238
Epigenetic Databases.....	240
Epigenetic Tools in Cancer	240
Future Directions	243
References.....	244
CHAPTER 16 Integrative Epigenomics of Prostate Cancer.....	247
<i>Madonna Peter, Shivani Kamdar and Bharati Bapat</i>	
Prostate Cancer: An Overview	247
Genomic Alterations in PCa.....	247
Epigenomic Alterations in PCa.....	248
DNA Methylation	249
DNA Hydroxymethylation	249
Histone Modifications.....	250
microRNA and Long Noncoding RNA.....	250
Rationale for Integrative Analysis.....	251
Emerging Integrative Analysis Tools Utilized in PCa.....	252
Future Directions and Potential Applications for PCa	255
Concluding Remarks	256
Acknowledgments.....	256
References.....	257
CHAPTER 17 Network Analysis of Epigenetic Data for Bladder Cancer	265
<i>Bor-Sen Chen</i>	
Introduction.....	265
Materials and Methods	269
Data Preprocessing of Omics Data	269
Construction of the Stochastic Regression Models for the IGEN System.....	270
Identification of the TF Regulatory Ability a_{ij} , the miRNA Repression Ability c_{li} , and the Protein Interaction Ability d_{jk} and Their Statistical Significance Testing	271
Principal Genome-Wide Network Projection	273
Design of a Multiple Drug Combination With Minimal Side Effects for the Treatment of Bladder Cancer.....	276

Results and Discussion	276
Construction of IGEN.....	276
Projection of the Core Network Biomarkers into Biological Processes and Signaling Pathways to Investigate Carcinogenic Mechanisms of Bladder Cancer	278
The Impact of Aging, Smoking, and miRNA and Epigenetic Regulation on Bladder Carcinogenesis Through the Core Network Biomarkers	279
miR1-2 and miR200b Mediate the Reduction of Cell Proliferation and Metastasis Through KPNA2 and ECT2, Respectively.....	280
The Smoking-Related Protein HSP90AA1 and DNA Methylation of ECT2 Mediate the Metastasis of Bladder Cancer	280
Functional Module Network Analysis in Bladder Carcinogenesis	281
Two Separate Drug Combinations for Treating Stage 1 and Stage 4 Bladder Cancer Cells With Minimal Side Effects	283
Conclusion	285
References.....	286
Further Reading	288
CHAPTER 18 Epigenome-Wide Analysis of DNA Methylation in Colorectal Cancer.....	289
<i>Nurul-Syakima Ab Mutalib, Rashidah Baharuddin and Rahman Jamal</i>	
Introduction.....	289
Approaches to Analyze DNA Methylation in Colorectal Cancer	291
Epigenome-Wide Analysis of DNA Methylation in Colorectal Cancer	292
DNA Methylation Biomarkers in Colorectal Cancer	292
Blood-Based DNA Methylation Biomarkers	292
Stool-Based DNA Methylation Biomarkers	295
Prognostic Biomarkers	296
Computational Tools for DNA Methylation	296
Workflow for DNA Methylation Analysis in CRC.....	299
Conclusion	303
Acknowledgment	304
References.....	304
Further Reading	310
CHAPTER 19 Integrative Omic Analysis of Neuroblastoma	311
<i>Kamalakannan Palanichamy</i>	
Introduction.....	311
Neuroblastoma	311
Omics: Genomics, Transcriptomics, Proteomics, Epigenomics, and Metabolomics	312

Integrative Omics.....	313
Tools for NGS Data Analysis and Integrative Omics.....	313
Workflow.....	313
Neuroblastoma Omics	316
Transcriptome and Epigenome.....	319
Integrative Omics.....	320
Network Modeling, Reverse Engineering Modeling, and Dynamic Modeling	321
Machine Learning-Based modeling	321
Summary and Future Directions	322
References.....	322
CHAPTER 20 Computational Analysis of Epigenetic Modifications in Melanoma.....	327
<i>Ming Tang and Kunal Rai</i>	
Introduction.....	327
DNA Modifications.....	328
Histone Modifications and Chromatin States	330
Higher-Order Chromatin Structure	332
Nucleosome Positioning	333
Future Perspective	334
References.....	334
CHAPTER 21 DNA Methylome of Endometrial Cancer	343
<i>Golnaz Asaadi Tehrani</i>	
Introduction.....	343
Molecular Signaling Pathways of Endometrial Carcinoma.....	345
PI3/AKT/mTOR	345
MAPK/ERK.....	346
WNT/β-Catenin	346
VEGF/VEGFR	346
HER-2/neu	347
Epigenetic Alterations in Endometrial Carcinoma	347
Enzyme Digestion-Based Methods	347
Affinity Enrichment-Based Methods.....	348
Bisulfite Conversion-Based Methods	348
DNA Mismatch Repair Genes.....	350
Steroid Receptor Genes.....	354
Tumor Suppressor Genes.....	354
Other Related Genes.....	356

microRNA Aberrant Methylation in Endometrial Carcinoma	356
TS-miRNAs Involved in Endometrial Cancer With Their Function	
Including miR-129-2, miR-152, miR-124, miR-126, miR-137, and miR-491.....	357
DNA Methylation Machinery in Endometrium	358
Application of DNA Hypermethylation for Treatment	359
Future Directs and Conclusion.....	360
References.....	361
Further Reading	364
CHAPTER 22 Epigenetics and Epigenomics Analysis for Autoimmune Diseases.....	365
<i>Bhawna Gupta, Kumar Sagar Jaiswal, Arup Ghosh and Sunil Kumar Raghav</i>	
Study Design and Data Acquisition Methods.....	367
Microarray-Based Detection.....	368
Next-Generation Sequencing.....	370
Epigenetic Changes in Autoimmune Diseases	373
Rheumatoid Arthritis	373
Systemic Lupus Erythematosus.....	375
Multiple Sclerosis	375
Type 1 Diabetes	376
Analyzing Epigenetic Changes in Autoimmune Diseases.....	376
DNA Methylation	376
Histone Modification Analysis.....	380
miRNA and Targets Prediction	382
Epigenetic Databases	384
HIstome.....	385
MethylomeDB	385
MethBase	386
miRWalk2.0	386
Roadmap Epigenomics	386
Conclusion	386
References.....	387
CHAPTER 23 Computational Epigenetics in Lung Cancer	397
<i>S. Babichev, V. Lytvynenko, M. Korobchynskyi and I. Sokur</i>	
Introduction.....	397
Conceptual Basis of the Objective Clustering Inductive Technology	398
Affinity Metric and Clustering Quality Criteria to Estimate the Proximity of Gene Expression Profiles.....	399

Simulation of the Objective Clustering Process Using Lung Cancer Patients' Gene Expression Profiles	405
Practical Implementation of SOTA and DBSCAN Clustering Algorithms Within the Framework of the Objective Clustering Inductive Technology.....	408
Results of the Simulation and Discussion.....	411
Hybrid Model of Cluster—Bicluster Analysis of Gene Expression Profiles	413
Conclusions.....	415
References.....	416
Index	419

This page intentionally left blank

Contributors

S. Babichev

Jan Evangelista Purkyně University in Usti nad Labem, Usti nad Labem, Czech Republic; Kherson National Technical University, Kherson, Ukraine

Rashidah Baharuddin

UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

Ankush Bansal

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, India

Bharati Bapat

Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada; Division of Urology, University of Toronto, Toronto, ON, Canada

Baidehi Basu

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Aditi Chandra

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Raghunath Chatterjee

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Bor-Sen Chen

Lab of Control and Systems Biology, National Tsing Hua University, Hsinchu, Taiwan

Javed Hussain Choudhury

Department of Biotechnology, Assam University, Silchar, India

Yashmin Choudhury

Department of Biotechnology, Assam University, Silchar, India

Huang Kuo Chuan

Department of Nursing, Ching Kuo Institute of Management and Health, Keelung, Taiwan

Ho-Ryun Chung

Epigenomics, Max Planck Institute for Molecular Genetics, Berlin, Germany; Institute for Medical Bioinformatics and Biostatistics, Philipps-Universität Marburg, Marburg, Germany

Raima Das

Department of Biotechnology, Assam University, Silchar, India

Shantanab Das

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Bishal Dhar

Department of Biotechnology, Assam University, Silchar, India

Thorsten Dickhaus

Institute for Statistics, University of Bremen, Bremen, Germany

Ivanka Dimova

Department of Medical Genetics, Medical University Sofia, Sofia, Bulgaria

Arup Ghosh

Institute of Life Sciences, Bhubaneswar, India

Sankar Kumar Ghosh

Department of Biotechnology, Assam University, Silchar, India; University of Kalyani, Nadia, India

Sharad Ghosh

Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, India

Bhawna Gupta

School of Biotechnology, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Kumar Sagar Jaiswal

School of Biotechnology, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Rahman Jamal

UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

Shivani Kamdar

Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada

M. Korobchynskyi

Military-Diplomatic Academy named Eugene Bereznyak, Kyiv, Ukraine

Manish Kumar

Department of Biotechnology, Assam University, Silchar, India

Sharbadeb Kundu

Department of Biotechnology, Assam University, Silchar, India

Ruhina S. Laskar

International Agency for Research on Cancer (IARC), Lyon, France

Shaheen Laskar

Department of Biotechnology, Assam University, Silchar, India

Stephen L. Lessnick

Center for Childhood Cancer and Blood Diseases, Nationwide Children's Hospital Research Institute, Columbus, OH, United States; Division of Pediatric Hematology/Oncology/BMT, The Ohio State University College of Medicine, Columbus, OH, United States

Xia Li

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Yongsheng Li

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Simon Lin

Research Information Solutions and Innovation, Nationwide Children's Hospital, Columbus, OH, United States

Jiandong Liu

Department of Pathology and Laboratory Medicine, Department of Medicine, McAllister Heart Institute, University of North Carolina, Chapel Hill, NC, United States

V. Lytvynenko

Kherson National Technical University, Kherson, Ukraine

Rosy Mondal

Institute of Advanced Study in Science and Technology (IASST), Guwahati, India

Nurul-Syakima Ab Mutalib

UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

Kamalakkannan Palanichamy

Department of Radiation Oncology, The Ohio State University College of Medicine and Comprehensive Cancer Center, Columbus, OH, United States

Madonna Peter

Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada

Li Qian

Department of Pathology and Laboratory Medicine, Department of Medicine, McAllister Heart Institute, University of North Carolina, Chapel Hill, NC, United States

Sunil Kumar Raghav

Institute of Life Sciences, Bhubaneswar, India

Kunal Rai

Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, United States

Tingting Shao

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Tiratha Raj Singh

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, India

I. Sokur

Kherson Regional Oncology Dispancer, Kherson, Ukraine

Siti Aishah Sulaiman

UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

Deqiang Sun

Center for Epigenetics & Disease Prevention, Institute of Biosciences and Technology, Texas A&M University College of Medicine, Houston, TX, United States

Fazlur Rahaman Talukdar

International Agency for Research on Cancer (IARC), Lyon, France

Ming Tang

Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, United States

Cenny Taslim

Center for Childhood Cancer and Blood Diseases, Nationwide Children's Hospital Research Institute, Columbus, OH, United States

Golnaz Asaadi Tehrani

Molecular Genetics, Department of Genetics, Zanjan Branch, Islamic Azad University, Zanjan, Iran

Sarah Amandine Caroline Voisin

Genetics, Exercise and Performance, Institute for Health and Sport, Victoria University, Victoria, Australia

Loo Keat Wei

Department of Biological Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Malaysia

Juan Xu

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Qi Zhang

Department of Statistics, University of Nebraska—Lincoln, Lincoln, NE, United States

Yang Zhou

Department of Pathology and Laboratory Medicine, Department of Medicine, McAllister Heart Institute, University of North Carolina, Chapel Hill, NC, United States

COMPUTATIONAL EPIGENETICS AND DISEASE

1

Loo Keat Wei

Department of Biological Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Malaysia

INTRODUCTION

Epigenetics represents a rapidly growing and promising field for the discovery of novel disease biomarkers and understanding the pathophysiology of complex diseases. Epigenetic modifications regulate gene expression and gene activity without altering the underlying DNA sequence, but instead modifying the chromatin structure via DNA methylation, histone modifications, miRNAs, and non-coding RNAs [1]. These epigenetic mechanisms play important roles in embryonic development, transcriptional regulation, chromatin structure, genomic imprinting, and maintenance of genome integrity. While epigenetic changes are required for normal development and cell function, they can also be responsible for disease initiation and progression, especially cancer. Technological advances such as high-throughput technologies (e.g., next-generation sequencing [NGS] and microarray) and modern bioinformatics tools have enabled the profiling and mapping of large-scale epigenomic data [1]. Thus, computational approaches are required as part of the epigenomic research, especially during experimental design, data visualization, hypothesis validation, and result interpretation. Moreover, a computational modeling is required to facilitate the integration of variable data sources, including differentially methylated regions, miRNA binding, chromatin modifications, gene expressions, genetic variations, genomic regions, phenotypic characteristics, etc. Although the field of computational epigenetics is still in its infancy, the potential payoffs are enormous. It is possible to understand the mechanistic basis of human diseases by using computational approaches, even without a deep understanding of the fundamental pathophysiologic mechanisms behind the illness. By writing this book, we aim to provide theoretical insight, summarize practical implications, and draw attention to the emerging area of computational epigenetics and disease.

COMPUTATIONAL APPROACHES IN DNA METHYLATION

DNA methylation is one of the most intensely studied epigenetic modifications in humans. A methyl group is covalently added at the fifth position of cytosine (C) to form 5-methylcytosine (5mC), which is catalyzed by DNA methyltransferases (DNMTs). DNMTs are a group of enzymes that involved in the regulation of DNA methylation patterns, especially during normal development and diseases [2]. For instance, DNMT3a and DNMT3b play important roles in de novo methylation and

embryonic development, while DNMT1 maintains DNA methylation patterns during gene duplication and mitosis. Methyl-CpG-binding domain proteins (MBDs) recruit the specific components of the epigenetic machinery to read and interpret the genetic information encoded by the methylated DNA. DNA methylation can occur in the repetitive genomic regions, including satellite DNA and parasitic elements (e.g., long interspersed transposable elements [LINEs], short interspersed transposable elements [SINES], and endogenous retroviruses), which contain CpG dinucleotides for cytosine to be methylated. In humans, methylation of cytosine occurs predominantly at 5'-CpG-3' dinucleotides, and to a lesser extent at non-CpG sites (e.g., CpA, CpT, and CpC). The CpG dinucleotides are highly concentrated in CpG islands (CGIs), which are often located in the gene promoters, near the transcription start sites, and the enhancer regions [3–5]. CGIs are typically unmethylated and may undergo dynamic methylation changes during development, differentiation, and disease [5,6]. Methylated or unmethylated CGIs could affect the gene expression patterns through regulation of chromatin structure and transcription factor binding [7]. Therefore, it is crucial to measure the differential DNA methylation in the context of CG. Numerous approaches have been proposed to study DNA methylation, including bisulfite PCR sequencing, PyroMark CpG assay, Illumina's Infinium Methylation assay, quantitative MethyLight assay, luminometric methylation assay, methylated DNA immunoprecipitation (MeDIP), MeDIP coupled with high-throughput sequencing (MeDIP-seq), methyl-CpG-binding domain coupled with high-throughput sequencing (MBD-seq), methylation-sensitive restriction enzyme sequencing (MRE-seq), reduced representation bisulfite sequencing (RRBS), and whole genome bisulfite sequencing (WGBS) [8–11].

Bisulfite sequencing remains the gold standard method for the detection of DNA methylome, due to the increasing throughput of NGS technologies and the decrement in cost. The mapping and alignment of bisulfite reads from NGS (e.g., RRBS, Agilent SureSelect Human Methyl-Seq, NimbleGen SeqCap Epi CpGiant, and whole genomic bisulfite sequencing) are more complicated than the regular sequence reads. However, this massive task can become less burdensome via computational tools, which can be filtered and quality controlled by using BALM, Bismark, BRAT-nova, BS-seeker, BSMAP, MAQ, MOABS, MACAU, MEDIPS, RMAP, PASH, TAMEBS, WALT, etc. [1]. Bisulfite treatment converts the unmethylated cytosines to uracils, and subsequently recognized as thymines in the sequencing reads. The degree of DNA methylation can be calculated from the frequency of cytosines and thymines at a specific CpG locus, by aligning the raw reads against cytosines in the reference genomic sequence [1]. In brief, wild card aligners (e.g., BSMAP, RMAP, and Pash 3.0) substitute cytosines with IUPAC letter “Y” and then align with hashing extension method, in order to match to thymines in the bisulfite reads [1]. Alternatively, three-letter aligners (e.g., Bismark, BS-seeker, and BRAT-nova) can be used to convert all cytosines to lower case “t” in both reference sequence and reads, followed by short read alignment (e.g., Bowtie or Bowtie 2) based on the three-letter code of DNA (A, G, and T) [1]. Upon obtaining the processed data, DNA methylation regions can be highly predictive based on the transcriptional activity of downstream genes, transcription start sites, transcription factor binding sites, presence or absence of TATA box, and/or RNA polymerase II occupancy on DNA [3]. Such computational predictions [3] are useful, particularly where experimental data are still lacking [11,12], which represent the first step toward quantitative analysis of DNA methylation data. When no a priori knowledge is available on a candidate gene methylation, it is more acceptable to assess the DNA methylated regions comprising a number of cytosines or known as “CpG island.” Although several statistical methods have been applied in the detection of differential DNA methylated regions [13], Fisher's exact test or paired nonparametric tests

are the most common methods for comparing the methylation levels of the cytosines within the regions of interest. The false discovery rate is required to be corrected for multiple testing, based on the Benjamini–Hochberg procedure. Alternatively, probabilistic and more unbiased methods such as Hidden Markov Models (HMM) can be used for this segmentation problem. Additionally, a multivariate statistical model has been proposed for analyzing epigenetic data [14]. Such approaches are much more realistic than marginal models, in order to optimize the interpretation of the resulting epigenetic data.

COMPUTATIONAL APPROACHES IN HISTONE MODIFICATIONS

In addition to DNA methylation, histone modifications are also widely studied epigenetic mechanisms. DNA is wrapped around by an octamer of histone core to form nucleosomes, and subsequently organized into chromatin. Each nucleosome is composed of two copies of four histone proteins H2A, H2B, H3, and H4. Overall structure of chromatin can be altered through the posttranslational modifications of histone N-terminal tails, such as methylation, phosphorylation, acetylation, ubiquitination, SUMOylation, ADP ribosylation, biotinylation, deamination, and proline isomerization [15]. Notably, histone acetylation, methylation, phosphorylation, and ubiquitination are involved in gene activation, whereas methylation, ubiquitination, SUMOylation, biotinylation, deamination, and proline isomerization are involved in gene repression. These histone modifications act as the docking sites for chromatin to recruit histone chaperones and nucleosome remodelers, and subsequently alter the chromatin architecture for transcriptional activity and gene expression [16]. Typically, high levels of acetylation and trimethylated H3K4, H3K36, and H3K79 are detected in the actively transcribed euchromatin [15]. On the other hand, heterochromatin is characterized by low levels of acetylation and high levels of H3K9, H3K27, and H4K20 methylation [15]. Histone acetylation is regulated by the action of two antagonistic enzymes, histone acetyltransferases (HATs) and histone deacetylases (HDACs). HATs catalyze the transfer of an acetyl group to the ε-amino group of lysine side chains on histone tails, whereas HDACs reverse lysine acetylation by removing the acetyl functional group from lysine residues [17]. Histone phosphorylation mainly occurs on serine, threonine, and tyrosine residues within the N-terminal histone tails. All the four nucleosomal histone tails have acceptor sites which can be phosphorylated by a number of protein kinases and dephosphorylated by phosphatases [18]. Histone methylation takes place on the side chains of lysine and arginine residues. Notably, lysines can be mono-, di-, or trimethylated by histone lysine methyltransferases, whereas arginines can be either mono- or dimethylated by arginine N-methyltransferase [19]. Histone modifications can be detected using chromatin immunoprecipitation (ChIP) with deep sequencing (ChIP-seq), ChIP with DNA microarray (ChIP-chip), and ChIP with quantitative polymerase chain reaction (ChIP-qPCR) [1,20].

ChIP-seq utilizes high-throughput DNA sequencing to detect transcription factor binding and histone modifications. The initial step for ChIP data processing is the mapping of sequence reads to the reference genome. This step is usually carried out using specific software provided by NGS platforms (e.g., Illumina Genome Analyzer/HiSeq 2000/MiSeq, Applied Biosystems SOLiD Analyzer, etc.) as well as open-source alignment software (e.g., BWA, Bowtie, etc.) [1,21]. In order to analyze ChIP-seq data, a variety of peak calling methods have been developed. Typically, data of transcription factor binding sites may yield narrow ChIP-Seq peaks (sharp peaks), whereas histone modifications lead to broad regions of interests (broad peaks). The underlying algorithms for peak callers are based on several features such as the shape of peaks matters (e.g., sharp, broad, and mixed), the experimental design of

ChIP-Seq, the GC content bias, and the consistency of biological replicates in ChIP-seq experiments [21]. Chung [21] has discussed few peak calling methods, namely MACS, MACS2, spp, MOSAiCS, and GEM. After peak calling, DESeq2, edgeR, DiffBind, ChIPComp, DBChIP, MAnorm, Homer, macs2bdgdiff, and RSEG are commonly used for differential ChIP-seq analysis [21,22].

COMPUTATIONAL APPROACHES IN miRNAs

miRNAs are short single-stranded noncoding RNAs ranging from 18 to 25 nucleotides long, which are located in the intron, or intergenic region, and/or untranslated region (UTR) of the genome [23]. Biosynthesis of miRNAs involves two-stage process, with two different types of RNase III-type enzymes as the intermediates [24]. miRNAs are initially derived from longer transcripts called primary miRNA (pri-miRNA) containing one or more hairpin structures. Such hairpins are critical for recognizing and cleaving by the nuclear RNase III enzyme Drosha and produce an approximately 70-nucleotide-long hairpin precursor miRNA (pre-miRNA). pre-miRNA is subsequently cleaved by the second RNase III enzyme, Dicer, into approximately 22-nucleotide miRNA. These mature miRNAs can interact with multiple mRNA targets through partial sequence complementation at the 3'-untranslated region (3'-UTR) or 5'-UTR of the transcripts, leading to a complicated miRNA-mediated gene regulatory network [25,26]. The miRNA–mRNA base pairing may result in the degradation or blocking of mRNA translation [27]. Therefore, miRNA expression levels are inversely correlated with the corresponding mRNA expression levels. Conventional methods for miRNA detection may include northern blotting, reverse transcription-polymerase chain reaction (RT-PCR), microarrays, NGS nanoparticle-derived probes, isothermal amplification, electrochemical methods, and others [28,29].

The computational prediction and identification of novel miRNA genes remain a challenge in the field of epigenetics. Majority of the computational methods for miRNA identification are divided into both comparative and noncomparative algorithms [23]. Among them, the main miRNA features used by different computational tools are based on their sequence complementarity, evolutionary conservation of putative target sites, hairpin-shaped stem-loop secondary structure, and minimal free energy folding [27,30]. Numerous computational tools have been developed to identify and validate novel miRNAs such as miRscan, miRFinder, miPred, miRAnalyzer, miRCat, miREval, MiReNA, miRTRAP, TargetScan, miRanda, DIANA Tools, miRDeep and its updated version [27,31]. These methods incorporate different algorithms with either scoring, rule-based, machine-learning classification of the hairpin features or their combination [31]. Since a large number of computational tools are available for the identification and prediction of miRNA targets, it is crucial to understand the basic concepts of these algorithms before selecting the miRNA tool that best fits the research objectives.

COMPUTATIONAL EPIGENETICS IN METABOLIC AND CARDIAC DISORDERS

The role of DNA methylation in obesity is being increasingly recognized, through candidate gene and epigenome-wide association studies [32]. Regardless of the variety of DNA methylation profiling techniques such as Illumina arrays, MeDIP-seq, Me-DIP chip, and RRBS, much computational efforts have been made possible in the data preprocessing, filter and normalization pipeline, and statistical analysis with R software [32]. A total of 68 packages contained algorithms for preprocessing and

downstream analysis of DNA methylation data, including algorithms for cell-type deconvolution, feature selection, as well as pathway, integrative, and system-level analysis [32].

Epigenetics is a possible molecular link between environmental factors and type 2 diabetes mellitus [33]. The epigenetic regulation of type 2 diabetes mellitus has been explored in pancreatic islets, by using whole-genome DNA methylation analysis. Following bisulfite conversion, Infinium Human-Methylation450 BeadChip has been applied by interrogating 482,421 CpG sites and 3091 non-CpG sites [33]. Moreover, Illumina HiSeq2500 NGS technology is used in the epigenomic study of type 2 diabetes mellitus to generate high-quality paired-end 125 bp reads (Illumina version 4 chemistry) [33]. By using the computational epigenetic tool of Bismark, the methylation score for a particular cytosine can be calculated [33]. After that, methylation profile of patients with type 2 diabetes mellitus is smoothed, and differentially methylated regions are detected using the BSsmooth algorithm from Bioconductor bsseq package [33].

The epigenomic dynamics in heart development and cardiovascular diseases has been revealed by large-scale imputation of epigenomic data sets [34]. Epigenomic reprogramming may play important roles in both normal cardiac development and heart diseases [34]. Through the understanding of epigenomic changes with computational analysis of multi-omics data, the epigenomic signatures in cardiac development and heart diseases can be identified [34]. Recently, cardiomyocyte nuclei isolated from fetal, infant, adult, and end-stage heart failing human hearts have been used to generate high-coverage DNA methylomes by whole-genome bisulfite sequencing [34].

COMPUTATIONAL EPIGENETICS IN NEUROLOGICAL DISORDERS

Aberrant DNA methylation has been associated with various neurodegenerative and neuropsychiatric disorders. Alzheimer's disease is the most common neurodegenerative disorder characterized by an accumulation of amyloid beta plaques and aggregated hyperphosphorylated tau protein, neurofibrillary tangles, throughout the brain. In a recent computational epigenetic study, Alzheimer's disease interactome has been constructed, depends on several parameters such as degree band, similarity index, and identified Alzheimer's disease-related proteins [35]. In their study, regulatory network motifs and the patterns of epigenetic modifications are further explored [35]. A total of 22 genes and 11 miRNAs are computationally predicted from the network motifs, which may provide new insights into potential therapeutic targets for Alzheimer's disease [35]. Furthermore, epigenetic drug-target network has been constructed with the drugs associated with the proteins identified from epigenetic protein–protein interaction network [36]. As a result, 14 epigenetic repositioning drugs have overlapping epigenetic targets and miRNAs of Alzheimer's disease [36].

Parkinson's disease is the second most prevalent neurodegenerative disorder. The same group has investigated transcription factor (TF)-miRNA-mRNA regulatory network and miRNA co-expression network in Parkinson's disease [37]. A total of 14 interregulatory hub miRNAs and 18 co-expressed hub miRs are generated from both networks, respectively [37]. The roles of these 32 novel miRNAs in different molecular pathways of Parkinson's disease are further strengthened with hierarchical clustering analysis [37]. Additionally, the epigenetic regulatory network, namely mTF-miRNA-gene-gTF involving miRNA transcription factor (mTF), miRNA, gene, and gene transcription factor (gTF), as well as long noncoding RNA (lncRNA) mediated regulatory network involving miRNA, gene, mTF, and lncRNA are further constructed [38]. In brief, mTF-miRNA-gene-gTF regulatory network identified a novel feed-forward loop, whereas lncRNA-mediated regulatory network identified novel

lncRNAs of Parkinson's disease [38]. Both epigenetic regulatory networks can provide an overview of the cellular and molecular mechanisms underlying Parkinson's disease [38].

COMPUTATIONAL EPIGENETICS AND CANCER

Epigenetic changes such as DNA methylation, histone modifications, miRNA alterations, and chromatin structure have been established in several cancers, including cancer in oral, breast, head and neck, colon, gastric, prostate, ovarian, endometrial, bladder, neuroblastoma, and melanoma.

In search for the epigenetic markers of oral squamous cell carcinoma, numerous epigenomic techniques have been applied, including Illumina Golden Gate Methylation Array, Infinium Human-Methylation 450K array, HumanMethylation27 Bead Chip array, and Agilent 4×44 k Custom CGH microarray based methylated-CGI amplification method [39]. Subsequent analysis using Illumina Genome Studio software, BeadStudio Software, Partek Genomic Suite, MetaCore, and Wilcoxon rank sum test with 5% false discovery rate has been performed to identify the β values and differentially methylated probes [39]. Hierarchical agglomerative clustering using differentially methylated probes identified two distinct clusters, namely low- and high-CGI methylator phenotypes [39]. Unsupervised hierarchical clustering with Spotfire DecisionSite identified three separate clusters with higher methylation level in patients with oral squamous cell carcinoma [39]. Pathway analysis revealed hypermethylation in genes associated with cell adhesion, cell proliferation, growth regulation, and cell apoptotic pathways in oral cancer patients [39].

Epigenetic modifications have been shown to play essential roles in breast cancer. Computational epigenetic analysis can reveal distinct patterns of epi-modifications in breast cancer subtypes, particularly in the promoter regions [40]. In addition, a systemic analysis of competitive endogenous RNA (ceRNA) interactions may yield novel insights with regard to the biological networks involving in breast cancer [40]. Recently, Xu et al. [40] published a ceRNA network for each breast cancer subtype, whereby it is depending on the significance of both positive co-expression and the miRNAs identified from the miRNA dysregulatory network. From their study, a total of 29 critical subtype-specific ceRNA hubs have been found to be associated with different breast cancer subtypes [40].

A computational network-based model has been developed for genetic and epigenetic interdependencies observed at different stages in colorectal cancer [41]. This multilayered framework dynamics integrated genetic and epigenetic events, gene relationships, and cancer stage levels, by visualizing the data from StatEpigen database and incorporating hypermethylation, hypo-methylation, gene expression levels, and mutations of different genes corresponded to this disease [41]. The developed network model has been tested on a case with colorectal cancer, *carcinoma in situ* [41]. The findings indicated that the progression rate of colorectal cancer is higher for a small and closely associated network of genes than for a larger and less-connected set [41]. Therefore, the development and progression of colorectal cancer are largely dependent on genetic and epigenetic interdependencies as described in the network model [41].

Integrative epigenomic approaches have been applied in prostate cancer using several computational tools. For instance, Epidaurus is a specified bioinformatics tool that aggregates the epigenomic data sets obtained from RNA-seq, MeDIP-seq, ChIP-seq, MNase-seq, and DNase-seq and subsequently integrates the aggregated data to reveal the relevance and differences between epigenetic modifications [42]. Model-based analysis of regulation of gene expression (MARGE) uses H3K27ac ChIP-seq data to predict gene expression and transcription factor binding, which has three main

functions: MARGE-potential, MARGE-express, and MARGE-cistrome [42]. RegNetDriver is a computational tool that can identify prostate tumor regulatory drivers via the integrative analysis of genetic (e.g., single nucleotide variants, structural variants, etc.) and epigenetic (e.g., DNA methylation, histone modifications, and chromatin organization) data sets. This computational framework revealed that the differential gene expression of *FAS*, *FAM3B*, and *TNFSF13* is regulated by both genetic and epigenetic alterations [42].

Furthermore, an Integrated Genetic and Epigenetic Network (IGEN) system has been developed for the analysis of bladder cancer, based on three coupling regression models that characterize protein–protein interaction, transcription regulation, miRNA regulation, and DNA methylation [43]. The IGEN applied system identification method and principal genome-wide network projection based on principal component analysis to identify core network biomarkers in bladder carcinogenesis [43]. By assessing the GO, NCBI, and KEGG databases, the functional roles of the core network biomarkers are classified into three pathways, including SUMOylation, ubiquitination, and proteasome pathway, tumor necrosis factor signaling pathway, and endoplasmic reticulum signaling pathway [43]. Based on the connection differences of the core network biomarkers between different cellular stages, multiple drug combinations have been proposed for treating stage 1 and stage 4 bladder cancer [43].

CONCLUSIONS

The emerging field of computational epigenetics is moving from a hypothesis-driven approach toward a holistic data-driven modeling approach. The computational tools for DNA methylation, histone modifications, transcription factor binding, nucleosome positioning, and chromosomal organization have become increasingly important to the study of diseases. Furthermore, integrative analysis of multi-omics data may contribute greatly to our understanding of epigenetic modifications and transcriptional regulations at the systemic level and shed some light on the epigenomic's involvement in health and disease.

ACKNOWLEDGMENT

We acknowledge the financial support from a Fundamental Research Grant Scheme (UTARRF 6200/LF3).

REFERENCES

- [1] Wei LK, Au A. Computational epigenetics. In: Handbook of epigenetics. 2nd ed. 2017. p. 167–90.
- [2] Robertson KD. DNA methylation and human disease. *Nat Rev Genet* August 2005;6(8):597.
- [3] Wei K, Sutherland H, Camilleri E, Haupt LM, Griffiths LR, Gan SH. Computational epigenetic profiling of CpG islets in MTHFR. *Mol Biol Rep* 2014;41(12):8285–92.
- [4] Liyanage VR, Jarmasz JS, Murugesan N, Del Bigio MR, Rastegar M, Davie JR. DNA modifications: function and applications in normal and disease States. *Biology* October 22, 2014;3(4):670–723.
- [5] Jeziorska DM, Murray RJ, De Gobbi M, Gaentzsch R, Garrick D, Ayyub H, Chen T, Li E, Telenius J, Lynch M, Graham B. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci USA* September 5, 2017;114(36):E7526–35.

- [6] Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* May 1, 2014;6(5):a019133.
- [7] Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev.* May 15, 2011;25(10):1010–22.
- [8] Zuo T, Tycko B, Liu TM, Lin HJ, Huang TH. Methods in DNA methylation profiling. *Epigenomics* December 2009;1(2):331–45.
- [9] Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* January 6, 2016;5(1):3.
- [10] Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. *Epigenet Chromatin* December 2016;9(1):26.
- [11] Wei LK, Sutherland H, Au A, Camilleri E, Haupt LM, Gan SH, Griffiths LR. Methylenetetrahydrofolate reductase CpG islands: epigenotyping. *J Clin Lab Anal* 2016;30(4):335–44.
- [12] Wei LK, Sutherland H, Au A, Camilleri E, Haupt LM, Gan SH, et al. A potential epigenetic marker mediating serum folate and vitamin B12 levels contributes to the risk of ischemic stroke. *BioMed Res Int* 2015;2015:167976.
- [13] Bock C. Analysing interpreting DNA methylation data. *Nat Rev Genet* 2012;13(10):705–19.
- [14] Dickhaus T. Statistical approaches for epigenetic data analysis. In: Computational epigenetics and disease; 2018.
- [15] Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Research* March 2011;21(3):381.
- [16] Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* November 2014;15(11):703.
- [17] Yang XJ, Seto E. Lysine acetylation: codified crosstalk with other posttranslational modifications. *Mol. Cell* August 22, 2008;31(4):449–61.
- [18] Rossetto D, Avvakumov N, Côté J. Histone phosphorylation: a chromatin modification involved in diverse nuclear events. *Epigenetics* October 13, 2012;7(10):1098–108.
- [19] Smith BC, Denu JM. Chemical mechanisms of histone lysine and arginine modifications. *Biochim Biophys Acta* January 31, 2009;1789(1):45–57.
- [20] Kimura H. Histone modifications for human epigenome analysis. *J Hum Genet* July 2013;58(7):439.
- [21] Chng HR. Computational methods for epigenomics analysis. In: Computational epigenetics and disease; 2018.
- [22] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings Bioinf* November 1, 2016;17(6):953–66.
- [23] Gomes CP, Cho JH, Hood LE, Franco OL, Pereira RW, Wang K. A review of computational tools in microRNA discovery. *Front Genet* May 15, 2013;4:81.
- [24] Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: synthesis, mechanism, function, and recent clinical trials. *Biochim Biophys Acta Mol Cell Res* November 1, 2010;1803(11):1231–43.
- [25] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* January 1, 2009;19(1):92–105.
- [26] Valinezhad Orang A, Safaralizadeh R, Kazemzadeh-Bavili M. Mechanisms of miRNA-mediated gene regulation from common downregulation to mRNA-specific upregulation. *Int J Genomics* 2014;2014.
- [27] Riffó-Campos ÁL, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: what to choose? *Int J Mol Sci* December 9, 2016;17(12):1987.
- [28] Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. *Nat Rev Genet* May 2012;13(5):358.
- [29] Persano S, Guevara ML, Wolfram J, Blanco E, Shen H, Ferrari M, Pompa PP. Label-free isothermal amplification assay for specific and highly sensitive colorimetric miRNA detection. *ACS Omega* September 30, 2016;1(3):448–55.

- [30] Mutalib NS, Jamal R. Computational tools for microRNA target prediction. In: Computational epigenetics and disease; 2018.
- [31] Kang W, Friedländer MR. Computational prediction of miRNA genes from small RNA sequencing data. *Front. Bioeng. Biotechnol.* January 26, 2015;3:7.
- [32] Voisin S. Bioinformatic and biostatistic methods for DNA methylome analysis of obesity. In: Computational epigenetics and disease; 2018.
- [33] Dimova I. Epigenomics of diabetes mellitus (epigenetic regulations in diabetes mellitus). In: Computational epigenetics and disease; 2018.
- [34] Zhou Y, Liu JD, Qian L. Epigenomic reprogramming in cardiovascular disease. In: Computational epigenetics and disease; 2018.
- [35] Chatterjee P, Roy D. Insight into the epigenetics of Alzheimer's disease: a computational study from human interactome. *Curr Alzheimer Res* December 1, 2016;13(12):1385–96.
- [36] Chatterjee P, Roy D, Rathi N. Epigenetic drug repositioning for Alzheimer's disease based on epigenetic targets in human interactome. *J Alzheim Dis* January 1, 2018;61(1):53–65.
- [37] Chatterjee P, Bhattacharyya M, Bandyopadhyay S, Roy D. Studying the system-level involvement of microRNAs in Parkinson's disease. *PLoS One* April 1, 2014;9(4):e93751.
- [38] Chatterjee P, Roy D, Bhattacharyya M, Bandyopadhyay S. Biological networks in Parkinson's disease: an insight into the epigenetic mechanisms associated with this disease. *BMC Genomics* December 2017; 18(1):721.
- [39] Chatterjee R, Das S, Chandra A, Basu B. Epigenome-wide DNA methylation profiles in oral cancer. In: Computational epigenetics and disease; 2018.
- [40] Xu J, Li YS, Shao TT. Computational epigenetics for breast cancer. In: Computational epigenetics and disease; 2018.
- [41] Roznová IA, Ruskin HJ. A computational model for genetic and epigenetic signals in colon cancer. *Interdiscipl Sci Comput Life Sci* September 1, 2013;5(3):175–86.
- [42] Peter M, Kamdar S, Bapat B. Integrative epigenomics of prostate cancer. In: Computational epigenetics and disease; 2018.
- [43] Chen BS. Network analysis of epigenetic data for bladder cancer. In: Computational epigenetics and disease; 2018.

This page intentionally left blank

COMPUTATIONAL METHODS FOR EPIGENOMIC ANALYSIS

2

Ho-Ryun Chung^{1,2}

¹*Epigenomics, Max Planck Institute for Molecular Genetics, Berlin, Germany;* ²*Institute for Medical Bioinformatics and Biostatistics, Philipps-Universität Marburg, Marburg, Germany*

INTRODUCTION

The epigenome comprises covalent modifications of DNA and histone proteins that change and/or reflect chromatin structure and function. In contrast to the constant genome, the epigenome differs between cell types. The epigenetic differences are established during development. Moreover, environmental factors impact on the epigenome leading to changes in cellular function, which in turn contribute to the etiology of diseases.

Epigenomics aims at annotating the epigenome. Such an annotation can be used to unravel functional elements, such as promoters and enhancers, in a cell-type-specific, that is, activity, dependent manner. Moreover, it helps to segment the genome into active and repressed domains. Thus, an epigenomic annotation paves the way for a mechanistic understanding of the cell-type-specific transcriptional program.

Another important aspect of epigenomics is the identification of epigenomic differences between cell types and/or conditions, such as healthy or diseased. Here, the epigenome is interrogated to unravel the changes in the usage of functional elements or positional shifts of boundaries between active and repressed domains. Such an analysis should deepen our understanding about the molecular mechanisms that lead to disease-related changes in the cellular phenotype and may uncover novel avenues for diagnosis and treatment.

Covalent modifications of DNA and histone proteins are measured by approaches that use high-throughput sequencing (1) to localize these modifications in the genome and (2) to determine their occupancy in cell populations. While DNA cytosine methylation is certainly an important epigenetic modification, this chapter will deal only with computational methods to analyze histone modification data, which are generated by chromatin immunoprecipitation followed by sequencing (ChIP-seq; [1]). ChIP-seq recovers the genomic position of histone modifications and their abundance. During ChIP specific antibodies against a histone modification precipitate chromatin fragments carrying the histone modifications. After ChIP the associated DNA is purified and sequenced. The DNA sequences (reads) of these fragments are aligned to a reference genome. In this way both the localization and the abundance of histone modifications can be inferred from the reads' positions and their number along the genome.

This chapter exemplifies integrative analyses of histone modification data. We will explain why a proper normalization against a control sample is required to identify ChIP-enriched regions for histone modifications. Based on these insights, we will show how to identify ChIP-enriched regions. Further, we will demonstrate how to identify functional elements, such as promoters and enhancers, and repressed domains using three chromatin segmentation approaches. Finally, we will exemplify an analysis to unravel chromatin state differences between conditions.

UNBIASED DETECTION OF CHIP-ENRICHMENT

ChIP-seq enriches chromatin fragments that harbor a certain histone modification or protein. It is the principal approach to map histone modifications to the genome. Given ChIP-seq data we want to identify regions occupied by a histone modification: a problem referred to as “peak calling”. Intuitively, a high number of ChIP reads at a given genomic region signals ChIP enrichment—the more ChIP reads the higher the population occupancy. However, due to systematic biases, such as copy number variations and mapping artifacts, the number of ChIP reads is not a direct measure of ChIP enrichment [2–5]. To mitigate these biases the number of ChIP reads is usually compared to a suitable control, for example, an unspecific ChIP using an antibody against IgG or the input to the ChIP. This comparison is rendered difficult due to the variation in sequencing depth and the effect of enrichment during ChIP on the overall read distribution along the genome. In fact, a meaningful comparison between ChIP and control requires normalization, that is, a base line of no ChIP-enrichment, to call ChIP-enrichment. Ideally, normalization should transform the data such that the average ChIP-enrichment in base line regions of no ChIP-enrichment [6–8] is approximately set to unity. Thus, normalization requires the identity of base line regions of no ChIP-enrichment. Naturally, these nonenriched regions are just the inverse of the enriched regions, whose identification requires normalization. In this sense, normalization and ChIP-enrichment calling are the same problem.

The ChIP reads fall into two categories: (1) reads from target regions and (2) reads from background regions. Sequencing depth normalization implicitly assumes that most of ChIP reads are from background regions. However, this assumption is more and more violated if the achieved ChIP enrichment and/or the number of target regions increase. As a consequence the sequencing-depth-estimated normalization factor is too high leading to an overestimation of the background. To illustrate this effect, we simulated two scenarios: a “peaky” enrichment regime, where only a few regions are enriched ([Fig. 2.1](#), left) and a “broad” enrichment regime, where many (consecutive) regions are enriched ([Fig. 2.1](#), right). We fixed the sequencing-depth to a $10\times$ fragment coverage (this corresponds to ~ 160 million uniquely nonduplicated reads in the human genome) in both ChIP and control, set the ChIP-enrichment to tenfold, and the occupancy to 100% ([Fig. 2.1A](#)). Already at the level of read counts it becomes apparent that the target regions show much higher read counts in the “peaky” than in the “broad” case ([Fig. 2.1B](#)). This effect cannot be attributed to a differential sequencing-depth nor to a differential ChIP-enrichment as they are identical in both scenarios. A scatterplot of control versus ChIP read counts reveals that the reduced number of reads in target region in the “broad” case leads to less separation between target and background regions ([Fig. 2.1C](#)). Moreover, the background estimated by the sequencing depth (red line in [Fig. 2.1C](#)) is higher than the background estimated by the background regions (green line). This is less problematic in the “peaky” case because the target regions (orange) are well separated from the background (black) and the difference between the red and the

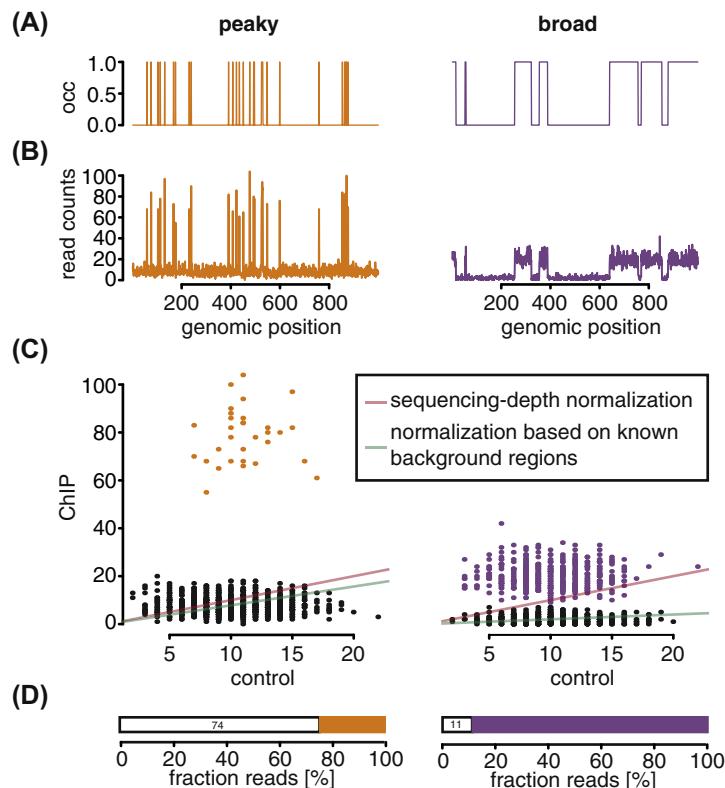


FIGURE 2.1 Proper Normalization is Required for Broadly Distributed Histone Modifications.

Simulated data for a “peaky” (left) and a “broad” case (right) using 1,000 bins. In both simulated ChIPs the enrichment was set to tenfold and for each ChIP as well as the control 10,000 reads were sampled, corresponding to a 10× coverage. (A) Target region occupancy was set to 100%, with 31 target regions for the “peaky” case (orange) and 443 for the broad case (purple). (B) Read counts along the 1,000 bins. (C) Scatterplots for control (x-axis) and ChIP (y-axis) read counts per bin. The orange (purple) dots indicate target regions for the “peaky” (“broad”) case. The red line indicates the background estimated by sequencing-depth normalization and the green line the background estimated on the basis of nontarget regions. (D) Fraction of reads falling into nontarget (open rectangle) and target regions (filled rectangles, orange for the “peaky” case, purple for the “broad” case).

green line is not that large. However, it becomes a problem in the “broad” case. There are target regions (purple) that are below the red line and therefore constitute false negatives. All target regions are above the green line and the background regions (black) above this line can be found in low coverage regions, where the effect size is small and the statistical power low, that is, they are not significantly different from the background. Finally, the larger difference between the red and the green line indicates that the implicit assumption of sequencing-depth normalization that most reads are from background regions is violated. Indeed, while in the “peaky” case ~75% of the reads come from background regions, it is only 11% in the “broad” case (Fig. 2.1D).

Nonetheless, some peak callers account only for sequencing-depth difference in the ChIP- and control samples, for example, MACS2 [9] or DFilter [10]. Others address this problem by “guessing” the background regions based on ad hoc assumption on the data, for example, CisGenome [11], SPP [12] and MUSIC [13]. To our knowledge there are only three methods that take a systematic, data-driven approach to normalization: NCIS [6,8], SES [8], and normR [14].

As illustrated above, a realistic background model is pivotal to uncover ChIP-enriched regions if the number of target regions is high. In our experience this is typically the case when assaying heterochromatic histone modifications, such as H3K9me3 or H3K27me3 that cover large consecutive regions of the genome (domains). The overestimation of the background by the sequencing-depth normalization leads to a diminished sensitivity, that is, many real target regions remain unidentified.

We demonstrate the adverse effect of sequencing-depth normalization by comparing MACS2 (sequencing-depth normalization by linearly scaling the larger to the smaller sequencing depth; [9,15]) and normR (data-driven normalization on inferred background regions; [14]) on a real data set. We ran both algorithms on an H3K4me3 and H3K27me3 data set from the cell line K562 from the Encode Project [16] to uncover ChIP-enriched regions. We ran the call peaks subprogram of MACS for H3K4me3, and for H3K27me3 we additionally set the “broad” option. normR was run using standard setting in both cases. normR uses fixed-sized nonoverlapping bins along the genome. In order to facilitate a comparison we mapped the MACS2 peaks to these bins. We give a Venn diagram (Fig. 2.2A and B) that displays the number of genomic bins that were called enriched (normR) or contain a peak (MACS2) and the overlap between the two methods.

normR identified approximately twice as many H3K4me3-enriched regions (Fig. 2.2A) and approximately thrice as many H3K27me3-enriched regions (Fig. 2.2B) compared to MACS2 at an FDR of 5%. H3K4me3 is usually enriched at few regions in the genome corresponding to promoters [17] or DNA hypomethylated CpG islands [14]. Although the achieved ChIP-enrichment is usually high for H3K4me3, the low number of target regions warrants a high signal-to-noise ratio such that the overestimation of the background poses only a little problem as expected. Even while overestimating the background MACS2 finds many H3K4me3-enriched regions. However, normR finds more H3K4me3 regions, which may originate from targets, where the occupancy and hence the enrichment is lower (Fig. 2.2C).

In differentiated cells H3K27me3 usually forms large domains of facultative heterochromatin. Even if the achieved ChIP-enrichment is high, the high number of target regions leads to an unfavorable signal-to-noise ratio such that the overestimation of the background becomes a problem. The overestimation of the background by MACS2 leads to much fewer called by identified H3K27me3-enriched regions. The 66% regions only identified by normR show a comparable \log_2 ChIP over control ratio to the regions found by both normR and MACS2 (Fig. 2.2D), indicating that they are indeed H3K27me3 targets. By contrast, the regions specific to MACS2 show a much lower \log_2 ChIP over control ratio, sometimes even lower than the genome-wide average \log_2 ChIP over control ratio (red line), indicating a higher proportion of false positives.

In case of H3K4me3 ~9.5 million reads had been assigned to the nonoverlapping 500 base pair bins. In case of H3K27me3 ~12.2 million reads had been assigned. Are these numbers of reads sufficient to warrant a comprehensive calling of H3K4me3 and H3K27me3 target regions? In order to test this, we combined two (biological) replicates for H3K4me3 and H3K27me3 and rerun MACS2 and normR. In case of H3K4me3 doubling the sequencing-depth does not yield substantially more target regions for normR and MACS2 (Fig. 2.2E and G), indicating that for H3K4me3 a single

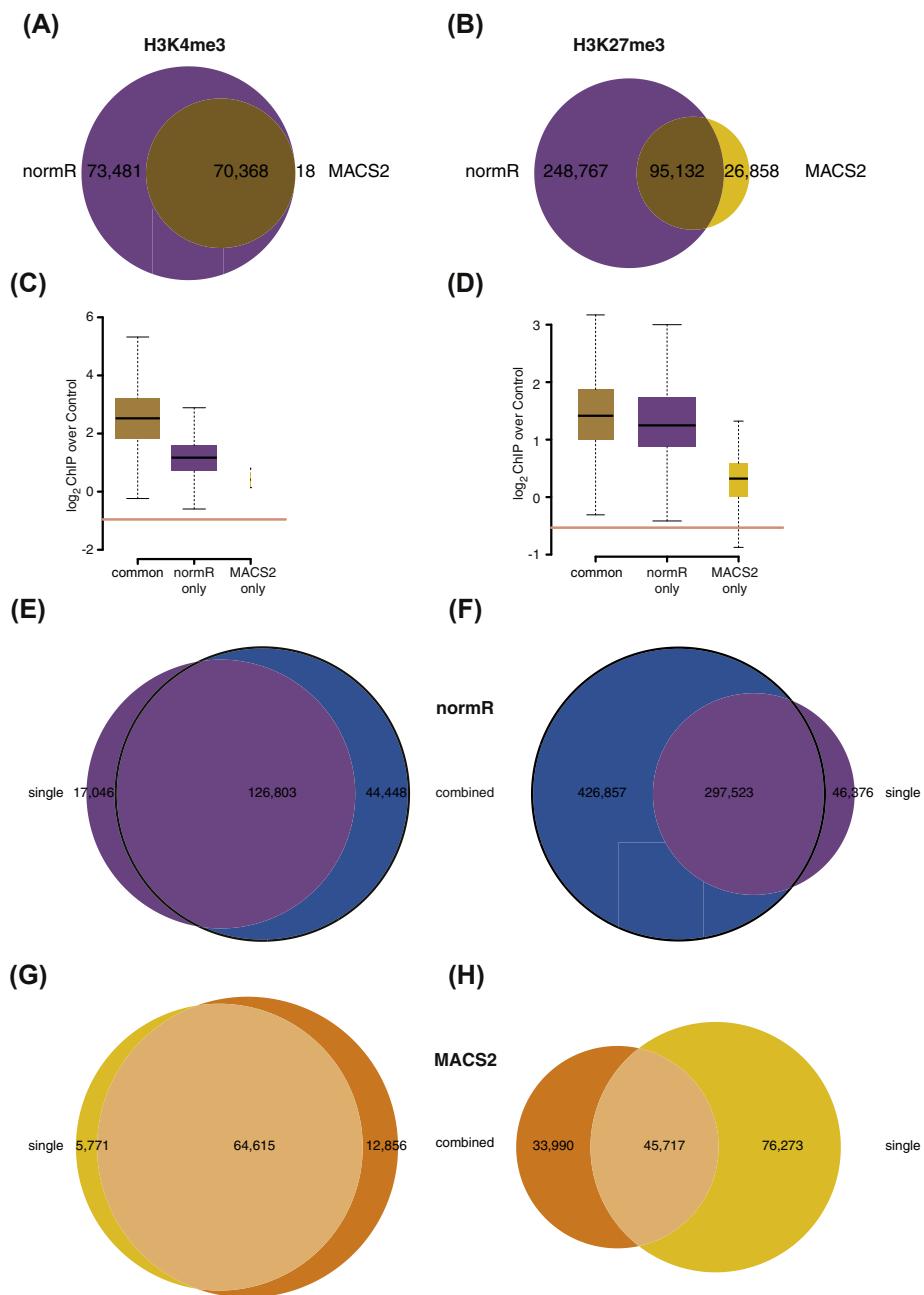


FIGURE 2.2 Proper Normalization Helps Identifying Broadly Distributed Histone Modifications.

The MACS2 calls were mapped back to the nonoverlapping 500 base pair windows used in normR. There were 6,072,654 such 500 base pair windows covering chromosomes 1–22, X, and the mitochondrial genome. If such a window overlapped 250 or more bases of a MACS2 peak it was called positive. (A) Overlap between normR and MACS2 windows for H3K4me3. (B) Overlap between normR and MACS2 windows for H3K27me3. (C) \log_2 H3K4me3 ChIP over control ratio for common (brown), normR-only (purple), and MACS2-only (yellow) 500 base pair windows. The box-widths are scaled to the set size. (D) \log_2 H3K27me3 ChIP over control ratio for common (brown), normR-only (purple), and MACS2-only (yellow) 500 base pair windows. The box-widths are scaled to the set size. (E–H) Overlap between target regions for H3K4me3 (E and G) and H3K27me3 (F and H) using one replicate versus combining two replicates called by normR (E and F) or MACS2 (G and H).

replicate with \sim 10 million reads is already close to saturation. In case of H3K27me3, however, doubling the sequencing depth yields many more target regions in the normR analysis (Fig. 2.2F; 343,899 vs. 724,380), indicating that \sim 12 million reads and possibly also 20 million reads are still far away from saturation. For MACS2 we found that instead of increasing the number target regions the number of 500 base pair windows actually decreased when doubling the sequencing depth.

A recent study highlighted the features that allow some tools to perform better than others [18]. The ChIP peak calling or ChIP enrichment calling problem can be subdivided into two subtasks: (1) the detection of candidate regions and (2) the determination of the statistical significance of the enrichment in these candidate regions. Little interest is paid to the problem of proper normalization, which in our opinion is pivotal to call ChIP-enriched regions. That this problem is more or less ignored stems from the general focus on the “peaky” ChIP-seq enrichment regime, where a proper normalization is not that critical. However, for factors that show a “broad” ChIP-seq enrichment regime a proper normalization is pivotal for the comprehensive identification of ChIP-enriched regions.

There are some peak callers that implement unique features that render them especially suited for certain tasks. For example, GEM combines the identification of ChIP-seq peaks with the discovery of DNA binding site motif [19], which renders GEM especially suited for the identification of *in vivo* transcription factor binding sites with corresponding DNA binding site motif instances. Bayesian Change-Point (BCP; [20]) leverages on recent advances in infinite-state Hidden Markov Models to model the posterior means of read densities and exhaustively searches change points between different posterior means, rendering BCP well suited for the identification of low, domain-like ChIP-enrichment regimes as expected for broad histone modifications. In our experience MACS2 is well suited for factors exhibiting a “peaky” ChIP-enrichment regime but less suited for a “broad” enrichment regime. normR in our hands is well suited for factors with “peaky” and “broad” enrichment regimes.

SEGMENTATION OF THE EPIGENOME INTO CHROMATIN STATES

Most “interesting” genomic features, such as promoters, enhancers, transcription units, and heterochromatic domains, exhibit characteristic and recurrent histone modification patterns referred to as chromatin states [21,22]. For example, the combination of H3K4me3 and H3K27ac occurs at active promoters and H3K4me1 together with H3K27ac is associated to enhancers [23]. These recurrent patterns somewhat decrease the information content in each histone modification track because it is possible to predict missing tracks using the others [24]. However, this also indicates that a joint analysis of all histone modification tracks may help to mitigate problems arising from noisy data. An integrative analysis of histone modification tracks leads to a more robust identification of genomic regions acting as promoters, enhancers, and insulators as well as a demarcation of actively transcribed regions and heterochromatic domains.

Such an integrated analysis of histone modification ChIP-seq data is referred to as chromatin segmentation, which has been made popular by ChromHMM [21]. ChromHMM follows an unsupervised machine learning approach to learn both the chromatin states and their genomic location. In this way, genomic features, such as promoters and enhancers, can be robustly identified despite the inevitable noise in the ChIP-seq measurement and the uncertainty connected to the computational identification of ChIP targets. Moreover, chromatin segmentation leads to a significant reduction in data complexity, easing and enabling the interpretation of epigenomic data.

ChromHMM uses a binarized input, such that for each nonoverlapping 200 base pair bin along the genome every histone modification is either present or absent. To assign the presence or absence of a histone modification ChromHMM uses a threshold on the ratio of ChIP over control reads falling into these bins. ChromHMM models the binarized histone modification data by a multivariate Bernoulli distribution ignoring correlations between histone modifications. Due to its unsupervised training it uncovers histone modification patterns, that is, chromatin states, occurring repeatedly in the data in an unbiased manner.

In the ChromHMM segmentation for histone modification ChIP-seq experiments in the K562 cell line of the Encode consortium [16] there is one chromatin state that accounts for more than half of the genome ([Fig. 2.3A](#), middle, pink). This state is characterized by the absence of all assayed histone modifications ([Fig. 2.3D](#)). This ChromHMM background state may be an artifact of the binarization step, thus corresponding to bins where most of the read counts were below the binarization threshold rather than representing a well-defined chromatin state. A more comprehensive chromatin segmentation can be achieved using EpiCSEg [22]. In contrast to ChromHMM EpiCSEg does not binarizes the histone modification ChIP-seq data but uses the read counts directly. EpiCSEg models the read counts by a negative multinomial distribution. This leads to more sensitivity and enables also the identification of chromatin states with different overall strength mirroring the activity levels of the underlying elements.

The aforementioned ChromHMM background regions correspond to three EpiCSEg states ([Fig. 2.3A](#)). One of them has very low read counts for all histone modifications ([Fig. 2.3B](#)) indicating that they indeed contain no usable information. The other two chromatin states are associated with heterochromatic environments enriched with H3K27me3 and H3K9me3, respectively ([Fig. 2.3B](#)).

The ChromHMM as well as the EpiCSEg segmentations depend strongly on the genomic context, that is, they capture and represent the most important functional elements and biological processes acting on chromatin ([Fig. 2.3C and E](#)). For example, there is a state peaking exactly at the transcriptional start site. Another prominent one peaks downstream of the transcription start site. It becomes apparent that the promoter chromatin state in the EpiCSEg segmentation ([Fig. 2.3C](#), yellow line) is more focused on the transcription start site than the corresponding one of ChromHMM ([Fig. 2.3E](#), yellow line). Moreover, there are many more TSS flagged with the promoter state in the EpiCSEg compared to the ChromHMM segmentation.

In our opinion ChromHMM and EpiCSEg have their individual advantages. The former allows for a better comparability between chromatin segmentations from different experiments and conditions. Furthermore, as ChromHMM has been the standard tool for the Encode, NIH Roadmap and IHEC consortia a new ChromHMM segmentation can be readily compared to a vast number of existing ones. EpiCSEg allows for a more detailed and comprehensive chromatin segmentation. It is less frequently used, such that comparison to chromatin segmentations of other data sets requires computation of an EpiCSEg segmentation first. The comparison between chromatin segmentations by EpiCSEg will be more difficult because it may not be easy to establish the correspondence between the identified chromatin states (see below). This is due to the modeling of read counts rather than a binarized input. The read counts are much more dependent on the experiment through variations in sequencing-depth and ChIP-efficiency than a preprocessed binarized input that mitigates these variations.

In order to overcome this problem of EpiCSEg and to augment the sensitivity of ChromHMM, we propose to use normR as preprocessing step to identify enriched bins for the histone modifications.

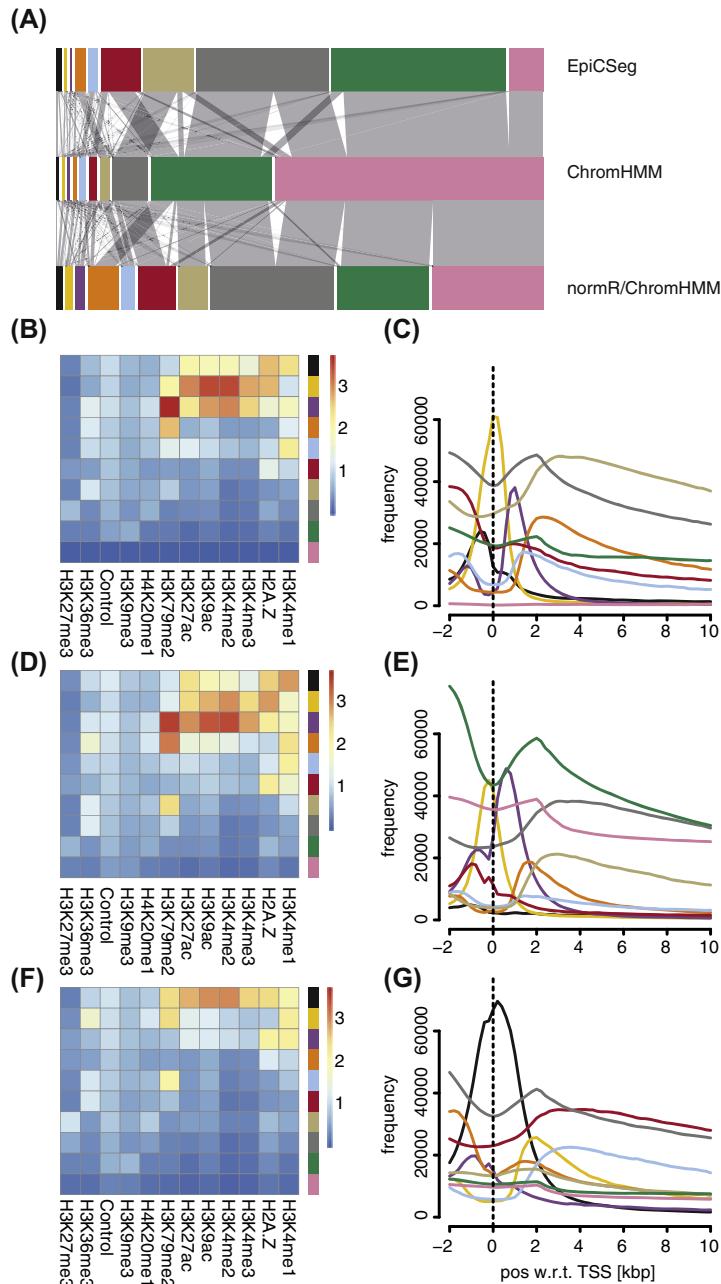


FIGURE 2.3 Chromatin Segmentation.

(A) Comparison of region calls for 10 distinct chromatin states for EpiCSEg (top), ChromHMM (middle), and normR/ChromHMM. The 10 states have distinct colors, where the color between segmentations does not necessarily have to match. The grey shaded areas depict where the chromatin states of ChromHMM end up in EpiCSEg and normR/ChromHMM, respectively. (B, D, and F) $\text{Log}_e(\text{counts} + 1)$ for the states 1 to 10. For EpiCSEg (B), ChromHMM (D), and normR/ChromHMM (F). (C, E, and G) Distribution of the 10 states around the transcriptional start site of Gencode v19 [32] annotated transcriptional start sites for EpiCSEg (C), ChromHMM (E), and normR/ChromHMM (G).

The comparison of such a normR/ChromHMM analysis to ChromHMM and EpiCSEG, respectively, indicates that by preprocessing the histone modification data with normR and then using ChromHMM achieves sensitivity similar to EpiCSEG (Fig. 2.3A, F, and G).

Unsupervised chromatin segmentation eases and enables an integrative analysis of histone modification data. In the future, it may be of interest to integrate DNA methylation data, from for example whole genome bisulfite sequencing, as well as chromatin accessibility data, such as DNase- or ATAC-seq, into the segmentation. Furthermore, chromatin segmentation may benefit from an at least partially supervised training approach to obtain more specific chromatin state assignments.

THE DIFFERENTIAL EPIGENOME

The identification of differences between the epigenomes of healthy versus diseased samples is one of the major goals in biomedical research. Once such differences are identified they may serve as biomarkers for diagnostic purposes. Furthermore, they may help unraveling molecular mechanisms that contribute to the etiology of diseases. Epigenomic differences can be established on the basis of individual histone modifications as well as on the basis of chromatin state changes. For the identification of differences on the basis of individual histone modifications commonly used tools are HistoneHMM [25], Chipdiff [26], Diffreps [27], RSeg [28], Odin [29], and Pepr [30]. The main problem to detect differences between conditions is the variability in ChIP efficiency. For example a dismal ChIP-enrichment in one condition may lead to the false identification of many qualitative as well as quantitative changes. In part this can be avoided by considering replicates, which enable the comparison of the between-condition to the within-condition variance. To our knowledge no systematic study has been performed to objectively assess the performance of the aforementioned tools. Moreover, we strongly believe that comparisons using all available data at the level of chromatin states are more meaningful.

For the identification of chromatin state changes, we will exemplify an approach to identify reproducible chromatin state changes. In order to identify chromatin state changes between conditions we propose the use of the combination of normR preprocessing and ChromHMM (see above). We think that EpiCSEG is less well suited for this purpose (see above). It is difficult to establish a one-to-one correspondence between the read count patterns of histone modifications from distinct experiments, for example, healthy versus diseased, because the read distributions are experiment-specific (variations in ChIP efficiency and sequencing-depth).

As mentioned above, we greatly reduced the size of the ChromHMM chromatin state of no or little signal and can annotate a larger fraction of the epigenome if we use normR instead of the ChromHMM binarization. We aim at obtaining changes in chromatin states (rather than single histone modifications) between conditions, which may be instrumental in discovering novel biomarkers for a disease as well as novel molecular mechanisms/targets for treatments.

To illustrate such an analysis, we compared the K562 and GM12878 cell line at the level of chromatin states. We used histone modification ChIP-seq data from the Encode consortium [16]. In order to achieve maximal comparability, we concatenated the four epigenomes, that is, two biological replicates for K562 and GM12878, respectively. The resulting input matrix has therefore four times the number of 500 base pair bins of a single genome with the same number of histone modifications, namely H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K79me2,

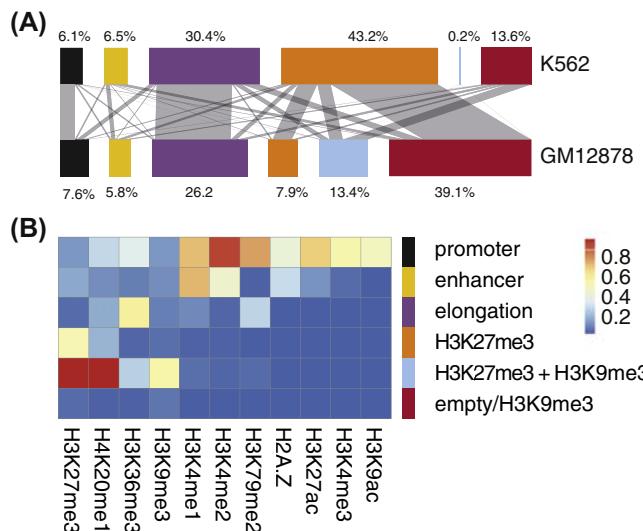


FIGURE 2.4 Identification of Differential Chromatin States.

(A) Comparison of region calls for six distinct chromatin states jointly learned from two replicates each for K562 and GM12878. Shown are regions that had the same chromatin state in both replicates. We removed regions that were assigned to the empty/H3K9me3 state in K562 and GM12878. (B) The probability to find an enriched histone modification in the six different states. A tentative assignment to a possible function of the chromatin state is given at the right.

H4K20me1, and H2A.Z. It contains for each histone modification either a zero indicating no enrichment or a one indicating enrichment. We ensured for each histone modification and each sample an equal number of enriched bins.

After chromatin segmentation using six chromatin states we selected bins that had the same state in both replicates in K562 and GM12878, respectively, and discarded the rest. Overall there is a good correspondence between K562 and GM12878 at the level of promoter-like and elongation states (Fig. 2.4, black and purple). There is a considerable change at the level of enhancer-like states as expected for different cell lines (Fig. 2.4, yellow). Most changes, however, are related to the heterochromatic states (Fig. 2.4, orange, blue, and red). In particular the H3K27me3 + H3K9me3 state (Fig. 2.4, blue) is present in GM12878 but almost absent in K562. Most of the regions that are annotated by this state in GM12878 either become H3K27me3 only (Fig. 2.4, orange) or empty/H3K9me3 (Fig. 2.4, red) in K562. One explanation for such a qualitative change is that the H3K9me3 ChIP in GM12878 had a better quality than in K562. We think that this is less likely as the number of H3K9me3 enriched bins was much higher in K562 than in GM12878 (448,279 vs. 93,842) after controlling for the Irreproducible Discovery Rate [31] at 1%. Alternatively, the qualitative difference between GM12878 and K562 regarding the H3K27me3 + H3K9me3 state may be related to their distinct origins. K562 cells were derived from a patient with chronic myelogenous leukemia, while GM12878 cells were immortalized using the Epstein–Barr virus transformation. Thus, this epigenomic difference may be due to different routes to establish immortalized cell lines.

Taken together we think that the combination of normR preprocessing and ChromHMM chromatin segmentation is a valuable tool to identify chromatin state changes between conditions. Such an integrated analysis may pave the way to identify meaningful changes in the epigenome that are involved in the etiology of diseases.

REFERENCES

- [1] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* June 8, 2007;316(5830):1497–502.
- [2] Vega VB, Cheung E, Palanisamy N, Sung W-K. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS One* 2009;4(4):e5241.
- [3] Jain D, Baldi S, Zabel A, Straub T, Becker PB. Active promoters give rise to false positive “Phantom Peaks” in ChIP-seq experiments. *Nucleic Acids Res* August 18, 2015;43(14):6959–68.
- [4] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* November 2014;15(11):709–21.
- [5] Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci USA* November 12, 2013;110(46):18602–7.
- [6] Liang K, Keles S. Normalization of ChIP-seq data with control. *BMC Bioinf* 2012;13:199.
- [7] Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei C-L, et al. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* May 1, 2010;26(9):1199–204.
- [8] Diaz A, Park K, Lim DA, Song JS. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol* 2012;11(3). Article 9.
- [9] Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, et al. Model-based analysis of chip-seq (MACS). *Genome Biol* 2008;9(9):R137.
- [10] Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol* June 16, 2013;31(7):615–22.
- [11] Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotech* November 2008;26(11):1293–300. Nature Publishing Group.
- [12] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* December 2008;26(12):1351–9.
- [13] Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* 2014;15(10):474.
- [14] Kinkley S, Helmuth J, Polansky JK, Dunkel I, Gasparoni G, Fröhler S, et al. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells. *Nat Commun* 2016;7:12514.
- [15] Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* September 2012;7(9):1728–40.
- [16] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* September 6, 2012;489(7414):57–74.
- [17] Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* July 13, 2007;130(1):77–88.
- [18] Thomas R, Thomas S, Holloway AK, Pollard KS. Features that define the best ChIP-seq peak calling algorithms. *Brief Bioinform* May 1, 2017;18(3):441–50.
- [19] Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* 2012;8(8):e1002638.

- [20] Xing H, Mo Y, Liao W, Zhang MQ. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS Comput Biol* 2012;8(7):e1002613.
- [21] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* August 2010;28(8):817–25.
- [22] Mammana A, Chung H-R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol* 2015;16:151.
- [23] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* March 2007;39(3):311–8.
- [24] Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* April 2015;33(4):364–76.
- [25] Heinig M, Colomé-Tatché M, Taudt A, Rintisch C, Schafer S, Pravenc M, et al. histoneHMM: differential analysis of histone modifications with broad genomic footprints. *BMC Bioinform* 2015;16:60.
- [26] Xu H, Wei C-L, Lin F, Sung W-K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* October 15, 2008;24(20):2344–9.
- [27] Shen L, Shao N-Y, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One* 2013;8(6):e65598.
- [28] Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* March 15, 2011;27(6):870–1.
- [29] Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics* December 15, 2014;30(24):3467–75.
- [30] Zhang Y, Lin Y-H, Johnson TD, Rozek LS, Sartor MA. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics* September 15, 2014;30(18):2568–75.
- [31] Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* September 1, 2011;5(3):1752–79. Institute of Mathematical Statistics.
- [32] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res* September 2012;22(9):1760–74.

STATISTICAL APPROACHES FOR EPIGENETIC DATA ANALYSIS

3

Thorsten Dickhaus

Institute for Statistics, University of Bremen, Bremen, Germany

INTRODUCTION

Establishing reliable statistical models for epigenetic data is often hard, for instance due to the rapid technological developments of epigenetic measurements (see, e.g., Ref. [1] for a novel technology to quantify tumor-infiltrating T-lymphocytes). Each new technology leads to specific characteristics of the resulting data and their distributions, making it very difficult to verify parametric model assumptions. This is the reason why we propose nonparametric statistical models for epigenetic data. Such models do not require strong assumptions about the data-generating mechanisms. Thus, the resulting data analysis methods can easily be transferred from one technological platform to the other.

A second issue is raised by the (often strong) dependencies among epigenetic measurements. These dependencies can have different reasons, for instance (i) biological reasons such as linkage disequilibrium for markers with a small epigenetic distance, (ii) functional reasons such as the presence of cell type specific markers in blood samples, and (iii) application-based reasons such as the consideration of different (derived) epigenetic parameters referring to the same measurements. This is why we propose multivariate statistical models for epigenetic data. Such models implicitly account for the aforementioned dependencies. Thus, they are much more realistic than marginal models, and they can exploit the dependency structure in the statistical analysis. This is often important for an optimal performance of the resulting data analysis methods, especially for applications with high-dimensional data.

Based on the proposed models, we derive statistical methodology (multiple tests) for the detection of coordinate-wise distributional differences in multivariate epigenetic two-group data. This is particularly relevant in the disease context, because it can be used as a first step in deriving a diagnostic model for the disease under consideration based on epigenetic profiles.

The rest of the material is structured as follows. [The second section](#) deals with the statistical modeling of multivariate epigenetic data. In [the third section](#), the statistical methodology under the proposed model class is elaborated. [The fourth section](#) is devoted to the analysis of real data, and we conclude with a discussion of potential extensions of our proposed approach in [the fifth section](#).

STATISTICAL MODELING

We consider a specific class of statistical models, namely, two-group models with stochastically independent, multivariate observables. This means that we observe d -dimensional random vectors, where the dimension $d \in \mathbb{N}$ refers to the number of epigenetic loci, markers, or parameters, respectively, which are under consideration. In the remainder, we will use the term “coordinate” l to refer to a particular locus, parameter, or marker, respectively, where $1 \leq l \leq d$.

Model 2.1. *We denote the two experimental groups by A and B, and we consider $N \in \mathbb{N}$ observational units with n_A observables in group A and n_B in group B, such that $N = n_A + n_B$. We assume that all N observables are stochastically independent and that the observations in group $i \in \{A, B\}$ are realizations of independent and identically distributed (i.i.d.) d -dimensional random vectors*

$\mathbf{X}_{ik} = (X_{ik}^{(1)}, \dots, X_{ik}^{(d)})^\top$, where the index $i \in \{A, B\}$ denotes the group and $1 \leq k \leq n_i$ indexes the k -th observational unit within group i , while the superscript $1 \leq l \leq d$ denotes the coordinate. The random vectors are assumed to follow the group-specific distribution (law) $\mathcal{L}(\mathbf{X}_{A1}) = P$ or $\mathcal{L}(\mathbf{X}_{B1}) = Q$, respectively.

The assumption of stochastically independent observables corresponds to the assumption of having a representative sample from the target population of interest, meaning that the observational units have been sampled at random. The assumption of identical distributions within each group $i \in \{A, B\}$ means that the observational units in group i are homogeneous with respect to the epigenetic measurements to be analyzed. The reason for considering multivariate distributions P and Q is that such models are capable of exploiting the (often strong) dependencies among the coordinates in the statistical analysis (see our respective remarks in the Introduction section). In the multiple testing context, this leads to an optimization of statistical power to detect coordinate-wise group differences. This is due to the fact that a multiplicity correction has to be performed, which can often be relaxed under strong dependency (see e.g., Refs. [2,3] or Section 3.2 of Ref. [4]). Finally, notice that Model 2.1 is a nonparametric model, because the distributions P and Q are not restricted to parametric families (for example, the family of normal distributions), but are arbitrary distributions on \mathbb{R}^d . In [the next section](#), we will only impose mild qualitative assumptions such as the existence of finite second moments (variances and covariances) under P and Q .

Let us discuss two specific examples to which our basic Model 2.1 applies.

Example 2.1 (Identifying differentially methylated cytosine phosphate guanine (CpG) loci). *Consider an epigenetic methylation data set comprising d CpG loci. For each locus l , each group i , and each observational unit k in group i , a methylation ratio (occasionally referred to as β value) is defined as*

$$X_{ik}^{(l)} = \frac{M_{ik}^{(l)}}{M_{ik}^{(l)} + U_{ik}^{(l)}}, \quad (3.1)$$

where $M_{ik}^{(l)}$ ($U_{ik}^{(l)}$) is an intensity value for the amount of methylated (unmethylated) cells, assuming that suitable preprocessing steps have been performed prior to the statistical analysis. In previous literature the family of beta distributions has been considered as a model for the distribution of $X_{ik}^{(l)}$ (see, e.g., Ref. [5]). However, often bimodality and skewness are encountered in CpG methylation data, questioning this parametric assumption. Notice also that numerator and denominator in (3.1) are

highly dependent. As we are not aware of a parametric model capturing these distributional characteristics, we propose a nonparametric approach as in Ref. [6].

Example 2.2 (Group differences for interrelated immune relevant parameters) . As a second example, consider the comparison of human colorectal tissue in cancer patients and healthy controls. In Section “Association of immune cell counts with cancer” of Ref. [7], we analyzed data from a study in which three highly dependent immune relevant parameters were measured utilizing novel epigenetic markers based on methylation signatures in tissue. Since no prior information about distributional properties of these marker data was at hand, our nonparametric approach was applied, only making use of our basic model assumptions. In this context, $d = 3$ refers to the three interrelated parameters, and the groups A and B refer to patients and controls, respectively. The data for observational unit k in group i are in this example of the following structure:

$X_{ik}^{(1)}$ = the number of regulatory T cells,

$X_{ik}^{(2)}$ = the total number of T cells,

$X_{ik}^{(3)}$ = the cellular ratio of immune tolerance.

STATISTICAL METHODOLOGY

The aim of the statistical analysis is to detect coordinates for which the (marginal) distributions of the observables differ between the two groups. In the disease context, the group indicates the disease status (healthy/diseased) of each observational unit (typically, a study participant). The detection of distributional differences between the groups will be formalized by means of a multiple statistical hypothesis test problem, meaning that d statistical tests (one per coordinate) have to be performed simultaneously on the basis of one and the same data set. Coordinates for which the null hypothesis of no distributional difference between the groups is rejected may be considered in a subsequent analysis in order to build a prognostic (regression) model for the disease at hand based on epigenetic features.

FORMULATION OF MULTIPLE TEST PROBLEMS

We denote by F_i the cumulative distribution function (cdf) of \mathbf{X}_{i1} with marginal cdfs $F_i^{(l)}$ for each coordinate $1 \leq l \leq d$, where $i \in \{A, B\}$. We are interested in testing two families of marginal hypotheses, say $\mathcal{H} = (H_l : 1 \leq l \leq d)$ and $\mathcal{H}' = (H'_l : 1 \leq l \leq d)$. The family \mathcal{H} corresponds to marginal homogeneity in the sense of Ref. [8]. This means, one is interested in testing which of the coordinate-specific marginal distributions are the same in both groups A and B and which differ, i.e.,

$$H_l: F_A^{(l)} = F_B^{(l)} \quad \text{versus} \quad K_l: F_A^{(l)} \neq F_B^{(l)}.$$

While considering \mathcal{H} may appear straightforward, it is worth noticing that the rejection of a particular null hypothesis H_l in favor of the corresponding alternative K_l does not yield any information about the type of distributional difference between the groups in coordinate l . Many such types are possible such as (i) the observations in group A tend to larger (or smaller) values than the observations in group B, (ii) the observations in group A tend to be more (or less) spread out than the observations in group B, or (iii) the distribution of the observables in group A appears to be more (or less) skewed than the distribution of the observables in group B. In many applications, distributional

differences of type (i) are of utmost importance. For instance, in the context of Example 2.1 the goal of the statistical analysis will often be to detect coordinates l for which the methylation intensity is systematically (and significantly) larger in one of the two groups. However, such a conclusion cannot be drawn from the rejection of H_l .

Therefore, we consider a second family \mathcal{H}' of hypotheses, which corresponds to finding a particular type of coordinate-specific distributional differences. Namely, one is interested in detecting coordinates in which there are group differences in the central tendencies of the marginal distributions. To this end, recall the definition of the relative effect in the sense of Ref. [9].

Definition 3.1 (Relative effect). *Let X_A and X_B denote two stochastically independent random variables which are defined on a common probability space with probability measure \mathbb{P} . Assume that X_A and X_B have nondegenerate distributions and denote the normalized version of their cdf, as considered in Ref. [10], by F_A and F_B , respectively. Then, the relative effect of F_A with respect to F_B is defined as*

$$p_{AB} = \mathbb{P}(X_A < X_B) + \frac{1}{2}\mathbb{P}(X_A = X_B) = \int F_A dF_B.$$

For a d -variate distribution the relative effects can be defined coordinate-wise for each $1 \leq l \leq d$ by

$$p_{AB}^{(l)} = \int F_A^{(l)} dF_B^{(l)}.$$

Let $\mathbf{p}_{AB} = (p_{AB}^{(1)}, \dots, p_{AB}^{(d)})^T$ denote the vector of marginal relative effects in the latter case. The functional $p_{AB}^{(l)}$ is capturing central tendencies in coordinate l , that is, whether realizations from one of the group-specific marginal distributions in coordinate l are tending to larger values than the ones from the other. Hence, we let

$$H_l : p_{AB}^{(l)} = 1/2$$

with two-sided alternatives $K'_l : p_{AB}^{(l)} \neq 1/2$, for $1 \leq l \leq d$.

Let $S \subseteq \{1, \dots, d\}$. In the remainder, we make use of the notation

$$H_S = \bigcap_{l \in S} H_l, \quad H_0 = H_{\{1, \dots, d\}} = \bigcap_{l=1}^d H_l,$$

and refer to H_0 as the global hypothesis in \mathcal{H} . An analogous notation applies for intersection hypotheses in \mathcal{H}' .

TEST STATISTICS AND THEIR LIMITING NULL DISTRIBUTIONS

For the univariate nonparametric two-sample problem, that is, for testing one particular hypothesis H_l , Wilcoxon's rank sum test (or, equivalently, the Mann–Whitney U test) is commonly applied. We make use of multivariate generalizations described in Ref. [11] (for testing \mathcal{H}) and in Ref. [12] (for testing \mathcal{H}').

Definition 3.2 (Mann–Whitney U -statistic). *For $1 \leq l \leq d$, we let*

$$U^{(l)} = \frac{1}{n_A n_B} \sum_{j=1}^{n_A} \sum_{k=1}^{n_B} \phi^{(l)}(\mathbf{X}_{Aj}, \mathbf{X}_{Bk})$$

with $\phi^{(l)}(\mathbf{X}_{Aj}, \mathbf{X}_{Bk}) = \mathbb{I}\{\mathbf{X}_{Aj}^{(l)} > \mathbf{X}_{Bk}^{(l)}\}$, where \mathbb{I} denotes the indicator function.

According to Theorem 2 (iii*) of Ref. [11], the d -dimensional random vector $\mathbf{U}_N = \sqrt{N} \left(U^{(1)} - \frac{1}{2}, \dots, U^{(d)} - \frac{1}{2} \right)$ is, under mild regularity assumptions, asymptotically normally distributed under H_0 with mean zero and covariance matrix Σ . Hence, a suitable test statistic W_N^U for testing H_0 is given in the following corollary.

Corollary 3.1 (Theorem 9.1 in Ref. [11]). *Let $\widehat{\Sigma}$ be a consistent estimator of Σ . Assuming that $\det(\Sigma) > 0$ it holds that*

$$W_N^U = N \left(\mathbf{U}_N - \frac{1}{2} \mathbf{1}_d \right)^T \widehat{\Sigma}^{-1} \left(\mathbf{U}_N - \frac{1}{2} \mathbf{1}_d \right)$$

is under H_0 asymptotically chi-square distributed with d degrees of freedom as $N \rightarrow \infty$, where $\mathbf{1}_d = (1, \dots, 1)^T \in \mathbb{R}^d$.

Based on Corollary 3.1 and letting χ_d^2 denote the chi-square distribution with d degrees of freedom, a test at asymptotic ($N \rightarrow \infty$) level α for H_0 is given by $\varphi_0 = \mathbb{I}\{W_N^U > c_{\alpha,d}\}$, where the critical value $c_{\alpha,d}$ is chosen as the $(1 - \alpha)$ -quantile of χ_d^2 , for $\alpha \in (0, 1)$.

The empirical counterpart of the vector \mathbf{p}_{AB} of relative effects is denoted by $\widehat{\mathbf{p}}_{AB} = (\widehat{p}_{AB}^{(1)}, \dots, \widehat{p}_{AB}^{(d)})^T$ with $\widehat{p}_{AB}^{(l)} = \int \widehat{F}_A^{(l)} d\widehat{F}_B^{(l)}$, $1 \leq l \leq d$, where $\widehat{F}_i^{(l)}$, given by $\widehat{F}_i^{(l)}(x) = n_i^{-1} \sum_{k=1}^{n_i} \frac{1}{2} \left(\mathbb{I}_{(-\infty, x]}(X_{ik}^{(l)}) + \mathbb{I}_{(-\infty, x]}(X_{ik}^{(l)}) \right)$, denotes the normalized version of the empirical cdf in group $i \in \{A, B\}$ pertaining to coordinate l . Theorem 3.3 in Ref. [12] yields mild conditions for asymptotic normality of $\mathbf{T}_N = \sqrt{N}(\widehat{\mathbf{p}}_{AB} - \mathbf{p}_{AB})$, where the limiting d -variate normal distribution has mean zero and covariance matrix V . In analogy to Corollary 3.1, we thus obtain a test statistic W_N for testing the global hypothesis $H'_0 : \mathbf{p}_{AB} = \mathbf{1}_d/2$ in \mathcal{H}' .

Corollary 3.2 *Assuming V to be positive definite and the availability of a consistent estimator \widehat{V}_N of V , it follows that, under $H'_0 : \mathbf{p}_{AB} = \mathbf{1}_d/2$, the statistic*

$$W_N = N \left(\widehat{\mathbf{p}}_{AB} - \frac{1}{2} \mathbf{1}_d \right)^T \widehat{V}_N^{-1} \left(\widehat{\mathbf{p}}_{AB} - \frac{1}{2} \mathbf{1}_d \right)$$

is asymptotically χ_d^2 distributed as $N \rightarrow \infty$.

Remark 3.1. *Consistent estimators $\widehat{\Sigma}$ (see Corollary 3.1) and \widehat{V}_N (see Corollary 3.2) have been discussed in Ref. [7].*

MULTIPLE TEST PROCEDURES: CLOSURE PRINCIPLE

Corollaries 3.1 and 3.2 only provide us with a test statistic for the (single) global hypothesis H_0 or H'_0 , respectively, while our initial goal was to test every individual hypothesis H_l or H'_l , respectively, where $1 \leq l \leq d$. To this end, it is essential to notice that the latter corollaries remain valid if the respective full

d -dimensional vector \mathbf{U}_N or \mathbf{T}_N , respectively, is replaced by a subvector which only contains the indices in a given subset $S \subseteq \{1, \dots, d\}$.

In the assertions of the corollaries, only the degrees of freedom of the asymptotic χ^2 -distributions have to be changed from d to $|S|$ if such a subvector is under consideration.

This fact allows us to apply the closure principle (see Ref. [13]). The general idea behind this method is to add to the null hypotheses of interest in \mathcal{H} (or \mathcal{H}') all their intersections H_S or H'_S , respectively, where $S \in 2^{\{1, \dots, d\}} \setminus \emptyset$, with $2^{\{1, \dots, d\}}$ denoting the powerset of $\{1, \dots, d\}$. Even if these intersection hypotheses are not of scientific interest, they are tested auxiliarly in order to provide a multiplicity correction. Namely, a closed test procedure tests every such intersection hypothesis at level $\alpha \in (0, 1)$ by a level α test φ_S or φ'_S , respectively. The adjustment for multiplicity is then performed via the decision rule that only those coordinate-specific hypotheses H_l or H'_l , respectively, are rejected for which all intersection hypotheses H_S (H'_S) with $l \in S$ have been rejected by $\varphi_S(\varphi'_S)$. Thus, the price to pay for the multiplicity of the problem is that one has to perform $2^d - 1$ tests. A concise description of this principle, together with further references, can for instance be found in Section 3.3 of Ref. [4]. It yields strong control of the so-called family-wise error rate (FWER), meaning that the probability of at least one individual type I error is bounded by α .

In our case, based on Corollaries 3.1 and 3.2, we have to perform $2^d - 1$ chi-square tests $\varphi_S(\varphi'_S)$. However, it is possible to speed up computations by exploiting the stochastic representation of a χ^2 -distributed random variable and precomputing terms which appear repeatedly in several of the test statistics corresponding to the tests $(\varphi_S)_S$ or $(\varphi'_S)_S$, where S varies among all nonempty subsets of $\{1, \dots, d\}$.

FINITE SAMPLE MODIFICATION: STUDENTIZED PERMUTATION APPROACH

Corollaries 3.1 and 3.2 rely on asymptotic considerations (i.e., limit theorems) with the sample size N tending to infinity. However, in practice (especially in the disease group) there will often be limitations regarding the obtainable sample size, questioning the applicability of statistical methods based on asymptotics.

To address this issue, it is possible to utilize a different null distribution (instead of χ_d^2) for the test statistic W_N or W_N^U , respectively, which is based on a Studentized permutation approach (see e.g., Refs. [14–16] and references therein). In an algorithmic manner, we may describe this approach for W_N as follows.

Algorithm 3.1

1. Denote by $\mathbf{X} = (\mathbf{X}_{A1}, \dots, \mathbf{X}_{An_A}, \mathbf{X}_{B1}, \dots, \mathbf{X}_{Bn_B})$ the originally observed data matrix with values in $\mathbb{R}^{d \times N}$, where each column corresponds to one observational unit.
2. Let π denote an arbitrary, but fixed permutation of $\{1, \dots, N\}$, and consider the matrix $\mathbf{X}^\pi = (\mathbf{X}_{A1}^\pi, \dots, \mathbf{X}_{An_A}^\pi, \mathbf{X}_{B1}^\pi, \dots, \mathbf{X}_{Bn_B}^\pi)$ which contains the permuted column vectors from \mathbf{X} . Typically, there will be some vectors among the first n_A columns of \mathbf{X}^π which have indeed originated from group B (i.e., which have occupied a column in \mathbf{X} with first index B). Furthermore, let $\hat{\mathbf{p}}_{AB}^\pi$ denote the estimator of the vector of relative effects based on the permuted data set \mathbf{X}^π . Analogously, let \hat{V}_N^π denote the estimator of V mentioned in Remark 3.1, but now applied to \mathbf{X}^π .
3. Let

$$W_N^\pi = N(\hat{\mathbf{p}}_{AB}^\pi - \mathcal{P}_{AB}^\pi)^T (\hat{V}_N^\pi)^{-1} (\hat{\mathbf{p}}_{AB}^\pi - \mathcal{P}_{AB}^\pi),$$

where $\mathcal{P}_{AB}^\pi = \tau(1 + n_A/n_B)1_d/2 + [1 - \tau(1 + n_A/n_B)]\hat{\mathbf{p}}_{AB}$ with $\tau = \tau(\pi, n_A, n_B)$ denoting the fraction of observations from group B within the first n_A columns of \mathbf{X}^π .

4. Approximate the null distribution of W_N by the permutation distribution of W_N^π , that is, by the discrete distribution induced by letting π be uniformly distributed on all $N!$ possible permutations of the set $\{1, \dots, N\}$, while keeping the data X fixed.

Analogously, the respective permutation distribution can be used instead of $\chi^2_{|S|}$ in order to calibrate each test ϕ'_S in the closed test procedure described in Section 3.3 for type I error control at level α . Furthermore, we may proceed in an analogous manner in the case that W_N^U is considered instead of W_N .

By means of computer simulations, it has been demonstrated in Ref. [7] that the permutation approach often leads to a much more accurate calibration of the multiple tests than the asymptotic chi-square approach, especially if N is small to moderate. This means that the permutation approach keeps the FWER under control, even for finite sample sizes, while the chi-square approach is asymptotic in nature and is prone to violate the FWER in the case of small or moderately large samples. Also, the advantage of the permutation-based calibration becomes more and more pronounced with growing dimension d . However, the necessity to carry out the permutations increases the computational effort, in addition to the blowup of the family of null hypotheses caused by the closed testing approach. Thus, in the case of large d our proposed methodology may be infeasible, depending on the available computational resources.

Remark 3.2

- (a) In practice, it is not necessary to carry out all $N!$ permutations explicitly. First, notice that all such permutations are equivalent for which the same n_A observables occupy the first n_A columns in \mathbf{X}^π . Second, the exact permutation distribution can be approximated by a Monte Carlo variant, where only a limited number of randomly chosen permutations of $\{1, \dots, N\}$ are taken into account.
- (b) Our simulations have also indicated that the statistical power of the proposed nonparametric approach is competitive with that of parametric approaches. Hence, especially if the data are not explicitly modeled prior to the statistical analysis (based on expert knowledge, for example), we recommend the nonparametric approach, because it is robust against model misspecification.

REAL DATA ANALYSIS

The UK Ovarian Cancer Population Study (see Ref. [17]) aimed at detecting differentially methylated CpG loci between ovarian cancer cases and healthy controls (GEO accession number GSE19711; see our Example 2.1 for a description of the data structures). To this end, 274 healthy controls were compared with 131 untreated, confirmed ovarian cancer cases. After quality control, data of 264 controls and 124 cases remained for the statistical analysis. Since the total number of considered loci in the study would be too large in order to carry out the multivariate test procedures described in Section 3, and since strong control of the family-wise error rate is considered as a too stringent criterion for large d , we mimicked a two-stage study design consisting of a screening stage (where all loci are under consideration) and a confirmation stage (where only a limited number of loci, which have been

Table 3.1 Multiplicity-Adjusted *P* Values of the Tests for Relative Effects for the Loci Selected at the Screening Stage, for Data Taken From [17]

Locus	<i>cg00645579</i>	<i>cg00974864</i>	<i>cg02679745</i>	<i>cg08044694</i>	<i>cg09134726</i>
χ^2	0.0046	0.0002	0.0002	0.0002	0.0002
Perm	0.0126	0.0029	0.0029	0.0029	0.0029
Locus	<i>cg09303642</i>	<i>cg09305224</i>	<i>cg20070090</i>	<i>cg24427660</i>	<i>cg24777950</i>
χ^2	0.0002	0.0047	0.0001	0.0002	0.0002
Perm	0.0029	0.0146	0.0029	0.0076	0.0029

The *p*-values are based on the asymptotic χ^2 multiple test (χ^2) and the multiple permutation test (Perm), respectively, in combination with the closure principle. The multiplicity-adjusted *p*-value for locus 1 denotes the smallest significance level such that H'_{l1} is rejected for the actually observed data. The permutation test was carried out as a Monte Carlo permutation test employing 9,999 randomly chosen permutations of $\{1, \dots, N\}$, together with the identity permutation.

selected in the screening stage, are tested). To guarantee independence of the data from screening and from confirmation stage, we randomly split the original sample, such that the screening data set comprised 176 controls and 84 cases, while the confirmation data set consisted of the remaining 88 controls and 40 cases. By means of simple univariate analyses, we selected $d = 10$ loci based on the screening data to be carried over to the confirmation stage. In this confirmation stage, the 10 selected loci were tested for a relative effect being unequal to 1/2 based on asymptotic critical values from the limit distribution (χ^2) and permutation-based critical values (Perm), as described in the third section. Since the number $d = 10$ of null hypotheses to be tested simultaneously is rather small here, control of the FWER is an appropriate type I error criterion in this context. In Table 3.1, the results are presented as multiplicity-adjusted *P* values. For locus $1 \leq l \leq 10$, the multiplicity-adjusted *P* value denotes the smallest FWER level such that H'_{l1} is rejected by the respective multiple test procedure. It becomes apparent that the more accurate control of the FWER achieved by Perm leads to a reduced statistical power (in terms of larger *P* values), compared to χ^2 . However, in this example all 10 candidate CpG sites have a multiplicity-adjusted *P* value below 5% for both methods. This is an FWER level which is often chosen in practice.

DISCUSSION

We have presented a statistical methodology by means of which group differences in epigenetic data can reliably be detected (see, in addition to the data analysis in Section 4, Refs. [18] and [19], for example), without having to rely on (often restrictive) parametric model assumptions. Essentially, the only assumptions made in Model 2.1 refer to the sampling scheme, namely, that the observables are sampled randomly and independently, and that the two considered groups are distributionally homogeneous with respect to the epigenetic profiles in each group.

There are several potential extensions of the proposed approach. First, it is straightforward to generalize the methodology to more than two groups (corresponding, for example, to different stages of severity of the disease under consideration). In a nutshell, the nonparametric tests of Wilcoxon- and Mann–Whitney type have in such a case to be replaced by tests of Kruskal–Wallis type (see, e.g., Section 3 of Ref. [20]). Second and less straightforwardly, it may be worthwhile to consider the case of

a large dimension d in depth, in particular with respect to control of the false discovery rate (FDR; see Ref. [21]), which is nowadays a quasi-standard type I error criterion for large-scale multiple test problems. Changing the type I error criterion from FWER control to FDR control typically leads to an increase in statistical power and more rejections (at least on average). On the other hand, the FDR is defined as the expected proportion of type I errors among all rejections. This means that controlling the FDR typically implies to allow for a few individual type I errors, while FWER control aims at avoiding any individual type I errors with probability at least $1 - \alpha$. Resampling-based FDR control has been considered in Refs. [22–24] and [25], among others. However, their proposed methods would have to be adapted to the nonparametric Model 2.1. A permutation-based approach for multiple testing in epigenome-wide association studies has recently been proposed by Saffari et al. [26].

ACKNOWLEDGMENTS

The author is grateful to Sven Olek, Konstantin Schildknecht, Karsten Tabelow, and all other cooperation partners at Epiontis GmbH for inspiring discussions. Special thanks are due to Loo Keat Wei for the fruitful cooperation in this book project.

REFERENCES

- [1] Sehouli J, Loddenkemper C, Cornu T, Schwachula T, Hoffmüller U, Grütkau A, Lohneis P, Dickhaus T, Gröne J, Kruschewski M, et al. Epigenetic quantification of tumor-infiltrating Tlymphocytes. *Epigenetics* 2011;6(2):236–46.
- [2] Dickhaus T, Stange J. Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. *Calcutta Statistical Association Bulletin* 2013;65(257–260):123–44.
- [3] Stange J, Dickhaus T, Navarro A, Schunk D. Multiplicity- and dependency-adjusted p -values for control of the family-wise error rate. *Stat Probab Lett* 2016;111:32–40.
- [4] Dickhaus T. Simultaneous statistical inference. With applications in the life sciences. Berlin: Springer; 2014.
- [5] Houseman EA, Christensen BC, Yeh R-F, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinf* 2008;9(1):365.
- [6] Chen Z, Huang H, Liu Q. Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data. *BMC Bioinf* 2014;15:142.
- [7] Schildknecht K, Olek S, Dickhaus T. Simultaneous statistical inference for epigenetic data. *PLoS One* 2015; 10(5). e0125587/1–e0125587/1.
- [8] Jelizarow M, Cieza A, Mansmann U. Global permutation tests for multivariate ordinal data: alternatives, test statistics and the null dilemma. *J. R. Stat. Soc. Ser. C. Appl. Stat* 2015;64(1):191–213. <https://doi.org/10.1111/rssc.12070>.
- [9] Brunner E, Munzel U. Nichtparametrische Datenanalyse. Unverbundene Stichproben. 2nd ed. Heidelberg: Springer Spektrum; 2013.
- [10] Ruymgaart F. A unified approach to the asymptotic distribution theory of certain midrank statistics. Statistique non paramétrique asymptotique, Actes Journ. statist., Rouen/France 1979, Lect. Notes Math. 1980;821:1–18.
- [11] Sugiura N. Multisample and multivariate nonparametric tests based on U statistics and their asymptotic efficiencies. *Osaka J Math* 1965;2(2):385–426. <http://projecteuclid.org/euclid.ojm/1200691466>.

- [12] Brunner E, Munzel U, Puri ML. The multivariate nonparametric Behrens–Fisher problem. *J Stat Plann Inference* 2002;108(1):37–53.
- [13] Marcus R, Peritz E, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976;63:655–60.
- [14] Janssen A. Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Stat Probab Lett* 1997;36(1):9–21.
- [15] Neubert K, Brunner E. A studentized permutation test for the nonparametric Behrens-Fisher problem. *Comput. Statist. Data Anal* 2007;51(10):5192–204. <https://doi.org/10.1016/j.csda.2006.05.024>.
- [16] Pauly M, Asendorf T, Konietzschke F. Permutation-based inference for the AUC: a unified approach for continuous and discontinuous data. *Biom J* 2016;58(6):1319–37. <https://doi.org/10.1002/bimj.201500105>.
- [17] Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwenter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* April 2010;20(4):440–6.
- [18] Kleen TO, Yuan J. Quantitative real-time PCR assisted cell counting (qPACC) for epigenetic - based immune cell quantification in blood and tissue. *J. Immunother. Cancer* 2015;3:46.
- [19] Yuan J, Hegde PS, Clynes R, Foukas PG, Harari A, Kleen TO, Kvistborg P, Macallie C, Maecker HT, Page DB, Robins H, Song W, Stack EC, Wang E, Whiteside TL, Zhao Y, Zwierzina H, Butterfield LH, Fox BA. Novel technologies and emerging biomarkers for personalized cancer immunotherapy. *J. Immunother. Cancer* 2016;4:3.
- [20] Dickhaus T, Royen T. A survey on multivariate chi-square distributions and their applications in testing multiple hypotheses. *Statistics* 2015;49(2):427–54. <https://doi.org/10.1080/02331888.2014.993639>.
- [21] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol* 1995;57(1):289–300.
- [22] Romano JP, Shaikh AM, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST* 2008;17(3):417–42.
- [23] Troendle JF. Stepwise normal theory multiple test procedures controlling the false discovery rate. *J Stat Plann Inference* 2000;84(1–2):139–58.
- [24] Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plann Inference* 1999;82(1–2):171–96.
- [25] Dudoit S, van der Laan MJ. Multiple testing procedures with applications to genomics. Springer Series in Statistics. New York, NY: Springer; 2008.
- [26] Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, Dudbridge F. Estimation of a significance threshold for epigenome-wide association studies. *Genet. Epidemiol.* 2018;42(1):20–33.

BIOINFORMATICS METHODOLOGY DEVELOPMENT FOR THE WHOLE GENOME BISULFITE SEQUENCING

4

Deqiang Sun

*Center for Epigenetics & Disease Prevention, Institute of Biosciences and Technology,
Texas A&M University College of Medicine, Houston, TX, United States*

INTRODUCTION

DNA methylation, an epigenetic modification affecting organization and function of the genome, plays a critical role in both normal development and disease. Until recently, the only known DNA methylation was 5-methylcytosine (5mC) at CpG dinucleotides, which is generally associated with transcriptional repression [1]. In 2009, another form of DNA methylation termed 5-hydroxymethylcytosine (5hmC) [2] was found to be involved in active demethylation [3] and gene regulation [4]. Understanding the functional role of DNA methylation requires knowledge of its distribution in the genome [5,6]. Bisulfite conversion of unmethylated Cs to Ts followed by deep sequencing (BS-seq) has emerged as the gold standard to study genome-wide DNA methylation at single-nucleotide resolution. The most popular protocols include RRBS (reduced representation bisulfite sequencing) [7] and WGBS (whole genome bisulfite sequencing) [8] for the combination of 5mC and 5hmC, oxBS-seq (oxidative bisulfite sequencing) [9] for 5mC, and TAB-Seq (Tet-assisted bisulfite sequencing) [10] for 5hmC, respectively. After mapping BS-seq reads to the genome, the proportion of unchanged Cs is regarded as the absolute DNA methylation level. Due to random sampling nature of BS-seq, deep sequencing (e.g., >30 fold) is usually required to reduce the measurement error. Technological advances and reduced costs have seen a significant increase in interest in BS-seq among biologists. Currently, BS-seq is widely used by small laboratories to profile cell lines and animal models [11], as well as by large consortiums such as the NIH ENCODE, Roadmap Epigenomics, The Cancer Genome Atlas (TCGA), and European BLUEPRINT to profile thousands of cell populations. Hence, it is expected that BS-seq data will continue to grow exponentially. However, despite recent progress [7,12–14], computational methods designed for issues specific to BS-seq are much less developed than those for other sequencing applications such as ChIP-Seq and RNA-seq.

The most fundamental aspects of BS-seq data analysis include read mapping and differential methylation detection. We previously developed one of the most widely used BS mapping

programmed BSMAP [15]. After read mapping, the most common task is the identification of differentially methylated regions (DMRs) between samples, such as disease versus normal. Based on the biological question, DMRs can range in size from a single CpG (DMC: differentially methylated CpG) to tens of millions of bases. Although several statistical methods have been applied to DMR detection [13], among which Fisher's exact test *P*-value (FETP) method [16] is the most popular, several challenges remain to be addressed. (1) Statistical Power: most previous methods are very conservative in power and require deep sequencing (e.g., 30 fold). For example, Hansen [14] recently calculated that for single CpG methylation level “*even 30× coverage yields standard error as large as 0.09.*” As a compromise, many studies assumed that neighboring CpGs have similar methylation levels, thus can be combined together within a genomic region (e.g., 1 kb) to increase the statistical power [17]. For example, BSmooth [14] performs local smoothing followed by *t*-test for DMR detection. While this strategy may be applicable in many cases, regional average analysis will unfortunately miss low-CpG-density DMRs that are abundant in the genome and critical for gene expression, such as TFBSSs. Most TFBSSs are small (i.e., < 50 bp) as implied by high-resolution ChIP-seq and ChIP-exo experiments [18] and contain few or even a single CpG(s) that are in general differentially methylated compared to surrounding ones, thus are very likely to be “overlooked” by the regional average analysis. (2) Biological Significance: previous methods use *P*-value for statistical significance of DMR. This *P*-value metric only tells whether a region is differentially methylated, but does not directly measure the magnitude of the methylation difference. A similar problem also exists in gene expression profiling, where the *P*-value does not directly measure the expression fold-change [19]. Since sequencing depth in BS-seq experiments can fluctuate by an order of magnitude in different loci, a very small methylation difference, although not biologically meaningful, can easily return a significant *P*-value if the underlying sequencing depth is deep enough. On the other hand, the nominal methylation difference, that is, direct subtraction of two methylation ratios, suffers significantly from the random sampling error such that a large difference with low sequencing depth is not likely to be statistically meaningful. (3) Biological Variation is an essential feature of DNA methylation [20] and should be handled carefully to detect reproducible DMRs that represent the common characteristics of the sample group. However, most previous methods fail to account for biological variation between replicates and simply pool the raw data from replicates for DMR detection. Some of the resulting DMRs may have significant differences at the mean level but might not be reproducible between replicates, and hence are “false-positives.” To our knowledge, BSmooth [14] is the first replicate-aware program that accounted for biological variation using a modified *t*-test.

In response to these challenges, we developed a powerful differential methylation analysis algorithm termed MOABS: Model-based Analysis of Bisulfite Sequencing data. MOABS uses a beta-binomial hierarchical model to capture both sampling and biological variations, and accordingly adjusts observed nominal methylation difference by sequencing depth and sample reproducibility. The resulting credible methylation difference (CDIF) is a single metric that combines both biological and statistical significance of differential methylation. Using both simulated and real whole-genome BS-seq data from mouse brain tissues and stem cells, we demonstrate the superior performance of MOABS over other leading methods, especially at low sequencing depth. Furthermore, one practical challenge is that BS-seq data analysis is usually computationally intensive and requires multiple steps. We therefore seamlessly integrate several major BS-seq processing procedures into MOABS, including read mapping, methylation ratio calling, identification of hypo- or hypermethylated regions from one sample, and differential methylation from multiple samples. MOABS is implemented in

C++ with highly efficient numerical algorithms, and thus is at least 10 times faster than other popular packages. For example, it takes only 24 CPU hours to detect differential methylation from 2 billion aligned reads. Together, MOABS provides a comprehensive, accurate, efficient, and user-friendly solution for analyzing large-scale BS-seq data.

RESULTS

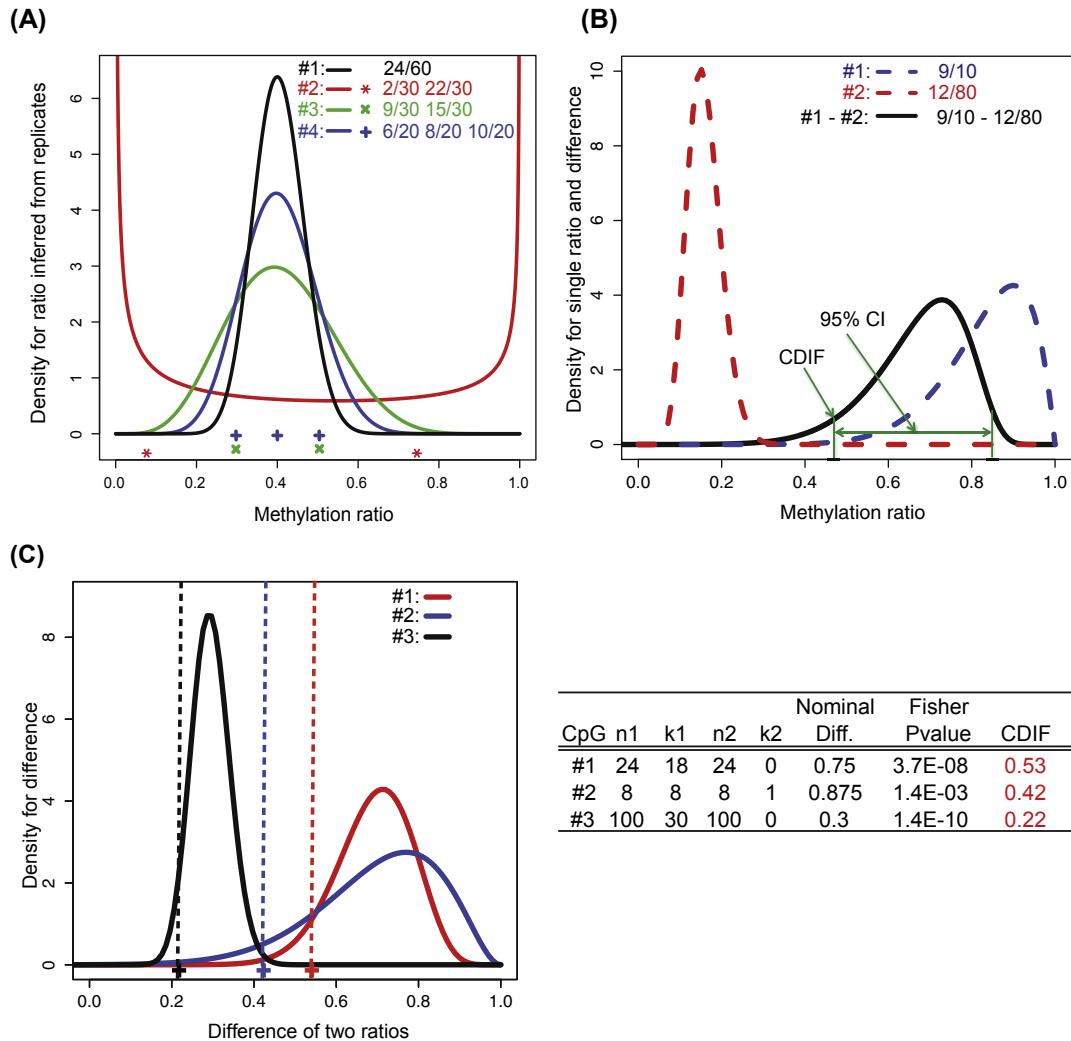
BETA-BINOMIAL HIERARCHICAL MODEL FOR BOTH SAMPLING AND BIOLOGICAL VARIATIONS

For a single CpG locus in the j th biological replicate of condition i , we denote the number of total reads, the number of methylated reads, and methylation ratio as n_{ij} , k_{ij} , and p_{ij} , respectively. For a typical two-group comparison, $i = 1, 2$ and $j = 1, 2, \dots, N$, where N is the number of replicates in each condition. n_{ij} and k_{ij} are observations from experiments, while p_{ij} is unknown with k_{ij}/n_{ij} as its nominal estimation. Given p_{ij} and n_{ij} , the number of methylated reads k_{ij} is characterized by the sampling variation from sequencing and can be modeled by a binomial distribution: $k_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$. The posterior distribution of the methylation ratio p_{ij} will then follow a beta distribution $\text{Beta}(\alpha_{ij}, \beta_{ij})$ and can be estimated using an empirical Bayes approach. The prior distribution will be estimated from the whole genome, in which most CpGs are either fully methylated or fully unmethylated, resulting in a bimodal distribution. The empirical Bayes approach will automatically incorporate such bimodal information in the methylation ratio estimation and hence increases the power of our analysis.

When biological replicates are available, we will refine the posterior distribution of p_{ij} with biological variation from the Bayesian perspective. Specifically, α_i and β_i will be treated as random variables with a prior distribution estimated from all the CpGs in the genome similar to the empirical Bayes priors. We will then use maximum likelihood approach to generate the posterior distribution of p_i . Typical posterior distributions of four CpGs are shown in Fig. 4.1A, in which all CpGs have the same average methylation ratios and the same total number of reads. Their methylation ratios would have identical beta distributions (black curve on CpG #1) if biological variation was not considered. Our method is able to adjust the posterior distribution of p_i based on observed biological variation. For example, highly variable replicates on CpG #2 result in a bimodal distribution, whereas reproducible replicates on CpG #3 lead to a normal-like distribution. Furthermore, increasing the number of reproducible replicates from two to three on CpG #4 will reduce the variation of the resulting posterior distribution. Taken together, the posterior distribution of the methylation ratio in condition i will be determined by its prior distribution, sequencing depth, and the degree of variation between replicates.

CREDIBLE METHYLATION DIFFERENCE (CDIF) IS A SINGLE METRIC FOR BOTH STATISTICAL AND BIOLOGICAL SIGNIFICANCE OF DIFFERENTIAL METHYLATION

We illustrate the idea of CDIF using a simple experimental design, in which only one sample ($N = 1$) is sequenced for each of the two conditions: $k_i \sim \text{Binomial}(n_i, p_i)$ and $p_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, 2$. The empirical Bayes priors α_i^0, β_i^0 of p_i will be estimated from all the CpGs in the genome by maximizing a marginal likelihood function using the quasi-Newton optimization method [21]. In this case, there is no

**FIGURE 4.1**

Overview of the MOABS algorithm. (A) Posterior distribution of methylation ratio inferred from biological replicates. Each curve represents the inferred methylation ratio beta distribution of a CpG. The symbols at the bottom indicate the observed methylation ratios of all replicates. The values on the top right corner indicate number of methylated reads over number of total reads in each replicate. (B) An example of credible methylation difference (CDIF). Dashed curves indicate inferred methylation ratio beta distributions from low (Sample #1) or high sequencing depth (Sample #2). The black curve is the exact distribution of the methylation difference between two samples. The CDIF is shown as the lower bound of the 95% confidence interval. (C) Ranking of three CpG examples by CDIF, FETP P-value, and nominal difference, that is, direct subtraction of two methylation ratios. The three curves are the exact distributions of methylation differences. The corresponding CDIF values are shown as vertical dashed lines.

biological variation, so $Beta(\alpha_i, \beta_i)$ will be only determined by the prior distribution and sequencing depth: $\alpha_i = k_i + \alpha_i^0$ and $\beta_i = n_i - k_i + \beta_i^0$. An example is shown in Fig. 4.1B. Due to low sequencing depth ($k_1 = 9; n_1 = 10$), Sample #1's beta distribution has higher variance than that of Sample #2 with high sequencing depth ($k_2 = 12; n_2 = 80$). The methylation ratio difference between two samples is denoted as $t = p_1 - p_2$. One immediate question is how to estimate the confidence interval $CI(a, b)$ of t . Many methods have been proposed but their merits have been subject to debate [22]. We therefore propose to use the exact numerical solution [23] to solve $CI(a, b)$. CDIF is then defined as the distance between 0 and the 95% $CI(a, b)$ (Fig. 4.1B):

$$CDIF \equiv \begin{cases} a, & \text{if } a \geq 0 \\ 0, & \text{if } a < 0 < b \\ b, & \text{if } b \leq 0 \end{cases}$$

In practice, CDIF represents the conservative estimation of the true methylation difference, that is, for 97.5% of chance the absolute value of true methylation difference is greater than or equal to that of CDIF. The CDIF value will be assigned to 0 if there is no significant difference. Constructed in this way, the CDIF value, if greater than the resolution defined as $\min(1/n_1, 1/n_2)$, guarantees a significant P -value from Fisher's exact test, and at the same time represents the magnitude of methylation difference. The sequencing depth will largely influence CDIF, since bigger n will make a smaller 95% CI of the methylation difference, normally resulting in greater CDIF value.

We believe CDIF is a better metric to capture the methylation difference than statistical P -value or nominal methylation difference. Three CpG examples are shown in Fig 4.1C. According to P -value 1.4e-10, CpG #3 is the most significant one. However, this significant P -value, which is largely driven by the high sequencing depth, does not correctly represent the actual biological difference of 0.3, which is the smallest among three CpGs. On the other hand, if we use nominal difference, CpG #2 would be the most significant. However, its low sequencing depth makes this high nominal difference unreliable. CDIF is able to penalize the nominal difference according to its statistical significance and ranks CpG #1 as the most significant followed by CpGs #2 and #3, although CpG #1 does not have the most significant P -value or nominal difference. Taken together, CDIF reaches a well balance between statistical and biological significance and gives a more stable and biological meaningful interpretation and ranking of differential methylation.

FUNCTIONS AND PERFORMANCE OF THE MOABS PIPELINE

We have implemented MOABS as a comprehensive software pipeline (Fig. 4.2A), including read alignment, quality control (QC), single sample analysis, and multiple sample comparative analysis. (1) The read alignment model is a wrapper of popular bisulfite mapping programs, such as BSMAP [15], which allows the trimming of low quality band adaptor sequences, as well as supports parallel computing on a cluster. (2) The QC module adjusts biases in PCR amplification, end-repair, bisulfite conversion failure, etc. [24]. In addition, it can also estimate bisulfite conversion rate based on cytosines in the non-CpG content. (3) Single sample analysis reports CpG or CpH methylation ratios with corresponding confidence intervals, detects hypo- or hypermethylated regions (e.g., *Trp53* gene in Fig. 4.2B) in the genome [25], and provides general statistics with descriptive figures (an example of the mouse methylome [25] is shown in Fig. 4.2C). (4) For multiple sample comparative analysis,

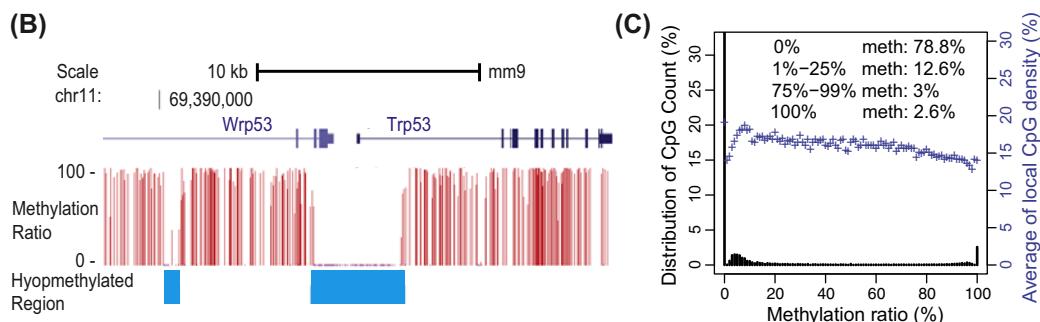
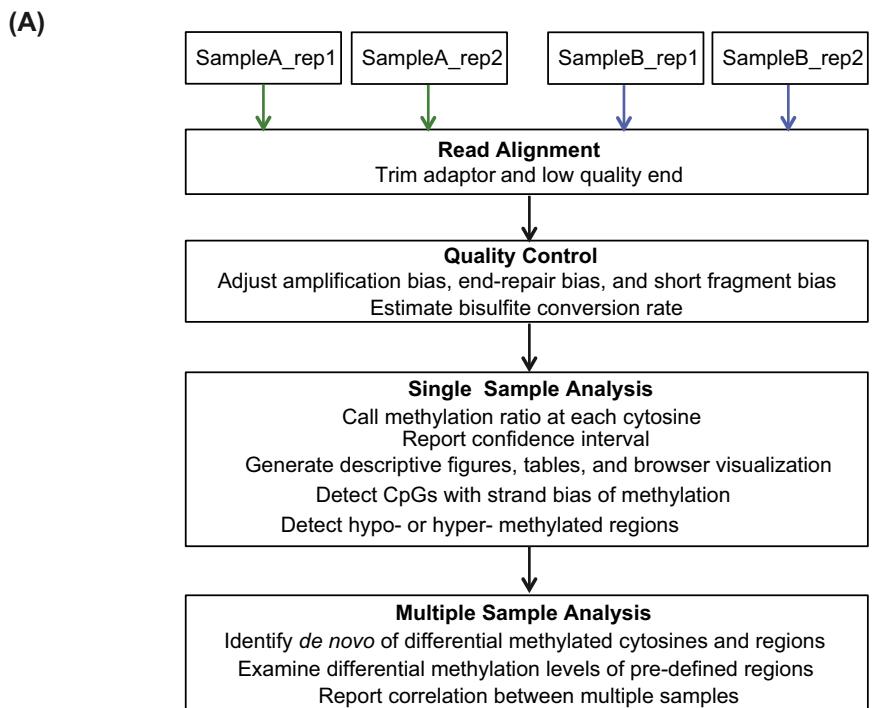


FIGURE 4.2

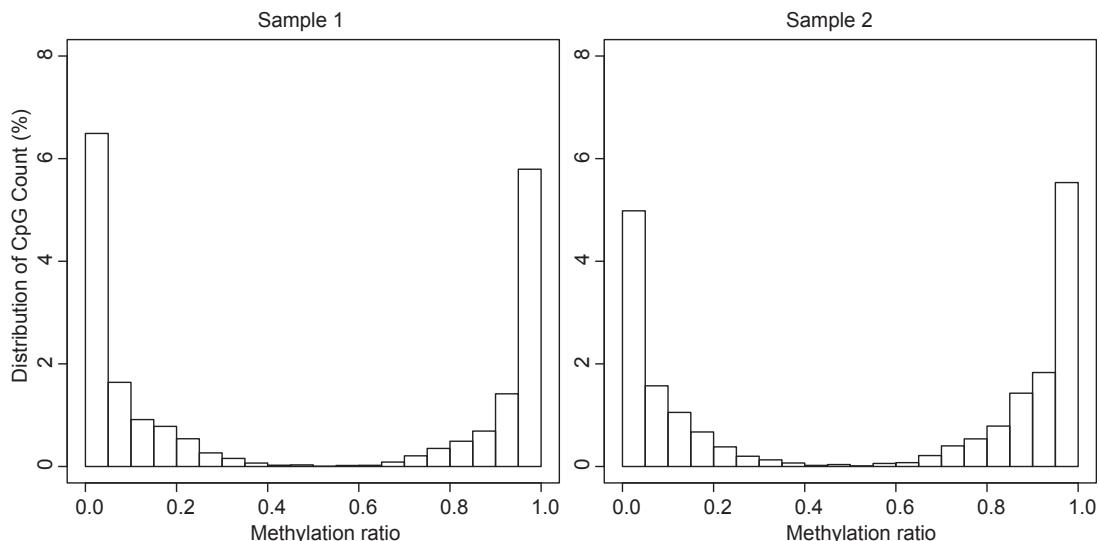
Overview of the MOABS software pipeline. (A) Comprehensive workflow of the MOABS pipeline. (B) An example of hypomethylated region. (C) A descriptive figure for global methylation distribution of a mouse methylome. The Y-axis on the left is percent of CpGs and the Y-axis on the right is the average of local CpG density at each specified methylation ratio.

MOABS detects de novo DMCs, which can be further grouped into DMRs using a hidden Markov model. MOABS can also examine the differential methylation levels of predefined regions, such as promoters.

All the modules are wrapped in a single master script such that users can specify the input BS-seq reads and run all the modules one by one automatically. The MOABS pipeline is developed using C++ with highly efficient numerical algorithms, native multiple threading, and cluster support so that multiple jobs can run in parallel on different computing nodes. Numerous mathematical and computational optimizations have made MOABS pipeline extremely efficient. For example, it takes only 1 h on 24 CPUs to detect differential methylation for approximately 30 million CpGs in the human genome based on 2 billion aligned reads, whereas other software packages can easily take more than 1 day for the same input data. In summary, MOABS is a comprehensive, accurate, efficient, and user-friendly solution for analyzing large-scale BS-seq data.

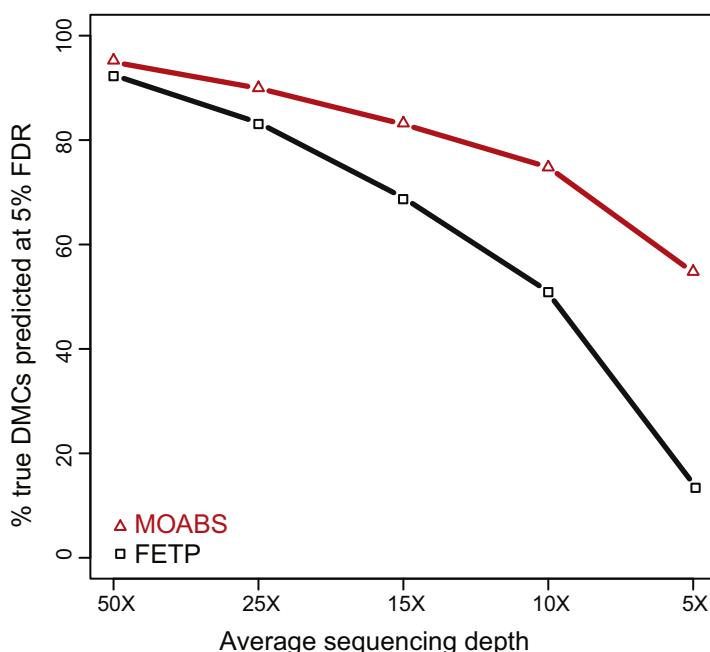
SIMULATED BS-SEQ DATA REVEAL THE SUPERIOR PERFORMANCE OF MOABS

To assess the performance of MOABS on differentially methylated CpGs (DMCs), we simulated 0.1 million true positive CpGs with large methylation difference and 0.9 million true negative CpGs ([Supplementary Fig. 4.1](#)) from an H1 methylome [16], and then compared MOABS with FETP at 5%



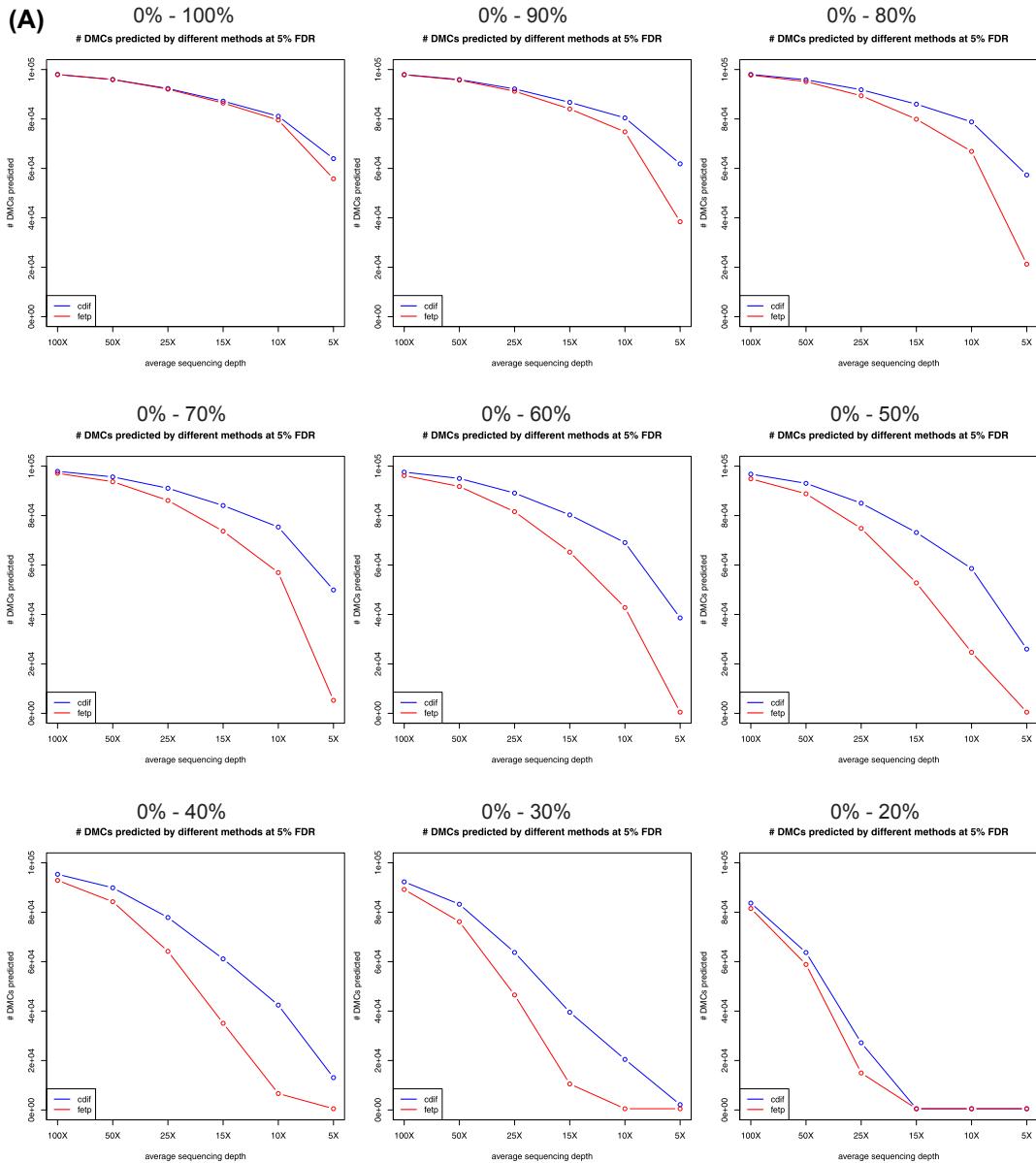
SUPPLEMENTARY FIGURE 4.1 Bimodal Distribution of Simulated Methylation Ratios.

Each bar represents the percent of CpGs in simulated methylome at each specified methylation ratio.

**FIGURE 4.3**

Comparison between MOABS and FETP in detecting DMCs. We simulated 1,000,000 CpGs in two samples with predefined true positive or true negative states. In both samples, 900,000 true negative CpGs were initially assigned the same methylation ratios. The density of the methylation ratios fits a bimodal distribution (Supplementary Fig. 4.1) frequently observed in real BS-seq data. The remaining true positive 100,000 CpGs were randomly assigned at low ratios [0, 0.25] in one sample and high ratios [0.75, 1] in the other sample, respectively. Each methylation ratio was then given a ± 0.05 fluctuation to simulate BS-seq errors. Sequencing depth is randomly sampled from 5 to 50 fold. The Y-axis shows the percentage of true DMCs predicted at 5% FDR.

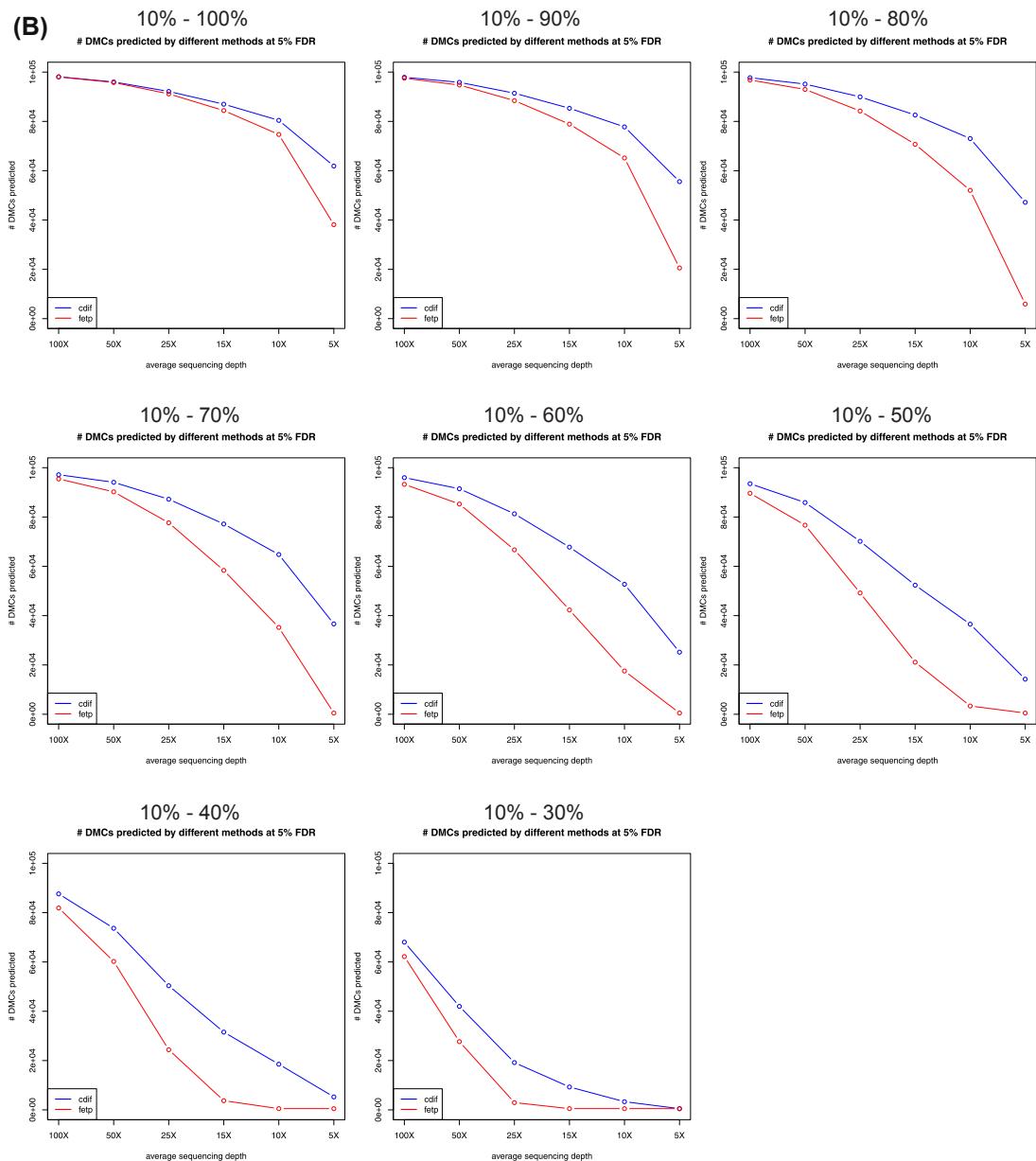
false discovery rate (FDR) (Fig. 4.3). Note that this evaluation is at single CpG resolution without local smoothing, and therefore BSmooth [14] cannot be used. The results indicate that MOABS clearly outperforms FETP with the most dramatic difference observed at low sequencing depth. For example, with sequencing depth at 5–10 fold, MOABS can recover 55%–75% true positives, while FETP only predicts 13%–51% true positives. To further evaluate the performance of MOABS at different methylation levels, we resimulated the 0.1 million true positive CpGs with different baseline methylation levels (0%–100%) and methylation differences (20%–100%). The results (Supplementary Fig. 4.2) indicate that MOABS is more accurate than FETP at any sequencing depth and at any methylation difference. Notably, the difference between the two methods is large when sequencing depth is low or when methylation difference is moderate (50%–70%). In contrast, the difference between methods is small when sequencing depth is high or when the methylation difference is either very high (80%–100%) or very low ($\sim 20\%$). Although FETP is well suited for the



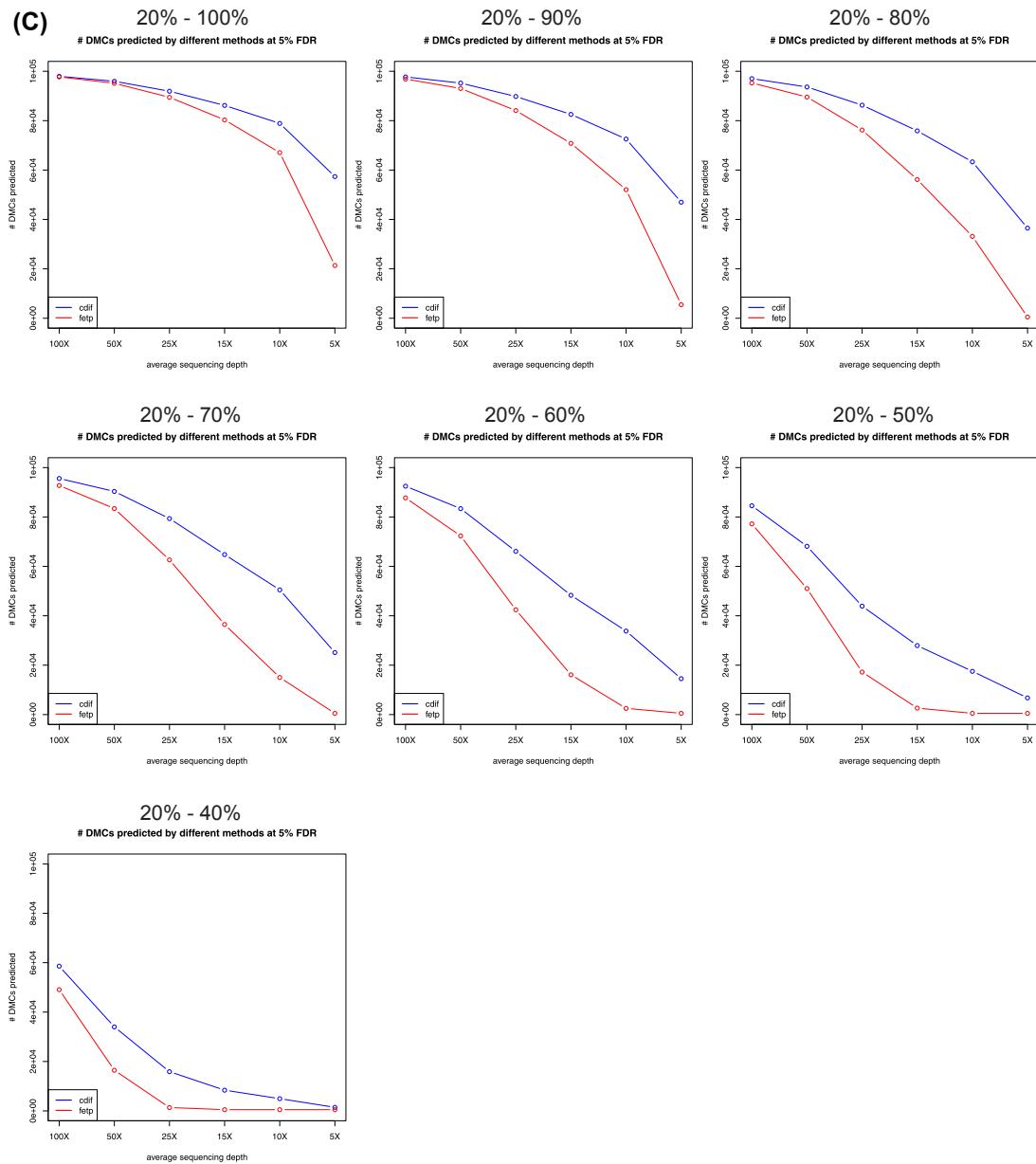
SUPPLEMENTARY FIGURE 4.2 DMC Detection Based on Simulated Data.

The Y-axis is the percent of DMCs predicted by different methods at 5% FDR. Each panel shows the results for DMC detection with different simulated methylation difference.

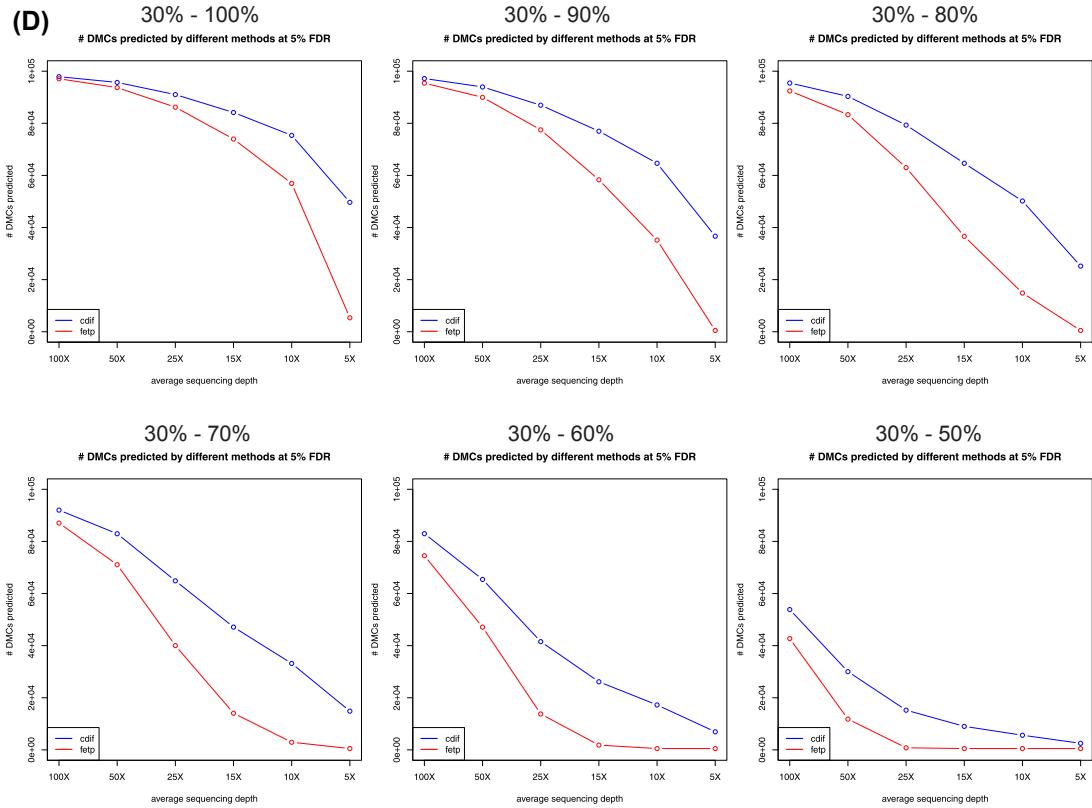
(B)



SUPPLEMENTARY FIGURE 4.2 Cont'd



SUPPLEMENTARY FIGURE 4.2 Cont'd

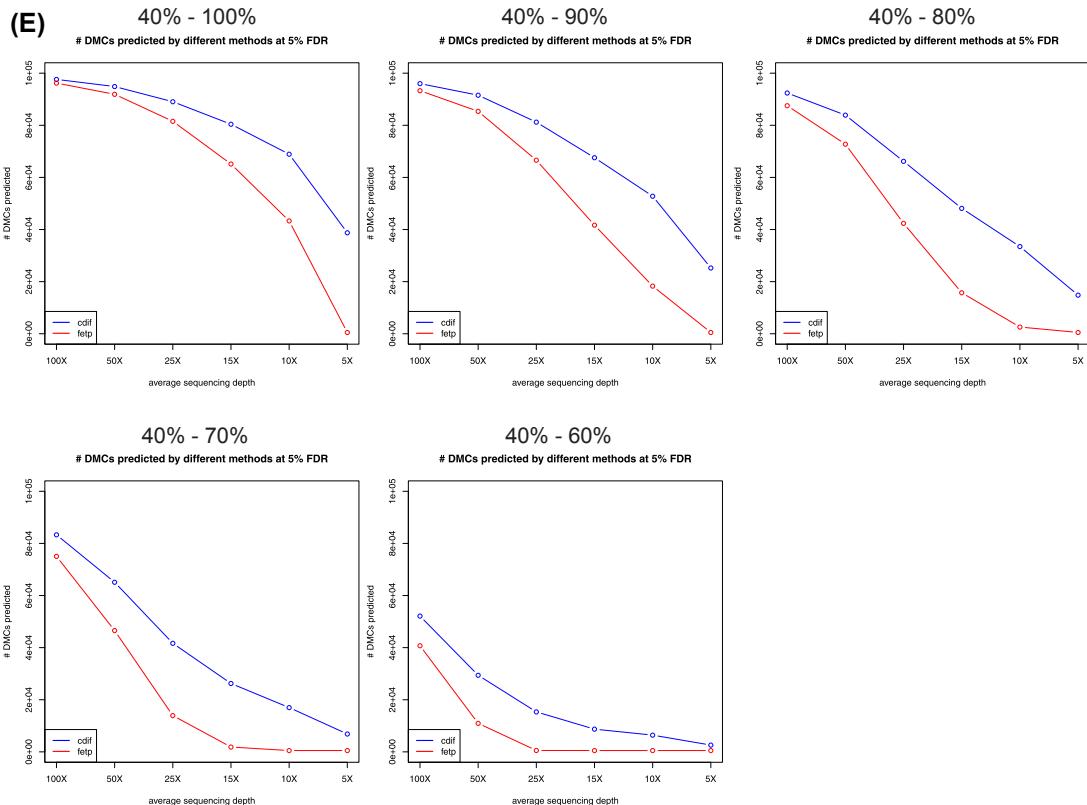


SUPPLEMENTARY FIGURE 4.2 Cont'd

analysis of discrete data, it has less power for DNA methylation, which by its nature is a continuous rather than discrete random variable. The improved power of MOABS results from the modeling of DNA methylation using a beta-binomial hierarchical model and the empirical Bayes approach to borrow information from all the CpGs in the genome.

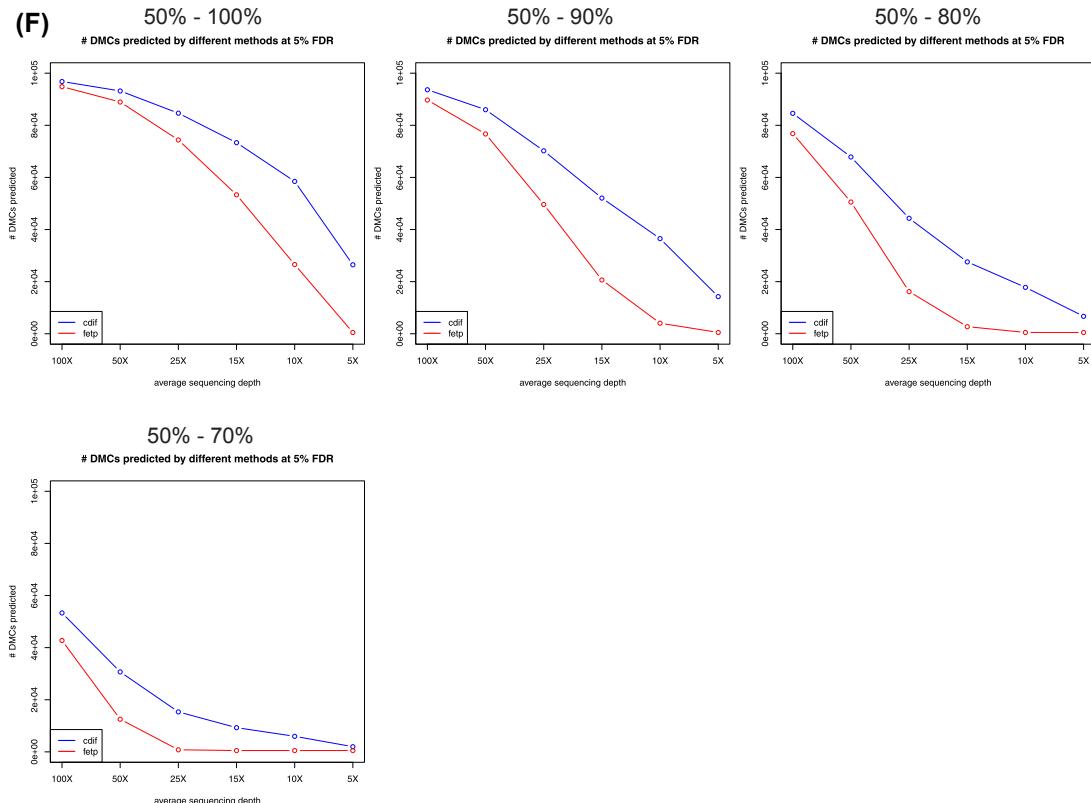
MOABS IMPROVES THE DETECTION OF ALLELE-SPECIFIC DNA METHYLATION

To assess how MOABS performs on DMRs for real BS-seq data, we compared MOABS with FETP and BSmooth [14] using allele-specific mouse methylomes [25], in which a list of well known imprinted DMRs can serve as gold standard for method evaluation. Xie et al. [25] used FETP followed by clustering of DMCs for DMR detection. They confirmed 32 known experimentally verified imprinted DMRs and reported 20 novel ones by pooling two biological replicates without considering sample variation. We noticed that two known DMRs (Ndn and Igf2r) are weak, exhibiting a very small methylation difference of approximately 10%. We also found that three novel DMRs they reported (Vwde, Casc1, and Nhlrc1) are differentially methylated in only one of the two replicates, and thus are likely to be false positives (Supplementary Fig. 4.3). Since the remaining 17 novel DMRs have yet to



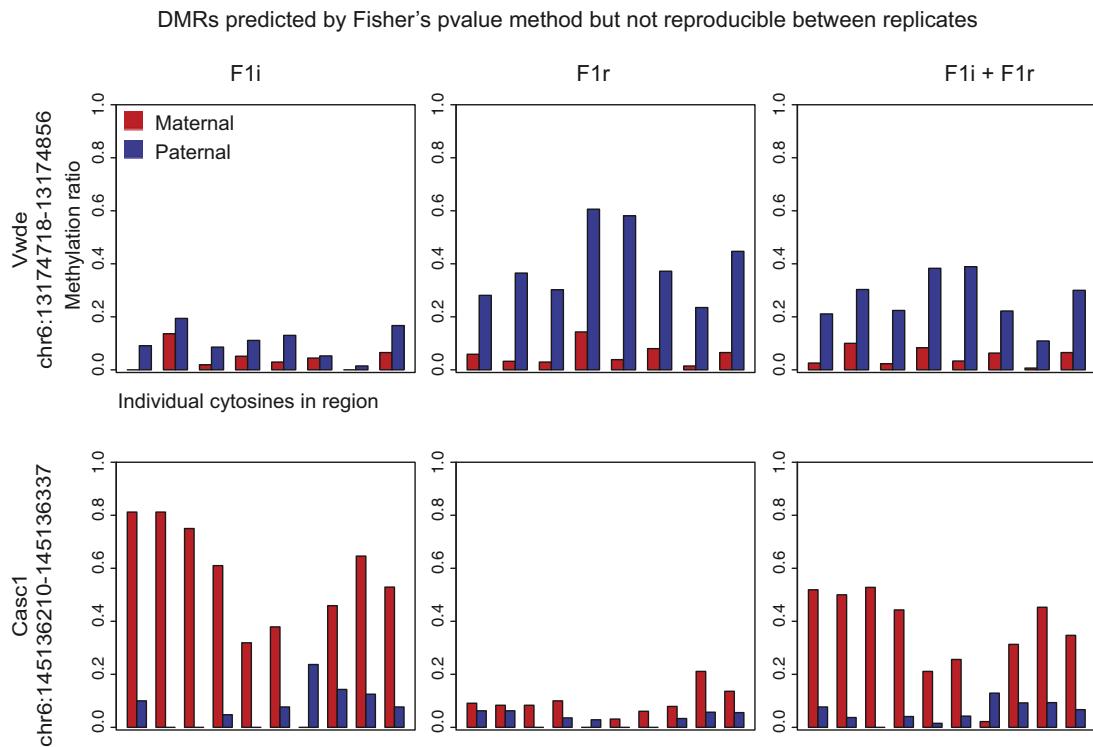
SUPPLEMENTARY FIGURE 4.2 Cont'd

be experimentally verified, we decided to remove them from our analysis. In our method evaluation, we used the 32 known DMRs as true positives and the remaining genome (with 17 reproducible novel DMRs removed) as true negatives. To allow for a fair comparison, we used the same method to calculate FDR for all three methods. In addition, we used the same procedure to cluster DMCs into DMRs for MOABS and FETP. The resulting ROC-like curves (Fig. 4.4A) clearly indicate that MOABS is superior to the other two methods. MOABS successfully reports all 32 known DMRs including the two weak ones at 11% FDR, and 4 “false positive” new DMRs (Cdh20, Trappc9, Pcdhb20, and Pfnd4). Manual inspection (Supplementary Fig. 4.4) confirms that these 4 “false positives” are indeed regions showing differential methylation in both replicates. Hence the 11% FDR of MOABS is significantly over estimated based on incomplete true positives. Interestingly, our FETP analysis predicts 7 new DMRs in addition to 32 known DMRs, suggesting additional filtering steps may have been performed in Xie et al. [25]. Among these 7 DMRs, 1 greatly overlaps with the new DMR Pcdhb20 reported by MOABS, while the other 6, including Vwde, Casc1, and Nhlc1, show poor correlation between replicates. Finally, the ROC-like curve indicates that BSmooth is less accurate than either FETP or MOABS.



SUPPLEMENTARY FIGURE 4.2 Cont'd

The 32 known DMRs can be easily detected by both MOABS and FETP mainly because they have large methylation differences and high read depth (54 fold), which is consistent with our simulation study. However, deep bisulfite sequencing of the mammalian genome is still quite expensive. This reality motivated us to test to what extent these known DMRs can still be recovered at a lower sequencing depth. The same previous procedure was applied to compare all three methods. The number of recovered known DMRs at 5% FDR is plotted at each sequencing depth from random sampling (Fig. 4.4B). We observe that the lower sequencing depth, the greater performance difference between MOABS and FETP. For example, when the depth is at 11 fold, MOABS recovers roughly 90% of known DMRs, while FETP only detects 78% of DMRs. When the depth is further lowered to 3.1 fold, MOABS can still recover roughly 70% of known DMRs, while FETP detects 44% DMRs. Interestingly, BSmooth's performance is largely independent of sequencing depth, probably because it was designed mainly for low sequencing depth. Indeed, at a low depth of 3.1 fold, BSmooth outperforms FETP. However, at sequencing depth higher than 3.1 fold, BSmooth has a lower sensitivity than the other two methods. Collectively, we conclude that MOABS is superior in DMR detection, especially at low sequencing depth.

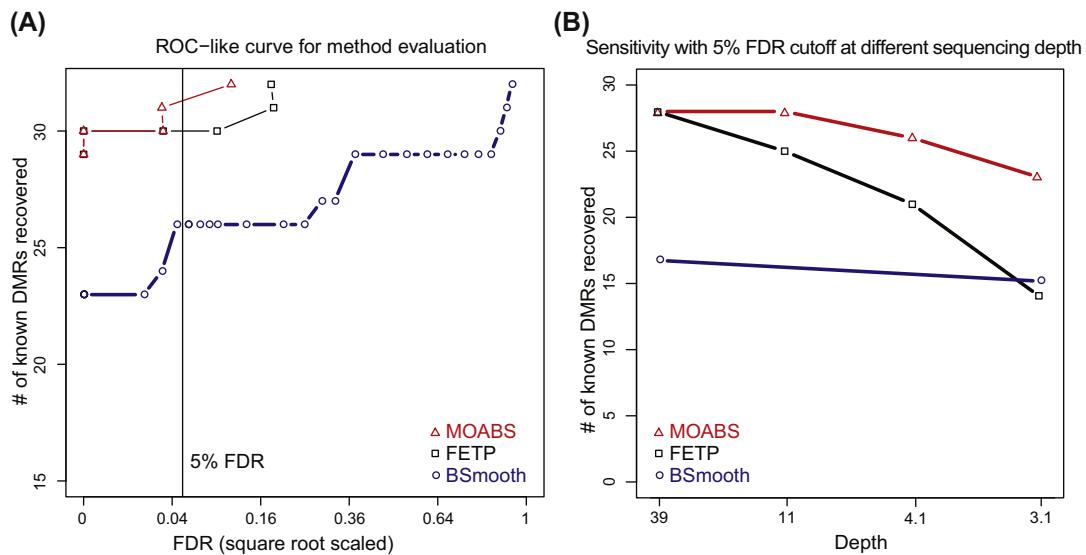


SUPPLEMENTARY FIGURE 4.3 Nonreproducible DMRs Predicted by FETP.

The first and second rows are for Vwde and Casc1 DMRs, respectively. The first, second, and third columns show the methylation information for replicate F1i, F1r, and direct combination of two replicates, respectively. Each bar represents methylation ratio of a CpG in the DMR. Red bar shows maternal methylation and blue bar shows paternal methylation. Vwde DMR shows differential methylation in replicate F1r but not in F1i; Casc1 DMR shows differential methylation in replicate F1i but not in F1r. The reproducibility information is lost by direct combination of replicates. The Nhlrc1 DMR (chr13:47106177-47106300) is not shown due to too many CpGs.

MOABS RELIABLY REVEALS DIFFERENTIAL METHYLATION UNDERLYING TFBSS

Since the previous benchmark is based on a small number of experimentally verified DMRs, we sought to further evaluate the performance of MOABS based on larger scale data sets. The link between differential methylation and TFBSSs provides such a good system. TFBSSs are usually hypomethylated compared to surrounding genome background; therefore, a tissue-specific TFBSS is expected to be a tissue-specific hypomethylated DMR (hypo-DMR). To test this hypothesis, we performed deep (46 fold) WGBS of the mouse hematopoietic stem cell (HSC) and compared the HSC methylome with that of a publicly available mouse embryonic stem cell (ESC) [26]. The HSC-specific hypo-DMRs were then compared with approximately 58,000 *in vivo* ChIP-seq TFBSSs of 10 major HSC-specific TFs [27], including Erg, Fli1, Gata2, Gfi1b, Lmo2, Lyl1, Meis1, Pu.1, Runx1, and Scl. Fig. 4.5A illustrates the hypo-methylation of a TFBSS in the *Runx2* gene. At the center of the TFBSS co-bound by Runx1, Gata2, and Scl,

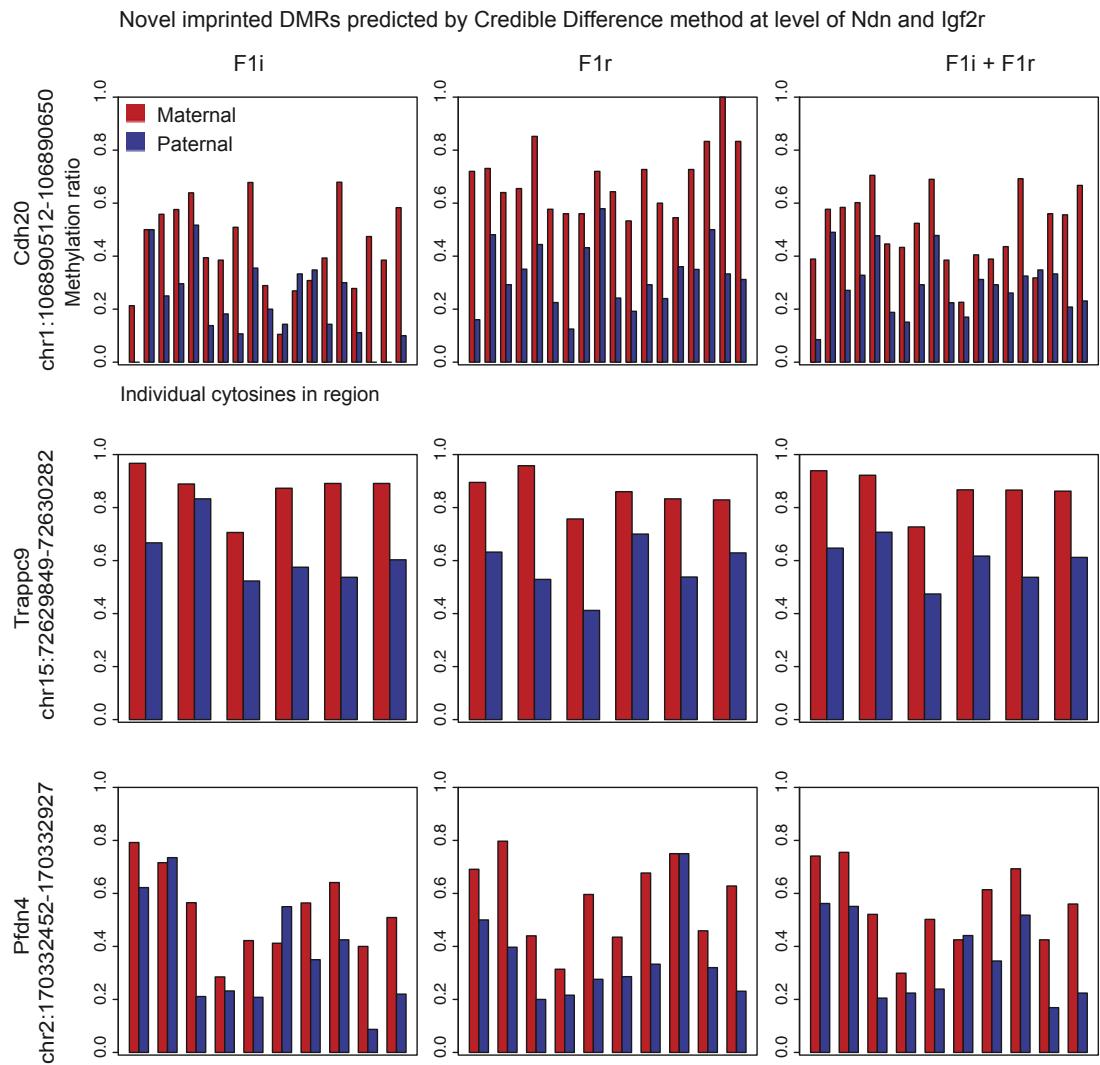
**FIGURE 4.4**

MOABS improves the detection of allele-specific DNA methylation. (A) The Y-axis shows the number of known DMRs recovered by three different methods. (B) Sensitivity (Y-axis) at 5% FDR with different sequencing depth (X-axis).

there are two CpGs fully methylated in mouse ESC but unmethylated in HSC, while the surrounding regions are almost fully methylated in both HSC and ESC. [Supplementary Fig. 4.5](#) shows more examples of tissue-specific hypo-DMR coupled with tissue-specific TFBSs. Such TFBS-associated hypomethylated regions are usually very small and abundant in the genome. Using Runx1 as an example, 71% of the 4793 Runx1 TFBSs show hypomethylation, while the remaining TFBSs are either fully methylated or have no underlying CpGs. Together, ~34% of TFBS associated hypomethylated regions contain no more than three CpGs with a median length of 51 bp ([Fig. 4.5B](#)). Furthermore, 14% of such regions even have a single CpG. For such small DMRs, single CpG level differential analysis is essential since regional averaging is very likely to overlook most of them.

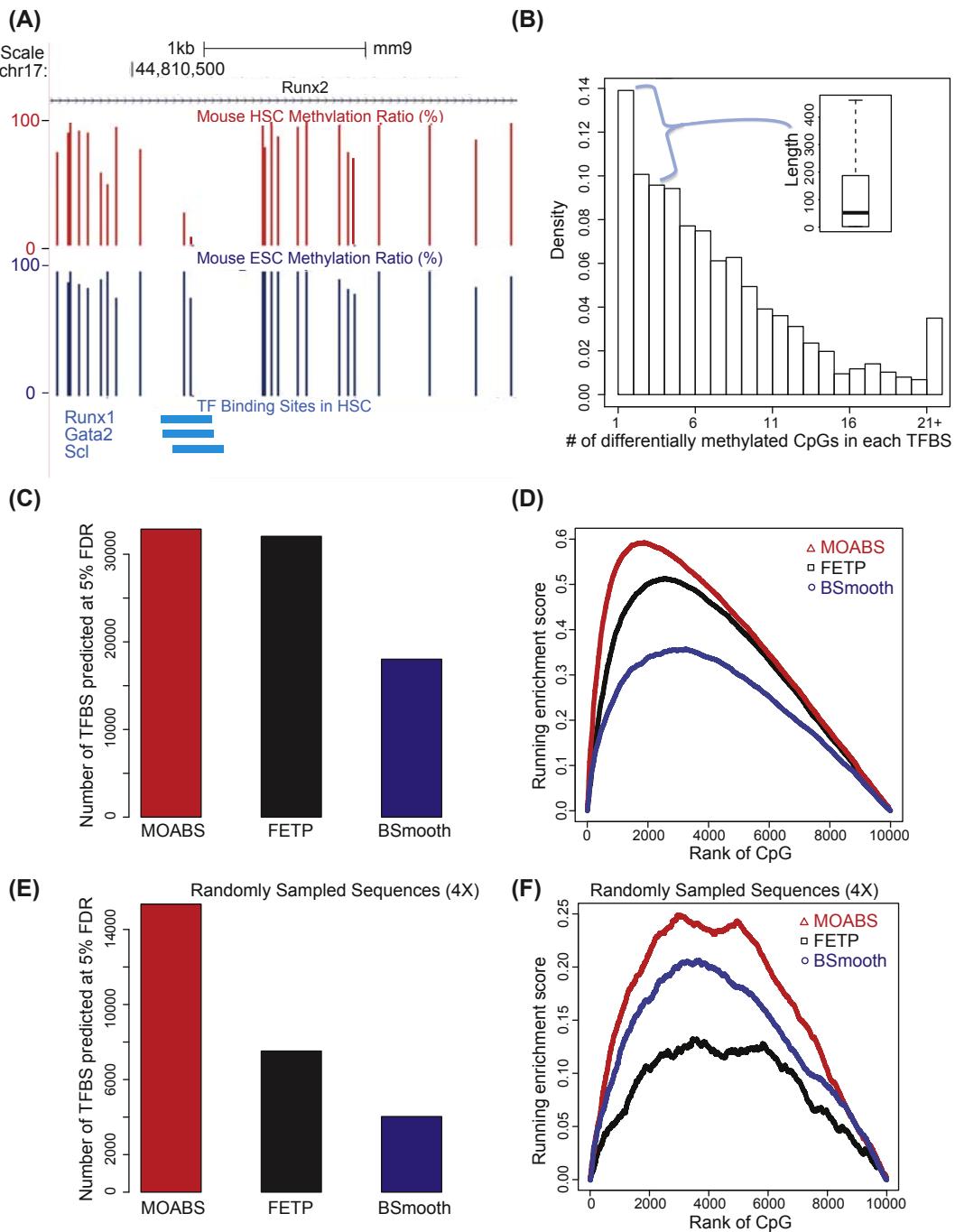
We then used TFBSs to evaluate DMC detection assuming tissue-specific TF binding is associated with tissue-specific hypomethylation. For a fair comparison, we calculated FDR for each method based on a method-specific null distribution obtained through permutation of read sample labels. At an FDR of 5%, MOABS, FETP, and BSmooth predicted 32,867, 32,047, and 18,021 differentially methylated TFBSs respectively ([Fig. 4.5C](#)). We also used a method similar to gene set enrichment analysis (GSEA) [28] to test enrichment of TFBS moving down the lists of DMCs ranked by different methods. MOABS shows the highest enrichment score ([Fig. 4.5D](#)) of TFBS. For example, with the same 4000 most significant DMCs, MOABS recovers 1000 TFBSs while FETP only predicts ~600 TFBSs (i.e., 40% less).

In this instance, the sequencing depth is sufficient to enable MOABS and FETP to recover very similar number of TFBSs. However, when we randomly sampled reads to a depth of fourfold, MOABS recovered many more TFBSs (15,349) than FETP (7520) and BSmooth (4028) ([Fig. 4.5E](#)). Again, at

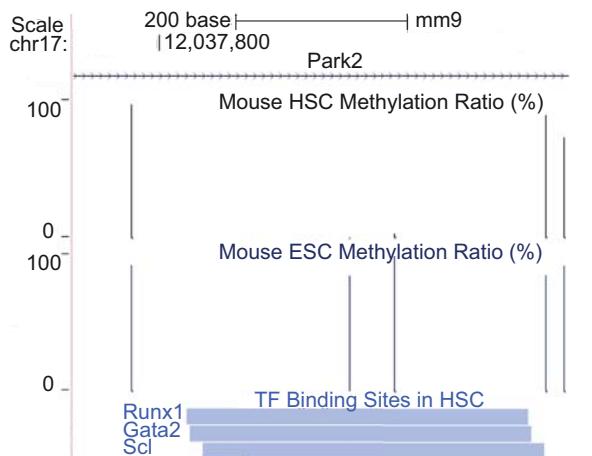


SUPPLEMENTARY FIGURE 4.4 Novel Imprinted DMRs Predicted by MOABS at Level of Ndn and Igf2r.

The first, second, and third rows are for Cdh20, Trappc9, and Pfdn4 DMRs respectively. The first, second, and third columns show the methylation information for replicate F1i, F1r, and direct combination of two replicates respectively. Each bar represents methylation ratio of a CpG in the DMR. Red bar shows maternal methylation and blue bar shows paternal methylation. The differential methylation is reproducible in both replicates. The nominal difference is small but they are at the same level of known DMRs Ndn and Igf2r.

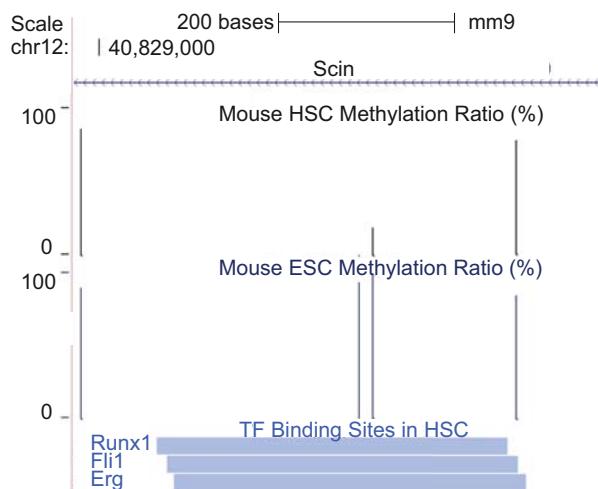
**FIGURE 4.5**

MOABS reveals differential methylation underlying TFBSs. (A) UCSC genome browser illustration of one TF binding site. The tracks from top to bottom are genomic positions, RefSeq Gene, HSC methylation, ESC



SUPPLEMENTARY FIGURE 4.5 Examples of DMCs Associated With TFBS.

Park2 and *Scin* show the small hypomethylated region surrounded by full methylation. The two tracks are methylation ratio in HSC and ESC with each bar denoting the methylation ratio on each CpG. The bottom track marks the positions of binding regions of TFs in mouse HSC.



methylation, and TFBS. For each CpG, an upward bar denotes the methylation ratio. (B) Distribution of the number of DMCs underlying TFBSs. The inserted boxplot indicates the length distribution of TFBSs with 1–3 DMCs. (C) Number of differentially methylated TFBSs predicted by different methods at 5% FDR. (D) Running enrichment scores for TFBSs. All the CpGs are ranked by each method. The score increases if the CpG is in a TFBS or decreases if not. Only 10,000 CpGs are sampled to make this plot, as indicated by the X-axis. The 10,000 times of random shuffle of TFBSs determined *P*-values of the maximum enrichment score to be 1.4E-3, 1.6E-3, and 4E-3 for MOABS, FETP, and BSmooth respectively. (E) and (F) Same as (C) and (D) with 4× sequencing depth by random sampling. The 10,000 times of random shuffle of TFBSs determined *P*-values of the maximum enrichment score to be 2.9E-2, 5.1E-2, and 9.2E-2 for MOABS, FETP, and BSmooth respectively.

this low sequencing depth, MOABS not only recovers two- to threefold more TFBSs, but also exhibits more significant score of TFBS enrichment in the most significant DMCs. In both high and low sequencing depths, BSsmooth recovers fewer TFBSs mainly because its smoothing function easily ignores small region with a few CpGs. Together, using tissue-specific *in vivo* TFBSs, we demonstrate that MOABS can better recover differential methylation in small regulatory regions with a few CpGs, especially at low sequencing depth (e.g., fourfold).

MOABS DETECTS DIFFERENTIAL 5hmC IN ES CELLS USING RRBS AND oxBS-SEQ

To demonstrate the broad utility of MOABS, we analyzed 5hmC data using RRBS and oxBS-seq [9]. RRBS measures both 5mC and 5hmC together while oxBS-seq [9] detects 5mC directly. The 5hmC level can then be inferred by the difference between RRBS and oxBS-seq of the same sample. The 5hmC level is often very small (e.g., at 5%) and hence its detection requires hundreds of fold coverage using FETP [9]. Our simulation study indicates that MOABS can significantly reduce the depth requirement (Fig. 4.6A). For example, to detect 5hmC at 5% when 5mC is at 0%, MOABS requires 80-fold coverage while FETP needs ~200-fold. However, when the 5mC level is close to 50%, significantly more reads will be needed for both methods (~120-fold for MOABS and >500-fold for FETP). The differential 5hmC between two samples can be inferred by the difference of two CDIF values, each of which is the difference between RRBS and oxBS-seq of the same sample. The similar numerical approach can then be used to infer the distribution of the difference of difference between two beta distributions, which are used to model BS-seq data. Fig. 4.6B shows an example, in which 5hmC is measured by both RRBS and oxBS-seq in two samples. FETP shows more significant *P*-value for 5hmC in Sample #1 than in #2, whereas MOABS CDIF is bigger in Sample #2 than in #1. However, the significance of FETP on Sample #1 is largely driven by the high sequencing depth, thus does not

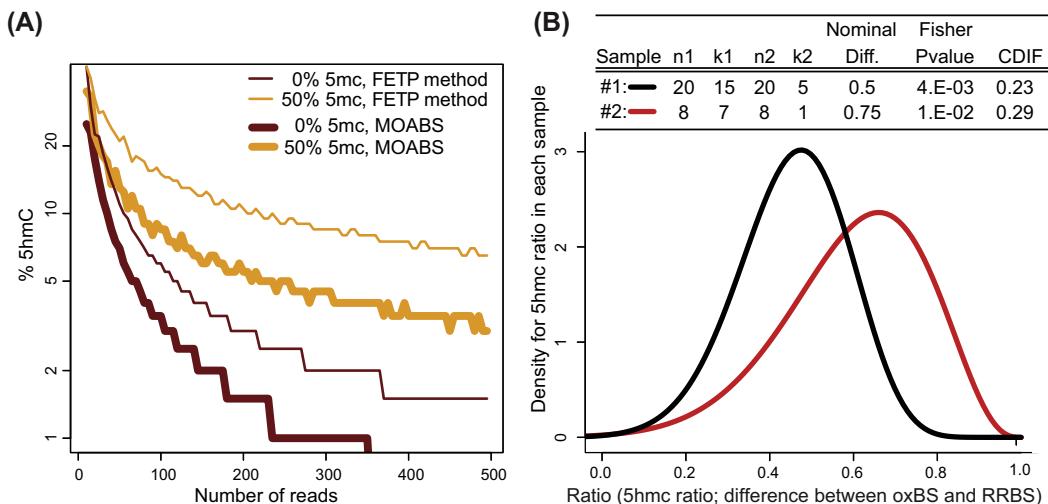


FIGURE 4.6

MOABS detects differential 5hmC using RRBS and oxBS-seq. (A) Simulation study of 5hmC detection from oxBS-seq and RRBS. Each point on curves represents the smallest number of reads (X-axis) needed to detect a 5hmC ratio (Y-axis) at specified 5mC ratio (indicated as colors). The thin and thick curves represent FETP and MOABS, respectively. (B) Beta distribution of 5hmC ratio in each sample.

correctly represent the actual biological difference. In contrast, MOABS CDIF reaches a balance between statistical and biological significance and gives a biologically meaningful differential 5hmC at a CDIF value of 0.06 (0.29–0.23).

When applied to RRBS and oxBs-seq data derived from ES cell lines with different passages [9], MOABS reported 299 genes with decreased 5hmC and 125 genes with increased 5hmC ([Supplementary Table 4.1](#)) in promoters in the later passage P20, which is consistent with the mass

Supplementary Table 4.1 Genes Associated With Decreased or Increased 5hmC in Later Passage

Decreased 5hmC		
1190002F15Rik	Auts2	Crb2
1300018I17Rik	BC037032	Crocc
1500002O20Rik	BC046331	Crtc3
1700014N06Rik	BC048403	Csgalnact1
2200002D01Rik	Bcl11b	Csrp2bp
2210408I21Rik	Bcl2l10	Ctnna3
2410004A20Rik	Bcl2l12	Cyp24a1
3830431G21Rik	Bend7	Cyp2c38
4930415O20Rik	Bri3	Cyp2c53-ps
4930474G06Rik	Bsg	D19Wsu162e
4930526L06Rik	Btbd17	Dab2ip
4931429I11Rik	C87414	Dennd1a
4932414N04Rik	Cacna1b	Dgkb
5730528L13Rik	Cacna1c	Diap2
7530420F21Rik	Cadm2	Dlc1
9330020H09Rik	Camk4	Dlgap1
9830001H06Rik	Cartpt	Dlgap2
Acot8	Cast	Dmd
Adam22	Ccdc63	Duox2
Adamts20	Cchcr1	Dus2l
Agbl1	Ccin	E130309F12Rik
Agxt2l2	Cdh4	E430025E21Rik
Ahdc1	Cdk5r1	Ebf3
Ak5	Celf1	Echdc2
Ank3	Cgnl1	Efcab5
App	Chd9	Egflam
Arhgef18	Chrm2	Eif3a
Ascc3	Cilp2	Eif4ebp3
Asph	Cirbp	Elk3
Ass1	Ckap2	Elp4
Astn1	Clpp	Emid2
Atf7ip	Col14a1	Eml1
Atmin	Col23a1	Enox1
AU019990	Cpe	Ephb4
AU022751	Cpsf3	Erbb4

Supplementary Table 4.1 Genes Associated With Decreased or Increased 5hmC in Later Passaged—cont'd

Erbb4	Hist1h2ba	Mier3
Esrrb	Hist1h2bq	Mir5046
Etv3	Hist1h2br	Mnx1
Etv6	Hnrnpab	Mtap1a
Extl3	Il6ra	Mtap7d1
F730043M19Rik	Immp2l	Myo7a
Fam110b	Itpk1	Nacad
Fam169b	Kcnc2	Nbas
Fam73a	Kcnh7	Ndufc2
Fam83h	Kcnq1	Npr2
Fanc1	Kcnq1ot1	Nwd1
Fbxo46	Kdm5a	Odz1
Fgf14	Khyn	Ophn1
Fgfr1	Kif19a	Oscar
Fgr	Kifc3	Otud7a
Filip1	Klk9	Oxct1
Foxi2	L3mbtl4	Palm
Galnt7	Larp4b	Parg
Galnt16	Lcorl	Pde3a
Gata3	Lemd3	Pdgfa
Gjb3	Letmd1	Pdgfc
Gli1	Lhfp12	Pgs1
Gm14405	Lin7a	Phkb
Gm15545	Lipk	Pkhd1
Gm17384	Lrrc10	Pla2r1
Gm2083	Lrrc7	Plagl1
Gm694	Lrrc8d	Pld3
Gng12	Lrsam1	Plekha7
Gpat2	Luzp2	Plekhm2
Gpr1	Lyst	Pmp22
Gpr173	Macf1	Poll
Gpr19	Macrod2	Polrmt
Gria2	Mamstr	Prdm16
Grid1	Map3k9	Prdm5
H1f0	Mapre2	Prkce
Hcn2	Mettl7a1	Prkczt
Hist1h2aa	Mfap4	Psmd3
Hist1h2ao	Mfhas1	Ptk7
Hist1h2ap	Mib2	Ptpn14

Ptpn20	Slain2	Tmem120a
Ptprn2	Slc1a2	Tmem132d
Ralgapa2	Slc30a6	Trdn
Rasip1	Slit3	Trim52
Rassf3	Smarca2	Ube2h
Relt	Sntg1	Ubtf
Rhbdf2	Snx21	Uhrf1bp1
Rhox13	Spag1	Unc5b
Rhox2e	Spen	Upf1
Ring1	Srbd1	Usp25
Rnf43	Stab2	Vav2
Robo3	Stard6	Wdfy1
Rpl12	Sumf2	Wdpcp
Rps6ka4	Supt3h	Wiz
Rps6ka6	Syde1	Wwc1
Rspo1	Syt16	Yif1b
Rufy3	Syt15	Zbtb7c
Ryk	Tbc1d30	Zdhhc17
Sall1	Tbxas1	Zfp217
Scarf2	Tcea1	Zfp362
Scn1a	Tcf7l1	Zfp366
Serinc5	Tfeb	Zfp42
Sfmbt2	Thbs3	Zfp442
Sh3rf3	Timm23	Zfp629
Sidt2	Tle6	Zfyve26
Sirt4	Tlx1	Zp3r
Increased Shmc		
0610009L18Rik	Agpat4	Col4a5
1300010F03Rik	Apc2	Cox7b2
1700008F21Rik	Arhgap24	Cpt1b
2610307P16Rik	Arhgap6	Cxadr
2900008C10Rik	Bahcc1	D030025P21Rik
4930443O20Rik	Bai3	Dcc
4930467D21Rik	BC090627	Dmbx1
5830403M04Rik	Bcor	Dnahc2
9130221F21Rik	Bmp7	Dnm3
A130022J15Rik	Camk2b	Dpyd
Abca13	Caps2	E2f8
Abcc3	Catsperb	Eda
Ablim1	Ccdc6	Emcn
Actg1	Cd46	Emr4
Adamts11	Chkb	Epha6

Supplementary Table 4.1 Genes Associated With Decreased or Increased 5hmC in Later Passaged—cont'd

Epha7	Naa11	Rnf17
Esrra	Nob1	Robo2
Ev1	Nosip	Rpl31
Fgd4	Odz2	Rsrc1
Fhit	Pard6b	Runx1
Fitm1	Pcdh11x	Rxra
Fyn	Pcdha1	Scaf1
Gabrg1	Pcdha10	Shank1
Galnt13	Pcdha11	Shisa4
Galnt2	Pcdha12	Skint9
Gfra2	Pcdha2	Slc45a1
Gnaq	Pcdha3	Stk3
Grm6	Pcdha4	Stx16
Hk1	Pcdha4-g	Tacr3
Igf1r	Pcdha5	Tceanc
Igsf21	Pcdha6	Tdrd1
Il1rl1	Pcdha7	Tmem132b
Ino80d	Pcdha8	Tmem150c
Itpr1	Pcdha9	Tnik
Kcnq4	Pgam1	Unc5d
Krt4	Poln	Wdr96
Magi2	Prkdc	Wnt7b
Mctp1	Rab30	Wwox
Metrn	Rbfox3	Xpo7
Mmp9	Rbm47	Zfp345
Mpdz	Rhof	Zic5
Myom2	Rn45s	

spectrometry data [9] that show overall reduced 5hmC in later passage. This result implies that the epigenetic stability of ES cells is impacted by prolonged in vitro culture. This is an important issue for both the safety and efficacy of stem cell-derived tissues in cell-replacement therapies as well as the appropriate interpretation of experimental models. Monoallelic gene expression, including genomic imprinting, is primarily regulated through epigenetic mechanisms and thus can serve as a useful model of epigenetic stability. As expected, our analysis identified five imprinted genes with decreased 5hmC (*Plagl1*, *Sfmbt2*, *Gpr1*, *Kcnq1*, and *Kcnq1ot1*) as well as one imprinted gene with increased 5hmC (*Pcdha4-g*).

The role of 5hmC in disease remains unclear. A recent study suggests that genome-wide loss of 5hmC is an epigenetic feature of neurodegenerative Huntington's disease [29]. The authors identified 559 genes with decreased 5hmC in the diseased mice compared to healthy controls. A considerable fraction of these disease-specific genes were uncovered in our differential 5hmC analysis in ES cells. This included 26 of 299 and 11 of 125 genes (overlapping *P*-value < 8e-5) with decreased and

increased 5hmC, respectively. These results suggest that one potential consequence of decreased epigenetic stability over time in ES cells is the acquisition of pathological epimutations.

The observed bias toward loss of 5hmC in ES cells upon long-term culture may also suggest stem cell properties, such as pluripotency, are affected. Ficz and colleagues [30] showed that knockdown of Tet1/Tet2 in mouse ES cells downregulates epigenetic reprogramming and pluripotency-related genes such as Esrrb, Klf2, Tcf1, Zfp42, Dppa3, Ecat1, and Prdm14. Decreased expression was concomitant with both decreased 5hmC and increased 5mC at the gene promoters. In our differential 5hmC analysis in ES cells, we observed decreased 5hmC at three of these genes: Ecat1, Esrrb, and Zfp42. Together, we conclude that MOABS can be used effectively to infer differential 5hmC using RRBS and oxBs-seq.

DISCUSSION

While progress in next-generation sequencing allows increasingly affordable BS-seq experiments, the resulting data generated pose significant and unique bioinformatics challenges. The lack of efficient computational methods is the major bottleneck that prevents a broad adoption of such powerful technologies. In response to this challenge, we developed MOABS, an accurate, comprehensive, efficient, and user-friendly pipeline for BS-seq data analysis. The MOABS analysis is novel and significant in two major aspects: (1) MOABS CDIF value provides an innovative strategy to combine statistical *P*-value and biological difference into a single metric, which will bring biological relevance to the interpretation of the DNA methylation data. It not only represents a novel concept for differential methylation, but also for any genome-wide assay in general, such as gene expression. (2) MOABS does not sacrifice resolution with low sequencing depth. By relying on the beta-binomial hierarchical model and empirical Bayes approach, MOABS has enough power to detect single-CpG-resolution differential methylation in low-CpG-density regulatory regions, such as TFBSSs, with as low as 4–10 fold. The low-depth BS-seq experimental design enables remarkable cost reduction per sample. With the same budget, we would recommend low-depth (e.g., 10 fold) BS-seq on more biological samples, which in most scenarios will provide greater biological insights than high-depth BS-seq on fewer samples.

In summary, as DNA methylation is increasingly recognized as a key regulator of genomic function, deciphering its genome-wide distribution using BS-seq in numerous samples and conditions will continue to be a major research interest. MOABS significantly increases the speed, accuracy, statistical power, and biological relevance of the BS-seq data analysis. We believe that MOABS's superior performance will greatly facilitate the study of epigenetic regulation in numerous biological systems and disease models.

METHODS

DISTRIBUTION FOR DIFFERENCE OF TWO BINOMIAL PROPORTIONS

In the Supplementary we show that a methylation ratio p inferred from k methylated cytosines out of n total reads follows a beta distribution from the Bayesian perspective. The probability density function is

$$f(p; n, k) = Be(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp}, \quad (4.1)$$

where $\alpha = k + \alpha_0$ and $\beta = n - k + \beta_0$ if $Be(\alpha_0, \beta_0)$ is priori distribution for p . We also give formulas to numerically calculate the confidence interval for the single binomial proportion p under observed (n, k) .

The methylation ratio difference at a defined genomic locus from two biological samples is the difference of two binomial proportions $p_1 - p_2$. Many methods have been proposed to estimate the confidence interval of $p_1 - p_2$ and their merits have been subject to decades of considerable debate [22,31–36]. No comprehensive comparison of methods is currently available. This motivated us to turn to the direct and exact numerical calculation of confidence interval from Bayesian perspective.

Let $t = p_1 - p_2$, where p_i is the proportion for the sample i with observation n_i and k_i . Since the joint probability density of such observation is $f(p_1, n_1, k_1)f(p_2, n_2, k_2)$, the PDF for t is

$$f(t) = \int_0^1 dp_2 f_1(p_2 + t) f_2(p_2) = \int_0^1 dp_1 f_1(p_1) f_2(p_1 - t), \quad (4.2)$$

where $f_i(p_i) \equiv f(p_i; n_i, k_i)$. Boundary conditions like the proportional area condition and minimal length condition can be applied to get unique solutions for (a, b).

DISTRIBUTION FOR DIFFERENCE OF DIFFERENCE

Let $t = p_1 - p_2$, where p_i is the proportion for the assay i with observation n_i and k_i . In the ox-BS experiments, p_2 is the oxBS methylation ratio and p_1 is the RRBS methylation ratio, and t is the 5hmC methylation ratio. Since the joint probability density of such observation is $f(p_1, n_1, k_1)f(p_2, n_2, k_2)$, the PDF for t is

$$f(t) = \int_0^1 dp_2 f_1(p_2 + t) f_2(p_2) = \int_0^1 dp_1 f_1(p_1) f_2(p_1 - t), \quad (4.3)$$

where $f_i(p_i) \equiv f(p_i; n_i, k_i)$.

Let $t' = p'_1 - p'_2$, where ' denotes the other sample. To be clear, call the two samples S and S' . In general, we want to know the difference of the two 5hmC ratios, that is, $t - t'$. Let $x = t - t'$, we can immediately obtain the distribution of difference of 5hmC ratio between two samples by

$$f(x) = \int_{-1}^1 f(t) f'(t - x) dt = \int_{-1}^1 f(t' + x) f'(t') dt', \quad (4.4)$$

where $f(t)$ and $f'(t')$ are the distributions of 5hmC ratio for sample S and S' respectively. After distribution of difference of 5hmC ratio between two samples is obtained, similarly confidence interval, credible difference, and similarity test P -value can be calculated.

DISTRIBUTION FOR MEASUREMENTS WITH REPLICATES

Here we use the exact numerical approach to calculate the distribution of p at observance of (m_i, l_i) with m_i as total count for replicate i and l_i as methylated count for replicate i . Let us start with two replicates. We try to fit this unknown distribution of (m_1, l_1) at observance (m_2, l_2) and $f(p; \alpha, \beta)$ into a beta distribution $f(p; \alpha, \beta)$. The parameter estimation is based on the following formula

$$P(k_i; n_i, \alpha, \beta) = \int_0^1 f(k_i, n_i, p) f(p; \alpha, \beta) dp, \quad (4.5)$$

where $P(k_i; n_i, \alpha, \beta)$ is the probability to observe (n_i, k_i) under the beta distribution $f(p; \alpha, \beta)$, and $f(k_i; n_i, p)$ is the binomial distribution, that is, the probability to observe (n_i, k_i) under a specific true ratio p . For N number of replicates, (α, β) may be estimated by maximizing the log-likelihood function

$$\log L(\alpha, \beta) = \sum_{i=1}^N \log \left(C_{n_i}^{k_i} \frac{B(\alpha + n_i, \beta + k_i - n_i)}{B(\alpha, \beta)} \right), \quad (4.6)$$

where the expression inside log is the probability $P(k_i; n_i, \alpha, \beta)$ defined in Eq. (4.5) and $B(\alpha, \beta)$ is the beta function.

ACKNOWLEDGMENTS

We are grateful to Wei Xie for sharing the mouse methylome data, and Grant A. Challen for critical reading of this manuscript. This work was supported by CPRIT RP110471-C3 and NIH R01HG007538 (to WL).

REFERENCES

- [1] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012; 13(7):484–92.
- [2] Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, NY)* 2009;324(5929):930–5.
- [3] He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 2011;333(6047):1303–7.
- [4] Song CX, Yi C, He C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* 2012;30(11):1107–16.
- [5] Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 2010;11(3): 191–203.
- [6] Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 2010;11(3):204–20.
- [7] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008; 454(7205):766–70.
- [8] Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 2008;452(7184):215–9.
- [9] Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 2012;336(6083):934–7.
- [10] Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 2012;149(6):1368–80.
- [11] Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Vasanthakumar A, Gu H, Xi Y, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* 2012;44(1):23–31.

- [12] Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012; 13(10):R87.
- [13] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;13(10):705–19.
- [14] Hansen KD, Langmead B, Irizarry RA. BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;13(10):R83.
- [15] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 2009;10: 232.
- [16] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009; 462(7271):315–22.
- [17] Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods* 2010;7(2): 133–6.
- [18] Rhee Ho S, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 2011;147(6):1408–19.
- [19] Feng J, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* 2012.
- [20] Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* 2010;107(Suppl. 1):1757–64.
- [21] Newcombe MMNRG. Score intervals for the difference of two binomial proportions. *Methodologic Notes On Score Intervals*
- [22] Kawasaki Y. Comparison of exact confidence intervals for the difference between two independent binomial proportions. *Adv Appl Stat* 2010;15(2):157–70.
- [23] Quan DJBH. Exact confidence limits for binomial proportions—Pearson and Hartley revisited 1990;39: 391–7.
- [24] Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, Li W. BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics* 2013.
- [25] Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* 2012;148(4): 816–31.
- [26] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 2011;480(7378):490–5.
- [27] Hannah R, Joshi A, Wilson NK, Kinston S, Gottgens B. A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Exp Hematol* 2011;39(5):531–41.
- [28] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;34(3):267–73.
- [29] Wang F, Yang Y, Lin X, Wang JQ, Wu YS, Xie W, Wang D, Zhu S, Liao YQ, Sun Q, et al. Genome-wide loss of 5-hmC is a novel epigenetic feature of Huntington’s disease. *Hum Mol Genet* 2013.
- [30] Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 2011; 473(7347):398–402.

- [31] Newcombe MMNRG. Score intervals for the difference of two binomial proportions. Methodologic Notes On Score Intervals.
- [32] Pradhan V, Tathagata B. Confidence interval of the difference of two independent binomial proportions using weighted profile likelihood. Commun Stat Simulat Comput 2008;37(4):645–59.
- [33] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998;17(8).
- [34] Santner TJ, Pradhan V, Senchaudhuri P, Mehta CR, Tamhane A. Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions. Comput Stat Data Anal 2007;51(12): 5791–9.
- [35] Tamhane A, Coe PR. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. Commun Stat Simulat Comput 1993;22(4).
- [36] Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22.

SUPPLEMENTARY METHODS

METHYLATION RATIO OF ONE LOCUS FOLLOWS A BETA DISTRIBUTION

In bisulfite sequencing, one cytosine locus is sequenced n times. Out of n reads, k reads show cytosine and $n - k$ reads show thymine as a result of bisulfite conversion of unmethylated cytosine. The methylation ratio of this locus, p , is inferred from the pair (n, k) . In other words, a population of size s and true proportion p are sampled n times with observed success k . Given s, p, n , the probability of obtaining k successes is probability of obtaining k from pool of sp successes and obtaining $n - k$ failures from pool of $s(1 - p)$ failures. Population size s is usually the number of cells and can be considered as infinity, resulting each of n trials as an independent event. So, the probability of obtaining k obeys binomial distribution,

$$P(k; p, n) = \lim_{s \rightarrow \infty} \frac{C_{sp}^k C_{s(1-p)}^{n-k}}{C_s^n} = C_n^k p^k (1-p)^{n-k}, \quad (1.1)$$

where P is the probability and C_n^k is binomial coefficient. It is also the probability density distribution function $f(k; p, n) = C_n^k p^k (1-p)^{n-k}$ because it is a discrete distribution.

This is a function of k and we want to estimate the proportion p . Since each trial is independent and binomial, the inferred true proportion is called binomial proportion. Here and throughout this article, we estimate it regarding it as a random variable, that is, from the Bayesian perspective. Under the uniform prior for p in $(0, 1)$, the probability of $p_0 \in (p - \frac{dp}{2}, p + \frac{dp}{2})$ is

$$\begin{aligned} dP(p; n, k) &= \frac{f(k; n, p) dp}{\sum_0^1 f(k; n, p) dp} \\ &= \frac{C_n^k p^k (1-p)^{n-k} dp}{\int_0^1 C_n^k p^k (1-p)^{n-k} dp} \\ &= \frac{\int_0^1 p^k (1-p)^{n-k} dp}{\int_0^1 p^k (1-p)^{n-k} dp}. \end{aligned}$$

And hence the PDF (probability density function) of p is

$$f(p; n, k) = \frac{dP(p; n, k)}{dp} = \frac{p^k(1-p)^{n-k}}{\int_0^1 p^k(1-p)^{n-k} dp}, \quad (1.2)$$

which is recognized as beta distribution PDF

$$Be(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp} \quad (1.3)$$

with $\alpha = k + 1$ and $\beta = n - k + 1$.

So for the methylation analysis, k follows binomial distribution (1.1), and the methylation ratio for this locus, p follows beta distribution (1.2). Under a more general prior distribution like beta distribution $p \sim Be(\alpha_0, \beta_0)$, distribution for p is

$$f(p; n, k) = Be(p; \alpha, \beta),$$

with $\alpha = k + \alpha_0$ and $\beta = n - k + \beta_0$.

CI FOR SINGLE BINOMIAL PROPORTION

One immediate question is what is the CI (confidence interval) of the methylation ratio. In 2008 Pires and Amado [1] compared 20 methods of interval estimators for single binomial proportion. These estimators are either in analytical form with asymptotic approximations, or in numerical solutions. Since the sequencing depth could vary from one to hundreds fold, and the methylation ratio of most loci is close to either 0 or 1, the validity of asymptotic approximation becomes questionable. We used the exact numerical method described by Brenner and Quan [2]. It is a *Bayesian* confidence interval under uniform prior. The confidence interval (a, b) for proportion p is straightforward when its distribution is known.

$$\int_a^b f(p; k, n) dp = 1 - \alpha, \quad (2.1)$$

where f is beta distribution PDF and α is type-I error, usually at 0.05.

We propose a physically meaningful “proportional area condition,” that is, requiring the two-sided tail areas being proportional to two areas under the p distribution curve separated by the mean,

$$\frac{P(p < a)}{P(p > b)} = \frac{P(p < \mu)}{P(p > \mu)}, \quad \text{where } \mu = \frac{k}{n}. \quad (2.2)$$

The usual choice of minimal length condition also satisfies the needs well. In the C++ and R source code two other alternative conditions are made available for general use except for DNA methylation because these two conditions, symmetric width $b - u = u - a$ and symmetric area $P(p > b) = P(p < a)$, need additional processing for the abundant situations when $p = 0$ or 1.

The minimal length condition is equivalent to

$$\begin{cases} f(a; k, n) = f(b; k, n), k \neq 0 \\ a = 0, b = 1 - \alpha^{\frac{1}{n+1}}, k = 0 \end{cases}. \quad (2.3)$$

Combining Eq. (2.1) with (2.2) or (2.2) with (2.3), (a, b) can be uniquely solved.

The methylation ratios of millions of cytosines in genome often obey a bimodal distribution. We may use this bimodal distribution as the prior distribution. The influence of a nonuniform prior distribution, for example a U-shape prior $p \sim Be(0.5, 0.5)$, will in general make 0% methylation CI narrower but 50% methylation CI wider and have more influence at low depth than high depth.

CI FOR DIFFERENCE OF TWO BINOMIAL PROPORTIONS IN DETAIL

We showed that methylation ratio p with k methylated cytosines out of n total reads follows beta distribution from the Bayesian perspective. The probability density function is

$$f(p; n, k) = Be(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp}, \quad (3.1)$$

where $\alpha = k + \alpha_0$ and $\beta = n - k + \beta_0$ if $Be(\alpha_0, \beta_0)$ is prior distribution for p . We also give formulas to numerically calculate the confidence interval for the single binomial proportion p under observed (n, k) .

The question is the difference of two binomial proportions, for example, the methylation ratio difference of the same genomic locus from two biological samples. Many methods have been proposed to estimate the confidence interval of $p_1 - p_2$. Newcombe [3] compared 11 methods, including 9 asymptotic methods and 2 exact methods, and concluded that the Wilson [4] score method with modifications has superior performance. Santner et al. [5] in a small-sample study compared the score method with other four exact methods and arrived at an opposite conclusion where score method is worst and the CT method [6] has best small-sample performance. However, Nurminen and NewCombe replied with disagreement [7]. Much of the debates come from different evaluation criteria, for example, whether coverage probability is minimum or average at $1 - \alpha$ and whether minimum CI length or symmetric tail area is looked for. Pradhan and Banerjee [8] proposed a weighted likelihood method and concluded it is better than score method. Kawasaki [9] compared several exact methods and recommended some revisions. The various methods discussed in each comparison article are just a portion of all available methods. There does not exist a comprehensive comparison of currently available methods. That motivated us to turn to the direct and exact numerical calculation of confidence interval from Bayesian perspective.

Let $t = p_1 - p_2$, where p_i is the proportion for the sample i with observation n_i and k_i . Since the joint probability density of such observation is $f(p_1; n_1, k_1)f(p_2; n_2, k_2)$, the PDF for t is

$$f(t) = \int_0^1 dp_2 f_1(p_2 + t) f_2(p_2) = \int_0^1 dp_1 f_1(p_1) f_2(p_1 - t), \quad (3.2)$$

where $f_i(p_i) \equiv f(p_i; n_i, k_i)$.

The probability

$$\begin{aligned}
 P(p_1 - p_2 > d) &= P(t > d) = \int_d^1 dt f(t) \\
 &= \int_d^1 dt \int_0^1 dp_1 f_1(p_1) f_2(p_1 - t) \\
 &= \int_d^1 dp_1 f_1(p_1) \int_d^1 dt f_2(p_1 - t) \\
 &= \int_d^1 dp_1 f_1(p_1) \int_{p_1-1}^{p_1-d} dy f_2(y) \\
 &= \int_d^1 dp_1 f_1(p_1) \int_0^{p_1-d} dy f_2(y) \\
 P(p_1 - p_2 > d) &= \int_0^1 dp_1 f_1(p_1) I_2(p_1 - d)
 \end{aligned} \tag{3.3}$$

where substitution of variable $p_1 - t = y$ is made and $I_2(x)$ is cumulative distribution function for beta distribution function $f_2(x)$.

Suppose the confidence interval for t is (a, b) ,

$$\begin{aligned}
 1 - \alpha &= P(t > a \& t < b) \\
 &= P(t > a) + P(t < b) - 1 \\
 &= P(t > a) - P(t > b)
 \end{aligned} \tag{3.4}$$

Similar conditions as in the single proportion case, like the proportional area condition and minimal length condition, can be applied to get unique solutions for (a, b) .

IDENTIFICATION OF DMCs FOR TWO OR MORE SAMPLES

Previously methods define a DMC by requiring methylation ratio difference and Fisher's exact P -value, all reach some threshold values. Now, the CDIF alone is good enough to define and rank DMCs. In MOABS, the default criterion for DMC is:

$$|v| > |v_0|, \tag{4.1}$$

where v_0 is either arbitrary or determined by controlling FDR, estimated by permutation of sample labels, to be 5% (or other arbitrary cutoffs). This condition may be extended to multiple samples:

$$v = \max\{v_{ij}\} \tag{4.2}$$

where v_{ij} denotes the credible difference between sample i and sample j .

IDENTIFICATION OF DMRs FOR TWO SAMPLES BY SIMPLY GROUPING DMCs

After DMCs are identified from methylome, one may simply group DMCs into a DMR. One need specify the maximum gap distance between two DMCs, and how many nondifferential CpGs are allowed in a DMR. The minimal number of DMCs can be determined by controlling FDR to be 5%. The NULL distribution for FDR calculation is obtained by shuffling the coordinates of all CpGs in the genome followed by DMR calling using the same method.

IDENTIFICATION OF DMRs FOR TWO SAMPLES BY HIDDEN MARKOV MODEL

Here, we propose a first-order hidden Markov model approach to combine neighboring CpGs into DMR. The state of i^{th} cytosine is denoted as S_i where S_i can take three hidden states for a two-sample comparison:

$$\begin{aligned} S_0: & \text{ hypo-methylation state if } p_2 - p_1 < -v_0; \\ S_1: & \text{ no difference state if } |p_2 - p_1| < v_0; \\ S_2: & \text{ hyper-methylation state if } p_2 - p_1 > v_0; \end{aligned} \quad (6.1)$$

where v_0 is a preset parameter and marks the characteristic threshold of difference for underlying data set. In MOABS, this parameter is determined in DMC scan stage by controlling FDR, estimated by permutation of sample labels, to be 5%. We model the neighbor correlation by first-order Markov chain

$$\Pr(S_i) = \Pr(S_i|S_{i-1}), \quad (6.2)$$

which means that the state of site i is directly influenced by previous site $i-1$.

Each observation for each site is a combination of four numbers from two samples: $x = (n_1, k_1, n_2, k_2)$. In this problem, we are given the observation sequence from all sites and we want to find the HMM that maximizes the probability of observation sequence. The HMM is characterized by initial state π_0 , transition probability matrix $A = \Pr(S_i|S_{i-1})$, and emission probability matrix $B = \Pr(x_i|S_i)$.

The initial state π_0 can just takes value S_1 , though its value does not matter since there are millions of CpGs in the genome. By assuming a site is in one of the three states, the emission probability for the i^{th} site to observe $x = (n_1, k_1, n_2, k_2)$ when the state of the site is S_i can be derived as

$$\Pr(n_1, k_1, n_2, k_2|s_i) = \frac{\iint_{S_i} dp_2 dp_1 f(k_1; n_1, p_1) f(k_2; n_2, p_2)}{\int_0^1 f(k_1; n_1, p_1) dp_1 \int_0^1 f(k_2; n_2, p_2) dp_2} \quad (8.3)$$

Since there are millions of sites and there is a high chance of repeated observations, MOABS uses a lookup table to avoid repeated computation of numerical integrations. The state transition probability matrix can be trained using the forward-backward algorithm. In the training process, the initial state and the emission probability matrix are fixed while the state transition probability is the only model variable. Since the training is computationally intensive, MOABS may choose only a subset of all cytosine sites in the genome, like first one million sites in chromosome 19 or locus provided by users. After the change of likelihood of the model is smaller than a given threshold or maximum number of iterations is reached, the optimal hidden state for each site is obtained. Consecutive sites with S_0 (or S_1) states are merged as hypo-DMR (or hyper-DMR).

IDENTIFICATION HYPOMETHYLATED REGIONS FROM ONE SAMPLE

Similar to DMR detection, MOABS used a two-state first-order hidden Markov model (HMM) to detect highly methylated and lowly methylated regions from a single sample. Random shuffle of all the CpGs in the genome, followed by the same procedure, generates a NULL distribution to control the FDR.

SUPPLEMENTARY REFERENCES

- [1] Pires AM, Amado C. Interval estimators for a binomial proportion: comparison of 20 methods. *REVSTAT* 2008;6.
- [2] Brenner DJ, Quan H. Exact confidence limits for binomial proportions—Pearson and Hartley revisited. 1990;39:391–97.
- [3] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of 11 methods. *Stat Med* 1998;17.
- [4] Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927;22.
- [5] Santner TJ, Pradhan V, Senchaudhuri P, Mehta CR, Tamhane A. Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions. *Comput Stat Data Anal* 2007;51:5791–5799.
- [6] Coe PR, Tamhane AC. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Commun Stat Simul Comput* 1993;22.
- [7] Newcombe RG, Nurminen MM. Score intervals for the difference of two binomial proportions. *Methodologic Notes on Score Intervals*.
- [8] Pradhan V, Tathagata B. Confidence interval of the difference of two independent binomial proportions using weighted profile likelihood. *Commun Stat Simul Comput* 2008;37:645–659.
- [9] Kawasaki Y. Comparison of exact confidence intervals for the difference between two independent binomial proportions. *Adv Appl Stat* 2010;15:157–170.

DATA ANALYSIS OF ChIP-SEQ EXPERIMENTS: COMMON PRACTICE AND RECENT DEVELOPMENTS

5

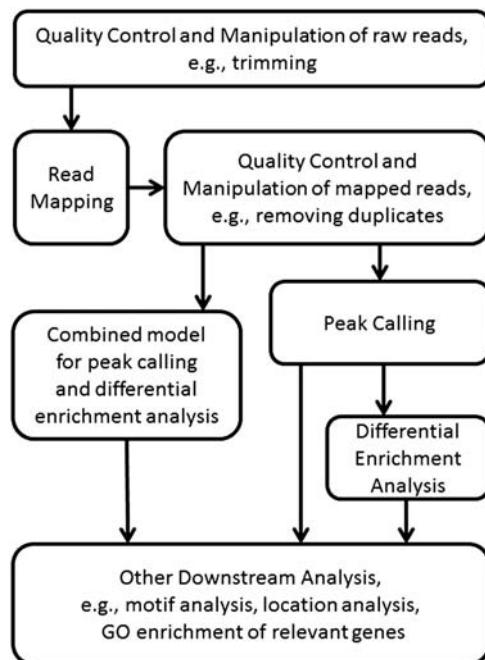
Qi Zhang

Department of Statistics, University of Nebraska—Lincoln, Lincoln, NE, United States

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is the mainstream method for the genome-wide investigation of histone modifications and transcription factor (TF)—DNA interactions. A web lab experiment protocol for ChIP-Seq typically involves the following steps. It starts with crosslinking DNA with the protein of interest, then DNA are sheared by sonication, and the fragments associated with the protein or other biological signals of interest are enriched by immunoprecipitation using the antibody specific to the targets under investigation. After purification and size selection, the remaining fragments are sequenced using a platform such as Illumina. The bioinformatics analysis pipeline (Fig. 5.1) of the sequenced short reads usually starts with mapping the reads to the reference genome. Then the ChIP-Seq peaks, that is, the genomic regions enriched for ChIP-Seq reads, could be identified based on the mapped reads. After that, various downstream analyses could be performed, such as differential enrichment of ChIP-Seq across two or more conditions, allele-specific enrichment of ChIP-Seq signal, sequence motif analysis, and functional annotation of the peaks.

The targets of ChIP-Seq experiments are very diverse, which will impact the specifics of the data analysis protocol. For example, transcription factor binding sites typically yield narrow ChIP-Seq peaks (narrow peaks), while many types of histone modification lead to broad regions of interest (broad peaks) [1,2]. Nevertheless, the workflow shown in Fig. 5.1 generally applies to most types of ChIP-Seq data. Additionally, it can also handle some other NGS data types with similar structure, such as DNase-Seq [3] and ATAC-Seq [4] for detecting open-chromatin regions.

The goal of this chapter is to provide the readers a broad review of the common practice and the recent development in the data analysis involving ChIP-Seq, and suggest tools and new perspectives. In what follows, we will discuss the important issues and review popular tools for each step of ChIP-Seq data analysis, and then review the all-in-one analysis pipelines and the issues and tools for the allele-specific enrichment analysis in ChIP-Seq, an emergent complementary analysis for ChIP-Seq data.

**FIGURE 5.1**

Standard data analysis workflow for ChIP-Seq.

THE DESIGN OF CHIP-SEQ

Many factors in ChIP-Seq experiment influence the choice of the analysis methods and results, and can be more or less controlled by the researcher, namely the sequencing coverage, read length, the choice of pair-end (PE) or single-end (SE), and whether or not to include a control sample.

The sequencing depth has the most well-known impact on the analysis results [5–8]. Generally speaking, the numbers of detected peaks increase as the coverage increases. When ChIP-Seq data with sufficient high coverage are available, the necessary coverage can be determined based on saturation analysis [5,6]. Such saturation point has been reported for organisms such as fruit fly, but there appears to be no such point for human and mouse based on the currently available data. For human sample, recommendations on the number of mapped reads needed range from 20 million to 60 million reads per study [6,9].

The read length and whether PE or SE reads will be generated can also be decided by the researcher. Using comparable mapping rules, longer reads and PE reads typically lead to lower genome-wide coverage with higher accuracy, and higher coverage in the repetitive regions [8]. PE reads and long reads are generally believed to be more accurate than SE reads and short reads, especially if the quality in the 3' ends is reasonably good.

Another aspect of the design of ChIP-Seq is whether or not to include a control sample [9]. In ChIP-Seq experiments, the coverage is not uniform genome-wide even if there is no biological signal of interest (e.g., TF–DNA binding, histone modification). This is because the chromatin accessibility during fragmentation is uneven, and the reads will be enriched in the open-chromatin regions even if they are not associated with the signal of interest. Thus it is desirable to have a control sample where the experiment was done exactly as the ChIP-Seq experiment except that there is no enrichment for the specific ChIP target. Control samples, if properly utilized, can reduce the false positives in peak identification due to the bias caused by the chromatin structure.

THE QUALITY OF ChIP-SEQ DATA

It is important to understand the quality of the ChIP-Seq data before the final analysis for a study. There are many aspects of the quality that need to be considered. We remark that some of the following quality control (QC) measures are defined for the raw reads, while the others are for the mapped reads.

Position-wise content and quality of the reads. If the sequence content at certain positions (usually at 5' end) is dominated by one or two letters, it could be part of the adaptor that needs to be further trimmed. In many cases, the quality towards 3' may become low, and trimming this reads, either directly or during the read mapping step, may be helpful. FASTQC [10] is the most common software for such tasks.

Library complexity. Library complexity can be measured by the nonredundant fraction (NRF), the fraction of nonredundant reads in all mapped reads, and the PCR bottleneck coefficient (PBC), the proportion of mapped genomic locations with exactly one uniquely mapped read. Large NRF and PBC imply good library complexity. Take PBC for example: ENCODE [9] advocates that “Provisionally, 0–0.5 is severe bottlenecking, 0.5–0.8 is moderate bottlenecking, 0.8–0.9 is mild bottlenecking, while 0.9–1.0 is no bottlenecking.” [11]

Signal-to-noise ratio. The ChIP-Seq reads from genomic regions without the targeted biological signal could be considered as “noise,” and a measure of the signal-to-noise ratio helps the researchers understand how informative their data are. One intuitive measure for this is the proportion of reads from the peaks. But this measure depends on the peak caller used. An alternative measure without peak calling is the cross-correlation profile measured by the normalized strand cross-correlation coefficient (NSC) and relative strand cross-correlation coefficient (RSC). However, these measures could be sensitive to the experimental condition and the other biological effects, and the results from distinct cell types or experiments may be different. Additionally, these measures are more appropriate for single-end reads targeting at biological signals that result in narrow ChIP-Seq peaks. The cross-correlation measures and the PBC can be calculated by the R package SPP [5].

MAPPING ChIP-SEQ READS

The first step for the analysis of sequencing reads is to map the reads to the relevant reference genome. For this task, Bowtie [12] and BWA [13] remain the dominant tools. In our previous study, we found that the alignment rates and accuracy of Bowtie and BWA are similar when their alignment rules are comparable and if only the uniquely mapped reads (uni-reads) are kept [8]. Thus the practical differences between them are largely from the difference in configuration and the handling of multiply

mapped reads (multi-reads). Bowtie—v mode is the most transparent alignment rule, as the only criterion applied is the number of mismatches. When the read-length is long (>75 bp), the quality towards the 3' ends of the reads could become low, thus it may be helpful trimming the low quality base pairs at 3' ends. It can be achieved by the “-q” argument in BWA. Comparing with trimming before mapping, this argument allows different trimming for different reads, which will increase the mapping rate. There are many other popular mapping tools such as Bowtie2 [14] and GEM [15].

In many situations, a nonnegligible proportion of reads can be mapped to multiple locations of the genome. This issue is particularly important for short single-end reads, for organisms with a significant amount of long repetitive elements in their genome such as segmental duplication regions, and when the regulatory role of such regions is potentially of interest. Many software deals with this issue. For example, CSEM [16] allocates the multi-reads proportionally based on the uni-read counts and an EM algorithm. It has been shown that including multi-reads can increase the mapping coverage and the number of identified peaks. Unsurprisingly, these additional peaks are enriched in repetitive regions, which illustrates that the properly processed multi-reads provide new biological insights into gene regulation. When applicable, multi-read allocation can be further improved using data-integration based approaches by incorporating additional information from relevant data type such as the copy number variation, DNase-seq, and other ChIP-Seq data [17,18]. The input of these tools is the preliminary read mapping result that contains all candidate locations of the multi-reads, and the external information to be incorporated, and the output is a refined alignment result that reports the allocation proportion at each candidate location for each multi-read. Similar algorithm has also been applied to the multi-read allocation problem for data from newer technologies such as CLIP-Seq [19].

The reference genome for one species is never the most accurate representation of the genomic sequences of all individuals in this species. For human and other vertebrates, their genomes are diploid, which means that each subject has two copies of genomic sequence, and we call each copy an allele. There may be even differences between the two alleles within the same individual. The genomic differences between individuals and between the alleles are called structural variations (SV), for example, single nucleotide polymorphisms (SNPs), short insertions and deletions (indels). They have huge impact in read mapping, especially in the immediate neighborhood of such SVs [20–23]. When mapping the reads from the minor alleles (the alleles that are different from the reference genome due to SVs), SVs will be considered as mismatches, thus reduce the coverage on the minor allele. It does not only reduce the total coverage at this genomic location, but also create a minor allele-specific bias in ChIP-Seq signal, which will bias the allele-specific enrichment analysis. RefEditor [23] is a tool for building personalized genome for NGS read mapping in general. It is particularly useful when the genotyping information is available for the samples under consideration. In the literature of allele-specific enrichment analysis, allelic-unbiased mapping is particularly important, and we will continue this discussion in the corresponding section.

PEAK CALLING

There could be more than 100 methods for peak calling from ChIP-Seq. The models underlie these methods and the features of these algorithms are very different. For example, SPP [5] and the original version of MACS [24] only utilize the ChIP-Seq data themselves, and the control sample when it is available. MOSAiCS [25] uses the mappability scores and GC content for a better estimate of the

background model, especially when there is no control sample. GEM [26] also takes advantage of the sequence information in the candidate regions to improve the resolution of the peaks. It is notoriously hard to compare all peak calling algorithms, or choose the “best” for a specific application, which is further complicated by the numerous postpublication developments of these software. Earlier comparisons tend to focus on the description of the differences in number of peaks, and partial external validation using ChIP-chip, qPCR, sequence motif, etc., when such information is available and relevant [27–29].

We do not intend to compare the peak callers comprehensively, or suggest specific algorithms in this chapter. Instead, we will discuss many issues need to be considered in peak calling. First, the shape of peaks matters. Algorithms such as SPP has an underlying model that is more appropriate for narrow peaks, and the hidden Markov structure in MOSAiCS-HMM [30] makes it more suitable for broad peaks. Second, the design of the ChIP-Seq experiment needs to be taken into consideration. Some earlier peak callers cannot properly handle paired-end reads, while the newer development of MACS2 (the developing version of MACS) and MOSAiCS can. Third, if the target is known to be sequence specific, it may helpful to incorporate the sequence information using peak callers such as GEM and the newer development of MOSAiCS and MACS2. Fourth, when there is no control sample, tools such as MOSAiCS incorporate variables such as GC content and mappability in its model to correct the unevenness of the background due to the open-chromatin structure. SPP can also be extended with a GC bias correction, and the results are shown to be less biased than using control samples [31]. Fifth, the consistency across replicates is desired. Most of these peak callers were not originally designed for replicated experiments. It was a common practice that the replicates were pooled before peak calling. Alternatively, a “robust” peak list can be decided in a postprocessing manner from the peak lists from the individual replicates. The most naïve way of doing so is overlapping the peak lists obtained from different replicates. Irreproducible discovery rate (IDR) revolutionized the practice of ChIP-Seq peak calling with replicates, as it explicitly models the ranking consistency across a pair of replicates. ENCODE consortium has been making a significant amount of effort in promoting IDR in ChIP-Seq data analysis, and have coupled many peak callers with IDR [9].

DIFFERENTIAL ENRICHMENT DETECTION

Just like differential gene expression analysis, detecting the differential enrichment in ChIP-Seq signal between two or more conditions is a common downstream analysis after peak calling. There is no dominant differential enrichment (DE) analysis tool specifically for ChIP-Seq. In fact, DESeq2 [32] and edgeR [33], two popular software for differentially expressed gene detection, are often used in DE analysis for ChIP-Seq as well. In DE analysis, the peaks detected from the two conditions are being merged first to form a list of candidate regions for DE in ChIP-Seq. Then edgeR or DESeq2 can detect DE in ChIP-Seq, treating each such candidate region as a “gene.” There are many tools developed for ChIP-Seq that share a similar spirit, for example, DiffBind [34], ChIPComp [35], DBChIP [36]. Their input includes a predefined list of peaks as the candidate regions, and the read files, and they output which of these candidate regions show a significant differential enrichment in ChIP-Seq signal. Many other more flexible approaches are window based and/or combine the steps of peak calling and differential enrichment in one single model (PePr [37], MACS2 bdgdiff). Thus they do not require calling the peaks ahead of time. Differential enrichment detection and peak calling can be viewed as a data

integration problem, that is, analyzing multiple related ChIP-Seq experiments. Thus the models for ChIP-Seq data integration can also be applied. For example, jMOSAICs [38] was originally designed for the integrative analysis of multitype ChIP-Seq data and segmenting the genome based on the chromatin states, but can also be used for peak calling and differential binding detection. MultiGPS [39] shared a similar spirit, but it was originally designed for detecting condition-specific binding by the joint modeling of the same type of ChIP-Seq data under multiple conditions.

ALL-IN-ONE DATA ANALYSIS PIPELINES FOR CHIP-SEQ

A standard workflow of ChIP-Seq data analysis usually includes quality control, read mapping, and peak calling. It could be followed by differential binding detection if multiple conditions are being compared. These pipelines may also provide various functionalities of other downstream analyses such as sequence motif analysis, annotation by genes, gene set enrichment analysis, and various visualizations. Next, we will introduce the all-in-one pipelines by the following three categories: (1) stand-alone software; (2) online platform; and (3) packages in a single computational platform (e.g., R packages).

HICCHIP [40] is stand-alone software for ChIP-Seq data analysis, which incorporates many open-source software for ChIP-Seq analysis. It examines the reads quality using FastQC, maps the reads by BWA, and processes the mapped reads for additional quality control. Then it calls peaks using SICER [41] and MACS2, and investigate the cross-replicate consistency by IDR. It also provides additional functionality of de novo motif discovery using MEME [42], GO enrichment analysis, etc. It can be run on either a single Linux machine, or on a Sun Grid Engine (SGE) cluster. Many other ChIP-Seq data analysis pipelines, such as cisGenome [44], start from the mapped reads and also provide modules for motif analysis and gene-based annotation.

Even though the cost of computing has been dropping down in the recent decade, it could still be troublesome or even prohibitive for many biological researchers to learn and install all software needed for their ChIP-Seq data analysis. Many computational platforms also host servers with a wide range of tools installed and a web-based interface, so that the biological researchers can analyze their data via mouse-clicking [43]. Galaxy [45] is one of the most successful open-source web-based platforms for biomedical data analysis, and it provides a wide variety of tools along with recommended workflow for ChIP-Seq data analysis. The biomedical researchers can either use the default settings of the recommended pipeline or customize their ChIP-Seq data analysis protocol by choosing other tools through the Galaxy web-based interface. Many institutions host their own bioinformatical infrastructure powered by the Galaxy open-source framework, among which Nebula [46] and Cistrome [47] are two web-based services dedicated to ChIP-Seq data analysis. Other open-source computational infrastructures of a similar nature include CyVerse [48], the rebranded iPlant [49] for serving wider community, and Crunch [50], a recent automated ChIP-Seq data analysis pipeline. While the free data analysis pipelines are the mainstream in academia, commercial solutions for cloud computing may appeal some researchers. For example, the newest ENCODE ChIP-Seq data analysis pipeline can be run on DNAnexus (<https://www.dnanexus.com>), a commercial cloud-based computational platform targeting at the genomics research community. These web-based bioinformatics analysis pipelines have empowered the genomics researchers by making sophisticated data analysis accessible to those who have little background in bioinformatics and programming, and have had the

potential of making data analysis reproducible. But the majority of the researchers have not taken full advantage of such features and do not make the exact analysis available on the web-based platform they use or share their data.

R [51] is the most popular data analysis software in statistics, and in many scientific fields. It has also been gaining its popularity in genomics research. R provides an extremely flexible and interactive computational environment that allows the researchers try various tools with different options in a single environment by loading the R packages from CRAN, Bioconductor [52], or the developers' personal page. In a ChIP-Seq analysis pipeline, there could be many R packages for each step [53]. For example, the peak caller MOSAiCS provides a Bioconductor package and SPP provides a package on its developer's web page. There are also multiple R/Bioconductor packages for differential binding analysis, such as DiffBind, DBChIP, csaaw [54], edgeR, or DESeq2. Traditionally, R is known to be memory consuming for large data set. Partially due to this, R is more popular in the downstream analysis of ChIP-Seq after obtaining the mapped reads and even the peak files, and the majority of the R-based ChIP-Seq data analysis pipelines focus on the downstream analysis and visualization, and tend to call external command line tools for the preprocessing and import their results [55,56]. Recently, the decreasing price in computational resources including physical memory has made it possible to conduct ChIP-Seq data analysis from the unmapped reads within R environment only. Park et al. [57] proposed one such pipeline based on Bioconductor packages. For narrow peak calling from one ChIP-Seq sample and one control, the authors reported that it took about 1.2 h (with 7 cores) on read mapping, 1.2 h on QC (with 7 cores), 13 min on peak calling with MOSAiCS, and 2 min on annotation and visualization of the peaks. While it is plausible to perform ChIP-Seq data analysis from raw reads to the downstream analysis, it may still be more efficient to start from the peaking calling in R using the mapped and quality checked reads.

BEYOND THE STANDARD PIPELINE: ALLELIC-IMBALANCE DETECTION FROM ChIP-SEQ

Standard NGS analysis pipelines ignore the differences in DNA sequence among individuals by mapping reads to the reference genome, which is usually sufficient for genome-wide profiling of transcription factor binding or histone modification. A recent trend in epigenetics, especially in the research related to complex disease, is to analyze the ChIP-Seq data in its genetic context and detect the ALlelic-Imbalance (ALI) in ChIP-Seq [58]. ALI detection from ChIP-Seq is a complementary analysis to the standard computational pipeline, and it associates the genotype with the ChIP-Seq signal with internal control (by the other allele). We remark that ALI could happen anywhere in the genome, but it can be observed from ChIP-Seq data only at heterozygous structural variations such as heterozygous SNPs. In the literature, ALI has been shown to be useful in providing the potential molecular mechanistic interpretation of GWAS hits.

ALI detection from ChIP-Seq includes the following three steps: (1) Read Alignment: determining the genomic location and the allele of each read; (2) Counting: determine the count of legitimate mapped reads at each heterozygous SNP from each allele; and (3) Statistical Testing: rank the candidate SNPs and determine which of them have ALI. The first two steps are bioinformatical in nature, while the last step is statistical.

For read mapping and counting, recall that the SV that is different from the reference allele will be considered as read error, which will unfairly reduce the coverage from the nonreference allele and bias the ALI detection. Thus an allelic “unbiased” mapping strategy is critical. The simplistic way of correcting the mapping bias is using an “N”-masked genome, in which the SNP locations in reference genome are substituted by letter “N” [20]. Then the reads from both alleles will have similar chances to be mapped at this SNP location. However, this approach cannot handle short indels. Another strategy is to build the personalized diploid genome using the genotype information first, then map the reads to the two alleles separately, and assign the origin allele based on the location and quality of the two sets of the alignment results [20,23]. When the genotype is accurately phased, this should be the most accurate strategy in theory. But some resent studies suggest that the personalized genome approach cannot resolve the allelic mapping bias completely, as the chances of being uniquely mapped on each allele for a read could be different [22]. One of these papers proposed a two-stage strategy WASP [22] for unbiased allele-specific mapping. Using this method, the reads are mapped to the reference genome first. Then the mapped reads that overlap with SNPs are kept, and their bases at the SNP locations are altered to the other allele combinations different from the reference. Finally, the modified reads are remapped. Only the reads that are mapped to the same location before and after being modified are kept. While WASP seems to reduce the allelic bias in mapping, it may also reduce the coverage, especially in regions with close-by SNPs. Another issue with all the above tools is that the genotype information needs to be provided, which could be unrealistic in many applications. Many researchers have bypassed this by calling heterozygous structural variants directly from ChIP-Seq data [59,60]. It will miss the heterozygous loci at which one allele is completely silenced, but may also reduce the false positives by discarding the homozygous loci that could have been falsely labeled as heterozygous by genotype calling based on WGS.

The statistical setup for testing ALI from ChIP-Seq is straightforward. Many software simply use binomial test (e.g., [21]) in which the total read count that covers an SNP is n , the count from one allele (e.g., the maternal allele) is x , and the goal is to test whether the true proportion from that allele (p) is 0.5 or not. However, the binomial test ignores the natural variation in allelic frequency when there is no ALI, and thus is overpowered when the total count is large. Many alternative statistical methods have been proposed for modeling such overdispersion, and improve ALI detection, such as the beta-binomial test in AlleleDB [61], NPBin [58], a nonparametric empirical Bayes model, and the combined likelihood ratio test in WASP. For the analysis of a single sample, NPBin makes minimal assumption on the distribution of the unobserved true allelic frequency, and has good accuracy and cross-replicate consistency. When there are a large number of samples, WASP has the potential of boosting the detection power and robustness by its joint modeling framework. It is widely known that ALIs detected in the regions with copy number variation (CNV) are more likely to be false positives, as if the two alleles have different copy numbers, the background allelic frequency is altered. The common practice in dealing this issue is to discard the SNPs with CNV. However, in cancer research, CNVs are abundant and potentially informative. Thus it is unwise to ignore these genomic regions. Many tools incorporate the background allele frequency (e.g., from external gDNA data) into the model for correction at each SNP [62–64], but the SNP identification and the allele frequency estimation could be wrong as well. BaalChIP [64] and iASeq [65] detect ALI from multiple data sets jointly, which may improve the power and accuracy.

SUMMARY

In this chapter, we reviewed the common steps, tools, and pipelines for ChIP-Seq analysis, and how the design and the quality of the ChIP-Seq data may influence the choice of tools and the results. Beyond the standard practice of ChIP-Seq data analysis, we also reviewed an emergent type of analysis, the allele-specific enrichment analysis in ChIP-Seq. Another emergent type of genomic data analysis is the integrative analysis. We have covered many tools for the integrative analysis of multiple ChIP-Seq data, such as jMOSAiCS and MultiGPS for joint peak calling, and BaalChIP and iASeq for joint ALI detection.

REFERENCES

- [1] Bailey T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* 2013;9(11):e1003326.
- [2] Pepke S, Wold B, Mortazavi A. Computation for chip-seq and RNA-seq studies. *Nat Methods* 2009;6: S22–32.
- [3] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;2010(2). pdb.prot5384.
- [4] Buenrostro JD, et al. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;109. 21.29.1-9.
- [5] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008;26(12):1351–9.
- [6] Jung YL, et al. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res* 2014;42(9):e74.
- [7] Sims D, et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014; 15(2):121–32.
- [8] Zhang Q, et al. Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC Bioinformatics* 2016;17(1):96.
- [9] Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;22(9):1813–31.
- [10] Andrews S. FASTQC: a quality control tool for high throughput sequence data. 2010. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [11] Consortium E. Encode quality Metrics. 2012. Available from: <https://genome.ucsc.edu/ENCODE/qualityMetrics.html#definitions>.
- [12] Langmead B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10(3):R25.
- [13] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;25(14):1754–60.
- [14] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.
- [15] Marco-Sola S, et al. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 2012;9(12):1185–8.
- [16] Chung D, et al. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput Biol* 2011;7(7):e1002111.
- [17] Zeng X, et al. Perm-seq: mapping protein-DNA interactions in segmental duplication and highly repetitive regions of genomes with prior-enhanced read mapping. *PLoS Comput Biol* 2015;11(10):e1004491.

- [18] Zhang Q, Keleş S. CNV-guided multi-read allocation for ChIP-seq. *Bioinformatics* 2014;30(20):2860–7.
- [19] Zhang Z, Xing Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res* 2017;45(16):9260–71.
- [20] Krueger F, Andrews SR. SNPsSplit: allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res* 2016;5.
- [21] Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011;7(1):522.
- [22] Van De Geijn B, et al. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* 2015;12(11):1061–3.
- [23] Yuan S, et al. One size doesn't fit all-RefEditor: building personalized diploid reference genome to improve read mapping and genotype calling in Next generation sequencing studies. *PLoS Comput Biol* 2015;11(8):e1004448.
- [24] Zhang Y, et al. Model-based analysis of chip-seq (MACS). *Genome Biol* 2008;9(9):R137.
- [25] Kuan PF, et al. A statistical framework for the analysis of ChIP-Seq data. *J Am Stat Assoc* 2011;106(495):891–903.
- [26] Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* 2012;8(8):e1002638.
- [27] Koohy H, et al. A comparison of peak callers used for DNase-Seq data. *PLoS One* 2014;9(5):e96303.
- [28] Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One* 2010;5(7):e11471.
- [29] Laajala TD, et al. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* 2009;10(1):618.
- [30] Chung D, Zhang Q, Keleş S. MOSAiCS-HMM: a model-based approach for detecting regions of histone modifications from ChIP-seq data. In: Statistical analysis of Next generation sequencing data. Springer; 2014. p. 277–95.
- [31] Teng M, Irizarry RA. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. *Genome Res* 2017;27(11):1930–8.
- [32] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
- [33] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139–40.
- [34] Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. R Package Version 2011;100.
- [35] Chen L, et al. ChIPComp: a novel statistical method for quantitative comparison of multiple ChIP-seq datasets. 2016.
- [36] Liang K, Keleş S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* 2011;28(1):121–2.
- [37] Zhang Y, et al. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics* 2014;30(18):2568–75.
- [38] Zeng X, et al. jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol* 2013;14(4):R38.
- [39] Mahony S, et al. An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput Biol* 2014;10(3):e1003501.
- [40] Yan H, et al. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. *BMC Bioinformatics* 2014;15(1):280.
- [41] Zang C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 2009;25(15):1952–8.
- [42] Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011;27(12):1696–7.

- [43] Wei LK, Au A. Computational epigenetics. In: *Handbook of epigenetics*. 2nd ed. 2017. p. 167–90.
- [44] Ji H, et al. Using CisGenome to analyze ChIP-chip and ChIP-seq data. *Curr Protoc Bioinformatics* 2011. Chapter 2:Unit2.13.
- [45] Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;11(8):R86.
- [46] Boeva V, et al. Nebula—a web-server for advanced ChIP-seq data analysis. *Bioinformatics* 2012;28(19):2517–9.
- [47] Liu T, et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* 2011;12(8):R83.
- [48] Merchant N, et al. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol* 2016;14(1):e1002342.
- [49] Goff SA, et al. The iPlant collaborative: cyberinfrastructure for plant biology. *Front Plant Sci* 2011;2.
- [50] Berger S, et al. Crunch: completely automated analysis of chip-seq data. *bioRxiv* 2016:042903.
- [51] Team RC. R language definition. Vienna, Austria: R foundation for statistical computing; 2000.
- [52] Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5(10):R80.
- [53] de Santiago I, Carroll T. Analysis of chip-seq data in R/Bioconductor. In: *Chromatin immunoprecipitation*. Springer; 2018. p. 195–226.
- [54] Lun AT, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res* 2015;44(5):e45.
- [55] Cormier N, Kolisnik T, Bieda M. Reusable, extensible, and modifiable R scripts and Kepler workflows for comprehensive single set ChIP-seq analysis. *BMC Bioinformatics* 2016;17(1):270.
- [56] Deepayan Sarkar RG, Michael L, Zizhen Y. Chipseq: Chipseq: A Package for Analyzing Chipseq Data. 2017. [p. R package].
- [57] Park S-J, et al. A chip-seq data analysis pipeline based on Bioconductor packages. *Genomics Inform* 2017;15(1):11–8.
- [58] Zhang Q, Keleş S. An empirical Bayes test for allelic-imbalance detection in ChIP-seq. *Biostatistics* 2017.
- [59] Ni Y, et al. Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. *BMC Genet* 2012;13(1):46.
- [60] Maurano MT, et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* 2015;47(12):1393–401.
- [61] Chen J, et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* 2016;7.
- [62] Bailey SD, et al. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics* 2015;31(18):3057–9.
- [63] Younesy H, et al. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* 2013;30(8):1172–4.
- [64] de Santiago I, et al. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome Biol* 2017;18(1):39.
- [65] Wei Y, et al. iASEq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics* 2012;13(1):681.

This page intentionally left blank

COMPUTATIONAL TOOLS FOR MICRORNA TARGET PREDICTION

6

Nurul-Syakima Ab Mutalib, Siti Aishah Sulaiman, Rahman Jamal

UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

INTRODUCTION

Since their discovery in *Caenorhabditis elegans* over 2 decades ago [1], microRNAs (miRNAs) have been one of the most exhilarating biological discoveries in RNA regulation. miRNAs are short (~19–25 nucleotides in length) single-stranded RNAs which do not code for any protein and control the expression of protein-coding genes posttranscriptionally [2]. miRNAs are ubiquitous; one-third of mammalian miRNAs are embedded in the introns of protein-coding genes [3] while the rest can be found in the 3' untranslated region (3' UTR) of protein-coding genes [4] as well as exons and introns of noncoding genes [5]. miRNA biogenesis has been comprehensively reviewed [6,7], whereby the majority of mammalian miRNAs are mainly transcribed by RNA polymerase II [8]. Based on computational predictions, 60%–90% of human protein-coding genes are targeted by miRNAs through conserved sequence matching between the 5' region of miRNA, dubbed as the “seed region,” and the 3' UTR of mRNA [9,10].

The development of next-generation sequencing (NGS) platforms has greatly expanded the discovery of miRNAs which is evident in the sharp increase of miRBase entries [11,12]. The latest release (v21) of the database contains 28,645 entries representing hairpin precursor miRNAs, expressing 35,828 mature miRNA products in 223 species. In humans, 1881 precursors and 2588 mature miRNAs have been identified, with 296 precursors designated as high confidence [12]. As the number of miRNAs increases, the quest to identify their validated targets is also increasing. Precise identification of miRNA targets is essential in order to correctly infer their roles.

Prior to the amplification of various miRNA target prediction tools, miRNA targets were examined manually and subsequently confirmed by labor intensive, time-consuming techniques including luciferase reporter assay, gene expression analysis, RNA ligase mediated-5' rapid identification of cDNA ends (5' RLM-RACE), and immunoprecipitation of the RNA-induced silencing complex (RISC) components [13]. A review by Thomson et al. [13] summarizes the strengths and caveats of the experimental methods used for miRNA target identification. The idea to develop *in silico* target prediction algorithms came from the observation that miRNAs generally have a targeting pattern which led to the discovery of the first targets for the let-7 and lin-4 miRNAs [14].

The development of computational miRNA prediction tools has indeed revolutionized miRNA research; however, these tools should be used with caution due to the high false-positive rate.

Nevertheless, computational miRNA prediction tools can help to reduce the requirement for experimental validation. A single miRNA can regulate multiple targets; likewise, the same target could also be regulated by many miRNAs [15]. Large numbers of validated miRNA targets have been experimentally identified; yet the information is rather scattered. DIANA-TarBase v7.0, a database that has cataloged the published and experimentally validated miRNA:target interactions, contains more than 500,000 validated interactions via manual curation [16]. As of October 2017, Tarbase v8.0 is already accessible [17] and houses 1,077,279 experimentally validated interactions curated from 1165 publications, which is equivalent to 9- to 250-fold more entries than any other related databases.

Due to their widespread targets, the involvement of miRNAs is anticipated in many cellular and biological processes. Disturbance in miRNA biogenesis or dysregulated expression could lead to human diseases. Indeed, miRNAs have been proven to be involved in nearly all types of cellular pathways, from normal development [18,19], immunity [20,21], diabetes [22], diabetic nephropathy [23], atherosclerosis [24], cardiovascular disease [25], and carcinogenesis [26–31]. Given their participation in gene regulation as well as disease processes, their clinical utility is now being actively investigated. The potential of miRNAs in therapeutic applications is also being explored [7,32]. Several miRNAs have entered clinical trials as therapeutic agents, for instance miR-16 mimics for the treatment of mesothelioma and non-small cell lung cancer [33], miR-29 mimics for scleroderma treatment [34], anti-miR-155 for cutaneous T cell lymphoma and mycosis fungoides [35], and anti-miR-103/107 for patients with type 2 diabetes and nonalcoholic fatty liver diseases [36,37]. The current applications of miRNA as therapeutic tools in the clinic have been already comprehensively reviewed elsewhere [7,38] and are beyond the scope of this chapter. Collectively, these tiny molecules can be considered as the master regulator of human health and diseases.

PRINCIPLES OF MICRORNA TARGET PREDICTION

Plant and animal miRNAs have distinct targeting patterns. In plant, each miRNA and its target mRNA exhibit an almost-perfect sequence complementarity, which subsequently leads to the cleavage of the double-stranded RNA (dsRNA) [39]. This almost-perfect complementarity makes miRNA target prediction much simpler in plants, and therefore the issues of lack of accuracy and false positives caused by computational prediction methods are greatly reduced [40,41]. In contrast, animal miRNAs, particularly human, show partial sequence complementarity with their targets. The seed region within the structure of the miRNAs which is approximately six to eight nucleotides in length plays a crucial role in target regulation [42]. As mentioned previously, one of the key mechanisms of miRNA action is the interaction of their seed sequences at the 5' end of miRNA with 3' UTR of mRNA transcribed from the target genes, typically resulting in repression of the target mRNA in question [43]. However, there is growing evidence that showed human miRNA binding with the 5' UTR [44] as well as coding sequence (CDS) of the mRNA [45,46], further adding to the complexity of human miRNA target prediction.

miRNA regulation is influenced by various factors and as our understanding of miRNAs evolves, scientists are presented with a new observation; the proposed targets can also mutually control the level and function of miRNAs [47]. Various approaches are available for identification of miRNA target sites, but the presence of a miRNA-binding site alone is inadequate to accurately predict target regulation. Most of the available prediction tools require Watson–Crick pairing with the targeted site; yet any algorithms solely dependent on simple base-pairing rules will produce high false-positive rates [42].

There are many principles being used for computational miRNA target prediction. The algorithms combine several features to increase prediction efficiency, including (1) seed sequence complementarity [15,48]; (2) evolutionary conservation status [48]; (3) free energy [49]; (4) target-site accessibility [50,51]; (5) target-site abundance [52]; (6) pattern-based approach [9]; (7) local AU flanking content [53]; and (8) G–U wobble [54] (Fig. 6.1). However, these principles are confined to human interpretation of interaction between the miRNAs and their targets [55]. Therefore, it is believed that machine learning algorithm which utilized neither seed information nor conservation status has the potential to further increase prediction accuracy. Approximately 6% of human miRNA prediction tools incorporates machine learning algorithm [56] either independently or in combination with other principles. It is important to understand the basis of target prediction algorithms before deciding which tool(s) to use. The description of each of the principles will be outlined below (Fig. 6.1).

SEED SEQUENCE COMPLEMENTARITY

Seed sequence complementarity is defined as a perfect match of miRNA seed region to the target mRNA without any gap [15,48]. The majority of prediction tools scan the 3' UTR of the target genes when searching for complementarity, but other tools have offered target-site searching in the 5' UTRs as well as protein-coding regions. However, 3' UTRs are poorly characterized portions of the human genome [57], therefore increasing the complexity of target prediction. Estimation of poorly defined 3' UTR can be performed by taking the downstream flanking sequence from the stop codon with an average corresponding to the 3' UTR length [58]. This approach may partly resolve the problem of uncharacterized 3' UTRs, however, the accuracy of the prediction will be affected as well as introducing false-negative results.

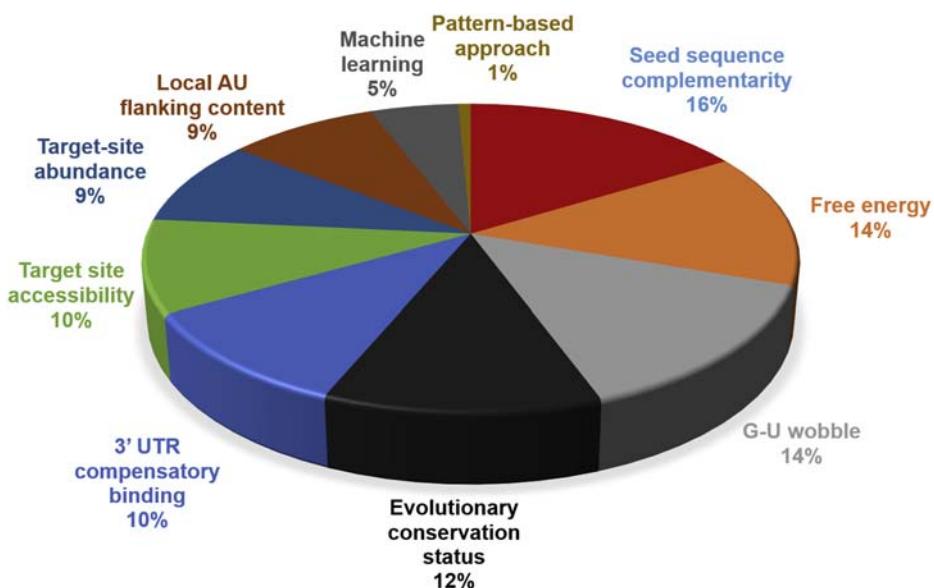


FIGURE 6.1

The principles of miRNA target prediction and the distribution of its frequency in existing tools.

FREE ENERGY

The thermodynamic properties of miRNA–mRNA binding can be measured by calculating the free energy (ΔG) of the putative binding. Through measurement of minimum free energy (MFE), it is possible to assess how robust is the binding between the miRNA and its target mRNA. The energy is represented as negative real value in kcal/mol unit [59]. A good miRNA:mRNA binding is reflected by a small free energy [2,49], suggesting that the predicted binding is stable and true hybridization is likely to occur. The threshold for acceptable MFE value varies according to researchers' preferences. Liu et al. [60] used a threshold of -15 kcal/mol while other groups set the threshold at maximum of -8.5 kcal/mol [61] or -17 kcal/mol [62]. The approximate free energy of the miRNA–mRNA duplex can be estimated by various RNA folding programmes [63–65], and Vienna package is the most widely used in most miRNA target prediction tools [65]. A step-by-step protocol on how to utilize Vienna webserver (RNAfold) to perform MFE prediction, RNA–RNA duplex prediction, or noncoding RNA detection is comprehensively described by Gruber et al. [66].

G–U WOBBLE

G–U wobble or sequence-based imperfections refer to the condition when G nucleotide is allowed to pair with a U nucleotide instead of a C nucleotide [54]. The presence of a G–U wobble in the miRNA seed region could potentially interfere with the miRNA suppression ability [42,54]. The consequence of G–U mismatch on miRNA:target binding is still controversial and there have been conflicting studies. Some have suggested that the presence of a G–U wobble would eliminate the miRNA targeting capacity, while others managed to identify functional target sites containing G–U mismatch in the miRNA seed [67,68]. The contradictory findings were then later cleared when a study concluded that imperfect seed pairing with a G–U wobble to the target may still be functionally relevant [69].

EVOLUTIONARY CONSERVATION STATUS

A binding site is considered “conserved” when it is preserved across species. A “conserved” sequence can exist in any region within the miRNA sequence, with the seed region being the most conserved area followed by the 3' end of the miRNA [59]. A higher conservation score is alleged to reflect a more dependable prediction [10], however, this assumption is also debatable. Even though this approach helps to reduce false-positive results, it may also result in losing targets which are less conserved or increasing the false positives by including conserved regions which are not the actual miRNA targets. Moreover, a study showed that at least 30% of the experimentally validated target sites are nonconserved, suggesting that the conservation of the miRNA target site alone is inadequate [70]. On the other hand, crosslinking immunoprecipitation (CLIP) data revealed that around 30% of the conserved miRNA targets in a knockdown experiment did not lead to increase expression of the targets [71], signifying that the conserved status does not guarantee a target site to be functional.

3' UTR COMPENSATORY BINDING

There are three types of miRNA interaction with target mRNA which includes 5'-dominant canonical, 5'-dominant seed-only, and 3'-compensatory [53]. These three interactions have distinct features: The canonical sites have perfect complementarity at both the 5' and 3' of the miRNA with a specific

loop in the middle, whereas the dominant seed sites have perfect seed match only to 5' UTR and poor 3' complementarity. On the other hand, compensatory sites have G:U-wobbles in the 5' region but compensate through perfect complementarity at the 3' end [53], and hence are denoted as 3' UTR compensatory binding sites. The miRNAs are shown to have distinct inclination; some miRNAs prefer to bind to the canonical binding sites while others could have the tendency to bind to the 3' UTR compensatory sites [72]. Binding to this 3' UTR compensatory site can complement seed pairing to increase target recognition and compensate for seed mismatch [42]. Therefore, a target prediction tool that incorporates 3' UTR compensatory binding principle is thought to be more sensitive.

TARGET-SITE ACCESSIBILITY

Although the total MFE is considered as a good criterion for selecting candidate binding sites, it is not the most efficient [51]. miRNAs perform their regulatory function by binding to the mRNA targets, therefore the target sites need to be accessible for the binding to occur. Target-site accessibility is a measure of the ease of miRNA binding to target mRNA and is influenced by the secondary structure of mRNA [50,51]. The target mRNA is assumed to form a secondary structure after the transcription process and it is plausible that this structure influences target recognition [73], which can interrupt the miRNA's ability to bind to a target site [50]. To facilitate the binding, at least the seed region of the miRNA and the target site within the mRNA should be accessible [51] and this feature is as important as the seed matching in assessing the efficiency of miRNA-mediated regulation of targets.

TARGET-SITE ABUNDANCE

Target-site abundance reflects the quantity of binding sites of miRNA in a given mRNA target [52]. Certain algorithms consider the number of target sites for prediction [74] based on the hypothesis that multiple target sites in the same 3' UTR can hypothetically improve translational inhibition [75]. However, more recent studies suggested that the binding sites will be more active if the miRNA has lower target-site abundance [52] as it will dilute miRNA activity [76]. Garcia and colleagues [52] showed that miR-23 in HeLa cells exhibits low target repression proficiency via reporter assay as well as array data and this suggests that it is contributed by a weak predicted seed-pairing stability (SPS) and high target-site abundance [52]. Incorporation of target-site abundance information can improve performance of target prediction tools.

LOCAL AU FLANKING CONTENT

Local AU content refers to the number of A and U nucleotides flanking the corresponding seed region of the miRNA [10,53]. Through microarray analysis, functional miRNA target binding sites are found to be favorably located within a locally AU enriched context [76]. This could be explained by the weaker mRNA secondary structure in the surrounding area, which increases the accessibility to the seed region [76], facilitates the miRNA:mRNA binding, and subsequently enhances translational suppression. In addition, via site-depletion analysis, local AU content influences not only destabilization of mRNA but also protein expression [76]. Abundant AU content within 3' UTRs also correlates with a greater density of conserved miRNA complementary sites [77]. The presence of local AU flanking also provides the explanation on why other prediction tools solely based on 3' matching were less efficient in identifying effective targets [76].

MACHINE LEARNING

Majority of miRNA target prediction tools depend on the conserved seed match and it has been proven that this requisite might be too rigid, subsequently leading to substantial number of false negatives [78]. Usually the machine learning-based target prediction tools were trained on multiple species such as mouse, human, and fruitfly, thus making it applicable to a wide range of species. Machine learning method is also developed without the evolutionary conservation requirement; therefore, it is optimal for analyzing unconserved genomic sequences [78].

PATTERN-BASED APPROACH

This approach is based on patterns of statistically significant miRNA motifs created after a sequence analysis of known mature miRNAs performed using Teiresias algorithm [79]. Firstly, it examines the putative reverse complement sites within the mRNA of interest and followed by identification of targeting miRNAs which are likely to bind to these sites [9]. Since this approach does not depend on the cross-species conservation status, it permits the discovery of miRNA binding sites that are not conserved and identification of sites which will be targeted by novel miRNAs [9].

MICRORNA TARGET PREDICTION TOOLS

Experimental miRNA target validations are arduous and time-consuming. Hundreds of genes could be targeted by a single miRNA and the validation of hundreds of targets is definitely not cost-effective. Prediction of miRNA targets via computational approaches is providing researchers with a basic target selection and will in turn reduce the cost. Majority of the tools enable a user-selected threshold for filtering of the false positives, an added value to increase the accuracy and sensitivity of the prediction. Most of the tools combine multiple principles of target prediction in the quest to increase the accuracy. In recent years, the number of combinatorial miRNA prediction tools which is thought to further reduce the false positives is also increasing. To date, there are at least 60 miRNA target prediction tools and this number could be quite intimidating as not all of the tools offer the same features nor use similar principles of prediction. Tools4miR [56], a platform which collects more than 170 various miRNA analysis tools, offers many advantages to researchers. It enables the users to filter the available tools based to their specific research preferences. The selected computational tools for miRNA target prediction are described below.

miRanda is developed based on a comparison of miRNA complementarity to 3' UTR as well as coding sequence (CDS) regions and performs scoring for the likelihood of mRNA downregulation using mirSVR [80]. miRanda recognizes the target using seed sequence complementarity, evolutionary conservation, free energy, and G–U wobble permitted in the seed region [80]. Despite being cited over 3000 times, this tool has not been updated since August 2010 and the mature miRNA sequences used were from miRBase v15.0 (April 2010 release). TargetScan is a web-based server that predicts miRNA targets by screening for conserved and nonconserved sites [81]. It identifies mRNA targets in the 3' UTR and CDS based on seed sequence complementarity, evolutionary conservation, free energy, target-site accessibility, target-site abundance, 3' compensatory pairing, G–U wobble, and local AU content [81]. In addition, the latest 7.1 version also uses updated miRNA families [82,83]. TargetScan was last updated in June 2016, making it as one of the most regularly updated tools.

Target-site accessibility is the main feature of miRNA target prediction by PITA [73]. It is a parameter-free model that measures the changes between the free energy gained from the formation of miRNA:target complexes and the energy cost in unpairing the target in order to make the binding site accessible [73]. In addition to target-site accessibility, PITA also relies on seed sequence complementarity, target-site abundance, and G–U wobble [73]. However, prediction using PITA is restricted to 3' UTR only. On the contrary, rna22 is a pattern-based approach to discover miRNA binding sites and parallel miRNA:target complexes regardless of a cross-species sequence conservation status in any region of the genome, not limited only to 3' UTR [9]. rna22 also enables the identification of potential miRNA binding sites even when the targeting miRNA is yet to be discovered [9].

miRDB [84] is an online database for predicted miRNA targets and functional annotations. The web server interface also permits submission of user's sequences for miRNA target prediction [84], offering the flexibility to study any customized miRNAs or target mRNAs of interest. The targets are predicted by MirTarget [85], which was built by analyzing high-throughput profiling data obtained from CLIP-RNA ligation sequencing studies. An added feature of miRDB is the functional miRNA annotation, whereby the developers have put in a great effort to combine computational analyses with literature curation that led to the discovery of 568 functional miRNAs in humans [84]. This algorithm relies on seed sequence complementarity, machine learning, evolutionary conservation, free energy, target-site accessibility, 3' compensatory pairing, and local AU content [84]. The latest v5.0 of miRDB was updated in August 2015 and utilizes miRNA from the most recent miRBase v21.

DIANA-microT-CDS is the fifth version of the microT algorithm that integrates a machine learning approach to identify binding sites in both CDS and 3' UTRs [86,87]. This web-based target prediction algorithm features seed sequence complementarity, machine learning, evolutionary conservation, free energy, target-site accessibility, target-site abundance, 3' compensatory pairing, G–U wobble, and local AU content [86]. For the advanced users, DIANA-microT-CDS also offers an in-house Taverna plug-in for a more extensive miRNA analysis from high-throughput experiments such as microarrays or NGS [86]. However, DIANA-microT-CDS was last updated in July 2012 and incorporates miRBase v18.

STarMir webserver [88,89] uses a logistic prediction model approach built upon miRNA binding information from CLIP studies [60]. This tool permits the users to submit their miRNA and target mRNA sequences for prediction of binding sites [89]. To predict miRNA binding sites, STarMir takes into account seed sequence complementarity, free energy, target-site accessibility, 3' compensatory pairing, G–U wobble, and local AU content as a measure of confidence for each of seed or seedless sites in all three regions of an mRNA (3' UTR, CDS, and 5' UTR) [89]. This tool was last updated in July 2014; however, the miRNA sequences are obtained from an internal database developed using miRBase v20 [88]. Targets predicted by STarMir are assembled into STarMirDB [90], a recent database application module based on Sfold RNA package [91,92].

Most computational algorithms predict targets for a single miRNA, however, another tool, probabilistic identification of combinations of target sites (PicTar) [93], identifies the targets for both single and combinations of miRNAs. To improve prediction specificity and accuracy, PicTar also depends on tissue-specific target prediction scores from four mammalian tissues; however, it only scans the 3' UTR for target searching [93]. It relies on seed sequence complementarity, evolutionary conservation, free energy, and target-site abundance [93]. PicTar was last updated in March 2007 and prediction for targets in human is based on hg19. RNAhybrid predicts multiple putative miRNA binding sites in large target RNAs by searching for the most energetically favorable hybridization

sites [93,94]. It was last updated in August 2013 and predicts targets based on seed sequence complementarity, free energy, 3' compensatory pairing, G–U wobble, and target-site abundance.

High-throughput RNA-seq experiments have significantly transformed the miRNomics field by enabling the generation of an unparalleled amount of data. Apart from the well-established tools, newer analysis suites which provide complete analysis workflow to accommodate new sequencing technologies are emerging. For instance, miARma-Seq (miRNA-Seq And RNA-Seq Multiprocess Analysis) was quite recently developed to perform differential expression, target prediction, and functional analysis for not only miRNAs but also mRNAs and circular RNAs [95]. This analysis suite is robust as it can be used for any sequenced organism [86]. miRNA target prediction is achieved through miRGate [96], a database containing novel predicted miRNA:target mRNA pairs that are calculated using well-established algorithms such as miRanda [80], TargetScan [81], PITA [73], RNAhybrid [94], and MicroTar [97]. Moreover, miRGate also integrates experimentally validated miRNA:targets from miRecords [98], Tarbase v6 [99], OncomirDB [100], and miRTarBase v4.5 [101].

miRNA sequencing has also revealed the presence of miRNA isoforms (isomiRs), but nearly all of the target prediction tools only provide prediction for the canonical miRNAs. Therefore, a tool that can detect, annotate, and predict isomiR target is in great demand. Hence, DeAnnIso (Detection and Annotation of IsomiRs from sRNA sequencing data) was developed in 2016 [102]. DeAnnIso is able to detect isomiRs from the FASTQ files, perform differential expression, and predict isomiR targets using miRanda [62] or RNAhybrid [94].

Earlier in this chapter, we described eight fundamental principles used for miRNA target prediction. Recently, seven additional new features of miRNA target sites have been identified via combination of four distinct machine learning approaches to the crosslinking ligation and sequencing of hybrids (CLASH) data [103]. By merging these new features with the previous ones, TarPmiR was recently developed [103] and is shown to be able to predict more than 74% bona fide miRNA target sites in the training data set.

Collectively, a selective list of the computational miRNA target prediction tools with the type of input and output files is provided in Table 6.1. Fig. 6.2 illustrates a simplistic workflow to guide the new researchers on the selection of tools while Table 6.2 summarizes the principles of the selected tools.

CONCLUSION AND FUTURE DIRECTION

Computational prediction tools are not without limitations. The role of miRNAs in normal and pathological conditions has been overestimated due to issues related to the false positives which have been overlooked [130]. There is a need to utilize the power of integrated analysis, where miRNAs and mRNAs from the same source are profiled simultaneously, followed by correlation analysis. For example, MAGIA2 [117] enables the users to upload their respective paired or unpaired miRNA and mRNA expression profiles, performs target prediction based on multiple algorithms, and overlaps the predicted target with the user-supplied mRNA data. Correlation will be analyzed and provided in positive or negative value, assisting the user in generating more reliable results. In addition, integrated platforms that combine multiple computational tools could possibly produce more accurate results than a single algorithm, provided that the principles behind them are different. Combinatorial approaches such as MMIA [131] and miRwalk [132], which utilize multiple miRNA target prediction algorithms, are anticipated to improve prediction sensitivity, coverage, and accuracy in the future.

Table 6.1 Computational Tools for Human miRNA Target Prediction

Tool	Description	Input	Output ^a	Link/Homepage/ Source References
BCmirO	A supervised machine learning tool that combines the prediction of TargetScan, miRanda, PicTar, mirTarget, PITA, and Diana microT	miRNA selection from drop-down list	A list of predicted targets with BCmirO	http://compgenomics.utsa.edu/gene/gene_1.php [104]
BioVLAB-MMIA-NGS	A tool for miRNA: mRNA integrated analysis from microarray and NGS data. It incorporates target prediction tools which are based on TargetScan, PITA, miRSVR, and PMTED [105]	Linear value of expression data (microarray) or FASTQ files (RNA-seq)	A miRNA:mRNA integrated analysis result with negative correlation values	http://epigenomics.snu.ac.kr/biovlab_mmia_ngs/ [106]
ComiR	A supervised machine learning tool to predict whether a given mRNA is targeted by a set of miRNAs. It combines four prediction algorithms (miRanda, PITA, TargetScan, and mirSVR)	Expression value of miRNA(s)	A ranked list of predicted genes based on the target probability computed through the support vector machine (SVM) model	http://www.benoslab.pitt.edu/comir/ [107]
CPSS 2.0 (Computational Platform for analysis of Small RNA deep Sequencing data)	An online tool for complete annotation and functional analysis. Eight miRNA prediction tools which are miRanda, MicroCosm [108], microT_v3.0, MirTarget2, miRNAMap, TargetScan, TargetSpy, and RNAhybrid	FASTA format or FASTA files compressed in *.gz format	A ranked list of predicted genes based on the total score and total energy	[109]

Continued

Table 6.1 Computational Tools for Human miRNA Target Prediction—cont'd

Tool	Description	Input	Output ^a	Link/Homepage/ Source References
CSmiRTar	Collects computationally predicted targets of 2588 human miRNAs from four existing databases (miRDB, TargetScan, microRNA.org , and DIANA-microT)	miRNA ID(s) or gene symbol(s)	List of targets with normalized score	http://cosbi.ee.ncku.edu.tw/CSmiRTar/ [110]
DeAnnIso	A tool for detection and target prediction of isomiRs	FASTA files	List of predicted targets with fold enrichment and <i>P</i> value	http://mcg.ustc.edu.cn/bsc/deanniso/ [102]
DIANA-microT-CDS	A supervised machine learning tool which performs miRNA target prediction using different thresholds and metaanalysis statistics, followed by pathway enrichment analysis. It is specifically trained on a positive and a negative set of miRNA recognition elements (MREs) located in both the 3' UTR and CDS region	miRNA ID(s)	List of predicted targets with miRNA target gene (miTG) prediction score	http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index [86]
EIMMo3	A target prediction tool based on Bayesian target prediction algorithm that automatically infers the phylogenetic distribution of functional sites for each miRNA, and assigns a	miRNA ID	List of predicted targets with probability value of conserved site, expected number of sites, and number of binding sites	http://www.clipz.unibas.ch/EIMMo3/index.php [111]

	posterior probability to each putative target site. This tool also provides insights into the evolution of target sites for individual miRNAs			
GUUGle	An algorithm to search for RNA–RNA interactions under RNA base-pairing rules (Watson–Crick and G–U pairs)	FASTA file of a set of RNA sequences (miRNA and target mRNA)	A single or a set of mRNAs with complementary binding site	https://bibiserv.cebitec.uni [112]
homoTarget	A prediction tool that combines pattern recognition neural network (PRNN) and principle component analysis (PCA) while incorporating 12 structural, thermodynamic, and positional features of miRNA:mRNA binding sites	miRNA(s) and mRNA(s) sequence in FASTA format	A single or a set of mRNAs with complementary binding site	http://lbb.ut.ac.ir/dynamic/uploads/soft/homoTarget.rar [113]
iMir	A tool for comprehensive analysis of small RNA-Seq data, which include target prediction based on miRanda and TargetScan	FASTQ files	A list of predicted targets with mirSVR score and aggregate preferentially conserved targeting (PCT) value	http://www.labmedmolge.unisa.it/italiano/home/imir [114]
IntaRNA	A program for prediction of interactions between mRNA target sites for given miRNA	FASTA file of a set of RNA sequences (miRNA and target mRNA)	Summary of the best 100 predicted interactions	http://rna.informatik.uni-freiburg.de/IntaRNA/Input.jsp [115]

Continued

Table 6.1 Computational Tools for Human miRNA Target Prediction—cont'd

Tool	Description	Input	Output ^a	Link/Homepage/ Source References
MAGI	A web service for miRNA-Seq data analysis which incorporates miRanda for miRNA target identification	FASTQ files	A list of predicted targets with mirSVR score	http://magi.ucsd.edu [116]
MAGIA2	A supervised machine learning tool for target prediction through integrated miRNA–target expression. Eight different algorithms were incorporated (Microcosm, miRanda, DIANA-microT, miRDB, PicTar, PITA, RNA22, and TargetScan)	Expression value of miRNA(s) and mRNA(s) in tab-delimited matrices (samples on the columns and mRNA or miRNAs on the rows)	miRNAs, mRNA, and transcription factors functional interactions with correlation score and q values	http://gencomp.bio.unipd.it/magia2/start/?sid=f566bc45e0b03d7fb0bc16bebfc4ac5 [117]
miARma-Seq (miRNA-Seq And RNA-Seq Multiprocess Analysis)	A tool for identification of mRNAs, miRNAs and circular RNAs that includes miRGate for target prediction	FASTQ files	Tabulated file (excel compatible) with the predicted targets and the statistical values of the prediction	http://miarmaseq.idoproteins.com/ [95]
MicroInspector	A web-based tool for identification of miRNA binding sites in a target RNA sequence	GENBANK or TAIR sequence accession number. Alternatively, the user can enter the sequence in the box as provided	List of predicted target(s) with miRNA binding sites in the 3' UTR sequence and free energy	http://bioinfo.uni-plovdiv.bg/microinspector/ [118]
MicroTar	A target prediction tool based on miRNA-target complementarity and RNA duplex energy prediction	FASTA file of a set of RNA sequences (miRNA and target mRNA)	A list of predicted target(s) with the <i>P</i> value derived from negative normalized free energy and extreme value distributions	http://tiger.dbs.nus.edu.sg/microtar/index.html [97]

miRanda	An algorithm to identify potential target sites for miRNAs in the genomic sequence via local alignments of miRNA: UTR and assessment of the thermodynamic folding energy	miRNA ID(s)	List of predicted targets with mirSVR score	http://www.microrna.org/microrna/home.do [80]
miRcode	A web-based tool that predicts miRNA targets based on seed match and conservation score	A single miRNA family and/or gene symbol	A list of predicted target(s) with seed location and conservation scores	http://mircode.org/ [119]
miRDB	A supervised machine learning database for miRNA targets predicted by MirTarget. In addition to precompiled prediction data, it allows submission of user-provided sequences for custom miRNA target prediction	miRNA ID for precompiled prediction or miRNA/target mRNA sequence for custom prediction	A list of predicted target(s) with target rank and score	http://mirdb.org/miRDB/ [84]
mirDIP v4.1	A tool that combines 152 million miRNA–target predictions across 30 resources	miRNA ID(s) or gene symbol	A list of target(s) with integrative score and score class	http://ophid.utoronto.ca/mirDIP/ [120]
miRGate	A database containing the predicted and experimentally validated miRNA–mRNA target pairs. The experimentally validated interactions are compiled from Tarbase, miRTarbase, miRecords, and OncomirDB	miRNA ID(s) and/or gene symbol	A list of predicted target(s) or interaction (if both miRNAs and mRNA are used). Genomic agreement score is provided for the users to rank the prediction	http://mrgate.bioinfo.cnio.es/miRGate/ [96]

Continued

Table 6.1 Computational Tools for Human miRNA Target Prediction—cont'd

Tool	Description	Input	Output ^a	Link/Homepage/ Source References
miRrror Suite	A supervised machine learning platform that combines TargetScan, MicroCosm implemented in miRBase, PicTar, DIANA-MicroT, PITA, EIMMO-MirZ, miRanda-based microRNA.org , TargetRank, miRDB, and TarBase for target prediction	A list of at least two miRNAs or genes	A list of predicted target(s) or targeting miRNA(s) with miRIS internal score	http://www.proto.cs.huji.ac.il/mirror/ [121]
miRTar	An integrated system which provides seven scenarios to identify putative miRNA target sites in the gene	miRNA sequence in FASTA, miRBase accession number, and Ensembl ID/gene symbol	Minimum free energy (MFE) and alignment score of the miRNA:mRNA	http://mirtar.mbc.nctu.edu.tw/human/index.php [122]
miRTarCLIP	A tool for mining of miRNA:target sites from CLIP-Seq and PAR-CLIP sequencing data	Sequence Read Archive (.sra), FASTQ, or FASTA files	A summary of predicted miRNA binding sites with scores from TargetScan, target-site locations, target-gene annotations, and types of seed region	http://mirtarclip.mbc.nctu.edu.tw/ [123]
mirTools 2.0	A supervised machine learning tool for miRNA-seq analysis which includes miRanda and RNAhybrid for target prediction. Additionally, user can also select another six target prediction tools or databases (TargetScan, TargetSpy, miRNAMap, microT v4.0, MicroCosm, and MirTarget2)	FASTA files or mapping results (.bam)	A list of predicted miRNA:mRNA and their functional annotation with GO, the Kyoto encyclopedia of genes and genomes (KEGG) pathway, and the protein–protein interaction network	http://www.wzgenomics.cn/mr2_dev/index.php [124]

MultiMiTar	A supervised machine learning tool for prediction of miRNA:mRNA interaction	A list of miRNA:mRNA pairs (for known miRNAs). For novel miRNA, its sequence can be used as input miRNA selection from drop-down list	A list of miRNA:mRNA pairs with decision value	http://www.isical.ac.in/~bioinfo_miul/multimitar.htm [125]
PicTar (probabilistic identification of combinations of target sites)	A tool for identifying common targets of miRNAs using genome-wide alignment statistics of eight vertebrate genomes		A list of predicted target(s) with PicTar score	http://pictar.mdc-berlin.de/ [93]
PITA	A parameter-free model for miRNA:target interaction that calculates the difference between the free energy gained from the formation of the miRNA:target duplex	miRNA(s) or UTR(s) sequence in FASTA format	A list of miRNA:mRNA pairs with $\Delta\Delta G$ score	https://genie.weizmann.ac.il/pubs/mir07/index.html [73]
RNA22	A pattern-based approach to discover miRNA binding sites and parallel miRNA:target complexes regardless of a cross-species sequence conservation status in any region of the genome	Up to 50 miRNA sequence(s) and a single target sequence in FASTA format	A list of miRNA:mRNA pairs with folding energy and <i>P</i> value	https://cm.jefferson.edu/rna22/ [9]
RNAhybrid	A tool for determining the minimum free energy hybridization of RNA sequence(s) and miRNA(s)	miRNA(s) or mRNA(s) sequence in FASTA format	Minimum free energy (MFE) and <i>P</i> value of the miRNA:mRNA pairs	https://bibiserv.cebitec.uni [94]
STarMir	A web-based tool for prediction of miRNA binding sites on a target RNA	miRNA ID(s) and/or gene symbol. Alternatively, users can upload the sequence in FASTA format	A list of miRNA:mRNA pairs with logistic probability	http://sfold.wadsworth.org/cgi-bin/starmirtest2.pl [88,89]
submiRine	A software for predicting miRNA target-site variants (miR-TSVs) from clinical genomic data sets	miRNA(s) or 3' UTR sequence in FASTA format	A list of miRNA:mRNA pairs with sum of log-scaled probabilities (SLP) score	https://research.nhgri.nih.gov/software/SubmiRine/index.shtml [126]

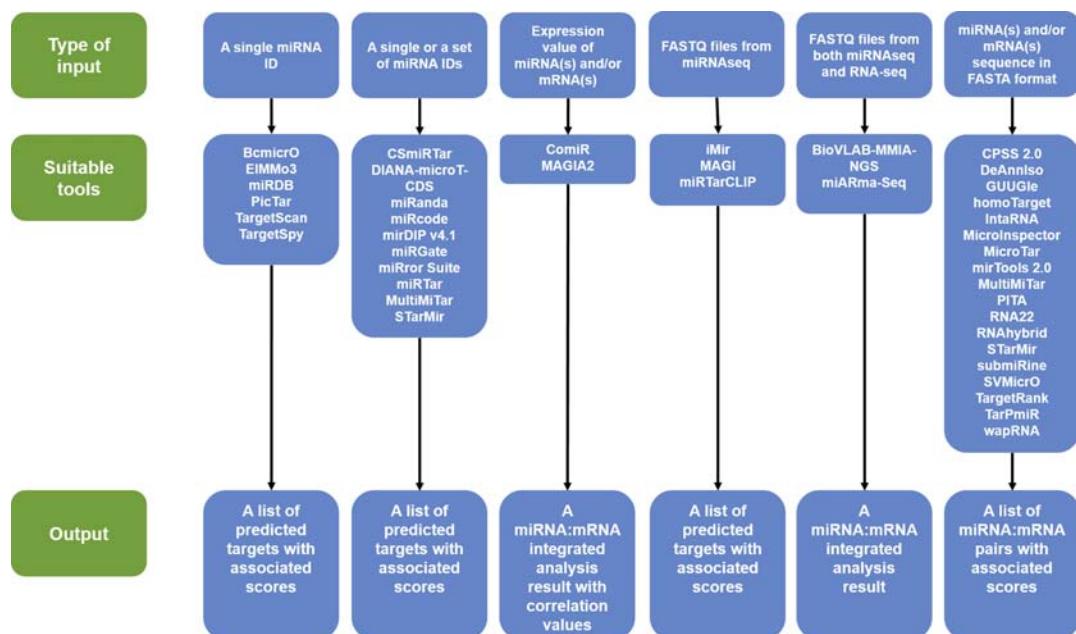
Continued

Table 6.1 Computational Tools for Human miRNA Target Prediction—cont'd

Tool	Description	Input	Output ^a	Link/Homepage/ Source References
SVMicrO	A supervised machine learning miRNA target prediction algorithm which assumes a two-stage structure comprises of a site support vector machine (SVM) and a UTR-SVM	miRNA(s) or 3' UTR sequence in FASTA format	A list of miRNA:mRNA pairs with F score	http://compgenomics.utsa.edu/svmicro.html [127]
TargetRank	A target prediction tool which relies on the scoring of the seed matches in the 3' UTR	miRNA(s) sequence in FASTA format or miRNA ID	A list of predicted targets with TargetRank score	http://hollywood.mit.edu/targetrank/ [128]
TargetScan	A web-based server that predicts miRNA targets by screening for conserved and nonconserved sites	miRNA ID	A list of predicted targets with aggregate preferentially conserved targeting (PCT) value	http://www.targetscan.org/vert_71/ [81]
TargetSpy	A supervised machine learning tool that predicts target sites independently from the seed match. It is suitable for predicting targets in the unconserved genomic sequences	miRNA ID	A list of predicted targets with TargetSpy score and free energy	http://webclu.bio.wzw.tum.de/targetspy/index.php?search=true [78]
TarPmiR	A supervised machine learning based on random-forest approach which combines 13 features for miRNA target-site prediction	miRNA(s) and/or mRNA(s) sequence in FASTA format	A list of miRNA:mRNA pairs with binding site and probability	http://hulab.ucf.edu/research/projects/miRNA/TarPmiR/ [103]
wapRNA	An analysis tool for mRNA-seq and miRNA-seq data which integrates miRanda and RNAHybrid for miRNA target prediction	Sequence file (CSFASTA), or quality file(QUAL) from Solid platform FASTA files for Solexa/Illumina platform	A list of miRNA:mRNA pairs	http://waprna.big.ac.cn/ [129]

Plant and bacterial miRNA target prediction tools are excluded from the list.

^aRefers to output from target prediction analysis only.

**FIGURE 6.2**

The selection of suitable tools based on the input files available.

Table 6.2 Principles of Prediction of Selected Tools

Tool	Principles of Prediction
BcmicrO	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status • 3' UTR compensatory binding
BioVLAB-MMIA-NGS	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status
ComiR	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status • 3' UTR compensatory binding
CPSS 2.0 (Computational Platform for analysis of Small RNA deep Sequencing data)	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning <ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site accessibility • Target-site abundance • Local AU content <ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning <ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning

Continued

Table 6.2 Principles of Prediction of Selected Tools—cont'd

Tool	Principles of Prediction
CSmiRTar	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status • 3' UTR compensatory binding
DeAnnIso	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble
DIANA-microT-CDS	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status • 3' UTR compensatory binding
EIMMo3	<ul style="list-style-type: none"> • Seed sequence complementarity • G—U wobble
GUUGle homoTarget	<ul style="list-style-type: none"> • Seed sequence complementarity • Seed sequence complementarity • Free energy • G—U wobble
iMir	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status
IntaRNA	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status
MAGI	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy
MAGIA2	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status • 3' UTR compensatory binding
miARma-Seq (miRNA-Seq And RNA-Seq Multiprocess Analysis)	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • 3' compensatory pairing
MicroInspector	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy
MicroTar	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy
miRanda	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble
miRcode	<ul style="list-style-type: none"> • Seed sequence complementarity
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Evolutionary conservation status • 3' UTR compensatory binding • Target-site abundance • Target-site accessibility • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Evolutionary conservation status
	<ul style="list-style-type: none"> • G—U wobble • Local AU content • Pattern-based approach
	<ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site abundance • Local AU content
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance
	<ul style="list-style-type: none"> • G—U wobble • Evolutionary conservation status
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Evolutionary conservation status • Target-site accessibility • Target-site abundance • Pattern-based approach
	<ul style="list-style-type: none"> • G—U wobble • Target-site abundance • G—U wobble • Target-site accessibility • Evolutionary conservation status • Pattern-based approach
	<ul style="list-style-type: none"> • Evolutionary conservation status

Table 6.2 Principles of Prediction of Selected Tools—cont'd

Tool	Principles of Prediction
miRDB	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status
mirDIP v4.1	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status • 3' UTR compensatory binding
miRGate	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • 3' compensatory pairing
miRrror Suite	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status • 3' UTR compensatory binding
miRTar	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status
miRTarCLIP	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status
mirTools 2.0	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble • Evolutionary conservation status
MultiMiTar	<ul style="list-style-type: none"> • Seed sequence complementarity • G–U wobble • 3' UTR compensatory binding • Target-site accessibility
PicTar (probabilistic identification of combinations of target sites)	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy
PITA	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble
RNA22	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy
RNAhybrid	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G–U wobble
	<ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site accessibility • Local AU content
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Evolutionary conservation status • Target-site accessibility • Target-site abundance • Pattern-based approach
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site accessibility • Target-site abundance • Local AU content
	<ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site abundance • Local AU content
	<ul style="list-style-type: none"> • 3' UTR compensatory binding • Target-site accessibility • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Target-site abundance • Local AU content • Machine learning
	<ul style="list-style-type: none"> • Evolutionary conservation status • Target-site abundance
	<ul style="list-style-type: none"> • Target-site accessibility • Target-site abundance

Continued

Table 6.2 Principles of Prediction of Selected Tools—cont'd

Tool	Principles of Prediction
STarMir	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble
submiRine	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • Evolutionary conservation status • 3' UTR compensatory binding
SVMicrO	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status • 3' UTR compensatory binding
TargetRank	<ul style="list-style-type: none"> • Seed sequence complementarity • Evolutionary conservation status
TargetScan	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status
TargetSpy	<ul style="list-style-type: none"> • Free energy • 3' UTR compensatory binding
TarPmiR	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • Evolutionary conservation status • 3' UTR compensatory binding
wapRNA	<ul style="list-style-type: none"> • Seed sequence complementarity • Free energy • G—U wobble • Evolutionary conservation status • 3' UTR compensatory binding • Target-site abundance

REFERENCES

- [1] Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993;75(5):843–54.
- [2] Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell* 1993;75(5):855–62.
- [3] Ying SY, Chang CP, Lin SL. Intron-mediated RNA interference, intronic microRNAs, and applications. *Methods Mol Biol* 2010;629:205–37. https://doi.org/10.1007/978-1-60761-657-3_14.
- [4] Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 2004;10(12):1957–66. <https://doi.org/10.1261/rna.7135204>.
- [5] Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. Identification of mammalian microRNA host genes and transcription units. *Genome Res* 2004;14(10A):1902–10.
- [6] Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. *Nat Rev Cancer* 2015;15(6):321–33. <https://doi.org/10.1038/nrc3932>.

- [7] Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov* 2017;16(3):203–22. <https://doi.org/10.1038/nrd.2016.246>.
- [8] Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 2004;23(20):4051–60. <https://doi.org/10.1038/sj.emboj.7600385>.
- [9] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 2006;126(6):1203–17. <https://doi.org/10.1016/j.cell.2006.07.031>.
- [10] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009;19(1):92–105. <https://doi.org/10.1101/gr.082701.108>.
- [11] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 2011;39:D152–7. <https://doi.org/10.1093/nar/gkq1027>.
- [12] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;42:D68–73. <https://doi.org/10.1093/nar/gkt1181>.
- [13] Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Res* 2011;39(16):6845–53. <https://doi.org/10.1093/nar/gkr330>.
- [14] Mazière P, Enright AJ. Prediction of microRNA targets. *Drug Discov Today* 2007;12(11–12):452–8. <https://doi.org/10.1016/j.drudis.2007.04.002>.
- [15] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120:15–20. <https://doi.org/10.1016/j.cell.2004.12.035>.
- [16] Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res* 2015;43(Database issue):D153–9. <https://doi.org/10.1093/nar/gku1215>.
- [17] http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=tarbasev8. [Accessed 23.10.2017].
- [18] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 2007;129(7):1401–14. <https://doi.org/10.1016/j.cell.2007.04.040>.
- [19] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. *Nucleic Acids Res* 2016;44(8):3865–77. <https://doi.org/10.1093/nar/gkw116>.
- [20] Mehta A, Baltimore D. MicroRNAs as regulatory elements in immune system logic. *Nat Rev Immunol* 2016;16(5):279–94. <https://doi.org/10.1038/nri.2016.40>.
- [21] Baumjohann D, Ansel KM. MicroRNA-mediated regulation of T helper cell differentiation and plasticity. *Nat Rev Immunol* 2013;13(9):666–78. <https://doi.org/10.1038/nri3494>.
- [22] Guay C, Regazzi R. Circulating microRNAs as novel biomarkers for diabetes mellitus. *Nat Rev Endocrinol* 2013;9(9):513–21. <https://doi.org/10.1038/nrendo.2013.86>.
- [23] Allison SJ. Diabetic nephropathy: a lncRNA and miRNA megacluster in diabetic nephropathy. *Nat Rev Nephrol* 2016;12(12):713. <https://doi.org/10.1038/nrneph.2016.151>.
- [24] Schober A, Nazari-Jahantigh M, Weber C. MicroRNA-mediated mechanisms of the cellular stress response in atherosclerosis. *Nat Rev Cardiol* 2015;12(6):361–74. <https://doi.org/10.1038/nrccardio.2015.38>.
- [25] McManus DD, Freedman JE. MicroRNAs in platelet function and cardiovascular disease. *Nat Rev Cardiol* 2015;12(12):711–7. <https://doi.org/10.1038/nrccardio.2015.101>.
- [26] Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nat Rev Cancer* 2006;6(11):857–66. <https://doi.org/10.1182/blood-2007-07-098749>.
- [27] Nurul-Syakima AM, Yoke-Kqueen C, Sabariah AR, Shiran MS, Singh A, Learn-Han L. Differential microRNA expression and identification of putative miRNA targets and pathways in head and neck cancers. *Int J Mol Med* 2011;28(3):327–36. <https://doi.org/10.3892/ijmm.2011.714>.

- [28] Iorio MV, Croce CM. Causes and consequences of microRNA dysregulation. *Cancer J* 2012;18(3):215–22. <https://doi.org/10.1097/PPO.0b013e318250c001>.
- [29] Lee M, Kim EJ, Jeon MJ. MicroRNAs 125a and 125b inhibit ovarian cancer cells through post-transcriptional inactivation of EIF4EBP1. *Oncotarget* 2016;7(8):8726–42. <https://doi.org/10.18632/oncotarget.6474>.
- [30] Ibrahim FF, Jamal R, Syafruddin SE, Ab Mutalib NS, Saidin S, MdZin RR, et al. MicroRNA-200c and microRNA-31 regulate proliferation, colony formation, migration and invasion in serous ovarian cancer. *J Ovarian Res* 2015;8:56. <https://doi.org/10.1186/s13048-015-0186-7>.
- [31] Ab Mutalib NS, Othman SN, Mohamad Yusof A, Abdullah Suhaimi SN, Muhammad R, Jamal R. Integrated microRNA, gene expression and transcription factors signature in papillary thyroid cancer with lymph node metastasis. *PeerJ* 2016;4:e2119. <https://doi.org/10.7717/peerj.2119>.
- [32] Matsui M, Corey DR. Non-coding RNAs as drug targets. *Nat Rev Drug Discov* 2017;16(3):167–79. <https://doi.org/10.1038/nrd.2016.117>.
- [33] van Zandwijk N, Pavlakis N, Kao SC, Linton A, Boyer MJ, Clarke S, et al. Safety and activity of microRNA-loaded minicells in patients with recurrent malignant pleural mesothelioma: a first-in-man, phase 1, open-label, dose-escalation study. *Lancet Oncol* 2017;18(10):1386–96. [https://doi.org/10.1016/S1470-2045\(17\)30621-6](https://doi.org/10.1016/S1470-2045(17)30621-6).
- [34] <https://clinicaltrials.gov/ct2/show/NCT02603224> [Accessed 23.10.2017].
- [35] <https://clinicaltrials.gov/ct2/show/NCT02580552> [Accessed 23.10.2017].
- [36] <https://clinicaltrials.gov/ct2/show/NCT02826525> [Accessed 23.10.2017].
- [37] <https://clinicaltrials.gov/ct2/show/NCT02612662> [Accessed 23.10.2017].
- [38] Li Z, Rana TM. Therapeutic targeting of microRNAs: current status and future challenges. *Nat Rev Drug Discov* 2014;13(8):622–38. <https://doi.org/10.1038/nrd4359>.
- [39] Rogers K, Chen X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell* 2013;25(7): 2383–99. <https://doi.org/10.1105/tpc.113.113159>.
- [40] Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 2011; 39(Web Server issue):W155–9. <https://doi.org/10.1093/nar/gkr319>.
- [41] Dai X, Zhuang Z, Zhao PX. Computational analysis of miRNA targets in plants: current status and challenges. *Brief Bioinform* 2011;12(2):115–21. <https://doi.org/10.1093/bib/bbq065>.
- [42] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;136(2):215–33. <https://doi.org/10.1016/j.cell.2009.01.002>.
- [43] Gulyaeva LF, Kushlinskiy NE. Regulatory mechanisms of microRNA expression. *J Transl Med* 2016;14: 143. <https://doi.org/10.1186/s12967-016-0893-x>.
- [44] Li G, Wu X, Qian W, Cai H, Sun X, Zhang W, et al. CCAR1 5' UTR as a natural miRancer of miR-1254 overrides tamoxifen resistance. *Cell Res* 2016;26(6):655–73. <https://doi.org/10.1038/cr.2016.32>.
- [45] Schnall-Levin M, Rissland OS, Johnston WK, Perrimon N, Bartel DP, Berger B. Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Res* 2011;21(9): 1395–403. <https://doi.org/10.1101/gr.121210.111>.
- [46] Haussler J, Syed AP, Bilen B, Zavolan M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* 2013;23(4):604–15. <https://doi.org/10.1101/gr.139758.112>.
- [47] Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet* 2012;13(4):271–82. <https://doi.org/10.1038/nrg3162>.
- [48] Lewis BP, Shih I, Jones-Rhoades MW, et al. Prediction of mammalian microRNA targets. *Cell* 2003;115: 787–98.
- [49] Yue D, Liu H, Huang Y. Survey of computational algorithms for MicroRNA target prediction. *Curr Genomics* 2009;10:478–92.

- [50] Mahen EM, Watson PY, Cottrell JW, Fedor MJ. mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biol* 2010;8(2):e1000307. <https://doi.org/10.1371/journal.pbio.1000307>.
- [51] Marín RM, Vanícek J. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res* 2011;39(1):19–29. <https://doi.org/10.1093/nar/gkq768>.
- [52] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* 2011;18(10):1139–46. <https://doi.org/10.1038/nsmb.2115>.
- [53] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 2010;11(8):R90. <https://doi.org/10.1186/gb-2010-11-8-r90>.
- [54] Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. *Genes Dev* 2004;18(5):504–11. <https://doi.org/10.1101/gad.118440>.
- [55] Liu H, Yue D, Zhang L, Bai Z, Lei X, Gao SJ, Huang Y. A machine learning approach for miRNA target prediction. *IEEE Int Workshop Genomic Signal Process Stat* 2008;2008:1–3.
- [56] Lukasik A, Wójcikowski M, Zielenkiewicz P. Tools4miRs—one place to gather all the tools for miRNA analysis. *Bioinformatics* 2016;32(17):2722–4. <https://doi.org/10.1093/bioinformatics/btw189>.
- [57] Kotagama K, Babb CS, Wolter JM, Murphy RP, Mangone M. A human 3'UTR clone collection to study post-transcriptional gene regulation. *BMC Genom* 2015;16:1036. <https://doi.org/10.1186/s12864-015-2238-1>.
- [58] Hamzeiy H, Allmer J, Yousef M. Computational methods for microRNA target prediction. *Methods Mol Biol* 2014;1107:207–21. https://doi.org/10.1007/978-1-62703-748-8_12.
- [59] Riffó-Campos ÁL, Riquelme I, Brebi-Mievile P. Tools for sequence-based miRNA target prediction: what to choose? *Int J Mol Sci* 2016;17(12):E1987. <https://doi.org/10.3390/ijms17121987>.
- [60] Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, et al. CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res* 2013;41(14):e138. <https://doi.org/10.1093/nar/gkt435>.
- [61] Ragan C, Zuker M, Ragan MA. Quantitative prediction of miRNA-mRNA interaction based on equilibrium concentrations. *PLoS Comput Biol* 2011;7(2):e1001090. <https://doi.org/10.1371/journal.pcbi.1001090>.
- [62] John B, Sander C, Marks DS. Prediction of human microRNA targets. *Methods Mol Biol* 2006;342:101–13.
- [63] Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999;288(5):911–40.
- [64] Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 1999;49(2):145–65.
- [65] Lorenz R, Bernhart SH, Höner Zu SC, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;6:26. <https://doi.org/10.1186/1748-7188-6-26>.
- [66] Gruber AR, Bernhart SH, Lorenz R. The ViennaRNA web services. *Methods Mol Biol* 2015;1269:307–26. https://doi.org/10.1007/978-1-4939-2291-8_19.
- [67] Vella MC, Reinert K, Slack FJ. Architecture of a validated microRNA::target interaction. *Chem Biol* 2004;11(12):1619–23.
- [68] Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* 2009;10(2):141–8. <https://doi.org/10.1038/nrm2619>.
- [69] Wang X. Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics* 2014;30(10):1377–83. <https://doi.org/10.1093/bioinformatics/btu045>.
- [70] Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA* 2006;12(2):192–7. <https://doi.org/10.1261/rna.2239606>.

- [71] Xu J, Zhang R, Shen Y, Liu G, Lu X, Wu CI. The evolution of evolvability in microRNA target sites in vertebrates. *Genome Res* 2013;23(11):1810–6. <https://doi.org/10.1101/gr.148916.112>.
- [72] Elefant N, Altuvia Y, Margalit H. A wide repertoire of miRNA binding sites: prediction and functional implications. *Bioinformatics* 2011;27(22):3093–101. <https://doi.org/10.1093/bioinformatics/btr534>.
- [73] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nat Genet* 2007;39(10):1278–84. <https://doi.org/10.1038/ng2135>.
- [74] Min H, Yoon S. Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med* 2010;42(4):233–44.
- [75] Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. Target mRNA abundance dilutes microRNA and siRNA activity. *Mol Syst Biol* 2010;6:363. <https://doi.org/10.1038/msb.2010.24>.
- [76] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 2007;27(1):91–105.
- [77] Robins H, Press WH. Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. *Proc Natl Acad Sci U S A* 2005;102(43):15557–62.
- [78] Sturm M, Hackenberg M, Langenberger D, Frishman D. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinf* 2010;11:292. <https://doi.org/10.1186/1471-2105-11-292>.
- [79] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 1998;14(1):55–67.
- [80] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. *PLoS Biol* 2004;2(11):e363. <https://doi.org/10.1371/journal.pbio.0020363>.
- [81] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 2015;4. <https://doi.org/10.7554/eLife.05005>.
- [82] Chiang HR, Schoenfeld LW, Ruby JG, Auyueung VC, Spies N, Baek D, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* 2010;24(10):992–1009. <https://doi.org/10.1101/gad.1884710>.
- [83] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, et al. A Uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet* 2015;49:213–42. <https://doi.org/10.1146/annurev-genet-120213-092023>.
- [84] Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res* 2015;43(Database issue):D146–52. <https://doi.org/10.1093/nar/gku1104>.
- [85] Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics* 2016;32(9):1316–22. <https://doi.org/10.1093/bioinformatics/btw002>.
- [86] Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res* 2013;41(Web Server issue):W169–73. <https://doi.org/10.1093/nar/gkt393>.
- [87] Reczko M, Maragakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. *Bioinformatics* 2012;28(6):771–6. <https://doi.org/10.1093/bioinformatics/bts043>.
- [88] Kanoria S, Rennie W, Liu C, Carmack CS, Lu J, Ding Y. STarMir tools for prediction of microRNA binding sites. *Methods Mol Biol* 2016;1490:73–82. https://doi.org/10.1007/978-1-4939-6433-8_6.
- [89] Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, et al. STarMir: a web server for prediction of microRNA binding sites. *Nucleic Acids Res* 2014;42(Web Server issue):W114–8. <https://doi.org/10.1093/nar/gku376>.
- [90] Rennie W, Kanoria S, Liu C, Mallick B, Long D, Wolenc A, et al. STarMirDB: a database of microRNA binding sites. *RNA Biol* 2016;13(6):554–60. <https://doi.org/10.1080/15476286.2016.1182279>.

- [91] Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 2003;31(24):7280–301.
- [92] Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 2004;32(Web Server issue):W135–41. <https://doi.org/10.1093/nar/gkh449>.
- [93] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. *Nat Genet* 2005;37(5):495–500.
- [94] Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA* 2004;10(10):1507–17. <https://doi.org/10.1261/rna.5248604>.
- [95] Andrés-León E, Núñez-Torres R, Rojas AM. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 2016;6:25749. <https://doi.org/10.1038/srep25749>.
- [96] Andrés-León E, Gómez-López G, Pisano DG. Prediction of miRNA-mRNA interactions using miRGate. *Methods Mol Biol* 2017;1580:225–37. https://doi.org/10.1007/978-1-4939-6866-4_15.
- [97] Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinf* 2006;7(Suppl. 5):S20.
- [98] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 2009;37(Database issue):D105–10. <https://doi.org/10.1093/nar/gkn851>.
- [99] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 2012;40(Database issue):D222–9. <https://doi.org/10.1093/nar/gkr1161>.
- [100] Wang D, Gu J, Wang T, Ding Z. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics* 2014;30(15):2237–8. <https://doi.org/10.1093/bioinformatics/btu155>.
- [101] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 2014;42(Database issue):D78–85. <https://doi.org/10.1093/nar/gkt1266>.
- [102] Zhang Y, Zang Q, Zhang H, Ban R, Yang Y, Iqbal F, et al. DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data. *Nucleic Acids Res* 2016;44(W1):W166–75. <https://doi.org/10.1093/nar/gkw427>.
- [103] Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics* 2016;32(18):2768–75. <https://doi.org/10.1093/bioinformatics/btw318>.
- [104] Yue D, Guo M, Chen Y, Huang YA. Bayesian decision fusion approach for microRNA target prediction. *BMC Genom* 2012;13(Suppl. 8):S13. <https://doi.org/10.1186/1471-2164-13-S8-S13>.
- [105] Sun X, Dong B, Yin L, Zhang R, Du W, Liu D, Shi N, Li A, Liang Y, Mao L. PMTED: a plant microRNA target expression database. *BMC Bioinf* 2013;14:174. <https://doi.org/10.1186/1471-2105-14-174>.
- [106] Chae H, Rhee S, Nephew KP, Kim S. BioVLAB-MMIA-NGS: microRNA-mRNA integrated analysis using high-throughput sequencing data. *Bioinformatics* 2015;31(2):265–7. <https://doi.org/10.1093/bioinformatics/btu614>.
- [107] Coronello C, Benos PV. ComiR: combinatorial microRNA target prediction tool. *Nucleic Acids Res* 2013;41(Web Server issue):W159–64.
- [108] Wan C, Gao J, Zhang H, Jiang X, Zang Q, Ban R, Zhang Y, Shi QCPSS. 2.0: a computational platform update for the analysis of small RNA sequencing data. *Bioinformatics* 2017;33(20):3289–91. <https://doi.org/10.1093/bioinformatics/btx066>.
- [109] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008;36(Database issue):D154–8.
- [110] Wu WS, Tu BW, Chen TT, Hou SW, Tseng JT. CSmiRTar: condition-Specific microRNA targets database. *PLoS One* 2017;12(7):e0181231. <https://doi.org/10.1371/journal.pone.0181231>. eCollection 2017.

- [111] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinf* 2007;8:69.
- [112] Gerlach W, Giegerich R. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics* 2006;22(6):762–4.
- [113] Ahmadi H, Ahmadi A, Azimzadeh-Jamalkandi S, Shoorehdeli MA, Salehzadeh-Yazdi A, Bidkhori G, et al. HomoTarget: a new algorithm for prediction of microRNA targets in *Homo sapiens*. *Genomics* 2013;101(2):94–100. <https://doi.org/10.1016/j.ygeno.2012.11.005>.
- [114] Giurato G, De Filippo MR, Rinaldi A, Hashim A, Nassa G, Ravo M, et al. iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinf* 2013;14:362. <https://doi.org/10.1186/1471-2105-14-362>.
- [115] Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res* 2017. <https://doi.org/10.1093/nar/gkx279> [Epub ahead of print].
- [116] Kim J, Levy E, Ferbrache A, Stepanowsky P, Farcas C, Wang S, et al. MAGI: a Node.js web service for fast microRNA-Seq analysis in a GPU infrastructure. *Bioinformatics* 2014;30(19):2826–7. <https://doi.org/10.1093/bioinformatics/btu377>.
- [117] Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA²: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res* 2012;40(Web Server issue):W13–21. <https://doi.org/10.1093/nar/gks460>.
- [118] Rusinov V, Baev V, Minkov IN, Tabler M. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res* 2005;33(Web Server issue):W696–700.
- [119] Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 2012;28(15):2062–3. <https://doi.org/10.1093/bioinformatics/bts344>.
- [120] Tokar T, Pastrello C, Rossos AEM, Abovsky M, Hauschild AC, Tsay M, Lu R, Jurisica I. mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Res* 2018;46(D1):D360–70. <https://doi.org/10.1093/nar/gkx1144>.
- [121] Friedman Y, Karsenty S, Linial M. miRrror-Suite: decoding coordinated regulation by microRNAs. *Database* 2014;2014:bau043. <https://doi.org/10.1093/database/bau043>. Print 2014.
- [122] Hsu JB, Chiu CM, Hsu SD, Huang WY, Chien CH, et al. miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinf* 2011;12:300. <https://doi.org/10.1186/1471-2105-12-300>.
- [123] Chou CH, Lin FM, Chou MT, Hsu SD, Chang TH, Weng SL, et al. A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. *BMC Genom* 2013;14(Suppl. 1):S2. <https://doi.org/10.1186/1471-2164-14-S1-S2>.
- [124] Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, et al. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol* 2013;10(7):1087–92. <https://doi.org/10.4161/rna.25193>.
- [125] Mitra R, Bandyopadhyay S. MultiMiTar: a novel multi objective optimization based miRNA-target prediction method. *PLoS One* 2011;6(9):e24583. <https://doi.org/10.1371/journal.pone.0024583>.
- [126] Maxwell EK, Campbell JD, Spira A, Baxevanis AD. SubmiRine: assessing variants in microRNA targets using clinical genomic data sets. *Nucleic Acids Res* 2015;43(8):3886–98. <https://doi.org/10.1093/nar/gkv256>.
- [127] Liu H, Yue D, Chen Y, Gao SJ, Huang Y. Improving performance of mammalian microRNA target prediction. *BMC Bioinf* 2010;11:476. <https://doi.org/10.1186/1471-2105-11-476>.
- [128] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 2007;13(11):1894–910.
- [129] Zhao W, Liu W, Tian D, Tang B, Wang Y, Yu C, et al. wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics* 2011;27(21):3076–7. <https://doi.org/10.1093/bioinformatics/btr504>.

- [130] Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. microRNA target prediction programs predict many false positives. *Genome Res* 2017;27(2):234–45. <https://doi.org/10.1101/gr.205146.116>.
- [131] Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res* 2009; 37(Web Server issue):W356–62. <https://doi.org/10.1093/nar/gkp294>.
- [132] Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat Methods* 2015;12(8):697. <https://doi.org/10.1038/nmeth.3485>.

FURTHER READING

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* 2005;123(6):1133–46.

This page intentionally left blank

INTEGRATIVE ANALYSIS OF EPIGENOMICS DATA

7

Cenny Taslim¹, Stephen L. Lessnick^{1,2}, Simon Lin³

¹*Center for Childhood Cancer and Blood Diseases, Nationwide Children's Hospital Research Institute, Columbus, OH, United States;* ²*Division of Pediatric Hematology/Oncology/BMT, The Ohio State University College of Medicine, Columbus, OH, United States;* ³*Research Information Solutions and Innovation, Nationwide Children's Hospital, Columbus, OH, United States*

INTRODUCTION

Understanding the functional consequences of genome-wide variation and how it affects complex human diseases remains a critical challenge for the scientific community. With the advent of next-generation sequencing, a wealth of experimental information from diverse “omics” data has been generated enabling us to comprehensively study alterations in the sequences of DNA and how epigenetic modifications alter chromatin structure, modulating gene activation or repression. Gene regulation is a complex process that involves genetic and epigenetic mechanisms. A site-specific transcription factor (TF) is thought to interact with regulatory elements to govern the activity of target genes [1]. The binding of these regulatory factors is restricted by complex chromatin structure. Epigenetic modifications such as histone modifications alter chromatin structure and DNA accessibility, thereby regulating gene expression. For example, promoters are typically associated with histone H3 lysine 4 trimethylation (H3K4me3), while enhancers (distal regulatory elements) often show histone H3 lysine 4 monomethylation (H3K4me1) [2] (Fig. 7.1). State-of-the-art next-generation sequencing (NGS) technologies such as ChIP-seq (chromatin immunoprecipitation sequencing) can provide genome-wide profiles of the localization of transcription factors, global histone modification patterns, and chromatin accessibility depending on the antibody used to pull down the protein. In ChIP-seq experiment, a region in the genome enriched with many sequence reads (also called a peak) represents the location where the target protein binds on the DNA. ChIP-seq peaks can be identified using peak calling algorithms such as MACS2 [3]. For a review of ChIP-seq analysis workflow and other peak calling algorithms, see Ref. [4]. Another NGS technology, RNA-seq (RNA sequencing), provides genome-wide transcriptional activity (Fig. 7.1). Peaks from RNA-seq experiments represent gene expression level and can be analyzed using tools such as DESeq2 [5]. Conesa et al. have laid out a guideline for RNA-seq data analysis [6]. A rigorous mechanistic understanding of transcriptional regulation is needed to integrate the combinatorial effect of transcription factors and their target genes,

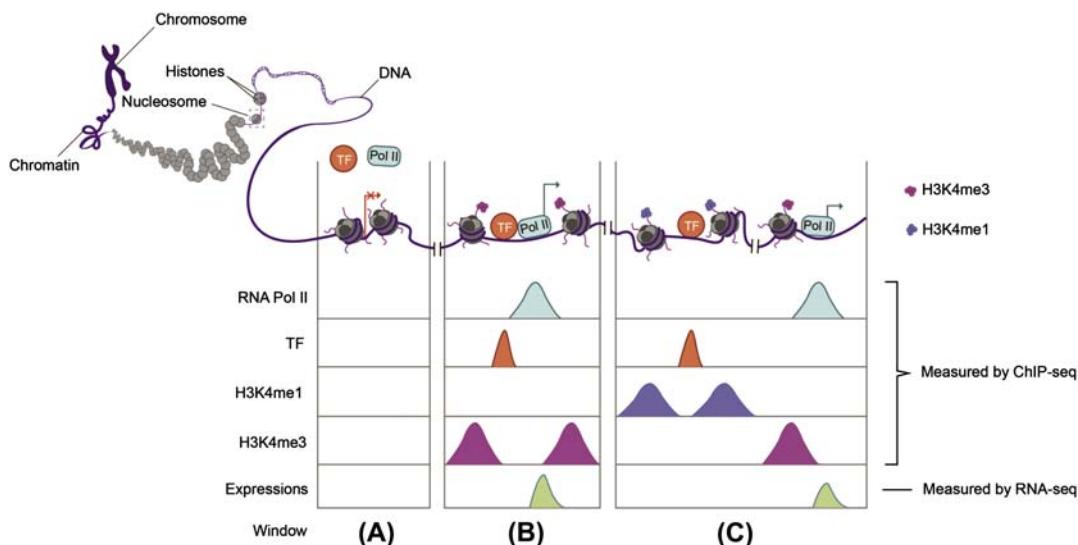


FIGURE 7.1 Chromatin structure and histone modification at regulatory elements.

Each rectangle represents a genomic window (segment). Peaks in each row are regions showing reads enrichment at the genomic locations where the proteins are bound in ChIP-seq experiment or where the genes are expressed in RNA-seq experiment. Peak height represents the binding intensity or expression level in ChIP-seq and RNA-seq experiment, respectively. (A) Closed chromatin restricting binding of TF resulting in repression of target gene. (B) Accessible promoter regions can be marked by H3K4me3 histone modification, allowing binding of TF which activates target gene. (C) Active enhancer marked by H3K4me1 is bound by TF that activates the transcription of distal gene.

as well as the chromatin organization of their regulatory elements [1]. Several large research projects, including The Encyclopedia of DNA Elements (ENCODE) [7], The Cancer Genome Atlas (TCGA) [8], and the NIH Roadmap Epigenomics Mapping Consortium [9], provide expansive resources of large-scale multiomics data sets that are publicly available for researchers. Many computational and statistical methods have been implemented to perform integrative analysis of multiomics data [2,10–15]. Some studies focus on analyzing multiple histone modifications, others focus on integration of multiple transcription factor binding sites, and there are those that also integrate gene expression analysis. However, there is no universal method that is used to integrate all these multiomics data.

This chapter provides a general guideline and overview of methodologies that have been implemented to jointly analyze multiple omics data generated using ChIP-seq and RNA-seq technologies. We will focus on methods for integrative analysis of transcription factors, histone modifications, and gene expression. This chapter does not intend to provide a detailed description of the methods or a comparison of the methods. Instead it attempts to provide general overview on how these machine learning/statistical methods were applied to jointly analyze multiple NGS data sets and answer different biological questions. For more detailed implementation of the analysis tools presented here, please refer to the following references [10–17]. It is important to note that there are many other methods developed to integrate multiple omics data that are not presented here. Table 7.1 lists all the computational tools described in this chapter.

Table 7.1 A Curated List of Computational Tools and Libraries

Methods Used	Description	Software Availability/URL	References
Multinomial logistics regression and hierarchical clustering	Semisupervised classification of differentially regulated genes into four classes	glmnet (R package)	[16]
Logistics regression	Supervised classification of differentially regulated genes into two classes with distinct regulation mechanism	glm (R package)	[17]
Finite mixture model	Identification of novel and alternative promoter regions based on shapes of transcription factors and histone modification binding patterns	MATLAB (global optimization toolbox)	[14]
Bayesian mixture model	Identification of differentially expressed genes by modeling both gene expression and histone modification binding using Bayesian mixture model	epigenomix (R package)	[12]
ChromHMM	Unsupervised method to learn and characterize chromatin states using hidden Markov model	Java-based ChromHMM	[15]
RFECS	Discriminant random forest method used to identify genome-wide enhancers	In-house MATLAB scripts	[13]
Self-organizing map (SOM)	Method based on artificial neural network to perform high-dimensional clustering of multiple transcription factor binding sites studying patterns of colocalization	Kohonen (R package)	[11]
DeepSEA	Prediction of transcription factors and histone marks binding in multiple cell types and the effect of single nucleotide variants on binding using deep learning based algorithm	http://deepsea.princeton.edu/job/analysis/create/	[10]

QUALITY CONTROL AND DATA PREPROCESSING

Effective integration and analysis of multiple omics data are expected to provide insights into the complex molecular mechanisms of diseases. Unfortunately, technical heterogeneity (or batch effects) such as experiment times, reagents, sequencing depth, sequencing quality, etc. can confound the biological signal and lead to incorrect conclusions [18]. Therefore, all data need to be of high quality and processed uniformly. Uniformity in terms of read length and run type (single- or paired-end) is also important. Common quality control metrics such as number of reads, GC content, and base-calling quality scores are typically measured for all types of raw sequencing reads using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Postmapping quality metrics specific for ChIP-seq and RNA-seq experiments will be discussed below.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become the go-to method for identifying genome-wide protein–DNA interactions (e.g., transcription factor or histone protein binding). Eliminating or replacing poor quality data is important for successful data integration. Several postmapping quality metrics have been developed and standards are emerging [19,20]. Currently, it is recommended for each experiment to have two or more biological replicates with matching input control experiments to ensure reliability and reproducibility. Concordance across replicates can be assessed using irreproducible discovery rate (IDR) [21]. Several other criteria have been developed to assess the quality of experiments: (1) Library complexity which signifies low amount of DNA or problems with library construction can be assessed using nonredundant fraction (NRF) and/or PCR bottlenecking coefficients (PBC) and (2) Signal-to-noise ratio can be used to assess the degree of noise in the input control, while signal in the ChIP sample can be evaluated by cross-correlation analysis and fraction of reads in peaks (FRiP) [4,19,20]. Different standards are applied depending on whether the target protein binds at specific locations that are punctate (i.e., transcription factor) or at longer chromatin domains (i.e., histone modification) [19]. Additionally, ChIP quality assessment in cells under which the protein is perturbed (i.e., performed in an siRNA, CRISPR, or genetic knockout background) or in cells where repressive marks are present should be interpreted differently [22]. The ENCODE Consortium has developed standards for these quality metrics including the recommended sequencing read depth and number of aligned reads that they keep up-to-date to ensure high quality data (<https://www.encodeproject.org/data-standards/>).

RNA-seq is a well-established approach to quantify gene expression genome-wide. Quality assessment of RNA-seq experiment is critical to ensure downstream analysis of biologically relevant data. As with ChIP-seq experiments, two or more biological replicates are recommended. Typically, concordance between replicates is measured at the gene level with a Spearman correlation [6]. Several quality metrics have been proposed to assess RNA-seq data: (1) Capture efficiency metric, which measures the percentage of reads that map to the intended target region, can be used to identify low sample quality, library contamination, or inefficient removal of ribosomal RNA; (2) GC content and 3'–5' bias can help identify nonuniform coverage of transcripts [23,24]. Other quality control metrics such as detection of aberrant splicing have also been implemented in RSeQC [25].

In addition to the quality metrics described above, when consolidating multiple ChIP-seq data for integrated analysis, it is also important to avoid artificial differences due to sequencing depth and mappability [26]. To avoid these artificial differences, all histone and transcription factor ChIP-seq data should have comparable sequencing depth, or uniformly subsampled. Read counts should be standardized. Additionally, reads map to multiple locations should be discarded and regions associated with repetitive regions should be excluded [2]. Finally, dimension reduction approaches such as principal component analysis (PCA), which decomposes high-dimensional data into components that explain most of the variation in the data, can be used to identify batch effects or outliers [27].

RELATIONSHIP BETWEEN HISTONE MODIFICATION PATTERN, TRANSCRIPTION FACTOR BINDING, AND mRNA EXPRESSION LEVEL

A critical challenge in understanding disease mechanisms is elucidating the underlying regulation of gene expression. The widespread application of the next-generation sequencing method has generated large-scale genome-wide molecular profiles, enabling researchers to study the combinatorial effects of

chromatin organization, regulatory elements, and transcription factor binding that govern gene expression and underlie cell state changes [7,28].

Transcription factors (TFs) bind to specific DNA motifs in the promoter or enhancer region and regulate transcription either through proximity to the transcription start site (TSS) or higher-order chromatin looping [29,30]. Epigenetic histone modifications are also known to regulate gene expression by either reorganizing the local chromatin structure to control accessibility of TF binding sites [31] or by recruiting other chromatin remodelers [32]. Thus, nucleosomes in the vicinity of transcriptional regulatory elements (e.g., enhancer or promoter regions) typically contain histones with specific posttranslational modifications (PTM), such as H3 lysine 4 mono- and trimethylation (H3K4me1/H3K4me3) (Fig. 7.1). Dysregulation of gene expression caused by aberrant TF or histone modification has been directly associated with many diseases including Ewing sarcoma [16], prostate cancer [17], and various other disorders [33]. Here, we focus on methods for integrative analysis of ChIP-seq (for measuring TF binding and histone modification) and RNA-seq (for measuring mRNA expression). In this section, we review several existing statistical methods used to investigate the relationship between histone modification, TF binding, and gene expression.

REGRESSION ANALYSIS

To understand the effect of TF binding and chromatin regulation on gene expression, Tomazou et al. first identified the differentially regulated genes, comparing cell lines with two different conditions [16]. The expression of these genes was then correlated to the histone binding intensities at their promoter regions. For each histone mark, the maximum peak score within a 1–5-kb window (depending on the type of histone mark) around the TSS was assigned, resulting in a matrix of genes × histone modification. A weighted discretized version of this matrix was then used as input to perform unsupervised clustering. In a semisupervised approach, 40–50 genes from four distinct clusters were selected based on unsupervised clustering and expert classification. These genes were then used to build a multinomial logistic regression to model the four clusters. This model was used to predict the cluster assignment of remaining genes. Using this approach, the authors identified four clusters of genes: Cluster 1, genes associated with four active promoter marks but not repressive mark; Cluster 2, genes associated with two active promoter marks; Cluster 3, genes associated with one active promoter mark; Cluster 4, genes associated with repressive mark in the presence or absence of active promoter marks. Genes in the same group that respond differently to a transcription factor were associated with distinct molecular functions. For example, a gene identified in Cluster 1 whose expression is anticorrelated with a transcription factor is illustrated in Fig. 7.2. In their implementation, genes were clustered based on both supervised (i.e., expert classification) and unsupervised (i.e., hierarchical clustering) approaches.

Taslim et al. [17] used a logistic regression model to perform a supervised classification to study the effect of distance and combinatorial binding of transcription factors, and active histone modification on regulation of gene expression. Specifically, a logistic model was built using variables such as distance from regulated genes to transcription factor binding sites and the presence of overlap between several transcription factors and active histone mark binding sites, with expression of regulated genes as response. Initially, models with up to six variables were considered, followed by selection of a model with the optimal specificity and sensitivity in a fivefold cross-validation. Bayes factor like criterion was used to perform classification on the regulated genes. Bayes factor is defined as a ratio of

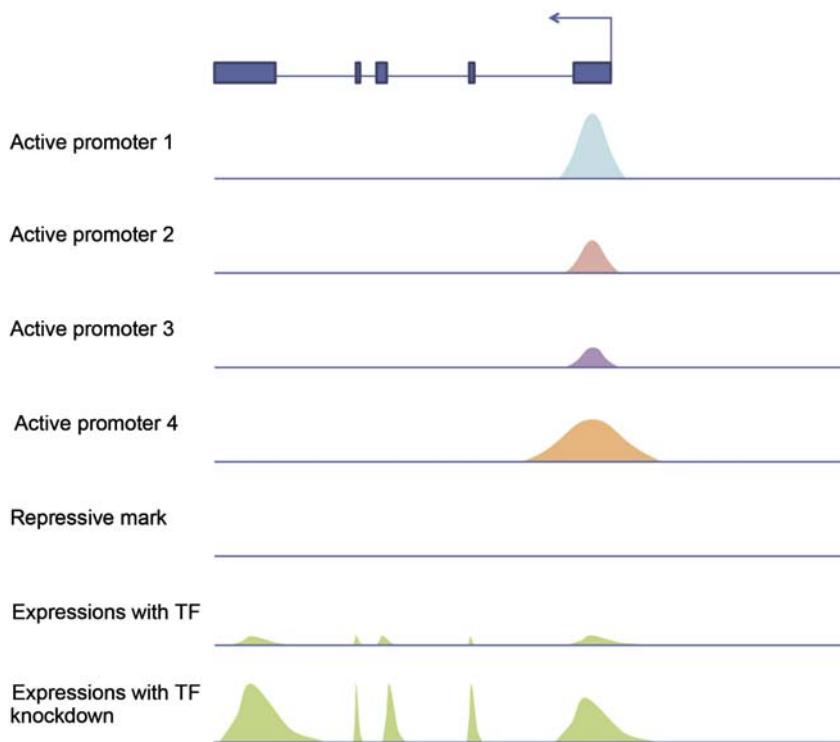


FIGURE 7.2 Illustration of a gene characterized by four active promoter marks which has negative correlation with TF.

Gene model is depicted on top with exons represented by blue boxes and introns as blue lines.

probability of null and alternative hypothesis. It provides an assessment that represents how well a hypothesis predicts the empirical data relative to the alternative [34]. Using this approach, the authors identified a group of treatment-responsive genes which has a relatively short long-range regulation mechanism facilitated by a combination of transcriptional factors (including a TF that regulates chromatin structure) and histone modifications.

MIXTURE MODEL

In order to identify novel and alternative promoter regions, Taslim et al. investigated the shapes of RNA polymerase II and active histone modification binding patterns [14]. Combining binding patterns of histone modification that has been shown to be enriched in active promoter regions with transcriptional factor such as RNA polymerase II (Pol-II) and RNA-seq expression provide more specific identification of patterns within promoter regions. Furthermore, phenomenon such as RNA Pol-II stalling, in which the Pol-II binds at the promoter but the gene is not transcribed, may display binding patterns specific to their functions. Therefore, modeling the shapes of these binding patterns may provide additional insights. Specifically, normalized ChIP-seq read counts of Pol-II and active histone mark in the 10-kb regions around TSSs were segmented into 100-bp bins. These binding patterns were

then grouped using K-means clustering to identify the four classes of promoter binding patterns. A mixture model with two double-exponential and uniform components was fitted to the averaged binding patterns within a group. The double-exponential component is designed to capture the shape of Pol-II and active histone mark that are unimodal and bimodal, respectively. The uniform component is used to capture the tails of Pol-II binding profile. The mixture model was fitted using a generalized pattern search algorithm [35] with Kullback–Leibler distance metrics [36]. This model of combined patterns representing binding signatures for promoter regions was then used to search for unknown genomic regions that may represent promoters of unknown genes or known genes with novel alternative promoters. The authors also showed that genes whose promoters display the combined binding patterns were associated with higher expression compared to those who do not. Fig. 7.3, box 2 shows an example of a gene which displays the combined binding patterns of active histone and TF.

Joint analysis on histone ChIP-seq and RNA-seq expression data was performed by Klein et al. using Bayesian mixture model to identify genes with significant differences in both expression and histone modification bindings comparing two different conditions [12]. For each differential gene, expressions and histone modifications were integrated into a correlation variable (Z) which was

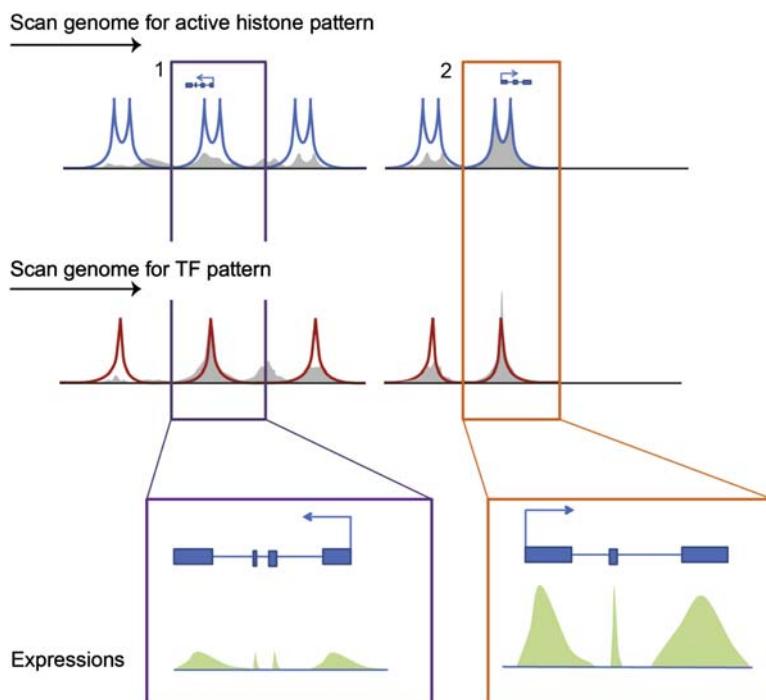


FIGURE 7.3 Illustration of the scanning process to find regions with combined TF and active histone patterns.

The models of active histone and TF patterns are shown as blue and red curves, respectively. These fitted patterns were used to scan the genome. The actual binding data are shown as gray areas, top row for active histone and bottom row for TF. Box 1 shows a gene associated with TF but not active histone pattern. This gene is shown to have less expression than the gene in box 2 which displays both TF and active histone patterns.

calculated by multiplying the standardized difference of expressions and binding intensities in two distinct conditions. The distribution of Z is then fitted using Bayesian mixture model with normal and exponential distributions. Active genes associated with increased binding of active histone marks are considered to be in the same direction and called positive component. Repressive genes associated with increased binding of repressive histone marks are thought to be in the opposite direction and called negative component. Both of these components are modeled by exponential distribution. The normal component represents genes that do change in expressions and histone bindings. The proportions of the mixture model are modeled with a Dirichlet prior distribution. Klein et al. used a Bayesian approach in which informative prior distributions need to be selected [12]. The model is fitted using Gibbs sampling which is based on Markov Chain Monte Carlo (MCMC) techniques [37,38], and model estimates were obtained by averaging the posterior distribution. Using this approach, for example, activated (repressed) genes associated with increased binding of active (repressive) histone marks can be identified.

IDENTIFICATION OF FUNCTIONAL REGULATORY REGIONS

Posttranslational histone modifications have been shown to change nucleosome dynamics allowing or blocking DNA–protein binding and thereby affecting the transcriptional activity of target genes [39,40]. There are a wide variety of posttranslational histone modifications, and it was suggested that multiple histone modifications act in a combinatorial fashion to specify distinct functional regulatory regions [41]. Here, we describe two methods that have been developed to integrate multiple histone modification ChIP-seq data, to identify and define functional regulatory regions such as active promoters, enhancers, repressed promoters, etc.

ChromHMM [15], which is based on hidden Markov model (HMM), was developed to model the combinatorial patterns of multiple chromatin elements. The goal is to capture the similarities of chromatin patterns across different cell types giving an unbiased classification of the genome into functional regions (i.e., chromatin states are modeled as latent variables in HMM). In the implementation of HMM, the genome was segmented into 200-bp regions (default value) that roughly approximate nucleosome sizes and are used as the “time” axis. Thus as input, multiple histone modification and general TF (and chromatin accessibility assays such as DNase-seq and FAIRE-seq) data were normalized and converted into a binary presence or absence call for each mark at 200-bp resolution [42]. ChromHMM was trained with two-stage nested parameter initialization using Euclidean distance for pruning the states. Models with up to 30 hidden states from multiple observed combinations of chromatin marks and 200 training iterations were completed to learn the parameters of the final HMM. Biologically these hidden states may represent different genomic features such as strong/weak enhancers, repressed/active promoters, open chromatin, etc. The number of states chosen was a compromise between capturing all the combinatorial complexity of chromatin marks and models that are useful for interpretation of biologically meaningful functional regions. The trained model was then used to compute the posterior probability of each state for each segment. Each region was then labeled using the state with the maximum posterior probability. The states discovered by the model were then annotated by the authors using large-scale systematic data-mining effort. An example of a state identified by ChromHMM is “poised promoter” regions that are associated with both activating and repressing histone modifications. In this state, genes are silenced but poised for activation.

RFECS (Random Forest based Enhancer identification from Chromatin States) was developed to integrate multiple histone modification profiles to identify enhancers. Enhancers are distal regulatory elements where transcription factors bind to and regulate gene expression [43]. Transcriptional regulations via enhancer elements are thought to be cell type-specific and dynamically regulated during development [44,45]. Identification of enhancers has been challenging due to a lack of common features and their locations that are often distal from their target genes. Rajagopal et al. [13] implemented a discriminant random forest based algorithm to identify genome-wide enhancers. RFECS was also used to identify the most informative set of histone modifications that can accurately predict enhancers. For training, only histone modifications that provide largely nonredundant information were selected and used as features. RFECS was constructed using binary classification trees. Specifically, each histone profile in bins of 100-bp surrounding binding sites of a protein called p300 known to be recruited to enhancers was used to train enhancer class. Similarly, histone profile around non-p300 binding sites was used to train nonenhancer class. At each node in the tree, a random subset of features was selected and parametric multivariate linear discriminant analysis (LDA) was applied to split the tree nodes. The optimal number of trees was determined by examining the area under the curve (AUC) generated using fivefold cross-validation. The final class predictions (enhancer vs. nonenhancer) were decided by majority voting. Variable importance (i.e., importance of individual histone modifications) was estimated using out-of-bag measure implemented in MATLAB® by Rajagopal et al. [13]. The most informative sets of histone modification for enhancer identification were chosen based on the ordering of variable importance across multiple replicates and cell types.

ASSOCIATION BETWEEN MULTIPLE TRANSCRIPTION FACTORS USING SELF-ORGANIZING MAP (SOM)

Different TFs need to co-operate at specific loci to perform complex regulation of their target genes. Investigating the colocalization of a combination of transcription factors to achieve complex and accurate regulation of target genes is challenging due to the high dimensionality of the data. For 128 TFs, there are more than 10^{38} possible combinations. In order to take into account the full combinatorial co-binding of many TFs, Xie et al. [11] used a type of artificial neural network called self-organizing map (SOM) to perform high-dimensional clustering based on the combinations of up to 128 TF binding profiles. An artificial neural network is inspired by the neural network in the brain, and as such, consists of layers of fully connected neurons that nonlinearly transformed input [46]. SOM provides a way to reduce the dimensionality of data using a compression technique known as vector quantization that also retains the topological properties [47]. In the study of TF colocalization, a block defined as the maximum overlap of TF binding sites is discretized as either bound or not bound for each TF and used as input to SOM. Each neuron in SOM is associated with weights the same dimension as the input and a position in the lower-dimensional map space. The weight for each neuron and its position (and also of its neighbors') is updated through unsupervised learning process that decays over time. The end result is a map comprised of neurons with distinct colocalization patterns at different genomic loci. Neurons with similar colocalization patterns will be in close proximity within the map space. This lower-dimensional map allows for clustering of regions which have the same TFs bound and therefore also identifies regions bound by high number of TFs (hotspot regions).

PREDICTION OF CHROMATIN AND TRANSCRIPTION BINDING SITES DIRECTLY FROM DNA SEQUENCES USING DEEP LEARNING

Genetic variations associated with disease often appear in noncoding parts of the genome. Accumulating evidence has shown that these variants occur in regulatory elements affecting chromosome structure and the modulation of genes [48]. Unfortunately, functional characterization of these non-coding variants remains a challenge. Classical machine learning methods such as hidden Markov model (ChromHMM) and random forest (RFECS) described in previous sections have been used to identify and annotate these regulatory regions [13,15]. However, these methods do not operate directly on the DNA sequences and therefore require preprocessing features (in non-deep learning approach, typically binding regions were first associated with presence or absence of a binding factor and then used as input).

Recent advances in artificial neural network lead to the development of state-of-the-art networks with large number of hidden layers known as deep networks. A deep network takes raw data at the lower (input) layer and automatically learns more abstract features that are needed for classification (or regression) through multiple levels of representation learning. These abstract features were obtained by combining and performing nonlinear transformation of outputs from previous layers [49]. These layers of features, which are automatically learned from data, can discover highly complicated structures in high-dimensional data. Deep learning has been successfully applied to image analysis and speech recognition with record breaking results [50,51].

Zhou et al. developed DeepSEA which uses a deep learning architecture called convolutional neural network (CNN) to learn DNA regulatory sequences (motifs) for predicting protein binding sites [10]. Rather than preprocessing relevant features, the network learns regulatory motifs directly from DNA sequences, enabling prediction of binding sites with single nucleotide sensitivity. The model learnt can then be used to improve noncoding variant interpretation. Zhou et al. start by training the model using binding site information from 919 features (125 DNase, 690 TFs, and 104 histones in multiple cell types) from ENCODE and Roadmap Epigenomics projects [2,7]. Each training sample is a 1000-bp sequence centered on 200-bp bin of binding sites. Each 1000-bp sequence (with at least one binding event) is paired with a labeled vector of 919 features (classes). A class is labeled 1 if more than half of 200-bp bin overlap with the peak regions, or 0 otherwise. As a deep learning model capable of learning directly from DNA sequences, the 1000-bp sequence is represented by a 1000×4 binary matrix corresponding to the four nucleotides (Fig. 7.4). Data sets were split into nonoverlapping training, validation, and testing sets. Validation sets were used to select hyperparameters for the network. DeepSEA consists of sequential, alternating convolution and pooling layers which allow learning the sequence motifs at increasing spatial scales. These layers are followed by a fully connected layer which integrates information from the full-length sequences and a final sigmoid layer that outputs the probability of each input sequence belonging to each individual class (Fig. 7.4).

The initial convolution layer can be considered as learning a set of position weight matrices (PWMs) [52] that minimizes prediction error. A PWM represents the common sequences (motif) where the target protein is bound. This layer also computes PWM scores with a 1 step moving window and thus provides invariance to small sequence shifts. Higher level layers receive input from lower layers and therefore learn more complex patterns from larger spatial ranges. DeepSEA shares learned predictive regulatory sequences across all classes. This means, for example, a regulatory sequence that

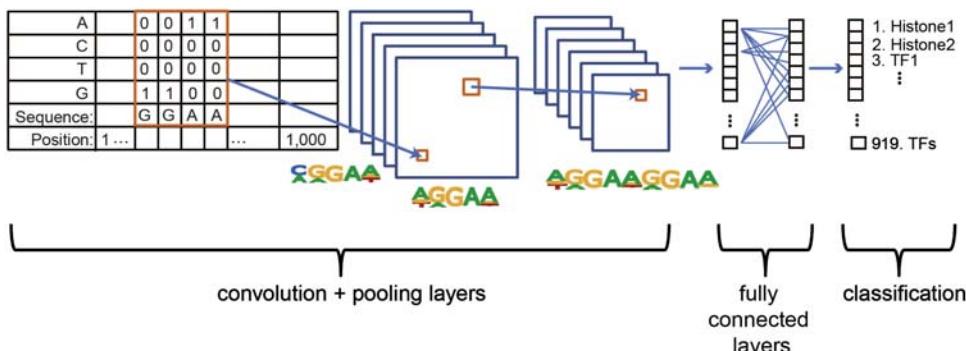


FIGURE 7.4 Illustration of DeepSEA algorithm.

DNA sequences in each peak region are represented as a 1000×4 binary matrix corresponding to the four nucleotides and used as input to DeepSEA. The initial convolution layer can be thought of learning a set of optimal position weight matrices that represents the local binding motif. Higher level layers receive input from lower layers and therefore learn more complex motifs at increasing spatial scales. These layers are followed by fully connected layers which integrate motif information from the full-length sequences and a final layer that outputs the class prediction.

is effective for classification of a specific TF can be used to predict the bindings of another physically interacting protein. On the holdout test sequences, DeepSEA was able to perform classification with high accuracy (median AUC ranges from 0.856 to 0.958). In order to identify functional SNPs, the DeepSEA model was first used to compute 919 class prediction for every pair of sequences carrying either the reference or the alternative allele. DeepSEA predictions for each SNP and evolutionary conservation scores were used to train boosted logistic regression classifiers. This functional prediction on SNPs was shown to outperform previous methods [10].

DISCUSSION

As technology advances and more data become available, integrative analysis will become critical as it provides a comprehensive view on the complexity of diseases' molecular mechanisms. The workflow to learn functional relationships from molecular data involves the following general steps: data cleaning, preprocessing, features extraction, model fitting or clustering, and finally evaluation of the results. Due to the high dimensionality of the data and typically limited number of samples, integrative analysis requires novel statistical and computational solutions. Many methods have been proposed but there is no single method that is universally applicable. Generally, these methods are applied to either make a prediction or to discover the underlying relationship between the variables of interest. For example, a researcher may be interested to understand what factors regulate a gene regulation and how they influence this modulation or simply to predict whether or not a gene is going to be up-/down-regulated. There are trade-offs associated with these goals. Optimizing prediction accuracy often comes from decreased interpretability of the model. Model-based methods (e.g., logistics regression, mixture model), although simple and usually fast computationally, require assumptions about the underlying relationships between input and output variables. Model-free algorithms do not make

strong assumptions about the underlying function; however, they are slower to compute and are typically more difficult to interpret. In both types of algorithms, selecting the most informative features are critical for performance and this is essentially the bottleneck, especially for high-dimensional omics data. State-of-the-art deep learning algorithm enables automatic features extraction that is scalable to high-dimensional data, making it a very attractive algorithm for application in genomics. However, deep learning requires a large amount of well-annotated data for the system to successfully learn features automatically from the data. Early applications in genomics have been to study the variation between regions, segmenting the genome to obtain large data sets for training. All of these methods can be a powerful complement to each other. We expect further improvement will help provide deeper understanding and eventually help solve important biological problems.

ACKNOWLEDGMENTS

The authors would like to thank Kirsten M. Johnson for critical reading of the manuscript and Trisha D. Shah for creating Fig. 7.1.

REFERENCES

- [1] Voss TC, Hager GL. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat Rev Genet* 2013;15:69–81.
- [2] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.
- [3] Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* 2012;7: 1728–40.
- [4] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform* 2017;18:279–90.
- [5] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014. <https://doi.org/10.1101/002832>.
- [6] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17:13.
- [7] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
- [8] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The cancer genome Atlas pan-cancer analysis project. *Nat Genet* 2013;45:1113–20.
- [9] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;28:1045–8.
- [10] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12:931–4.
- [11] Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. Dynamic trans-acting factor colocalization in human cells. *Cell* 2013;155:713–24.
- [12] Klein H-U, Schäfer M, Porse BT, Hasemann MS, Ickstadt K, Dugas M. Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics* 2014;30:1154–62.
- [13] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* 2013;9:e1002968. http://enhancer.ucsd.edu/renlab/RFECS_enhancer_prediction/.

- [14] Taslim C, Lin S, Huang K, Huang TH-M. Integrative genome-wide chromatin signature analysis using finite mixture models. *BMC Genomics* 2012;13:S3.
- [15] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215–6. <http://compbio.mit.edu/ChromHMM/>.
- [16] Tomazou EM, Sheffield NC, Schmidl C, Schuster M, Schöenegger A, Datlinger P, et al. Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. *Cell Rep* 2015;10:1082–95.
- [17] Taslim C, Chen Z, Huang K, Huang TH-M, Wang Q, Lin S. Integrated analysis identifies a class of androgen-responsive genes regulated by short combinatorial long-range mechanism facilitated by CTCF. *Nucleic Acids Res* 2012;40:4754–64.
- [18] Goh WWB, Wang W, Wong L. Why batch effects Matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;35:498–507.
- [19] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;22:1813–31.
- [20] Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* 2013;9:e1003326.
- [21] Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 2011;5:1752–79.
- [22] Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. *G3* 2014;4:209–23.
- [23] Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Funct Genomics* 2017;16:194–204.
- [24] DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire M-D, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 2012;28:1530–2.
- [25] Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;28:2184–5.
- [26] Wei LK, Au A. Computational epigenetics. In: *Handbook of Epigenetics*. 2nd ed. 2017. p. 167–90.
- [27] Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;17:628–41.
- [28] Rivera CM, Ren B. Mapping human epigenomes. *Cell* 2013;155:39–55.
- [29] Farnham PJ. Insights from genomic profiling of transcription factors. *Nat Rev Genet* 2009;10:605–16.
- [30] Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 2013;14:390–403.
- [31] Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell* 2007;128:669–81.
- [32] Kouzarides T. Chromatin modifications and their function. *Cell* 2007;128:693–705.
- [33] Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010;28:1057–68.
- [34] Raftery AE. Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall; 1996. p. 163–88.
- [35] Audet C, Dennis JE. Analysis of generalized pattern searches. *SIAM J Optim* 2002;13:889–903.
- [36] Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;22:79–86.
- [37] Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *J Am Stat Assoc* 2001;96:161–73.
- [38] Ishwaran H. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* 2000;87:371–90.
- [39] Wierer M, Mann M. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Hum Mol Genet* 2016;25:R106–14.
- [40] Bowman GD, Poirier MG. Post-translational modifications of histones that influence nucleosome dynamics. *Chem Rev* 2015;115:2274–95.

- [41] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008;40:897–903.
- [42] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2013;41:827–41.
- [43] Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 2011; 144:327–39.
- [44] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009;459:108–12.
- [45] Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* 2012;44:148–56.
- [46] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol* 2016;12:878.
- [47] Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982;43:59–69.
- [48] Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet* 2015;24:R102–10.
- [49] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- [50] Mohamed A-R, Sainath TN, Dahl G, Ramabhadran B, Hinton GE, Picheny MA. Deep Belief Networks using discriminative features for phone recognition. In: 2011 IEEE International Conference on Acoustics, speech and signal processing (ICASSP); 2011. <https://doi.org/10.1109/icassp.2011.5947494>.
- [51] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;60:84–90.
- [52] Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* 2006;22:e454–63.

DIFFERENTIAL DNA METHYLATION AND NETWORK ANALYSIS IN SCHIZOPHRENIA

8

Huang Kuo Chuan*Department of Nursing, Ching Kuo Institute of Management and Health, Keelung, Taiwan*

INTRODUCTION

Schizophrenia (SCZ) is the most complex and multidimensional disorder with high genetic predisposition. The contribution of genetic factors to the etiology of SCZ has gained increasingly focus on disease variants, which could explain only small portion of susceptibility. Besides, the environmental factors have been proved to play more important role in the disease mechanism. In recent years, several studies focused on the environment interactions to genetic alterations that bridge the gap between genetic expression alteration and environmental insults. In the majority of studies, epigenetic mechanism is the promising field in psychiatry medicine, which includes the potential links to the genetic expression architecture and environmental exposures. Epigenetic modifications of DNA methylation, histone modifications, and noncoding RNA provide important clues about disease mechanisms contributing to dysregulated expression of neurodevelopmental and metabolic genes in SCZ brain [1]. SCZ and other neurodevelopmental disorders are associated with abnormalities in both multiple genetic and epigenetic mechanisms. They result in dysregulation of altered gene expression during neural development. Polymorphisms and copy number variants in SCZ risk genes also contribute to the high heritability of the disease, such as polymorphisms in *COMT*, *BDNF*, and *FKBP5* genes, which might influence the outcomes of SCZ [2], but environmental factors that lead to epigenetic modifications may either reduce or exacerbate the expression of cell molecular functions and eventually result in behavioral phenotype changes associated with SCZ [3].

METHODOLOGY FOR DNA METHYLATION

There are many methylation databases regarding of DNA methylation in the research of cancer epigenetics. Few are found in schizophrenic brain methylation. The list below shows each prominent methylation database in brain development.

1. MethylomeDB [4]: a database of DNA methylation profiles of the brain.
2. MethBank [5]: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data.

3. DiseaseMeth version 2.0 [6]: a major expansion and update of the human disease methylation database.
4. PD_NGSAtlas [7]: a reference database combining next-generation sequencing epigenomic and transcriptomic data for psychiatric disorders (out of maintenance).

Another group of research database includes the most comprehensive genetic research in common psychiatric disease.

1. SZDB [8]: a database for SCZ genetic research.
2. SZGR 2.0 [9]: a one-stop shop of SCZ candidate genes.
3. SchizConnect [10]: mediating neuroimaging databases on SCZ and related disorders for large-scale integration (<http://schizconnect.org/>).

To comprehend, the underlying pathways, which are driven and affected by the genetic pathway databases could provide gene list influenced by selected pathways. The data in pathway databases were prepared in a variety of formats that could be easily used for systematic analysis.

1. NDEx [11]: a community resource for sharing and publishing of biological networks (old version name: PID, out of maintenance).
2. PANTHER version 11 [12,13]: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements.
3. KEGG [14,15]: Kyoto encyclopedia of genes and genomes.
4. Pathway Commons: access and discover data integrated from public pathway and interactions databases; a Web resource for biological pathway data.
5. Reactome pathway database: It is a free, open-source, curated, and peer reviewed pathway database. The goal is to provide intuitive bioinformatics tools for the visualization, interpretation, and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology, and education.

DNA methylation is an important epigenetic modification involved in many biological processes. Bisulfite treatment coupled with high-throughput sequencing provides an effective approach for studying genome-wide DNA methylation at base resolution.

Bisulfite sequencing analysis has been the standard method used by biological and medical researchers to detect cytosine methylation profiles in genomic DNA at the single-nucleotide level. It is a standard method for DNA methylation profile analysis, is widely used in basic and clinical studies. This method is limited, however, by the time-consuming data analysis processes required to obtain accurate DNA methylation profiles. The important methylation analysis tools are as follows:

1. Bismark [16]: Bismark is a combination of bisulfite treatment of DNA and high-throughput sequencing (BS-Seq), which can capture a snapshot of a cell's epigenomic state by revealing its genome-wide cytosine methylation at single-base resolution. It is a flexible tool for the time-efficient analysis of BS-Seq data.
2. BS-Seeker2 [17]: BS-Seeker2 is a full pipeline for mapping bisulfite sequencing data and generating DNA methylomes. It improves mapping ability over existing aligners by using local alignment.

METHYLATION SCHIZOPHRENIA NETWORK

1. It is a DNA methylation network interaction measure and detection of network oncomarkers [16].
 2. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood methylome [17].
 3. DNA methylation in SCZ in different patient-derived cell types [18].
 4. DNA methylation gene network dysregulation in peripheral blood lymphocytes of SCZ patients [19].
-

NOVEL PREDICTION APPLICATIONS

1. Prediction of SCZ in IEEE by fusing networks and mathematical models from SNPs, DNA methylation, and fMRI data [20].
2. SCZ interactome with 504 novel protein–protein interactions [21] (Table 8.1, Fig. 8.1).

The analytic flow chart illustrates the step-by-step workflow and databases in different stages, with corresponding algorithm and integrated disease gene databases by the biological modulation network analysis. The most updated protein interaction databases including BioGRID, BIND, MINT, MIPS, DIP, IntACT, Human Protein Reference Database (HPRD), and protein complex databases such as CORUM and PCDq for integrated application and analysis. For pathway enrichment analysis, the NDEX (PID), KEGG, Reactome, CellMap, and HumanCyc are included to be updated for pathway analysis.

CANDIDATE GENES IN SCHIZOPHRENIA

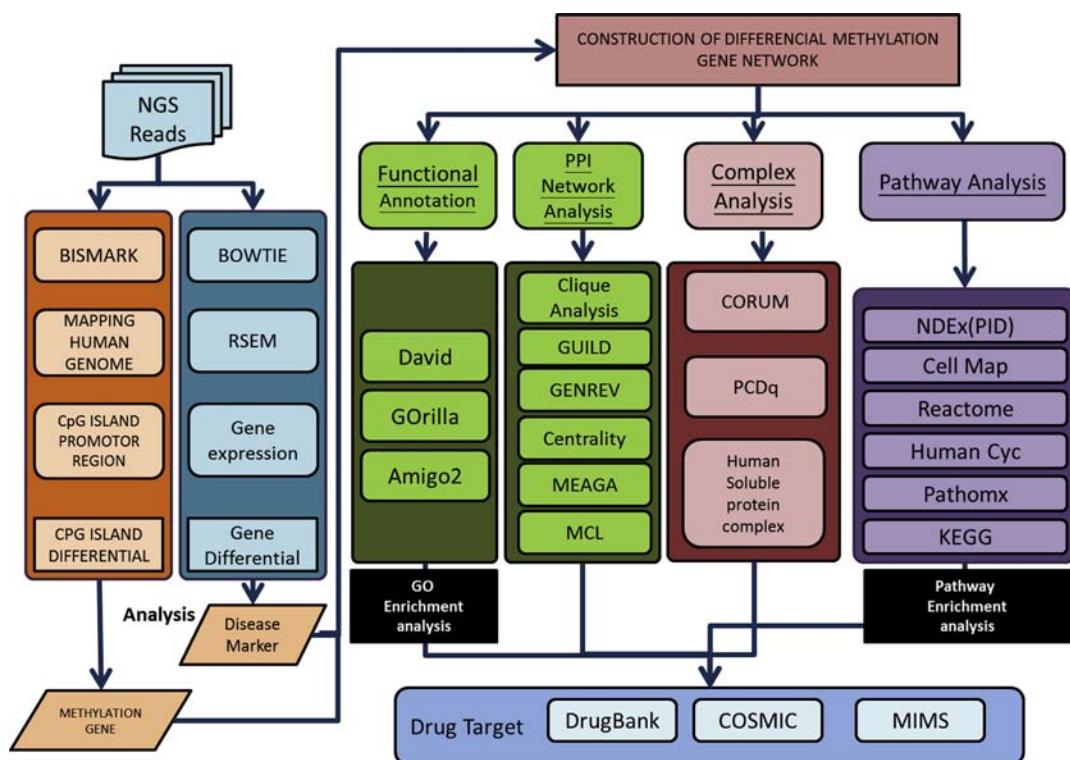
The preliminary study of schizophrenic candidate gene revealed that the overlapped candidate genes in different studies are low, that is, only 18%. It implicated that the outcome and phenotype of SCZ cannot be only explained by genetic variants. Gathering schizophrenic candidate gene databases from SZGene, Huang et al., Sellmann et al., Ayalew M et al., Wu et al., there are 15 genes including *DISC1*, *MTHFR*, *GAD1*, *COMT*, *BDNF*, *SYN2*, *RGS4*, *PRODH*, *NOS1*, *MAP6*, *HTR2A*, *DTNBPI*, *DRD4*, *DRD2*, *APOE*, which are enlisted as candidate genes in at least four different studies and validated to be associated with the cause of SCZ. Some of them are *UBE2L3*, *SEZ6L2*, and *RIMS3*.

SDMGs AND DISEASE MECHANISM OF SCHIZOPHRENIA

A total of 1689 genes (4% of *Homo sapiens* genes) are found to be differentially methylated in SCZ (schizophrenia differential methylation genes, SDMGs) from 38,000 *Homo sapiens* genes. A percentage of 39.6 of them are from promotor regions. A total of 123 genes coexist with differential expression and differential methylation. The different methylation profile in promotor regions may be the etiology for SCZ. To illustrate the interaction of each differentially methylated gene, a schizophrenic differential methylation network (SDMN) was constructed. The SDMN was generated by QQPI of SDMGs with underlying pathway enrichment. Among those SDMGs, there are

Table 8.1 The List of Next-Generation Sequencing of Transcriptome and Methylome From NCBI SRA Database of Schizophrenia

Accession	Description	Sample Type	Library Strategy	Library Source	No.	Year	Department
PRJNA 182544	Whole transcriptome analysis of postmortem human hippocampus dentate gyrus granule cells	Human hippocampus dentate gyrus granule cells	RNA-Seq	TRANSCRIPTOMIC	79	27-Nov-2012	Psychiatry and Behavioral Sciences, University of Washington
PRJNA 200967	RNA-seq in neurons derived from iPSCs in controls and patients with schizophrenia and 22q11 del	iPSC	RNA-Seq	TRANSCRIPTOMIC	8	1-May-2013	Behavioral Genetics, Psychiatry, Albert Einstein College of Medicine
PRJNA 219443	Interindividual variability contrasts with regional homogeneity in the human brain DNA methylome.	Human brain samples	MBD-Seq	GENOMIC	59	18-Sep-2013	Human Genetics Unit, MRC
PRJNA 235930	The DNA methylome and transcriptome of different brain regions in schizophrenia and bipolar disorder	<i>Homo sapiens</i> , brain disease epigenomics	RNA-Seq MeDIP-Seq ChIP-Seq	TRANSCRIPTOMIC GENOMIC GENOMIC	35 36 1	19-Jan-2014	Harbin medical university
PRJNA 260454	Genome-wide methylome analyses reveal novel epigenetic regulation patterns in schizophrenia and bipolar disorder	<i>Homo sapiens</i> , blood, epigenomics	RNA-Seq MeDIP-Seq	TRANSCRIPTOMIC GENOMIC	3 10	7-Sep-2014	Harbin Medical University

**FIGURE 8.1**

Analytic flow chart

10 hypermethylated promoters in SDMN including GNA13, CAPNS1, GABPB2, GIT2, LEFTY1, NDUFA10, MIOS, MPHOSPH6, PRDM14, and RFWD2. They represent the key roles in modulating specific regulatory functions in SCZ. The 10 schizophrenic hypermethylated genes discovered by SDMN are associated with biological functions such as cell structure, energy metabolism, mitochondrial function, GABA metabolism, signaling transduction and zinc fingers, as well as G-protein signaling, neurodevelopment, platelet activity, and thrombosis. The pivotal relationships have been validated in the previous study between the genetic methylation and the cause of SCZ [22–24]. Epigenetic mechanisms, especially DNA methylation, can mediate these interactions and may also trigger long-lasting adaptations in developmental programs that increase the risk of SCZ [25]. However, little is known about how methylation profile modulates the disease phenotype. These schizophrenic hypermethylated genes may have vital roles in the etiology of SCZ.

CORRESPONDING PATHWAYS AND SCHIZOPHRENIA

The pathway enrichment analysis may indicate the biological functions influenced by SDMGs. It could reveal the potential disease mechanism and novel therapeutic strategy for SCZ. There

are 29 corresponding pathways with FDR-adjust *P* value <.05 found in enrichment analysis from SDMGs, which may implicate the underlying disease mechanisms and characteristics for SCZ under the regulatory role of SDMGs. Top-ranked pathways with FDR *P* value <.05 are TGF_beta_receptor, pyrimidine metabolism, metabolic pathways, WNT pathway, folate biosynthesis, nicotinate and nicotinamide metabolism, and purine metabolism. The top-ranked pathways such as TNF alpha, PDGFR-beta signaling, TGF beta receptor, VEGFR1 and VEGFR2 signaling, regulation of telomerase, hepatocyte growth factor receptor signaling, ErbB1 downstream signaling, and mTOR signaling pathways, may be the key players in the symptoms of SCZ. Among these pathways, tumor necrosis factor alpha (TNF- α) is a cytokine product. Its primary role is the regulation of immune cells with biological functions of apoptotic cell death and inhibition of tumorigenesis and viral replication. Dysregulation of TNF- α production may cause negative symptoms of psychosis and SCZ [66, 67]. Platelet-derived growth factor receptors (PDGF-R) are cell surface tyrosine kinase receptors. Its subunits -A and -B are important factors, which regulate cell proliferation, cellular differentiation, cell growth, and neuronal development. The genes for platelet-derived growth factor beta (PDGFB) and PDGFB receptor (PDGFBR) may be important in the pathology of SCZ through interacting with the DRD2/DRD4 and NMDA receptors [68]. It should be noticed that PDGFBR mRNA transcripts are significantly increased in postmortem brains of schizophrenic patients [69].

SCHIZOPHRENIA AND EPIGENETIC REVIEW

The methylation status in promoters or gene-coding region, including *CYP3A4*, *CYP2D6*, *ABCB1*, *HTR2A*, and *DRD2*, was noted with risperidone drug response-related genes [26]. Transcription factors such as myocyte-specific enhancer factor 2C (MEF2C), or multiple regulators of the open chromatin mark, methyl-histone H3-lysine 4, are associated with the genetic risk architectures of common psychiatric disease and alterations in chromatin structure and function in diseased brain tissue [27]. Epigenetic mechanisms, especially DNA methylation, can mediate these interactions and may also trigger long-lasting adaptations in developmental programs that increase the risk of major depressive disorders (MDDs) and SCZ [25]. Only a small fraction of psychoses could be easily explained by genetics, but this screening in clinical practice is important as it can lead to therapeutic challenge or genetic counseling. Nowadays, it is clear that the pathophysiology of the psychoses can only be understood by an integrative approach taking into account the interaction between genes and the environment. Epigenome is stable but could be modified by environmental factors. Several epigenetic mechanisms have been studied in psychosis, in particular, the DNA methylation, the modification of histones, and the microRNA. All of these mechanisms are under regulation by genetic factors and variants in these epigenetic-involved genes and cofactors have been also associated with SCZ [28].

Metagenomic studies show that bioactive nutrients and gut microbiota can alter either DNA methylation or histone signatures through a variety of mechanisms. Much of this interplay may be moderated by epigenetic changes. Similar to genetic mutations, epigenetic modifications such as DNA methylation, histone modifications, and RNA interference can influence gene expression and therefore may cause behavioral and neuronal changes observed in mental disorders [29].

FINDINGS HIGHLIGHT THE SIGNIFICANCE OF ANTIPSYCHOTIC DRUGS ON DNA METHYLATION IN SCHIZOPHRENIA PATIENTS

All DNA methylation differences may not necessarily represent the cause of the disease; rather some may result from the effect of antipsychotics, medication, and stress tolerance. Patient-specific pathways affected by differential DNA methylation are responsible for the disease [30].

Epigenetic mechanisms are involved in the regulation of neural differentiation as well as in functional processes related to memory consolidation, learning, or cognition during healthy lifespan. On the other side of the coin, many neurodegenerative diseases are associated with epigenetic dysregulation. The reversible nature of epigenetic factors and, especially, their role as mediators between the genome and the environment make them exciting candidates as therapeutic targets [31].

DNA methylation of gene promoter regions represses transcription and is a mechanism via which environmental risk factors could affect cells during development in individuals at risk for SCZ. All cell types had distinct, statistically significant SCZ-associated differences in DNA methylation and linked gene expression, with Gene Ontology analysis showing that the differentially affected genes clustered in networks associated with cell growth, proliferation, and movement, functions known to be affected in SCZ patient-derived cells. Understanding the role of epigenetics in cell function of the brain in SCZ is likely to be complicated by similar cell type differences in intrinsic and environmentally induced epigenetic regulation [18].

Genome-wide association studies (GWAS) have remarkably advanced insight into the genetic basis of SCZ. Still, most of the functional variance in disease risk remains unexplained. Hence, there is a growing need to map genetic variability-to-genes-to-functions for understanding the pathophysiology of SCZ and the development of better treatments. Genetic variation can regulate various cellular functions including DNA methylation, an epigenetic mark with important roles in transcription and the mediation of environmental influences. Methylation quantitative trait loci (meQTLs) are derived by mapping levels of DNA methylation in genetically different, genotyped individuals and define loci at which DNA methylation is influenced by genetic variation. Recent evidence points to an abundance of meQTLs in brain tissues whose functional contributions to development and mental diseases are still poorly understood. Fetal meQTLs reside in regulatory domains affecting methylome reconfiguration during early brain development and are enriched in loci identified by GWAS for SCZ. Moreover, fetal meQTLs are preserved in the adult brain and could trace early epigenomic deregulation during vulnerable periods. Overall, these findings highlight the role of fetal meQTLs in the genetic risk for and in the possible neurodevelopmental origin of SCZ [32].

Early-life adversity is a major risk factor for MDD/SCZ and can trigger persistent genome-wide changes in DNA methylation at genes important to early, but also to mature, brain function, including neural proliferation, differentiation, and synaptic plasticity, among others. Moreover, genetic variations controlling dynamic DNA methylation in early life are thought to influence later epigenomic changes in SCZ. Epigenetic mechanisms, especially DNA methylation, can mediate these interactions and may also trigger long-lasting adaptations in developmental programs that increase the risk of MDDs and SCZ. Genetic variants influencing DNA methylation are also enriched in risk variants from GWAS on SCZ supporting a role in neurodevelopment. Overall, epigenomic responses to early-life adversity appear to be controlled to different degrees by genetics in MDD/SCZ, even though the

potential reversibility of epigenomic processes may offer new hope for timely therapeutic interventions in MDD/SCZ [25].

Genome-wide profiling efforts have given informative insights into biological processes; however, considering the wealth of variation, the major challenge still remains in their meaningful interpretation. In particular, sequence variation in noncoding contexts is often challenging to interpret. Here, data integration approaches for the identification of functional genetic variability represent a possible solution. Functional linkage analysis integrating genotype and expression data determined regulatory quantitative trait loci and proposed causal relationships. In addition to gene expression, epigenetic regulation, and specifically DNA methylation, was established as highly valuable surrogate mark for functional variance of the genetic code. Epigenetic modification has served as powerful mediator trait to elucidate mechanisms forming phenotypes in health and disease [33].

A molecular interaction network can be viewed as a network in which genes with related functions are connected. Therefore, at a systems level, connections between individual genes in a molecular interaction network can be used to infer the collective functional linkages between biologically meaningful gene sets. When an observed gene set is not enriched by known biological processes, traditional enrichment-based interpretation methods cannot produce functional insights, but GSLA can still evaluate whether those genes work in concert to regulate specific biological processes [34].

DNA methylation is an important epigenetic regulator of gene expression. Recent studies have revealed widespread associations between genetic variation and methylation levels. However, the mechanistic links between genetic variation and methylation remain unclear. SNPs that change predicted TF-binding affinities are significantly enriched for associations with DNA methylation at nearby CpGs [35].

REFERENCES

- [1] Akbarian S. Epigenetic mechanisms in schizophrenia. *Dialogues Clin Neurosci* 2014;16(3):405–17.
- [2] Misiak B, et al. Interactions between variation in candidate genes and environmental factors in the etiology of schizophrenia and bipolar disorder: a systematic review. *Mol Neurobiol* 2017.
- [3] Shorter KR, Miller BH. Epigenetic mechanisms in schizophrenia. *Prog Biophys Mol Biol* 2015;118(1–2):1–7.
- [4] Xin Y, et al. MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res* 2012;40(Database issue):D1245–9.
- [5] Zou D, et al. MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res* 2015;43(Database issue):D54–8.
- [6] Xiong Y, et al. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res* 2017;45(D1):D888–95.
- [7] Zhao Z, et al. PD_NGSAtlas: a reference database combining next-generation sequencing epigenomic and transcriptomic data for psychiatric disorders. *BMC Med Genom* 2014;7:71.
- [8] Wu Y, Yao YG, Luo XJ. SZDB: a database for schizophrenia genetic research. *Schizophr Bull* 2017;43(2):459–71.
- [9] Jia P, et al. SZGR 2.0: a one-stop shop of schizophrenia candidate genes. *Nucleic Acids Res* 2017;45(D1):D915–24.
- [10] Wang L, et al. SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *Neuroimage* 2016;124(Pt B):1155–67.
- [11] Pillich RT, et al. NDEx: a community resource for sharing and publishing of biological networks. *Meth Mol Biol* 2017;1558:271–301.

- [12] Mi H, et al. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 2016;44(D1):D336–42.
- [13] Mi H, et al. PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 2017;45(D1):D183–9.
- [14] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27–30.
- [15] Du J, et al. KEGG-PATH: kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst* 2014;10(9):2441–7.
- [16] Bartlett TE, Olhede SC, Zaikin A. A DNA methylation network interaction measure, and detection of network oncomarkers. *PLoS One* 2014;9(1):e84573.
- [17] Davies MN, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol* 2012;13(6):R43.
- [18] Vitale AM, et al. DNA methylation in schizophrenia in different patient-derived cell types. *NPJ Schizophr* 2017;3:6.
- [19] Auta J, et al. DNA-methylation gene network dysregulation in peripheral blood lymphocytes of schizophrenia patients. *Schizophr Res* 2013;150(1):312–8.
- [20] Su-Ping D, et al. Predicting schizophrenia by fusing networks from SNPs, DNA methylation and fMRI data. *Conf Proc IEEE Eng Med Biol Soc* 2016;2016:1447–50.
- [21] Ganapathiraju MK, et al. Schizophrenia interactome with 504 novel protein-protein interactions. *NPJ Schizophr* 2016;2:16012.
- [22] Ovenden ES, et al. DNA methylation and antipsychotic treatment mechanisms in schizophrenia: progress and future directions. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2017;81:38–49.
- [23] Rivollier F, et al. Epigenetics of schizophrenia: a review. *Encephale* 2014;40(5):380–6.
- [24] Teranova N, et al. DNA methylation in peripheral tissue of schizophrenia and bipolar disorder: a systematic review. *BMC Genet* 2016;17:27.
- [25] Hoffmann A, et al. Epigenomics of major depressive disorders and schizophrenia: early life decides. *Int J Mol Sci* 2017;18(8).
- [26] Shi Y, et al. Combined study of genetic and epigenetic biomarker risperidone treatment efficacy in Chinese Han schizophrenia patients. *Transl Psychiatry* 2017;7(7):e1170.
- [27] Javidfar B, et al. The epigenomics of schizophrenia, in the mouse. *Am J Med Genet B Neuropsychiatr Genet* 2017;174(6):631–40.
- [28] Chaumette B, Kebir O, Krebs MO. [Genetics and epigenetics of schizophrenia and other psychoses]. *Biol Aujourd’hui* 2017;211(1):69–82.
- [29] Alam R, Abdolmaleky HM, Zhou JR. Microbiome, inflammation, epigenetic alterations, and mental diseases. *Am J Med Genet B Neuropsychiatr Genet* 2017;174(6):651–60.
- [30] Melka MG, et al. Insights into the origin of DNA methylation differences between monozygotic twins discordant for schizophrenia. *J Mol Psychiatry* 2015;3(1):7.
- [31] Delgado-Morales R, et al. Epigenetic mechanisms during ageing and neurogenesis as novel therapeutic avenues in human brain disorders. *Clin Epigenet* 2017;9:67.
- [32] Hoffmann A, Ziller M, Spengler D. The future is the past: methylation QTLs in schizophrenia. *Genes (Basel)* 2016;7(12).
- [33] Heyn H. A symbiotic liaison between the genetic and epigenetic code. *Front Genet* 2014;5:113.
- [34] Zhou X, et al. Human interactome resource and gene set linkage analysis for the functional interpretation of biologically meaningful gene sets. *Bioinformatics* 2013;29(16):2024–31.
- [35] Banovich NE, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* 2014;10(9):e1004663.

This page intentionally left blank

EPIGENOME-WIDE DNA METHYLATION AND HISTONE MODIFICATION OF ALZHEIMER'S DISEASE

9

Ankush Bansal, Tiratha Raj Singh*Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, India*

BACKGROUND

The term *epigenetics* was coined in the 1940s by the scientist C. H. Waddington. Waddington presented epigenetics as a branch of science that deals with the relations between genes and resulting phenotype by considering in-between associations among them [1]. Such general ideas and thoughts were later presented as a branch of biology that understands the changes in genes without changes in the DNA sequences [2], although, such epigenetic factors are considered as an important attributes to present the phenotypic behavior changes transferred to the daughter cells [3]. Earlier, any such issues with localization did not contribute in neuronal complexity understanding as neurons do not isolate and always remain in connection. But, neuroscientists have considered the epigenetic perspective for recognition of sensory systems. Currently, epigenetics is focused on understanding the auxiliary changes of the chromosomes that control and alter phenotype without modifying the genetic outlook of the model [4].

The question that comes to mind is why are these neuroscientists so interested in epigenetic factors and enthusiastic to start new epigenetic studies? One of the main points is that epigenetic mechanisms and associated factors are the best suited for integration of variable dependable results from various different sources. Epigenetics covers the very fundamental basis of neuroscience, giving a potential substrate to memory designation, and for articulating the theory of quality and condition communication related with numerous multifactorial infections, for example, Parkinson's and Alzheimer's diseases (AD), amyotrophic lateral and various scleroses, epilepsy [5]. Truly, it is realized that the epigenetic mechanism is vital in the procedures of learning and memory development [6–12], maturing [13], nourishment [14], and environmental factors [15] related with AD and can alter the epigenetic cosmetics and may in this manner add to the pathophysiology of AD [16–18].

EPIGENETICS ASSOCIATION WITH THE NERVOUS SYSTEM

The nervous system is a very precise system in which a huge number of cells are collected in various arrangements with epigenetic and expression profiles that are related with specific functions [19–21].

There are various specifications of epigenetics with respect to the nervous system, which includes control of different cellular mechanisms, e.g., three out of four genes exchange information [22], most spliced variants are identified [23–25], and most miRNAs are utilized [26–27]. There is no doubt that it is the nervous system where the patterns instead of number determine the most noteworthy level of heterogeneity, with approximately 70% of genes being expressed in under 20% of the cells of the whole brain [28]. Factors that contribute to this intricacy include the transcription machinery, which faces considerable challenges in the nervous system and is very sensitive to epigenetic perturbations.

It has been noticed that the significance of epigenetics in the functioning of the nervous system is underscored by the way that changes in epigenetic genes may cause extreme mental distortion [29,30]. Changes in the genes that act as epigenetic markers, for example, NSD1, NSD2, DNMT1, or CBP, may cause hereditary sensory autonomic neuropathy with dementia (HSAN1), Sotos, Wolf-Hirschhorn, and Rubinstein-Taybi syndromes, separately. Therefore, changes in genes that expel epigenetic marks, for KDM5C and MeCP2, are responsible for integration, SWI/SNF proteins are related with X-connected mental hindrance, Rett syndrome and Coffin-Siris syndrome, individually [5,29,31,32].

In such cases, one of the most important discoveries highlighting the significance of epigenetics in the working of the cerebrum has revealed that neuronal action alters DNA methylation and histone modifications. Further, these processes affect learning and memory, which rely upon these epigenetic changes [12,33–36]. For example, neuronal action depends on DNMT3A2, TET1, and TET3 [34,37–39]. As well, a few HDAC inhibitors, for example, valproic corrosive, sodium butyrate, and others, potentiate learning and memory development in various standards and animals models [7], and also in various neurological infections, for example, Alzheimer's, Parkinson's, and Huntington's diseases [40]. In this manner, it is clear that neuronal action, and in addition, learning and memory, draw on and to some degree rely upon various epigenetic players, and that epigenetic changes disable the brain's ordinary working as well as are related with numerous neurological illnesses, including AD.

EPIGENETIC MECHANISMS IN AD

At the submolecular level, it is well known that epigenetics involves two primary mechanisms: the immediate methylation of the DNA, and the change of the proteins that bundle the DNA, called as histones. Chromatin remodelers and noncoding RNAs can take part in the direction of the chromatin, yet, they are not absolutely considered epigenetic mechanisms. Here, we initially portray the functioning of these two epigenetic changes as a rule, before moving to their participation in neuroscience, and specifically, in AD.

EPIGENETIC CHANGES IN AD

AD is the fundamental driver of dementia in the Western world, where it affects 17% of individuals older than 65 and half of those older than 85 [41]. It is realized that hereditary and nonhereditary elements influence AD. Uncommon transformations in three genes—APP, PSEN1, and PSEN2—are related to 1% of AD [42] and other continuous hereditary variations, for example, APOE-E4 can represent up to 20% of aggregate instances of the disease [43]. Altogether, the heritability for AD is assessed to account for between one-half and 66% of aggregate AD cases [44], the other one-third/half

being attributed to nonhereditary hazard factors in which epigenetic mechanisms are concerned connected to diabetes mellitus, hypertension, stoutness, physical dormancy, melancholy, smoking, and low instructive fulfillment [45].

A major risk factor for AD is aging itself, since AD just shows up in late adulthood, and the increasing risk of the infection doubles after age 65 [46]. Significantly, epigenetic components have likewise been recommended to be a noteworthy power of aging [13,47,48] and comparable epigenetic modifications have additionally been portrayed in AD [16–18]. In any case, before illustrating the proof of an epigenetic suggestion in AD, we specify that this inquiry has been undertaken from various specialized points of view, and that the acquired outcomes clearly depend on the trial approaches and the procedures utilized. Some of these examinations depended on cell lines, others on animal models, and many of them on human postmortem tissue. Similarly, various procedures have been utilized for determining the levels of DNA methylation, which include the utilization of DNA methylation delicate confinement proteins, antibodies that particularly perceive DNA methylation alterations, as well as DNA methylation by bisulfite sequencing [16–18].

EPIGENETIC MODIFICATIONS

DNA METHYLATION

DNA methylation has hitherto been the most considered epigenetic alteration. It principally comprises of the expansion of a methyl group at cytosines that go before guanines (alleged CpG dinucleotides). These dinucleotides are underrepresented in the genome and have a tendency to gather in CpG-thick areas (alleged CpG islands, or CGI) in spite of the fact that around 95% of CpGs are scattered through all the genome without demonstrating any sort of accumulation. In general, CpGs in non-CGI and CGIs have a tendency to be completely methylated and nonmethylated, separately, with an irrelevant measure of CpGs being incompletely methylated [49,50].

For many decades, DNA methylation has been viewed as an epigenetic characteristic of suppression, since fundamental analyses have demonstrated that the genomic inclusion of exogenous DNA brings about dynamic translation just with nonmethylated DNA [51–53], and since CGI hypermethylation has been intermittently connected with the hushing of tissue-particular qualities and X inactivation [54]. In fact, the presence of CGI together with the bimodal example of DNA methylation has evoked qualities that can be turned on and off by controlling the DNA methylation of their CGIs. Nonetheless, it is currently evident that actually around 70% of annotated genes contain CGI areas in their promoters [55] and the greater part of them are nonmethylated. CGIs favor the entrance of the DNA polymerase and gene expression by constituting an inflexible structure that confuses the wrapping of DNA and nucleosome situating [56]. In this manner, CGIs are not only platforms for controlling gene expression by DNA methylation but their impact relies upon the nearby sequence and three-dimensional chromatin structure.

Critically, DNA isn't arbitrarily distributed in the nucleus but is related to histones framing the nucleosomes [57]. The dispersion and the compaction of these nucleosomes decides the chromatin structure, and in this way the entrance of the transcriptional machinery to the DNA [57]. Generally, DNA methylation in gene promoters is related to bringing down levels of expression [58], whereas in gene bodies it favors gene expression [59].

HYPOMETHYLATION IN AD

In general, the utilization of cell lines is independent of the system utilized, which proposes that AD is related to low levels of DNA methylation. For example, the glioblastoma cell line H4 harboring the Swedish transformation of APP (K670M/N671L two-fold change isolating in a Swedish family), which causes an expansion in A β production [60–62]. Study has demonstrated a general inclination toward hypomethylation as measured by DNA microarrays following bisulfite transformation [63]. Therefore, treatment of the neuronal-like cell line SH-SY5Y with molded media acquired from cells harboring the Indiana change (V717F transformation recognized by a collection of Indiana), related with higher A β levels, initiated general DNA hypomethylation as measured by DNA methylation-sensitive antibodies [64]. In general, brain microvascular endothelial cells subjected to production of large amounts of A β indicated low levels of DNA methylation as measured by high-performance liquid chromatography [65]. In comparison, IMR-32 neuroblastoma cells subjected to elevated amounts of synthetic A β do not indicate huge modifications in DNA methylation as measured by DNA microarrays [66].

The investigation of humans postmortem has not tackled this obvious inconsistency. Use of the antibodies that perceive methylated DNA, lost DNA methylation has been seen in the entorhinal cortex [67] and the hippocampus of posthumous samples of AD [47]. As well, different examinations by using a similar strategy have found no distinctions in the entorhinal cortex [68] or even pickup of DNA methylation in the frontal cortex [69]. In a similar way, ELISA 5-methylcytosine (5mC) tests of the entorhinal cortex of AD patients [68] and, additionally, DNA methylation microarrays in frontal cortex [70], have not indicated noteworthy DNA methylation contrasts.

HYDROXYMETHYLATION IN AD

A comparably confounding situation additionally appears from investigation of the DNA hydroxymethylation in AD. Larger amounts of DNA hydroxymethylation have been observed in 3xTg-AD mice that harbor APPSwedish, PSEN1 M146L, and P301L. It has been found that TAU changes affect A β arrangements, TAU phosphorylation and brings down its level in the human frontal, entorhinal, and cerebral cortex [69,71]. In addition, no critical contrasts were seen in entorhinal cortex utilizing 5-hydroxymethylcytosine (5hmC) particular ELISA measures [68].

It is important to specify that, in these examinations, the amount of DNA methylation and hydroxymethylation changes, and the quantity of tests performed are generally small, and the outcomes can undoubtedly be affected by contrasts in the investigated areas [21,32,72–75]. Therefore, if DNA methylation and hydroxymethylation contrasts are available in AD, these are probably few or to be related areas of the genome. The investigation of infection dissonant twins has been urgent for disentangling the epigenetic segment of regular maladies [76]. Shockingly, just a solitary couple of monozygotic twins conflicting for AD have been examined up until this point.

GENE-WISE DNA METHYLATION CHANGES IN AD

Attempts to decide if particular hereditary locales or specific genes are modified in AD have at first centered on genes already connected with the disease—APP, PSEN1, and TAU—and no definitive confirmations have been found in these examinations.

Regardless of a few reports proposing a hypomethylation in the promoter of APP in the worldly cortex of AD [77] and aging [78], examinations utilizing higher sample numbers have not possessed the capacity to discover contrasts in frontal cortex, parietal cortex, and hippocampus of AD patients [79]. Comparable discoveries have been acquired in SK-N-BE neuroblastoma cell line utilizing vitamin B₆ and B₁₂ inadequate media [80]. PSEN1 hypomethylation has not been seen in frontal cortex and hippocampus of AD tests [81,82]. Lastly, no critical contrasts in DNA methylation in the frontal cortex or the hippocampus of postmortem AD tests have been seen in the promoter of TAU [82]. In this manner, it appears that these three established AD-related qualities are not epigenetically dysregulated in AD at the DNA methylation level, which may demonstrate that DNA methylation changes do not assume a part in AD, or that hereditary and nonhereditary types of AD may be the aftereffects of adjustments in an alternate subset of qualities. As an outcome, unprejudiced genome-wide screening is, additionally, beginning to be performed.

GENOME-WIDE DNA METHYLATION ALTERATIONS IN AD

Unfortunately, genome-wide case-control examinations have not been more convincing, with practically each and every investigation detailing an alternate subset of changed qualities that may reflect that present methodologies are as yet juvenile. Accordingly, the mix of all of the inclusive techniques with longitudinal investigations of AD patients and mouse models yields more steady information. Two unique qualities that have been accounted for hypermethylated are free gatherings of Sorbin and SH3 Domain Containing 3 (SORBS3) [83] and Ankyrin 1 [72,84].

DNA REPAIR AND METHYLATION IN AD

DNA repair outcomes were acquired during a time subordinate DNA methylation examination and an all-inclusive DNA methylation screening in two diverse AD mouse models—APPswe/dE9 and 3xTg-AD—and later found in the frontal cortex of humans in postmortem AD tests, and additionally, from two broad DNA methylation screenings in a few human cerebrum areas in individual differentially AD-influenced tests. By all accounts, the hypermethylation of the quality insulin-like growth factor binding protein 7 (IGFBP7), which is maintained by predictable changes in DNA methylation in the APPPS1-21 AD mouse, is shown to harbor the Swedish APP transformation in blend with the L166P PSEN1 transformation, and in human frontal cortex tests [85]. Finally, the hypermethylation of dual specificity phosphatase 22 (DUSP22), also to ANK1, corresponds with the seriousness of the malady and was exhibited alter TAU phosphorylation and cell suitability in vitro [83].

Regardless, it must be noticed that these connections do not really reflect a causal connection with the disease, and may even be the result of auxiliary adjustments. This is especially essential for the ones seen in mouse models, since these models have as of now a hereditary inclination to grow AD pathology. Additionally, another restriction of these investigations is that they didn't recognize diverse cells that are as of now modified in AD (being a neurodegenerative disease related with a noticeable gliosis and a particular loss of neurons; [86]), and that exhibit unmistakable epigenetic profiles. In this way, these outcomes ought to be considered since they will be required for approval of other investigations.

HISTONE MODIFICATIONS

HISTONE ACETYLATION CHANGES IN AD

In spite of DNA methylation, histone changes have been less concentrated in AD, and confirmations connecting histone adjustment modifications with AD are mostly aberrant. The few investigations that have discovered that few histone deacetylase (HDAC) inhibitors apply a defensive impact in AD, enhancing dendritic spine thickness, and encouraging learning and memory development in various mouse models of the disease, in spite of the fact that the exact systems by which the HDAC inhibitors work need to be resolved. Besides, HDAC2 was observed to increase with age in mice and people [47], in APP/PS1 [87], p25/Cdk5—harboring the Cdk5 activator p25 transgene that initiates TAU phosphorylation and neurodegeneration [88]—and 5xFAD AD mouse models—harboring the Swedish, I171V Florida, and V717I London APP mutations in mix with the M146L and L286L PSEN1 changes with incited A β arrangement and neurodegeneration [89]—and also in the hippocampus and entorhinal cortex of posthumous human AD tests. It has also been demonstrated that HDAC2 can differentially tie and direct the declaration of a few learning and neuroplasticity-related qualities, yet its viral-intervened exhaustion or its particular pharmacological restraint is adequate for reestablishing the synaptic and subjective shortages seen in p25/Cdk5 mice [33]; Wagner et al., 2015. Accordingly, there is convincing confirmation that HDAC2 is expanded in aging and AD, and most likely involved in the related psychological decay, in spite of the fact that it ought to be specified that an abatement of HDAC2 in AD patients has been additionally found by other research [67].

Despite these confirmations, it is as yet not clear whether basal histone acetylation is adjusted in AD. One conceivable clarification may be that rather than a modification of the basal levels of histone acetylation, AD may be more related to the inadequacy of adjusting the epigenetic designs in specific conditions, for example, learning and memory arrangement, in which HDAC inhibitors that increase in histone acetylation would “prime” the levels of histone acetylation and, subsequently, of quality movement [33,90]. To support this view, the basal levels of H4K12ac in maturing stage remains consistent; however, when mice are subjected to learning and memory ideal models as youthful creatures they can expand these levels but matured mice cannot [91]. In Tg2576 AD mice—harboring the Swedish APP transformation in blend with the M146V PSEN1 alternations, which brings about more elevated amounts of A β production [92]—the global levels of H4 acetylation are not modified. On the other hand, in spite of the facts of past theory, it could be conceivable that histone acetylation changes happen just in specific loci, which could be more affected by HDAC inhibitors, without reflecting general propensities in the mass chromatin. To better understand these situations, genome-wide screenings of histone adjustments are beginning to be embraced and will provide better insights in the near future.

GENE-WISE HISTONE ALTERATIONS IN AD

The likelihood that in AD particular genes may be posttranslationally changed on their histones has recently begun to be addressed, and to our knowledge, just two investigations in the p25/Cdk5 AD mouse demonstrate this point. In 2012, a few neuroplasticity-related qualities were accounted for as hypoacetylated and quelled in p25/Cdk5 mice [93] and, as of late, the index of deregulated qualities and posttranslational changes has been largely amplified [94]. An interesting finding of this

examination was further reliable improvement of dynamic imprints (H3K27ac and H3K4me3) in enhancers and promoters of safe and jolt reaction capacities combined with a particular decline in neural connection and learning-related capacities [95]. Similar to reported DNA methylation alterations in AD, these outcomes likely reflect the two changes in cell organization and cell-sort particular changes related with AD pathology, in this manner requiring cell-particular approvals for a superior assessment of their criticalness in AD.

EPIGENOMICS

MOLECULAR MECHANISMS LINKING GENOMIC RISK FACTORS TO AD

Mutations in APP, PSEN1, and PSEN2 assumed a critical part in propelling our comprehension of AD. APP is a transmembrane protein that, when cleaved, yields the amyloid- β [76] peptide that speaks to a basic pathologic sign of AD. APP cleavage is accomplished by an arrangement of three secretases, α , β , and γ [61]. β -secretase is the protein BACE-1. Cleavage with this secretase is essential for creating the amyloid- β peptide. α -Secretase cuts inside the amyloid- β peptide succession keeping the age of the amyloid- β protein [61]. By differentiating, γ -secretase severs somewhere else and leaves the amyloid- β section in place, permitting the age of the amyloid- β peptide. Presenilin-1 and -2 are the two segments of γ -secretase and prompt the improvement of AD.

There is broad confirmation that APOE is associated with amyloid probably by means of freedom of amyloid- β [33,61]. It is likely, be that as it may, that at least two components are included. Notwithstanding the way that APOE ϵ 4 is common enough to make it a down-to-earth genotype-particular focus for tranquilize improvement and over many years of research, APOE ϵ 4 remains an exploration apparatus and has exceptionally restricted clinical utility [33]. Identifying and approving helpful focuses in the pathway(s) connecting these genomic variations to AD is among the most pressing issues in the field. The connection of the other hereditary hazard variables to great AD pathology has recently been the subject of examination. Methodologies are incorporating relationships with neuropathologic attributes at dissection or with imaging or biofluid biomarkers. The information to date proposes that a few but not the majority of the newly found genomic variations are identified with much AD pathology, for example, amyloid and that the impact sizes have a tendency to be small, as one may see with the quality of relationship in the GWAS. The outcomes are variable, however, incorporating CR1, ABCA7, CD2AP, CD33, PICALM, and SORL1 [61].

POLYMORPHISMS AND AD

Linkage to chromosome 19 brought about recognizable proof of two polymorphisms in the apolipoprotein E gene related with AD: increased risk of AD, and the APOE ϵ 2 haplotype is related with diminished risk [95]. There are currently nine loci containing vulnerability alleles that are viewed as affirmed by various diverse studies, including ABCA7, BIN1, CD33, CLU, CR1, CD2AP, EPHA1, MS4A6A-MS4A4E, and PICALM. The latest GWAS with almost 75,000 subjects recognized these, and additionally, SORL1 (which had beforehand been recommended to be a defenselessness locus), and 11 novel loci, including HLA-DRB5/HLA-DRB1, PTK2B, SLC24A4-RIN3, DSG2, MEF2C, INPP5D, NME8, ZCWPW1, CELF1, FERMT2, and CASS4. These biomarkers can be probable targets to understand the AD pathology, but GWAS studies only give an idea to trace down a few

biomarkers on the basis of expression analysis. There is a need for specific tissue or patient studies to provide more comprehensive information at the epigenetic level.

SYSTEMS LEVEL MODULES FOR AD

For investigation of the neurodiseases (NDs) from the network viewpoint, firstly, the interaction networks of disease proteins were modeled [96]. The expressed proteins of disease genes are the disease proteins that are linked to particular NDs, the list of disease genes related to the top neuro-diseases were obtained from the morbid map published in the Online Mendelian Inheritance in Man database. Sequentially the construction of the protein interaction network related to the top NDs was done and then the disease genes to disease proteins were mapped based on the mapping design of the UniProt, which is a database of protein sequences and their functional information [97]. The associations of those disease proteins were then derived by validation of the experimental interactions from the Interologous Interaction Database (i2d) database [98]. The homologous predicted protein interactions in the i2d database were expelled to amplify the dependability of protein interaction data. The ultimate interaction network of interest enclosed the disease proteins (nodes) and their direct interacting partners (edges). The network design was not directed and weighted as only the binary interactions were considered [99].

On the basis of the disease proteins network, the network of NDs with meta-nodes and meta-edges were modeled that represented the diseases and the relations among them, in the same way. A meta-node was understood as a bunch of the disease proteins associated to one disease [91]. A meta-edge linking one disease to the other was defined as a set of the paths connecting their disease proteins. The meta-edges were weighted by means of special score functions ri to reveal the potency of the connection involving two diseases, like the two meta-nodes, called as amyotrophic lateral sclerosis (ALS) and spinal muscular atrophy (SMA), involved all the proteins associated to ALS and SMA, correspondingly [37]. Every pair of proteins (one from ALS and one from SMA) was grouped by computing the paths connecting them together to recognize a meta-edge, called ALS-SMA, along with weight calculated by ri . The meta-edges are only visible if their score $ri > 0$. All these studies provide an opportunity to look at the coassociation between shared genes for many polygenic NDs.

A system-level transcriptional regulation and control model can help in identifying epigenetic factors involved in AD signaling. Weighted gene coexpression module-based analysis presents a robust method to determine key genes controlling functionality and modularity. Comparison on the basis of conserved patterns in AD and aging may result in understanding of associated complex of synaptic plasticity. DNA methylation and histone modification models are available for various types of cancer on the basis of cell line optimization, and the same models can be used for NDs by concentration-based optimization (Table 9.1; Scheme 9.1).

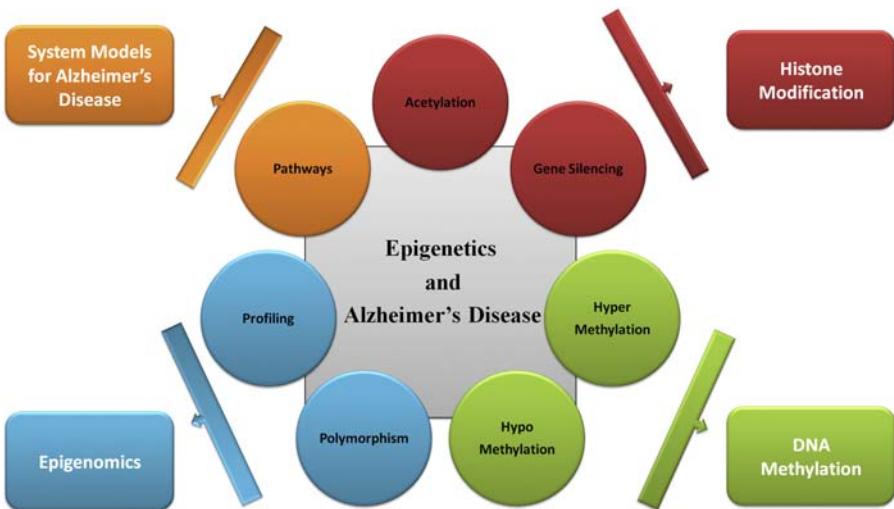
Table 9.1 Experimental Approaches for DNA Methylation

S.No.	Epigenetic Changes	Experimental/Tools	Strength	Weakness
1.	DNA Methylation	CHARM	Cost-effective, investigates CpG sites genome-wide without depending on proximity to genes or CpG islands	Resolution is moderate with selected regions in proximity to enzymes' recognition sites
2.		TAB-seq	This approach can distinguish 5hmC from 5mC	Substantial DNA degradation after bisulfite treatment and Tet enzyme with low efficiency might leave methylated residues unconverted. Also, high sequencing depth is required to detect 5hmC with low abundance
3.		scWGBS	Able to study methylome intrapopulation distribution	Low sequencing efficiency (~20 million reads typically required per cell). Cannot discriminate between 5mC and 5hmC
4.		WGBS	Evaluate methylation state of almost every CpG site	High cost. Substantial DNA degradation after bisulfite treatment. Cannot discriminate between 5mC and 5hmC
5.		Illumina's Infinium Methylation assay	Cost-effective. Does not require a large amount of input DNA	Human sample only. Coverage is highly dependent on the array design. Substantial DNA degradation after bisulfite treatment
6.		MeDIP	Cost-effective. No mutation introduced. Specific to 5mC/5hmC depending on the antibody specificity. More sensitive in regions with low CpG density than MBDCap-Seq	Biased toward hypermethylated regions. Does not identify individual 5mC sites. Inability to predict absolute methylation level

Continued

Table 9.1 Experimental Approaches for DNA Methylation—cont'd

S.No.	Epigenetic Changes	Experimental/Tools	Strength	Weakness
7.		MBDCap-Seq	Cost-effective. Allows the detection of DMRs within highly CpG-dense regions and regions with lower CpG density. MBD proteins can discriminate 5mC from 5hmC. No mutation introduced. More sensitive than MeDIP in regions with higher CpG density	Relatively low resolution. Biased toward hypermethylated regions
8.		RRBS	High CGI coverage. High sensitivity. Cost-effective compared to WGBS	May exhibit a lack of coverage at intergenic and distal regulatory elements. Substantial DNA degradation after bisulfite treatment. Limited to regions in proximity to enzymes' recognition sites. Cannot discriminate between 5mC and 5hmC
9.		scRRBS	Sensitivity is high and it can identify CpG targets. Can detect target CpG sites at high coverage with relatively low number of sequence reads	Substantial DNA degradation after bisulfite treatment. Cannot discriminate between 5mC and 5hmC. Provides relatively poor coverage for imprinting loci
10.	Histone Modification	Histone HMM	Histone HMM can be used for differential analysis of histone modifications with broad genomic footprints.	Approximate solution
11.		Chrom HMM	ChromHMM is based on a multivariate Hidden Markov Model that explicitly models the presence or absence of each chromatin mark.	Does not provide exact solution
12.		ChromaSig	Uncovers novel, functionally significant genomic elements	Probabilistic model-based solution

**SCHEME 9.1**

Computational tools for epigenetic analysis.

FUTURE DIRECTIONS

Technological enhancements will soon permit better cross-examination of the epigenome by expanding the extent of the genome that can be inspected viably in a high-throughput way. This will empower the execution of concentrates on the substantial scale that is required for illness-related examinations. This data will likewise be utilized to understand the subatomic systems that drive AD pathology—mapping the epigenomic highlights that identify with AD recognizes districts of the genome in which transcriptional potential is modified in illness. To survey whether these progressions originate before early confirmation of Alzheimer pathology (for example, amyloid pathology), we have to assess their connection to other hazard factors, for example, hereditary and experiential elements. Be that as it may, early examinations propose that hereditary variables seem to have to great extent autonomous impacts in connection to epigenomic factors.

The methylation information and reference chromatin framework from various cortical and subcortical tissues are a magnificent beginning assessment of epigenetic changes in various areas of the aging brain. In coming years, further examinations utilizing better advancements will improve the information that we have to see how chromatin structure impacts the part of weakness variations and whether it might intervene in a portion of the experiential hazard factors related with AD powerlessness. DNA methylation is probably going to remain the sign of decision, yet the investigation of histone protein alterations will be important to comprehend the fine engineering of loci involved in AD. This will be supplemented by ChIP-Seq questions focusing on correlated translation factors that direct quality articulation inside helplessness loci. These outcomes will impact the identification and refinement of the atomic systems that prompt AD and are affected by

AD pathology. In spite of its many difficulties, epigenomic investigations of AD will assume a vital part in the exploration group's endeavors to depict the arrangement of occasions driving from well-being to dementia.

REFERENCES

- [1] Waddington CH. The epigenotype. 1942. *Int. J. Epidemiol.* 2012;41:10–3. <https://doi.org/10.1093/ije/dyr184>.
- [2] Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. *Genes Dev* 2009;23:781–3. <https://doi.org/10.1101/gad.1787609>.
- [3] Bansal A, Ramana J. TCGDB: A Compendium of Molecular Signatures of Thyroid Cancer and Disorders. *J. Cancer Sci. Ther.* 2015;7. <https://doi.org/10.4172/1948-5956.1000350>.
- [4] Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 2003;33(Suppl):245–54. <https://doi.org/10.1038/ng1089>.
- [5] Urdinguio RG, Sanchez-Mut JV, Esteller M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *Lancet Neurol* 2009;8:1056–72. [https://doi.org/10.1016/S1474-4422\(09\)70262-5](https://doi.org/10.1016/S1474-4422(09)70262-5).
- [6] Fraga MF, Ballestar E, Paz MF, . Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu Y-Z, Plass C, Esteller M. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A* 2005;102:10604–9. <https://doi.org/10.1073/pnas.0500398102>.
- [7] Gräff J, Tsai L-H. Histone acetylation: molecular mnemonics on the chromatin. *Nat. Rev. Neurosci* 2013;14: 97–111. <https://doi.org/10.1038/nrn3427>.
- [8] Guzman-Karlsson MC, Meadows JP, Gavin CF, Hablitz JJ, Sweatt JD. Transcriptional and epigenetic regulation of Hebbian and non-Hebbian plasticity. *Neuropharmacology* 2014;80:3–17. <https://doi.org/10.1016/j.neuropharm.2014.01.001>.
- [9] Jarome TJ, Thomas JS, Lubin FD. The epigenetic basis of memory formation and storage. *Prog. Mol. Biol. Transl. Sci.* 2014;128:1–27. <https://doi.org/10.1016/B978-0-12-800977-2.00001-2>.
- [10] Levenson JM, Sweatt JD. Epigenetic mechanisms in memory formation. *Nat. Rev. Neurosci.* 2005;6: 108–18. <https://doi.org/10.1038/nrn1604>.
- [11] Woldemichael BT, Bohacek J, Gapp K, Mansuy IM. Epigenetics of memory and plasticity. *Prog. Mol. Biol. Transl. Sci.* 2014;122:305–40. <https://doi.org/10.1016/B978-0-12-420170-5.00011-8>.
- [12] Zovkic IB, Guzman-Karlsson MC, Sweatt JD. Epigenetic regulation of memory formation and maintenance. *Learn. Mem.* Cold Spring Harb. N. 2013;20:61–74. <https://doi.org/10.1101/lm.026575.112>.
- [13] Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, Puca AA, Sayols S, Pujana MA, Serra-Musach J, Iglesias-Platas I, Formiga F, Fernandez AF, Fraga MF, Heath SC, Valencia A, Gut IG, Wang J, Esteller M. Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U. S. A.* 2012;109:10522–7. <https://doi.org/10.1073/pnas.1120658109>.
- [14] Cooney CA, Dave AA, Wolff GL. Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *J. Nutr.* 2002;132:2393S–400S.
- [15] Anway MD, Cupp AS, Uzumcu M, Skinner MK. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* 2005;308:1466–9. <https://doi.org/10.1126/science.1108190>.
- [16] Bennett DA, Yu L, Yang J, Srivastava GP, Aubin C, De Jager PL. Epigenomics of Alzheimer's disease. *Transl. Res. J. Lab. Clin. Med.* 2015;165:200–20. <https://doi.org/10.1016/j.trsl.2014.05.006>.
- [17] Cacabelos R, Torrellas C. Epigenetic drug discovery for Alzheimer's disease. *Expert Opin. Drug Discov.* 2014;9:1059–86. <https://doi.org/10.1517/17460441.2014.930124>.
- [18] Coppedè F. The potential of epigenetic therapies in neurodegenerative diseases. *Front. Genet.* 2014;5:220. <https://doi.org/10.3389/fgene.2014.00220>.

- [19] Ko Y, Ament SA, Eddy JA, Caballero J, Earls JC, Hood L, Price ND. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proc. Natl. Acad. Sci. U. S. A.* 2013;110: 3095–100. <https://doi.org/10.1073/pnas.1222897110>.
- [20] Sanchez-Mut JV, Aso E, Panayotis N, Lott I, Dierssen M, Rabano A, Urdinguio RG, Fernandez AF, Astudillo A, Martin-Subero JI, Balint B, Fraga MF, Gomez A, Gurnot C, Roux J-C, Avila J, Hensch TK, Ferrer I, Esteller M. DNA methylation map of mouse and human brain identifies target genes in Alzheimer's disease. *Brain J. Neurol.* 2013;136:3018–27. <https://doi.org/10.1093/brain/awt237>.
- [21] Xin Y, Chanrion B, Liu M-M, Galfalvy H, Costa R, Ilievski B, Rosoklija G, Arango V, Dwork AJ, Mann JJ, Tycko B, Haghghi F. Genome-wide divergence of DNA methylation marks in cerebral and cerebellar cortices. *PLoS One* 2010;5:e11357. <https://doi.org/10.1371/journal.pone.0011357>.
- [22] Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanović D, Geschwind DH, Mane SM, State MW, Sestan N. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 2009;62:494–509. <https://doi.org/10.1016/j.neuron.2009.03.027>.
- [23] Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ. An alternative-exon database and its statistical analysis. *DNA Cell Biol* 2000;19:739–56. <https://doi.org/10.1089/104454900750058107>.
- [24] Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* 2002;30:3754–66.
- [25] Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biol* 2004;5:R74. <https://doi.org/10.1186/gb-2004-5-10-r74>.
- [26] Bansal A, Singh TR, Chauhan RS. A novel miRNA analysis framework to analyze differential biological networks. *Sci. Rep.* 2017;7:14604. <https://doi.org/10.1038/s41598-017-14973-x>.
- [27] Cai C, Lin P, Cheung K-H, Li N, Levchook C, Pan Z, Ferrante C, Boulianee GL, Foskett JK, Danielpour D, Ma J. The presenilin-2 loop peptide perturbs intracellular Ca²⁺ homeostasis and accelerates apoptosis. *J. Biol. Chem.* 2006;281:16649–55. <https://doi.org/10.1074/jbc.M512026200>.
- [28] Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, Chen L, Chen L, Chen T-M, Chin MC, Chong J, Crook BE, Czaplinska A, Dang CN, Datta S, Dee NR, Desaki AL, Desta T, Diep E, Dolbeare TA, Donelan MJ, Dong H-W, Dougherty JG, Duncan BJ, Ebbert AJ, Eichele G, Estin LK, Faber C, Facer BA, Fields R, Fischer SR, Fliss TP, Frenzley C, Gates SN, Glattfelder KJ, Halverson KR, Hart MR, Hohmann JG, Howell MP, Jeung DP, Johnson RA, Karr PT, Kawal R, Kidney JM, Knapik RH, Kuan CL, Lake JH, Laramee AR, Larsen KD, Lau C, Lemon TA, Liang AJ, Liu Y, Luong LT, Michaels J, Morgan JJ, Morgan RJ, Mortrud MT, Mosqueda NF, Ng LL, Ng R, Orta GJ, Overly CC, Pak TH, Parry SE, Pathak SD, Pearson OC, Puchalski RB, Riley ZL, Rockett HR, Rowland SA, Royall JJ, Ruiz MJ, Sarno NR, Schaffnit K, Shapovalova NV, Sivisay T, Slaughterbeck CR, Smith SC, Smith KA, Smith BI, Sodt AJ, Stewart NN, Stumpf K-R, Sunkin SM, Sutram M, Tam A, Teemer CD, Thaller C, Thompson CL, Varnam LR, Visel A, Whitlock RM, Wohnoutka PE, Wolkey CK, Wong VY, Wood M, Yaylaoglu MB, Young RC, Youngstrom BL, Yuan XF, Zhang B, Zwingman TA, Jones AR. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 2007;445:168–76. <https://doi.org/10.1038/nature05453>.
- [29] Berdasco M, Esteller M. Genetic syndromes caused by mutations in epigenetic genes. *Hum. Genet.* 2013; 132:359–83. <https://doi.org/10.1007/s00439-013-1271-x>.
- [30] Kumar A, Bansal A. Integrated bioinformatics analysis of differentially expressed genes (degs) of alzheimer's disease (ad) datasets from gene expression omnibus (GEO). *Alzheimers Dement. J. Alzheimers Assoc.* 2017;13:P953. <https://doi.org/10.1016/j.jalz.2017.06.1270>.
- [31] Gasser SM, Li E. Epigenetics and disease: pharmaceutical opportunities. Preface. *Prog. Drug Res. Fortschritte Arzneimittelforschung Progres Rech. Pharm.* 2011;67:v–viii.

- [32] Sanchez-Mut JV, Huertas D, Esteller M. Aberrant epigenetic landscape in intellectual disability. *Prog. Brain Res.* 2012;197:53–71. <https://doi.org/10.1016/B978-0-444-54299-1.00004-2>.
- [33] Gräff J, Rei D, Guan J-S, Wang W-Y, Seo J, Hennig KM, Nieland TJF, Fass DM, Kao PF, Kahn M, Su SC, Samiei A, Joseph N, Haggarty SJ, Delalle I, Tsai L-H. An epigenetic blockade of cognitive functions in the neurodegenerating brain. *Nature* 2012;483:222–6. <https://doi.org/10.1038/nature10849>.
- [34] Guo JU, Su Y, Zhong C, Ming G, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* 2011;145:423–34. <https://doi.org/10.1016/j.cell.2011.03.022>.
- [35] Levenson JM, O’Riordan KJ, Brown KD, Trinh MA, Molfese DL, Sweatt JD. Regulation of histone acetylation during memory formation in the hippocampus. *J. Biol. Chem.* 2004;279:40545–59. <https://doi.org/10.1074/jbc.M402229200>.
- [36] Miller CA, Sweatt JD. Covalent modification of DNA regulates memory formation. *Neuron* 2007;53:857–69. <https://doi.org/10.1016/j.neuron.2007.02.022>.
- [37] Li X, Wei W, Zhao Q-Y, Widagdo J, Baker-Andresen D, Flavell CR, D’Alessio A, Zhang Y, Bredy TW. Neocortical Tet3-mediated accumulation of 5-hydroxymethylcytosine promotes rapid behavioral adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 2014;111:7120–5. <https://doi.org/10.1073/pnas.1318906111>.
- [38] Oliveira AMM, Hemstedt TJ, Bading H. Rescue of aging-associated decline in Dnmt3a2 expression restores cognitive abilities. *Nat. Neurosci.* 2012;15:1111–3. <https://doi.org/10.1038/nn.3151>.
- [39] Rudenko A, Dawlaty MM, Seo J, Cheng AW, Meng J, Le T, Faull KF, Jaenisch R, Tsai L-H. Tet1 is critical for neuronal activity-regulated gene expression and memory extinction. *Neuron* 2013;79:1109–22. <https://doi.org/10.1016/j.neuron.2013.08.003>.
- [40] Zhang R, Lu J, Kong X, Jin L, Luo C. Targeting epigenetics in nervous system disease. *CNS Neurol. Disord. Drug Targets.* 2013;12:126–41.
- [41] Alzheimer’s Association. 2016 Alzheimer’s disease facts and figures. *Alzheimers Dement. J. Alzheimers Assoc.* 2016;12:459–509.
- [42] Chouraki V, Seshadri S. Genetics of Alzheimer’s disease. *Adv. Genet.* 2014;87:245–94. <https://doi.org/10.1016/B978-0-12-800149-3.00005-6>.
- [43] Mayeux R, Stern Y. Epidemiology of Alzheimer disease. *Cold Spring Harb. Perspect. Med.* 2012;2. <https://doi.org/10.1101/cshperspect.a006239>.
- [44] Ertekin-Taner N. Genetics of Alzheimer’s disease: a centennial review. *Neurol. Clin.* 2007;25:611–67. <https://doi.org/10.1016/j.ncl.2007.03.009>. v.
- [45] Kivipelto M, Mangialasche F. Alzheimer disease: To what extent can Alzheimer disease be prevented? *Nat. Rev. Neurol.* 2014;10:552–3. <https://doi.org/10.1038/nrneurol.2014.170>.
- [46] Kawas C, Gray S, Brookmeyer R, Fozard J, Zonderman A. Age-specific incidence rates of Alzheimer’s disease: the Baltimore Longitudinal Study of Aging. *Neurology* 2000;54:2072–7.
- [47] Chouliaras L, van den Hove DLA, Kenis G, van Draanen M, Hof PR, van Os J, Steinbusch HWM, Schmitz C, Rutten BPF. Histone deacetylase 2 in the mouse hippocampus: attenuation of age-related increase by caloric restriction. *Curr. Alzheimer Res.* 2013;10:868–76.
- [48] Chouliaras L, van den Hove DLA, Kenis G, Keitel S, Hof PR, van Os J, Steinbusch HWM, Schmitz C, Rutten BPF. Prevention of age-related changes in hippocampal levels of 5-methylcytidine by caloric restriction. *Neurobiol. Aging*. 2012;33:1672–81. <https://doi.org/10.1016/j.neurobiolaging.2011.06.003>.
- [49] Bansal A, Srivastava PA. Transcriptomics to Metabolomics: A Network Perspective for Big Data. *Http://servicesigi-Glob.-1-5225-2607-0ch008* 2018:188–206. <https://doi.org/10.4018/978-1-5225-2607-0.ch008>.
- [50] Vinson C, Chatterjee R. CG methylation. *Epigenomics* 2012;4:655–63. <https://doi.org/10.2217/epi.12.55>.
- [51] Pollack Y, Stein R, Razin A, Cedar H. Methylation of foreign DNA sequences in eukaryotic cells. *Proc. Natl. Acad. Sci. U. S. A.* 1980;77:6463–7.

- [52] Stein R, Razin A, Cedar H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl. Acad. Sci. U. S. A.* 1982;79:3418–22.
- [53] Wigler M, Levy D, Perucho M. The somatic replication of DNA methylation. *Cell* 1981;24:33–40.
- [54] Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z, Cedar H. Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.* 2009;16:564–71. <https://doi.org/10.1038/nsmb.1594>.
- [55] Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;25:1010–22. <https://doi.org/10.1101/gad.2037511>.
- [56] Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* 2009;138:114–28. <https://doi.org/10.1016/j.cell.2009.04.020>.
- [57] Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008;132:887–98. <https://doi.org/10.1016/j.cell.2008.02.022>.
- [58] Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* 2012;22:2497–506. <https://doi.org/10.1101/gr.143008.112>.
- [59] Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 2007;130:77–88. <https://doi.org/10.1016/j.cell.2007.05.042>.
- [60] Citron M, Oltersdorf T, Haass C, McConlogue L, Hung AY, Seubert P, Vigo-Pelfrey C, Lieberburg I, Selkoe DJ. Mutation of the beta-amyloid precursor protein in familial Alzheimer's disease increases beta-protein production. *Nature* 1992;360:672–4. <https://doi.org/10.1038/360672a0>.
- [61] Haass C, Lemere CA, Capell A, Citron M, Seubert P, Schenk D, Lannfelt L, Selkoe DJ. The Swedish mutation causes early-onset Alzheimer's disease by beta-secretase cleavage within the secretory pathway. *Nat. Med.* 1995;1:1291–6.
- [62] Mullan M, Crawford F, Axelman K, Houlden H, Lilius L, Winblad B, Lannfelt L. A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N-terminus of beta-amyloid. *Nat. Genet.* 1992;1:345–7. <https://doi.org/10.1038/ng0892-345>.
- [63] Sung HY, Choi EN, Ahn Jo S, Oh S, Ahn J-H. Amyloid protein-mediated differential DNA methylation status regulates gene expression in Alzheimer's disease model cell line. *Biochem. Biophys. Res. Commun.* 2011;414:700–5. <https://doi.org/10.1016/j.bbrc.2011.09.136>.
- [64] Hodgson N, Trivedi M, Muratore C, Li S, Deth R. Soluble oligomers of amyloid- β cause changes in redox state, DNA methylation, and gene transcription by inhibiting EAAT3 mediated cysteine uptake. *J. Alzheimers Dis. JAD.* 2013;36:197–209. <https://doi.org/10.3233/JAD-131011>.
- [65] Chen K-L, Wang SS-S, Yang Y-Y, Yuan R-Y, Chen R-M, Hu C-J. The epigenetic effects of amyloid-beta(1–40) on global DNA and neprilysin genes in murine cerebral endothelial cells. *Biochem. Biophys. Res. Commun.* 2009;378:57–61. <https://doi.org/10.1016/j.bbrc.2008.10.173>.
- [66] Taher N, McKenzie C, Garrett R, Baker M, Fox N, Isaacs GD. Amyloid- β alters the DNA methylation status of cell-fate genes in an Alzheimer's disease model. *J. Alzheimers Dis. JAD.* 2014;38:831–44. <https://doi.org/10.3233/JAD-131061>.
- [67] Mastroeni D, Grover A, Delvaux E, Whiteside C, Coleman PD, Rogers J. Epigenetic changes in Alzheimer's disease: decrements in DNA methylation. *Neurobiol. Aging.* 2010;31:2025–37. <https://doi.org/10.1016/j.neurobiolaging.2008.12.005>.
- [68] Lashley T, Gami P, Valizadeh N, Li A, Revesz T, Balazs R. Alterations in global DNA methylation and hydroxymethylation are not detected in Alzheimer's disease. *Neuropathol. Appl. Neurobiol.* 2015;41:497–506. <https://doi.org/10.1111/nan.12183>.

- [69] Coppiepers N, Dieriks BV, Lill C, Faull RLM, Curtis MA, Dragunow M. Global changes in DNA methylation and hydroxymethylation in Alzheimer's disease human brain. *Neurobiol. Aging.* 2014;35:1334–44. <https://doi.org/10.1016/j.neurobiolaging.2013.11.031>.
- [70] Bakulski KM, Dolinoy DC, Sartor MA, Paulson HL, Konen JR, Lieberman AP, Albin RL, Hu H, Rozek LS. Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. *J. Alzheimers Dis. JAD.* 2012;29:571–88. <https://doi.org/10.3233/JAD-2012-111223>.
- [71] Condliffe D, Wong A, Troakes C, Proitsi P, Patel Y, Chouliaras L, Fernandes C, Cooper J, Lovestone S, Schalkwyk L, Mill J, Lunnon K. Cross-region reduction in 5-hydroxymethylcytosine in Alzheimer's disease brain. *Neurobiol. Aging.* 2014;35:1850–4. <https://doi.org/10.1016/j.neurobiolaging.2014.02.002>.
- [72] Davies MN, Volta M, Pidsley R, Lunnon K, Dixit A, Lovestone S, Coarfa C, Harris RA, Milosavljevic A, Troakes C, Al-Sarraj S, Dobson R, Schalkwyk LC, Mill J. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol* 2012;13:R43. <https://doi.org/10.1186/gb-2012-13-6-r43>.
- [73] Hernandez DG, Nalls MA, Gibbs JR, Arepalli S, van der Brug M, Chong S, Moore M, Longo DL, Cookson MR, Traynor BJ, Singleton AB. Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum. Mol. Genet.* 2011;20:1164–72. <https://doi.org/10.1093/hmg/ddq561>.
- [74] Ladd-Acosta C, Pevsner J, Sabuncian S, Yolken RH, Webster MJ, Dinkins T, Callinan PA, Fan J-B, Potash JB, Feinberg AP. DNA methylation signatures within the human brain. *Am. J. Hum. Genet.* 2007;81: 1304–15. <https://doi.org/10.1086/524110>.
- [75] Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* 2011;39:e9. <https://doi.org/10.1093/nar/gkq1015>.
- [76] Bell JT, Spector TD. A twin approach to unraveling epigenetics. *Trends Genet. TIG.* 2011;27:116–25. <https://doi.org/10.1016/j.tig.2010.12.005>.
- [77] West RL, Lee JM, Maroun LE. Hypomethylation of the amyloid precursor protein gene in the brain of an Alzheimer's disease patient. *J. Mol. Neurosci. MN.* 1995;6:141–6. <https://doi.org/10.1007/BF02736773>.
- [78] Tohgi H, Utsugisawa K, Nagane Y, Yoshimura M, Genda Y, Ukitsu M. Reduction with age in methylcytosine in the promoter region -224 approximately -101 of the amyloid precursor protein gene in autopsy human cortex. *Brain Res. Mol. Brain Res.* 1999;70:288–92.
- [79] Yoshihikai S, Sasaki H, Doh-ura K, Furuya H, Sakaki Y. Genomic organization of the human amyloid beta-protein precursor gene. *Gene* 1990;87:257–63.
- [80] Fuso A, Nicolia V, Pasqualato A, Fiorenza MT, Cavallaro RA, Scarpa S. Changes in Presenilin 1 gene methylation pattern in diet-induced B vitamin deficiency. *Neurobiol. Aging.* 2011;32:187–99. <https://doi.org/10.1016/j.neurobiolaging.2009.02.013>.
- [81] Wang S-C, Oelze B, Schumacher A. Age-specific epigenetic drift in late-onset Alzheimer's disease. *PloS One* 2008;3:e2698. <https://doi.org/10.1371/journal.pone.0002698>.
- [82] Barrachina M, Ferrer I. DNA methylation of Alzheimer disease and tauopathy-related genes in postmortem brain. *J. Neuropathol. Exp. Neurol.* 2009;68:880–91. <https://doi.org/10.1097/NEN.0b013e3181af2e46>.
- [83] Sanchez-Mut JV, Aso E, Panayotis N, Lott I, Dierssen M, Rabano A, Urdinguio RG, Fernandez AF, Astudillo A, Martin-Subero JI, Balint B, Fraga MF, Gomez A, Gurnot C, Roux J-C, Avila J, Hensch TK, Ferrer I, Esteller M. DNA methylation map of mouse and human brain identifies target genes in Alzheimer's disease. *Brain J. Neurol.* 2013;136:3018–27. <https://doi.org/10.1093/brain/awt237>.
- [84] Bennett DA, Yu L, Yang J, Srivastava GP, Aubin C, De Jager PL. Epigenomics of Alzheimer's disease. *Transl. Res. J. Lab. Clin. Med.* 2015;165:200–20. <https://doi.org/10.1016/j.trsl.2014.05.006>.

- [85] Agbemenyah HY, Agis-Balboa RC, Burkhardt S, Delalle I, Fischer A. Insulin growth factor binding protein 7 is a novel target to treat dementia. *Neurobiol. Dis.* 2014;62:135–43. <https://doi.org/10.1016/j.nbd.2013.09.011>.
- [86] Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT. Neuropathological alterations in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* 2011;1:a006189. <https://doi.org/10.1101/cshperspect.a006189>.
- [87] Gonzalez-Zuñiga M, Contreras PS, Estrada LD, Chamorro D, Villagra A, Zanlungo S, Seto E, Alvarez AR. c-Abl stabilizes HDAC2 levels by tyrosine phosphorylation repressing neuronal gene expression in Alzheimer's disease. *Mol. Cell.* 2014;56:163–73. <https://doi.org/10.1016/j.molcel.2014.08.013>.
- [88] Sun W, Poschmann J, Cruz-Herrera Del Rosario R, Parkash NN, Hajan HS, Kumar V, Ramasamy R, Belgard TG, Elangovan B, Wong CCY, Mill J, Geschwind DH, Prabhakar S. Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* 2016;167:1385–97. <https://doi.org/10.1016/j.cell.2016.10.031>. e11.
- [89] Oakley H, Cole SL, Logan S, Maus E, Shao P, Craft J, Guillozet-Bongaarts A, Ohno M, Disterhoft J, Van Eldik L, Berry R, Vassar R. Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: potential factors in amyloid plaque formation. *J. Neurosci. Off. J. Soc. Neurosci.* 2006;26:10129–40. <https://doi.org/10.1523/JNEUROSCI.1202-06.2006>.
- [90] Gräff J, Tsai L-H. Histone acetylation: molecular mnemonics on the chromatin. *Nat. Rev. Neurosci.* 2013;14: 97–111. <https://doi.org/10.1038/nrn3427>.
- [91] Peleg S, Sananbenesi F, Zovoilis A, Burkhardt S, Bahari-Javan S, Agis-Balboa RC, Cota P, Wittnam JL, Gogol-Doering A, Opitz L, Salinas-Riester G, Dettenhofer M, Kang H, Farinelli L, Chen W, Fischer A. Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science* 2010; 328:753–6. <https://doi.org/10.1126/science.1186088>.
- [92] Chishti MA, Yang DS, Janus C, Phinney AL, Horne P, Pearson J, Strome R, Zuker N, Loukides J, French J, Turner S, Lozza G, Grilli M, Kunicki S, Morissette C, Paquette J, Gervais F, Bergeron C, Fraser PE, Carlson GA, George-Hyslop PS, Westaway D. Early-onset amyloid deposition and cognitive deficits in transgenic mice expressing a double mutant form of amyloid precursor protein 695. *J. Biol. Chem.* 2001;276: 21562–70. <https://doi.org/10.1074/jbc.M100710200>.
- [93] Seo J, Kritskiy O, Watson LA, Barker SJ, Dey D, Raja WK, Lin Y-T, Ko T, Cho S, Penney J, Silva MC, Sheridan SD, Lucente D, Gusella JF, Dickerson BC, Haggarty SJ, Tsai L-H. Inhibition of p25/Cdk5 Attenuates Tauopathy in Mouse and iPSC Models of Frontotemporal Dementia. *J. Neurosci.* 2017;37: 9917–24. <https://doi.org/10.1523/JNEUROSCI.0621-17.2017>.
- [94] Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai L-H, Kellis M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 2015;518:365–9. <https://doi.org/10.1038/nature14252>.
- [95] Jindal K, Bansal A. APOE ϵ 2 is Associated with Milder Clinical and Pathological Alzheimer's Disease. *Ann. Neurosci.* 2016;23:112. <https://doi.org/10.1159/000443572>.
- [96] Hasegawa M, Smith MJ, Goedert M. Tau proteins with FTDP-17 mutations have a reduced ability to promote microtubule assembly. *FEBS Lett* 1998;437:207–10.
- [97] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004;32:D115–9. <https://doi.org/10.1093/nar/gkh131>.
- [98] Kumar A, Singh TR. A New Decision Tree to Solve the Puzzle of Alzheimer's Disease Pathogenesis Through Standard Diagnosis Scoring System. *Interdiscip. Sci. Comput. Life Sci.* 2017;9:107–15. <https://doi.org/10.1007/s12539-016-0144-0>.

- [99] Panigrahi PP, Singh TR. Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: revealed association with aging and other diseases. *J. Theor. Biol.* 2013;334:109–21. <https://doi.org/10.1016/j.jtbi.2013.06.013>.

FURTHER READING

Goñi J, Esteban FJ, de Mendizábal NV, Sepulcre J, Ardanza-Trevijano S, Agirrezabal I, Villoslada P. A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Syst. Biol.* 2008;2:52. <https://doi.org/10.1186/1752-0509-2-52>.

10

EPIGENOMIC REPROGRAMMING IN CARDIOVASCULAR DISEASE

Yang Zhou, Jiandong Liu, Li Qian

*Department of Pathology and Laboratory Medicine, Department of Medicine, McAllister Heart Institute,
University of North Carolina, Chapel Hill, NC, United States*

INTRODUCTION

Cardiovascular diseases (CVDs) have emerged as one of the leading causes of illness and death in the world and have been producing numerous health and economic burdens worldwide. As reported by American Heart Association (AHA), around 92.1 million adults in the United States currently have at least one type of CVDs [1]. There are multiple types of CVDs that involve heart or blood vessels with different causes and pathological characteristics [2]. Despite advances in the prevention and treatment of CVDs, the death rate attributable to CVDs continues to rise mainly because the therapeutic strategies for CVDs, especially for heart failure, are limited and inadequate. Therefore, deeper understanding of molecular mechanisms underlying CVDs and new technologies are needed to discover more efficacious therapeutic approaches.

Considering the inherent low proliferative and regenerative capacity of mammalian cardiomyocytes (CMs), the most straightforward strategy for treating heart failure is to replenish functional CMs or replace the malfunctioned CMs in order to recover heart function [3–5]. Recently, the revolutionary work in the field of stem cell biology and cardiac regenerative medicine has progressed rapidly to deepen our understanding of cardiac development and open the new path to cardiac regeneration. Generation of autologous CMs via induced pluripotent stem cell (iPSC) reprogramming followed by differentiation and direct reprogramming from fibroblasts holds great promise as an alternative strategy for heart regeneration and disease modeling [6,7]. In studies over the past decade, iPSC reprogramming has been successfully achieved through the ectopic expression of master pluripotent transcription factors in various types of somatic cells from both murine and human [8–10]. The efficient differentiation of iPSCs into functional CMs mimics cardiac differentiation during early embryonic stage, providing not only the platform to dissect underlying mechanisms of cardiac development and diseases in patients, but also the source of CMs for potential utility in cell therapy [11–16]. More recently, inspired by iPSC reprogramming, functional induced cardiomyocytes (iCMs) have been derived from cardiac fibroblasts via forced expression of key cardiac transcription factors both in vitro and in vivo [17–24]. Direct cardiac reprogramming offers an appealing approach as it could accomplish *in situ* cardiac regeneration for regenerative therapy for heart diseases.

Epigenetics is typically defined as the regulatory mechanisms of gene activity that are not due to alterations in DNA sequence. Epigenome means genome-wide epigenetic regulations, including DNA methylation, posttranslational modification of histone, chromatin remodeling, higher-order DNA organization, and noncoding RNA alterations, all of which are heritable and sequence-independent. Epigenetic dynamics in the local and global organization of chromatin has been recognized as critical regulators to precisely determine the transcriptome of a cell [25]. The understanding of epigenomes will be able to explain how identical genomes in diverse types of cells within an individual organism produce such varied transcriptomes specific to each cell type. Moreover, results from studies in animal models clearly demonstrate that not only genetic factors, but also epigenomic variability leads to phenotypic variability and influence disease susceptibility [26,27]. Thus, the epigenomic regulations and misregulations underlying normal development and diseases hold promise to develop innovative biomarkers and therapies for CVDs.

With the rapid accumulation of the large-scale mapping of epigenomic and related data in CVD study, it has been highlighted that the use of computational tools is the core for data collecting, cleaning, clustering, modeling, and predicting, which gives rise to complex and comprehensive epigenomic information that is inaccessible using traditional approaches [28]. In addition, to facilitate the data analysis, massive data sources for epigenetic research have been collected and classified as invaluable databases [29], such as NIH Roadmap Epigenomics database [30], DNA methylation databases MethDB [31] and MethPrimerDB [32], VISTA human enhancer database [33], 3D-genome Interaction Viewer and database (3DIV) [34], Human Enhancer Disease Database (HEDD) [35]. Such tools and information allow for the systematic study of relationship between epigenomics and diseases. Herein, we focus on the recent computational analyses-based studies on landscapes and dynamics of chromatin modifications and structures in normal cardiomyocyte differentiation, CVD development, and heart regeneration, in particular direct cardiac reprogramming.

DECIPHER HISTONE CODES OF CM TRANSCRIPTION

In eukaryotic nucleus, the genomic DNA is wrapped around histone proteins in nucleosomes, which are the fundamental repeating structural units of chromatin. Each nucleosome is composed of about 146 base pairs of DNA wrapped around eight histones, called histone octamer, which contains two copies each of the histone proteins H2A, H2B, H3, and H4. The histones possess a diverse array of posttranslational modifications on specific residues along their N-terminal “tails”. To date, the histone modifications include acetylation, methylation, ubiquitination, and SUMOylation of lysine (K) residues, phosphorylation of serine (S) and threonine (T) residues, methylation of arginine (R) residues, ADP ribosylation, deamination, and isomerization of proline (P) [36,37]. The different histone modifications maintained and altered via corresponding enzymatic systems have been proposed to act sequentially or in combination to form a “histone code” that is read by other histone binding proteins or readers to bring about distinct genome activities and gene regulation [38]. Nowadays, numerous posttranslational modifications of histones have been documented and revealed with critical roles in mediating the genome function and gene activity in response to upstream signaling pathways [39]. For instance, H3K4me3 at promoters and transcription start sites, H3K27ac at enhancers and promoters, H3K36me3 at transcribed gene bodies are associated with transcription activation [40]. In contrast, H3K27me and H3K9me2/3 are related to gene repression and heterochromatin formation [40].

In addition, new methods harnessing the power of next-generation sequencing technology have been developed to interrogate chromatin dynamics at the genome-wide scale to reveal the link between epigenomic status and gene regulation, such as Chromatin ImmunoPrecipitation assay (ChIP-seq) and Assay for Transposase Accessible Chromatin (ATAC-seq), following with the generation of computational tools to interpret, visualize, and annotate this genome-wide information [41–44]. Thus, increasing enthusiasm of deciphering histone codes in the epigenome has been triggered.

IDENTIFY CHROMATIN MODIFICATION LANDSCAPES AND DYNAMICS DURING HEART DEVELOPMENT

Both gene expression and epigenetic signatures are highly cell type specific. A map of tissue- or cell type-specific promoter and enhancer regions has been drawn in the mouse genome to demonstrate and utilize the tight link of chromatin state and transcriptional activity [45,46]. In mouse heart, cardiac-specific promoters are identified by enrichment of H3K4me3 or Pol II binding, while cardiac-specific enhancers are defined based on the presence of H3K4me1 or H3K27ac outside promoter [46]. In pericentriolar material 1 (PCM1) positive mouse CM nuclei, the activity of mRNA is highly correlated with occupancy of H3K4me1, H3K4me3, H3K27ac, and H3K36me3, but inversely correlated with H3K27me3. Among these histone modifications, H3K27ac is the most predictive one for deducing transcriptional activity [45]. Furthermore, epigenome-wide analysis of H3K36me3 patterns could facilitate the identification of cardiac gene isoforms expressed in CMs [45].

In contrast to the relatively stable genome, the epigenome is very dynamic during development and differentiation in order to establish and maintain cell type-specific gene expression based on cellular identity and function. During cardiac differentiation, cell morphology and function are changed sequentially, as a result of alterations in gene expression as well as dramatic changes in the epigenetic landscape, which is required for appropriate cell fate differentiation [47,48]. In particular, during mouse embryonic stem cell (ESC) differentiation into CMs, a pattern of chromatin state transition, in which H3K4me1 enrichment is prior to enrichment of H3K4me3 and RNA Pol II on the promoter region of genes, is associated with later gene activation [48]. Also, analysis of occupancy of H3K4me1 and/or H3K27ac at distal enhancer regions has led to the identification of numerous putative cardiac fate-specific enhancers, which are undergoing rapid transitions between poised and active states at each stage of differentiation [48]. Similarly, genome-wide analysis of chromatin modifications along the time course of cardiac differentiation from human ESCs showed stage-specific changes in H3K4me3 and H3K27me3 levels [47]. Notably, a cardiac-specific chromatin signature has been identified to discriminate master regulatory factors of CM differentiation from CM structure proteins involved in muscle contraction and energy production [47]. A majority of cardiac transcription factors and members of key signaling pathways had increased active chromatin modifications (H3K4me3 and H3K36me3) and decreased repressive chromatin modification (H3K27me3), while genes encoding cardiac structure proteins showed a similar increase in active chromatin modification but no H3K27me3 deposition at any time [47]. This cardiac specific chromatin signature is also able to predict new regulators for appropriate human cardiac development [47]. Interestingly, a recent paper took advantage of iPSC reprogramming to study the epigenetic remodeling of cis-regulatory elements in cardiac development and diseases [49]. They performed massive ChIP-seq experiments for histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3) in different types of somatic cells derived from the same human fetal heart and their respective iPSCs. Defining with enrichment of H3K27ac

and H3K4me1, cell type-specific enhancer elements were identified in human iPSCs and heart cells, and functionally validated in transgenic mouse embryos and human cells.

Taken together, epigenome-wide analyses of histone modifications in combination with the transcriptome data not only provide new insights into the role of histone modifications and transcription factors in regulating cardiac-specific gene expression, but also identify stage-specific enhancers and promoters along heart development.

DYNAMICS OF REGULATORY *cis*-ELEMENTS IN HEART DISEASE

In addition to chromatin dynamics in normal heart development, a large number of regulatory *cis*-elements defined by a distinguished pattern of histone modifications have been identified in heart diseases [50–52]. Comparison of massive ChIP-seq and RNA-seq data between adult CMs isolated from hypertrophic hearts induced by transverse aortic constriction (TAC) and normal hearts demonstrated that pressure-overload hypertrophy is associated with changes in multiple types of histone modifications on a wide array of genes involved in cardiac functions [52]. Consistently, distinct global distribution of H3K36me3 and H3K4me3 enrichment correlated with mRNA abundance in cardiac *cis*-regulatory elements has been profiled in cardiomyopathic and normal hearts from rat models and human tissues [50,51]. Moreover, the epigenome study in mouse revealed that during cardiac hypertrophy, the epigenomic reprogramming occurs both on enhancers associated with normal heart development and on *cis*-elements specifically active in pathogenesis-wide [52].

It is also of great importance to understand whether and how the epigenomic dynamics contributes to the development of cardiomyopathy. The effect of histone modifications on cardiac diseases has been determined when they were removed or activated by manipulation of histone modification enzymes and administration of small molecules (see review in Ref. [53]). Such studies support the notion that epigenome-wide histone dynamics plays critical roles in heart failure pathogenesis (see review in Ref. [54]). Recently, epigenomic analyses showed that hyperacetylated chromatin induced by excessive activation of a histone acetyltransferase p300 is involved in pathological cardiac hypertrophy [55], while inhibition of p300 attenuates hypotrophic phenotype [56]. Overexpression of JMJD2, a demethylase of histone H3K9me3 and H3K36me3, exaggerated TAC-treated cardiac hypertrophy in CMs. Conversely, mice with JMJD2 loss were partially protected from pressure overload [57]. JMJD2 was shown to be associated with removal of H3K9me3 in specific loci of prohypertrophic genes [57]. In addition, a major component of polycomb repressive complex 2 (PRC2), EZH2, which catalyzes histone mark H3K27me3 for chromatin silencing, plays an important role during heart development and in the adult CMs [58,59].

Accumulating evidence implicates a potential mechanism by which altered chromatin signatures are recognized by specific DNA binding factors, or readers, to further regulate corresponding gene activity [60]. For instance, BET bromodomain reader protein was found to be critical transcriptional coactivator in activating pathologic genes that drive CM hypertrophy and heart failure progression. BET proteins recognize acetyllysine, which marks disease-specific genes in the genome, to increase the binding of Pol II complexes promoting chromatin remodeling, transcriptional initiation, and elongation [60]. Therefore, early administration of BET bromodomain selective inhibitor JQ1 blocks pathological hypertrophy in mice during pressure overload [60,61]. Collectively, dynamic changes in histone marks at the *cis*-regulatory regions of a genome allow DNA binding factors to activate and maintain specific gene expression with spatiotemporal precision during normal and pathological development.

DNA METHYLATION DURING HEART DEVELOPMENT AND IN DISEASE

The major form of DNA methylation refers to the addition of a methyl group to the 5th position on the cytosine ring of DNA (5-methylcytosine, 5mC). Mammalian DNA methylation at CpG dinucleotides is mainly catalyzed by three known DNA methyltransferases (DNMT1, DNMT3A, and DNMT3B) [62]. DNMT1 mainly involves in methylation maintenance, while DNMT3A and DNMT3B are de novo DNMTs that primarily methylate unmethylated DNA independent of DNA replication [63]. Reversely, removal of DNA methylation is carried out by the ten eleven translocation (TET) family catalyzing the conversion of 5mC to DNA to 5-hydroxymethylcytosine (5hmC) [64,65]. DNA methylation provides a critical epigenetic means for defining and maintaining cellular identity by regulating gene regulatory elements such as promoters and enhancers [66,67]. In general, DNA methylation is associated with gene repression, since DNA methylation is able to prevent the binding of transcriptional machinery and of transcription factors directly or influence chromatin structure [68,69]. There are a growing number of DNA methylation profiling technologies, which based on how 5mC is distinguished from cytosine, including commonly used reduced representation bisulfite sequencing (RRBS) [70], whole-genome bisulfite sequencing (WGBS) [71], methylated DNA immunoprecipitation sequencing (MeDIP-Seq) [72], and hydroxymethylated DNA immunoprecipitation sequencing (hMeDIP-Seq) [73].

DNA METHYLATION IS ORCHESTRATED IN NORMAL HEART

The study of DNA methylation in heart development and disease is at its early stage. DNA methylation has been increasingly recognized as a highly dynamic process [74,75], and highly associated with cell type-specific gene expression during development [76]. Whole-genome bisulfite sequencing of adult mouse heart tissues at base-pair resolution firstly provides a map of heart-specific differentially methylated regions (DMR), which are predominantly regulatory elements enriched with transcription factor binding motifs [76]. Furthermore, the DNA methylomes from nuclei of PCM-1 positive CMs in neonatal and adult mouse hearts were generated [74] and compared with those from ESCs [77] and whole heart tissues [76]. Interestingly, a highly dynamic pattern of DNA methylation at cardiac enhancers and gene bodies occurs during CM development and maturation [74]. In particular, fetal genes gain methylation and adult genes lose methylation on their gene bodies during postnatal CM maturation until adulthood, resulting in postnatal isoform switch of sarcomeric genes. Of note, postnatal methylation of fetal genes was dependent on the presence of de novo DNA methyltransferases Dnmt3a/b [74]. Most recently, high-coverage DNA methylomes have been generated by WGBS in FACS-sorted CM nuclei isolated from fetal, infant, adult, and end-stage failing human hearts [78]. Distinct mCpG patterns were identified in distal regulatory and genic regions and grouped into partially methylated regions (PMR), low methylated regions (LMR), and unmethylated regions (UMR), which are characterized with low level of active histone marks, enhancer histone mark H3K4me1, and promoter mark H3K4me3, respectively. They also found that during normal prenatal development and postnatal maturation, CM transcriptome is shaped by a highly dynamic interplay between mCpG on gene body and histone modifications. These results highlight the involvement of dynamic DNA methylation on gene bodies in CM maturation at postnatal stage. Investigation of DNA methylome in postnatal CMs in the absence of DNMT3a/b will address the issues whether DNA methylation is required for heart maturation. The underlying molecular mechanisms need further investigation.

Furthermore, the effect of DNA methylation on heart development can be suggested in studies of mice lacking DNMT. A general knockout of *Dnmt3b* leads to embryonic lethality between E13.5 and E16.5 with ventricular septal defects [63]. *Dnmt3a* homozygous knockout mice die at postnatal 3 weeks [63]. Whereas, CM-specific *Dnmt3a* and/or *Dnmt3b* knockout mice were generated in independent laboratories and showed complicated results [79,80]. α MHC-driven CM-specific loss of *Dnmt3b* resulted in compromised systolic function, widespread interstitial fibrosis, and myo-sarcomeric disarray [80]. In contrast, *Myl7*-driven CM-specific deletion of catalytic domains of both *Dnmt3a* and *Dnmt3b* showed no significant difference in CM function and CM response to pressure overload induced by TAC [79]. Moreover, in vitro RNAi-induced knockdown of *Dnmt3a* but not *Dnmt3b* via siRNA in cultured mouse embryonic CMs leads to sarcomere disassembly, and decrease in contractility and cytosolic calcium signaling [81]. Since these studies only showed restricted expression changes and methylation ablation in knockout or knockdown cells, it is still elusive whether DNA methyltransferases play pivotal roles for normal heart function. Combinational analyses of transcriptome and DNA methylome upon depletion of DNMTs will be valuable for a better understanding of DNA methylation in CMs.

DNA METHYLATION IS POTENTIAL THERAPEUTIC TARGET IN HEART DISEASE

Based on comparative analyses of DNA methylation in normal and diseased hearts, DNA methylation is increasingly recognized as a fundamental epigenetic modification associated with cardiac diseases. As compared with postnatal CMs isolated from normal hearts, failing CMs partially resemble DNA methylation patterns in neonatal mouse CMs rather than those in adult CMs [74]. Moreover, two recent studies have reported differential DNA methylation signatures in hearts from patients with cardiomyopathy [51,82]. For the first time, genome-wide cardiac DNA methylation on human dilated cardiomyopathy in patients was generated by methylation chip and compared with controls from healthy patients [82]. The authors found that genes with differential DNA methylation were significantly enriched in pathways related to cardiac disease. In addition, they identified novel candidate regulators *LY75* and *ADORA2A* via alteration of DNA methylation and mRNA expression for dilated cardiomyopathy, and evaluated the function of these genes in zebrafish model [82]. In the other study, compared with normal human hearts, DNA methylation maps generated by methylated DNA immunoprecipitation sequencing (MeDIP-seq) in end-stage failing hearts revealed that methylation changes in promoter CpG islands and gene bodies of genes that play critical roles in myocardial stress response, but not in intergenic CpG islands and enhancer CpG islands [51]. Intriguingly, loss of methylation in promoter regions is only correlated with upregulated genes in cardiomyopathic hearts, while no significant changes of methylation at the promoter of downregulated genes. However, different results were found in the comparative analysis of DNA methylome in nonfailing and heart failing human CMs, purified with specific CM marker phospholamban (PLN) [78]. Differentially expressed pathological genes show minimal alterations in mCpG of genic and cis-regulatory regions. Instead, they found that single-nucleotide polymorphisms (SNPs) associated with cardiac disease traits are highly enriched in low methylated cis-regulatory regions of human CMs. Collectively, apart from the identification of a specific DNA methylation dynamics of heart disease, we learned two more lessons from these studies. First, since different CVDs have disease-specific DNA methylation signatures [51,74,82], it is reasonable to hypothesize that DNA methylation signatures in CVDs can serve as potential diagnostic biomarkers and therapeutic targets. Most recently, in a multiomics study,

several epigenetic loci have been identified to be significantly associated with dilated cardiomyopathy and might be potential epigenetic biomarkers for heart failure [83]. Second, through epigenome-wide analysis of DNA methylation combined with maps of histone marks, novel disease-related genes have been identified with functional relevance and will be potentially druggable targets in CVDs.

DNA HYDROXYMETHYLATION REGULATES GENE EXPRESSION IN CARDIAC DEVELOPMENT AND HYPERTROPHY

Although the role of DNA hydroxymethylation is not fully understood, studies have suggested that 5hmC is not only an intermediate product in the active DNA demethylation of 5mC, but also recognized as another stable epigenetic mark on DNA to regulate gene expression in several cell types [84–88]. 5hmC is found to be located on gene body and associated with active transcription in multiple cells [84–88]. Whereas, enrichment of 5hmC was also found on the transcription start sites of repressed but poised genes, whose promoters carry bivalent histone methylation marks, H3K4me3 and H3K27me3 in ESCs [86,87]. Moreover, it is widely reported that the occupancy of 5mC and 5hmC in the genome is highly correlated [84,89]. Recently, the base-resolution analysis of 5hmC in the CM genome dissects the role of DNA hydroxymethylation in cardiac development and hypertrophy [90]. The genome-wide 5hmC distribution was determined by hydroxymethylated DNA immunoprecipitation (hMeD-IP) coupled with high-throughput sequencing in CMs isolated from embryonic, neonatal, adult, and TAC-induced hypertrophic hearts. The majority of 5hmC was located at introns and intergenic regions. However, accumulation of 5hmC on the gene body is strongly correlative with active cardiac gene expression during development, while accompanied with loss of 5hmC on intergenic regions. Interestingly, in line with the finding in DNA methylome during heart development [74], cardiac hypertrophy leads to a shift of 5hmC modification towards a neonatal-like pattern. Furthermore, appropriate 5hmC distribution and gene expression in CM require the presence of TET2. Taken together, DNA hydroxymethylation is largely reprogrammed in heart development, as well as cardiac hypertrophy. Further studies need to elucidate the balance between 5mC and 5hmC for gene regulation in cardiac physiology and disease.

CHROMATIN CONFORMATION IN CARDIOMYOCYTES

As mentioned above, besides modifications on histone proteins and DNA, epigenomic regulation also includes changes of chromatin structures, which reflects dynamic accessibility of genetic information and inter- and intrachromosomal communication [91]. Using newly developed chromosome conformation capture technology (such as 3C, 4C, 5C, Hi-C) and related computational tools [92], the 3-dimensional organization of chromatin has been investigated at high resolution in the whole genome [91,93]. Vast amounts of genome-wide interaction data discovered that chromosomes consist of discrete topologically associating domains (TADs), genome compartments, chromatin looping and interactions between cis-elements [94], which are defined by boundary binding of CTCF [95]. The loss of CTCF, which is the critical insulator for proper chromatin structures, disrupts morphogenesis and maturation of embryonic hearts before death at embryonic day 12.5 [96]. Interestingly, another α MHC-driven CM-specific loss of CTCF leads to cardiomyopathy even worse than TCA-induced phenotype [97]. Then Hi-C and high-throughput sequencing were applied to adult CMs from normal, TCA-treated, and CTCF-CKO hearts to investigate the contribution

of chromatin structure in healthy and diseased CMs [97]. In heart failure models, the large-scale alterations in chromatin structure have been identified. Notably, pressure overload and CTCF depletion selectively altered chromatin looping near genes associated with disease, especially influenced the interaction between enhancers and genes [97]. This study provides a valuable resource for further investigation of epigenome dynamics when combined with many other data sets such as DNA methylome, histone mark ChIP-seq and ATAC-seq in cardiac development and diseases.

RAPID CHROMATIN SWITCH DURING SOMATIC REPROGRAMMING

Underlying mechanisms of direct cardiac reprogramming are not clear, yet we know that cardiac reprogramming process is associated with extensive epigenetic changes [98–101]. Recently, we investigated the occupancy of H3K4me3, H3K27me3 and DNA methylation on specific cardiac and fibroblast-related loci and showed temporal changes of epigenetic status during cardiac reprogramming, in which cardiac loci were rapidly bound with active marks, while loci associated with fibroblast fate were gradually labeled with repressive marks [102]. Moreover, decrease of repressive mark H3K27me3 by downregulating PRC2 complex components or pharmaceutical inhibitors appeared to be promotive for the initiation of iCM reprogramming mediated by a microRNA combination of miR-1, miR-133, miR-208, and miR-499 [98]. Besides, we and others identified major epigenetic barriers to iCM reprogramming through loss-of-function and gain-of-function screens of epigenetic factors [99,101]. As the key component of polycomb repressive complex 1 (PRC1), Bmi1 represses cardiac gene expression during Mef2c/Gata4/Tbx5-induced iCM reprogramming through histone repressive mark H2AK119ub on cardiac loci. Removal of Bmi1 deactivates cardiac gene expression, especially Gata4, leading to successful reprogramming with only two factors, Mef2c and Tbx5 [101]. Meanwhile, Liu et al. found that overexpression of Men1 reduced reprogramming efficiency. Men1 is the coactivator of methyltransferase Mll1, which is one of the “writer” proteins of H3K4 methylation. Consistently, inhibition of Mll1 by small molecules enhanced iCM generation [99]. However, the dynamic epigenomic landscape highly associated with cardiac fate conversion and maintenance remains not clarified.

Nevertheless, during direct neuronal reprogramming, epigenomic changes mediated by pioneer transcription factors have been investigated and showed critical roles of epigenomic dynamics in the achievement of induced neuronal cells (iNs) [103]. The higher order chromatin architecture is relatively less clear, but loose and condensed chromatin structures reflecting DNA accessibility to regulatory factors and complexes are thought to have a pronounced impact on gene regulation [39] and control of cell fate [104]. Recently, the chromatin accessibility changes of direct reprogramming of fibroblasts into iNs have been studied in a genome-wide fashion and showed the rapid chromatin switch mediated by Ascl1 throughout the course of iN reprogramming [103]. The analysis of ATAC-seq data indicated that the majority of genomic loci are affected from as early as 12 hours to 5 days post-infection, while little chromatin remodeling occurs in sorted iN expressing neuronal reporter gene TauEGFP at day 5 and later stages. In combination with ChIP-seq data for Ascl1 and RNA-seq data for the corresponding time points during iN reprogramming [105], network analysis of ATAC-seq data identified several novel critical transcription factors Zfp238, Sox8, and Dlx3. Each of them is able to generate iNs in combination with another iN reprogramming factor Myt1l [103]. Although it is still unknown if the epigenomic landscape change is a common feature between different direct reprogramming systems, these findings highlight the importance of chromatin state switch from donor cell type to the target one and might be valuable for future translation of direct reprogramming for regenerative medicine.

CONCLUSION

Epigenomic reprogramming is engaged during normal heart development and cardiac diseases. All of the above findings including altered histone modifications, global DNA methylation, and the plastic genome structures support the notion that epigenome offers a new perspective in the control of gene regulation, with a promising application to CVD therapy. With a rapidly growing number of epigenomes being determined in normal and diseased cells, one of the main challenges we will face is how to better integrate epigenomic data with other omics data, like transcriptome, proteome, and metabolome to gain useful biological insight. More effective bioinformatics and standardized computational tools are urgently needed. Further efforts to combinational omics analyses will provide new mechanisms of chromatin regulation at different chromatin levels and inspire new treatment strategies for heart failure therapy. Meanwhile, comprehensive decoding of epigenetic patterns in somatic cell reprogramming could facilitate understanding of the underlying mechanism of cell fate conversion and clinical translation, and ultimately pave the way for the development of personalized medicine for CVDs.

REFERENCES

- [1] Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, De Ferranti SD, Floyd J, Fornage M, Gillespie C, Isasi CR, Jim'nez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, MacKey RH, Matsushita K, Mozaffarian D, Mussolino ME, Nasir K, Neumar RW, Palaniappan L, Pandey DK, Thiagarajan RR, Reeves MJ, Ritchey M, Rodriguez CJ, Roth GA, Rosamond WD, Sasson C, Towfighi A, Tsao CW, Turner MB, Virani SS, Voeks JH, Willey JZ, Wilkins JT, Wu JHY, Alger HM, Wong SS, Muntner P. Heart disease and stroke Statistics'2017 update: a report from the American heart association. *Circulation* 2017;135:e145–603.
- [2] Mendis S, Puska P, Norrvig B. Global atlas on cardiovascular disease prevention and control. *World Health Org* 2011;2–14.
- [3] Laflamme MA, Murry CE. Heart regeneration. *Nature* 2011;473:326–35.
- [4] Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN, Sadek HA. Transient regenerative potential of the neonatal mouse heart. *Science* 2011;331:1078–80.
- [5] Xin M, Olson EN, Bassel-Duby R. Mending broken hearts: cardiac development as a basis for adult heart regeneration and repair. *Nat Rev Mol Cell Biol* 2013;14:529–41.
- [6] Lee CY, Kim R, Ham O, Lee J, Kim P, Lee S, Oh S, Lee H, Lee M, Kim J, Chang W. Therapeutic potential of stem cells strategy for cardiovascular diseases. *Stem Cells Int* 2016;2016:4285938.
- [7] Srivastava D, DeWitt N. In-vivo cellular reprogramming: the next generation. *Cell* 2016;166:1386–96.
- [8] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 2007;131:861–72.
- [9] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;126:663–76.
- [10] Takahashi K, Yamanaka S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nat Rev Mol Cell Biol* 2016;17:183–93.
- [11] Bellin M, Casini S, Davis RP, Aniello CD, Haas J, Oostwaard DW, Tertoolen LGJ, Jung CB, Elliott DA, Welling A, Laugwitz K, Moretti A, Mummery CL. Isogenic human pluripotent stem cell pairs reveal the role of a KCNH2 mutation in long-QT syndrome. *EMBO J* 2013;32:3161–75.

- [12] Dambrot C, Passier R, Atsma D, Mummery CL. Cardiomyocyte differentiation of pluripotent stem cells and their use as cardiac disease models. *Biochem J* 2011;434:25–35.
- [13] Davis RP, Casini S, van den Berg CW, Hoekstra M, Remme CA, Dambrot C, Salvatori D, Oostwaard DW, Wilde AAM, Bezzina CR, Verkerk AO, Freund C, Mummery CL. Cardiomyocytes derived from pluripotent stem cells recapitulate electrophysiological characteristics of an overlap syndrome of cardiac sodium channel disease. *Circulation* 2012;125:3079–91.
- [14] Galdos FX, Guo Y, Paige SL, Vandusen NJ, Wu SM, Pu WT. Cardiac regeneration: lessons from development. *Circ Res* 2017;120:941–59.
- [15] Karakikes I, Ameen M, Termglinchan V, Wu JC. Human induced pluripotent stem cell-derived cardiomyocytes: insights into molecular, cellular, and functional phenotypes. *Circ Res* 2015;117:80–8.
- [16] Kattman SJ, Witty AD, Gagliardi M, Dubois NC, Niapour M, Hotta A, Ellis J, Keller G. Stage-specific optimization of activin/nodal and BMP signaling promotes cardiac differentiation of mouse and human pluripotent stem cell lines. *Cell Stem Cell* 2011;8:228–40.
- [17] Fu JD, Stone NR, Liu L, Spencer CI, Qian L, Hayashi Y, Delgado-Olguin P, Ding S, Bruneau BG, Srivastava D. Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. *Stem Cell Rep* 2013;1:235–47.
- [18] Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* 2010;142:375–86.
- [19] Nam Y-J, Song K, Luo X, Daniel E, Lambeth K, West K, Hill JA, DiMaio JM, Baker LA, Bassel-Duby R, Olson EN. Reprogramming of human fibroblasts toward a cardiac fate. *Proc Natl Acad Sci USA* 2013;110:5588–93.
- [20] Qian L, Huang Y, Spencer CI, Foley A, Vedantham V, Liu L, Conway SJ, Fu J, Srivastava D. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* 2012;485:593–8.
- [21] Song K, Nam Y-J, Luo X, Qi X, Tan W, Huang GN, Acharya A, Smith CL, Tallquist MD, Neilson EG, Hill JA, Bassel-Duby R, Olson EN. Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature* 2012;485:599–604.
- [22] Wang G, McCain ML, Yang L, He A, Pasqualini FS, Agarwal A, Yuan H, Jiang D, Zhang D, Zangi L, Geva J, Roberts AE, Ma Q, Ding J, Chen J, Wang D-Z, Li K, Wang J, Wanders RJA, Kulik W, Vaz FM, Laflamme MA, Murry CE, Chien KR, Kelley RI, Church GM, Parker KK, Pu WT. Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat Med* 2014;20:616–23.
- [23] Ye L, Chang YH, Xiong Q, Zhang L, Somasundaram P, Lepley M, Swingen C, Su L, Wendel JS, Guo J, Jang A, Rosenbush D, Greder L, Dutton JR, Zhang J, Kamp TJ, Kaufman DS, Ge Y, Zhang J. Cardiac repair in a porcine model of acute myocardial infarction with human induced pluripotent stem cell-derived cardiovascular cells. *Cell Stem Cell* 2014;15:750–61.
- [24] Zhang J, Wilson GF, Soerens AG, Koonce CH, Yu J, Palecek SP, Thomson JA, Kamp TJ. Functional cardiomyocytes derived from human induced pluripotent stem cells. *Circ Res* 2009;104:e30–41.
- [25] Cantone I, Fisher AG. Epigenetic programming and reprogramming during development. *Nat Struct Mol Biol* 2013;20:282–9.
- [26] Gluckman PD, Hanson MA, Buklijas T, Low FM, Beedle AS. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *Nat Rev Endocrinol* 2009;5:401–8.
- [27] Maunakea AK, Chepelev I, Zhao K. Epigenome mapping in normal and disease states. *Circ Res* 2010;107:327–39.
- [28] Lim SJ, Tan TW, Tong JC. Computational Epigenetics: the new scientific paradigm. *Bioinformation* 2010;4:331–7.
- [29] Galperin MY, Rigden DJ, Fernández-Suárez XM. The 2018 nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res* 2018;43:D1–5.

- [30] Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, Cohen RF, Schuler GD. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res* 2011;39:D908–12.
- [31] Grunau C. MethDB—a public database for DNA methylation data. *Nucleic Acids Res* 2001;29:270–4.
- [32] Pattyn F, Hoebeek J, Robbrecht P, Michels E, De Paepe A, Bottu G, Coornaert D, Herzog R, Speleman F, Vandesompele J. methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. *BMC Bioinf* 2006;7:496.
- [33] Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser - a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;35:D88–92.
- [34] Yang D, Jang I, Choi J, Kim M-S, Lee AJ, Kim H, Eom J, Kim D, Jung I, Lee B. 3DIV: a 3D-genome interaction viewer and database. *Nucleic Acids Res* 2017;46:D52–7.
- [35] Wang Z, Zhang Q, Zhang W, Lin J-R, Cai Y, Mitra J, Zhang ZD. HEDD: human enhancer disease database. *Nucleic Acids Res* 2017;46:D113–20.
- [36] Kouzarides T. Chromatin modifications and their function. *Cell* 2007;128:693–705.
- [37] Peterson CL, Laniel M-A. Histones and histone modifications. *Curr Biol* 2004;14:R546–51.
- [38] Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;403:41–5.
- [39] Fischle W, Wang Y, Allis CD. Histone and chromatin cross-talk. *Curr Opin Cell Biol* 2003;15:172–83.
- [40] Chen T, Dent SYR. Chromatin modifiers and remodelers: regulators of cellular differentiation. *Nat Rev Genet* 2014;15:93–106.
- [41] Bardet AF, He Q, Zeitlinger J, Stark A. A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc* 2012;7:45–61.
- [42] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;2015. 21.29.1-21.29.9.
- [43] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10:669–80.
- [44] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform* 2016;17:953–66.
- [45] Preissl S, Schwaderer M, Rauf A, Hesse M, Grüning BA, Köbele C, Backofen R, Fleischmann BK, Hein L, Gilsbach R. Deciphering the epigenetic code of cardiac myocyte transcription. *Circ Res* 2015;117:413–23.
- [46] Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. *Nature* 2012;488:116–20.
- [47] Paige SL, Thomas S, Stoick-Cooper CL, Wang H, Maves L, Sandstrom R, Pabon L, Reinecke H, Pratt G, Keller G, Moon RT, Stamatoyannopoulos J, Murry CE. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* 2012;151:221–32.
- [48] Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* 2012;151:206–20.
- [49] Zhao M, Shao N-Y, Hu S, Ma N, Srinivasan R, Jahanbani F, Lee J, Zhang SL, Snyder MP, Wu JC. Cell type-specific chromatin signatures underline regulatory DNA elements in human induced pluripotent stem cells and somatic cells. *Circ Res* 2017;121:1237–50.
- [50] Kaneda R, Takada S, Yamashita Y, Choi YL, Nonaka-Sarukawa M, Soda M, Misawa Y, Isomura T, Shimada K, Mano H. Genome-wide histone methylation profile for heart failure. *Gene Cells* 2009;14:69–77.
- [51] Movassagh M, Choy MK, Knowles DA, Cordeddu L, Haider S, Down T, Siggins L, Vujic A, Simeoni I, Penkett C, Goddard M, Lio P, Bennett MR, Foo RSY. Distinct epigenomic features in end-stage failing human hearts. *Circulation* 2011;124:2411–22.

- [52] Papait R, Cattaneo P, Kunderfranco P, Greco C, Carullo P, Guffanti A, Vigano V, Stirparo GG, Latronico MVG, Hasenfuss G, Chen J, Condorelli G. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. *Proc Natl Acad Sci USA* 2013;110:20164–9.
- [53] Gillette TG, Hill JA. Readers, writers, and erasers: chromatin as the whiteboard of heart disease. *Circ Res* 2015;116:1245–53.
- [54] Di Salvo TG, Haldar SM. Epigenetic mechanisms in heart failure pathogenesis. *Circ Heart Fail* 2014;7: 850–63.
- [55] Wei JQ, Shehadeh LA, Mitrani JM, Pessanha M, Slepak TI, Webster KA, Bishopric NH. Quantitative control of adaptive cardiac hypertrophy by acetyltransferase p300. *Circulation* 2008;118:934–46.
- [56] Morimoto T, Sunagawa Y, Kawamura T, Takaya T, Wada H, Nagasawa A, Komeda M, Fujita M, Shimatsu A, Kita T, Hasegawa K. The dietary compound curcumin inhibits p300 histone acetyltransferase activity and prevents heart failure in rats. *J Clin Investig* 2008;118:868–78.
- [57] Zhang QJ, Chen HZ, Wang L, Liu DP, Hill JA, Liu ZP. The histone trimethyllysine demethylase JMJD2A promotes cardiac hypertrophy in response to hypertrophic stimuli in mice. *J Clin Investig* 2011;121: 2447–56.
- [58] Delgado-Olguín P, Huang Y, Li X, Christodoulou D, Seidman CE, Seidman JG, Tarakhovsky A, Bruneau BG. Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis. *Nat Genet* 2012;44:343–7.
- [59] He A, Ma Q, Cao J, Von Gise A, Zhou P, Xie H, Zhang B, Hsing M, Christodoulou DC, Cahan P, Daley GQ, Kong SW, Orkin SH, Seidman CE, Seidman JG, Pu WT. Polycomb repressive complex 2 regulates normal development of the mouse heart. *Circ Res* 2012;110:406–15.
- [60] Anand P, Brown JD, Lin CY, Qi J, Zhang R, Artero PC, Alaiti MA, Bullard J, Alazem K, Margulies KB, Cappola TP, Lemieux M, Plutzky J, Bradner JE, Haldar SM. BET bromodomains mediate transcriptional pause release in heart failure. *Cell* 2013;154:569–82.
- [61] Spiltoir JI, Stratton MS, Cavasin MA, Demos-Davies K, Reid BG, Qi J, Bradner JE, McKinsey TA. BET acetyl-lysine binding proteins control pathological cardiac hypertrophy. *J Mol Cell Cardiol* 2013;63:175–9.
- [62] Bestor TH. The DNA methyltransferases of mammals. *Hum Mol Genet* 2000;9:2395–402.
- [63] Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 1999;99:247–57.
- [64] Ito S, Dalessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* 2010;466:1129–33.
- [65] Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 2009;324:930–5.
- [66] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012; 13:484–92.
- [67] Schübeler D. Function and information content of DNA methylation. *Nature* 2015;517:321–6.
- [68] Bird AP. CpG-Rich islands and the function of DNA methylation. *Nature* 1986;321:209–13.
- [69] Kass SU, Pruss D, Wolffe AP. How does DNA methylation repress transcription? *Trends Genet* 1997;13: 444–9.
- [70] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;454:766–70.
- [71] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462:315–22.

- [72] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavaré S, Beck S. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 2008;26:779–85.
- [73] Song CX, Szulwach KE, Fu Y, Dai Q, Yi C, Li X, Li Y, Chen CH, Zhang W, Jian X, Wang J, Zhang L, Looney TJ, Zhang B, Godley LA, Hicks LM, Lahn BT, Jin P, He C. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* 2011;29:68–75.
- [74] Gilsbach R, Preissl S, Grüning BA, Schnick T, Burger L, Benes V, Würch A, Bönisch U, Günther S, Backofen R, Fleischmann BK, Schübeler D, Hein L. Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease. *Nat Commun* 2014;5:5288.
- [75] Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 2008;9:465–76.
- [76] Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, Ren B. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* 2013;45:1198–206.
- [77] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 2011;480:490–5.
- [78] Gilsbach R, Schwaderer M, Preissl S, Grüning BA, Kranzhöfer D, Schneider P, Nührenberg TG, Mulero-Navarro S, Weichenhan D, Braun C, Dreßen M, Jacobs AR, Lahm H, Doenst T, Backofen R, Krane M, Gelb BD, Hein L. Distinct epigenetic programs regulate cardiac myocyte development and disease in the human heart *in vivo*. *Nat Commun* 2018;9:391.
- [79] Nührenberg TG, Hammann N, Schnick T, Preißl S, Witten A, Stoll M, Gilsbach R, Neumann FJ, Hein L. Cardiac myocyte de novo DNA methyltransferases 3a/3b are dispensable for cardiac function and remodeling after chronic pressure overload in mice. *PLoS One* 2015;10:e0131019.
- [80] Vujic A, Robinson EL, Ito M, Haider S, Ackers-Johnson M, See K, Methner C, Figg N, Brien P, Roderick HL, Skepper J, Ferguson-Smith A, Foo RS. Experimental heart failure modelled by the cardiomyocyte-specific loss of an epigenome modifier, DNMT3B. *J Mol Cell Cardiol* 2015;82:174–83.
- [81] Fang X, Poulsen RR, Wang-Hu J, Shi O, Calvo NS, Simmons CS, Rivkees SA, Wendler CC. Knockdown of DNA methyltransferase 3a alters gene expression and inhibits function of embryonic cardiomyocytes. *FASEB J* 2016;30:3238–55.
- [82] Haas J, Frese KS, Park YJ, Keller A, Vogel B, Lindroth AM, Weichenhan D, Franke J, Fischer S, Bauer A, Marquart S, Sedaghat-Hamedani F, Kayvanpour E, Köhler D, Wolf NM, Hassel S, Nietsch R, Wieland T, Ehlermann P, Schultz JH, Dösch A, Mereles D, Hardt S, Backs J, Hoheisel JD, Plass C, Katus HA, Meder B. Alterations in cardiac DNA methylation in human dilated cardiomyopathy. *EMBO Mol Med* 2013;5:413–29.
- [83] Meder B, Haas J, Sedaghat-Hamedani F, Kayvanpour E, Frese K, Lai A, Nietsch R, Scheiner C, Mester S, Bordalo DM, Amr A, Dietrich C, Pils D, Siede D, Hund H, Bauer A, Holzer DB, Ruhparwar A, Mueller-Hennen M, Weichenhan D, Plass C, Weis T, Backs J, Wuerstle M, Keller A, Katus HA, Posch AE. Epigenome-wide association study identifies cardiac gene patterning and a novel class of biomarkers for heart failure. *Circulation* 2017;136:1528–44.
- [84] Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 2011;473:398–404.
- [85] Ivanov M, Kals M, Kacevska M, Barragan I, Kasuga K, Rane A, Metspalu A, Milani L, Ingelman-Sundberg M. Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. *Genome Biol* 2013;14:R83.

- [86] Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S, Rao A. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 2011;473:394–7.
- [87] Szulwach KE, Li X, Li Y, Song CX, Wu H, Dai Q, Irier H, Upadhyay AK, Gearing M, Levey AI, Vasanthakumar A, Godley LA, Chang Q, Cheng X, He C, Jin P. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci* 2011;14:1607–16.
- [88] Tsagaratou A, Aijo T, Lio C-WJ, Yue X, Huang Y, Jacobsen SE, Lahdesmaki H, Rao A. Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. *Proc Natl Acad Sci USA* 2014;111:E3306–15.
- [89] Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B, He C. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 2012;149:1368–80.
- [90] Greco CM, Kunderfranco P, Rubino M, Larcher V, Carullo P, Anselmo A, Kurz K, Carell T, Angius A, Latronico MVG, Papait R, Condorelli G. DNA hydroxymethylation controls cardiomyocyte gene expression in development and hypertrophy. *Nat Commun* 2016;7:12418.
- [91] Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 2013;14:390–403.
- [92] Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat Methods* 2017;14:679–85.
- [93] Wit EDE, Laat WDe. A decade of 3C technologies-insights into nuclear organization. *Genes Dev* 2012;26:11–24.
- [94] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485:376–80.
- [95] Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 2006;7:703–13.
- [96] Gomez-Velazquez M, Badia-Careaga C, Lechuga-Vieco AV, Nieto-Arellano R, Tena JJ, Rollan I, Alvarez A, Torroja C, Caceres EF, Roy A, Galjart N, Delgado-Olguin P, Sanchez-Cabo F, Enriquez JA, Gomez-Skarmeta JL, Manzanares M. CTCF counter-regulates cardiomyocyte development and maturation programs in the embryonic heart. *PLoS Genet* 2017;13:e1006985.
- [97] Rosa-Garrido M, Chapski DJ, Schmitt AD, Kimball TH, Karbassi E, Monte E, Balderas E, Pellegrini M, Shih TT, Soehalim E, Liem D, Ping P, Galjart NJ, Ren S, Wang Y, Ren B, Vondriska TM. High-resolution mapping of chromatin conformation in cardiac myocytes reveals structural remodeling of the epigenome in heart failure. *Circulation* 2017;136:1613–25.
- [98] Dal-Pra S, Hodgkinson CP, Mirotsou M, Kirste I, Dzau VJ. Demethylation of H3K27 is essential for the induction of direct cardiac reprogramming by miR combo. *Circ Res* 2017;120:1403–13.
- [99] Liu L, Lei I, Karatas H, Li Y, Wang L, Gnatovskiy L, Dou Y, Wang S, Qian L, Wang Z. Targeting Mll1 H3K4 methyltransferase activity to guide cardiac lineage specific reprogramming of fibroblasts. *Cell Discov* 2016;2:16036.
- [100] Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, Wu X, Stack EC, Loda M, Liu T, Xu H, Cato L, Thornton JE, Gregory RI, Morrissey C, Vessella RL, Montironi R, Magi-Galluzzi C, Kantoff PW, Balk SP, Liu XS, Brown M, Varambally S, Cao R, Chen H, Tu SW, Hsieh JT, Lee ST, LaJeunesse D, Shearn A, Strutt H, Cavalli G, Paro R, Culig Z, Yu YP, Varambally S, Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL, Cao R, Zhang Y, Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D, Joshi P, Chen S, Kaneko S, Cha TL, Wei Y, Gao T, Furnari F, Newton AC, Sparmann A, Lohuizen M, van Nikoloski G, Ntziachristos P, Kuzmichev A, Jenuwein T, Tempst P, Reinberg D, Wang H, Wang Q, Ni M, Johnson WE, Li C, Rabinovic A, Bolstad BM, Irizarry RA, Astrand M, Speed TP, Smyth GK, Sandberg R, Larsson O, Xu J, He HH, Zhang Y, Wang Q, Li G, Tepper CG, Lin DL, Shin H,

- Liu T, Manrai AK, Liu XS, Rhodes DR, Yu J, Ji H, Vokes SA, Wong WH, Meyer CA, He HH, Brown M, Liu XS, Varambally S, Shatkina L, Cao R, Zhang Y. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* 2012;338:1465–9.
- [101] Zhou Y, Wang L, Vaseghi HR, Liu Z, Lu R, Alimohamadi S, Yin C, Fu J-D, Wang GG, Liu J, Qian L. Bmi1 is a key epigenetic barrier to direct cardiac reprogramming. *Cell Stem Cell* 2016;18:382–95.
- [102] Liu Z, Chen O, Zheng M, Wang L, Zhou Y, Yin C, Liu J, Qian L. Re-patterning of H3K27me3, H3K4me3 and DNA methylation during fibroblast conversion into induced cardiomyocytes. *Stem Cell Res* 2016;16:507–18.
- [103] Wapinski OL, Lee QY, Chen AC, Li R, Corces MR, Ang CE, Treutlein B, Xiang C, Baubet V, Suchy FP, Sankar V, Sim S, Quake SR, Dahmane N, Wernig M, Chang HY. Rapid chromatin switch in the direct reprogramming of fibroblasts to neurons. *Cell Rep* 2017;20:3236–47.
- [104] Guo C, Morris SA. Engineering cell identity: establishing new gene regulatory and chromatin landscapes. *Curr Opin Genet Dev* 2017;46:50–7.
- [105] Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, Giresi PG, Ng YH, Marro S, Neff NF, Drechsel D, Martynoga B, Castro DS, Webb AE, Südhof TC, Brunet A, Guillemot F, Chang HY, Wernig M. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* 2013;155:621–35.

This page intentionally left blank

BIOINFORMATIC AND BIOSTATISTIC METHODS FOR DNA METHYLOME ANALYSIS OF OBESITY

11

Sarah Amandine Caroline Voisin

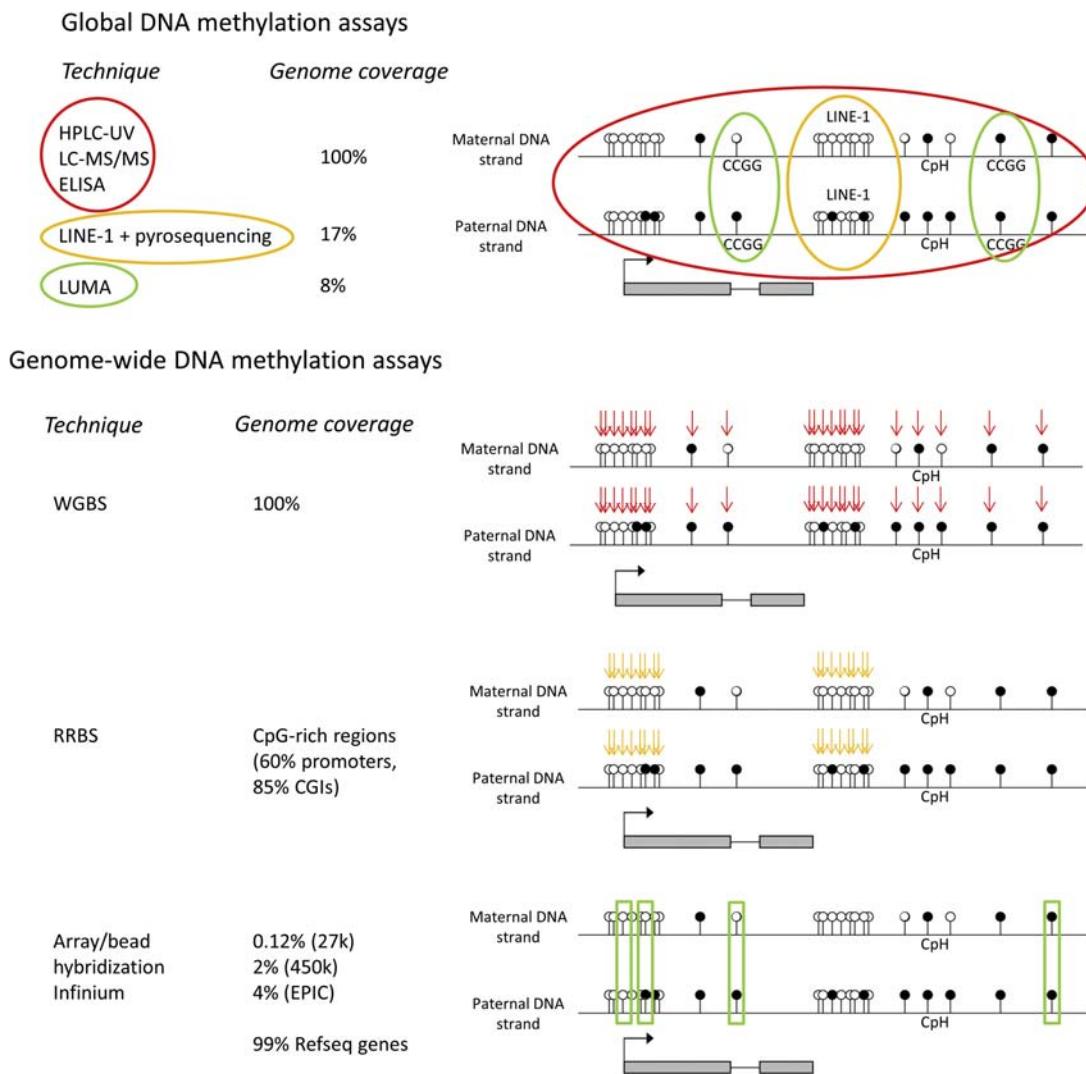
Genetics, Exercise and Performance, Institute for Health and Sport, Victoria University, Victoria, Australia

The past few years have seen a marked increase in the number of studies investigating the relationship between DNA methylation and obesity, particularly since the American Medical Association classified obesity as a disease [1]. We have gained interesting insights into the importance of DNA methylation in the context of obesity, but we are now aware of the biological, bioinformatical, and statistical limitations of these studies. By nature, epigenomic studies, especially the ones performed at the genome-wide level, combine many of the difficulties of genetic studies, with difficulties inherent to epigenomics [2–5]. This chapter is based on my doctoral thesis [6]. Some of the questions may arise when researchers plan to analyze their wet-lab data.

WHICH DNA METHYLATION ASSESSMENT TECHNIQUE SHOULD I USE?

The information generated by methylation studies is highly dependent on the technique used. More than 30 different assays have been developed using PCR, endonuclease digestion, affinity enrichment, or bisulfite conversion to look at global DNA methylation levels, region-specific DNA methylation, and genome-wide DNA methylation (spanning the entire genome) [7,8]. All of these techniques have their own advantages, caveats, and bioinformatics challenges that are important to keep in mind when interpreting results [8]. Techniques looking at global DNA methylation levels give information on the percentage of 5mC in the sample, but not on the percentage of 5mC at specific locations in the genome. On the contrary, techniques looking at genome-wide DNA methylation will give the percentage of 5mC at known locations in the genome. The genome coverage of these techniques varies greatly, from 8% to 100% for global DNA methylation assays, and from 0.12% to 100% for site-specific DNA methylation assays (Fig. 11.1).

For instance, high performance liquid chromatography ultraviolet (HPLC-UV) and whole-genome bisulfite sequencing (WGBS) have ~100% genome coverage and represent the gold standard of global and region-specific DNA methylation assessment, respectively. Another important aspect to keep in mind is that DNA samples are usually derived from a pool of different cells that may vary in their DNA methylation levels [8]. Therefore, very recent technical advances have made it possible to assess genome-wide DNA methylation levels of individual cells using single-cell bisulfite sequencing [9,10], and the bioinformatics community is currently working on tools and packages to provide the

**FIGURE 11.1**

Coverage of several global and genome-wide DNA methylation assays.

scientific community with the appropriate methods to analyze the data. Many techniques can read the DNA methylation level of a target sequence on individual DNA strands (e.g., reduced representation bisulfite sequencing (RRBS), WGBS, methylated DNA immunoprecipitation (MeDIP)-seq) (Fig. 11.1) but the more popular Infinium beadarrays read a DNA methylation level that has

been averaged over many DNA molecules (Fig. 11.1). Allele-specific DNA methylation can yield interesting information on the co-occurrence of DNA methylation on the same DNA strand, and potential insights into the function and regulation of DNA methylation at the target sequence. Finally, it should be noted that none of the aforementioned techniques allow the discrimination of 5mC from 5hmC, but there are digestion- and antibody-based techniques available [7].

The choice of technology is eventually dependent on the question under study. For instance, single-cell DNA methylation gives unique insights into the cell-specific changes of DNA methylation in heterogeneous tissues such as blood, muscle, and brain. This is important to see whether obesity-related traits affect only one cell type within a given tissue and leave others unaffected, or if they have pan-cell effects. The choice of technology is also important for groups studying genomic imprinting in relation to obesity: only techniques such as RRBS, MeDIP-seq, and WGBS that look at allele-specific DNA methylation can display DNA methylation haplotypes (i.e., DNA methylation patterns that co-occur on one DNA strand but not the other).

One of the key questions to address relates to the genomic location to investigate, which is also linked to the choice of technology. Where in the genome should we look to find relevant differentially methylated positions (DMPs) and differentially methylated regions (DMRs) in relation to obesity? Given our limited knowledge on the function of DNA methylation outside of promoters, most candidate-gene studies have focused on gene promoters, and the first genome-wide DNA methylation chip developed by Illumina in 2008 (HumanMethylation 27k beadchip) also targeted gene promoters. While these regions are known to be involved in the regulation of developmentally expressed housekeeping genes and have an important role in the pathogenesis of cancer [3], we do not know whether they are prominent genomic loci in the pathogenesis of obesity. The development of the HumanMethylation 450k beadchip by Illumina in 2011 has extended the interrogation of DNA methylation sites within gene bodies and CpG-poor promoters and between genes. The recently launched Infinium MethylationEPIC chip is another extension of this chip, which improved the coverage of regulatory elements, including 58% of FANTOM5 enhancers and 7% distal and 27% proximal ENCODE regulatory elements [11]. There are numerous R packages that have been developed specifically to analyze data generated by those chips, and a comprehensive workflow is available on the Bioconductor website [12]. Interestingly, a large proportion of DMPs and DMRs identified by EWASs for obesity traits are in the intergenic regions, open sea, and enhancers, suggesting that regulatory regions may be prominent targets in the pathogenesis of obesity [13,14].

WHICH SOFTWARE AND DATA SETS SHOULD I USE TO ANALYZE DNA METHYLATION DATA IN THE CONTEXT OF OBESITY?

Regardless of the chosen DNA methylation technique (RRBS, Illumina arrays, MeDIP-seq, Me-DIP chip, etc.), recent coordinated efforts by the bioinformatics community have made it possible to preprocess, filter, normalize, and perform all kinds of statistical analyses on DNA methylation data with the R statistical software. In 2003, the open-source, open-development Bioconductor project was launched with the goal of providing tools for the analysis and comprehension of high-throughput genomic data, using the R programming language. It proved extremely successful and is now the

leading platform for the analysis of DNA methylation data, whether in the context of obesity or in other disease contexts [15]. Among the 1473 packages that are now on the website [16], 68 include algorithms to preprocess and analyze DNA methylation data. The excellent review by Teschendorff and Relton described in detail these algorithms and software packages for downstream analyses of DNA methylation data, including algorithms for cell type deconvolution, feature selection, as well as pathway, integrative, and system-level analysis [17].

It is fair to say that there is no gold standard for the preprocessing of DNA methylation data from Illumina beadchips, but there are a few important steps that should be implemented to increase the validity of results. First, it is important to perform a logit transformation of β values, which represent the percentage of methylation at a given CpG, into M values. β values range from 0 (no allele is methylated) to 1 (all alleles are methylated) and are notoriously heteroscedastic when they are close to 0 and 1. This is a problem for most statistical tests that assume homoscedasticity, but this can be avoided by using M values, defined as $M = \log_2\left(\frac{\beta}{1-\beta}\right)$, since M values are approximately homoscedastic [18]. Most studies conduct their analyses with M values and report their results with β values, since β values have a more straightforward biological interpretation. Second, it is of prime importance to account for the two different probe designs on the Illumina HumanMethylation 450k and EPIC chips, called type I and type II designs. Specifically, β values from type II probes are less accurate and reproducible than type I probes, and show different distributions [19]. It is possible to account for this difference in probe design by either analyzing the two types of probes separately, or by normalizing the methylation values directly with peak-based correction (PBC) [19], Beta-Mixture Quantile (BMQ) normalization [20], or Regression on Correlated Probes (RCP) [21]. There are only minor performance differences between those methods, but RCP seems to outperform them all and to be computationally effective [21]. Finally, samples are often run on different plates and at different locations on the plates, introducing known batch effects that could have dramatic consequences on the downstream analysis if the sample distribution on the plates is unequal between groups. It is possible to adjust for this batch effect, either by adding both the plate number and location on the plate as covariates in the statistical analysis, or by normalizing the methylation data using empirical Bayes methods (ComBat) [22], surrogate variable analysis (SVA) [23], functional normalization [24], Remove Unwanted Variation (RUVm) [25], and BEclear [26]. ComBat is a very popular method that is easy to implement, but RUVm is particularly beneficial for the analysis of very “messy” data sets such as those that seek to combine samples from multiple labs/studies [25].

Scientists should strive to combine multiple data sets from different studies, or to replicate their results in different cohorts to strengthen their findings. DNA methylation data are not as sensitive as genetic data and are more easily shared in the scientific community. More and more journals are now asking authors to deposit their raw and processed DNA methylation data on open-access repositories before publication is accepted. The Gene Expression Omnibus (GEO) [27] and the ArrayExpress [28] platforms contain several thousand DNA methylation data sets in humans, which constitute a valuable and underexploited treasure for research groups working on DNA methylation data. For instance, a large epigenome-wide association study (EWAS) of body mass index [14] used a data set from the GEO database to perform cross-tissue correlation analyses, which led to the discovery that methylation

loci are enriched for functional genomic features in multiple tissues. It can however be challenging to obtain phenotypic data on samples deposited on such open-access platforms, as authors tend to share the minimum amount of phenotypic information when uploading data sets. It then becomes a daunting task to contact every author of every data set, and efforts should be made to make this phenotypic information more accessible.

HOW DO I ANNOTATE MY DMRs TO SPECIFIC GENES?

One issue that has probably not been sufficiently addressed has to do with the interpretation of differences in DNA methylation profiles [29]. The function of DNA methylation is highly context-dependent: when 5mC has a role in gene regulation, it does not necessarily regulate the closest gene, and the regulated gene(s) can differ depending on the tissue and the developmental window. The annotation of CpGs to certain genes and regulatory regions is therefore key to understand which pathways are regulated by those CpGs. Once a DMP or a DMR is detected between conditions (e.g., obese vs. controls), how do we know which gene(s) it regulates? Is this annotation as straightforward as often mentioned in papers? A CpG that falls into the promoter of a gene or within a gene body is often automatically annotated to the gene in question, and a CpG that falls in an intergenic region is automatically annotated to the closest gene. However, we can very well imagine that a region would serve as an enhancer in a tissue and as a promoter in another tissue. We need more information on the underlying chromatin state, long-range interactions, and transcription factor binding in specific tissues to appropriately assign CpGs to certain genes. Performing the actual experiments to obtain that information (e.g., ChIP-seq of histone modifications or transcription factors, chromosome conformation capture or chromatin interaction analysis by paired-end tag sequencing) would be preferable but extremely time-consuming and costly. Instead, it is possible to use recent data generated by large consortia such as the Encyclopedia of DNA Elements (ENCODE) Project (2003), the Functional Annotation of the Mammalian Genome 5 (FANTOM5) Project (2014) [30], and the Roadmap Epigenomics Project (2015) [31], as well as older data generated by the Genome Institute of Singapore (2011) [32]. Recent papers have used advanced statistical methods to build on the data produced by these consortia to accurately identify promoter–enhancer associations in many cell lines and tissues [33,34]. For instance, by cross-referencing chromatin states (Roadmap Epigenomics), active enhancers (FANTOM5) in relevant tissues, and long-range interaction information (Genome Institute of Singapore), it was possible to identify which genes may be impacted by those obesity-associated SNPs that are methylation QTLs [35]. Cross-referencing data from large consortia is also useful for targeted analysis of DNA methylation, by identifying which CpGs are located in probable enhancers for genes of interest, as illustrated in a recent study that selected only those CpGs that were located in known enhancers for core circadian genes [36]. Yet, while the use of data generated by big consortia definitely helps to make sense of the differential methylation signal, there are limitations that are important to consider. It is now acknowledged that the epigenome is under strong genetic control [37]. Moreover, males and females show marked differences in their epigenome and they respond differently to environmental stimuli such as social behavior, nutrition, and chemical compounds [38–40]. The reference epigenomes generated by the Roadmap Epigenomics Project [31] and the active enhancers detected by the FANTOM5 Project [30] are often derived from pools of individuals or single individuals with various ethnicities and sexes, and there is no available information on their lifelong environmental exposures. It is therefore possible that the reference chromatin states and active enhancers do not accurately reflect the chromatin states and active

enhancers present in the sample under study. Nevertheless, the interindividual variability in DNA methylation for a given tissue is typically much smaller than the intertissue variability for a given individual [3,41]. Moreover, differential DNA methylation at enhancer elements, with concurrent changes in histone modifications and transcription factor binding, is common at the cell, tissue, and individual levels, whereas promoter methylation is more prominent in reinforcing fundamental tissue identities [42]. Therefore, using reference epigenomes and reference chromatin activities in different tissues is likely useful to have a general idea of chromatin activity in a tissue of interest, especially at promoters.

A number of studies have used gene enrichment tools to find significant enrichment in certain pathways. However, the gene enrichment tools that were initially developed for transcription microarray data may not be appropriate for methylation data. For instance, using gene set enrichment analysis (GSEA) on methylation data yields biased results because of differences in the numbers of CpG sites associated with different classes of genes and gene promoters [43]. Young et al. developed a method to account for this bias in RNA-seq data [44], and the ChAMP pipeline has implemented it for DNA methylation data [45].

WHAT DOES A DIFFERENCE OF 5% IN METHYLATION MEAN?

Even if they cover 99% of RefSeq genes, the Infinium beadchips interrogate less than 5% of all CpG sites in the human genome, so human methylome changes with obesity remain largely unknown. However, the widespread use of the Infinium beadarrays has allowed a direct comparison of EWASs with one another, and allowed the systematic replication of certain hits for obesity and type 2 diabetes (T2D) in different cohorts and tissues [46]. But what if obesity actually causes methylome changes that are very small but that target a very high number of genes and pathways, leading to large effects? Now that techniques like WGBS have been developed and become increasingly cheaper, we have the ability to assess the entire methylome. The question is: do we have the appropriate statistical tools to analyze it? Studies have already had to deal with the multiple testing problem and often report only a couple of significant DMRs to keep the false positive rate at a certain level. To circumvent this issue, it is possible to restrict the analysis to certain regions, but then the global picture of the methylome is lost. It is also possible to average the methylation levels of proximal CpG sites since they are often correlated, but if only a single CpG site is important for the regulation of gene transcription, its variations with obesity would be diluted and go unnoticed. There are now R packages such as *Bumphunter* [47], *ProbeLasso* [48], and *DMRcate* [49] that make use of methylation differences at single sites to find DMRs without loss of information, and it should be very useful in future EWASs. Moreover, the reported **effect sizes** of most studies are extremely small and often close to the technical variability of the Infinium beadchips ($\sim 1\%-5\%$ DNA methylation difference). Can a sample of only a few hundred individuals suffice to find robust DNA methylation differences between groups that are that small? More importantly, what do these DNA methylation differences actually mean? Are they biologically active?

The significance of an effect is just as important as the magnitude of this effect, but this magnitude cannot be inferred simply by looking at the most common statistic reported in papers: the *P* value. Typical DNA methylation changes detected in cross-sectional or prospective studies of obesity and related comorbidities have consistently found widespread but small DNA methylation changes ($<10\%$ DNA methylation difference) between groups. This means that only a tiny proportion of cells have a DNA methylation status that differs between groups at a given location. What does this

biologically mean? It may mean that obesity and related comorbidities impact plethora of genes at a low level, which results in large disturbances of pathways important for metabolism. But can a difference of DNA methylation in only a few cells cause a significant change in transcription and eventually, in phenotype? Surprisingly few papers have addressed that question, but it is an important one [50]. Murphy et al. examined the correlation between *IGF2* transcription and the DNA methylation level of the *IGF2* imprinted region, and they suggested that a change in DNA methylation as little as 1% at this DMR can lead to either a doubling or halving of transcription, depending on the direction of DNA methylation [51]. However, these estimations were based on pyrosequencing whose technical variability can go up to 5% DNA methylation difference between technical replicates [52], and using a very small sample size ($n = 41$) [51]. Not only do small effect sizes pose biological questions, but they also raise an important statistical problem. If effect sizes are that small, wouldn't we need very large sample sizes to detect DNA methylation differences that are significant at the genome-wide level? Tsai and Bell performed simulations to estimate power under the case–control and discordant MZ twin EWAS study designs, under a range of epigenetic risk effect sizes and conditions [53]. They found that to detect a mean DNA methylation difference of 7% at genome-wide significance with 80% power, 178 pairs of twins or 211 cases and 211 controls would be needed [53]. However, each study has its own specific design, often more complex than a simple case–control or twin design (with addition of covariates, repeated measures, 2×2 factorial design), which means that sample calculations for EWASs are not straightforward. We therefore recommend performing simulations that take into account each parameter of the study to estimate the required sample size. For example, a high variance in DNA methylation in cases or controls, as well as genetic and environmental variables influencing DNA methylation levels, is likely to inflate sample sizes [53]. Effect size may also be important on the clinical side, since we need biomarkers that can identify individuals at risk with high sensitivity and specificity. If DNA methylation differences between diseased individuals and controls are extremely small, it will be difficult to find a reliable biomarker, and we may need to use a combination of several biomarkers to increase sensitivity and specificity. Zeevi et al. took a brilliant machine-learning approach to identify individuals at high risk of showing high glucose peaks after eating certain foods and paved the way to personalized nutrition [54]. It would be extremely interesting to see research teams take a similar approach to test the reliability of the few epigenetic biomarkers identified to date (e.g., CpGs in *PHOSPHO1* and *ABCG1* as predictors of T2D).

HOW DO I KNOW WHETHER MY DMRs ARE A CAUSE OR A CONSEQUENCE OF OBESITY?

In observational studies, it is usually impossible to determine whether the methylated regions that correlate with obesity-related traits are a cause or simply a consequence of the metabolic disturbance. In light of what we know about the sensitivity of DNA methylation to various environments and the establishment of obesity, it is likely that some DNA methylation changes are slow, progressive, and result from accumulated repetitions of metabolic stress while others, established early by risk variants or adverse early life conditions, give an increased susceptibility to develop obesity. It is also conceivable that some of the DNA methylation changes caused by obesity would in turn confer an increased susceptibility to further develop obesity, thus fueling a vicious cycle. However, we currently have very little idea of whether this is true or not and what specific genomic regions or tissues are involved. Interventional studies, longitudinal studies, and randomized controlled trials are good designs to answer

questions of causality, but they are expensive and would need to be conducted over extended lengths of time to truly mimic the pathogenesis of obesity and to accurately reproduce human weight trajectories. Indeed, longitudinal studies, particularly in twins, have the ability to capture the dynamics of the epigenome during individual weight trajectories, to disentangle causality issues between the epigenome and metabolic disorders or mortality risk, and understand whether and how the body can memorize past metabolic disturbances [55]. A technique called Mendelian randomization that uses meQTL has been recently developed to help answer causality issues [56]. The idea behind it is to use genetic polymorphisms as instrumental variables, specifically genetic polymorphisms that influence the DNA methylation under study and that influence the obesity-related trait under study. It was successfully used to show that most DNA methylation signatures associated with obesity are actually a consequence of adiposity, rather than a cause [14]. It was also used to show that *HIF3A* methylation is also likely a consequence of obesity [46,57], and that maternal hyperglycemia is part of causal pathways influencing offspring *LEP* epigenetic regulation in newborns [58]. The rapid development of targeted epigenome editing via an adaptation of the clustered regulatory interspaced short palindromic repeat (CRISPR)-Cas system [59–61] or the transcription activator-like effector (TALE) protein [62] might allow researchers to cause site-specific DNA methylation and answer both questions of causality and effect sizes. However, it was recently demonstrated that DNA methylation is often insufficient to transcriptionally repress promoters [63], so more work is needed to determine whether those are effective strategies to answer causality questions.

HOW CAN I BE SURE THAT MY DMRs ARE NOT DUE TO DIFFERENCES IN CELL TYPE PROPORTIONS?

The human body is made of highly specialized tissues, and these tissues are themselves made of a variety of highly specialized cells. Importantly, environmental stimuli can change the relative proportion of these different cell types at various rates depending on the tissue. For instance, skeletal muscle is made of ~90%–100% muscle cells and ~1%–10% satellite cells that have a role in muscle repair [64]. Satellite cell content increases with long-term endurance and strength training [64,65] and declines after 2 weeks of bed rest [66]. Another tissue that undergoes drastic changes in cell type proportions is blood. Blood contains varying proportions of neutrophils, lymphocytes, monocytes, eosinophils, and basophils that have specific DNA methylation signatures [67]. It was recently shown that the estimated leukocyte population in whole blood changes just 160 min after a meal, and these changes in blood cell type proportion explained >99% of the detected DNA methylation changes [68]. Moreover, as we mentioned earlier, obesity is characterized by low-grade, chronic inflammation [69,70] whereby white blood cells are more abundant in the circulation [71–73] and they infiltrate the adipose tissue [70]. The issue of cell type composition of tissues is particularly worrying given the fact that reported DNA methylation changes with obesity are often <10% DNA methylation difference, in blood [74–90], liver [91,92], adipose tissue [87,90,93–96], and sperm [97]. How can we ensure that these DNA methylation changes are not an artefact and do not reflect a simple change in cell type composition?

Depending on the tissue under study, it may be directly possible to estimate the relative cell type proportions with histological techniques or flow cytometry, and this is the most reliable approach. For instance, the relative proportions of type I and type II fibers in skeletal muscle can be estimated with staining of fiber-type specific proteins during histochemistry [98], and the relative proportions of

blood cell types with flow cytometry [99]. However, flow cytometry requires a large amount of fresh blood and laborious antibody tagging [67]. Several bioinformatic techniques have been developed to capture and account for differences in cell type composition without having to resort to these laborious techniques: a method similar to regression calibration [67] and implemented in the *minfi* R package [100], FaST-LMM-EWASher [101], RefFreeEWAS [102], surrogate variable analysis (SVA) [103], and the recent ReFACTor, that is based on principal component analysis (PCA) [104]. Some of these techniques rely on reference methylomes of purified cells while others are reference-free, and they offer the scientific community a fast and easy way to detect true DNA methylation differences with high sensitivity and specificity [105]. In particular, SVA outperforms all other reference-free techniques and can account for other sources of heterogeneity between groups, such as genetic, environmental, demographic, and technical factors [103]. A last method consists in using a set of CpGs that are known to vary considerably between cell types and to perform a principal component analysis (PCA) on them; the top principal components (PCs) can then be included as covariates in the analysis to account for the cell type profile of individual samples [35,106]. However, these techniques have all been developed and used for blood, and it is unknown whether they can be applied to heterogeneous tissues other than blood. Therefore, given the immense amount of possible confounders in human EWAS, one of the aforementioned techniques should be routinely used regardless of the study design and investigated tissue, in order to increase the biological accuracy and reproducibility of analyses.

REFERENCES

- [1] Pollack A. A.M.A. Recognizes obesity as a disease. New York Times; 2013.
- [2] Heijmans BT, Mill J. Commentary: the seven plagues of epigenetic epidemiology. *Int J Epidemiol* February 23, 2012;41(1):74–8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3304528/>.
- [3] Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* August 2013;14(8):585–94.
- [4] Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. *Lancet* June 2014; 383(9933):1952–4.
- [5] Callaway E. Epigenomics starts to make its mark. *Nature* 2014;508(7494):22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24695296>.
- [6] Voisin S. Bioinformatic and biostatistic analysis of epigenetic data from humans and mice in the context of obesity and its complications. Digital comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine NV—1245. [Department of Neuroscience, Faculty of Medicine, Disciplinary Domain of Medicine and Pharmacy, Uppsala University]: Acta Universitatis Upsaliensis; 2016. Available from: <http://uu.diva-portal.org/smash/get/diva2:952297/FULLTEXT01.pdf>.
- [7] Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* 2016;5(1).
- [8] Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 2010;11. Available from: <http://doi.org/10.1038/nrg2732>.
- [9] Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* March 2017; 12(3):534–47.
- [10] Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* November 27, 2017;10(8):1386–97. Available from: <http://doi.org/10.1016/j.celrep.2015.02.001>.

- [11] Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;17(1):208. Available from: <http://doi.org/10.1186/s13059-016-1066-1>.
- [12] Maksimovic J, Phipson B, Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data. 2017. Available from: <https://www.bioconductor.org/help/workflows/methylationArrayAnalysis/#normalisation>.
- [13] Sayols-Baixeras S, Subirana I, Fernández-Sanlés A, Sentí M, Lluís-Ganella C, Marrugat J, et al. DNA methylation and obesity traits: an epigenome-wide association study. The REGICOR study. *Epigenetics* November 3, 2017;1–8. Available from: <http://doi.org/10.1080/15592294.2017.1363951>.
- [14] Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* December 21, 2016;541:81. Available from: <http://doi.org/10.1038/nature20784>.
- [15] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* February 2015;12(2):115–21.
- [16] Bioconductor. Available from: <https://www.bioconductor.org/>.
- [17] Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* November 13, 2017;19:129. Available from: <http://doi.org/10.1038/nrg.2017.86>.
- [18] Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinf* 2010;11(1):587. Available from: <http://www.biomedcentral.com/1471-2105/11/587>.
- [19] Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium methylation 450K technology. *Epigenomics* 2011;3. Available from: <http://doi.org/10.2217/epi.11.105>.
- [20] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013;29(2):189–96.
- [21] Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. *Bioinformatics* September 2016;32(17):2659–63.
- [22] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* January 2007;8(1):118–27.
- [23] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;28(6):882–3.
- [24] Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* December 2014;15(11):503. Available from: <http://doi.org/10.1186/s13059-014-0503-2>.
- [25] Maksimovic J, Gagnon-Bartsch JA, Speed TP, Oshlack A. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Res* September 18, 2015;43(16):e106. Available from: <http://doi.org/10.1093/nar/gkv526>.
- [26] Akulenko R, Merl M, Helms V. BEclear: batch effect detection and adjustment in DNA methylation data. *PLoS One* August 25, 2016;11(8):e0159921. Available from: <http://doi.org/10.1371/journal.pone.0159921>.
- [27] Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* January 2002;30(1):207–10.
- [28] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* January 2015;43(Database issue):D1113–6.
- [29] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* October 2012;13(10):705–19.
- [30] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;507(7493):455–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24670763>.

- [31] Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518. Available from: <http://doi.org/10.1038/nature14248>.
- [32] Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo ASM, et al. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Res* 2011;21(5):665–75.
- [33] Hait TA, Amar D, Shamir R, Elkon R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biol* 2018;19:56. Available from: <http://doi.org/10.1186/s13059-018-1432-2>.
- [34] Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J. Accurate promoter and enhancer identification in 127 ENCODE and Roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* 2017;12(1):e0169249.
- [35] Voisin S, Almen MS, Zheleznyakova GY, Lundberg L, Zarei S, Castillo S, et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. *Genome Med* 2015;7(1):103.
- [36] Cedernaes J, Osler ME, Voisin S, Broman J-E, Vogel H, Dickson SL, et al. Acute sleep loss induces tissue-specific epigenetic and transcriptional alterations to circadian clock genes in men. *J Clin Endocrinol Metab* July 13, 2015;100(9):E1255–61. Available from: <http://doi.org/10.1210/JC.2015-2284>.
- [37] Taudt A, Colome-Tatche M, Johannes F. Genetic sources of population epigenomic variation. *Nat Rev Genet* June 2016;17(6):319–32.
- [38] Gabory A, Attig L, Junien C. Developmental programming and epigenetics. *Am J Clin Nutr* December 2011;94(Suppl. 6):194S–52S.
- [39] Mamrut S, Avidan N, Staun-Ram E, Ginzburg E, Truffault F, Berrih-Aknin S, et al. Integrative analysis of methylome and transcriptome in human blood identifies extensive sex- and immune cell-specific differentially methylated regions. *Epigenetics* 2015;10(10):943–57.
- [40] Nugent BM, McCarthy MM. Epigenetic underpinnings of developmental sex differences in the brain. *Neuroendocrinology* 2011;93(3):150–8.
- [41] Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* November 2, 2015;10(11):1024–32. Available from: <http://doi.org/10.1080/15592294.2015.1100786>.
- [42] Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* 2013;23(9):1522–40. Available from: <http://genome.cshlp.org/content/23/9/1522%5Cn>, <http://genome.cshlp.org/content/23/9/1522.full.pdf%5Cn>, <http://www.ncbi.nlm.nih.gov/pubmed/23804400>.
- [43] Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* June 2013;29(15):1851–7.
- [44] Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11(2):R14. Available from: <http://doi.org/10.1186/gb-2010-11-2-r14>.
- [45] Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* 2014;30(3):428–30.
- [46] Richmond RC, Sharp GC, Ward ME, Fraser A, Lyttleton O, McArdle WL, et al. DNA methylation and BMI: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* May 2016;65(5):1231–44.
- [47] Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* February 2012;41(1):200–9.

- [48] Butcher LM, Beck S. Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods* January 2015;72:21–8.
- [49] Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras KV, Lord R, et al. De novo identification of differentially methylated regions in the human genome. *Epigenet Chromatin* 2015;8(1):1–16. Available from: <http://doi.org/10.1186/1756-8935-8-6>.
- [50] Ek WE, Rask-Andersen M, Johansson A. The role of DNA methylation in the pathogenesis of disease: what can epigenome-wide association studies tell? *Epigenomics* 2016;8:5–7.
- [51] Murphy SK, Adigun A, Huang Z, Overcash F, Wang F, Jirtle RL, et al. Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene* February 15, 2012;494(1):36–43. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3627389/>.
- [52] Vilahur N, Baccarelli AA, Bustamante M, Agramunt S, Byun H-M, Fernandez MF, et al. Storage conditions and stability of global DNA methylation in placental tissue. *Epigenomics* June 2013;5(3). <https://doi.org/10.2217/epi.13.29>.
- [53] Tsai P-C, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol* May 13, 2015;44(4):1429–41. Available from: <http://ije.oxfordjournals.org/content/early/2015/05/12/ije.dyv041.abstract>.
- [54] Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. *Cell* July 10, 2016;163(5):1079–94. Available from: <http://doi.org/10.1016/j.cell.2015.11.001>.
- [55] Bray MS, Loos RJF, McCaffery JM, Ling C, Franks PW, Weinstock GM, et al. NIH working group report—using genomic information to guide weight management: from universal to precision treatment. *Obesity (Silver Spring)* January 2016;24(1):14–22.
- [56] Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* February 2012;41(1):161–76.
- [57] Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Äässi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014;6736(13):1–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24630777>.
- [58] Allard C, Desgagne V, Patenaude J, Lacroix M, Guillemette L, Battista MC, et al. Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. *Epigenetics* 2015;10(4):342–51.
- [59] Vojta A, Dobrinić P, Tadić V, Boćkor L, Korać P, Julg B, et al. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res* March 2016;44(12):5615–28.
- [60] McDonald JI, Celik H, Rois LE, Fishberger G, Fowler T, Rees R, et al. Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation. *Biol Open* May 11, 2016. Available from: <http://bio.biologists.org/content/early/2016/05/03/bio.019067.abstract>.
- [61] Xu X, Tao Y, Gao X, Zhang L, Li X, Zou W, et al. A CRISPR-based approach for targeted DNA demethylation. *Cell Discov* May 3, 2016;2:16009. Available from: <http://doi.org/10.1038/celldisc.2016.9>.
- [62] Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, et al. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotech* December 2013;31(12):1137–42. Available from: <http://doi.org/10.1038/nbt.2726>.
- [63] Ford EE, Grimmer MR, Stolzenburg S, Bogdanovic O, de Mendoza A, Farnham PJ, et al. Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *bioRxiv* January 1, 2017. Available from: <http://biorxiv.org/content/early/2017/09/20/170506.abstract>.
- [64] Kadi F, Charifi N, Denis C, Lexell J, Andersen JL, Schjerling P, et al. The behaviour of satellite cells in response to exercise: what have we learned from human studies? *Pflügers Arch* 2005;451(2):319–27. Available from: <http://doi.org/10.1007/s00424-005-1406-6>.

- [65] Hoedt A, Christensen B, Nellemann B, Mikkelsen UR, Hansen M, Schjerling P, et al. Satellite cell response to erythropoietin treatment and endurance training in healthy young men. *J Physiol* 2016;594(3):727–43. Available from: <http://doi.org/10.1113/JP271333>.
- [66] Arentson-Lantz EJ, English KL, Paddon-Jones D, Fry CS. Fourteen days of bed rest induces a decline in satellite cell content and robust atrophy of skeletal muscle fibers in middle-aged adults. *J Appl Physiol* April 15, 2016;120(8):965–75. Available from: <http://jap.physiology.org/content/120/8/965.abstract>.
- [67] Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinf* 2012;13:86.
- [68] Rask-Andersen M, Bringeland N, Nilsson EK, Bandstein M, Olaya Bucaro M, Vogel H, et al. Postprandial alterations in whole-blood DNA methylation are mediated by changes in white blood cell composition. *Am J Clin Nutr* July 2016;104(2):518–25.
- [69] Gregor MF, Hotamisligil GS. Inflammatory mechanisms in obesity. *Annu Rev Immunol* 2011;29:415–45.
- [70] Esser N, Legrand-Poels S, Piette J, Scheen AJ, Paquot N. Inflammation as a link between obesity, metabolic syndrome and type 2 diabetes. *Diabetes Res Clin Pract* August 2014;105(2):141–50. Available from: <http://www.sciencedirect.com/science/article/pii/S0168822714001879>.
- [71] Fisch IR, Freedman SH. Smoking, oral contraceptives, and obesity. Effects on white blood cell count. *JAMA* November 1975;234(5):500–6.
- [72] Dixon JB, O'Brien PE. Obesity and the white blood cell count: changes with sustained weight loss. *Obes Surg* March 2006;16(3):251–7.
- [73] Xu X, Su S, Wang X, Barnes V, De Miguel C, Ownby D, et al. Obesity is associated with more activated neutrophils in African American male youth. *Int J Obes* January 2015;39(1):26–32. Available from: <http://doi.org/10.1038/ijo.2014.194>.
- [74] Carless M a, Kulkarni H, Kos MZ, Charlesworth J, Peralta JM, Göring HHH, et al. Genetic effects on DNA methylation and its potential relevance for obesity in mexican americans. *PLoS One* 2013;8(9):e73950. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3772804&tool=pmcentrez&rendertype=abstract>.
- [75] Xu X, Su S, Barnes Va, De Miguel C, Pollock J, Ownby D, et al. A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics* 2013;8(5):522–33. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3741222&tool=pmcentrez&rendertype=abstract>.
- [76] Almén MS, Nilsson EK, Jacobsson JA, Kalnina I, Klovins J, Fredriksson R, et al. Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. *Gene* 2014;548(1):61–7.
- [77] Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci Transl Med* 2010;2(49):49ra67.
- [78] Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Äüssi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* June 7, 2014;383(9933):1990–8. Available from: <http://www.sciencedirect.com/science/article/pii/S0140673613626744>.
- [79] Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, et al. Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. *Circulation* August 2014;130(7):565–72.
- [80] Almén MS, Schiöth HB, Fredriksson R, Moschonis G, Chrousos GP, Jacobsson JA, et al. Genome wide analysis reveals association of a FTO gene variant with epigenetic changes. *Genomics* 2012;99:132–7.
- [81] Al Muftah WA, Al-Shafai M, Zaghlool SB, Visconti A, Tsai P-C, Kumar P, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. *Clin Epigenetics* 2016;8:13.
- [82] Mamtani M, Kulkarni H, Dyer TD, Goring HHH, Neary JL, Cole SA, et al. Genome- and epigenome-wide association study of hypertriglyceridemic waist in Mexican American families. *Clin Epigenetics* 2016;8:6.

- [83] Mansego ML, Milagro FI, Zulet MA, Moreno-Aliaga MJ, Martinez JA. Differential DNA methylation in relation to age and health risks of obesity. *Int J Mol Sci* 2015;16(8):16816–32.
- [84] Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *Am J Hum Genet* July 2015;97(1):75–85.
- [85] Kulkarni H, Kos MZ, Neary J, Dyer TD, Kent JWJ, Gorling HHH, et al. Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Hum Mol Genet* September 2015;24(18):5330–44.
- [86] Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol* July 2015;3(7):526–34.
- [87] Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou Y-H, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* August 2015;24(15):4464–79.
- [88] Ding X, Zheng D, Fan C, Liu Z, Dong H, Lu Y, et al. Genome-wide screen of DNA methylation identifies novel markers in childhood obesity. *Gene* July 2015;566(1):74–83.
- [89] Ollikainen M, Ismail K, Gervin K, Kyllonen A, Hakkarainen A, Lundbom J, et al. Genome-wide blood DNA methylation alterations at regulatory elements and heterochromatic regions in monozygotic twins discordant for obesity and liver fat. *Clin Epigenetics* 2015;7(1):39. Available from: <http://www.clinicalepigenticsjournal.com/content/7/1/39>.
- [90] Rönn T, Volkov P, Gillberg L, Kokosar M, Perfilieva A, Jacobsen AL, et al. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Hum Mol Genet* July 1, 2015;24(13):3792–813. Available from: <http://hmg.oxfordjournals.org/content/24/13/3792.abstract>.
- [91] Kirchner H, Sinha I, Gao H, Ruby MA, Schonke M, Lindvall JM, et al. Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Mol Metab* March 2016; 5(3):171–83.
- [92] Nilsson E, Matte A, Perfilieva A, de Mello VD, Kakela P, Pihlajamaki J, et al. Epigenetic alterations in human liver from subjects with type 2 diabetes in parallel with reduced folate levels. *J Clin Endocrinol Metab* November 2015;100(11):E1491–501.
- [93] Nilsson E, Jansson PA, Perfilieva A, Volkov P, Pedersen M, Svensson MK, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* September 2014;63(9):2962–76.
- [94] Guénard F, Tchernof A, Deshaies Y, Pérusse L, Biron S, Lescelleur O, et al. Differential methylation in visceral adipose tissue of obese men discordant for metabolic disturbances. *Physiol Genomics* March 15, 2014;46(6):216–22. Available from: <http://physiolgenomics.physiology.org/content/46/6/216.abstract>.
- [95] Pietiläinen KH, Ismail K, Jarvinen E, Heinonen S, Tummers M, Bollepalli S, et al. DNA methylation and gene expression patterns in adipose tissue differ significantly within young adult monozygotic BMI-discordant twin pairs. *Int J Obes (Lond)* April 2016;40(4):654–61.
- [96] Dahlman I, Sinha I, Gao H, Brodin D, Thorell A, Ryden M, et al. The fat cell epigenetic signature in post-obese women is characterized by global hypomethylation and differential DNA methylation of adipogenesis genes. *Int J Obes (Lond)* June 2015;39(6):910–9.
- [97] Donkin I, Versteyhe S, Ingerslev LR, Qian K, Mechta M, Nordkap L, et al. Obesity and Bariatric Surgery drive epigenetic variation of spermatozoa in humans. *Cell Metab* April 4, 2016;23(2):369–78. Available from: <http://doi.org/10.1016/j.cmet.2015.11.004>.
- [98] Scott W, Stevens J, Binder-Macleod SA. Human skeletal muscle fiber type classifications. *Phys Ther* November 2001;81(11):1810–6.
- [99] Brown M, Wittwer C. Flow cytometry: principles and clinical applications in hematology. *Clin Chem* August 1, 2000;46(8):1221–9. Available from: <http://www.clinchem.org/content/46/8/1221.abstract>.

- [100] Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics* 2014;30(10):1363–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24478339>.
- [101] Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. *Nat Meth* March 2014;11(3):309–11. Available from: <http://doi.org/10.1038/nmeth.2815>.
- [102] Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* January 21, 2014;30(10):1431–9. Available from: <http://bioinformatics.oxfordjournals.org/content/early/2014/01/21/bioinformatics.btu029.abstract>.
- [103] Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 2007;3:1724–35.
- [104] Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods* March 28, 2016;13:443. Available from: <http://doi.org/10.1038/nmeth.3809>.
- [105] Kaushal A, Zhang H, Karmaus WJJ, Wang JSL. Which methods to choose to correct cell types in genome-scale blood-derived DNA methylation data? *BMC Bioinf* October 23, 2015;16(Suppl. 15):P7. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4625103/>.
- [106] Lemire M, Zaidi SHE, Ban M, Ge B, Aissi D, Germain M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun* February 26, 2015;6. Available from: <http://doi.org/10.1038/ncomms7326a>.

This page intentionally left blank

EPIGENOMICS OF DIABETES MELLITUS

12

Ivanka Dimova

Department of Medical Genetics, Medical University Sofia, Sofia, Bulgaria

Diabetes mellitus is a serious health problem, whose frequency worldwide is steadily increasing and has become a pandemic. According to the International Diabetes Federation (IDF) data by 2015, 415 million adults would have diabetes, with the number expected to increase to 642 million by 2040 [1]. Because of the systemic nature of the disease and its long-term complications, World Health Organization reported that 1 in 20 deaths are due to diabetes. Moreover, in patients aged 35–64, the mortality of the disease increases to 1 in 10 cases.

Studies by classical physiological and biochemical methods, tissue and cell culturing, as well as gene-targeting approaches in mice and natural mutations in patients shed light on the molecular causes of type 2 diabetes and contribute to a more comprehensive understanding of the molecular mechanisms in the disease. Type 2 diabetes mellitus (T2DM) is characterized by impaired insulin activity and/or abnormal insulin secretion, as the earliest abnormality is insulin resistance—a defective condition in which insulin is unable to carry out their biological effects even with plasma concentrations that are effective in healthy subjects. Insulin resistance leads to a strong decrease in glucose uptake and glycogen synthesis in peripheral tissues, as well as to defective glucose output from the liver. The impaired antilipolytic insulin action stimulates triglycerides' degradation in adipose tissue and the generation of free fatty acids, which interfere with the insulin receptor signals. Changes in the serum levels of adipokines are also part of the process of insulin resistance. Before clinical presentation of T2DM, impaired glucose-decreasing function of insulin leads to an increase in blood glucose concentration, which stimulates insulin secretion and causes hyperinsulinemia (high blood levels of insulin)—at the beginning, this is able to compensate the insulin resistance. Diabetic condition occurs when the insulin secretion cannot compensate more for insulin resistance, and there is a fasting and postprandial hyperglycemia. The prospective studies suggest that prolonged hyperglycemia associates with a risk of vascular complications—macrovasculopathy (cardiovascular complications such as myocardial infarction and stroke) and microvasculopathy (diabetic nephropathy or retinopathy).

Today, it is generally accepted that the main causes of insulin resistance in type 2 diabetes are defective postsinsulin receptor signaling pathways, resulting in all the affected metabolic functions of insulin [2]. Impaired signaling pathways involve tyrosine dephosphorylation, imbalance of serine/threonine phosphorylation or insulin receptor internalization. A number of molecules are involved in these processes, such as free fatty acids, interleukin-6, TNF-alpha, most of which are associated with adipose tissue. Some transcription factors have also been associated with insulin resistance, such as peroxisome proliferator-activated receptor gamma (PPARG) and peroxisome proliferator-activated receptor gamma coactivator 1 alpha (PGC-1 α).

The function of β cells of the pancreas have essential role in the progression of type 2 diabetes. Beta-cell gene expression defects seen in monogenic forms of diabetes (MODY) or secondary β -cell deterioration (caused by glucotoxicity, elevated free fatty acids, cytokines, and/or mitochondrial dysfunction) may be involved in the pathophysiology of type 2 diabetes. How exactly the gene expression is changing during the course of the disease continues to be the subject of a scientific research because the relationship between the metabolic and molecular changes is more than evident.

The genetic basis for developing T2DM has been recognized for a long time. The concordance of T2DM in monozygotic twins is $\sim 70.0\%$ compared with $20.0\%-30.0\%$ in dizygotic twins, and a sibling of an affected individual has about three times higher risk for developing the disorder than the general population [3]. The decades of research into the genetic causes of T2DM have culminated with a succession of large genome-wide association studies (GWAS). Despite their power and cost, they have identified genetic variants that increase T2DM risk by only $10.0\%-30.0\%$ [4–6]. The incidence of T2DM has increased dramatically over the past decades [7], which is a too short period for accumulation of considerable alterations in the human genome. Therefore, it is likely that environmental factors, such as diet and sedentary lifestyle, might play a significant role in the development of the disease.

The role of epigenetic factors in the gene–environment interactions pointed to epigenetics as a possible molecular link between environmental factors and T2DM. Previous studies have shown that epigenetic mechanisms can predispose individuals to the diabetic phenotype. Conversely, the altered homeostasis in T2DM, such as prolonged hyperglycemia, dyslipidemia, and increased oxidative stress could also cause epigenetic changes associated with the development of disease complications [8].

BASICS OF EPIGENETICS

Environmental factors, both external and internal, play an important role in determining the function and the variability of endocrine axes [9]. Most cells have a different degree of phenotypic plasticity, and on the other hand, the phenotype determined by the genotype depends on external factors [10]. This shows that nongenetic factors can also determine the variability of endocrine functions and the risk of developing a disease [9]. Epigenetic modifications are defined as inherited changes in the function of genes that occur without any change in the nucleotide sequence [11,12]. They are potentially reversible [13,14]. Inherited/sporadic epimutations or epigenetic dysregulation in the endocrine gland or its hormonal target organs may result in disease progression [9]. Mechanisms known to affect the epigenome are DNA methylation, histone modification, and aberrant expression of microRNAs (miRNAs)—Fig. 12.1 [15]. DNA methylation is the addition of a methyl group to the 5'end of the cytosine pyrimidine ring and mainly affects cytosine in cytosine phosphate guanine (CpG) dinucleotides [16]. CpG dinucleotides are poorly represented in the mammalian genome ($1\%-2\%$) but clustered in CpG islands in promoter regions of genes. Hypermethylation of these islands leads to transcriptional attenuation [9].

DNA methylation is performed via the activity of the enzyme methyltransferases. Two groups of DNA methyltransferases exist: DNMT1, which accurately restore the DNA methylation pattern in the newly synthesized strand during the replication (maintained methylation), and DNMT3a and DNMT3b, which are responsible for the alteration in DNA methylation [17]. The process of demethylation of DNA is still poorly studied; an overview on that matter has been carried out by Patra et al. [18]. Methylation targets in the vertebrates are the cytosine residues in the CG dinucleotides of DNA,

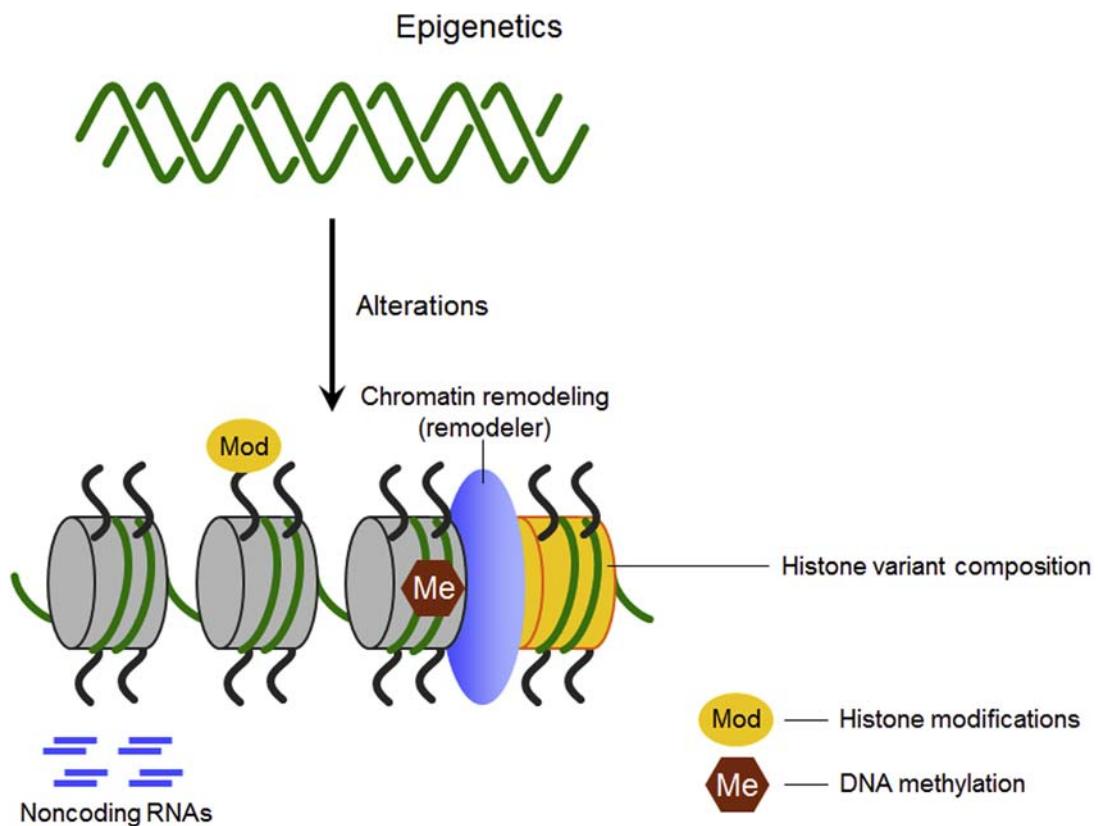


FIGURE 12.1

Basic epigenetic modifications.

as their methylation results in the suppression of the gene transcription. The mechanism of suppression is related to the specific connection of certain proteins such as MBD2 to the methylated sites, which attracts the histone deacetylases and other corepressors. It is well known that histone acetylation is of crucial importance for the active transcription. In such case, the reverse process of deacetylation, driven by the DNA methylation, will suppress the gene expression.

Histones are special proteins that help in packing the DNA into nucleosomes. Posttranslational modifications of histones are performed at their N-terminal end, and this determines the accessibility of DNA for transcription. Specific changes in histones are associated with the activation or suppression of gene activity [19]. Histones are modified by specific enzymes including histone acetyltransferases (HATs), deacetylases, and methyltransferases [20]. In most cases, histone changes occur along with DNA methylation [21].

miRNAs are a class of small noncoding RNAs and act as posttranslational regulators of gene expression [22]. They bind to target mRNA and inhibit translation of proteins [23]. Thus, one miRNA can bind to several mRNAs, and one mRNA can be regulated by multiple miRNAs. miRNAs are small

noncoding RNAs that bind to the 3' untranslated region of target mRNAs and downregulate their translation to protein or degrade the mRNAs. miRNAs play critical roles in many different cellular processes, including metabolism, apoptosis, differentiation, and development (Kim VN et al., 2006; Bentwich I. et al., 2005). miRNAs are processed sequentially from primary miRNA transcripts to pre-miRNAs to mature miRNAs—[Fig. 12.2](#). The primary miRNAs are transcribed by RNA polymerase II in the nucleus and are usually several kilobases in length. The primary miRNAs processed by Drosa/DGCR8/Pasha “microprocessor” are cleaved into ~70 to 100 nucleotide hairpin pre-miRNAs, which are rapidly exported into cytoplasma. Subsequently, pre-miRNAs are cleaved by cytoplasmic RNase

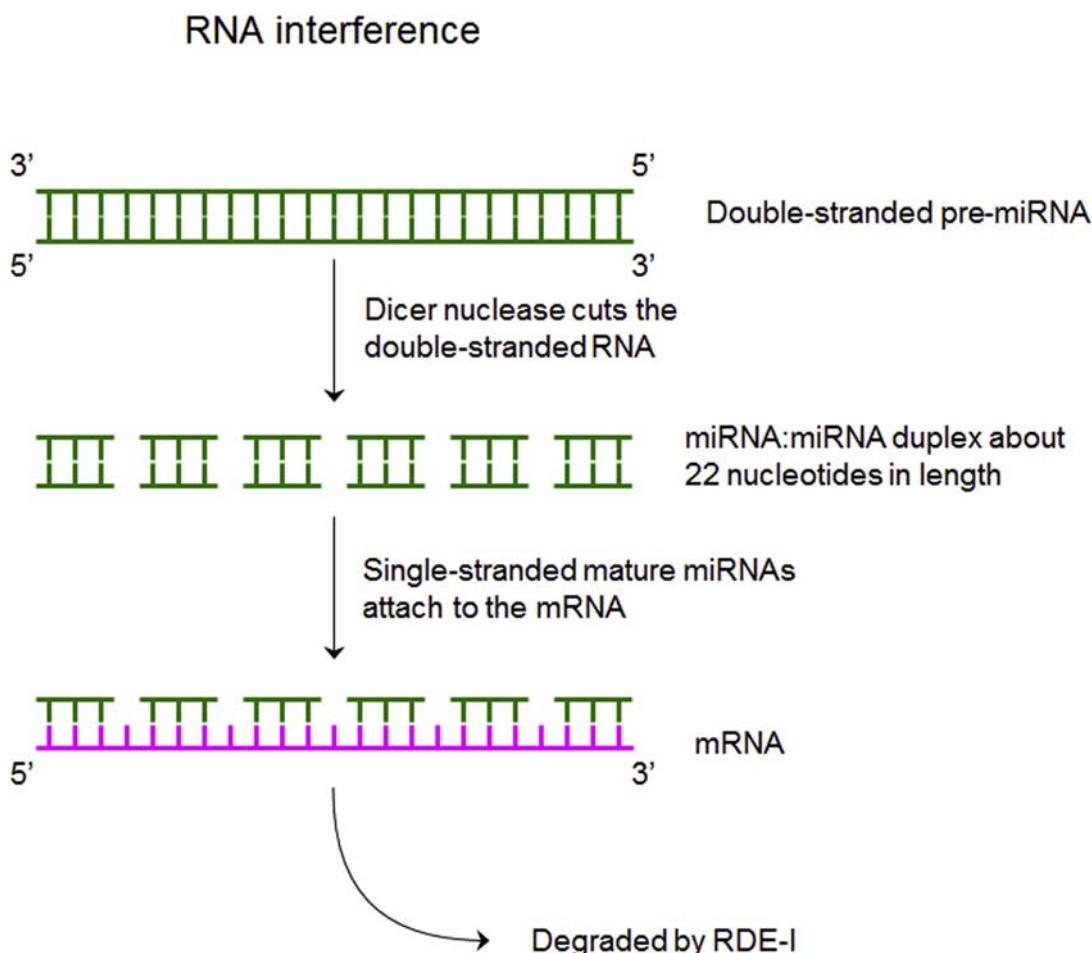


FIGURE 12.2

RNA-induced silencing complex—initially double-stranded pre-miRNA is formed, then mature miRNA is part of an active RNA-induced silencing complex (RISC) containing Dicer and many other associated proteins. Only one strand of the duplex is viable and become functional miRNA that target different mRNA populations.

III Dicer enzyme into ~22 nucleotide miRNA duplex. Based on the last miRBase release 16, more than 1110 human miRNAs were registered (<http://microrna.sanger.ac.uk/>) so far. miRNAs have important role in regulating the expression of signaling molecules, such as cytokines, growth factors, proapoptotic, and antiapoptotic genes.

Epigenetic regulation of gene expression is a fine mechanism, and its disturbance can lead to pathological conditions. Much of the variability of endocrine functions and susceptibility to endocrine disorders are determined by epigenetics.

EPIGENETIC REGULATION IN TYPE 2 DIABETES MELLITUS

T2DM is a multifactorial metabolic disease, which is influenced by both genetic and environmental factors. In recent years, GWAS have identified and confirmed a number of genetic variants for predisposition to type 2 diabetes. These findings, however, do not explain more than 10% of the total genetic risk for the disease, which raises the need for new studies to clarify the relationship between genetics and clinical manifestations of diabetes [24–26]. One of the reasons for the lack of information on the heredity of diabetes is the fact that epigenetic factors are most likely to be involved in the complex interaction between genes and the environment. Studies on epigenetic factors in the genesis of the disease are still not sufficient. Therefore, epigenetic studies could shed light on the understanding of the pathogenesis of type 2 diabetes [8,27,28]. Epigenetics is generally defined as inheritable changes in the gene function, which are not due to changes in the primary structure of DNA. The modifications include DNA methylation, histone transformations, and miRNAs and may explain why cells with identical DNA can acquire a different phenotype and differentiate into different types. One of the most commonly used indicators of epigenetic effects is the analysis for levels of DNA methylation at CpG sites in the genes [29]. It is assumed that variations in DNA methylation determine individual susceptibility to type 2 diabetes [8].

Insulin resistance is an example of epigenetic dysregulation. In this condition, normal or elevated insulin levels are insufficient for normal glycemic response of the target organs—liver, skeletal muscle, and fat, and it is observed in patients with prediabetes and type 2 diabetes. Insulin resistance in the liver is an important characteristic of both type 2 diabetes and aging. The glucokinase is a key enzyme for glucose absorption in the liver, and its activity decreases in the liver of diabetic patients [30]. Mutations in the glucokinase gene (*GCK*) can cause a monogenic form of diabetes (maturity-onset diabetes of the young [MODY]). Besides, the comparison of the liver of old and young rats demonstrated reduced levels of glucokinase expression and enzyme activity together with an increased DNA methylation of the glucokinase promoter. When culturing hepatocytes of old rats along with chemical demethylation of DNA, significant increase in the glucokinase expression was observed, which presupposes the important role of DNA methylation in the age-dependent regulation of this gene.

It was found that the diabetes affects the epigenetic status of some genes. The peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PPARGC1A) coordinates the mitochondrial oxidative metabolism in many tissues [31]. The DNA methylation in the *PPARGC1A* promoter is increased in the pancreatic islets in patients with type 2 diabetes compared with controls, and the expression of *PPARGC1A* was concomitantly reduced [32]. The expression of *PPARGC1A* positively correlates with the glucose-stimulated insulin secretion of the pancreatic islets [32], suggesting that epigenetic mechanisms can regulate the gene expression and hence the insulin secretion in the pancreas.

The environment certainly affects the epigenetic processes. The factors of the environment, which take part in the pathogenesis of the diabetes are well known—obesity, reduced physical activity, specific nutrition, and age. Age is an important factor, which increases the risk of type 2 diabetes. At the same time, the oxidation capacity and the mitochondrial function decrease both with the age and for patients with diabetes [33–38]. The mechanisms of these processes can be affected both by genetic factors and the environment [39–43]. Literature data indicate that aging of an individual changes the epigenetic status for key genes of the respiratory chain [44,45]. Cytochrome *c* oxidase polypeptide 7A1 (COX7A1), a part of complex 4 of the respiratory chain, which shows reduced expression in the muscles of patients with diabetes, is a target of the age-dependent DNA methylation [45,46]. The DNA methylation of *COX7A1* promoter increased, which negatively affected the *COX7A1* gene expression [45]. At the same time, it was found that the expression of *COX7A1* is associated with increased glucose assimilation [45]. These data demonstrate how the age affects the DNA methylation, the gene expression, and finally the metabolism. The interaction between nongenetic and epigenetic events may be affected by genetic factors. Thus, for instance, the introduction of polymorphism associated with DNA methylation site, CG dinucleotide, in the promoter of the gene for NADH:ubiquinone oxidoreductase subunit B6 (*NDUFB6*) results in increased DNA methylation—decreased gene expression—decreased metabolism with age progressing. This is an example for interaction between genetic (polymorphisms), epigenetic (DNA methylation), and nongenetic (age) factors in determining human metabolism.

Because epigenetic regulation of gene expression has tissue-specific manner, it is important to study the epigenetic changes of the main organ involved in the disease development—the pancreas. Several groups investigated DNA methylation pattern in pancreatic islets, using whole-genome DNA methylation analysis [47] identified 102 differentially methylated genes with concomitant differential expression, including genes for cyclin-dependent kinase inhibitor 1A (*CDKN1A*), phosphodiesterase 7B (*PDE7B*), Septin 9 (*SEPT9*), and exocyst complex component 3-like 2 (*EXOC3L2*), in T2DM islets compared with controls. They demonstrated by functional analyses that transcriptional suppression of *CDKN1A*, *PDE7B*, and *SEPT9* perturb insulin and glucagon secretion in clonal β and α cells, respectively, and *EXOC3L2* silencing reduced exocytosis. The authors provide with evidence that DNA methylation plays an important role in the pathogenesis and progression of T2DM. In this study, after bisulfite conversion, the authors used Infinium HumanMethylation450 BeadChip (Illumina, Inc., San Diego, CA, USA), which interrogates 482,421 CpG sites, 3091 non-CpG sites, and 65 random SNPs and covers 21,231 RefSeq genes. Methylation score for each analyzed site was calculated as a β value by using the fluorescence intensity ratio: β value = intensity of the methylated allele (M)/(intensity of the unmethylated allele (U)+intensity of the methylated allele (M)+100). DNA methylation β values range from 0 (completely unmethylated) to 1 (completely methylated).

Similar whole-genome DNA methylation study in pancreatic islets was performed by Volkov et al., whereby 457 genes with both methylation and significant expression changes were identified [48]. Here, the most sophisticated next-generation sequencing technology was applied, by means of Illumina HiSeq2500 using 125 bp long-paired end reads of Illumina type 4 chemistry. Using the Bismark methylation calling tool, the methylation value for a particular cytosine was calculated as number of reads that detect this cytosine in a methylated state divided by the total number of reads for this cytosine. Methylation profiles were then smoothed and differentially methylated regions (DMRs) were called using the BSsmooth algorithm from Bioconductor bsseq package. Among DMRs, not only genes with known islets function were found such as genes for pancreatic and duodenal homeobox 1

(*PDX1*), transcription factor 7 like 2 (*TCF7L2*), and adenylate cyclase 5 (*ADCY5*) but also new differentially methylated genes with functional influence on insulin secretion were revealed—such as genes for nuclear receptor subfamily 4 group A member 3 (*NR4A3*), parkin (*PARK2*), phosphotyrosine interaction domain containing 1 (*PID1*), solute carrier family 2 member 2 (*SLC2A2*), and suppressor of cytokine signaling 2 (*SOCS2*). Thus, epigenetic dysregulation provided with new information about the molecular regulation of pancreatic islets function.

Most of the studies reported *PDX1* as one of the most promising epigenetic islet markers—it is a transcription factor important for pancreas development and β-cell maturation and function [46,49,50]. Positive association was found between its epigenetically silenced mRNA levels and mRNA levels of insulin, as well as with the levels of HbA1c, suggesting epigenetic alterations in line with prolonged hyperglycemia. Indeed, looking for pancreatic epigenetic marker is an important step toward discovering of new molecular targets for treatment. However, the most useful clinical biomarker should be available in readily accessible tissues or cells, such as blood samples. There are several case-control studies for epigenetic alterations in candidate genes for epigenetic regulation in diabetes performed in blood samples [51–55]. DMRs were found in blood from patients with type 2 diabetes compared with controls in the genes for Centaurin-Delta-2 (*CENTD2*), alpha-ketoglutarate-dependent dioxygenase (*FTO*), potassium voltage-gated channel subfamily J member 11 (*KCNJ11*), transcription factor 7 like 2 (*TCF7L2*), and wolframin ER transmembrane glycoprotein (*WFS1*). The changes were found to appear prior the disease, suggesting the role of epigenetics in the initiation of T2DM and the possible predictive role of these epigenetic changes. One promising epigenetic blood marker for T2DM suggested from several studies is insulin-like growth factor binding protein 1 (*IGFBP1*) [56,57]. It had high levels of DNA methylation together with lower protein levels, even before the onset of the disease.

Essential role in the pathogenesis of T2DM plays the insulin resistance, especially in the liver and peripheral tissues such as skeletal muscle and adipose tissue. Whole-genome methylation analysis was performed in skeletal muscles of individuals with and without family history of T2DM, based on the fact that the disorder is associated with the reduced physical fitness. Differential methylation was observed for genes involved in MAPK, insulin, Wnt, and calcium signaling [58]. Interestingly, DNA methylation was changing in response to the 6 months exercise intervention, especially in the genes for myocyte enhancer factor 2A (*MEF2A*—exercise transcription factor), thyroid adenoma-associated protein (*THADA*—type 2 diabetes candidate gene), NADH:ubiquinone oxidoreductase subunit C2 (*NDUFC2*) (mitochondrial function), and *IL7* (cytokine). This pointed to the fact that epigenetic modifications are in strong relation to the external lifestyle factors. Genome-wide DNA methylation study in adipose tissue revealed the most differentially methylated CpG sites in *KCNQ1* and *TCF7L2*, which are very strongly associated with the disease [59]. Fat tissue analysis in monozygotic twins discordant for T2DM revealed the most significant methylation changes in the type 2 diabetes candidate genes *KCNQ1*, *NOTCH2*, *TCF7L2*, and *THADA* [60].

EPIGENETICS IN VASCULAR COMPLICATIONS OF TYPE 2 DIABETES MELLITUS

One of the most important events in the progression of diabetes is vascular inflammation accompanied by increased expression of genes for inflammation. Increased oxidative stress, dyslipidemia, and hyperglycemia were cited as factors in the development of diabetic complications. Recent studies

suggest that hyperglycemia may induce epigenetic modifications of genes involved in vascular inflammation. Recent publication in Journal of Clinical Investigation showed that DNA methylation plays a crucial role in endothelial gene expression changes, which induce atherosclerosis [61].

Nuclear factor-B (NF-Kb) is a transcription factor that regulates the expression of genes involved in inflammatory processes, including atherosclerosis and diabetes complications [62]. Poor glycemic control increases the activity of NF-Kb in monocytes and hence gene expression of inflammatory cytokines [63]. The regulation includes interaction between NF-Kb and HATs (e.g., CBP/p300), which leads to hyperacetylation of target genes, such as *TNF- α* and cyclooxygenase-2 [62]. Histone H3-lysine 4 methyltransferase SET7/9 also moves NF-Kb p65 subunit to gene promoters and thus it regulates proinflammatory genes [64]. Vascular smooth muscle cells of diabetic mice showed reduced levels of histone-H3-lysine-9 trimethylation (H3K9me3), and increased levels of histone-H3-lysine 4-dimethylation (H3K4me2) in the promoters of inflammatory genes, that is, *IL-6* and *MCP-1*, together with reduced levels of H3K9me3 in methyltransferase Suv39h1 and the histone demethylase—lysine-specific demethylase 1 (*LSD1*) [65,66]. The overexpression of Suv39h1 in vascular smooth muscle cells of diabetic mice improved the diabetic phenotype and the gene suppression of *SUV39H1* in normal human vascular smooth muscle cells increase the expression of inflammatory genes [66]. *NF-Kb* and *IL-6* represent genes with modified histone-H3-lysine-9 dimethylation in lymphocytes of patients with type 1 diabetes [67]. These observations suggest that hyperglycemia can induce epigenetic changes of proinflammatory genes associated with changes in gene expression and subsequent vascular inflammation. At the same time, it was found that a good glycemic control for 3–5 years in diabetic patients did not reduce the risk of vascular complications [68,69]. One possible explanation is that the effect of hyperglycemia can be long-lasting and the epigenetic modifications, induced by hyperglycemia, may persist for more than 5 years. Mean levels of glucose, measured in the course of the disease, can explain only a portion of the variation in the risk of developing diabetic complications. It was suggested that temporary exposure to hyperglycemia can induce permanent epigenetic changes and changes in NF- κ B-regulated gene expression and increase the risk of vascular complications in the long time period [70]. For example, the temporary exposure to hyperglycemia (16 h) induced epigenetic changes in the promoter of *NF-Kb* p65 subunit and, therefore, in p65 expression and NF-Kb activity in aortic endothelial cells. These changes persisted for 6 days despite culturing with normal glucose levels. It has been further shown that it is possible both histone methylase (SET7) and histone demethylase (*LSD1*) to regulate epigenetic changes in the promoter of *NF-Kb* p65, induced by transient hyperglycemia [71]. In fact, epigenetic modifications caused by transient hyperglycemia can explain hyperglycemic memory, discussed in many epidemiological studies. In the future, drugs affecting epigenetic mechanisms such as HDAC inhibitors could be applied for the treatment of diabetic complications [72,73]. In support of this idea, a recent study shows that myocardial infarction and ischemia induce HDAC activity in the heart [74]. The use of HDAC inhibitors in chemical myocardial infarction reduces infarction area and cell death [74].

EPIGENETICS AND CANCER DEVELOPMENT IN TYPE 2 DIABETES MELLITUS

The epidemiological data represent a significantly increased risk of various forms of cancer in patients with diabetes. Type 2 diabetes and cancer have many common risk factors, but the potential biological

link between the two socially significant diseases has not been studied. When examined at the cellular level, both diabetes and cancer are genetic diseases, caused by altered gene expression program. DNA methylation is associated with cancer development. Although aging is associated with a gene-specific hypermethylation, many tissues in mammals demonstrate global DNA hypomethylation and reduced expression levels of methyltransferases (DNMT1 and DNMT3a) with age [30,75–78]. There is a global hypomethylation in repeated DNA sequences, which may lead to genomic instability on aging. Advanced age is associated also with hypomethylation of specific genes, such as proto-oncogenes, thereby increasing susceptibility to cancer, especially if it is combined with hypermethylation of tumor suppressor genes. Studies of the effects of aging on the genomic epigenetic profile guide and assist the understanding of the molecular mechanisms in the pathogenesis and complications of type 2 diabetes. Recent data show an increased incidence of cancer in patients with type 2 diabetes—the risk is increased for liver cancer, pancreatic, lung, and uterus [79], as well as it is adverse factor for the development of colon cancer [80]. Several meta-analyses show that T2DM patients are at increased risk from cancer, as follows: liver cancer 2.5 times higher [81]; endometrial 2.1 [82]; pancreatic 1.82 [83]; urinary bladder 1.43 [84]; kidney 1.42 [85]; colorectal 1.3 [86], and breast cancer 1.2 times higher risk [87]. The data are mainly epidemiological and histological, and they do not explain the causes and molecular mechanisms. One possible explanation is the influence of epigenetic modifications in genes is important for oncogenesis.

Recently, we have performed analysis for promoter methylation of 8 tumor suppressor genes (*ATM*, *BRCA1*, *CDKN1a*, *Mlh1*, *Msh2*, *Rara*, *Tp53*, *Xpc*) in blood samples of patients with T2DM compared with controls with normal glucose tolerance. Briefly, we used Human Stress & Toxicity PathwayFinder EpiTect Methyl II Signature PCR Array (Qiagen Sciences Inc., Germantown, MD, USA). The method is based on detection of remaining input DNA after cleavage with a methylation-sensitive (these restriction enzymes are not able to cleave methylated cytosine residues, leaving methylated DNA intact—for example, *AatII*, *PstI*, *MspI*) and a methylation-dependent restriction enzyme (specifically cleaves DNA containing methyl cytosine—for example, *McrBC*). These enzymes digest unmethylated and methylated DNA, respectively. After digestion, the remaining DNA in each individual enzyme reaction is quantified by RT-PCR using primers that flank a promoter (gene) region of interest. The relative fractions of methylated and unmethylated DNA are subsequently determined by comparing the amount in each digest with that of a mock (no enzymes added) digest using a ΔCt method. Unmethylated results represent the fraction of input genomic DNA containing no methylated CpG sites in the amplified region of a gene. Methylated represents fraction of input genomic DNA containing two or more methylated CpG sites in the targeted region of a gene.

Because of the inverse relationship between the Ct value and the amount of DNA and because of the duplicate amount of the product in each PCR cycle, the amount of DNA in each reaction is determined as follows:

$$C_{Mo} = 2^{-C_T(M_0)}; C_{Ms} = 2^{-C_T(M_s)}; C_{Md} = 2^{-C_T(M_d)}; C_{Msd} = 2^{-C_T(M_{sd})}$$

The DNA fraction is calculated by normalizing the amount of DNA in the reaction to the amount of digestible DNA. The latter is defined as the difference between the amount of total DNA (determined by the mock reaction) and the amount of digestive-resistant DNA (as determined by the reactions with the two enzymes).

Unmethylated (UM) DNA fraction:

$$F_{UM} = \frac{C_{Md}}{C_{Mo} - C_{Msd}} = \frac{2^{-C_T(Md)}}{2^{-C_T(Mo)} - 2^{-C_T(Msd)}}$$

Hypermethylated (HM) DNA fraction:

$$F_{HM} = \frac{C_{Ms}}{C_{Mo} - C_{Msd}} = \frac{2^{-C_T(Ms)}}{2^{-C_T(Mo)} - 2^{-C_T(Msd)}}$$

Intermediate-methylated (IM) DNA fraction:

$$F_{IM} = 1 - F_{HM} - F_{UM}$$

Methylated (M) DNA fraction:

$$F_M = F_{HM} + F_{IM}$$

[Fig. 12.3](#) represents the average increase in the percentage of methylated DNA fraction in the genes analyzed between patients and controls. The highest increase (by more than 10 times) was detected for promoter methylation of *BRCA1* (increase by 18 times), *Msh2* gene (increase by 12 times), and *CDKN1a* (increase by 10 times). The first two genes predispose to breast/ovarian and colon/endometrial cancer, respectively. It is considered that there is a strong link between aberrant methylation of the *BRCA1* in white blood cells and breast cancer-related molecular changes, which indicate the potential predisposition of *BRCA1* dysmethylation carriers for developing breast cancer [88].

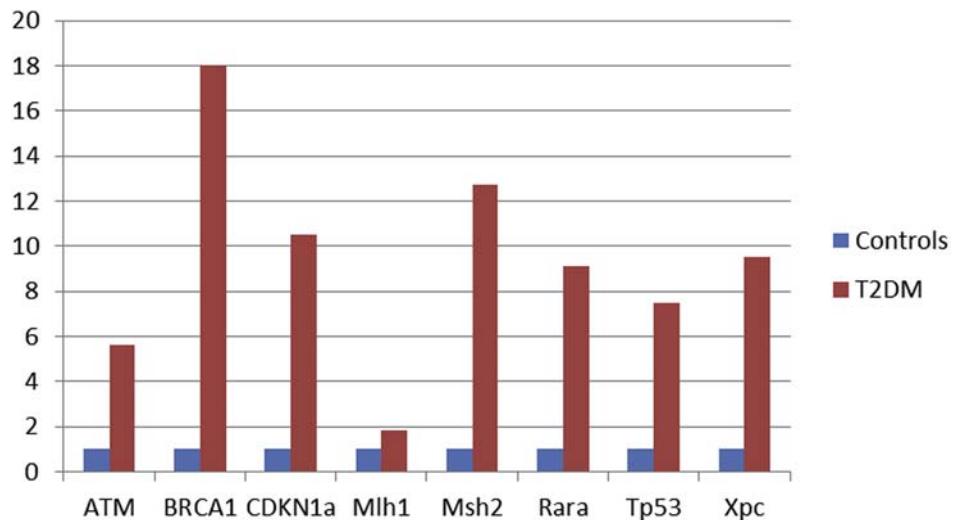


FIGURE 12.3

The average increase in the fraction of methylated DNA in promoters of 8 tumor suppressor genes between T2DM patients and controls with normal glucose tolerance.

ROLE OF MICRORNAs (miRNAs) IN TYPE 2 DIABETES MELLITUS

The miRNAs regulate gene expression through mRNA degradation or translational repression, so there would be a negative correlation between miRNA expression and its target mRNA expression level. Because of the role of miRNAs in regulating the expression of signaling molecules, such as cytokines, growth factors, proapoptotic, and antiapoptotic genes, it has been demonstrated that miRNAs can play an essential role in the pathogenesis of diabetes mellitus [89,90]. Another important issue concerns the role of miRNAs in regulating the insulin secretion by β cells of pancreas and in insulin resistance of peripheral tissues.

Functional role of miRNAs in diabetes is expressed by one of the following mechanisms: (i) **negative regulation of β -cell survival** by increasing the proapoptotic genes such as Trp53 and Bax through decreased activity of their targeting *miR-200* family [91] or decreasing the expression of antiapoptotic genes such as *Bcl-2*, targeted by *miR-34a* [92]; (ii) **inhibition of β -cell proliferation** by activation of a number of growth-inhibiting genes—for example, decreased *miR-375* expression activates Cadm1 (which negatively regulates the G1/S transition and represses cell growth) [93] and *miR-24* inhibits β -cell proliferation and insulin secretion by binding to two maturity-onset diabetes of the young genes, *Hnf1 α* and *Neurod1* [94]; (iii) **determinant in modulating insulin sensitivity**—by downregulation of insulin receptor (*INSR*) through an increasing of *miR-195* and *miR-15b* in the livers of obese T2DM model animals [95,96] or repression of *IRS-1* expression (which is the key molecule in the insulin-signaling pathway in peripheral tissues transmitting the signals from the *INSR* to the downstream enzymes) by increasing *miR-29a* expression [97], and (iv) **essential for cell differentiation**—for example, four islet-specific miRNAs (*miR-7, miR-375, miR-34a*, and *miR-146a*) exhibit distinct expression patterns during the differentiation of human embryonic stem cells into islet-like cell clusters [98,99].

A systematic study of dysregulated miRNA in T2DM reviewed 59 independent studies, selected only from human patient samples to investigate the functional involvement of miRNAs in human T2DM pathology [90]. The authors identified 158 dysregulated miRNAs in seven different major sample types. The miRNA expression has been investigated by means of quantitative PCR and/or microarray-based qPCR. The most common is the $2^{-\Delta\Delta Ct}$ method (relative analysis). This method gives information on the difference in expression levels between two samples and is calculated using the following equations:

$$\Delta Ct \text{ probe} = Ct \text{ target gene} - Ct \text{ endogenous control}$$

$$\Delta Ct \text{ control} = Ct \text{ target gene} - Ct \text{ endogenous control}$$

$$\Delta Ct \text{ sample} - \Delta Ct \text{ control} = \Delta\Delta Ct$$

$$RQ = 2^{-\Delta\Delta Ct}$$

In skeletal muscles, there were 29 miRNAs with decreased expression and 31 with increased expression compared with healthy controls. Common results were found between studies for 16 affected miRNAs in serum and plasma. There were also some common changes between skeletal muscle and whole blood, for example, *miR-100-5p*, *miR-126-3p*, and *miR-144-3p*. The mostly affected miRNA in pancreatic islets was *miR-375*. Other dysregulated miRNAs in the islets (such as *miR-7-5p*, *-369-5p*, *-129-3p*, *-136-5p*, *-187-3p*, *-589-5p*, *-224-5p*, *-655-3p*, *-495-3p*) affect the expression of *IRS1*, *IRS2*, *AKT1*, *PPARA*, *MAPK9*, *MAPK10*, *STAT3*, *PPKAG2*, *ACSL3*, and *ACSL4*, which are important genes involved in insulin signaling and type 2 diabetes pathways [100,101]. The affected miRNAs in adipose tissues were *miR-17-5p*, *-155-5p*, *-125b-5p*, *-30e-5p*, *-27a-5p*, *-221-3p*, *-199a-5p*, *-130b-3p*, *-181a-5p*, *-29a*, *-29b*, which interact with multiple transcription factors, such as *PPARs* (peroxisome proliferator-activated receptors), including *PPARG*, also known as *PPAR γ* , and adipocyte-enriched genes (*GLUT4* [also known as *SLC2A4*], *SOCS1*, *SOCS3*, *GRB2*, *INSR*, and *PPARG*), to regulate many aspects of the lipid and glucose metabolisms [102]. The dysregulated miRNAs and their interacting mRNA targets may provide new insights into the T2DM pathology and provide new disease monitoring and management tools.

In all these studies, the information from validated miRNA target databases was used together with miRNA target prediction algorithms for identifying the functional involvement of miRNAs in the progression of T2DM. The list of putative interacting targets for the miRNAs has been generated using the miRSystem (available online: <http://mirsystem.cgm.ntu.edu.tw/index.php>) Web server, which provides information from seven miRNA target prediction algorithms, DIANA-microT, miRanda, mirBridge, PicTar, PITA, RNA22, and TargetScan. Using this approach (based on validated and predicted miRNA targets), the authors performed pathway enrichment analysis, identifying various pathways associated with metabolic processes (carbohydrate and lipid metabolism), cell–cell communications (focal adhesion, tight junction), cell growth and death (apoptosis and cell cycle), signal transduction (JAK-STAT, MAPK, TGF- β , Wnt, cytokine–cytokine receptor interaction, and neurotrophin signaling), immune response (leukocyte transendothelial migration, T-cell receptor signaling, nodlike receptor signaling, toll-like receptor signaling, and chemokine signaling), insulin signaling, and type 2 diabetes signaling.

FUTURE PERSPECTIVES AND EPIGENETIC DRUGS

Much of the excitement surrounding epigenetics relates to the promise of therapies that alter the epigenetic code, activating or silencing disease-related genes. Although the majority of such treatments are still hypothetical or experimental, several epigenetic drugs that reactivate tumor suppressor genes by removing methylation marks have already received US Food and Drug Administration (FDA) approval [103,104]. These studies and therapies highlight the medical promise of mapping and understanding the role of DNA methylation.

Epigenetic modifications that modulate the activity of the genes are added and removed by epigenetic enzymes in the cells. These same enzymes also interact with a large number of activating and repressing factors, structural proteins and other enzymes, and thus regulate many cellular processes. Even a slight change in the activity of epigenetic enzymes induces drastic changes in epigenetic modifications and hence in the activity of target genes. The purpose of epigenetic therapy is to restore

normal gene regulation through adding or removing epigenetic modifications to histones and DNA or by inhibitors of miRNAs that perform regulatory function.

Because of the potential reversibility of epigenetic signals, an increasing number of epigenetic modulators (synthetic and of natural origin) pass preclinical and clinical trials. Currently, four have received the FDA approval as anticancer drugs. Two of them, which are the DNA methyltransferase inhibitors—5-azacytidine (azacitidine, Vidaza; Aza) and 5-aza-2'-deoxycytidine (Decitabine, Dacogen; DAC)—are approved for the treatment of myelodysplastic syndromes. Treating with these two preparations causes global and topical removal of the methyl groups from the DNA. This “releases” tumor suppressor genes and allows their activation. Activation of these genes leads to the initiation of multiple antitumor processes in cells, including the prevention of metastases, induction of differentiation, and cell death. The result is tumor regression.

Epigenetic drugs offer the opportunity to simultaneously target several epigenetic changes in cells. However, it is necessary to precise how this global, nonspecific effect that they have could pose risks for a potential increase in genomic instability. Limiting the negative effects of therapy will be achieved by examining how a particular epigenetic drug affects a particular type of cells and the accumulated knowledge will help determine when and how it will be used. Moreover, for the optimal effect of epigenetic intervention in the future, it is expected that the genetic and epigenetic status of the particular patient can be investigated. A personalized approach will help select the best epigenetic therapy or combine it with other drugs. The prediction software programs for miRNA–mRNA interactions, such as miRSystem, miRGator (available online: www.mirgator.kobic.re.kr/), and miR-PathDB (available online: mpd.bioinf.uni-sb.de/), as well as Web databases for expression and methylation pattern of different tissues/organs should be integrated in the global assessment of gene regulation and function.

CONCLUSION

The increasing wide prevalence of diabetes and its serious complications result in significant adverse health effects and high mortality. The insidious course of diabetes and its subtle vascular complications, along with the limited current methods of detection, prompted the need for new predictive, diagnostic, and prognostic biomarkers. The recent advances in molecular medicine provided with evidence that epigenetic factors are very important in the pathogenesis of diabetes and its complications and they interplay with genetic variants, metabolic factors, and environment. Epigenetic modifications transmit the effects of hyperglycemia on the vascular system. Using high-throughput technologies, an increasing number of differential DNA methylation and miRNA profiles have been identified for diabetes and its complications in different target tissues. Therefore, these epigenetic signatures appear as potential diagnostic and predictive biomarkers of the disease, as well as targets of treatment.

The interaction between the genome and the epigenome is a complex process. The epigenetic regulation may differ between species and tissues. Most studies so far lack validation and need larger replicative confirmation, along with standardization. Overall, the future of epigenetic diagnosis and therapy for diabetes and its complications seems promising. A better understanding of the complex mechanisms underlying deregulation of gene expression and more well-designed studies will facilitate translation into clinical practice.

REFERENCES

- [1] Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice* 2017;128:40–50.
- [2] Fernandez-Mejia C. Molecular basis of type 2 diabetes. *Mol Endocrinol* 2006;87–108.
- [3] Hari Kumar KVS, Modi KD. Twins and endocrinology. *Indian J Endocrinol Metab* 2014;18(Suppl. 1): S48–52.
- [4] Bramswig NC, Kaestner KH. Epigenetics and diabetes treatment: an unrealized promise? *Trends Endocrinol Met* 2012;23(6):286–91.
- [5] Kommoju U, Reddy BM. Genetic etiology of type 2 diabetes mellitus: a review. *Int J Diabetes Dev Ctries* 2011;31(2):51–64.
- [6] Lyssenko V, Laakso M. Genetic screening for the risk of type 2 diabetes: worthless or valuable?. *Research Support, Non-U.S. Gov't Review Diabetes Care* 2013;36(Suppl. 2):S120–6.
- [7] Zimmet P, Alberti KGMM, Shaw J. Global and societal implications of the diabetes epidemic. *Nature* 2001; 414(6865):782–7. <https://doi.org/10.1038/414782a>.
- [8] Ling C, Groop L. Epigenetics: a molecular link between environmental factors and type 2 diabetes. *Diabetes* 2009;5(12):2718–25. <https://doi.org/10.2337/db09-1003>.
- [9] Zhang X, Ho SM. Epigenetics meets endocrinology. *J Mol Endocrinol* 2011;46(1):R11–32.
- [10] Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;447(7143):433–40.
- [11] Bird A. Perceptions of epigenetics. *Nature* 2007;447(7143):396–8.
- [12] Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. *Cell* 2007;128(4):635–8.
- [13] Bannister AJ, Kouzarides T. Reversing histone methylation. *Nature* 2005;436(7054):1103–6.
- [14] Weaver IC, Champagne FA, Brown SE, Dymov S, Sharma S, Meaney MJ, et al. Reversal of maternal programming of stress responses in adult offspring through methyl supplementation: altering epigenetic marking later in life. *J Neurosci* 2005;25(47):11045–54.
- [15] Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol* 2005;45:629–56.
- [16] Ooi SK, O'Donnell AH, Bestor TH. Mammalian cytosine methylation at a glance. *J Cell Sci* 2009;122(Pt 16):2787–91.
- [17] Petersen KF, Befroy D, Dufour S, Dziura J, Ariyan C, Rothman DL, DiPietro L, Cline GW, Shulman GI. Mitochondrial dysfunction in the elderly: possible role in insulin resistance. *Science* 2003;300: 1140–2.
- [18] Patra SK, Patra A, Rizzi F, Ghosh TC, Bettuzzi S. Demethylation of (cytosine-5-C-methyl) DNA and regulation of transcription in the epigenetic pathways of cancer development. *Canc Metastasis Rev* 2008; 27:315–34.
- [19] Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu Rev Biochem* 2009;78: 273–304.
- [20] Miremadi A, Oestergaard MZ, Pharoah PD, Caldas C. Cancer genetics of epigenetic genes. *Hum Mol Genet* 2007;R28–49. 16 Spec No 1.
- [21] Kondo Y. Epigenetic cross-talk between DNA methylation and histone modifications in human cancers. *Yonsei Medical Journal* 2009;50(4):455–63.
- [22] Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. Identification of mammalian microRNA host genes and transcription units. *Genome Research* 2004;14(10A):1902–10.
- [23] Cannell IG, Kong YW, Bushell M. How do microRNAs regulate gene expression? *Biochem Soc Trans* 2008;36(Pt 6):1224–31.

- [24] Ahlqvist E, Ahluwalia TS, Groop L. Genetics of type 2 diabetes. *Clin Chem* 2011;5(2):241–54. <https://doi.org/10.1373/clinchem.2010.157016>.
- [25] Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci* 2010;5:59–77.
- [26] Imamura M, Maeda S. Genetics of type 2 diabetes: the GWAS era and future perspectives. *Endocr J* 2011;5(9):723–39. <https://doi.org/10.1507/endocrj.EJ11-0113>.
- [27] Drong AW, Lindgren CM, McCarthy MI. The genetic and epigenetic basis of type 2 diabetes and obesity. *Clin Pharmacol Ther* 2012;5(6):707–15. <https://doi.org/10.1038/clpt.2012.149>.
- [28] Kirchner H, Osler ME, Krook A, Zierath JR. Epigenetic flexibility in metabolic regulation: disease cause and prevention? *Trends Cell Biol* 2013;5(5):203–9. <https://doi.org/10.1016/j.tcb.2012.11.008>.
- [29] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;5(7271):315–22. <https://doi.org/10.1038/nature08514>.
- [30] Jiang MH, Fei J, Lan MS, Lu ZP, Liu M, Fan WW, Gao X, Lu DR. Hypermethylation of hepatic Gck promoter in ageing rats contributes to diabetogenic potential. *Diabetologia* 2008;51:1525–33.
- [31] Puigserver P, Spiegelman BM. Peroxisome proliferator-activated receptor-gamma coactivator 1 alpha (PGC-1 alpha): transcriptional coactivator and metabolic regulator. *Endocr Rev* 2003;24:78–90.
- [32] Ling C, Del Guerra S, Lupi R, Rönn T, Granhall C, Luthman H, Masiello P, Marchetti P, Groop L, Del Prato S. Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion. *Diabetologia* 2008;51:615–22.
- [33] Kelley DE, He J, Menshikova EV, Ritov VB. Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes* 2002;51:2944–50.
- [34] Ling C, Poulsen P, Carlsson E, Ridderstråle M, Almgren P, Wojtaszewski J, Beck-Nielsen H, Groop L, Vaag A. Multiple environmental and genetic factors influence skeletal muscle PGC-1alpha and PGC-1beta gene expression in twins. *J Clin Invest* 2004;114:1518–26.
- [35] Ling C, Poulsen P, Simonsson S, Rönn T, Holmkvist J, Almgren P, Hagert P, Nilsson E, Mabey AG, Nilsson P, Vaag A, Groop L. Genetic and epigenetic factors are associated with expression of respiratory chain component NDUFB6 in human skeletal muscle. *J Clin Invest* 2007a;117:3427–35.
- [36] Oootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;34:267–73.
- [37] Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, Miyazaki Y, Kohane I, Costello M, Saccone R, Landaker EJ, Goldfine AB, Mun E, DeFronzo R, Finlayson J, Kahn CR, Mandarino LJ. Co-ordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of PGC1 and NRF1. *Proc Natl Acad Sci U S A* 2003;100:8466–71.
- [38] Ritov VB, Menshikova EV, He J, Ferrell RE, Goodpaster BH, Kelley DE. Deficiency of subsarcolemmal mitochondria in obesity and type 2 diabetes. *Diabetes* 2005;54:8–14.
- [39] Bua E, Johnson J, Herbst A, Delong B, McKenzie D, Salamat S, Aiken JM. Mitochondrial DNA-deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers. *Am J Hum Genet* 2006;79:469–80.
- [40] Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Sunér D, Cigudosa JC, Urioste M, Benítez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 2005;102:10604–9.

- [41] Kujoth GC, Hiona A, Pugh TD, Someya S, Panzer K, Wohlgemuth SE, Hofer T, Seo AY, Sullivan R, Jobling WA, Morrow JD, Van Remmen H, Sedivy JM, Yamasoba T, Tanokura M, Weindruch R, Leeuwenburgh C, Prolla TA. Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* 2005;309:481–4.
- [42] Ling C, Wegner L, Andersen G, Almgren P, Hansen T, Pedersen O, Groop L, Vaag A, Poulsen P. Impact of the peroxisome proliferator activated receptor-gamma coactivator-1beta (PGC-1beta) Ala203Pro polymorphism on *in vivo* metabolism, PGC-1beta expression and fibre type composition in human skeletal muscle. *Diabetologia* 2007b;50:1615–20.
- [43] Ronn T, Poulsen P, Tuomi T, Isomaa B, Groop L, Vaag A, Ling C. Genetic variation in ATP5O is associated with skeletal muscle ATP5O mRNA expression and glucose uptake in young twins. *PLoS One* 2009;4: e4793.
- [44] Ronn T, Poulsen P, Hansson O, Holmkvist J, Almgren P, Nilsson P, Tuomi T, Isomaa B, Groop L, Vaag A, Ling C. Age influences DNA methylation and gene expression of COX7A1 in human skeletal muscle. *Diabetologia* 2008;51:1159–68.
- [45] Vaxillaire M, Froguel P. Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes. *Endocr Rev* 2008;29:254–64.
- [46] Kaneto H, Miyatsuka T, Kawamori D, et al. PDX-1 and MafA play a crucial role in pancreatic beta-cell differentiation and maintenance of mature beta-cell function. *Endocr J* 2008;55(2):235–52.
- [47] Dayeh T, Volkov P, Salö S, et al. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. In: Greally JM, editor. *PLoS genetics*, vol. 10(3); 2014. e1004160.
- [48] Volkov P, Bacos K, Ofori J, Esguerra JL, Eliasson L, Rönn T, Ling C. Whole-genome bisulfite sequencing of human pancreatic islets reveals novel differentially methylated regions in type 2 diabetes pathogenesis. *Diabetes* January 2017. <https://doi.org/10.2337/db16-0996>.
- [49] Yang BT, Dayeh TA, Kirkpatrick CL, et al. Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA(1c) levels in human pancreatic islets. *Diabetologia* 2011; 54(2):360–7.
- [50] Yang BT, Dayeh TA, Volkov PA, et al. Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with Type 2 diabetes. *Mol Endocrinol* 2012;26(7):1203–12.
- [51] Canivell S, Ruano EG, Siso-Almirall A, et al. Differential methylation of TCF7L2 promoter in peripheral blood DNA in newly diagnosed, drug-naïve patients with Type 2 diabetes. *PLoS One* 2014;9(6):e99310.
- [52] Del Rosario MC, Ossowski V, Knowler WC, Bogardus C, Baier LJ, Hanson RL. Potential epigenetic dysregulation of genes associated with MODY and Type 2 diabetes in humans exposed to a diabetic intrauterine environment: an analysis of genome-wide DNA methylation. *Metabolism* 2014;63(5):654–60.
- [53] Gu HF, Gu T, Hilding A, et al. Evaluation of IGFBP-7 DNA methylation changes and serum protein variation in Swedish subjects with and without Type 2 diabetes. *Clin Epigenet* 2013;5(1):20.
- [54] Gu T, Gu HF, Hilding A, et al. Increased DNA methylation levels of the insulin-like growth factor binding protein 1 gene are associated with Type 2 diabetes in Swedish men. *Clin Epigenet* 2013;5(1):21.
- [55] Toporoff G, Aran D, Kark JD, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Hum Mol Genet* 2012;21(2):371–83.
- [56] Gokulakrishnan K, Velmurugan K, Ganeshan S, Mohan V. Circulating levels of insulin-like growth factor binding protein-1 in relation to insulin resistance, Type 2 diabetes mellitus, and metabolic syndrome (Chennai Urban Rural Epidemiology Study 118). *Metabolism* 2012;61(1):43–6.
- [57] Petersson U, Ostgren CJ, Brudin L, Brismar K, Nilsson PM. Low levels of insulin-like growth-factor-binding protein-1 (IGFBP-1) are prospectively associated with the incidence of type 2 diabetes and impaired glucose tolerance (IGT): the Soderakra Cardiovascular Risk Factor Study. *Diabetes Metab* 2009; 35(3):198–205.

- [58] Nitert MD, Dayeh T, Volkov P, et al. Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with Type 2 diabetes. *Diabetes* 2012;61(12):3322–32.
- [59] Ronn T, Volkov P, Davegardh C, et al. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. *PLoS Genet* 2013;9(6):e1003572.
- [60] Nilsson E, Jansson PA, Perfilieva A, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* 2014;63(9):2962–76.
- [61] Dunn J, Qiu H, Kim S, Jjingo D, Hoffman R, Kim CW, Jang I, Son DJ, Kim D, Pan C, Fan Y, Jordan IK, Jo H. Flow-dependent epigenetic DNA methylation regulates endothelial gene expression and atherosclerosis. *J Clin Invest* July 1, 2014;124(7):3187–99.
- [62] Miao F, Gonzalo IG, Lanting L, Natarajan R. In vivo chromatin remodeling events leading to inflammatory gene transcription under diabetic conditions. *J Biol Chem* 2004;279:18091–7.
- [63] Shanmugam N, Reddy MA, Guha M, Natarajan R. High glucose-induced expression of proinflammatory cytokine and chemokine genes in monocytic cells. *Diabetes* 2003;52:1256–64.
- [64] Li Y, Reddy MA, Miao F, Shanmugam N, Yee JK, Hawkins D, Ren B, Natarajan R. Role of the histone H3 lysine 4 methyltransferase, SET7/9, in the regulation of NF- κ B-dependent inflammatory genes. Relevance to diabetes and inflammation. *J Biol Chem* 2008;283:26771–81.
- [65] Reddy MA, Villeneuve LM, Wang M, Lanting L, Natarajan R. Role of the lysine-specific demethylase 1 in the proinflammatory phenotype of vascular smooth muscle cells of diabetic mice. *Circ Res* 2008;103:615–23.
- [66] Villeneuve LM, Reddy MA, Lanting LL, Wang M, Meng L, Natarajan R. Epigenetic histone H3 lysine 9 methylation in metabolic memory and inflammatory phenotype of vascular smooth muscle cells in diabetes. *Proc Natl Acad Sci U S A* 2008;105:9047–52.
- [67] Miao F, Smith DD, Zhang L, Min A, Feng W, Natarajan R. Lymphocytes from patients with type 1 diabetes display a distinct profile of chromatin histone H3 lysine 9 dimethylation: an epigenetic study in diabetes. *Diabetes* 2008;57:3189–98.
- [68] Gerstein HC, Miller ME, Byington RP, Goff Jr DC, Bigger JT, Buse JB, Cushman WC, Genuth S, Ismail-Beigi F, Grimm Jr RH, Probstfield JL, Simons-Morton DG, Friedewald WT. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* 2008;358:2545–59.
- [69] MacMahon S, Chalmers J, Neal B, Billot L, Woodward M, Marre M, Cooper M, Glasziou P, Grobbee D, Hamet P, Harrap S, Heller S, Liu L, Mancia G, Mogensen CE, Pan C, Poultier N, Rodgers A, Williams B, Bompoint S, de Galan BE, Joshi R, Travert F. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2008;358:2560–72.
- [70] El-Osta A, Brasacchio D, Yao D, Pocai A, Jones PL, Roeder RG, Cooper ME, Brownlee M. Transient high glucose causes persistent epigenetic changes and altered gene expression during subsequent normoglycemia. *J Exp Med* 2008;205:2409–17.
- [71] Brasacchio D, Okabe J, Tikellis C, Balcerzyk A, George P, Baker EK, Calkin AC, Brownlee M, Cooper ME, El-Osta A. Hyperglycemia induces a dynamic cooperativity of histone methylase and demethylase enzymes associated with gene-activating epigenetic marks that coexist on the lysine tail. *Diabetes* 2009;58:1229–36.
- [72] Bieliauskas AV, Pfleum MK. Isoform-selective histone deacetylase inhibitors. *Chem Soc Rev* 2008;37:1402–13.
- [73] Szyf M. Epigenetics, DNA methylation, and chromatin modifying drugs. *Annu Rev Pharmacol Toxicol* 2009;49:243–63.
- [74] Granger A, Abdullah I, Huebner F, Stout A, Wang T, Huebner T, Epstein JA, Gruber PJ. Histone deacetylase inhibition reduces myocardial ischemia-reperfusion injury in mice. *Faseb J* 2008;22:3549–60.
- [75] Burzynski SR. Aging: gene silencing or gene activation? *Med Hypotheses* 2005;64:201–8.

- [76] Issa JP. Age-related epigenetic changes and the immune system. *Clin Immunol* 2003;109:103–8.
- [77] Youssef EM, Estecio MR, Issa JP. Methylation and regulation of expression of different retinoic acid receptor beta isoforms in human colon cancer. *Canc Biol Ther* 2004;3:82–6.
- [78] Zhang Z, Deng C, Lu Q, Richardson B. Age-dependent DNA methylation changes in the ITGAL (CD11a) promoter. *Mech Ageing Dev* 2002;123:1257–68.
- [79] Kajüter H, Geier AS, Wellmann I, Krieg V, Fricke R, Heidinger O, Hense HW. Cohort study of cancer incidence in patients with type 2 diabetes : record linkage of encrypted data from an external cohort with data from the Epidemiological Cancer Registry of North Rhine-Westphalia. Article in German *Bundesgesundheitsblatt - Gesundheitsforsch - Gesundheitsschutz* January 2014;57(1):52–9.
- [80] Sharma A, Ng H, Kumar A, Teli K, Randhawa J, Record J, Maroules M. Colorectal cancer: histopathologic differences in tumor characteristics between patients with and without diabetes. *Clin Colorectal Canc* November 14, 2013. pii: S1533-0028(13) 00120–00125.
- [81] El-Serag HB, Hampel H, Javadi F. The association between diabetes and hepatocellular carcinoma: a systematic review of epidemiologic evidence. *Clin Gastroenterol Hepatol* 2006;4(3):369–80.
- [82] Friberg E, Orsini N, Mantzoros C, Wolk A. Diabetes mellitus and risk of endometrial cancer: a meta-analysis. *Diabetologia* 2007;50(7):1365–74.
- [83] Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, Barzi F, Woodward M. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. (Meta-Analysis). *Br J Canc* 2005;92(11):2076–83.
- [84] Larsson S, Orsini N, Brismar K, Wolk A. Diabetes mellitus and risk of bladder cancer: a meta-analysis. *Diabetologia* 2006;49(12):2819–23.
- [85] Larsson S, Wolk A. Diabetes mellitus and incidence of kidney cancer: a meta-analysis of cohort studies. *Diabetologia* 2011;54(5):1013–8.
- [86] Larsson SC, Orsini N, Wolk A. Diabetes mellitus and risk of colorectal cancer: a meta-analysis. *J Natl Cancer Inst* 2005;97(22):1679–87.
- [87] Larsson SC, Mantzoros CS, Wolk A. Diabetes mellitus and risk of breast cancer: a meta analysis. *Int J Canc* 2007;121(4):856–62.
- [88] Al-Moghrabi N, Nofel A, Al-Yousef N, Madkhali S, Bin Amer SM, Alaiya A, et al. The molecular significance of methylated BRCA1 promoter in white blood cells of cancer-free females [Research Support, Non-U.S. Gov't] *BMC Canc* 2014;14:830.
- [89] Feng J, Xing W, Xie L. Regulatory roles of MicroRNAs in diabetes. In: Pichler M, editor. International journal of molecular Sciences, vol. 17(10); 2016. p. 1729. <https://doi.org/10.3390/ijms17101729>.
- [90] He Y, Ding Y, Liang B, et al. A systematic study of dysregulated MicroRNA in type 2 diabetes mellitus. In: Cho WC, editor. International journal of molecular sciences, vol. 18(3); 2017. p. 456. <https://doi.org/10.3390/ijms18030456>.
- [91] Belgardt BF, Ahmed K, Spranger M, Latreille M, Denzler R, Kondratiuk N, von Meyenn F, Villena FN, Herrmanns K, Bosco D, et al. The microRNA-200 family regulates pancreatic β-cell survival in type 2 diabetes. *Nat Med* 2015;21:619–27.
- [92] Lin X, Guan H, Huang Z, Liu J, Li H, Wei G, Cao X, Li Y. Downregulation of Bcl-2 expression by miR-34a mediates palmitate-induced Min6 cells apoptosis. *J. Diabetes Res* 2014;1729.
- [93] Zhang W, Xie HY, Ding SM, Xing CY, Chen A, Lai MC, Zhou L, Zheng SS. CADM1 regulates the G1/S transition and represses tumorigenicity through the Rb-E2F pathway in hepatocellular carcinoma. *Hepatobiliary Pancreat Dis Int* 2016;15:289–96.
- [94] Zhu Y, You W, Wang H, Li Y, Qiao N, Shi Y, Zhang C, Bleich D, Han X. MicroRNA-24/MODY gene regulatory pathway mediates pancreatic β-cell dysfunction. *Diabetes* 2013;62:3194–206.
- [95] Yang WM, Jeong HJ, Park SW, Lee W. Obesity-induced miR-15b is linked causally to the development of insulin resistance through the repression of the insulin receptor in hepatocytes. *Mol Nutr Food Res* 2015;59: 2303–14.

- [96] Yang WM, Jeong HJ, Park SY, Lee W. Induction of miR-29a by saturated fatty acids impairs insulin signaling and glucose uptake through translational repression of IRS-1 in myocytes. *FEBS Lett* 2014a;588: 2170–6.
- [97] Yang WM, Jeong HJ, Park SY, Lee W. Saturated fatty acid-induced miR-195 impairs insulin signaling and glycogen metabolism in HepG2 cells. *FEBS Lett* 2014b;588:3939–46.
- [98] Nathan G, Kredo-Russo S, Geiger T, Lenz A, Kaspi H, Hornstein E, Efrat S. MiR-375 promotes redifferentiation of adult human β -cells expanded in vitro. *PLoS One* 2015;10:1729.
- [99] Shae A, Azarpira N, Karimi MH, Soleimani M, Dehghan S. Differentiation of human-induced pluripotent stem cells into insulin-producing clusters by microRNA-7. *Exp. Clin. Transplant* 2015;16:121–8.
- [100] Gilbert ER, Liu D. Epigenetics: the missing link to understanding β -cell dysfunction in the pathogenesis of type 2 diabetes. *Epigenetics* 2012;7:841–52.
- [101] Plaisance V, Waeber G, Regazzi R, Abderrahmani A. Role of microRNAs in Islet β -cell compensation and failure during diabetes. *J. Diabetes Res.* 2014;2014:618652.
- [102] Williams MD, Mitchell GM. MicroRNAs in insulin resistance and obesity. *Exp Diabetes Res* 2012;2012: 484696.
- [103] Kaminskas E, Farrell A, Abraham S, Baird A, Hsieh L-S, Lee S-L, Leighton JK, Patel H, Rahman A, Sridhara R. Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. *Clin Canc Res* 2005;11:3604–8.
- [104] Sharma S, Kelly T, Jones P. Epigenetics in cancer. *Carcinogenesis* 2010;31:27.

This page intentionally left blank

13

EPIGENETIC PROFILING IN HEAD AND NECK CANCER

Javed Hussain Choudhury¹, Sharbadeb Kundu¹, Fazlur Rahaman Talukdar², Ruhina S. Laskar², Raima Das¹, Shaheen Laskar¹, Bishal Dhar¹, Manish Kumar¹, Sharad Ghosh³, Rosy Mondal⁴, Yashmin Choudhury¹, Sankar Kumar Ghosh^{1,5}

¹Department of Biotechnology, Assam University, Silchar, India; ²International Agency for Research on Cancer (IARC), Lyon, France; ³Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, India; ⁴Institute of Advanced Study in Science and Technology (IASST), Guwahati, India; ⁵University of Kalyani, Nadia, India

INTRODUCTION

One of the fundamental questions regarding the diversity of phenotypes within a population is why monozygotic twins or cloned animals can have different phenotypes and disease susceptibility despite their identical DNA sequences; classic genetics is unable to explain these phenomena. However, the concept of epigenetics offers a partial explanation of these phenomena. In 1939, C. H. Waddington introduced “the causal interactions between genes and their products, which bring the phenotype into being.” Later on, the term *epigenetics* was described as the study of heritable changes in gene expression without any changes in the DNA sequences. Epigenetic gene patterns play a fundamental role in diverse biological development including embryonic changes, X-chromosome inactivation, and genetic imprinting [1,2]. Unlike genetic changes, epigenetic alterations are reversible, and the key processes involved in epigenetic regulation include DNA methylation, chromatin modification (covalent alteration in core histones), nucleosome positioning, and posttranslational gene expression regulation by noncoding RNAs. Epigenetic changes occur more often than genetic mutation and may persevere for the entire cell life and even for multiple generations. Disruptions of these epigenetic processes can cause aberrant gene expression and function, which may lead to initiation, development, and progression of cancer [3].

Head and neck cancer (HNC) is a broad term that refers to a heterogeneous group of malignancies that arise in the oral cavity, larynx, pharynx, nasal cavity, and paranasal sinuses. Globally, HNC is the sixth most frequent malignancy, accounting for more than 650,000 new cases and 350,000 deaths annually [4]. The development of HNC is a multistep process modulated by genetic, epigenetic, and environmental factors. The environmental risk factors such as tobacco smoking and chewing, in addition to HPV infection, may influence a wide range of genetic and epigenetic alterations that promote genomic and epigenetic instability and endorse tumor development. Epigenetics is a bridge between genotype and phenotype, a phenomenon that changes the ultimate outcome of a genetic locus

Table 13.1 Epigenetic Alteration and Molecular Changes in Cancer Cells

Epigenetic Alterations in Head and Neck Cancer	Molecular Changes Within Cells
DNA hypermethylation	DNA hypermethylation of CpG sites within the promoters of genes promote silencing of tumor suppressor genes, resulting in genomic instability, and thus endorse proliferation and development of tumor.
DNA hypomethylation	DNA hypomethylation may activate oncogenes and transposable elements and thus causing genomic instability.
Histone acetylation	Histone acetylation promotes loss of function or gain of function of tumor-promoting genes. It also promotes defects in DNA repair and checkpoint pathways.
Histone deacetylation	Histone deacetylation restrains tumor suppressor genes and promotes genomic instability.
Histone methylation	Histone methylation promotes loss of heritable patterns of gene expression.
MicroRNAs amplification in cancer	MicroRNA amplifications function as oncogenes and result in tumor formation. Whereas, microRNA deletions may function as tumor suppressors.
Loss of imprinting (LOI)	Promotes reactivation of silent allele expression of imprinted gene, causing increase of cell proliferation.
X-chromosome inactivation	X-chromosome inactivation is age related but its alteration may result in abnormal gene dosage and promote tumor growth.

without changing the underlying DNA sequence [5,6]. The key events involved in epigenetic regulation are DNA methylation, histone modification, and gene regulation by noncoding RNAs (Table 13.1). Any disruption of these three distinct epigenetic mechanisms leads to inappropriate gene expression, resulting in and development of head and neck cancer. The explanation of how epigenetic changes can alter gene expression has led to initiate human epigenome projects and epigenetic therapies for cancers. The goal of this chapter is to review epigenetic studies done so far on HNC and the techniques currently available for epigenetic profiling of HNC.

EPIGENETIC ALTERATIONS IN CANCER

The most extensively studied epigenetic alteration is DNA methylation. DNA methylation is a covalent modification of the DNA molecule itself in which a methyl group attaches to the fifth carbon of the cytosine ring of a CpG dinucleotide by the enzyme DNA methyltransferase (DNMT). There are mainly three DNMTs, viz. DNMT1, DNMT3a, and DNMT3b. DNMT3a and DNMT3b are de novo enzymes that target unmethylated CpG island to initiate methylation; whereas, DNMT1 maintains the existing methylation patterns. The alteration in DNA methylation was the first identified epigenetic marker associated with cancer. These alterations include hypermethylation and hypomethylation [7]. DNA hypermethylation is the gain of methylation at specific sites, mainly in promoter CpG sites. CpG sites are widely distributed in CpG-rich regions of the genome known as CpG islands. These CpG islands are located upstream from the promoter region of a gene at the 5' end [8]. These CpG islands are approximately 500 base pairs in length, form more than 55% of the nucleotides, and present in the promoter regions of 40%–50% of mammalian genes, and around 45,000 CpG islands are distributed in

the human genome. Aberrant DNA methylation of CpG islands causes gene silencing and thus plays an important role in carcinogenesis. CpG-island promoter hypermethylation can affect genes mainly involved in the tumor suppressor's pathway, the DNA repair system, the metastasis-related pathway, the metabolism of carcinogens, cell-to-cell interaction, cell cycle, apoptosis, and angiogenesis. In tumor cells, mainly, global hypomethylation is escorted by hypermethylation of CpG islands promoter that usually remains unmethylated in normal cells. This special pattern of individual CpG-islands promoter hypermethylation of different tumor suppressor genes is observed in most types of human cancers [2]. The hypermethylation pattern of gene is different for each type of cancer; for example, *BRCA1* (DNA repair gene) is found to be hypermethylated in breast and ovarian cancer, but not at other sites [9]. DNA hypomethylation is the loss of DNA methylation in genome-wide regions. DNA hypomethylation found in tumors was one of the first epigenetic alterations observed in human cancer. DNA hypomethylation can assist mitotic recombination, leading to deletions and translocations, and it can promote chromosomal rearrangements. Three probable mechanisms could explain the DNA hypomethylation in cancer development: generation of chromosomal instability, reactivation of transposable elements, and loss of imprinting. The loss of methyl groups from DNA can also interrupt genomic imprinting, for example, in mice models, with a loss of imprinting of IGF2 or overall defects in imprinting have an increased risk of cancer. In numerous cancer cells, promoter regions undergo demethylation and the normally repressed genes become expressed [1,10].

Another major epigenetic event involved in carcinogenesis is histone modification, which alters chromatin structure and plays an important role in gene regulation. Histones can undergo multiple posttranslational modifications by different enzymes, such as histone acetyltransferase, histone methyltransferase, histone deacetylase (and sirtuins) and histone demethylases, kinases, phosphatases, ubiquitin ligases, deubiquitinases, sumoligases, and proteases. Histone acetylation and methylation have direct effects on a variety of nuclear processes, such as gene transcription, DNA replication, DNA repair, and the chromosome rearrangement. However, effect of histone methylation depends on the type of amino acid and its position in the histone tail [11,12]. Usually, acetylation of histone is associated with transcriptional activation, but deacetylation of histone leads to repression for transcription and hence promotes gene silencing. The combination of both acetylation and histone modification is known as the "histone code," and significant cross talk between the histone code and DNA methylation together arbitrate gene silencing. The combination of the hypoacetylated and hypermethylated histones H3 and H4 can silence tumor-suppressor genes, despite the absence of hypermethylation of the CpG island expression patterns of histone-modifying enzymes differentiating tumor tissues from their normal counterparts, and they vary according to tumor types [1,13,14].

The noncoding RNAs are also known to play a crucial role in epigenetic alteration during cancer development. The small ncRNA includes microRNA (miRNA), small interfering RNA (siRNA), small nucleolar RNA (snoRNA), and PIWI-interacting RNA (piRNA). Among these noncoding RNAs, miRNAs (short, 22-nucleotide) are most extensively studied as they are very important to normal cell physiology, and alternation in their expression has been associated with several diseases, including cancer [15]. The miRNAs regulate gene expression by sequence-specific base pairing in the 3' untranslated regions of the target mRNA. Recent studies have explained that miRNA expression profile varies between normal tissues and tumor tissues and among tumor types [16,17]. DNA hypermethylation in the 5' regulatory region of miRNA is a mechanism that can account for the downregulation of miRNA in tumors. The methylation silencing of miR-124a causes activation of the cyclin d-kinase 6 oncogene (*CDK6*), which is one of the common epigenetic events in tumor formation [18].

DNA METHYLATION PROFILING IN HEAD AND NECK CANCER

DNA methylations in promoter of tumor-related genes are likely to be playing a vital role in various cancers' development. In past decades, there has been a rapid increase of interest in promoter hypermethylation studies in various human cancers including HNC. Many studies around the globe reveal that CpG island hypermethylation in the promoter region of genes (those involved in cell cycle regulation, apoptosis, DNA repair, and detoxification pathways) are associated with cancer development and progression (Table 13.2) [19,20]. Therefore, aberrant promoter methylation of CpG islands is part of the epigenetic alteration that will promise potential molecular biomarkers for prediction and detection of head and neck cancers. Promoter hypermethylation and subsequent silencing of numerous tumor suppressor genes has been found in head and neck cancers [21,22]. In recent years, many epigenetic studies have been done, covering a broad group of tumor-associated pathway genes, including *p14*, *p15*, *p16* (cell cycle control); *DAPK*, *p73*, *RASSF1* (apoptosis); *BRCA1*, *MLH1*, *MSH2*, *MGMT* (DNA repair); *ATM*, *GSTP1* (carcinogen metabolism); *ECAD*, *CDH1*, *EDNRB* (cell-cell adhesion); and *MINTs* (methylated in tumors) loci such as *MINT1*, *MINT2*, and *MINT31*. E-cadherin (*ECAD*), a transmembrane glycoprotein, is accountable for cell–cell adhesion, the altered expression of which is highly associated with regional metastasis in OSCC [23]. The *ECAD* gene hypermethylation frequency ranged between 7% and 46% [24]. In HNC, the promoter hypermethylation of *p16*, *DAPK*, *MGMT*, and *ECAD* are a frequently observed [25–29]. According to one of the studies, the prevalence of *p16*, *DAPK*, *ECAD*, and *RASSF1A* promoter methylation in HNSCC was 32.5%, 23.8%, 36.3%, 7.5%, respectively [21]. The death associated protein kinase (*DAPK*) gene is associated with loss of apoptosis and cell immortality, and its reduced expression has been associated with metastasis in different cancers. In HNC, 27% of *DAPK* promoter hypermethylation has been observed [28]. Recent studies also reported that *DAPK* and

Table 13.2 The Epigenetic Studies (DNA Methylation) in Head and Neck Cancer and Techniques Used

Technique	Methylated Genes/Loci	Sample
MSP	<i>p16</i> , <i>DAPK</i> , <i>RASSF1</i> , <i>BRAC1</i> , <i>GSTP1</i> , <i>ECAD</i> , <i>MLH1</i> , <i>MINT1</i> , <i>MINT2</i> , and <i>MINT31</i>	Tissue
BeadChip and qMSP	<i>HOXA9</i> , <i>NID2</i> , <i>GATA4</i> , <i>KIF1A</i> , <i>EDNRB</i> , <i>DCC</i> , <i>MCAM</i> , <i>CALCA</i>	Tissue
BeadChip	<i>S100A8</i>	Tissue
MSP	<i>EBNA1</i> , <i>LMP1</i> , <i>RASSF1A</i> , <i>DAPK</i> , <i>ITGA9</i> , <i>P16</i> , <i>WNT7A</i> , <i>CHFR</i> , <i>CYB5R2</i> , <i>WIFI</i> , <i>RIZ1</i> , <i>FSTL1</i>	Tissue
qMSP	<i>TIMP3</i> , <i>DCC</i> , <i>DAPK</i> , <i>CCNA1</i> , <i>AIM1</i> , <i>MGMT</i> , <i>CDH1</i> , <i>HIC1</i>	Saliva
Pyrosequencing	<i>P16</i>	Tissue
BS	<i>CCNA1</i> , <i>DAPK</i> , <i>MGMT</i> , <i>SFRP1</i> , <i>TIMP3</i>	Tissue
BS and MSP	<i>CDK10</i>	Blood
COBRA	<i>ALU</i>	Saliva
qMSP	<i>DCC</i> , <i>EDNRB</i>	Saliva
MSP	<i>RASSF1A</i> , <i>p16</i> , <i>DAPK1</i>	Saliva
MSP	<i>P16</i> , <i>DAPK</i> , <i>RARB</i> , <i>CDH1</i> , <i>RASSF1A</i>	Tissue
MSP	<i>P16</i>	Tissue

p16 aberrant hypermethylation was associated with poor prognosis in oral cancers [30,31]. Another frequently studied hypermethylated gene is *RASSF1A*, which belongs to the Ras association family (RASSF) of proteins involved in the Ras/PI3K/AKT pathways. The methylated loci in tumors (MINT) family CpG islands are associated with tumors at several sites; however, their functions are indecisive as they are not situated near any known genes. Many studies found aberrant promoter methylation of *MINT1*, *MINT2*, and *MINT31* in HNCs [5,32,33]. The frequency of methylation of *CDKN2A* (*p16*) has been commonly investigated and reported in various cancers; it is also the mostly studied gene in HNC. The reported incidence of hypermethylation of *p16* ranges from 23% to 76% in oral cancer [24]. A recent review article demonstrated that the methylation status of *p¹⁶INK4A* acts as a promising candidate biomarker for predicting clinical outcome of oral cancer, particularly for recurrence-free survival [34]. A study reported that, in HNC, promoter methylation of *RASSF1A* was 42.9% in cell lines and 15% in primary tumors but not found in the normal control DNA. The study also observed a significant inverse association between *RASSF1A* promoter methylation and HPV infection ($P = .038$) [35]. A recent study identified *NOL4* and *IRX1* as a highly specific promoter methylated gene associated with HNC and may have potential as a biomarker for HNC [36]. Another recent study highlights the importance of assessing tumor suppression genes at the genomic and epigenomic level to identify key pathways in HNC deregulated by simultaneous promoter methylation and somatic mutations using whole-genome sequencing [37]. One of our studies found significantly high levels of hypermethylation in *p16*, *DAPK*, *ECAD*, *RASSF1*, *MINT1*, *MINT2*, and *MINT31* in HNC tumor tissues compared to normal counterparts, reflecting the possible involvement of epigenetic alteration toward the development and progression of HNC [5]. The hypermethylation profiles of gene promoters are diverse for each type of cancer, and this variation depends on the selection of detection method and use of multiple gene panels. On the basis of hypermethylation pattern of multiple tumor-related genes/loci panel, we can stratify HNC into different subgroups with distinct molecular characteristics (Fig. 13.1).

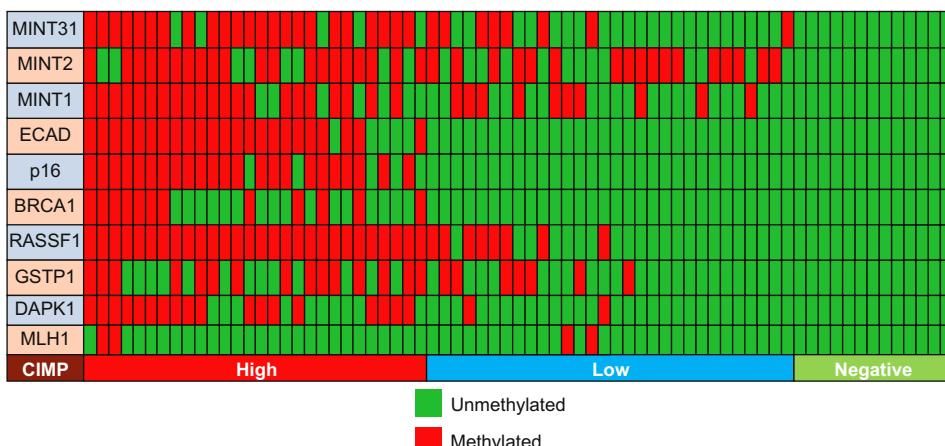


FIGURE 13.1

Each red rectangle represents methylated genes, while green rectangles represent unmethylated genes. Five or more methylated genes in an HNC tissue represents CIMP-high group, CIMP-low group is denoted by less than five methylated genes, and CIMP-negative represents no methylated genes.

TECHNIQUES AVAILABLE FOR EPIGENETIC PROFILING OF HNC

Epigenetic profiling mostly focuses on determination of DNA methylation status because DNA methylations are relatively stable and well established in cancer development. DNA methylation involves the covalent addition of a methyl group to the fifth position of cytosine within CpG sites in the promoter regions of genes. In early days, methylation profiling was restricted to determining the DNA methylation of a few genes. However, the methylation profiling was upgraded to the genome-wide level due to the development of next-generation sequencing and microarray hybridization technology. The most recent advancement in epigenomic profiling is epigenome-wide association studies (EWASs) via methylation array, whole-genome bisulfite sequencing (WGBS), reduced-representation bisulfite sequencing (RRBS), and nanopore-based single-molecule real-time sequencing technology (SMRT). All these advanced techniques allow detection of DNA methylation in real time on a large scale [20,38].

Various methods used for methylation analysis, such as methylation-specific PCR (MSP), combined bisulfite restriction analysis (COBRA), real-time qMSP or MethylLight, bisulfate sequencing, and pyrosequencing are all based on bisulfite conversion of genomic DNA (Fig 13.2). Bisulfite treatment of genomic DNA converts (by deamination) unmethylated cytosine (C) into uracil (U), while leaving methylated cytosine (C^m) unchanged. The uracil (U) finally converts to thymine (T) in a following polymerase chain reaction (PCR) [39]. This bisulfite treatment method is also used to study in EWASs via methylation array and WGBS.

METHYLATION SPECIFIC PCR

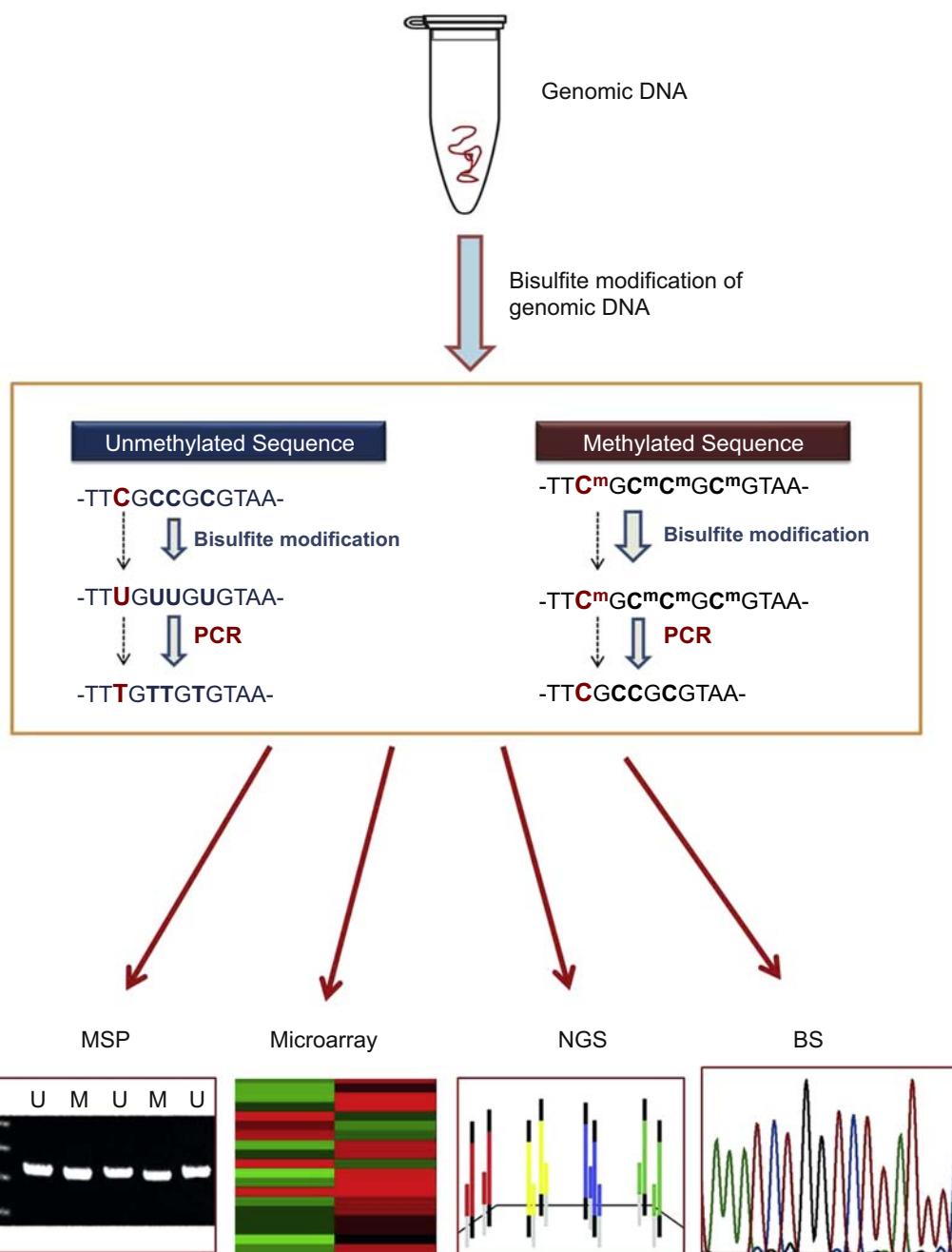
This technique is very simple and commonly used to detect methylation status of a specific gene of interest by performing PCR with specific primers for unmethylated or methylated sequences. MSP is sensitive and specific for methylation of any sites of CpG in a CpG island. Unmodified DNA or DNA incompletely reacted with bisulfite can also be distinguished, since marked sequence differences exist between these DNAs. In MSP, two separate sets of primers are designed for detection of methylation. After bisulfite treatment, the two strands of DNA are no longer complementary; therefore, a primer can be designed for either of the modified strands. PCR amplified band generated by using specific sets of primers will decide the methylation status of the DNA sample. PCR primers can be designed at any position, thus MSP has the flexibility of selecting a large genomic segment. The disadvantage of the technique is the high rate of false-negative or -positive results and requires careful determination of the number of PCR cycles performed [40].

COMBINED BISULFITE RESTRICTION ANALYSIS ASSAY

This assay is a combination of three techniques: bisulfite treatment genomic DNA, PCR amplification, and restriction digestion. COBRA assay makes it possible to analyze site-specific differences in methylation patterns. This assay has high sensitivity and very low possibility of false-positive results, compared to the other site-specific methylation techniques like MSP [41].

BISULFITE SEQUENCING

In DNA methylation studies, bisulfite sequencing (BS-Seq) is considered to be the “gold standard” technique. The sodium bisulfite treatment of DNA mediates the deamination of unmethylated cytosine

**FIGURE 13.2**

Commonly used methods for DNA methylation analysis, such as methylation specific PCR (MSP), real-time qMSP or MethylLight, bisulfate sequencing (BS), next-generation sequencing (NGS) and BeadChip (microarray) are all based on bisulfite conversion of genomic DNA.

into uracil, while methylated cytosine residues remain unaffected, and determined by subsequent PCR amplification and Sanger sequencing. Comparing the sequences of untreated DNA sample to the bisulfite treatment sample allows detection of the methylated cytosines. With the arrival of next-generation sequencing (NGS), we can extend DNA methylation study across the entire genome. However, post-NGS sequence alignment becomes a tricky job because of genome complexity created by bisulfite treatment. Thus, it is crucial to ensure complete conversion of nonmethylated cytosines, as the estimated level of DNA methylation depends on it.

PYROSEQUENCING

This technique is suitable for low-throughput projects. Primers are designed for PCR amplification and short-read pyrosequencing reaction (~100 bp). The status of methylation for each CpG region within the sequenced site is detected based on the signal intensities for incorporated dATP and dGTP. This method is quantitative and is able to detect even small variations in methylation (down to 5%). It is a good technique for heterogeneous cancer samples, where only a small fraction of cells has a differentially methylated gene [42].

WHOLE GENOME BISULFITE SEQUENCING

WGBS is analogous to whole genome sequencing, excluding bisulfite conversion. It is the most inclusive of all accessible epigenetic profiling methods. In WGBS, genomic DNAs are purified and sheared into small fragments, and then fragmented DNAs are end repaired by adding adenine bases to the 3' end (A-tailing). Next, methylated adapters are ligated to the DNA fragments. The fragments are size selected before bisulfite treatment and PCR amplification, and the resulting library is sequenced. Due to high number of PCR cycles and inappropriate selection of a uracil-insensitive DNA polymerase, overrepresentation in the methylated DNA data occurs [43]. WGBS could be performed using any existing NGS platform. The limitations of WGBS could be minimized by reduced RRBS, where only a small part of the genome is sequenced [44]. In RRBS, enhancement of CpG-rich regions is attained by isolation of short fragments after enzyme digestion that identifies CpG sites. This method isolates ~85% of CpG islands in the human genome and then use in WGBS. Also, this procedure usually requires a very small amount of DNA. The enhancement for CpG-rich regions of interest could be performed before NGS.

The nanopore-based SMRT has been recently adopted for epigenetics research. SMRT allows detecting modified bases directly by monitoring the activity of DNA polymerase during the incorporation of various fluorescently labeled bases into complementary DNA strands. WGBS has become the standard epigenetic profiling method in major epigenome consortiums, such as NIH Roadmap, IHEC, ENCODE, Blueprint, e.t.c. [38,42].

ARRAY OR BEAD HYBRIDIZATION TECHNIQUES FOR EPIGENETIC PROFILING

DNA methylation microarray will be used to identify DNA methylation at cytosine positions across the human genome. Methylated DNA fragments of the genome, generally obtained by immunoprecipitation, could be used for hybridization with microarrays. High-throughput DNA methylation profiling technology (450K BeadChip array) similar to successful GWAS is now available, which allows

measurement of DNA methylation at 485,512 cytosine positions across the human genome, of which there are 482,421 CpG sites and 3091 non-CpG sites. Illumina's Infinium HumanMethylation450 BeadChip (HM450K) kits are available, which involves the bisulfite conversion of genomic DNA and PCR amplification, followed by the hybridization of the converted DNA to arrays containing pre-designed probes to differentiate between methylated and unmethylated cytosine. This BeadChip assay mostly focused on promoter regions of genes/loci, enhancer regulatory elements, and untranslated regions (3' UTRs) where prominent DNA methylation occurs. Each HM450K BeadChip can screen more than 450,000 methylation sites [45]. So far, 450K BeadChip array dominates epigenome-wide studies investigating the cancer methylome. However, the most recently Infinium MethylationEPIC BeadChip has been introduced in epigenome-wide studies, which covers more than 850K CpG methylation sites, including more than 90% of the 450K sites plus additional CpG sites in the enhancer regions identified by the FANTOM5 and ENCODE projects [46].

ENRICHMENT-BASED METHODS

DNA-methylation-specific antibodies, methyl-binding domain proteins, or restriction enzymes are used to enrich a fraction of hypermethylated (or hypomethylated/unmethylated) DNA fragments, and the enrichment of specific fragments is quantified by next-generation sequencing. The two key advantages of enrichment-based methods are the relatively low cost of achieving genome-wide coverage and the ability to distinguish between different forms of DNA methylation, for example, using antibodies that specifically recognize 5-hydroxymethylcytosine (5 hmC) but not 5 mC. However, these advantages come at the cost of relatively low resolution and high susceptibility to experimental biases [47]. Some important examples of this method are MeDIP-seq [48], where methylated DNA can be enriched using methylation-specific antibodies (in methylated DNA immunoprecipitation coupled with high-throughput sequencing, methyl-CpG-binding domain (MBD) proteins (in MBD sequencing [MBD-seq]) [49] or a restriction enzyme that specifically cuts methylated DNA (in methylation-dependent restriction enzyme sequencing, MethylRAD) [50]). Alternatively, unmethylated DNA can also be enriched using restriction enzymes that specifically cut unmethylated DNA (for example, in HpaII tiny fragment enrichment by ligation-mediated PCR coupled with sequencing (HELP-seq) [51]).

METHYLATED DNA IMMUNOPRECIPITATION

Among all of the above-mentioned enrichment-based methods, the most popular method that is currently being used is methylated DNA immunoprecipitation (MeDIP). This method exploits anti-methylcytosine antibody to immunoprecipitate DNA among methylated CpG sites [39]. The specific region of DNA enhanced by MeDIP can be assessed by using MeDIP-chip or high-throughput MeDIP sequencing (MeDIP-seq). MeDIP-seq can be a cost-effective method when there is no requirement of single-base resolution. An advantage of MeDIP-seq is that it requires very small amount of DNA (~ 1 ng) [52,53].

COMPUTATIONAL EPIGENETICS ANALYSIS

Computational epigenetics applies various bioinformatics methods to harmonize the experimental research works in epigenetics. From its definition, we may say that computational epigenetics

encompasses the development and application of some sophisticated bioinformatics methods/tools for solving different epigenetic questions, as well as computational data analysis and theoretical modeling in the context of epigenetics, which includes modeling of the effects of histone and CpG island methylation. In simpler terms, the major goals of computational epigenetics are (1) to foster our understanding of epigenetics and disease by computational means and (2) to develop advanced bioinformatics methods for the analysis and interpretation of large epigenome datasets.

As a consequence, from the computational point of view and the characteristics of the generated data, epigenomics is a very complex field, for two main reasons. First, epigenetics takes in a multi-layered set of regulatory signals that act coordinately and possibly in a combinatorial way to control fundamental biological processes, such as the output of gene expression patterns. Second, high-throughput sequencing-based epigenetics profiling techniques are widely adopted nowadays in this field, generating widespread yet complicated and massive genome-wide datasets. As a result, the contribution of scientists with computational skills (computer scientists, statisticians, physicists, and computational biologists) is considered an essential component of research institutes investing in this research field [54,55].

BIOINFORMATICS TOOLS FOR COMPUTATIONAL EPIGENOMICS

A large amount of data pertinent to epigenetic research is currently available in different scientific literature, molecular databases/repositories, and several case reports. Usually, scientific literature serves as the primary source of data, providing high-level descriptions of biological entities and processes. A compiled list of epigenetic data browsers, repositories (databases), and tools and resources that are commonly used by epigenetic researchers can be found at <http://epigenie.com/epigenetic-tools-and-databases/>. Apart from this, there are certain databases especially related to cancer, such as **MethDB**, containing information on 19,905 DNA methylation content data and 5382 methylation patterns for 48 species, 1511 individuals, 198 tissues and cell lines, and 79 phenotypes; **PubMeth**, containing over 5000 records on methylated genes in various cancer types; and **MeInfoText**, containing gene methylation information across 205 human cancer types and others, which have been reviewed elsewhere [56]. In addition, a concise list of software used for computational epigenetics/epigenomics can be found at <http://www.computational-epigenetics.de/software.php>.

METHODS FOR ANALYZING AND INTERPRETING THE DNA METHYLATION DATA

Among the various epigenetic alterations, DNA methylation is the only epigenetic mark for which a detailed mechanism of mitotic inheritance has been described [57]. Therefore, the technologies that are now available are mostly developed for studying DNA methylation pattern genome wide, at a high resolution and in a large number of samples [47]. As mentioned before, all these new methods create plenty of directions for epigenome research, but they also pose substantial challenges in terms of data processing, statistical analysis, and biological interpretation of observed differences [54]. Especially, in the case of cancer, where identification of cancer-specific differentially DNA-methylated regions (cDMRs) is very critical, as tumor heterogeneity is a major hurdle. A study suggested that these cDMRs might be generalized across cancer types and that the changes in the methylation pattern in these regions can distinguish cancer from normal tissue [58]. They also suggested a model for cancer

involving loss of epigenetic stability of well-defined genomic domains that underlies increased methylation variability in cancer that may contribute to tumor heterogeneity. In addition, keeping in mind the critical role of viral infection (e.g., HPV) in the pathophysiology of different cancer types (e.g., cervical, HNC, etc.), a recent study proposed a methodology that can provide a confined but significant insight into the presence, concentration, and types of methylated viral sequences in MBD-Seq data at low additional cost, where a priori knowledge of viral reference genome sequences is not available [59].

According to a recent study [60], for the identification of differentially methylated regions (DMRs) based on WGBS, per-sample coverage can be kept in the range of 5–15 ×, depending on the magnitude of methylation differences between the groups and whether a smoothing or single CpG-based DMR identification strategy is used. In order to identify long DMRs with large methylation differences, reducing the coverage down to 1 × or 2 × per sample is acceptable. At least two separate biological replicates should be considered for DMR analysis and they should be analyzed separately to increase power, as opposed to being pooled together for analysis. However, choosing an appropriate number of biological replicates is a complex issue influenced by the degree of within-group heterogeneity, the magnitude of between-group differences, and the presence of various confounding factors.

A complete list of software tools for the analysis, interpretation, and visualization of DNA methylation data is given and carefully reviewed [47], and from them, some are mentioned in the following sections.

Data Processing of Bisulfite-Sequencing Data

Bisulfite conversion of genomic DNA and subsequent PCR amplification gives rise to two PCR products and up to four potentially different DNA fragments for any given locus (Fig 13.3). Also, as cytosine methylation is not symmetrical, the two strands of DNA in the reference genome must be considered separately. BS-Seq mapping may therefore require up to four different strand alignments to

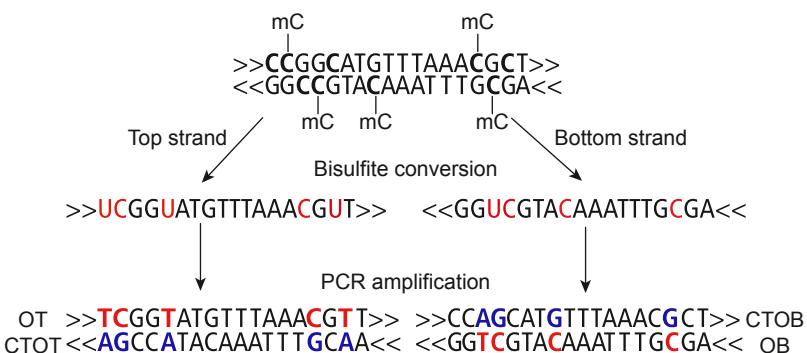


FIGURE 13.3

Mapping of BS-Seq reads to four possible bisulfite strands (OT/CTOT/OB/CTOB) is equivalent to mapping the bisulfite read and its reverse complementary read to both top (or forward)/bottom (or reverse) strands of the original reference sequence. *CTOB*, strand complementary to the original bottom strand; *CTOT*, strand complementary to the original top strand; *OB*, original bottom strand; *OT*, original top strand;

Figure adopted from Krueger F et al. Nat Methods 2012; 9(2):145–51.

be analyzed for each sequence [61]. Because of this complexity of BS-Seq alignments, standard sequence alignment software cannot be used.

The alignment of BS-Seq reads needs to account for the selective depletion of unmethylated Cs, but otherwise it can be carried out with short-read aligners that are similar to those used for chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) or genome-resequencing data. From this alignment, absolute DNA methylation levels can be calculated by determining the percentage of C's and T's among all reads aligned to each C in the reference genomic DNA sequence. In this context, two approaches have been developed, wild-card alignment and three-letter alignment [47]. During analysis, wild-card aligners (viz. BSMAP, GSNAP, Last, Pash, RMAP, RRBSMAP, and segemehl) replace C's in the genomic DNA sequence by the wild-card letter Y, which matches both C's and T's in the read sequence, or they modify the alignment scoring matrix in such a way that mismatches between C's in the genomic DNA sequence and T's in the read sequence are not penalized. In contrast, three-letter aligners (viz. Bismark, BRAT, BS-Seeker, and MethylCoder) simplify BS alignment by converting all C's into T's in the reads and for both strands of the genomic DNA sequence. This way, they can carry out the alignment exclusively on a three-letter alphabet (namely, A, G, and T) using a standard aligner, such as Bowtie. As compared to three-letter aligners, wild-card aligners are expected to achieve a higher genomic coverage, though they are at an increased risk of introducing bias toward higher methylation levels because the extra C's in a methylated sequencing read can raise the sequence complexity to a level that is sufficient for unique alignment to the genome, whereas the corresponding, unmethylated T-containing read is discarded owing to nonunique alignment.

In real BS-Seq data, the quality of base calls tends to drop as the read length increases. However, longer reads increase coverage but also increase the number of incorrect methylation calls. And so, base-call errors are found to be random, for which, the frequency for each base in a position with high error rates is expected to be around 25%. Now, the possible source of contamination that can lead to a change in base composition is the presence of (methylated) adaptor sequences. Such deviations of the average nucleotide distribution can usually be spotted in a base composition analysis. A way to thwart methylation miscalls/mismapping events due to the erroneous base call in the reads is to select stringent alignment parameters. Increasing the mapping stringency prevented mismatched sequences from aligning, thus reducing the number of erroneously inferred methylation states but at the cost of reduced mapping efficiency. A better way of diminishing incorrect methylation calls from such poor quality data is to trim off low-quality base calls before read alignment steps are carried out [61]. In this context, the Bis-SNP variant caller broadens the well-validated Genome Analysis Toolkit algorithm to bisulfite-sequencing data and thereby provides an important step in this direction [62]. It also removes a common error source in the analysis of DNA methylation data, as it can distinguish bisulfite-induced changes from genetic variants. This is possible because bisulfite-induced C to T variants exhibit a G on the opposing strand, whereas genetic C to T variants exhibit an A instead. Furthermore, Bis-SNP can directly infer accurate genotype information from bisulfite-sequencing data [47].

Data Processing of Bisulfite Microarray Data

For high-throughput profiling of bisulfite-converted DNA, specialized microarrays have been developed, which can quantify the DNA methylation levels of a predefined subset of C's, each of which is represented by dedicated probes on the microarray. The latest version of the Illumina Infinium assay comprises slightly more than 450,000 covered CpGs [45]. The genomic coverage of the Infinium assay

is more limited than that of most bisulfite-sequencing protocols (1.5% of CpGs in the human genome are present on the Infinium 450k microarray), but the compatibility with existing genotyping pipelines makes it an attractive assay for measuring DNA methylation in large sample populations. The bioinformatics processing of Infinium data primarily consists of image processing and data normalization. Image processing is almost always carried out using the vendor-provided Illumina BeadScan software, whereas quite a few options exist for normalizing the probe intensity data and for inferring absolute DNA methylation levels. The commercial Illumina GenomeStudio software provides a basic algorithm for signal normalization and background reduction using positive and negative control probes, and a similar algorithm is implemented in R/Bioconductor [63] as a part of the open-source packages minfi and methylumi. The main data normalization result gives a table of β -values (and, optionally, M values) that serves as the starting point for further downstream analyses. These β -values are conceptually equivalent to the absolute DNA methylation levels calculated from BS-Seq data, whereas M values are logically transformed β -values and exhibit a distribution that is better suited for use with some common statistical tests [64].

Data Processing of Enrichment-Based Data

Usually, next-generation sequencing of the DNA libraries counts the frequency of specific DNA fragments in each library and provides the raw data from which DNA methylation levels can be inferred. However, in the enrichment-based methods the DNA methylation information is not contained in the read sequence but in the enrichment or depletion of sequencing reads that map to specific regions of the genome. As a consequence, this method requires careful handling of any kind of systematic biasness in the data because any fluctuations in DNA-sequencing coverage will directly affect the DNA methylation measurement. Furthermore, to obtain absolute DNA methylation measurements, it is necessary to statistically correct for region-specific differences in CpG density [47].

The initial step in the analysis of enrichment-based DNA methylation data is reference genome alignment, which can be done using a standard aligner, such as the BWA or Bowtie. On the basis of this alignment, relative enrichment scores, the relative enrichment of DNA fragments from a given genomic region after comparing with the control experiment (sequencing of unenriched DNA), are calculated by extending the sequencing reads to the estimated DNA fragment size and counting the number of unique reads that overlap with each CpG or with the genomic regions of interest. These enrichment scores actually predict the regional DNA methylation levels, but the irregular distribution of CpGs throughout mammalian genomes heavily confounds the estimation of this score. For example, a region with a high CpG density and moderate levels of DNA methylation can give rise to higher enrichment scores than a region with a low CpG density but with high levels of DNA methylation, and a region without any aligning reads can result either from the absence of DNA methylation or from difficulties in sequencing or aligning reads originating from this region. Several algorithms have been developed for correcting this bias and for inferring absolute DNA methylation levels at a single-base resolution [47].

Data Visualization and Statistical Analysis

In the context of data integration and visualization, an online platform (DaVIE) was developed based on a database of DNA methylation experiments. This tool allows navigating through multiple DNA methylation experiments and integrating different data types, including ChIP-seq data [65]. While the experimental methods and the computational analysis of individual data types are compared and

perfected, scientists nowadays are investigating how to make connections between the various epigenetic layers/alterations that are surveyed [55]. In this context, different computational and experimental strategies are proposed that can be helpful for further investigating how different epigenetic layers and histone marks are interconnected [66].

CONCLUSION AND FUTURE PERSPECTIVES

The main objectives of epigenetic profiling are to explore pathological diagnosis at molecular level, characterization of tumor, and to develop appropriate therapies. The basic principle behind the development of epigenetic therapy is reversal of epigenetic alterations to restore cellular defense mechanisms against tumor development. Therefore, epigenetic therapies can only be productive if epigenetic information is fully explored. To achieve these goals, many national and international initiatives have been started. Recently, due to advancements of the high-throughput technologies like NGS, array hybridization, and nanopore sequencing, it is possible to generate more reads and sequencing of larger fragments, which will make subsequent bioinformatics analyses more easy, reliable, and cost-effective. Epigenomics studies are among the most exciting topics in biomedical science today. The epigenome acts as an interface between the environment and the genome. Different epigenetic mechanisms control DNA accessibility of a person throughout the lifetime. Epigenetic changes determine and contribute to interindividual discrepancy in gene expression, and thus it is crucial to explore epigenetic profiles of tumor-related genes in cancers. Among different epigenetic alterations, aberrant DNA methylation has been considered to be a potentially useful marker for multiple cancers. Aberrant DNA methylation profiling of different tumor-related genes such as *RASSF1A*, *p16*, *DAPK1*, *ECAD*, *MGMT*, etc. provide us ample sensitivity and specificity for the uncovering of HNC etiology. It is also observed that highly significant association between copy number and DNA methylation profiles shows that these modes of gene regulation are linked in head and neck cancer. Unlike genetic alterations, epigenetic changes are potentially reversible, and this feature makes them attractive targets for therapeutic intervention. Thus it is important to explore the uses of epigenetic pathways in the development of new to molecular diagnosis approaches and novel targeted therapies across the clinical field. In recent years, genome-wide association studies have been conducted in large scale to understand the genetic factors associated with complex diseases like cancers. However, GWAS is poorly equipped to reveal the specific mechanistic part involved in cancers. Therefore, biomedical researchers have gradually shifted toward EWASs to explore how epigenetic alteration functioned in etiology of cancer phenotype. High-throughput DNA methylation profiling technology (450K BeadChip array) similar to successful GWAS is now available that allows to measure DNA methylation at 485,512 cytosine positions across the human genome, of which there are 482,421 CpG sites and 3091 non-CpG sites. Methylation status screened by epigenome-wide methylation analysis will identify disease-associated genes or loci.

Recent developments in the knowledge of epigenetics of cancer have allowed the development of several inhibitors of DNA methyltransferase, such as 5-azacitidine and decitabine, and histone deacetylase, which are effectively used in the treatment of several malignancies. The NIH storehouse for clinical trials reports some trials involving epigenetic-based drugs in head and neck cancer treatment. Azacitidine and cisplatin have been tested in combined chemotherapy in advanced, recurrent, and metastatic squamous cell carcinoma of head and neck, but no data is available to date

about the outcome of these studies. The use of epigenetic inhibitors in association with traditional anticancer therapeutic agents looks very promising as a tool to improve the chemosensitivity of nonresponsive cancers. Therefore, the epigenetic profiling of head and neck cancers will not only help us to understand the molecular details behind cancer development but also provide novel strategies to develop new therapies.

REFERENCES

- [1] Esteller M. Epigenetics in cancer. *N Engl J Med* 2008;358(11):1148–59.
- [2] Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Canc* 2011;11(10):726–34.
- [3] Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 2010;31(1):27–36.
- [4] Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA A Cancer J Clin* 2005;55(2):74–108.
- [5] Choudhury J, Ghosh S. Promoter hypermethylation profiling identifies subtypes of head and neck cancer with distinct viral, environmental, genetic and survival characteristics. *PLoS One* 2015;10(6):e0129808.
- [6] Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. *Cell* 2007;128(4):635–8.
- [7] Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 1983;301(5895):89–92.
- [8] Ng HH, Bird A. DNA methylation and chromatin modification. *Curr Opin Genet Dev* 1999;9(2):158–63.
- [9] Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. *Canc Res* 2001;61(8):3225–9.
- [10] Feinberg AP, Vogelstein B. Hypomethylation of ras oncogenes in primary human cancers. *Biochem Biophys Res Commun* 1983;111(1):47–54.
- [11] Dillon N. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* 2006;14(1):117–26.
- [12] Mack GS. Epigenetic cancer therapy makes headway. *J Natl Cancer Inst* 2006;98(20):1443–4.
- [13] Feinberg AP. Cancer epigenetics takes center stage. *Proc Natl Acad Sci USA* 2001;98(2):392–4.
- [14] Ozdag H, Teschendorff AE, Ahmed AA, Hyland SJ, Blenkiron C, Bobrow L, et al. Differential expression of selected histone modifier genes in human solid cancers. *BMC Genom* 2006;7:90.
- [15] Mitra SA, Mitra AP, Triche TJ. A central role for long non-coding RNA in cancer. *Front Genet* 2012;3:17.
- [16] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. *Nature* 2005;435(7043):834–8.
- [17] Chen CZ. MicroRNAs as oncogenes and tumor suppressors. *N Engl J Med* 2005;353(17):1768–71.
- [18] Lujambio A, Ropero S, Ballestar E, Fraga MF, Cerrato C, Setien F, et al. Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. *Canc Res* 2007;67(4):1424–9.
- [19] Baylin SB, Esteller M, Rountree MR, Bachman KE, Schuebel K, Herman JG. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* 2001;10(7):687–92.
- [20] Ji X, Guan C, Jiang X, Li H. Diagnostic accuracy of DNA methylation for head and neck cancer varies by sample type and number of markers tested. *Oncotarget* 2016;7(48):80019–32.
- [21] Hasegawa M, Nelson HH, Peters E, Ringstrom E, Posner M, Kelsey KT. Patterns of gene promoter methylation in squamous cell cancer of the head and neck. *Oncogene* 2002;21(27):4231–6.
- [22] Yalniz Z, Demokan S, Suoglu Y, Ulusan M, Dalay N. Simultaneous methylation profiling of tumor suppressor genes in head and neck cancer. *DNA Cell Biol* 2011;30(1):17–24.
- [23] Tanaka N, Odajima T, Ogi K, Ikeda T, Satoh M. Expression of E-cadherin, alpha-catenin, and beta-catenin in the process of lymph node metastasis in oral squamous cell carcinoma. *Br J Canc* 2003;89(3):557–63.

- [24] Mascolo M, Siano M, Ilardi G, Russo D, Merolla F, De Rosa G, et al. Epigenetic deregulation in oral cancer. *Int J Mol Sci* 2012;13(2):2331–53.
- [25] Asokan GS, Jeelani S, Gnanasundaram N. Promoter hypermethylation profile of tumour suppressor genes in oral leukoplakia and oral squamous cell carcinoma. *J Clin Diagn Res* 2014;8(10):ZC09–12.
- [26] Kulkarni V, Saranath D. Concurrent hypermethylation of multiple regulatory genes in chewing tobacco associated oral squamous cell carcinomas and adjacent normal tissues. *Oral Oncol* 2004;40(2):145–53.
- [27] Maruya S, Issa JP, Weber RS, Rosenthal DI, Haviland JC, Lotan R, et al. Differential methylation status of tumor-associated genes in head and neck squamous carcinoma: incidence and potential implications. *Clin Canc Res* 2004;10(11):3825–30.
- [28] Sanchez-Cespedes M, Esteller M, Wu L, Nawroz-Danish H, Yoo GH, Koch WM, et al. Gene promoter hypermethylation in tumors and serum of head and neck cancer patients. *Canc Res* 2000;60(4):892–5.
- [29] Viswanathan M, Tsuchida N, Shanmugam G. Promoter hypermethylation profile of tumor-associated genes p16, p15, hMLH1, MGMT and E-cadherin in oral squamous cell carcinoma. *Int J Cancer* 2003;105(1):41–6.
- [30] Su PF, Huang WL, Wu HT, Wu CH, Liu TY, Kao SY. p16(INK4A) promoter hypermethylation is associated with invasiveness and prognosis of oral squamous cell carcinoma in an age-dependent manner. *Oral Oncol* 2010;46(10):734–9.
- [31] Supic G, Kozomara R, Jovic N, Zeljic K, Magic Z. Prognostic significance of tumor-related genes hypermethylation detected in cancer-free surgical margins of oral squamous cell carcinomas. *Oral Oncol* 2011; 47(8):702–8.
- [32] Toyota M, Ahuja N, Suzuki H, Itoh F, Ohe-Toyota M, Imai K, et al. Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. *Canc Res* 1999;59(21):5438–42.
- [33] Ogi K, Toyota M, Ohe-Toyota M, Tanaka N, Noguchi M, Sonoda T, et al. Aberrant methylation of multiple genes and clinicopathological features in oral squamous cell carcinoma. *Clin Canc Res* 2002;8(10):3164–71.
- [34] Al-Kaabi A, van Bockel LW, Pothen AJ, Willems SM. p16INK4A and p14ARF gene promoter hypermethylation as prognostic biomarker in oral and oropharyngeal squamous cell carcinoma: a review. *Dis Markers* 2014;2014:260549.
- [35] Dong SM, Sun DI, Benoit NE, Kuzmin I, Lerman MI, Sidransky D. Epigenetic inactivation of RASSF1A in head and neck cancer. *Clin Canc Res* 2003;9(10 Pt 1):3635–40.
- [36] Demokan S, Chuang AY, Pattani KM, Sidransky D, Koch W, Califano JA. Validation of nucleolar protein 4 as a novel methylated tumor suppressor gene in head and neck cancer. *Oncol Rep* 2014;31(2):1014–20.
- [37] Guerrero-Preston R, Michailidi C, Marchionni L, Pickering CR, Frederick MJ, Myers JN, et al. Key tumor suppressor genes inactivated by "greater promoter" methylation and somatic mutations in head and neck cancer. *Epigenetics* 2014;9(7):1031–46.
- [38] Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. *Epigenet Chromatin* 2016;9:26.
- [39] Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* 1992;89(5):1827–31.
- [40] Choudhury JH, Das R, Laskar S, Kundu S, Kumar M, Das PP, et al. Detection of p16 promoter hypermethylation by methylation-specific PCR. *Meth Mol Biol* 2018;1726:111–22.
- [41] Boyko A, Kovalchuk I. Analysis of locus-specific changes in methylation patterns using a COBRA (combined bisulfite restriction analysis) assay. *Meth Mol Biol* 2010;631:23–31.
- [42] Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* 2016;5(1).
- [43] Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc* 2015;10(3):475–83.
- [44] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005;33(18): 5868–77.

- [45] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98(4):288–95.
- [46] Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016;8(3):389–99.
- [47] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;13(10):705–19.
- [48] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 2005;37(8):853–62.
- [49] Aberg KA, McClay JL, Nerella S, Xie LY, Clark SL, Hudson AD, et al. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case-control samples. *Epigenomics* 2012;4(6):605–21.
- [50] Wang S, Lv J, Zhang L, Dou J, Sun Y, Li X, et al. MethylRAD: a simple and scalable method for genome-wide DNA methylation profiling using methylation-dependent restriction enzymes. *Open Biol* 2015;5(11).
- [51] Oda M, Glass JL, Thompson RF, Mo Y, Olivier EN, Figueroa ME, et al. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res* 2009;37(12):3829–39.
- [52] Clark C, Palta P, Joyce CJ, Scott C, Grundberg E, Deloukas P, et al. A comparison of the whole genome approach of MeDIP-seq to the targeted approach of the Infinium HumanMethylation450 BeadChip(R) for methylome profiling. *PLoS One* 2012;7(11):e50233.
- [53] Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, et al. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* 2012;7(4):617–36.
- [54] Bock C, Lengauer T. Computational epigenetics. *Bioinformatics* 2008;24(1):1–10.
- [55] Robinson MD, Pelizzola M. Computational epigenomics: challenges and opportunities. *Front Genet* 2015;6: 88.
- [56] Lim SJ, Tan TW, Tong JC. Computational Epigenetics: the new scientific paradigm. *Bioinformation* 2010; 4(7):331–7.
- [57] Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;16(1):6–21.
- [58] Hansen KD, Timp W, Bravo HC, Sabuncian S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011;43(8):768–75.
- [59] Mensaert K, Van Criekinge W, Thas O, Schuuring E, Steenbergen RD, Wisman GB, et al. Mining for viral fragments in methylation enriched sequencing data. *Front Genet* 2015;6:16.
- [60] Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* 2015;12(3):230–2. 1 p following 2.
- [61] Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* 2012;9(2):145–51.
- [62] Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* 2012;13(7):R61.
- [63] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5(10):R80.
- [64] Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinf* 2010;11:587.
- [65] Fejes AP, Jones MJ, Kobor MS. DaVIE: database for the visualization and integration of epigenetic data. *Front Genet* 2014;5:325.
- [66] de Pretis S, Pelizzola M. Computational and experimental methods to decipher the epigenetic code. *Front Genet* 2014;5:335.

This page intentionally left blank

EPIGENOME-WIDE DNA METHYLATION PROFILES IN ORAL CANCER

14

Raghunath Chatterjee, Shantanab Das, Aditi Chandra, Baidehi Basu*Human Genetics Unit, Indian Statistical Institute, Kolkata, India*

INTRODUCTION

Before the birth of epigenetics, study of genetics was thought to be the most significant phenomenon in terms of gene regulation. Eventually, it was observed that only genetics is not sufficient to describe the gene regulation. Cancer was also thought to be regulated by genetic mutations of oncogenes and tumor suppressor genes. But recent research advancement shows that epigenetic modifications give enough information on the mechanisms of neoplastic transformation. Oral squamous cell carcinoma (OSCC) is the eighth most common cancer in the world and the number of annual death is quite consistent for the last 30 years. Oral cancer is one of the most common malignancies in Southeast Asia, accounting for up to 30%–40% of all malignancies in India [1]. Most oral malignancies occur as squamous cell carcinomas (SCCs) and many OSCCs develop from premalignant conditions of the oral cavity [2]. Various premalignant conditions have been implicated in the development of oral cancer, including leukoplakia, erythroplakia, oral lichen planus, oral submucous fibrosis, discoid lupus erythematosus, and hereditary disorders such as dyskeratosis congenita and epidermolysis bullosa [3]. Despite the general accessibility of the oral cavity during physical examination, many malignancies are not diagnosed until late stages of the disease. Despite the significant improvements in therapeutic modalities in OSCC, 5-year survival rates are among the lowest of the major cancers and the main reason is ascribed to the lack of early detection.

Extensive studies of DNA methylation and histone modifications have helped to understand several fundamental biological processes, such as genomic imprinting, activation or silencing of transposons, cell differentiation, cell development, etc. [4–7]. Epigenetic mechanisms are generally heritable, but these modifications are not stable across cell differentiation and lead to lower fidelity of inheritance. It also propagates in both mitotic and meiotic generations [8]. Besides the genetic information encoded in DNA, packaging of DNA inside the nucleus also affects gene regulation [9]. In mammalian system, DNA methylation is a postreplication modification, where methyl group is attached covalently to the 5'-carbon position of cytosine base [4]. This event is mediated by different DNA methyltransferase and is predominant mainly in CpG dinucleotide context [6]. CpG dinucleotides are sometimes clustered in

0.5–4 kb regions of the genome and are termed as CpG islands (CGIs) [4]. Epigenetic silencing of genes by CGI methylation at the promoter regions of tumor suppressor genes in several cancers has been well established in the literature [7,10–13].

Since 1980s, when first the global hypomethylation was observed in primary cancer tissues and compared to the normal tissues [14], the role of DNA methylation in cancer has been studied extensively. In cancers, hypomethylation occurs at gene bodies, transposable elements, and repetitive sequences, and hypermethylation occurs at promoters, which leads to aberrant transcription initiation and genome instability.

EPIGENETIC REGULATION IN ORAL CANCER

Both genetic and epigenetic alterations and their complementary roles helped us to understand the molecular pathogenesis of oral cancer. Among genetic, epigenetic, and protein diagnostic markers of carcinogenesis, promoter hypermethylation is a critical step and easily diagnosable as compared to other biomarkers. DNA methylation profile can be a useful source for developing potential biomarker and may lead to better disease diagnosis [15]. Environmental exposure to genotoxic agents such as tobacco, alcohol, and smoke has been identified as risk factor towards the development of oral cancer. These genotoxic agents have been acting synergistically along with the epigenetic machinery towards oral cancer pathogenesis, for example, methylation of p16^{INK4A} and E-cadherin is associated with smoking [16].

Global DNA hypomethylation leads to tumorigenesis by several mechanisms and three of them are highly acceptable. First, demethylation of repeat elements (LINE, Alu) leads to chromosomal instability. Second, demethylation of silenced proto-oncogene promoters or endoparasitic elements leads to carcinogenesis. And third, DNA methylation alteration leads to loss of imprinting [11]. These mechanisms are very consistent for different tumor entities including OSCC. Epigenetic silencing as a result of promoter hypermethylation has been identified in different classes of genes including cell cycle-related genes *p16*, *p15*, *cyclin A1*, and *Rb* (retinoblastoma protein) in OSCC. DNA repair genes showing promoter methylation and associated with oral cancer are *O-6-methylguanine-DNA methyltransferase (MGMT)*, *MutL homolog 1 (MLH1)*, and *fragile histidine triad (FHIT)*. It is seen that *hMLH1* promoter methylation is an early phenomenon in oral cancer. Certain apoptotic genes (e.g., *death-associated protein kinase 1 (DAPK1)*, *tumor protein p73 (TP73)*, *Ras association (RalGDS/AF-6) domain family member 1 (RASSF1)*) and tumor suppressor genes (e.g., *deleted in colorectal carcinoma (DCC)*, *deleted in bladder cancer chromosome region candidate 1 (DBCCR1)*, *ataxia telangiectasia mutated (ATM)*) are also downregulated in oral cancer as a result of DNA hypermethylation. Nuclear transcriptional regulators such as retinoic acid receptor beta (*RARβ*) and cytoglobin are also controlled epigenetically in oral cancer. The changed expressions of retinoic acid receptors show correlation with *p16*, *p15*, and *p21* cell cycle regulators. It is reported that this correlation is strongly associated with the development, progression, and prognosis of oral cancer [17–19]. The *serpin family E member 1 (SERPINE1)* gene shows altered epigenetic profile in OSCC when compared to normal subjects [20]. These targeted analyses gave us some understanding about the role of methylation in OSCC development. But it is not sufficient as epigenome-wide methylation pattern will help to understand the genome-wide regulation of DNA methylation in OSCC development [21].

NEED FOR COMPUTATIONAL TOOLS IN EPIGENETICS STUDY

Epigenetic regulation is regarded as an early event in the development of oral carcinogenesis [22]. In recent times, the advancement of massively parallel next-generation sequencing (NGS) and microarray technology raises the possibilities to map DNA methylation in genome-wide scale for large number of samples [23]. These technologies bring tons of opportunities and enrich the epigenome research. At the same time, it poses critical challenges in terms of data processing, data analysis, and their biological interpretation [24]. The goal of International Human Epigenome Consortium (IHEC) perfectly describes the degree of bioinformatic challenges [25]. IHEC Data Portal (<http://epigenomesportal.ca/ihec>) provides access to more than 7000 reference epigenomic data sets for over 600 tissues [26,27]. To accomplish this project, alignment of 1 trillion sequencing reads to the human genome will be essential to detect cell type-specific DNA methylation patterns. To access the data, we need user-friendly epigenome browsers and analysis tools [28]. Nowadays the epigenetic research is shifted towards exploring the epigenetic basis of human diseases and as a result several methylomes in large case-control studies are already generated and many more are in process. Analysis of such data requires rigorous computational and statistical tools for identification of disease-associated differentially methylated regions [29].

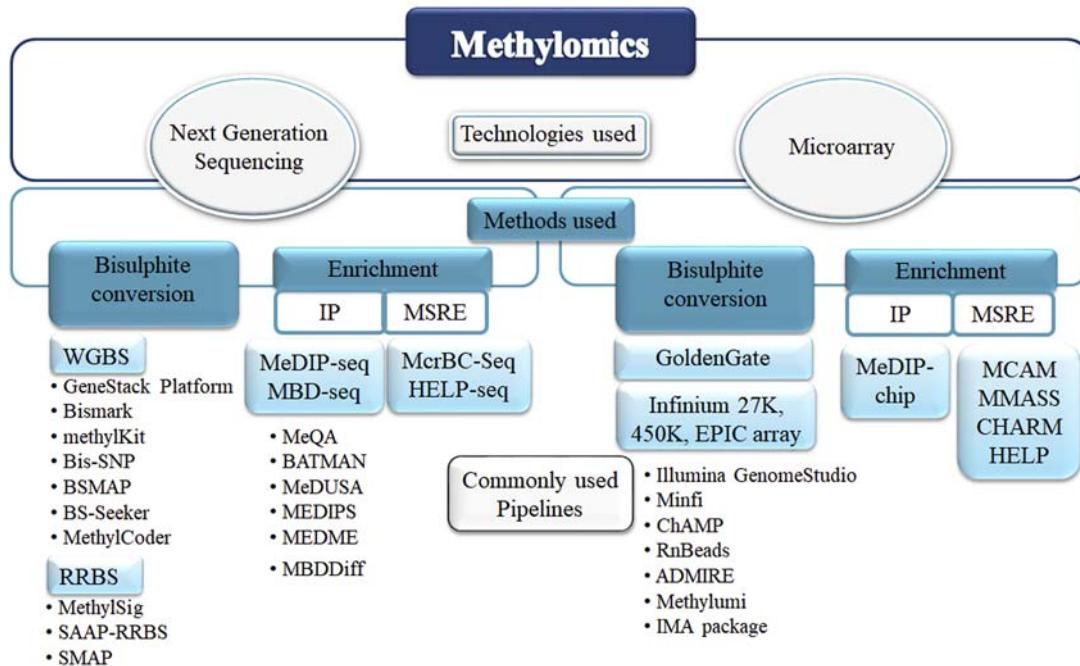
AVAILABLE METHODS AND COMPUTATIONAL TOOLS FOR ORAL CANCER METHYLOMICS

The study of genome-wide DNA methylation or methylomics is performed mainly either by microarray hybridization or next-generation sequencing (NGS) (Fig. 14.1). Both techniques are used either following sodium bisulfite conversion of DNA or following enrichment of methylated DNA. Later method includes enrichment of methylated DNA by methyl-specific antibodies and methylation-sensitive restriction enzyme (MSRE) digestion followed by size selection or by capturing CpG-rich regions using probe hybridizations. Sodium bisulfite-treated DNA samples are either subjected to NGS or hybridization with oligonucleotide probes. We have briefly discussed available tools for analyzing methylomics data, generated from each of these methods, and specifically focused on those tools that are used in oral cancer methylomics.

TOOLS FOR METHYLOMICS BY BISULFITE-SEQUENCING METHOD

Sodium bisulfite converts all unmethylated cytosines to thymines, while the methylated cytosines remain unaffected. Bisulfite-converted DNA samples are sequenced using NGS technologies. The degree of DNA methylation is calculated from the mapped sequence reads in the reference genome. The number of cytosine and thymine bases in all alignments is used to calculate the methylation level of each interrogated CpG [29].

For genome-wide DNA methylation analysis, whole genome bisulfite sequencing (WGBS) is cost inefficient for multiple samples as it requires significantly higher genomic coverages [30,31]. That is probably one of the reasons that WGBS has not been used in OSCC. Reduced representation bisulfite sequencing (RRBS) is another alternative, as it only covers CpG-rich regions of the genome [32].

**FIGURE 14.1**

Summarized representation of methyloomics data generation methods and commonly used pipelines. *ADMIRE*, Analysis of DNA Methylation in genomic REgions; *BATMAN*, BAyesian Tool for Methylation ANalysis; *ChAMP*, Chip Analysis Methylation Pipeline; *CHARM*, Comprehensive High-throughput Arrays for Relative Methylation; *HELP*, HpaII tiny fragment Enrichment by Ligation-mediated PCR; *HELP-Seq*, HpaII tiny fragment Enrichment by Ligation-mediated PCR coupled with Sequencing; *IP*, ImmunoPrecipitation; *MBDCap-Seq*, Methyl-CpG Binding Domain-based Capture and Sequencing; *MCAM*, Methylated CpG island Amplification with Microarray hybridization; *McrBC-Seq*, McrBC Digestion and Sequencing; *MeDIP-Seq*, Methyl-CpG ImmunoPrecipitation and Sequencing; *MEDME*, Modeling Experimental Data with MeDIP Enrichment; *MMASS*, Microarray-based Methylation Assessment of Single Samples; *MSRE*, Methylation-Sensitive Restriction Enzyme digestion; *RRBS*, Reduced Representation Bisulfite Sequencing; *WGBS*, Whole Genome Bisulfite Sequencing.

For alignment of the NGS reads, two methods are mainly used, namely wild-card alignment and three-letter alignment. BSMAP, RMAP, GSNAP, RRBSMAP, Last, Pash, and segemehl tools are mostly used wild-card aligners [33–39]. After successful alignment, the degree of methylation is calculated from the frequency of cytosines and thymines at a specific CpG locus. At this step the methylation call can be enhanced more by using local realignment and reanalyzing the sequence [40]. Bis-SNP variant caller uses the Genome Analysis Toolkit for bisulfite-sequencing data and also minimizes the errors in DNA methylation estimation by tracking the genetic variants [41].

TOOLS FOR METHYLOMICS BY BISULFITE-MICROARRAY METHOD

Specialized microarray techniques are developed for quantification of DNA methylation by probe hybridization. Illumina GoldenGate Assay for Methylation was used initially for profiling methylation status of 1536 CpG sites across the genome [42,43]. Methylation Cancer Panel I consisted of 1505 CpG sites across 807 genes, with 28.6% genes having one CpG site, 57.3% genes having two CpG sites, and 14.1% genes having three or more CpG sites per gene. For comprehensive genome-wide DNA methylation profiling, Illumina introduced Infinium HumanMethylation27 (27k) BeadChip array that can profile the methylation status of 27,578 CpGs across the human genome. Later on, the total genomic coverage of the array has been dramatically increased by the introduction of Infinium HumanMethylation450 (450k) BeadChip. It determines methylation status of 485,577 CpGs in the human genome. This assay identifies the methylation status by smoothly avoiding capture-associated biases. Most of the OSCC methylomics are generated using Illumina GoldenGate [44–46] and Infinium assays [47–50]. Illumina's recently developed Infinium Methylation EPIC BeadChip covers more than 850,000 CpG sites across the human genome. It has >90% of the original CpGs of HumanMethylation450 BeadChip and an additional 350,000 CpGs in the enhancer regions including 58% of FANTOM5 enhancers and 7% distal and 27% proximal ENCODE regulatory elements. Illumina Infinium assay is the most widely used bisulfite microarray technique and requires specific normalization methods [51]. Bioinformatics analysis of Infinium data consists of two major steps: image processing and data normalization. The image processing is always carried out in Illumina BeadsScan software that provides probe intensity data to check DNA methylation level, whereas normalization is carried out in Illumina GenomeStudio software [52]. Removal of batch effect is also the key step to process the data further in large case-control studies [53]. There are several R/Bioconductor packages for Infinium Methylation data handling. Methylumi and minfi are widely used for normalization, visualization, and analysis [54,55]. R-package RnBeads and ChAMP are also very helpful as they provide complete pipeline for data normalization, visualization, and differentially methylated region identification [56,57]. Although most of the OSCC methylomics identified differentially methylated probes from GoldenGate or Infinium assays, detection of differentially methylated regions (DMR) has not been performed in OSCC methylomics. DMR detection is primarily based on the comparison of locus-specific methylation differences between two groups. Several tools or packages, for example, comb-P [58], IMA package [52,59], EVORA package [60], Bumphunter [61], Probe-lasso [62], DMRcate [63], ChAMP [57] are available for DMR detection.

TOOLS FOR METHYLOMICS BY ENRICHMENT-BASED METHOD

Enrichment-based method has not been generally used for oral cancer methylomics yet, but it has widespread advantages for methylomics study. Enrichment-based method stands on the concept of enriching DNA in a methylation-specific way. Methylated DNA can be precipitated with antibodies raised against 5-methylcytosine, MBD, MeCP2, etc. The precipitated DNA is subjected to high-throughput sequencing. Methylated DNA immunoprecipitation sequencing (MeDIP-seq) is an enrichment-based method [64]. Methyl-CpG binding domain (MBD) based capture and sequencing (MBD-seq) uses proteins to capture methylated DNA in the genome [65]. Genomic DNA is first sonicated and incubated with tagged MBD proteins that can bind methylated cytosine. The protein-DNA complex is then precipitated with antibody-conjugated beads specific to the protein.

Enrichment of methylated DNA by methylation-sensitive restriction enzyme (e.g., McrBC, MspI) digestion followed by sequencing (McrBC-seq, RRBS) is also a potent enrichment-based method [66,67]. NGS reads are aligned to the reference genome using basic aligners such as Bowtie or BWA.

DNA METHYLATION DATA VISUALIZATION

To inspect DNA methylation level, successful visualization is one of the important criteria. For browser visualizations, genome-wide methylation data are converted into a specific file format. bigWig format can be used in UCSC Genome browser, Ensembl, or WashU Human Epigenome Brower [68]. DNA methylation level at each CpG locus is represented by the height of interspersed vertical bars in the browser. Desktop genome browsers such as Integrative Genomics Viewer (IGV) and Integrated Genome Brower (IGB) are also available for visualization of DNA methylation data [69,70].

In case-control studies, the objective of methylomics is to identify the differences in DNA methylation profiles between cases and control samples, and sometimes among the cases of different exposures or genetic background. Clustering has been used to distinguish the similarity and differences between the affected and normal individuals. Popularly used clustering methods for DNA methylation data analysis are hierarchical and partitioning clustering [71]. Both of these methods have some advantages and disadvantages over the other. Generally, partitioning clustering requires less computing time than hierarchical clustering. Partitioning clustering requires prior assumptions (e.g., number of clusters k), while hierarchical clustering only uses a similarity measure. Hierarchical clustering returns a subjective division of clusters but partitioning clustering results in exactly k clusters. These clustering methods can also be classified into supervised or unsupervised methods. The unsupervised method digs the data to find out relationships without any prior information. This method identifies groups of highly frequent and correlated data elements and compares them to search for significant variables. On the other hand, supervised clustering uses prior biological information for classification.

DNA METHYLOMICS IN ORAL CANCER

Genome-wide hypomethylation, aberrant promoter methylation, and inactivation of genes in OSCC cover a broad range of the cellular processes such as apoptosis, cell cycle control, signaling pathways, DNA repair mechanism pathways [44,46–50,72,73].

DNA METHYLATION BIOMARKER FOR OSCC

DNA methylation in OSCC plays a crucial role in disease development, prognosis, and regulation. On the other hand, DNA methylation as a biomarker for early identification of OSCC is another potential advancement (Table 14.1). In case-control studies, the objective of methylomics is to identify the differences in DNA methylation profiles between cases and control samples, and sometimes among the cases of different genetic background or environmental exposures. Genome-wide methylation pattern of 1505 CpG sites across 807 cancer-related genes was first performed using Illumina GoldenGate Methylation Array for preoperative saliva, postoperative saliva, associated cancer tissue of OSCC patients, and saliva from normal subjects [44]. This study constructed methylation classifier based on the methylated genes common to both preoperative saliva and tissue. Forty-one loci across 34 genes were methylated in preoperative saliva and tumors compared to postoperative and normal saliva samples.

Table 14.1 Epigenome-wide DNA Methylation Studies and Key Findings in Oral Cancer

Study Samples (Discovery Set)	Method for Genome-wide Methylation	Tools Used	Key Findings	References
OSCC tissue, pre- and postoperative saliva ($N = 13$) and normal saliva ($N = 10$)	Illumina GoldenGate Methylation Array	BeadStudio Software	Methylation array analysis of saliva can produce a set of cancer-related genes that are specific and can be used as a composite biomarker	[44]
Leukoplakia ($N = 4$), OSCC ($N = 4$), and normal ($N = 4$)	HumanMethylation 27K BeadChip array	Spotfire DecisionSite	Methylation statuses of <i>HOXA9</i> and <i>NID2</i> could serve as biomarkers	[47]
OSCC and paired normal tissue ($N = 44$)	Illumina GoldenGate Methylation Array	Methylumi, Partek Genomic Suite, and MetaCore	Epigenetic deregulation of NOTCH signaling in OSCC shows methylation signature for recurrence	[46]
OSCC, dysplastic and adjacent normal tissue ($N = 10$)	HumanMethylation 27K beadchip	BeadStudio Software	Hypermethylated genes associated in MAPK and WNT signaling pathways	[48]
Primary oral squamous cell carcinoma ($N = 20$) and unrelated normal ($N = 4$)	Illumina Infinium HumanMethylation 450K array	GenomeStudio software	Differentially methylated <i>p16</i> , <i>DDAH2</i> , <i>DUSP1</i> promoter	[49]
24 oral premalignant Samples, of which 12 did not and 12 developed OSCC	Agilent 4 \times 44k Custom CGH microarray	GenePix Pro 3.0	Promoter methylation status of <i>AGTR1</i> , <i>FOXI2</i> , and <i>PENK</i> and hypomethylation of LINE1 elements in OPLs	[73]
40 OSCC patients, 10 adjacent normal and 5 unrelated controls from Taiwanese population	Illumina GoldenGate Methylation Array	Not mentioned	Among top 20 CpG panels as OSCC biomarkers, <i>FLT4</i> , <i>KDR</i> , and <i>TFPI2</i> showed higher specificity and suggested to be potential candidates as biomarkers for early detection of buccal OSCC	[45]
11 well-differentiated OSCC and 10 adjacent normal tissue	Illumina Infinium HumanMethylation 450k BeadChip Array	RnBeads	Higher <i>CTLA4</i> expression in early stages is connected to better survival in OSCC patients	[50]

Hierarchical clustering of 41 gene loci using BeadStudio software identified a single cluster of postoperative saliva, and two extreme clusters of preoperative saliva. Nine sets of gene panels, each consisting of 4–10 genes, were developed for potential use in clinical trial and predicted as noninvasive early detection biomarkers with 62%–77% sensitivity and 83%–100% specificity [44].

A recent study on tissue samples of 40 OSCC patients, 10 adjacent normal and 5 unrelated controls from Taiwanese population used Illumina GoldenGate Methylation Cancer Panel to determine OSCC biomarkers [45]. They identified 34 CpG sites which could distinguish between normal and cancer samples. Hierarchical agglomerative clustering using Manhattan distance and complete linkage showed separate clustering of some OSCC samples, while other OSCC and normal samples were intermingled. They selected top 20 CpG panels as OSCC biomarkers, of which the panel of *Fms-related tyrosine kinase 4 (FLT4)* and *Achaete-scute homolog 1 (ASCL1)* had 100% specificity, 90% sensitivity, and 0.95 AUC. *FLT4* was validated by pyrosequencing and its expression was found to be 2.14-fold in normal compared to disease tissue. *FLT4*, *kinase insert domain receptor (KDR)*, and *tissue factor pathway inhibitor 2 (TFPI2)* showed 100% specificity and >0.7 AUC in leave-one-out cross-validation, and were therefore suggested to be potential candidates as biomarkers for early detection of buccal OSCC [45].

In search for the early detection biomarkers of OSCC, Guerrero-Preston et al. used Human-Methylation27 BeadChip array with normal, premalignant, and oral cancer tissues [47]. The authors used two-phase study design consisting of discovery screen and prevalence screens for the identification of differentially methylated promoters and deregulated pathways in OSCC patients. Unsupervised hierarchical clustering of log transformed β values using Spotfire DecisionSite identified three separate clusters with higher methylation level in OSCC compared to normal and leukoplakia. 301 and 143 hypermethylated tumor suppressor genes were identified in OSCC compared to normal and leukoplakia respectively. Pathway analysis showed enrichment of cell adhesion, cell proliferation, growth regulation, and cell death-associated pathways for hypermethylated genes. Around half of these hypermethylated genes were downregulated in OSCC. Using multiple selection criteria, eight hypermethylated genes were selected for further analysis. Four genes, namely *endothelin receptor type B (EDNRB)*, *homeobox A9 (HOXA9)*, *GATA binding protein 4 (GATA4)*, and *Nidogen 2 (NID2)*, were previously reported to be hypermethylated in non-OSCC/HNSCC tumors and the other four genes, namely *melanoma cell adhesion molecule (MCAM)*, *kinesin family member 1a (KIF1A)*, *DCC*, and *calcitonin related polypeptide alpha (CALCA)*, were reported to be hypermethylated in oral and prostate cancers. Six of the eight promoters were differentially methylated in validation cohort. Even though four of them were correlated with the clinical diagnosis, sensitivity, specificity, AUC, and methylation cutoffs were significant only for *NID2* and *HOXA9*. Further validation of these two genes showed differential methylation and perfect correlation with clinical diagnosis. This study also tested the panel of *HOXA9* and *NID2* on saliva samples from oropharyngeal and OSCC samples. This panel showed a sensitivity of 50%, but showed high specificity and AUC for OSCC saliva samples [47].

Another recent study determined genome-wide epigenetic signatures in OSCC using Illumina GoldenGate Methylation Array of 44 tumor and adjacent normal samples [46]. The initial analysis used methylumi followed by Partek Genomic Suite and MetaCore to identify the β values and differentially methylated probes. This study identified 48 differentially methylated probes that were concordant with previously published data [47]. The authors also identified differentially methylated genes among HPV (+) and HPV (−) patients, and patients with and without extracapsular spread. Hierarchical agglomerative clustering using differentially methylated probes identified two distinct

clusters, termed as low CGI methylator phenotype (low-CIMP) and high-CIMP. They observed association of poor prognosis with only high-CIMP group. This study identified a five-gene signature that was associated with recurrence. Particularly, epigenetic deregulation of NOTCH signaling was established as part of a methylation signature for recurrence in OSCC [46].

ADVANCEMENT IN DNA METHYLATION STUDY IN OSCC

To investigate further and to correlate the hypermethylation-induced genes with demographic, clinicopathological characteristics and survival rate of OSCC, Khor et al. applied bisulfite-treated DNA on the Illumina Infinium HumanMethylation 450K array [49]. They used Illumina GenomeStudio software for background normalization. Wilcoxon rank sum test with 5% FDR identified differentially methylated probes at 33 hypermethylated promoters. The data were then exported to Partek Genomics suite to determine the differentially methylated genes. *p16*, *dimethylarginine dimethylaminohydrolase 2 (DDAH2)*, and *dual-specificity phosphatase 1 (DUSP1)* promoters were significantly differentially methylated. Further validation using methylation-specific PCR (MSP) showed 78%, 80%, and 88% positivity for these three promoters respectively [49].

In the same year, Towle et al. used the HumanMethylation27 BeadChip assay with bisulfite-converted DNA and analyzed the data with BeadStudio Software [48]. This study used oral cancer, dysplasia, and adjacent normal tissues from OSCC patients. They found a significant higher number of hypermethylated CpGs for OSCC than dysplastic and adjacent normal samples. Signed-rank analysis showed that dysplastic tissues were more hypermethylated than hypomethylation. On the contrary, hypo- and hypermethylation were the same for OSCC tissues. Interestingly, only 2.09% of the total hypermethylation in dysplastic tissues further showed higher DNA methylation, and 8.33% showed lower DNA methylation in OSCC. However, the dysplasia samples were scattered between these two groups. The mild dysplastic samples were clustered with the adjacent normal, whereas the moderate and severe dysplasia were clustered with the OSCC group. Concordant methylation profiles for previously reported candidate genes were found to be more for OSCC compared to dysplastic samples. Pathway analysis of these differentially methylated genes showed involvement in MAPK and WNT signaling pathways [48].

Foy et al. profiled 24 oral premalignant lesions (OPLs) by a high-throughput Agilent $4 \times 44k$ Custom CGH microarray-based methylated-CGI amplification method [73]. Among these 24 samples, 12 were from patients who later developed OSCC whereas the other 12 patients did not develop OSCC. They identified 146 probes across 86 unique genes that were all hypermethylated in OPLs subsequently transforming to OSCC. Among these 86 genes, *angiotensin II receptor type 1 (AGTR1)*, *Forkhead Box I2 (FOXI2)*, *HOXA9*, *proenkephalin (PENK)*, and *Zic family member 1 (ZIC1)* were selected for further validation using pyrosequencing. Methylation of *AGTR1*, *FOXI2*, and *PENK* in patients who developed OSCC was significantly higher than that in patients who did not develop OSCC. Average percentage methylation of these three genes was used to calculate a methylation index. Significantly poor oral cancer-free survival was seen in patients with higher methylation index. The authors also determined LINE1 methylation using pyrosequencing. The low methylation level of LINE1 was associated with poor oral cancer-free survival. This study suggested that the promoter methylation of *AGTR1*, *FOXI2*, and *PENK* and hypomethylation of LINE1 in OPLs might be indicators of malignant transformation [73].

Another recent study using Illumina Infinium Human Methylation450 array identified a unique methylation signature among OSCC patients in India [50]. This study identified the contribution of methylation in OSCC-associated gene regulation and performed pathway and survival analysis to show the relatedness of these genes [50]. Around 25% of the differentially methylated promoters were hypermethylated and remaining were hypomethylated. Some of the novel differentially methylated targets were validated in a new cohort of OSCC patients. Different methylation pattern of hyper- and hypomethylation across CGIs, shores, and shelves suggested an independent epigenetic regulation of differential methylation in OSCC in these two parts of the genome. Comparison of the TCGA methylation data with that of the Indian patients identified a set of unique differentially methylated probes, associated with immune regulation among Indian OSCC patients. It was found that some of the carcinogen-metabolizing genes, such as *cytochrome P450 family 8 subfamily B member 1 (CYP8B1)* and *glutathione S-transferase A3 (GSTA3)*, were hypomethylated. The common differentially methylated probes among Indian patients and TCGA data set were probably those required for OSCC development, whereas the unique hypomethylated probes are because of the different oral habits among Indian patients. This study identified 134 hypomethylated genes associated in OSCC. Survival analysis with gene expression data showed significantly better survival for 17 hypomethylated genes. Ingenuity pathway analysis of uniquely hypomethylated promoters showed enrichment of *cytotoxic T-lymphocyte associated protein 4 (CTLA4)* signaling and *interleukin 9 (IL9)* signaling pathways. CTLA4 is expressed by regulatory T cells to prevent further T cell activation and thus inhibits anti-tumor immune response, but in this study the survival analysis shows that higher *CTLA4* expression in early stages is connected to better survival in OSCC patients. Enrichment of immune genes for the unique hypomethylated promoter suggested infiltration of lymphocytes in the tumor microenvironment among OSCC patients in India [50].

CONCLUSION

In oral cancer there was strong evidence of involvement of epigenetic regulation [74]. Several studies reported key hypermethylated and hypomethylated genes in the development of OSCC [17]. But for the advancement in OSCC methylomics, several downstream functional aspects are important to explore. Several computational methods came up with some key findings. Identification of epigenetically regulated OSCC-associated genes in early stage can lead to a better survival for OSCC patients and this finding may lead to a better prognosis of OSCC in future. Both genetic and epigenetic alterations in oral cancer development are reported in the literature. However, integration of all these pieces of information along with the oral habits and other environmental variables is required to understand the actual molecular mechanism for OSCC development, pathogenesis, and therapeutic management. Rigorous computational and statistical tools are required to integrate such complex and large amount of data to elucidate the functional significance of these genetic and epigenetic processes.

REFERENCES

- [1] Llewellyn CD, Johnson NW, Warnakulasuriya KA. Risk factors for squamous cell carcinoma of the oral cavity in young people—a comprehensive literature review. *Oral Oncol* 2001;37(5):401–18.
- [2] Silverman Jr S, Gorsky M, Lozada F. Oral leukoplakia and malignant transformation. A follow-up study of 257 patients. *Cancer* 1984;53(3):563–8.

- [3] Warnakulasuriya S, Johnson NW, van der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. *J Oral Pathol Med* 2007;36(10):575–80.
- [4] Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;16(1):6–21.
- [5] Chatterjee R, Vinson C. CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim Biophys Acta* 2012;1819(7):763–70.
- [6] Vinson C, Chatterjee R. CG methylation. *Epigenomics* 2012;4(6):655–63.
- [7] Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003;(Suppl. 33):245–54.
- [8] Ushijima T, et al. Fidelity of the methylation pattern and its variation in the genome. *Genome Res* 2003;13(5):868–74.
- [9] Dillon N. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* 2006;14(1):117–26.
- [10] Baylin SB, Jones PA. A decade of exploring the cancer epigenome—biological and translational implications. *Nat Rev Cancer* 2011;11(10):726–34.
- [11] Esteller M. Epigenetics in cancer. *N Engl J Med* 2008;358(11):1148–59.
- [12] Esteller M. Cancer epigenetics for the 21st century: what's next? *Genes Cancer* 2011;2(6):604–6.
- [13] Rodriguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. *Nat Med* 2011;17(3):330–9.
- [14] Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 1983;301(5895):89–92.
- [15] Bock C. Epigenetic biomarker development. *Epigenomics* 2009;1(1):99–110.
- [16] Magić Z, et al. DNA methylation in the pathogenesis of head and neck cancer. In: Dricu A, editor. Methylation—from DNA, RNA and histones to diseases and treatment. INTECH; 2012.
- [17] Radhakrishnan R, Kabekkodu S, Satyamoorthy K. DNA hypermethylation as an epigenetic mark for oral cancer diagnosis. *J Oral Pathol Med* 2011;40(9):665–76.
- [18] Gonzalez-Ramirez I, et al. hMLH1 promoter methylation is an early event in oral cancer. *Oral Oncol* 2011;47(1):22–6.
- [19] Shaw RJ, et al. Promoter methylation of P16, RARbeta, E-cadherin, cyclin A1 and cytoglobin in oral cancer: quantitative evaluation using pyrosequencing. *Br J Cancer* 2006;94(4):561–8.
- [20] Gao S, et al. Epigenetic alterations of the SERPINE1 gene in oral squamous cell carcinomas and normal oral mucosa. *Genes Chromosomes Cancer* 2010;49(6):526–38.
- [21] Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev Cell* 2010;19(5):698–711.
- [22] Kulkarni V, Saranath D. Concurrent hypermethylation of multiple regulatory genes in chewing tobacco associated oral squamous cell carcinomas and adjacent normal tissues. *Oral Oncol* 2004;40(2):145–53.
- [23] Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 2010;11(3):191–203.
- [24] Bock C, Lengauer T. Computational epigenetics. *Bioinformatics* 2008;24(1):1–10.
- [25] Satterlee JS, Schubeler D, Ng HH. Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol* 2010;28(10):1039–44.
- [26] Bujold D, et al. The international human epigenome consortium data portal. *Cell Syst* 2016;3(5):496–499.e2.
- [27] Stunnenberg HG, International Human Epigenome C, Hirst M. The international human epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* 2016;167(5):1145–9.
- [28] Wei LK, Au A. Computational epigenetics. In: Handbook of epigenetics. 2nd ed 2017. p. 167–90.
- [29] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;13(10):705–19.

- [30] Ziller MJ, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature* 2013; 500(7463):477–81.
- [31] De Meyer T, et al. Genome-wide DNA methylation detection by MethylCap-seq and Infinium Human-Methylation450 BeadChips: an independent large-scale comparison. *Sci Rep* 2015;5:15375.
- [32] Meissner A, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005;33(18):5868–77.
- [33] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf* 2009;10:232.
- [34] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010;26(7):873–81.
- [35] Frith MC, Mori R, Asai K. A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res* 2012;40(13):e100.
- [36] Coarfa C, et al. Pash 3.0: a versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *BMC Bioinf* 2010;11:572.
- [37] Smith AD, et al. Updates to the RMAP short-read mapping software. *Bioinformatics* 2009;25(21):2841–2.
- [38] Xi Y, et al. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* 2012;28(3):430–2.
- [39] Otto C, Stadler PF, Hoffmann S. Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics* 2012;28(13):1698–704.
- [40] Nielsen R, et al. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011; 12(6):443–51.
- [41] Liu Y, et al. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* 2012;13(7):R61.
- [42] Bibikova M, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 2006;16(3):383–93.
- [43] Fan JB, et al. Illumina universal bead arrays. *Methods Enzymol* 2006;410:57–73.
- [44] Schmidt CT, Schmidt BL. Methylation array analysis of preoperative and postoperative saliva DNA in oral cancer patients. *Cancer Epidemiol Biomarkers Prev* 2008;17:3603–11.
- [45] Li YF, et al. DNA methylation profiles and biomarkers of oral squamous cell carcinoma. *Epigenetics* 2015; 10(3):229–36.
- [46] Jithesh PV. The epigenetic landscape of oral squamous cell carcinoma. *Br J Cancer* 2013;370–9.
- [47] Guerrero-Preston R, et al. NID2 and HOXA9 promoter hypermethylation as biomarkers for prevention and early detection in oral cavity squamous cell carcinoma tissues and saliva. *Cancer Prev Res (Phila)* 2011;4(7): 1061–72.
- [48] Towle R, et al. Global analysis of DNA methylation changes during progression of oral cancer. *Oral Oncol* 2013;49(11):1033–42.
- [49] Khor GH, et al. DNA methylation profiling revealed promoter hypermethylation-induced silencing of p16, DDAH2 and DUSP1 in primary oral squamous cell carcinoma. *Int J Med Sci* 2013;10(12):1727–39.
- [50] Basu B, et al. Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India. *Clin Epigenetics* 2017;9:13.
- [51] Touleimat N, Tost J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 2012;4(3):325–41.
- [52] Wang D, et al. Comparison of different normalization assumptions for analyses of DNA methylation data from the cancer genome. *Gene* 2012;506(1):36–42.
- [53] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8(1):118–27.

- [54] Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5(10):R80.
- [55] Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;30(10):1363–9.
- [56] Assenov Y, et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* 2014; 11(11):1138–40.
- [57] Morris TJ, et al. ChAMP: 450k Chip analysis methylation pipeline. *Bioinformatics* 2014;30(3):428–30.
- [58] Pedersen BS, et al. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* 2012;28(22):2986–8.
- [59] Wang D, et al. IMA: an R package for high-throughput analysis of Illumina’s 450K Infinium methylation data. *Bioinformatics* 2012;28(5):729–30.
- [60] Xu X, et al. A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics* 2013;8(5):522–33.
- [61] Jaffe AE, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 2012;41(1):200–9.
- [62] Butcher LM, Beck S. Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods* 2015;72:21–8.
- [63] Peters TJ, et al. De novo identification of differentially methylated regions in the human genome. *Epigenet Chromatin* 2015;8:6.
- [64] Down TA, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 2008;26(7):779–85.
- [65] Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* 2010;38(2):391–9.
- [66] Wang X, et al. Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell* 2009;21(4):1053–69.
- [67] Nagarajan RP, et al. Methods for cancer epigenome analysis. *Adv Exp Med Biol* 2013;754:313–38.
- [68] Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462(7271):315–22.
- [69] Robinson JT, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29(1):24–6.
- [70] Freese NH, Norris DC, Loraine AE. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* 2016;32(14):2089–95.
- [71] Wilhelm-Benartzi CS, et al. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer* 2013;109(6):1394–402.
- [72] Mascolo M, et al. Epigenetic disregulation in oral cancer. *Int J Mol Sci* 2012;13(2):2331–53.
- [73] Foy JP, et al. New DNA methylation markers and global DNA hypomethylation are associated with oral cancer development. *Cancer Prev Res (Phila)* 2015;8(11):1027–35.
- [74] Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 2002;3(6): 415–28.

This page intentionally left blank

COMPUTATIONAL EPIGENETICS FOR BREAST CANCER

15

Juan Xu, Yongsheng Li, Tingting Shao, Xia Li

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

INTRODUCTION

Breast cancer is a complex disease involving both genetic and epigenetic alterations [1,2]. Despite the development of genetic studies in breast cancer, advances have been made in exploring and identifying the potential role of epigenetic regulation in breast cancer. Epigenetics is defined as the heritable or transient changes in gene expression but without accompanied by alterations in the DNA sequence. DNA methylation, histone modification, and noncoding RNA (ncRNA) regulation are the major types of epigenetic changes that have been observed in breast cancer (Fig. 15.1A). Understanding these epigenetic changes and their potential roles in contribution to breast cancer is very important for diagnosis, prognosis, and therapy of breast cancer.

Over the past few years, increasing studies have been carried out for studying the epigenetics in breast cancer, reflected by the exponential increase of literature (Fig. 15.1B). Many genes were identified to be dysregulated by epigenetic alterations [3], such as estrogen response genes (*ESR1* and *PR*) [4], cell cycle inhibitor genes (*p16* and *RASSF1A*), DNA repair genes (*BRCA1*) [5], and many others. However, it is still difficult to identify the comprehensive epigenetic altered gene list by traditional low-throughput experimental methods. Hopefully, with the development of sequencing technology, computational epigenetic methods have been developed to accelerate our understanding of the epigenetic regulation in breast cancer [6]. In this chapter, we mainly summarized the recent development of computational studies in breast cancer.

DNA METHYLATION IN BREAST CANCER

The role of DNA methylation in human breast cancer has been well characterized. DNA methylation is mediated by the DNA methyltransferases, including *DNMT1*, *DNMT3a*, and *DNMT3b* [7]. Hypermethylation of promoter regions and subsequent gene silencing play an important role in promoting cancer development. Early methods primarily examine a limited set of genes with methylation-sensitive PCR or methylation-sensitive restriction enzyme analysis to identify the genes that show increased or decreased methylation level in breast cancer [8]. However, these methods were biased to well-known genes. Advances in technologies have produced lots of DNA methylation data sets for investigating the genome-wide methylation pattern in breast cancer, including sequencing of

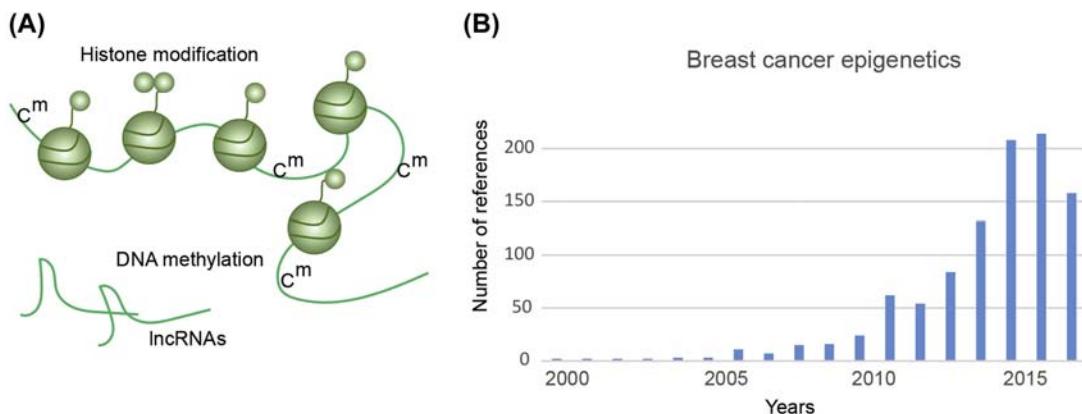


FIGURE 15.1 Epigenetics in breast cancer.

(A) The epigenetic involved DNA methylation, histone modification, and ncRNAs. (B) The number of literature for epigenetic studies in breast cancer.

bisulfite-converted DNA (BS-Seq) [9], methylated DNA immunoprecipitation (MeDIP) followed by sequencing [10], and dedicated Illumina 27K, 450K, and 850K arrays [11].

Since breast cancer involves multiple alterations in gene expression regulation, increasing studies are investigating a global mechanism of gene expression regulation. Pieces of evidence have shown that the gene expression is regulated by DNA methylation [12,13]. It is well known that two types of changes were observed in breast cancer: global hypomethylation and regional hypermethylation. The paradoxical coexistence of these two types of methylation changes implies that different processes are responsible for these changes. It is indicated that global hypomethylation is caused by the increase of demethylase activity while the regional hypermethylation results from local changes in chromatin structure [14]. The commonly used methods for identification of genes with methylation changes were based on *t*-test, Wilcoxon rank sum test, and some other methods specifically developed for methylation data. A significant amount of data has identified a list of genes (such as *p16*, *BRCA1*, *hMLH1*, and *HMSH2*) hypermethylated or hypomethylated in breast cancer. These methylation biomarkers were well correlated with breast cancer stage and have been proposed to be candidate diagnostic makers for breast cancer.

Breast cancer is a heterogeneous disease, involving both genetic and epigenetic alterations. The molecular background behind breast cancer is not well understood and at least four major subtypes with distinct expression patterns are identified, including basal-like, ERBB2+, luminal A, and luminal B. Epigenetic analyses have identified aberrant DNA methylation signatures that are related with molecular subtypes of breast cancer [15,16]. However, there are limited studies for identifying global methylation changes associated with breast cancer subtypes. Bediaga et al. analyzed DNA methylation in 806 cancer-related genes in breast cancer samples and recognized 15 CpG loci differentially methylated in breast cancer subtypes [15]. These studies suggest that DNA methylation is an important biomarker in breast cancer.

Despite growing appreciation of the importance of epigenetics of protein-coding genes in breast cancer, our understanding of epigenetic alterations of noncoding RNAs (ncRNAs) in breast cancer

remains poorly understood [17–19]. Recently, we explored the epigenetic alterations of ncRNAs in breast cancers using next-generation sequencing-based methylation data (Fig. 15.2), primarily focusing on the two most commonly studied ncRNA biotypes, long ncRNAs and miRNAs [20]. Widely aberrant methylation in the promoters of ncRNAs was observed, which was more frequent than that in protein-coding genes. Specifically, intergenic ncRNAs were observed to comprise a majority (51.45% of the lncRNAs and 51.57% of the miRNAs) of the aberrantly methylated ncRNA promoters. Most importantly, we summarized five patterns of aberrant ncRNA promoter methylation in the context of CpG islands (CGIs). We found that aberrant methylation occurred not only on CGIs, but also in regions flanking CGI and in CGI-lacking promoters. Integration with expression profile data sets enabled us to determine that the ncRNA promoter methylation events were associated with transcriptional changes. Furthermore, a panel of ncRNAs were identified as biomarkers that discriminated between disease phenotypes. This study represents a highly valuable public resource for understanding the epigenetic regulation of the breast cancer genome and for identifying lncRNAs and miRNAs as therapeutic targets.

HISTONE MODIFICATION IN BREAST CANCER

In addition to DNA methylation, histone modification is another type of well-known epigenetic modification to be altered in breast cancer, and loss of selected histone acetylation and methylation marks has recently been shown to play critical roles in breast cancer. Elsheikh et al. used immunohistochemistry to detect a series of histone lysine acetylation, lysine methylation, and arginine methylation marks in a series of human breast cancers ($n = 880$). Their analyses revealed low or absent H4K16ac in the majority (78.9%) of breast cancer samples, suggesting that this type of histone modification alteration may represent an early event of breast cancer [21]. In addition, there was a highly significant correlation between histone modification status, tumor biomarker phenotype, and clinical outcome. Specifically, high histone acetylation and methylation were associated with a favorable prognosis, and moderate to low levels of lysine acetylation and lysine and arginine methylation were observed in patients of poorer prognostic subtypes. This study identifies the presence of variations in global levels of histone marks in different classes of breast cancer and shows that these differences were associated with clinical significance.

All these observations point that broad epimodification alteration events emerge in cancer cells and offer partially different visions of how histone modifications are perturbed in various subtypes of breast cancer. However, the involvement of the histone modification in breast cancer and its contribution to the subtypes of the breast cancer are still poorly understood. One practical way to address the histone alterations in breast cancer is through conducting an analysis that integrates multiple types of epigenetic and transcriptomic data from the same individual cancer cells. The development of high-throughput sequencing technology provides us an opportunity to investigate the epigenetic mechanism in breast cancer subtypes at a high resolution. In one of our recent studies, we have portrayed and compared the epigenetic alterations of six types of histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, and H3K27me3) and DNA methylation between two commonly used cell lines that represent two different breast cancer subtypes—luminal and basal (Fig. 15.3). Widespread distinct patterns of epimodification alterations in breast cancer subtypes, particularly in the promoter regions, were revealed in a genome-wide analysis of epigenetic alterations between two breast cancer subtypes [22]. In addition, we found a combinatory effect

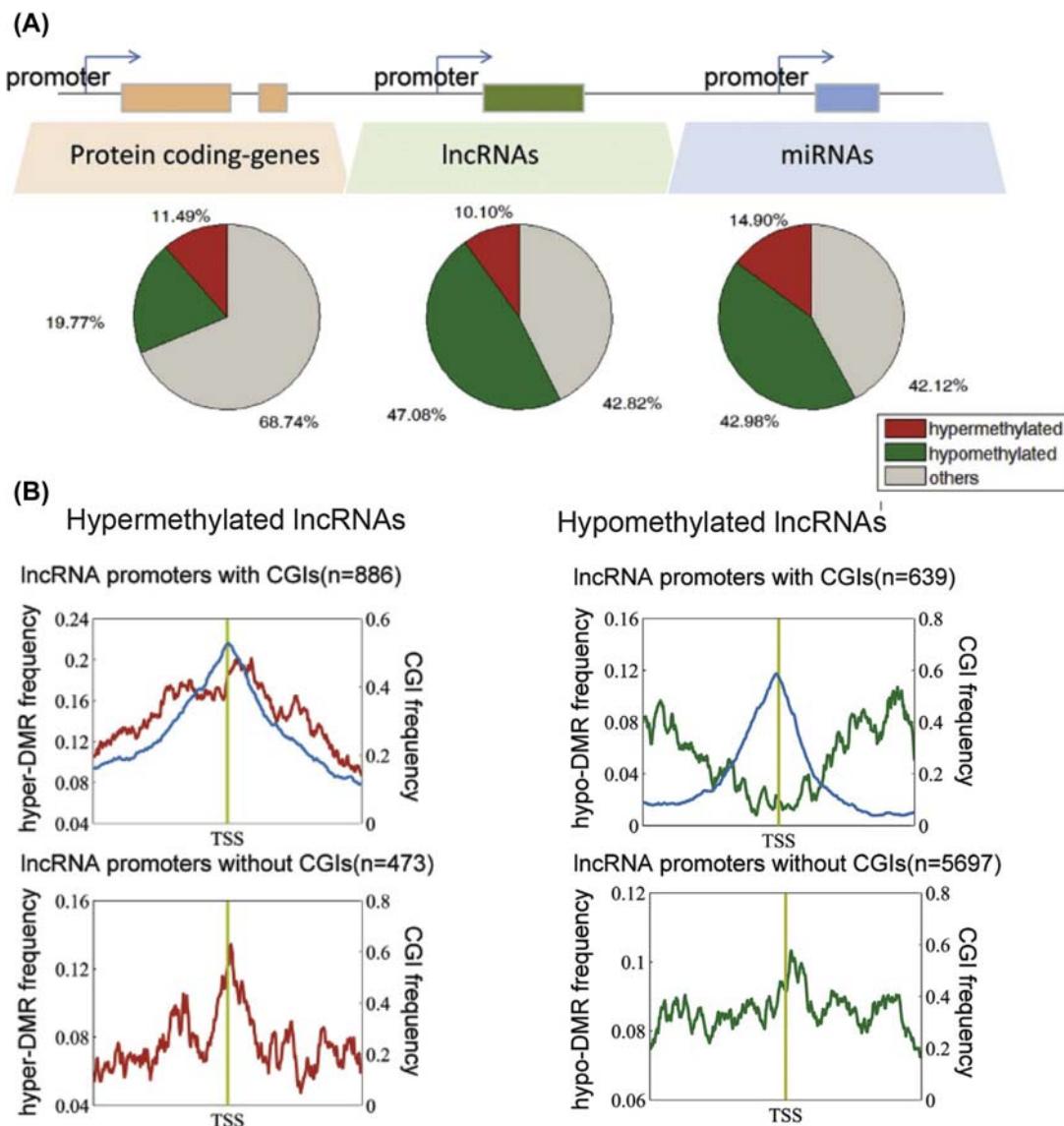


FIGURE 15.2 Aberrant methylation of ncRNAs in breast cancer.

(A) The proportion of aberrantly methylated lncRNAs, miRNAs, and coding genes. (B) The methylation pattern of ncRNAs in breast cancer.

of different epigenetic modifications on gene expression in both subtypes. Finally, we showed that alterations of epimodifications in the two breast cancer subtypes not only affect genes involved in common cancer biology, but also regulate genes participating in subtype-specific cancer hallmark functions. Through integrative analysis of different genetic and epigenetic data types, we have

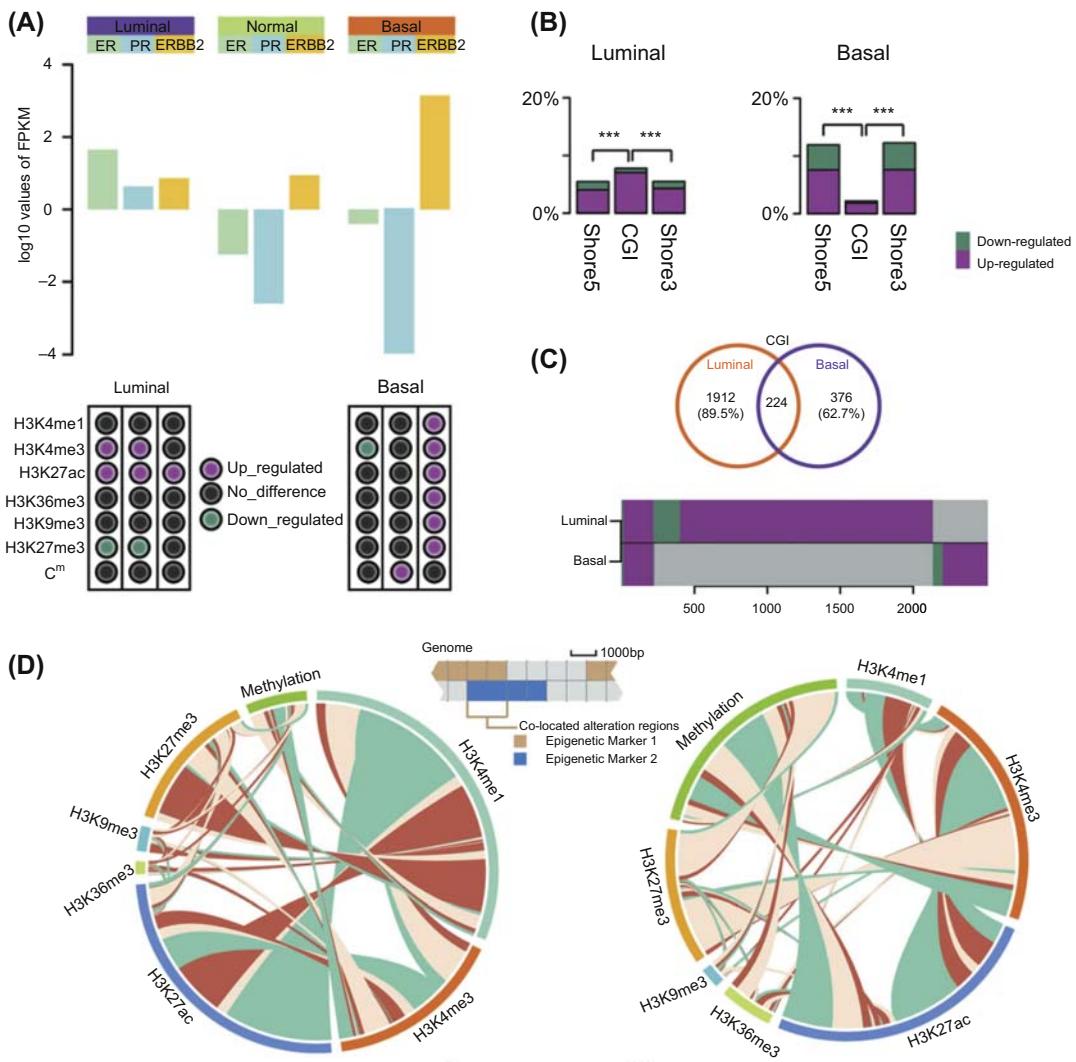


FIGURE 15.3 Epigenetic dysregulation of genes in breast cancer subtypes.

(A) The epigenetic dysregulation of marker genes in breast cancer subtypes. (B and C) Different methylation pattern of CGI in breast cancer subtypes. (C and D) Different combination of epigenetic regulation in breast cancer subtypes.

revealed distinct patterns of oncogenic pathway activation in different breast cancer subtypes and provided novel insights into subtype-specific therapeutic opportunities for breast cancer.

Besides protein-coding genes, systematic genomic studies have identified a broad spectrum of noncoding RNAs (ncRNAs) that are involved in breast cancer. However, our understanding of the epigenetic dysregulation of those ncRNAs (such as lncRNAs and miRNAs) remains limited [23,24].

Recently, we systematically analyzed the epigenetic alterations (including DNA methylation and histone modification) of miRNAs and lncRNAs in two breast cancer subtypes—luminal and basal (Fig. 15.4) [25]. Specifically, widespread epigenetic alterations of miRNAs and lncRNAs were observed in both breast cancer subtypes. In contrast to protein-coding genes, we found that the majority of epigenetically dysregulated ncRNAs were shared between these two subtypes, but a subset of transcriptomic and corresponding epigenetic changes of ncRNAs occurred in a breast cancer subtype-specific manner. In addition, these findings suggested that various types of epimodifications might synergistically regulate the expression of ncRNAs. Our observations further highlighted the complementary dysregulation roles of epimodifications, particularly for miRNA members within the same family. These epigenetic alterations produced the same directed expression alterations by diverse epimodifications. Functional enrichment analysis also revealed that these epigenetically dysregulated lncRNAs and miRNAs were significantly involved in several hallmarks of cancers. Finally, by analyzing the epigenetic modification-mediated miRNA regulatory networks, we revealed that cancer progression was associated with breast cancer specific miRNA-gene modules. This study enhances our understanding of the aberrant epigenetic patterns of ncRNA expression and provides novel insights into the functions of ncRNAs in breast cancer subtypes.

NONCODING RNA REGULATION IN BREAST CANCER

Noncoding RNAs (ncRNAs) are a class of RNAs that regulate gene expression transcriptionally and posttranscriptionally. Most of current studies have been focusing on two types of ncRNAs, including miRNA and lncRNAs [26,27]. miRNAs are a class of ~22-nucleotide long single-stranded noncoding RNAs that regulate gene expression by binding to miRNA response elements (MREs) on the RNAs. In addition, recent advances in tiling arrays and RNA deep sequencing (RNA-Seq) have revealed many thousands of long noncoding RNAs (lncRNAs) greater than 200 nucleotides (nt) in length. These ncRNAs affect many biological processes, including regulation of gene expression, genomic imprinting, and nuclear organization [28,29]. However, we have limited knowledge of the functions of ncRNAs in cancer.

Aberrant expression of ncRNAs has been observed in many types of cancer [30,31], including breast cancer and its subtypes. Blenkiron et al. detected different miRNA expression levels between the basal and luminal subtypes by profiling 309 miRNAs in 93 breast tumors [32]. In addition, de Rinaldis et al. also identified an miRNA signature formed by 46 miRNAs that could be used to differentiate between breast cancer subtypes [33]. Dvinge et al. obtained similar findings in their studies [34]. In a metaanalysis of independent trials, various subtype-specific miRNAs were identified, including luminal-A signature (let-7c, miR-10a, and let-7f), basal signature (miR-18a, miR-135b, miR-93, and miR-155), and HER2 signature (miR-142-3p and miR-150). Moreover, 453 miRNAs in 29 early stage breast cancer tumors were also profiled, and miRNA signatures that could be used to accurately predict the ER, PR, and HER2 status of breast tumor were identified.

In addition to the conventional miRNA-RNA regulation, increasing studies have shown that regulation among miRNA seed region and mRNA is not unidirectional [35], but that the pool of RNAs can crosstalk with each other through competing for miRNA binding. These competitive endogenous RNAs (known as ceRNAs) act as molecular sponges for an miRNA through their MRE, thereby regulating other target genes of the respective miRNAs [36]. Understanding this novel type of RNA crosstalk will lead to significant insight into regulatory networks and has been implied in human

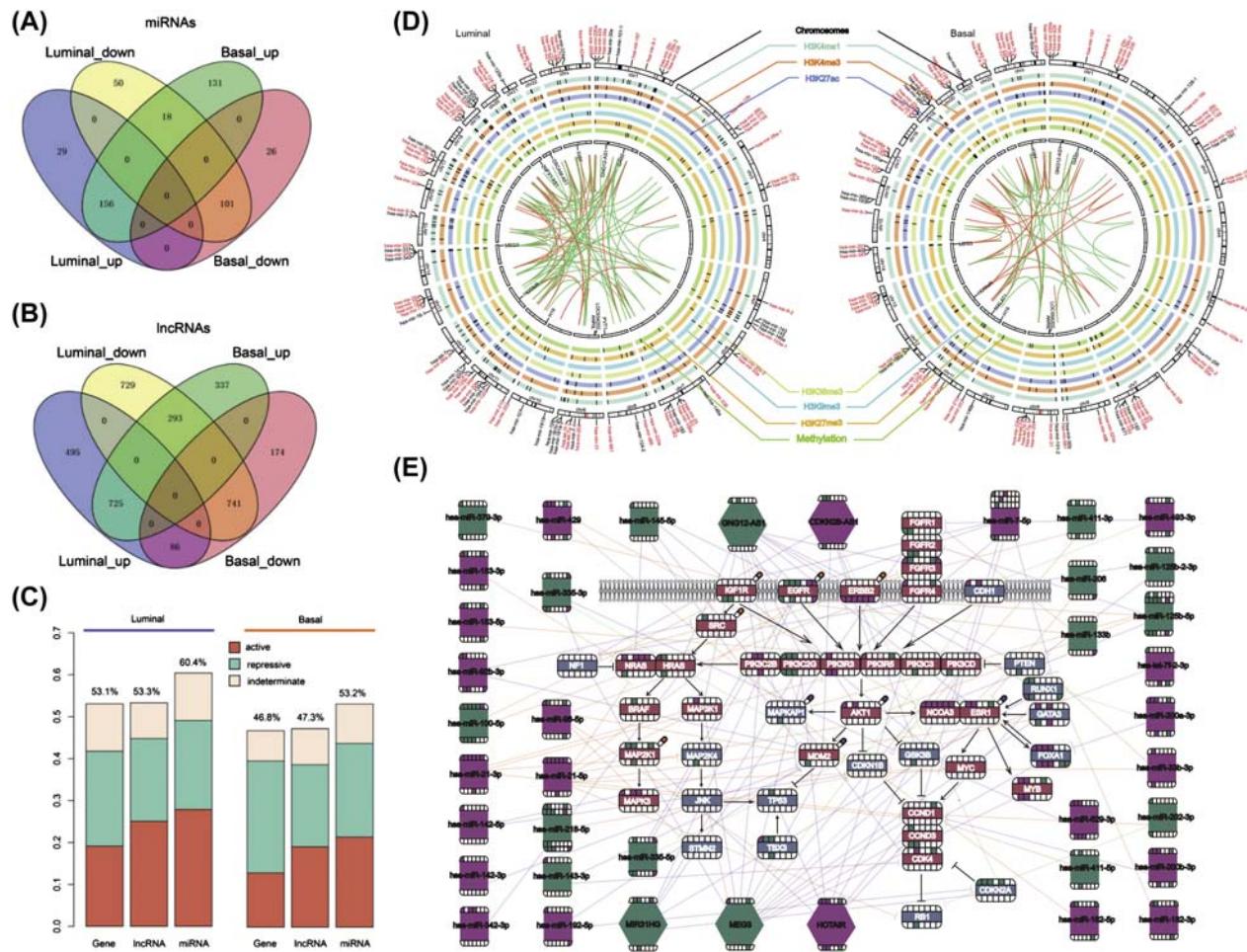


FIGURE 15.4 Epigenetic dysregulation of ncRNAs in breast cancer subtypes.

(A) The overlap of differentially expressed miRNAs in breast cancer subtypes. (B) The overlap of differentially expressed lncRNAs in breast cancer subtypes. (C) The proportion of epigenetically dysregulated genes in breast cancer subtypes. (D) The epigenetically dysregulated lncRNAs and miRNAs in breast cancer subtypes. (E) Epigenetically dysregulated ncRNAs and genes in breast cancer-related pathways.

breast cancer development and other complex diseases. For example, Li et al. found that a STARD13-correlated ceRNA network inhibits epithelial-mesenchymal transition (EMT) and metastasis in breast cancer [37]. However, this study only focused on the roles of individual ceRNA interactions and did not investigate the ceRNA interactions at a system level. Xu et al. conducted a systemic analysis of ceRNA interactions in breast cancer, but without considering the heterogeneity of breast cancer [26]. Moreover, several studies investigating specific ceRNA interactions in human cancer have focused on dysregulated RNAs that are aberrantly expressed during cancer initiation and progression [38]. Due to the heterogeneity of breast cancer, the transcriptome of breast cancer is more complex than what we expected [39,40]. Hence, a systemic analysis of ceRNA crosstalk among different breast cancer subtypes may yield novel insights regarding the interplay of various biological networks that are involved in breast cancer [41]. Recently, we generated a ceRNA network for each breast cancer subtype based on the significance of both positive co-expression and the shared miRNAs in the corresponding subtype miRNA dysregulatory network, which was constructed based on negative regulation between differentially expressed miRNAs and target genes (Fig. 15.5) [42]. All four subtype-specific ceRNA networks exhibited scale-free architecture and showed that common ceRNAs were located at the core of these networks. Furthermore, the common ceRNA hubs (nodes with higher number of connections) had greater connectivity than the subtype-specific hubs. Functional analysis of the common subtype ceRNA hubs highlighted critical factors involved in proliferation, MAPK signaling pathways, and tube morphogenesis. Subtype-specific ceRNA hubs highlighted uniquely subtype-specific pathways, like the estrogen response and inflammatory pathways in the luminal subtypes or the genes involved in the coagulation process that participates in the basal-like subtype. Finally, we also identified 29 critical subtype-specific ceRNA hubs that characterized different breast cancer subtypes. This study thus provides new insight into the common and specific subtype ceRNA interactions that define the different subtypes of breast cancer, and enhances our understanding of the pathology underlying different breast cancer subtypes.

EPIGENETIC DATABASES

With the development of high-throughput sequencing technology, various types of epigenetic modification data sets are generated for various types of cancer, including breast cancer. The cancer genomic projects such as The Cancer Genome Atlas (TCGA) provided the DNA methylation data sets for many cancer types. In addition, several databases are set up for specific functional studies of epigenetic in cancer. Here, we summarized the commonly used database in Table 15.1. These databases provided valuable resources as well as easy-to-use tools for viewing the epigenetic alterations in cancer.

EPIGENETIC TOOLS IN CANCER

With the increase of high-throughput sequencing data sets in epigenetics, lines of computational tools were developed to analyze these data for the identification of differentially methylated regions (DMR) or genomic regions with different histone modifications. Here, we summarized the commonly used tools in Table 15.2 and illustrated the usefulness of these tools.

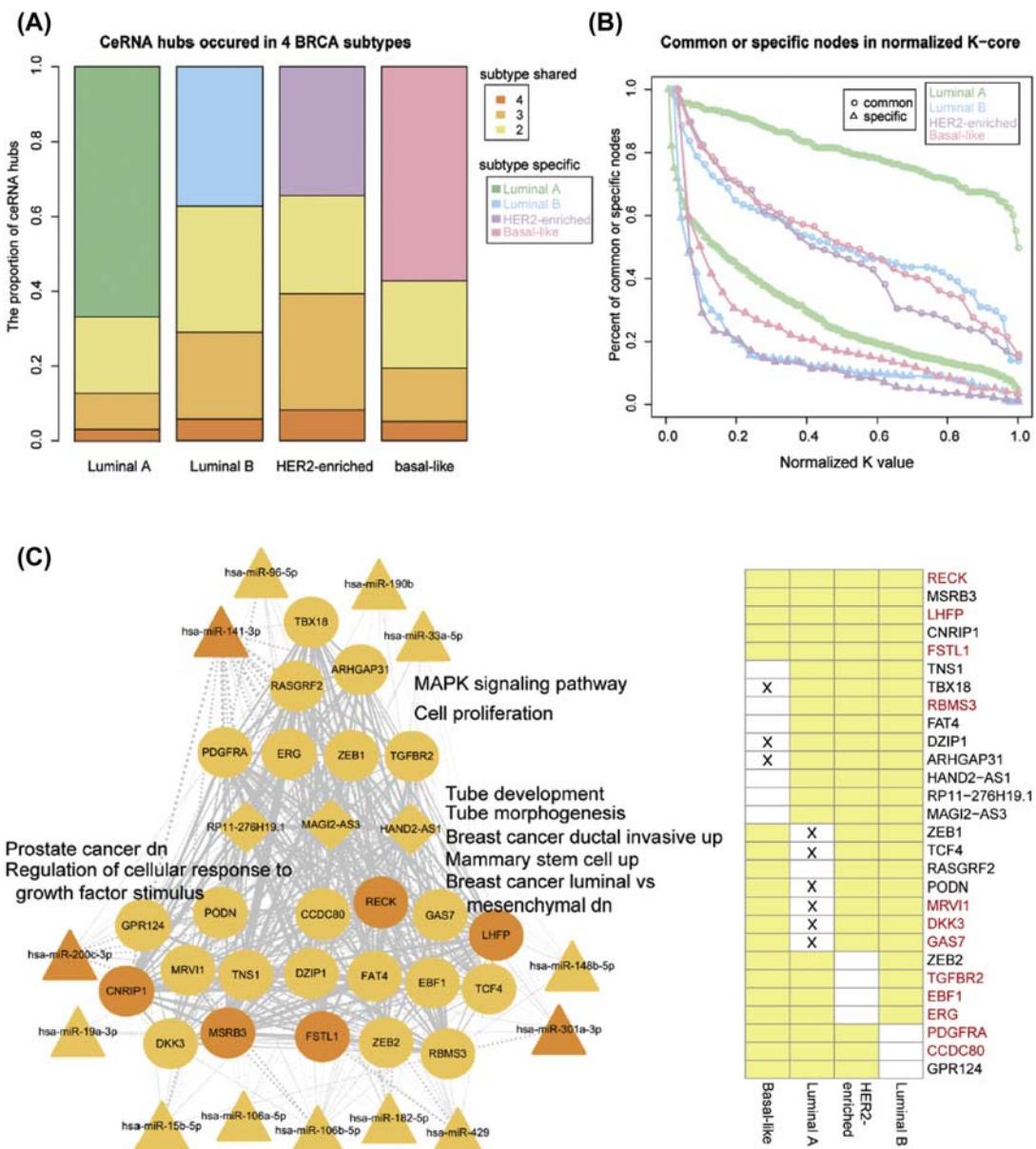


FIGURE 15.5 ceRNA regulation in breast cancer subtypes.

(A) The ceRNA hubs in breast cancer subtypes. (B) The number of ceRNA modules in breast cancer subtypes. (C) The representative ceRNA module in breast cancer.

Table 15.1 Epigenetic Databases for Cancer

Databases	Epigenetic	Description	Links	Ref.
MethHC	DNA methylation	Integration of a large collection of DNA methylation data and mRNA/microRNA expression profiles in human cancer	http://MethHC.mbc.nctu.edu.tw	[43]
Wanderer	DNA methylation	An interactive viewer to explore DNA methylation and gene expression data in human cancer	http://maplab.cat/wanderer	[44]
DiseaseMeth	DNA methylation	An interactive database that aims to present the most complete collection and annotation of aberrant DNA methylation in human diseases, especially various cancers	http://bioinfo.hrbmu.edu.cn/diseasemeth/	[45]
MethyCancer	DNA methylation	MethyCancer hosts both highly integrated data of DNA methylation, cancer-related gene, mutation and cancer information from public resources, and the CpG island (CGI) clones derived from our large-scale sequencing	http://methycancer.genomics.org.cn	[46]
NGSmethDB	DNA methylation	A repository for single-base whole-genome methylome maps for the best-assembled eukaryotic genomes	http://bioinfo2.ugr.es/NGSmethDB	[47]
ENCODE	Methylation/histone modification	A comprehensive parts list of functional elements in the human genome	https://www.encodeproject.org/	[48]
miRCancer	miRNA	Comprehensive collection of microRNA (miRNA) expression profiles in various human cancers which are automatically extracted from published literature in PubMed	http://mircancer.ecu.edu/	[49]
miR2Disease	miRNA	A manually curated database for microRNA deregulation in human disease	http://watson.compbio.iupui.edu:8080/miR2Disease/analysis.jsp	[50]
PanceRNADB	ceRNA	The mRNA related ceRNA—ceRNA landscape and significance across 20 major cancer types	http://www.bio-bigdata.com/pan-cernadb/	[26]

Table 15.2 Epigenetic Tools in Cancer

Tools	Description	Links	Ref.
QDMR	A quantitative approach to quantify methylation difference and identify DMRs from genome-wide methylation profiles by adapting Shannon entropy	http://fame.edbc.org/qdmr/	[51]
CpG_MP	Identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data	http://bioinfo.hrbmu.edu.cn/CpG_MP	[52]
methylKit	An R package for DNA methylation analysis and annotation from high-throughput bisulfite sequencing	https://github.com/al2na/methylKit/issues	[53]
RRBS-Analyser	A comprehensive genome-scale DNA methylation analysis server based on RRBS data	http://122.228.158.106/RRBSAnalyser/	[54]
eDMR	Comprehensive DMR analysis based on bimodal normal distribution model and weighted cost function for regional methylation analysis optimization	https://code.google.com/archive/p/edmr/issues	[55]
THOR	An HMM-based approach to detect and analyze differential peaks in two sets of ChIP-seq data from distinct biological conditions with replicates	http://www.regulatory-genomics.org/thor-2/basic-instruction/	[56]
HMCan	It is a hidden Markov model based tool developed to detect histone modification in cancer ChIP-seq data	http://www.cbrc.kaust.edu.sa/hmcan/	[57]
HHMD	A relatively comprehensive database for human histone modifications	http://bioinfo.hrbmu.edu.cn/hhmd	[58]
HistoneDB	Holds canonical histones and histone variants, their sequence, structural and functional features	https://www.ncbi.nlm.nih.gov/research/HistoneDB2.0/index.fcgi/help/	[59]

FUTURE DIRECTIONS

Breast cancer is related with the aberrant alterations of epigenetics, including DNA methylation, histone modification, and ncRNA regulation. Increasing evidence has demonstrated that targeting epigenetic changes in breast cancer is an exciting and evolving arena. Epigenetic-based therapy methods have the potential to change our current standard-of-care therapies in breast cancer. In this chapter, we have summarized the complex epigenetic regulatory alterations and pathways that play

key roles in breast cancer diagnosis and therapeutics. Particularly, it is promising for therapy of breast cancer is the combination of epigenetic therapy with the current cytotoxic and endocrine therapies in breast cancer. However, many uncertainties still remain, such as to how to best translate these epigenetic findings in the clinical arena, and how to use these identified epigenetic biomarkers in breast cancer. Given the epigenetic importance, more studies are needed to be ongoing to help define the optimal treatment schedules for breast cancer.

REFERENCES

- [1] Byler S, Goldgar S, Heerboth S, et al. Genetic and epigenetic aspects of breast cancer progression and therapy. *Anticancer Res* 2014;34:1071–7.
- [2] Wu Y, Sarkissyan M, Vadgama JV. Epigenetics in breast and prostate cancer. *Methods Mol Biol* 2015;1238:425–66.
- [3] Ambatipudi S, Horvath S, Perrier F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer* 2017;75:299–307.
- [4] Martinez-Galan J, Torres-Torres B, Nunez MI, et al. ESR1 gene promoter region methylation in free circulating DNA and its correlation with estrogen receptor protein expression in tumor tissue in breast cancer patients. *BMC Cancer* 2014;14:59.
- [5] Gong C, Fujino K, Monteiro LJ, et al. FOXA1 repression is associated with loss of BRCA1 and increased promoter methylation and chromatin silencing in breast cancer. *Oncogene* 2015;34:5012–24.
- [6] Cava C, Bertoli G, Castiglioni I. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. *BMC Syst Biol* 2015;9:62.
- [7] Robertson KD. DNA methylation, methyltransferases, and cancer. *Oncogene* 2001;20:3139–55.
- [8] Hernandez HG, Tse MY, Pang SC, et al. Optimizing methodologies for PCR-based DNA methylation analysis. *Biotechniques* 2013;55:181–97.
- [9] Krueger F, Kreck B, Franke A, et al. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* 2012;9:145–51.
- [10] Zhao MT, Whyte JJ, Hopkins GM, et al. Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. *Cell Reprogram* 2014;16:175–84.
- [11] Dedeurwaerder S, Defrance M, Calonne E, et al. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011;3:771–84.
- [12] Terry MB, McDonald JA, Wu HC, et al. Epigenetic biomarkers of breast cancer risk: across the breast cancer prevention continuum. *Adv Exp Med Biol* 2016;882:33–68.
- [13] Yan PS, Venkataramu C, Ibrahim A, et al. Mapping geographic zones of cancer risk with epigenetic biomarkers in normal breast tissue. *Clin Cancer Res* 2006;12:6626–36.
- [14] Ehrlich M. DNA hypomethylation in cancer cells. *Epigenomics* 2009;1:239–59.
- [15] Bediaga NG, Acha-Sagredo A, Guerra I, et al. DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Res* 2010;12:R77.
- [16] Szyf M. DNA methylation signatures for breast cancer classification and prognosis. *Genome Med* 2012;4:26.
- [17] McGuire A, Brown JA, Kerin MJ. Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. *Cancer Metastasis Rev* 2015;34:145–55.
- [18] Mulrane L, McGee SF, Gallagher WM, et al. miRNA dysregulation in breast cancer. *Cancer Res* 2013;73:6554–62.
- [19] Kristensen VN, Lingjaerde OC, Russnes HG, et al. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* 2014;14:299–313.

- [20] Li Y, Zhang Y, Li S, et al. Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. *Sci Rep* 2015;5:8790.
- [21] Elsheikh SE, Green AR, Rakha EA, et al. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. *Cancer Res* 2009;69:3802–9.
- [22] Li Y, Li S, Chen J, et al. Comparative epigenetic analyses reveal distinct patterns of oncogenic pathways activation in breast cancer subtypes. *Hum Mol Genet* 2014;23:5378–93.
- [23] Li Y, Chen H, Pan T, et al. LncRNA ontology: inferring lncRNA functions based on chromatin states and expression patterns. *Oncotarget* 2015;6:39793–805.
- [24] Li Y, Camarillo C, Xu J, et al. Genome-wide methylome analyses reveal novel epigenetic regulation patterns in schizophrenia and bipolar disorder. *Biomed Res Int* 2015;2015:201587.
- [25] Xu J, Wang Z, Li S, et al. Combinatorial epigenetic regulation of non-coding RNAs has profound effects on oncogenic pathways in breast cancer subtypes. *Brief Bioinformatics* 2018 Jan 1;19(1):52–64.
- [26] Xu J, Li Y, Lu J, et al. The mRNA related ceRNA-ceRNA landscape and significance across 20 major cancer types. *Nucleic Acids Res* 2015;43:8169–82.
- [27] Li Y, Chen J, Zhang J, et al. Construction and analysis of lncRNA-lncRNA synergistic networks to reveal clinically relevant lncRNAs in cancer. *Oncotarget* 2015;6:25003–16.
- [28] Li Y, Wang Z, Wang Y, et al. Identification and characterization of lncRNA mediated transcriptional dysregulation dictates lncRNA roles in glioblastoma. *Oncotarget* 2016;7:45027–41.
- [29] Xu J, Li CX, Li YS, et al. MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res* 2011;39:825–36.
- [30] Li Y, Xu J, Chen H, et al. Comprehensive analysis of the functional microRNA-mRNA regulatory network identifies miRNA signatures associated with glioma malignant progression. *Nucleic Acids Res* 2013;41:e203.
- [31] Xu J, Li CX, Lv JY, et al. Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol Cancer Ther* 2011;10:1857–66.
- [32] Blenkiron C, Goldstein LD, Thorne NP, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol* 2007;8:R214.
- [33] de Rinaldis E, Gazinska P, Mera A, et al. Integrated genomic analysis of triple-negative breast cancers reveals novel microRNAs associated with clinical and molecular phenotypes and sheds light on the pathways they control. *BMC Genome* 2013;14:643.
- [34] Dvinge H, Git A, Graf S, et al. The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* 2013;497:378–82.
- [35] Xu J, Feng L, Han Z, et al. Extensive ceRNA-ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues. *Nucleic Acids Res* 2016;44:9438–51.
- [36] Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature* 2014;505:344–52.
- [37] Li X, Zheng L, Zhang F, et al. STARD13-correlated ceRNA network inhibits EMT and metastasis of breast cancer. *Oncotarget* 2016 Apr 26;7(17):23197–211.
- [38] Huang M, Zhong Z, Lv M, et al. Comprehensive analysis of differentially expressed profiles of lncRNAs and circRNAs with associated co-expression and ceRNA networks in bladder carcinoma. *Oncotarget* 2016 Jul 26;7(30):47186–200.
- [39] DeSantis C, Ma J, Bryan L, et al. Breast cancer statistics, 2013. *CA Cancer J Clin* 2014;64:52–62.
- [40] Omberg L, Ellrott K, Yuan Y, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* 2013;45:1121–6.
- [41] Davidson EH. Emerging properties of animal gene regulatory networks. *Nature* 2010;468:911–20.
- [42] Chen J, Xu J, Li Y, et al. Competing endogenous RNA network analysis identifies critical genes among the different breast cancer subtypes. *Oncotarget* 2017;8:10171–84.

- [43] Huang WY, Hsu SD, Huang HY, et al. MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic Acids Res* 2015;43:D856–61.
- [44] Diez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenet Chromatin* 2015;8:22.
- [45] Xiong Y, Wei Y, Gu Y, et al. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res* 2017;45:D888–95.
- [46] He X, Chang S, Zhang J, et al. MethylCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res* 2008;36:D836–41.
- [47] Lebron R, Gomez-Martin C, Carpena P, et al. NGSmethDB 2017: enhanced methylomes and differential methylation. *Nucleic Acids Res* 2017;45:D97–103.
- [48] Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018 Jan 4;46(D1):D794–801.
- [49] Xie B, Ding Q, Han H, et al. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 2013;29:638–44.
- [50] Jiang Q, Wang Y, Hao Y, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;37:D98–104.
- [51] Zhang Y, Liu H, Lv J, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res* 2011;39:e58.
- [52] Su J, Yan H, Wei Y, et al. CpG_MP: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res* 2013;41:e4.
- [53] Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;13:R87.
- [54] Wang T, Liu Q, Li X, et al. RRBS-analyser: a comprehensive web server for reduced representation bisulfite sequencing data analysis. *Hum Mutat* 2013;34:1606–10.
- [55] Li S, Garrett-Bakelman FE, Akalin A, et al. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* 2013;14(Suppl. 5):S10.
- [56] Allhoff M, Sere K, Pires JF, et al. Differential peak calling of ChIP-seq signals with replicates with THOR. *Nucleic Acids Res* 2016;44:e153.
- [57] Ashoor H, Herault A, Kamoun A, et al. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics* 2013;29:2979–86.
- [58] Zhang Y, Lv J, Liu H, et al. HHMD: the human histone modification database. *Nucleic Acids Res* 2010;38:D149–54.
- [59] Draizen EJ, Shaytan AK, Marino-Ramirez L, et al. Histone DB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database (Oxford)* Mar 17;2016. pii: baw014.

INTEGRATIVE EPIGENOMICS OF PROSTATE CANCER 16

Madonna Peter^{1,2}, Shivani Kamdar^{1,2}, Bharati Bapat^{1,2,3}

¹*Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada;* ²*Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada;* ³*Division of Urology, University of Toronto, Toronto, ON, Canada*

PROSTATE CANCER: AN OVERVIEW

Prostate cancer (PCa) is the most common cancer and a major cause for cancer-related deaths among men [1], with an estimated 1.6 million cases and 366,000 deaths worldwide in 2015 alone [2]. Despite high incidence, largely due to prostate-specific antigen (PSA) based screening [3–5], the majority of patients are diagnosed with localized, low-risk PCa [6,7]. Therefore, distinguishing aggressive disease from indolent PCa remains an ongoing clinical challenge [8]. Current diagnostic and prognostication strategies/nomograms are often limited to clinical and histopathological grading systems, such as Gleason score, PSA levels, and TNM staging [9,10]. These clinical parameters are unable to fully capture the disease spectrum across PCa [11,12]. Furthermore, for the most lethal form of the disease, metastatic castration-resistant prostate cancer (mCRPC), clinical guidelines for optimal patient-specific treatment sequences are lacking [13]. As a result, there has been considerable effort to investigate underlying genetic and epigenetic molecular drivers [14] to derive biomarkers and novel therapeutic approaches. Indeed, large-scale genome-wide initiatives, such as The Cancer Genome Atlas network (TCGA) [15], have revealed molecular subtypes within PCa that could not be stratified by conventional clinical assessments alone. As a result, a complex interplay between genomic and epigenomic aberrations is emerging, warranting analysis tools that can integrate these data sets and infer the functional consequences of these interactions in PCa. In this chapter, we will first highlight key features of the PCa genome and epigenome as well as current understanding of their interaction, followed by an overview of recent cutting-edge tools developed for integrative analysis of the epigenome in PCa.

GENOMIC ALTERATIONS IN PCa

Whole exome and genome sequencing have been vital tools in deciphering genomic alterations in both primary and metastatic PCa lesions [15–19]. Collectively, several genes and mutation types have been implicated, including single-nucleotide variants (SNVs), copy number alterations (CNAs), gene fusions, and chromosomal rearrangements/structural variants (SVs) [20]. There is no stand-alone driver

gene that leads to PCa development, but rather subsets of recurrent mutational events in multiple genes [15]. Overall, primary PCa tumors exhibit more CNAs and chromosomal rearrangements than somatic SNVs [15,19]. Representative recurrent CNAs found within these tumors include *MYC* amplifications and loss of *NKX3.1*, *PTEN*, *RBI*, and *TP53* [16]. Additional mutations implicating genes such as *SPOP*, *FOXA1*, and *MEDI12* have also been reported in localized PCa [17,18].

The most common SVs (approximately 40%–50% or more of PCa tumors) are gene fusions involving the 5'-UTR of the androgen-induced transmembrane protease serine 2 gene, *TMPRSS2*, to members of the E26 transformation-specific (ETS) family of transcription factors, with *TMPRSS2-ERG* fusions being the most frequent [21–23]. Although *TMPRSS2-ERG* fusion status is not prognostic, it is linked to prevalence of complex genomic rearrangements in PCa [18,24,25]. Chromoplexy, or looped chromosomal rearrangement involving at least three loci, occurs more often in *TMPRSS2-ERG* fusion-positive tumors, usually at transcriptionally active sites [18]. In contrast, *TMPRSS2-ERG* fusion-negative tumors tend to favor chromothripsis—large numbers of double-stranded breaks and rearrangement occurring within localized regions—within inactive chromatin sites. This phenomenon is especially prevalent in tumors exhibiting *CHD1* deletion. Frequency of chromoplexy is estimated to be as high as 88%, while chromothripsis is observed in 13%–20% of PCa cases [18,19,26]. Kataegis, or localized hypermutation, is present in approximately 20% of tumors and more likely to occur in high-risk PCa cases; however, neither chromothripsis nor kataegis is significantly associated with biochemical recurrence following localized treatment (i.e., radical prostatectomy) [19].

In addition to clinical heterogeneity in disease outcome, genomic heterogeneity between patients sharing the same pathological scores and intrapatient heterogeneity due to the multifocal origin of PCa tumors further add to the complexity of the genomic landscape of PCa [27–29]. In contrast to localized PCa, advanced/mCRPC tumors are marked by increases in overall mutational rates and diversity in types of mutations [30]. As the mCRPC clinical state arises from resistance to androgen-deprivation therapy, mutations in the androgen receptor (AR) gene (i.e., amplifications) as well as genes involved in the PI3K, Wnt, and AR signaling pathways have been implicated [30,31].

In a recent publication by the TCGA consortium, seven genomic molecular classes based on mutational profiles were shown, encompassing 74% of all primary tumors investigated: *TMPRSS2* fusions (with *ERG*, *ETV1*, *ETV4*, or *FLI1*; 59% of cases); or *SPOP* or *FOXA1* or *IDH1* mutations (11%, 3%, and 1%, respectively) [15]. For the remaining 26% of tumors, additional nonrecurrent mutations were identified, and in some cases, no major genomic lesions were found. Furthermore, a study examining 200 patients via whole-genome sequencing found gene methylation status to exhibit more association with PCa recurrence than genomic aberrations [19]. Therefore, while these genomic studies are extensive and insightful, aberrations in the epigenome are also recognized as a major driver of PCa development and progression, contributing to the complex molecular etiology [32].

EPIGENOMIC ALTERATIONS IN PCa

Epigenetic instability is a major hallmark of all stages of PCa [32], ranging from altered DNA methylation to histone modifications and noncoding RNA.

DNA METHYLATION

DNA methylation is the most extensively studied epigenetic modification in PCa [33]. Advances in genome-wide methylation profiling technologies, such as the Infinium 450K array and next-generation sequencing strategies, have enabled comprehensive examination of tissue methylation in all stages of PCa [34]. Similar to other cancers, PCa is marked by region-specific (i.e., promoter) hypermethylation and global hypomethylation in regions such as long interspersed nuclear elements (LINE) and ALU repetitive elements, especially in advanced PCa [35,36]. Hypermethylation of the promoter region of *GSTP1* has been reported in approximately 90% of PCa tumors and is considered to be an early event in carcinogenesis, correlated with reduced expression [37–39]. Furthermore, promoter CpG island hypermethylation of several other genes, including *APC*, *RASSF1a*, *AOX1* and *HOXD3*, has been shown in PCa [40–43]. Aberrant expression of DNA methyltransferases (DNMT1, DNMT3a, DNMT3b), with higher expression in aggressive PCa, has also been reported [44]. Interestingly, beyond gene silencing, promoter methylation of certain genes could impact expression of specific transcriptional variants, such as favoring of the short isoform of *RASSF1* [45].

Several of these hypermethylation events have been explored as biomarkers, especially given their stability and detection in various biofluids, including urine and serum [46–48]. DNA methylation may be a more reliable marker in tissue than RNA expression markers [49]. Numerous studies have assessed the potential of methylation biomarkers [40,43,50–54]. Despite their promising clinical utility, independent validation studies that will address intra- and interindividual heterogeneity to develop robust and reliable markers for diagnosis and risk stratification are needed [55,56]. For instance, hypermethylation of promoter regions near specific genes such as *HOXD3*, *AOX1*, *HAPLN3*, and *PITX2* has been shown to be associated with biochemical recurrence [50,57,58]. In addition, promoter hypermethylation of *APC*, *HOXD3*, *TGFβ2*, and *RASSF1A* has been shown to be associated with PCa progression [54]. In advanced PCa, distinctive DNA methylation patterns were observed between adenocarcinoma-CRPC and a more aggressive subtype, neuroendocrine-CRPC [31].

DNA HYDROXYMETHYLATION

5-Hydroxymethylcytosine (5hmC), previously considered to be a transient intermediate in DNA demethylation, has recently been highlighted as a stable epigenetic mark showing global loss in solid tumors and hematological malignancies [59,60]. On a global scale, 5hmC loss may promote uncontrolled proliferation of tumor cells due to its involvement with cellular differentiation/development, and is associated with lower survival [15,59,61–64]. Furthermore, locus-specific distribution of 5hmC shows distinct effects on gene expression in PCa cells [65]. While exonic and CpG island regions show loss of 5hmC and decreased gene expression in PCa, intergenic 5hmC gain has been linked to decreased gene expression among anticancer genes and pathways [65].

Ten-eleven translocase (TET) enzymes, which oxidize 5mC to produce 5hmC, show loss of expression in PCa tissues, with *TET2* in particular being targeted for repression by the androgen-regulated microRNAs miR-29a and miR-29b [66,67]. Loss of *TET2*, in turn, is linked to PCa progression through dysregulation of key signaling pathways, such as AR and mTOR [67]. Similarly, mutations of *TET1* are present in approximately 14.9% of high-risk prostate cancer cases, with reductions in *TET1* mRNA levels significantly linked to lowered metastasis-free survival in patients [66]. Regulation of AR activity was found to be impaired in tumor tissues with *TET1* expression loss.

Intriguingly, intergenic gain of 5hmC in the PCa cell line 22Rv1 as compared to normal prostate RWPE-1 cells was also linked to upregulation of AR, establishing an important link between TET-mediated changes in 5hmC and prostate carcinogenesis [65,66].

HISTONE MODIFICATIONS

Altered histone modifications are also a major driver of transcriptional deregulation in PCa [68]. However, unlike DNA methylation studies, tissue-based site-specific analyses of histone modifications (i.e., via ChIP-seq) are often hindered by low input amounts, typically limiting analysis to PCa cell lines [32]. Alternatively, global histone measurements in PCa tissue have revealed dysregulation in several markers, such as histone acetylation (H3K9, H3K18, and H4K12) and methylation (H3K3me2, H3K4me2) [69]. Global loss of H3 and H4 acetylation levels have been reported [70], likely due to altered histone acetyltransferase (HAT) and histone deacetylase (HDAC) activity [71]. Interestingly, not all acetylation marks are reduced, as increased levels of H3K18Ac have been observed in PCa tissue [72]. The role of bromodomain and extraterminal (BET) proteins (i.e., BRD4), which bind to acetylated marks, and the efficacy of BET inhibition in PCa are currently being explored and reviewed elsewhere [68,73].

In PCa cell lines, loss of H3K27me polycomb repression complex (PRC) marks and switching to H3K9me has been observed [74]. Global loss of H3K27me was also shown in primary tumors, with further reductions in advanced PCa [75,76]. Despite loss of PRC marks, the histone methyltransferase EZH2 is often overexpressed in PCa [77]. In addition to its normal function, which acts to repress transcription [78], EZH2 could have potential PRC-independent functions in PCa and may act as a co-activator for AR-mediated transcriptional activity [76,79]. Finally, dysregulation of other histone methylation marks, such as increased H3K4me2 [72] and histone demethylases (HDMs) [68], including LSD1, has also been reported.

MICRORNA AND LONG NONCODING RNA

Dysregulation of noncoding RNA constitutes a key mechanism in PCa development, with several miRNA (miR) and lncRNA expression profiles linked to PCa specifically [80,81]. For example, downregulation of miR-34a results in increased migration, invasion, and metastasis in PCa cells due to upregulation of CD44 [82]. In contrast, upregulation of miR-375, miR-141, and/or miR-200b is correlated with tumor stage, Gleason score, and metastasis [83], demonstrating a link between miRNA dysregulation and PCa progression.

ERG rearrangement status is associated with specific miRNA expression signatures, highlighting a potential role for differential miR status in the molecular classification of PCa. For example, miR-221 is downregulated in *ERG* fusion-positive tumors compared to *ERG* fusion-negative tumors [84]. Similarly, miR-338-3p loss is correlated with disease progression and biochemical recurrence in *ERG* fusion-positive tumors specifically through dysregulation of the CXCR4 chemokine axis [85]. In one study, miR-26a exhibited interplay with both the methylome and the transcriptome of PCa, as miR-26a was not only downregulated by hypermethylation, but was found to correlate inversely with *EZH2* expression [86]. However, this link was unique to *ERG* fusion-negative tumors, implying that miR-26a downregulation is an alternative pathway to *ERG* fusion in causing *EZH2* upregulation in cancer.

In terms of metastatic PCa, reduction of miR-15 and miR-16 expressions, coupled with miR-21 upregulation, contributes to activation of TGF- β and Hedgehog signaling, enabling PCa cells to survive in bone marrow via RUNX2 and RANKL [87]. Loss of miR-466 is also associated with increased *RUNX2* expressions, favoring increased migration, invasion, and proliferation of PCa cells [88]. miR-21, which is regulated by AR activity [89], has also been connected to increased matrix metalloproteinase expression and migration in fibroblasts, enhancing the tumor microenvironment [90].

Long noncoding RNAs (lncRNAs) have also shown utility as biomarkers of PCa. Perhaps the most recognized of these, *PCA3*, shows overexpression in PCa and acts as a diagnostic biomarker in urine [91,92]. Mechanistically, *PCA3* forms a complex with the tumor suppressor *PRUNE2*, decreasing its expression through ADAR-based RNA editing [91]. Another well-known lncRNA, *MALAT1*, is highly expressed in PCa, promoting an invasive phenotype in cancer cells [92]. In contrast, lncRNA produced from within the *GAS5* gene, associated with apoptosis, shows lowered expression in PCa due to inhibition by mTOR signaling [93].

Similar to miRs, lncRNA dysregulation often acts in combination with other epigenetic mechanisms to exert oncogenic effects. For example, increased expression of *ANRIL* inhibit tumor suppressor *INK4b* through reduction of H3K27 methylation and direct binding to *INK4b* transcripts, affecting apoptosis and the DNA damage response [92]. *SChLAP1* expression antagonizes the SWI/SNF complex and is correlated with increasing Gleason score and metastasis [94]. lncRNAs and miRs can also co-ordinate their functions, as shown by a tumor-suppressive axis between the lncRNA *H19* and miR-675 that inhibit cell migration, with both components showing downregulation in metastatic PCa [95]. Finally, at a genomic level, SNPs linked to PCa risk are enriched in lncRNA regions of the genome [96], while the specific risk-related SNP rs7463708 interacts with the lncRNA *PCAT1* promoter to enhance *ONECUT2* promoter occupancy and AR recruitment [92].

RATIONALE FOR INTEGRATIVE ANALYSIS

Overall, it is well established that both genetic and epigenetic mechanisms are at play in the development and progression of PCa. Given the broad range of molecular events that have been documented, there are several challenges when trying to develop optimal biomarker panels and generating novel therapeutics. Going forward, an important question arises—how do these complex molecular events work together to create the diverse clinical phenotypes we observe in PCa? For instance, loss of *GSTP1* expression, as observed in the majority of PCa patients, may lead to increased susceptibility to DNA damage by oxidative stress [97]. *ETS* fusion status is associated with distinct methylation patterns, such as reduced LINE-1 repeat methylation in fusion-negative tumors compared to fusion-positive [45] and differential methylation of specific genes (i.e., *HOXD3* and *TBX15*) [53]. Hypermethylated regions in PCa are also enriched in *EZH2* [98] and the PRC marker H3K27me3 [99].

Recent studies have generated several matched genomic and epigenomic data sets [15,19,31] (for a list of currently available (epi)genome-wide data sets, see Table 16.1). For example, integrative analysis of the TCGA cohort provided key insights into epigenetic and genetic interactions. First, distinct methylation patterns were observed between genetic driver subtypes, such as ETS fusion-positive tumors [15]. The few tumors that harbored *IDH1* mutations also exhibited extreme,

Table 16.1 Summary of PCa Tumor Studies with Matched Genomic and Epigenomic Data Sets

Study	Primary/Metastatic	No. of Patients	Data Sets Available
TCGA [15]	Primary	498	Whole exome sequencing (WES) Whole genome sequencing (WGS) Infinium HumanMethylation 450K array RNA-seq (mRNA and miRNA) Reverse Phase Protein Array
Fraser et al. [19]	Primary	200	WGS/SNP microarray Infinium HumanMethylation 450K array mRNA microarray
Beltran et al. [31]	Metastatic	81	WES RNA-seq Enhanced reduced representation bisulfite sequencing (eRRBS)

genome-wide methylation changes. In addition, direct integration of methylation and expression data sets identified several epigenetically silenced genes. Fig. 16.1 summarizes known and potential genomic and epigenomic interactions in PCa.

Even in nongenome-wide studies, multiple interactions showing interplay between various types of epigenetic and genetic alterations have been described in PCa. The regulation of histone methyltransferase *EZH2* expression by either *ERG* fusion status or miR-26 [86,95], AR-signaling induced repression of 5hmC through miR-29a and miR-29b [66,67] and SNP-induced changes in binding at the *PCAT1* promoter [92] are all examples of how integrative analysis can provide new insights into PCa mechanisms.

EMERGING INTEGRATIVE ANALYSIS TOOLS UTILIZED IN PCA

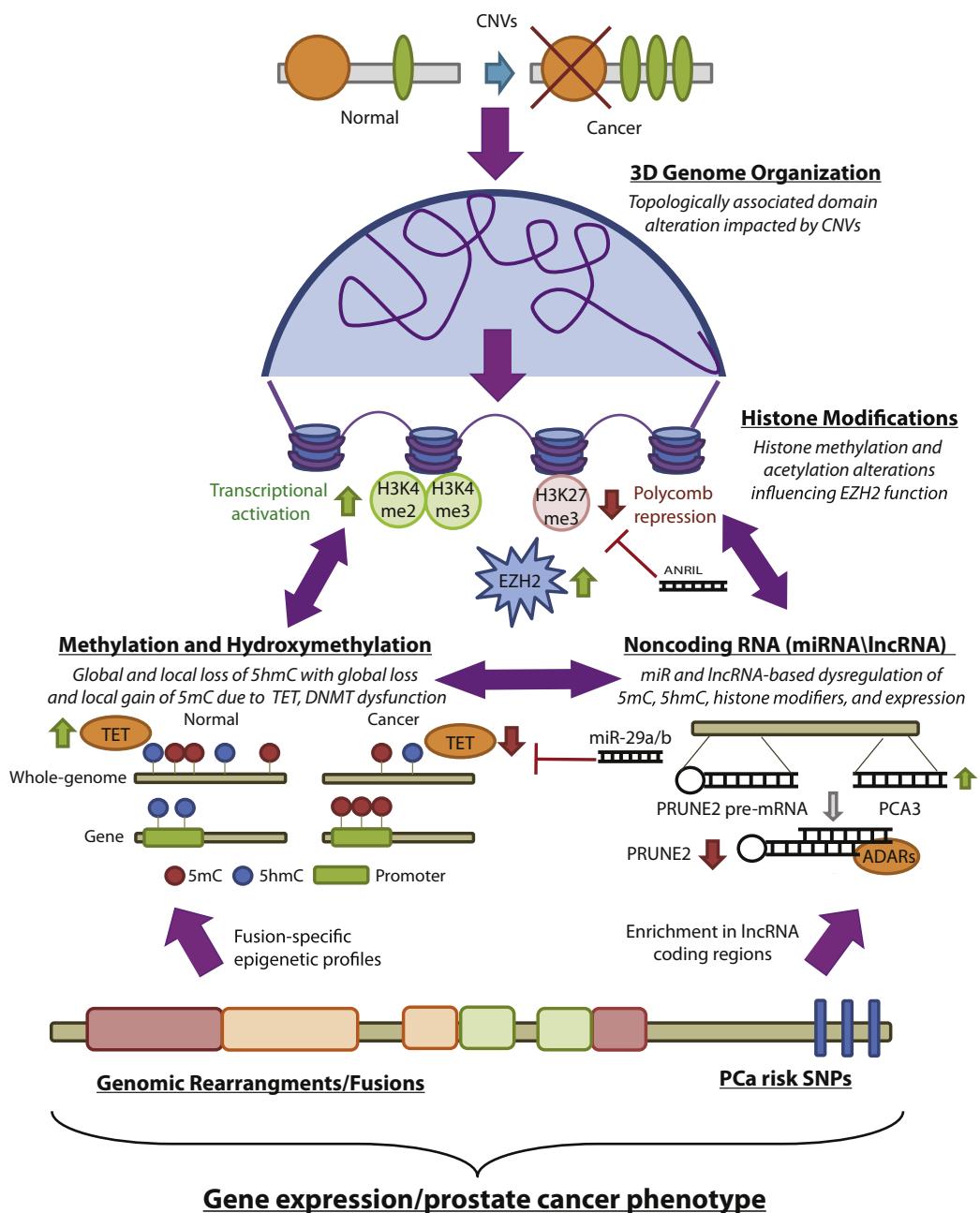
There is growing interest in the development of bioinformatic strategies that enable integrative analysis. Multiple studies have examined epigenetic interplay through specific targeting of a given epigenetic mark of interest. Emerging techniques from these studies have enabled these forms of integrative analysis on a genome-wide scale. Although many such pipelines are typically developed in-house through combining various computational methods, multiple published tool sets may be used to facilitate the discovery of epigenetic interactions. To date, most tools focus on chromatin modifications and DNA methylation; however, strategies to incorporate noncoding RNA are emerging. Below, we summarize recently developed tools combining genomic and epigenomic data sets.

Epidaurus [100] involves a two-step process that first performs aggregation analysis on individual genome-wide data sets (i.e., RNA-seq, MeDIP-seq, ChIP-seq, MNase-seq, and DNase-seq) to ascertain key genomic motifs, such as transcription factor (TF) binding sites. Subsequently, integration of aggregated data determines the association between different epigenetic modifications. For example, EZH2 has both transcriptional repressive and activating functions in PCa, exemplified by distinctive ChIP-seq profiles: ensemble peaks (enrichment of both EZH2 and H3K27me3) and solo peaks (EZH2 peaks that lack H3K27me3) [79]. Epidaurus was able to independently confirm the presence of these distinctive peaks, and by incorporating other published ChIP-seq data sets, demonstrated AR and RNA polymerase II co-localization in solo peaks as well as promoter (H3K4me2 and H3K4me3) and enhancer (H3K4me1) marks, indicating transcriptional activation. In another data set examining the effect of *FOXA1* knockdown on AR binding sites [101], Epidaurus implicated additional histone modifications and key genes known to be dysregulated in PCa, such as *KLK3* (PSA) and *EGFR*.

Gene signature association analysis (GS2A) [102] is a computational method that seeks novel TFs co-localizing within EZH2 solo peaks [79]. GS2A first identified 56 candidate genes that are regulated by EZH2 and bound to solo peak locations within CRPC cell lines (e.g., *FOXM1* and *CDK1*). Subsequently, GS2A screened for known transcriptional regulators in PCa, such as TFs, cofactors, and histone modifiers [103,104], as well as publicly available expression data sets [15,16], to find and rank association with candidate EZH2 collaborators. This refined potential transcriptional regulators in solo peaks to 35 candidate factors, with *E2F1* being the top-ranked TF. Experimental validation demonstrated co-localization of EZH2 and E2F1 in solo peaks. In addition, overexpression or knockdown of *E2F1* was shown to have reciprocal effects on EZH2 activated genes in PCa.

Model-based analysis of regulation of gene expression (MARGE) [105] uses H3K27ac ChIP-seq data, associated with active transcription, in order to predict changes in gene expression and TF binding. There are three main functions in the MARGE platform: MARGE-potential, MARGE-express, and MARGE-cistrome. MARGE-potential calculates and ranks the regulatory potential (RP) of H3K27ac on expression of specific genes by taking into account distance of this modification to transcriptional start sites (TSS), and has been shown to accurately predict changes in gene expression. Next, using logistic regression, MARGE-express is able to take multiple H3K27ac ChIP-seq data sets (i.e., 365 from human) from different cell types to identify cell-specific changes in expression, including prostate. MARGE-cistrome combines RPs identified from MARGE-express and 458 human DNase-seq data sets to predict other regulatory elements, such as TF binding sites. In order to validate whether MARGE can accurately predict expression perturbations from these publicly available data sets, knockdown of various TFs (i.e., *AR*, *E2F1*, *FOXA1*) was performed in PCa cell lines followed by H3K27ac ChIP-seq and RNA-seq. Overall, MARGE-express and MARGE-cistrome were shown to predict experimentally derived expression changes.

RegNetDriver [106] enables integrative analysis of genetic and epigenetic data sets. First, a prostate tissue-specific regulatory network was generated using DNase I hypersensitive site data sets from prostate epithelial cells and integrated with regulatory elements (i.e., TF binding sites, histone markers, promoters, and enhancers) from Encyclopedia of DNA Elements (ENCODE) [107] and the Roadmap Epigenomics Mapping Consortium (REMC) [108]. 612 TF binding motifs were implicated using the protein interaction quantification tool [109], which includes known PCa genes such as ETS family members *TP53*, and *MYC*. Following development of this regulatory network, the RegNetDriver platform next associates significantly altered SNVs and SVs as well as differentially methylated regions from genome-wide data sets [15,18,19,110]. Significant alterations are then combined with the

**FIGURE 16.1**

Summary of known and potential genomic and epigenomic interactions in PCa. Flowchart representing interactions between the genome and epigenome in prostate cancer, ranging from large-scale topological domain alterations to localized interactions between specific epigenetic and genomic alterations. *3D genome*

prostate regulatory network to highlight the combined impact of these mutations and epigenetic modifications on key PCa TF hubs. RegNetDriver revealed three cancer-related genes that are affected by both genetic and epigenetic aberrations (*FAS*, *FAM3B*, and *TNFSF13*). Several key TF hubs were distinguished by various genomic/epigenomic modifications, such as *ERG*, *TP53*, and *ERF* in SVs or *NR3C1* in differentially methylated regions.

3D genome analysis techniques: In addition to biochemical modifications of DNA and histones, spatial genomic organization also impacts gene expression. Through chromosome conformation capture-based sequencing technologies, such as Hi-C, the 3D structure of the genome can be divided into compartments known as topologically associated domains (TADs) [111,112]. These TADs are often large (hundreds of kilobases), include both genes and regulatory regions (i.e., enhancers), and are thought to be conserved across different cell lineages [113]. In a recent study, the organization of TADs was explored in PCa and normal prostate epithelial cell lines [114,115]. Generally, there were smaller but more TADs in PCa cell lines; however, 80% of TAD boundaries were maintained in both normal and cancer cells, with a subset of cancer-specific TAD boundaries. These cancer-specific TADs were associated with CNAs, suggesting interplay between genomic alterations and new TAD boundary formation. Differential TAD interactions between cancer and normal cell lines were further delineated using diffHiC [116] and chromHMM [117], identifying enrichment of several regulatory elements (i.e., enhancers, promoters, and CTCF binding sites) and association with known differentially expressed genes in PCa (TCGA RNA data [15]). To this end, a Hi-C data visualization program, Rondo, was created in order to integrate ChIP-seq and RNA-seq data. Alternatively, if Hi-C data are not available, chromatin compartments could potentially be inferred from methylation data sets and integrated with histone marker and gene expression data sets [118,119].

FUTURE DIRECTIONS AND POTENTIAL APPLICATIONS FOR PCA

Given the continually expanding genomic and epigenomic data sets in different stages of PCa (Table 16.1), we now have an opportunity to examine key relationships between different types of molecular drivers and to develop tools that facilitate integrative analysis. The ultimate goal is to have

organization: Copy number variations at boundaries between topologically associated domains (TADs) result in smaller-sized, but more numerous TADs in the cancer genome [114,115]. Histone modifications: Differential methylation and acetylation affect transcriptional activation or repression [68,76,78,79], while certain histone marks are suppressed by epigenetic mechanisms such as the lncRNA *ANRIL*, which represses H3K27me3 modification [92]. Methylation and hydroxymethylation: Whole-genome loss of 5mC and 5hmC, as well as local alterations in these marks at the promoters of tumor suppressors, is observed in cancer due to *DNMT* and TET enzyme dysregulation [35,36,44,59,60,66,67]. In turn, TET levels are negatively regulated by AR-induced miR-29a/b [67]. Noncoding RNA (miRNA/lncRNA): miRNA and lncRNAs interact heavily with histone modifications, 5mC/5hmC (as mentioned above), and gene expression. As an example, the repressive interaction of urinary biomarker lncRNA *PCA3* with *PRUNE2* through ADAR-based RNA editing is depicted [91]. Genomic rearrangements/fusions: Large-scale methylation and miRNA expression profile alterations have been linked to specific gene fusions (for example, loss of global 5hmC in *ERG* fusion-negative tumors specifically) [15]. PCa risk SNPs: Single-nucleotide polymorphisms can affect epigenetic modifications depending on their position, and also show differential distribution based on epigenetic status, with PCa risk SNPs showing enrichment in lncRNA coding regions [92,96].

the capacity to predict PCa phenotypes and clinical outcome. Currently available tools have shown tremendous promise, especially in deriving crucial transcriptional networks altered in PCa. By identifying key TFs, including androgen- and nonandrogen-regulated mechanisms, there is potential to discover novel therapeutic strategies and biomarker panels that were not apparent through assessment of genomic or epigenomic alterations alone. While promising, there are several limitations to currently available tools, as much focus has been placed on cell-line derived data sets and ChIP-based tissue analysis can be challenging.

In addition, we can draw from other tools that have been developed for other cancer types. *Epigenetic Module based on Differential Networks (EMDN)* [120] and *Functional Epigenetic Models (FEM)* [121], which integrate methylation array and expression data sets, can identify critical pathways exhibiting differential methylation in cancer. *Significance-based Modules Integrating the Transcriptome and Epigenome (SMITE)* [122] expands on the techniques used in FEM and EMDN by assigning gene scores to integrated genomic and epigenomic data without assumption of independence, and then performing module identification based on functional protein–protein interaction networks. *Methylation INTegration (Mint)* is a multifunctional tool enabling integration of DNA methylation and hydroxymethylation data to ascertain key patterns throughout the genome [123]. Alternatively, *Sigma²* is a visualization-based tool that allows for multidimensional integration of epigenetic and genetic data sets both visually and statistically, allowing (epi)genetic differences between two groups to be analyzed [124]. Similarly, specialized tools to model interactions between specific types of genetic and epigenetic alterations exist. For example, *MicroSNiPer* enables prediction of the effect a provided set of SNPs in the 3'-UTR of a gene will have on miRNA binding [125]. Combinations of these tools can be used to analyze data sets; however, future integration of their various functionalities would be instrumental in establishing a comprehensive picture of epigenetic and genetic interaction in PCa.

CONCLUDING REMARKS

Although epigenetic aberrations, including dysregulation of methylation, hydroxymethylation, miRNA/lncRNA, and histone remodeling, are hallmarks of prostate cancer, many of these changes act in concert with other epigenomic or genomic changes to exert oncogenic effects. The usage of integrative analysis pipelines such as Epidaurus, GS2A, MARGE, and RegNetDriver has enabled identification of these regulatory axes on a whole-genome level, furthering our understanding of PCa biology. Ultimately, further characterization of these epigenomic networks may possess clinical utility, identifying novel subtypes, diagnostic and prognostic biomarkers, and therapeutic targets.

ACKNOWLEDGMENTS

Dr. Bharati Bapat is funded by Prostate Cancer Canada (PCC) Movember Discovery Grant (No. D2014-10), Movember PCC TAG No. 2014-01 1417 and Astellas Prostate Cancer Innovation Fund (2017). Madonna Peter is supported by Canadian Institutes of Health Research Doctoral Research Award. Shivani Kamdar is supported by Ontario Student Opportunity Trust Fund Awards and Ontario Graduate Scholarship.

REFERENCES

- [1] Dy GW, Gore JL, Forouzanfar MH, Naghavi M, Fitzmaurice C. Global burden of urologic cancers, 1990–2013. *Eur Urol* 2017;71(3):437–46.
- [2] Global Burden of Disease Cancer C, Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the Global Burden of Disease Study. *J Am Med Assoc Oncol* 2017;3(4):524–48.
- [3] Hayes JH, Barry MJ. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. *J Am Med Assoc* 2014;311(11):1143–9.
- [4] Klotz L, Vesprini D, Sethukavalan P, Jethava V, Zhang L, Jain S, et al. Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. *J Clin Oncol* 2015;33(3):272–7.
- [5] Carroll PR, Parsons JK, Andriole G, Bahnson RR, Barocas DA, Castle EP, et al. NCCN clinical practice guidelines prostate cancer early detection, version 2.2015. *J Natl Compr Cancer Netw* 2015;13(12):1534–61.
- [6] Popiolek M, Rider JR, Andren O, Andersson SO, Holmberg L, Adami HO, et al. Natural history of early, localized prostate cancer: a final report from three decades of follow-up. *Eur Urol* 2013;63(3):428–35.
- [7] Penney KL, Stampfer MJ, Jahn JL, Sinnott JA, Flavin R, Rider JR, et al. Gleason grade progression is uncommon. *Cancer Res* 2013;73(16):5163–8.
- [8] Mohler JL, Armstrong AJ, Bahnson RR, D'Amico AV, Davis BJ, Eastham JA, et al. Prostate cancer, version 1.2016. *J Natl Compr Cancer Netw* 2016;14(1):19–30.
- [9] D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, Broderick GA, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *J Am Med Assoc* 1998;280(11):969–74.
- [10] Cooperberg MR, Broering JM, Carroll PR. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst* 2009;101(12):878–87.
- [11] Zhou P, Chen MH, McLeod D, Carroll PR, Moul JW, D'Amico AV. Predictors of prostate cancer-specific mortality after radical prostatectomy or radiation therapy. *J Clin Oncol* 2005;23(28):6992–8.
- [12] Freedland SJ, Humphreys EB, Mangold LA, Eisenberger M, Dorey FJ, Walsh PC, et al. Risk of prostate cancer-specific mortality following biochemical recurrence after radical prostatectomy. *J Am Med Assoc* 2005;294(4):433–9.
- [13] Attard G, Parker C, Eeles RA, Schroder F, Tomlins SA, Tannock I, et al. Prostate cancer. *Lancet* 2016;387(10013):70–82.
- [14] Khemlina G, Ikeda S, Kurzrock R. Molecular landscape of prostate cancer: implications for current clinical trials. *Cancer Treat Rev* 2015;41(9):761–6.
- [15] Cancer Genome Atlas Research N. The molecular taxonomy of primary prostate cancer. *Cell* 2015;163(4):1011–25.
- [16] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* 2010;18(1):11–22.
- [17] Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SP0P, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 2012;44(6):685–9.
- [18] Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. *Cell* 2013;153(3):666–77.
- [19] Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* 2017;541(7637):359–64.
- [20] Spratt DE, Zumsteg ZS, Feng FY, Tomlins SA. Translational and clinical implications of the genetic landscape of prostate cancer. *Nat Rev Clin Oncol* 2016;13(10):597–610.

- [21] Tomlins SA, Bjartell A, Chinnaian AM, Jenster G, Nam RK, Rubin MA, et al. ETS gene fusions in prostate cancer: from discovery to daily clinical practice. *Eur Urol* 2009;56(2):275–86.
- [22] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005;310(5748):644–8.
- [23] Mosquera JM, Mehra R, Regan MM, Perner S, Genega EM, Bueti G, et al. Prevalence of TMPRSS2-ERG fusion prostate cancer among men undergoing prostate biopsy in the United States. *Clin Cancer Res* 2009;15(14):4706–11.
- [24] Hoogland AM, Jenster G, van Weerden WM, Trapman J, van der Kwast T, Roobol MJ, et al. ERG immunohistochemistry is not predictive for PSA recurrence, local recurrence or overall survival after radical prostatectomy for prostate cancer. *Mod Pathol* 2012;25(3):471–9.
- [25] Xu B, Chevarie-Davis M, Chevalier S, Scarlata E, Zeizafoun N, Dragomir A, et al. The prognostic role of ERG immunopositivity in prostatic acinar adenocarcinoma: a study including 454 cases and review of the literature. *Hum Pathol* 2014;45(3):488–97.
- [26] Kloosterman WP, Koster J, Molenaar JJ. Prevalence and clinical implications of chromothripsis in cancer genomes. *Curr Opin Oncol* 2014;26(1):64–72.
- [27] Andreouli M, Cheng L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. *Hum Pathol* 2010;41(6):781–93.
- [28] Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet* 2015;47(4):367–72.
- [29] Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E, et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* 2015;47(7):736–45.
- [30] Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. *Cell* 2015;161(5):1215–28.
- [31] Beltran H, Prandi D, Mosquera JM, Benelli M, Puca L, Cyrta J, et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* 2016;22(3):298–305.
- [32] Yegnasubramanian S. Prostate cancer epigenetics and its clinical implications. *Asian J Androl* 2016;18(4):549–58.
- [33] Massie CE, Mills IG, Lynch AG. The importance of DNA methylation in prostate cancer development. *J Steroid Biochem Mol Biol* 2017;166:1–15.
- [34] Olkhov-Mitsel E, Bapat B. Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. *Cancer Med* 2012;1(2):237–60.
- [35] Yegnasubramanian S, Haffner MC, Zhang Y, Gurel B, Cornish TC, Wu Z, et al. DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer Res* 2008;68(21):8954–67.
- [36] Cho NY, Kim BH, Choi M, Yoo EJ, Moon KC, Cho YM, et al. Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. *J Pathol* 2007;211(3):269–77.
- [37] Lee WH, Morton RA, Epstein JI, Brooks JD, Campbell PA, Bova GS, et al. Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci U S A* 1994;91(24):11733–7.
- [38] Lee WH, Isaacs WB, Bova GS, Nelson WG. CG island methylation changes near the GSTP1 gene in prostatic carcinoma cells detected using the polymerase chain reaction: a new prostate cancer biomarker. *Cancer Epidemiol Biomarkers Prev* 1997;6(6):443–50.
- [39] Jimenez RE, Fischer AH, Petros JA, Amin MB. Glutathione S-transferase pi gene methylation: the search for a molecular marker of prostatic adenocarcinoma. *Adv Anat Pathol* 2000;7(6):382–9.

- [40] Kron K, Pethe V, Briollais L, Sadikovic B, Ozcelik H, Sunderji A, et al. Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays. *PLoS One* 2009;4(3):e4830.
- [41] Yegnasubramanian S, Kowalski J, Gonzalgo ML, Zahurak M, Piantadosi S, Walsh PC, et al. Hypermethylation of CpG islands in primary and metastatic human prostate cancer. *Cancer Res* 2004;64(6):1975–86.
- [42] Yegnasubramanian S, Wu Z, Haffner MC, Esopi D, Aryee MJ, Badrinath R, et al. Chromosome-wide mapping of DNA methylation patterns in normal and malignant prostate cells reveals pervasive methylation of gene-associated and conserved intergenic sequences. *BMC Genomics* 2011;12:313.
- [43] Geybels MS, Zhao S, Wong CJ, Bibikova M, Klotzle B, Wu M, et al. Epigenomic profiling of DNA methylation in paired prostate cancer versus adjacent benign tissue. *Prostate* 2015;75(16):1941–50.
- [44] Gravina GL, Ranieri G, Muži P, Marampon F, Mancini A, Di Pasquale B, et al. Increased levels of DNA methyltransferases are associated with the tumorigenic capacity of prostate cancer cells. *Oncol Rep* 2013;29(3):1189–95.
- [45] Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Res* 2011;21(7):1028–41.
- [46] Rogers CG, Gonzalgo ML, Yan G, Bastian PJ, Chan DY, Nelson WG, et al. High concordance of gene methylation in post-digital rectal examination and post-biopsy urine samples for prostate cancer detection. *J Urol* 2006;176(5):2280–4.
- [47] Zhao F, Olkhov-Mitsel E, van der Kwast T, Sykes J, Zdravic D, Venkateswaran V, et al. Urinary DNA methylation biomarkers for noninvasive prediction of aggressive disease in patients with prostate cancer on active surveillance. *J Urol* 2017;197(2):335–41.
- [48] Bastian PJ, Palapattu GS, Lin X, Yegnasubramanian S, Mangold LA, Trock B, et al. Preoperative serum DNA GSTP1 CpG island hypermethylation and the risk of early prostate-specific antigen recurrence following radical prostatectomy. *Clin Cancer Res* 2005;11(11):4037–43.
- [49] Paziewska A, Dabrowska M, Goryca K, Antoniewicz A, Dobruch J, Mikula M, et al. DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. *Br J Cancer* 2014;111(4):781–9.
- [50] Stott-Miller M, Zhao S, Wright JL, Kolb S, Bibikova M, Klotzle B, et al. Validation study of genes with hypermethylated promoter regions associated with prostate cancer recurrence. *Cancer Epidemiol Biomarkers Prev* 2014;23(7):1331–9.
- [51] Trock BJ, Brotzman MJ, Mangold LA, Bigley JW, Epstein JI, McLeod D, et al. Evaluation of GSTP1 and APC methylation as indicators for repeat biopsy in a high-risk cohort of men with negative initial prostate biopsies. *BJU Int* 2012;110(1):56–62.
- [52] Olkhov-Mitsel E, Siadat F, Kron K, Liu L, Savio AJ, Trachtenberg J, et al. Distinct DNA methylation alterations are associated with cribriform architecture and intraductal carcinoma in Gleason pattern 4 prostate tumors. *Oncol Lett* 2017;14(1):390–6.
- [53] Kron K, Trudel D, Pethe V, Briollais L, Fleshner N, van der Kwast T, et al. Altered DNA methylation landscapes of polycomb-repressed loci are associated with prostate cancer progression and ERG oncogene expression in prostate cancer. *Clin Cancer Res* 2013;19(13):3450–61.
- [54] Liu L, Kron KJ, Pethe VV, Demetashvili N, Nesbitt ME, Trachtenberg J, et al. Association of tissue promoter methylation levels of APC, TGFbeta2, HOXD3 and RASSF1A with prostate cancer progression. *Int J Cancer* 2011;129(10):2454–62.
- [55] Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep* 2014;8(3):798–806.
- [56] Aryee MJ, Liu W, Engelmann JC, Nuhn P, Gurel M, Haffner MC, et al. DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases. *Sci Transl Med* 2013;5(169):169ra10.

- [57] Haldrup C, Mundbjerg K, Vestergaard EM, Lamy P, Wild P, Schulz WA, et al. DNA methylation signatures for prediction of biochemical recurrence after radical prostatectomy of clinically localized prostate cancer. *J Clin Oncol* 2013;31(26):3250–8.
- [58] Uhl B, Gevensleben H, Tolkach Y, Sailer V, Majores M, Jung M, et al. PITX2 DNA methylation as biomarker for individualized risk assessment of prostate cancer in core biopsies. *J Mol Diagn* 2017;19(1):107–14.
- [59] Haffner MC, Chaux A, Meeker AK, Esopi DM, Gerber J, Pellakuru LG, et al. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* 2011;2(8):627–37.
- [60] Shukla A, Sehgal M, Singh TR. Hydroxymethylation and its potential implication in DNA repair system: a review and future perspectives. *Gene* 2015;564(2):109–18.
- [61] Yang H, Liu Y, Bai F, Zhang JY, Ma SH, Liu J, et al. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* 2013;32(5):663–9.
- [62] Chen Z, Shi X, Guo L, Li Y, Luo M, He J. Decreased 5-hydroxymethylcytosine levels correlate with cancer progression and poor survival: a systematic review and meta-analysis. *Oncotarget* 2017;8(1):1944–52.
- [63] Ficz G, Gribsen JG. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? *Genomics* 2014;104(5):352–7.
- [64] Mariani CJ, Madzo J, Moen EL, Yesilkanal A, Godley LA. Alterations of 5-hydroxymethylcytosine in human cancers. *Cancer* 2013;5(3):786–814.
- [65] Kamdar SN, Ho LT, Kron KJ, Isserlin R, van der Kwast T, Zlotta AR, et al. Dynamic interplay between locus-specific DNA methylation and hydroxymethylation regulates distinct biological pathways in prostate carcinogenesis. *Clin Epigenet* 2016;8:32.
- [66] Spans L, Van den Broeck T, Smeets E, Prekovic S, Thienpont B, Lambrechts D, et al. Genomic and epigenomic analysis of high-risk prostate cancer reveals changes in hydroxymethylation and TET1. *Oncotarget* 2016;7(17):24326–38.
- [67] Takayama K, Misawa A, Suzuki T, Takagi K, Hayashizaki Y, Fujimura T, et al. TET2 repression by androgen hormone regulates global hydroxymethylation status and prostate cancer progression. *Nat Commun* 2015;6:8219.
- [68] Baumgart SJ, Haendler B. Exploiting epigenetic alterations in prostate cancer. *Int J Mol Sci* 2017;18(5).
- [69] Seligson DB, Horvath S, Shi T, Yu H, Tze S, Grunstein M, et al. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature* 2005;435(7046):1262–6.
- [70] Ellinger J, Kahl P, von der Gathen J, Rogenhofer S, Heukamp LC, Gutgemann I, et al. Global levels of histone modifications predict prostate cancer recurrence. *Prostate* 2010;70(1):61–9.
- [71] Cang S, Feng J, Konno S, Han L, Liu K, Sharma SC, et al. Deficient histone acetylation and excessive deacetylase activity as epigenomic marks of prostate cancer cells. *Int J Oncol* 2009;35(6):1417–22.
- [72] Bianco-Miotto T, Chiam K, Buchanan G, Jindal S, Day TK, Thomas M, et al. Global levels of specific histone modifications and an epigenetic gene signature predict prostate cancer progression and development. *Cancer Epidemiol Biomarkers Prev* 2010;19(10):2611–22.
- [73] Lochrin SE, Price DK, Figg WD. BET bromodomain inhibitors—a novel epigenetic approach in castration-resistant prostate cancer. *Cancer Biol Ther* 2014;15(12):1583–5.
- [74] Gal-Yam EN, Egger G, Iniguez L, Holster H, Einarsson S, Zhang X, et al. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci U S A* 2008;105(35):12979–84.
- [75] Pellakuru LG, Iwata T, Gurel B, Schultz D, Hicks J, Bethel C, et al. Global levels of H3K27me3 track with differentiation in vivo and are deregulated by MYC in prostate cancer. *Am J Pathol* 2012;181(2):560–9.
- [76] Ngollo M, Lebert A, Daures M, Judes G, Rifai K, Dubois L, et al. Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression. *BMC Cancer* 2017;17(1):261.

- [77] Melling N, Thomsen E, Tsourlakis MC, Kluth M, Hube-Magg C, Minner S, et al. Overexpression of enhancer of zeste homolog 2 (EZH2) characterizes an aggressive subset of prostate cancers and predicts patient prognosis independently from pre- and postoperatively assessed clinicopathological parameters. *Carcinogenesis* 2015;36(11):1333–40.
- [78] Deb G, Thakur VS, Gupta S. Multifaceted role of EZH2 in breast and prostate tumorigenesis: epigenetics and beyond. *Epigenetics* 2013;8(5):464–76.
- [79] Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, et al. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* 2012;338(6113):1465–9.
- [80] Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. *Cancer Discov* 2011;1(5):391–407.
- [81] Bertoli G, Cava C, Castiglioni I. MicroRNAs as biomarkers for diagnosis, prognosis and theranostics in prostate cancer. *Int J Mol Sci* 2016;17(3):421.
- [82] Liu C, Kelnar K, Liu B, Chen X, Calhoun-Davis T, Li H, et al. The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Nat Med* 2011;17(2):211–5.
- [83] Bräse JC, Johannes M, Schlomm T, Falth M, Haese A, Steuber T, et al. Circulating miRNAs are correlated with tumor progression in prostate cancer. *Int J Cancer* 2011;128(3):608–16.
- [84] Gordianpour A, Stanimirovic A, Nam RK, Moreno CS, Sherman C, Sugar L, et al. miR-221 Is down-regulated in TMPRSS2:ERG fusion-positive prostate cancer. *Anticancer Res* 2011;31(2):403–10.
- [85] Bakkar A, Alshalalfa M, Petersen LF, Abou-Ouf H, Al-Mami A, Hegazy SA, et al. microRNA 338-3p exhibits tumor suppressor role and its down-regulation is associated with adverse clinical outcome in prostate cancer patients. *Mol Biol Rep* 2016;43(4):229–40.
- [86] Bornstein ST, Fischer A, Kerick M, Falth M, Laible M, Bräse JC, et al. Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov* 2012;2(11):1024–35.
- [87] Bonci D, Coppola V, Patrizii M, Addario A, Cannistraci A, Francescangeli F, et al. A microRNA code for prostate cancer metastasis. *Oncogene* 2016;35(9):1180–92.
- [88] Colden M, Dar AA, Saini S, Dahiya PV, Shahryari V, Yamamura S, et al. MicroRNA-466 inhibits tumor growth and bone metastasis in prostate cancer by direct regulation of osteogenic transcription factor RUNX2. *Cell Death Dis* 2017;8(1):e2572.
- [89] Ribas J, Ni X, Haffner M, Wentzel EA, Salmasi AH, Chowdhury WH, et al. miR-21: an androgen receptor-regulated microRNA that promotes hormone-dependent and hormone-independent prostate cancer growth. *Cancer Res* 2009;69(18):7165–9.
- [90] Sanchez CA, Andahur EI, Valenzuela R, Castellon EA, Fulla JA, Ramos CG, et al. Exosomes from bulk and stem cells from human prostate cancer have a differential microRNA content that contributes cooperatively over local and pre-metastatic niche. *Oncotarget* 2016;7(4):3993–4008.
- [91] Salameh A, Lee AK, Cardo-Vila M, Nunes DN, Efstatithiou E, Staquicini FI, et al. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proc Natl Acad Sci U S A* 2015;112(27):8403–8.
- [92] Malik B, Feng FY. Long noncoding RNAs in prostate cancer: overview and clinical implications. *Asian J Androl* 2016;18(4):568–74.
- [93] Pickard MR, Mourtada-Maarabouni M, Williams GT. Long non-coding RNA GAS5 regulates apoptosis in prostate cancer cell lines. *Biochim Biophys Acta* 2013;1832(10):1613–23.
- [94] Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* 2013;45(11):1392–8.
- [95] Zhu M, Chen Q, Liu X, Sun Q, Zhao X, Deng R, et al. lncRNA H19/miR-675 axis represses prostate cancer metastasis by targeting TGFB1. *FEBS J* 2014;281(16):3766–75.

- [96] Jin G, Sun J, Isaacs SD, Wiley KE, Kim ST, Chu LW, et al. Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis* 2011;32(11):1655–9.
- [97] Mian OY, Khattab MH, Hedayati M, Coulter J, Abubaker-Sharif B, Schwaninger JM, et al. GSTP1 Loss results in accumulation of oxidative DNA base damage and promotes prostate cancer cell survival following exposure to protracted oxidative stress. *Prostate* 2016;76(2):199–206.
- [98] Kirby MK, Ramaker RC, Roberts BS, Lasseigne BN, Gunther DS, Burwell TC, et al. Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns. *BMC Cancer* 2017;17(1):273.
- [99] Lin PC, Giannopoulou EG, Park K, Mosquera JM, Sboner A, Tewari AK, et al. Epigenomic alterations in localized and advanced prostate cancer. *Neoplasia* 2013;15(4):373–83.
- [100] Wang L, Huang H, Dougherty G, Zhao Y, Hossain A, Kocher JP. Epidaurus: aggregation and integration analysis of prostate cancer epigenome. *Nucleic Acids Res* 2015;43(2):e7.
- [101] Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 2011;474(7351):390–4.
- [102] Xu H, Xu K, He HH, Zang C, Chen CH, Chen Y, et al. Integrative analysis reveals the transcriptional collaboration between EZH2 and E2F1 in the regulation of cancer-related gene expression. *Mol Cancer Res* 2016;14(2):163–72.
- [103] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;10(4):252–63.
- [104] Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, et al. CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. *Bioinformatics* 2012;28(10):1411–2.
- [105] Wang S, Zang C, Xiao T, Fan J, Mei S, Qin Q, et al. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Res* 2016;26(10):1417–29.
- [106] Dhingra P, Martinez-Fundichely A, Berger A, Huang FW, Forbes AN, Liu EM, et al. Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. *Genome Biol* 2017;18(1):141.
- [107] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57–74.
- [108] Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518(7539):317–30.
- [109] Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014;32(2):171–8.
- [110] Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. *Nature* 2011;470(7333):214–20.
- [111] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485(7398):376–80.
- [112] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012;485(7398):381–5.
- [113] Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. *Mol Cell* 2016;62(5):668–80.
- [114] Achinger-Kawecka J, Taberlay PC, Clark SJ. Alterations in three-dimensional organization of the cancer genome and epigenome. *Cold Spring Harb Symp Quant Biol* 2016;81:41–51.
- [115] Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* 2016;26(6):719–31.

- [116] Lun AT, Smyth GK. diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* 2015;16:258.
- [117] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9(3):215–6.
- [118] Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 2015;16:180.
- [119] Simmonds P, Loomis E, Curry E. DNA methylation-based chromatin compartments and ChIP-seq profiles reveal transcriptional drivers of prostate carcinogenesis. *Genome Med* 2017;9(1):54.
- [120] Ma X, Liu Z, Zhang Z, Huang X, Tang W. Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics* 2017;18(1):72.
- [121] Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 2014;30(16):2360–6.
- [122] Wijetunga NA, Johnston AD, Maekawa R, Delahaye F, Ulahannan N, Kim K, et al. SMITE: an R/Bioconductor package that identifies network modules by integrating genomic and epigenomic information. *BMC Bioinformatics* 2017;18(1):41.
- [123] Cavalcante RG, Patil S, Park Y, Rozek LS, Sartor MA. Integrating DNA methylation and hydroxymethylation data with the mint pipeline. *Cancer Res* 2017;77(21):e27–30.
- [124] Chari R, Coe BP, Wedseltoft C, Benetti M, Wilson IM, Vucic EA, et al. SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes. *BMC Bioinformatics* 2008;9:422.
- [125] Barenboim M, Zoltick BJ, Guo Y, Weinberger DR. MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum Mutat* 2010;31(11):1223–32.

This page intentionally left blank

NETWORK ANALYSIS OF EPIGENETIC DATA FOR BLADDER CANCER

17

Bor-Sen Chen

Lab of Control and Systems Biology, National Tsing Hua University, Hsinchu, Taiwan

INTRODUCTION

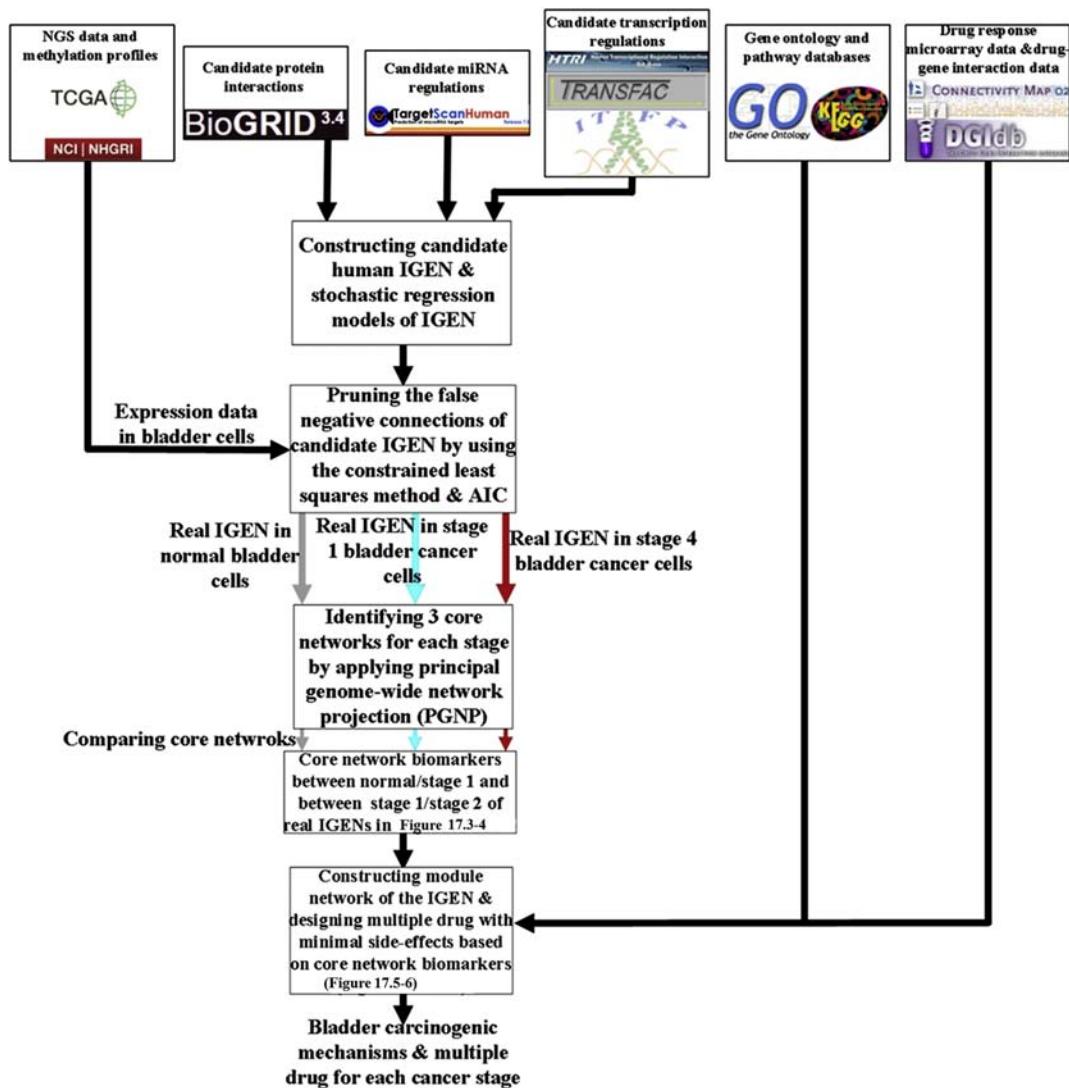
Bladder cancer is still one of the most common cancers worldwide. Single-gene markers have been proposed for improving cancer treatment [1]. However, single-gene markers cannot overcome treatment side effects because the markers are not implicated in genome-wide networks, and the analysis of a genome-wide network is a complicated issue from a systems biology perspective. The rapid development of molecular biology techniques has produced a great deal of high-throughput experimental data, including genome-wide microarray data, genome-wide methylation profiles, next-generation sequencing (NGS) data, microRNA (miRNA) profiles, genetic sequences, protein abundance data, and drug response genome-wide microarray data. These kinds of omics data provide an opportunity to design multiple drug combinations for the treatment of bladder cancer by applying the network biomarkers identified by systems biology method.

To date, genetic regulation systems, including protein–protein interaction networks (PPINs) and gene regulatory networks (GRNs), have been applied to analyze the functional mechanisms behind human aging and cancer [2,3]. We now know that epigenetic alterations are much more rapid and adaptive with regard to influencing genome-wide gene expression than genetic changes [4]. Rapid and slow response mechanisms, that is, epigenetic alterations and genetic changes, respectively, coordinate an efficient and robust system. Epigenetic regulation, including DNA methylation and histone modification, results in potentially reversible alterations in gene expression that do not involve permanent changes to the DNA sequence. miRNAs that are influenced by aberrant epigenetic regulation also mediate the regulation of gene expression [5]. It has been found that DNA methylation directly affects the binding affinities of miRNAs, RNA polymerase, and transcription factors (TFs) [6] and indirectly influences protein–protein interactions (PPIs) [7]. Methylation analysis of human genomic DNA in 12 tissues revealed that DNA methylation profiles are tissue specific [8]. Therefore, omics data and systems biology methods [9–11] are required to unravel the mechanisms underlying carcinogenesis from the complex molecular biology and design anticancer drugs for the treatment of bladder cancer.

The Human Genome Project (HGP) has identified 30,000–40,000 genes in human DNA, including miRNAs. The genes, proteins, and their associations, miRNA regulation, and DNA methylation constitute the integrated genetic and epigenetic genome-wide network (IGEN), which coordinates cellular responses. PPIN in human lung cancer [12] and GRN in human aging [13] of the genes with significant expression differences between cancer cells (or aged people) and normal cells (or young people) have been identified for the extraction of the core network biomarkers according to the estimated association abilities between TFs (or upstream proteins) and target genes (or target proteins). Aging is associated with cancer [14]. The association abilities estimated by the network models assume that the binding affinities of TFs (or upstream proteins) to target genes (or proteins) are the same. According to a recent study in primary human somatic and germline cells [6], the impact of the binding affinities of miRNAs, RNA polymerase, and TFs on gene expression is mediated by DNA methylation. According to the available genome-wide methylation profiles and NGS data for bladder cancer in The Cancer Genome Atlas (TCGA), the GRN model to identify the genome-wide IGEN can also characterize DNA methylation and miRNA regulation. In this chapter, we identified the IGENs in normal bladder cells and bladder cancer cells and then investigated the impact of epigenetic regulation and miRNA regulation on bladder carcinogenesis by comparing the IGEN in normal bladder cells with that in bladder cancer cells.

Although a genome-wide IGEN can be identified based on well-defined system identification techniques [2,12], the mean by which the core network biomarkers are extracted from the identified genome-wide network is still an important issue. The total association capabilities of a single node can affect the contribution it makes to its neighbors. However, the genome-wide IGEN including transcriptional gene regulations, miRNA regulations, and PPIs constitutes a genome-wide network structure. The contribution made by one node to its neighbors is not sufficient to explain its impact on a genome-wide scale network of bladder cells. In this chapter, we applied a principal genome-wide network projection (PGNP) based on principal component analysis (PCA) to identify core network biomarkers in bladder carcinogenesis, with the objective of extracting the most significant part from a genome-wide network structure. Because the drug response genome-wide microarray data are now available [15], we analyzed the drug response microarray data of the core network biomarkers to design multiple drug combinations with minimal side effects for bladder cancer treatment. Therefore, the identified core network biomarkers could provide an opportunity to design such drug combinations for bladder cancer treatment. Furthermore, it has been reported that aging (more than 45 years old) and smoking are two major risk factors for bladder carcinogenesis [16]. Therefore, we used the core network biomarkers to elucidate the cellular mechanisms by which aging and smoking elevate bladder cancer risk through epigenetic regulation, miRNA regulation, and signaling pathways.

According to the strategy shown in the flowchart (Fig. 17.1), we integrated omics data, including genome-wide methylation profiles, NGS expression data, miRNA profiles in TCGA, drug response genome-wide microarray data in the Connectivity Map (CMAP) [15], drug–gene interaction data in the Drug Gene Interaction Database (DGIdb) [17], miRNA–target gene association data in TargetScan [18], PPIs in BioGRID, transcription regulations in the Human Transcriptional Regulation Interactions database (HTRIdb) [19], the Integrated Transcription Factor Platform (ITFP) [20], and the TRANSFAC [21], biological processes and pathways in a gene ontology (GO) database, the National Center for Biotechnology Information (NCBI) Entrez Gene database, and the Kyoto Encyclopedia of Genes

**FIGURE 17.1**

Flowchart of the proposed method for constructing the core network biomarkers and identifying bladder carcinogenesis mechanisms.

and Genomes (KEGG) pathway database [22]. We used miRNA–target gene association data, PPIs, and transcription regulations to build the candidate IGEN for general molecular mechanisms. We then constructed a regression IGEN model to characterize the molecular mechanisms including miRNA regulation, PPIs, transcription regulation, and DNA methylation in cells. To prune the false-positive

connections in the candidate IGEN and identify the model parameters of the IGEN in the real human bladder cells, we used methylation profiles, NGS expression data, and miRNA profiles in normal bladder cells and stage 1 and stage 4 bladder cancer cells. We then applied the constrained least squares method and the Akaike information criterion (AIC) [23], a system order detection method, to prune the false-positive connections for obtaining the real IGENs in the three stages of human bladder carcinogenesis. The three genome-wide real IGENs in normal bladder cells and stages 1 and 4 bladder cancer cells were then projected into the three core networks of the three stages of bladder carcinogenesis, respectively. Because the core networks contain the identified signal transduction pathways, that is, the receptors and TFs of the core network can be directly or indirectly connected by the core proteins/TFs, the proteins/TFs, and the corresponding genes that participate in the identified signaling pathways of the core networks are considered as the core network biomarkers for normal and cancerous cells, respectively. The miRNAs with very different connections in regulating the genes of the core network biomarkers between two cells are also involved in the core network biomarkers. By comparing the identified connections of the IGENs, we investigated how the connection changes of the core network biomarkers from normal bladder cells to stage 1 bladder cancer cells and from stage 1 bladder cancer cells to stage 4 bladder cancer cells contribute to bladder carcinogenesis.

We also investigated how the module network of the core network biomarkers, including the KEGG pathways and biological processes, participates in bladder carcinogenesis. According to the information on the biological processes and signaling pathways in the GO database, the NCBI Entrez Gene database, and the KEGG pathway database, the roles of the TFs/proteins in the core network biomarkers are projected into three pathways: the SUMOylation, ubiquitination, and proteasome (SUP) pathway; the tumor necrosis factor (TNF) signaling pathway; and the endoplasmic reticulum (ER) signaling pathway. The roles of the downstream genes in the core network biomarkers are projected into three biological processes: cell proliferation, DNA repair, and metastasis. The module network, including the KEGG pathways, TFs, miRNAs, and biological processes, is connected according to the three identified IGENs in the three types of bladder cell. By comparing the connection changes of the module networks from normal bladder cells to stage 1 bladder cancer cells, and from stage 1 bladder cancer cells to stage 4 bladder cancer cells, we ultimately unraveled the cellular mechanisms behind bladder carcinogenesis and proposed two multiple drug combinations for treating stage 1 and stage 4 bladder cancers, respectively.

Additionally, to determine how the two major risk factors, aging and smoking, influence bladder carcinogenesis, we highlighted not only the significantly expressed genes between smokers and nonsmokers but also the significantly expressed genes between young (≤ 45 years old) and old (> 45 years old) people in the core network biomarkers of bladder carcinogenesis. Finally, we investigated the carcinogenic mechanism of human bladder cells by which the identified major factors, including downregulated miR-1-2, aging, and smoking, lead to the progression from normal bladder cells to stage 1 bladder cancer cells through the SUP and ER signaling pathways. The smoking-related protein HSP90AA1 and DNA methylation of *ECT2* mediates the progression from stage 1 bladder cancer cells to metastasis in stage 4 bladder cancer. Activated DNA repair and accumulated epigenetic alterations lead to the phenotypic changes of bladder cells from normal to cancerous, and from cancerous to metastatic cells owing to the immortality of cancer cells. Based on the core network

biomarkers in bladder carcinogenesis, a multiple drug combination comprising gefitinib, estradiol, yohimbine, and fulvestrant was designed for treating stage 1 bladder cancer with minimal side effects, whereas a multiple drug combination comprising gefitinib, estradiol, chlorpromazine, and LY294002 was designed for treating stage 4 bladder cancer with minimal side effects.

MATERIALS AND METHODS

According to the flowchart in Fig. 17.1, we constructed a candidate human IGEN by mining large databases, including BioGRID, TargetScan, HTRIdb, TRANSFAC, and ITFP. However, many false-positive and insignificant connections existed in the candidate human IGEN for normal and cancerous bladder cells. Using the NGS expression data, miRNA profiles, and the methylation profiles of normal and cancerous bladder cells in TCGA, we identified the association parameters of the network connections. We also applied AIC to detect the systems order, that is, the number of connections, and to delete the insignificant connections that were out of system order to prune the false-positive connections in the candidate IGEN and obtain the two real IGENs for normal and cancerous bladder cells, respectively. By applying PGNP to the two real IGENs in normal and cancerous cells, we first identified the core proteins/TFs that played a major role in the principal networks of the IGENs, constituting the core IGENs in normal and cancerous cells. To determine how the signaling cascades from the core receptor proteins to the core TFs participate in bladder carcinogenesis, the core proteins, which mediate the signal transductions from the core receptor proteins to the core TFs, and their corresponding genes were considered the core network biomarkers of the normal and cancerous cells. The miRNAs with very different connections in regulating the genes of the core network biomarkers between normal and cancerous cells were also involved in the core network biomarkers. Finally, by comparing the connection changes of the core network biomarkers from normal cells to stage 1 cancer cells, and from stage 1 cancer cells to stage 2 cancer cells, we investigated the cellular mechanisms of bladder carcinogenesis.

DATA PREPROCESSING OF OMICS DATA

We downloaded the genome-wide mRNA and miRNA NGS data and the methylation profiles from TCGA, including 17 samples for normal bladder cells, 348 samples for stage 1 bladder cancer cells, and 56 samples for stage 4, that is, metastatic stage, bladder cancer cells. The data also contained six samples for young (≤ 45 years old) people, 477 samples for old (> 45 years old) people, 98 samples for nonsmokers, and 323 samples for smokers. We used one-way analysis of variance (ANOVA) to identify significant differences in gene expression between smokers and nonsmokers, and between young and old people (P value $< .05$). We used the gene symbols of the human gene information data downloaded from the NCBI FTP site as standard human gene names to integrate the omics data, including NGS data, methylation profiles, drug response genome-wide microarray data in CMAP, drug–gene interaction DGIdb data, miRNA–target gene association data in TargetScan, PPIs in BioGRID, transcription regulations in HTRIdb, and ITFP and TRANSFAC data. We also used the GO database, the NCBI Entrez Gene database, and the KEGG pathway database to find the biological processes and pathways of each gene. We used Matlab’s text-file and string manipulation tools for text mining.

CONSTRUCTION OF THE STOCHASTIC REGRESSION MODELS FOR THE IGEN SYSTEM

The goal of the stochastic regression model is to characterize molecular mechanisms, including PPIs, transcription regulations, miRNA regulations, and epigenetic regulations via DNA methylation, by NGS data through detecting false positives of candidate IGENs in human cells. For the stochastic regression model of the gene regulatory subnetwork in the candidate human IGEN, including transcription regulations, miRNA regulations, and epigenetic regulations via DNA methylation, we identified the regulation capabilities of TFs and miRNAs in the GRN of the candidate IGEN. For the expression levels of the i th gene, its DNA methylation and its j th TF/protein and l th miRNA in the n th sample are denoted by $x_i(n)$, $m_i(n)$, $y_j(n)$, and $s_l(n)$, respectively. Then, the stochastic regression model of GRN is described by the following stochastic regression equation:

$$x_i(n) = \sum_{\substack{j \in \Omega_i \\ j \neq i}} a_{ij} M_i(n) y_j(n) + \sum_{l \in \delta_i} c_{li} M_i(n) x_i(n) s_l(n) + b_i M_i(n) + v_i(n), \quad (17.1)$$

for $i = 1, \dots, K$, $n = 1, \dots, N$,

where the repression ability from the l th miRNA to the i th gene $c_{li} \leq 0$; the basal level of the i th gene expression $b_i \geq 0$; $\Omega_i \subset \Omega \equiv \{1, \dots, K\}$; $\delta_i \subset \delta \equiv \{1, \dots, L\}$; $M_i(n) = 1/[1 + (m_i(n)/0.5)^2]$; Ω_i and δ_i denote the candidate regulations based on the databases of transcription regulation and miRNA—target association, respectively; a_{ij} indicates the regulatory ability from the j th TF $y_j(n)$ to the i th gene; $v_i(n)$ represents the stochastic noise due to the modeling residue and fluctuation in the i th gene; and K , L , and N are the total number of TFs, miRNAs, and data samples in the omics data, respectively. $M_i(n)$ denotes the effect of methylation $m_i(n)$ on the binding affinity of TFs, miRNAs, or RNA polymerase on the i th gene which also represents the impact of DNA methylation of the i th gene on the binding affinities of miRNAs, RNA polymerase, and TFs in the gene expression process. The effect on binding affinities $M_i(n)$, for $i = 1, \dots, K$, ranged between 0.2 and 1, whereas the expression range of the genome-wide DNA methylation $m_i(n)$, for $i = 1, \dots, K$, is between 0 and 1. If DNA methylation of the i th gene is close to 1, the effect on the binding affinity to the i th gene is close to 0.2, which implicates the impact of DNA methylation on the binding affinities of miRNAs, RNA polymerase, and TFs to be like an inhibitor. The i th mRNA expression results from transcription regulations

$$\sum_{j \in \Omega_i, j \neq i} a_{ij} M_i(n) y_j(n), \text{ miRNA repressions } \sum_{l \in \delta_i} c_{li} M_i(n) x_i(n) s_l(n), \text{ the mRNA basal expression } b_i M_i(n),$$

and the stochastic noise due to measurement and random fluctuations $v_i(n)$. In model (17.1), the TF regulations, miRNA regulations, and basal levels are all influenced by the DNA methylation $m_i(n)$ on the i th gene.

For the stochastic regression model of the miRNA regulatory subnetwork in the candidate IGEN, the expression levels of the l th miRNA and its i th target gene in the n th sample, denoted by $s_l(n)$ and $x_i(n)$, respectively, could be described by the stochastic regression model of miRNA regulatory network (MRN) as the following stochastic regression equation:

$$s_l(n) = \sum_{i \in \delta_l} c_{li} M_i(n) x_i(n) s_l(n) + M_l(n) z_l + e_l(n), \quad \text{for } l = 1, \dots, L, n = 1, \dots, N, \quad (17.2)$$

where the repression ability of the l th miRNA to the i th gene $c_{li} \leq 0$; the basal level of the l th miRNA expression $z_l \geq 0$; $\delta_l \subset \delta \equiv \{1, \dots, L\}$; δ_l denotes the candidate regulations of the l th miRNA based on the database of miRNA-target gene association; $e_l(n)$ represents the stochastic noise owing to the modeling residue and fluctuation in the l th miRNA. The l th miRNA expression in (17.2) results from miRNA–gene interactions $\sum_{i \in \delta_l} c_{li} M_i(n) x_i(n) s_l(n)$, the miRNA basal expression z_l , and the stochastic noise $e_l(n)$.

For the stochastic regression model of the PPI subnetwork in the candidate IGEN, the expression level of the j th protein and its k th connecting protein in n th sample, denoted by $y_j(n)$ and $y_k(n)$, respectively, could be described by the stochastic regression model of PPIN as the following stochastic regression equation:

$$y_j(n) = \sum_{k \in \Omega_j, k \neq j} d_{jk} y_k(n) + h_j + \omega_j(n), \quad \text{for } j = 1, \dots, K, \quad n = 1, \dots, N, \quad (17.3)$$

where the basal level of the j th protein expression $h_j \geq 0$; $\Omega_j \subset \Omega \equiv \{1, \dots, K\}$; Ω_j denotes the candidate interactions of the j th protein based on the PPI database; d_{jk} indicates the interaction ability of the k th protein to the j th protein; and $\omega_j(n)$ represents the stochastic noise owing to the modeling residue and fluctuation in the j th protein. The j th protein expression in (17.3) results from the rate of formation of the protein complex $y_k(n)y_j(n)$ proportional to the product of the concentration of each protein $\sum_{k \in \Omega_j, k \neq j} d_{jk} y_k(n) y_j(n)$, the protein basal expression $\omega_j(n)$ and the stochastic noise.

We proposed general stochastic regression models to characterize cellular mechanisms, including genetic and epigenetic regulations, in human cells. A number of parameters, including the TF regulatory ability a_{ij} , the miRNA repression ability c_{li} , and the protein interaction ability d_{jk} , needed to be estimated and were determined using the databases of PPI, miRNA–target gene association, and transcription regulation.

IDENTIFICATION OF THE TF REGULATORY ABILITY a_{ij} , THE MI RNA REPRESSION ABILITY c_{li} , AND THE PROTEIN INTERACTION ABILITY d_{jk} AND THEIR STATISTICAL SIGNIFICANCE TESTING

We used the mRNA and miRNA expression data from the NGS as the expression levels for $x_i(n)$ and $s_l(n)$, respectively, and used DNA methylation profiles as the expression level of $m_i(n)$ to identify the model parameters a_{ij} , c_{li} , d_{jk} , b_i , z_l , and h_j in (17.1)–(17.3). Because large-scale measurement of protein activities has yet to be realized and 73% of the variance in protein abundance can be explained by mRNA abundance [24], mRNA expression profiles were always used to substitute for the protein expression profiles. Therefore, we also applied mRNA expression levels in the NGS data as the expression levels of $y_j(n)$ to identify the parameters in (17.1)–(17.3). If the simultaneously measured genome-wide protein expression data and the mRNA expression data in each bladder cancer stage are available, the general models in (17.1)–(17.3) can also be applied to identify the real IGEN of the cancer more precisely. The regulatory parameters were identified by solving the constrained least square parameter estimation problem in the following, because the parameters in (17.1) have certain constraints, such as the nonpositive miRNA repressions and nonnegative basal levels.

To identify the parameters in (17.1), the stochastic regression model of GRN was rewritten as the following linear regression form:

$$\begin{aligned} x_i(n) &= [M_i(n)y_1(n) \quad \cdots \quad M_i(n)x_i(n)y_K(n)M_i(n)x_i(n)s_1(n) \quad \cdots \quad M_i(n)x_i(n)s_L(n) \quad M_i(n)], \\ \begin{bmatrix} a_{i1} \\ \vdots \\ a_{iK} \\ c_{1i} \\ \vdots \\ c_{Li} \\ b_i \end{bmatrix} + v_i(n) &= \phi_i(n)\theta_i^1 + v_i(n), \quad \text{for } j = 1, \dots, K, \quad n = 1, \dots, N, \end{aligned} \tag{17.4}$$

where $\phi_i(n)$ denotes the regression vector and θ_i^1 is the parameter vector of target gene i to be estimated. $x_i(n)$ and $\phi_i(n)$ are available in the omics data.

The regression model (17.4) at different data samples can be rearranged as follows:

$$\begin{bmatrix} x_i(1) \\ \vdots \\ x_i(N) \end{bmatrix} = \begin{bmatrix} \phi_i(1) \\ \vdots \\ \phi_i(N) \end{bmatrix}\theta_i^1 + \begin{bmatrix} v_i(1) \\ \vdots \\ v_i(N) \end{bmatrix}, \tag{17.5}$$

where N denotes the number of data samples in the NGS data of a bladder cancer stage.

For simplicity, we define the notations X_i , Φ_i , and V_i to represent (17.5) as follows:

$$X_i = \Phi_i\theta_i^1 + V_i. \tag{17.6}$$

The constrained least square parameter estimation problem of θ_i^1 is formulated as follows:

$$\begin{aligned} \min_{\theta_i^1} & \| \Phi_i\theta_i^1 - X_i \|_2^2 \\ \text{subject to} & \underbrace{\begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & -1 \end{bmatrix}}_{\text{subject to}} \theta_i^1 \leq \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \end{aligned} \tag{17.7}$$

This gives the constraints to force the miRNA repression c_{li} to be always nonpositive and the basal level b_i to be always nonnegative in (17.1); that is, $c_{li} \leq 0$ and $b_i \geq 0$. The constrained least square problem was solved using the active set method for quadratic programming [25,26].

Similarly, the stochastic regression model of the miRNA regulatory subnetwork in (17.2) was rewritten in the following regression form:

$$\begin{aligned} s_l(n) &= [M_1(n)x_1(n)s_l(n) \quad \cdots \quad M_K(n)x_K(n)s_l(n) \quad M_l(n)] \begin{bmatrix} c_{l1} \\ \vdots \\ c_{lK} \\ z_l \end{bmatrix} + e_l(n) \\ &= \vartheta_l(n)\theta_l^2 + e_l(n), \quad \text{for } j = 1, \dots, K, \quad n = 1, \dots, N, \end{aligned} \tag{17.8}$$

where $\vartheta_l(t)$ indicates the regression vector and θ_l^2 is the parameter vector to be estimated.

For simplicity, we define the notations S_l , ψ_l , and E_l to represent (17.8) as follows:

$$S_l = \psi_l \theta_l^2 + E_l. \quad (17.9)$$

The parameter identification problem is then formulated as follows:

$$\begin{aligned} & \min_{\theta_l^2} \|\psi_l \theta_l^2 - S_l\|_2^2 \\ \text{subject to } & \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -1 \end{bmatrix}}_K \theta_l^2 \leq \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \end{aligned} \quad (17.10)$$

This gives the constraint to force the miRNA repression c_{li} to be always nonpositive and the basal level z_l to be always nonnegative in (17.2); that is, $c_{li} \leq 0$ and $z_l \geq 0$. Finally, the protein model (17.3) uses the same way like above to make sure $h_j \geq 0$.

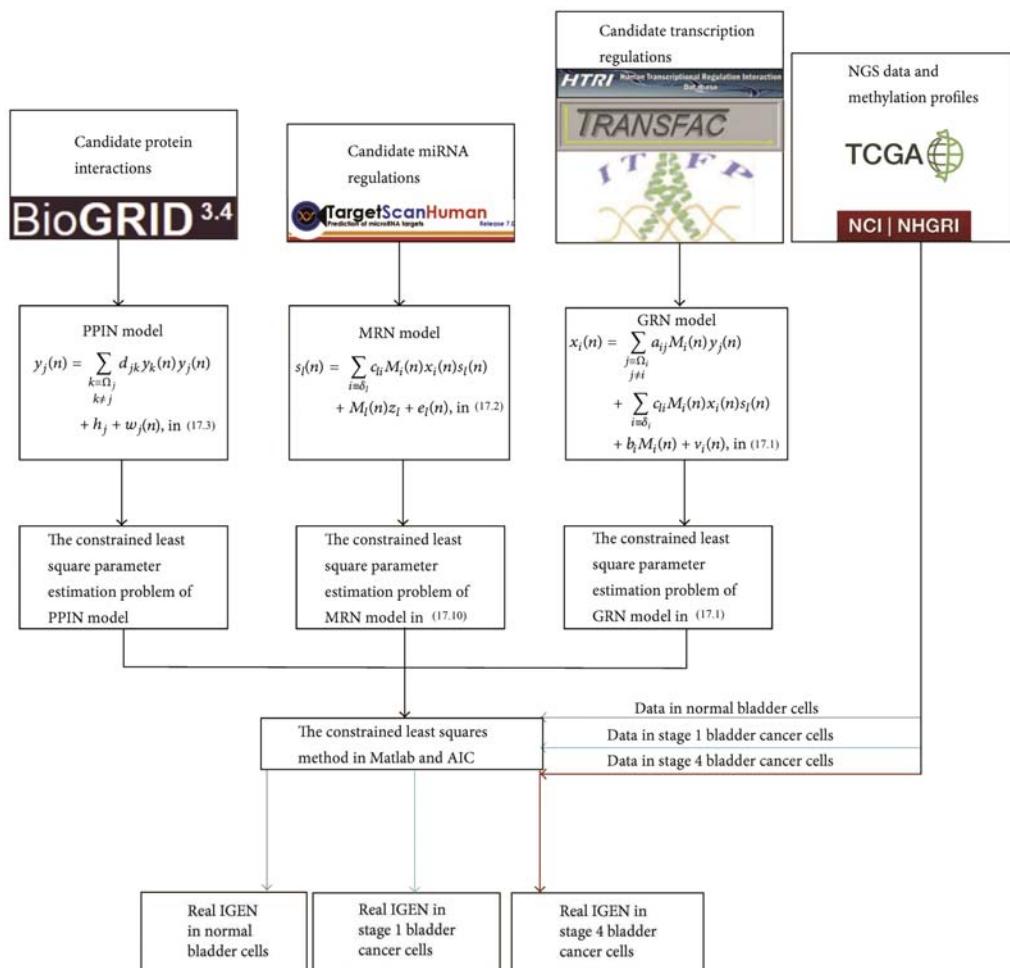
Furthermore, to extract the core network biomarkers from normal and cancerous cells, we first used NGS data and methylation profiles in the normal and stage 1 and 4 bladder cancer cells to identify an IGEN for normal bladder cells and a general IGEN for bladder cancer cells. The two identified IGENs were used to extract the core network bladder carcinogenesis. We then used the association parameters in the general IGEN of bladder cancer cells as the initial condition of the constrained least square parameter estimation and applied the data on stage 1 and 4 bladder cancer cells to identify the IGENs for stages 1 and 4, respectively. According to the three identified IGENs in normal bladder cells, and stage 1 and 4 bladder cancer cells, we determined the cellular mechanisms of the core network biomarkers in bladder carcinogenesis. The proposed methodology to identify the IGENs for normal bladder cells and stage 1 and 4 bladder cancer cells was summarized in the flowchart in Fig. 17.2.

By applying Student's *t*-test to the parameter estimation method [27], the *P* values for the estimated parameters, including the TF regulatory ability a_{ij} , the miRNA repression ability c_{li} , and the protein interaction ability d_{jk} , were calculated to determine the significance of the parameters. Additionally, to determine the significance of expression level and DNA methylation profile of a gene/miRNA between normal bladder cells and cancerous bladder cells, we applied one-way ANOVA to calculate the *P* value.

After the parameter identification problem had been solved, we identified the IGEN for each bladder cell type. For example, we identified the regulatory parameter $a_{RPS20,JUN} = 0.26$ from the TF JUN to the target gene RPS20 (*P* value $< .02$) in stage 4 bladder cancer cells, the interaction parameter $d_{HUWE1,ADRM1} = 1.2$ between the two proteins ADRM1 and HUWE1 (*P* value $< .005$) in stage 1 bladder cancer cells, and the coupling rate $c_{RPS20,MIR155} = -1.2$ between the miRNA miR155 and the mRNA RPS20 in stage 4 bladder cancer cells (*P* value $< .07$).

PRINCIPAL GENOME-WIDE NETWORK PROJECTION

After the identification of the IGENs in normal and cancer cells, we extracted the core network biomarkers of the IGENs based on the perspectives of the functional modules and pathways to reveal

**FIGURE 17.2**

Flowchart of the proposed methodology to identify the IGENs for normal bladder cells, and stages 1 and 4 bladder cancer cells.

the cellular mechanisms behind bladder carcinogenesis. To extract the core network biomarkers, including the core proteins, their corresponding genes, and their upstream miRNAs, from an IGEN on a genome-wide scale, we first decomposed the combined network matrix of the IGEN to left- and right-singular vectors and singular values based on singular value decomposition. The top left- and right-singular vectors with the top singular values constitute the principal network of the IGEN. The projection distance of each gene/protein/miRNA to these top singular vectors represents the significance of this gene/protein/miRNA in the IGEN. The genes/proteins/miRNAs with the top projection

distance ultimately constitute the core network biomarkers of the IGEN. Let the combined network matrix of the TF regulatory ability a_{ij} , the miRNA repression ability $c_{li}\sqrt{a^2 + b^2}$, and the protein interaction ability d_{jk} of the IGEN in (17.1)–(17.3) be represented by

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \\ c_{11} & \cdots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{L1} & \cdots & c_{LK} \\ d_{11} & \cdots & d_{1K} \\ \vdots & \ddots & \vdots \\ d_{K1} & \cdots & d_{KK} \end{bmatrix}. \quad (17.11)$$

By applying PGP, the matrix A is then be decomposed as follows:

$$\begin{aligned} A &= UDV^T \\ &= [u_1 \ \cdots \ u_K] \begin{bmatrix} d_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_K \end{bmatrix} [v_1 \ \cdots \ v_K]^T \\ &= \sum_{i=1}^K u_i d_i v_i^T, \end{aligned} \quad (17.12)$$

where $u_i, v_i \in \mathbb{R}^K$ are the i th left- and right-singular vectors of A , respectively. The diagonal entries of D are the K singular values of A in descending order, $d_1 \geq \cdots \geq d_K$.

$$E_m = \frac{d_m^2}{\sum_{m=1}^K d_m^2}. \quad (17.13)$$

We choose the top M singular vectors of V such that $\sum_{m=1}^M E_m \geq 0.85$, with the minimal M , so that the top M principal components contain 85% of the IGEN from an energy point of view. The principal projections of A to the top M singular vectors of V , or similarities, are defined as follows:

$$S(k, m) = a_k \cdot v_m^T, \quad \text{for } k = 1, \dots, (2K + L), \ m = 1, \dots, M, \quad (17.14)$$

where a_k and v_m^T denote the k th row vector of A and the m th singular vector of V , respectively. Furthermore, we defined the 2-norm distance from the target genes, miRNAs, and proteins/TFs to the top M singular vectors, respectively, as follows:

$$D(k) = \left[\sum_{m=1}^M [S(k, m)]^2 \right]^{1/2}, \quad \text{for } k = 1, \dots, (2K + L) \quad (17.15)$$

where $D(k)$ for $k = 1, \dots, K$, for $k = K + 1, \dots, K + L$, and for $k = K + L + 1, \dots, 2K + L$ are the 2-norm distances from the target genes, miRNAs, and proteins/TFs to the top M singular vectors, respectively. According to $D(k)$ for $k = K + L + 1, \dots, 2K + L$, we can identify the core proteins/TFs that play a major role in the principal networks of the IGENs, constituting the core IGENs in normal and cancer cells. The identified core proteins/TFs contain receptors that mediate the signaling cascades connected to core TFs. The core proteins, which participate in signal transduction from core receptors to core TFs, and their corresponding genes, were considered the core network biomarkers for normal and cancerous cells. The miRNAs with very different connections in regulating the genes of the core network biomarkers between two cells were also involved in the core network biomarkers.

DESIGN OF A MULTIPLE DRUG COMBINATION WITH MINIMAL SIDE EFFECTS FOR THE TREATMENT OF BLADDER CANCER

To design a multiple drug combination with minimal side effects for the treatment of bladder cancer based on the core network biomarkers of the IGEN, we considered two databases, CMAP and DGIdb. CMAP contains the genome-wide microarray data in response to 1327 drugs in five cell lines, whereas DGIdb comprises a drug–gene interaction database. Multiple drug therapy induces a genome-wide response. The strategy of multiple drug screening is that the multiple drugs should inhibit the highly expressed genes, activate the reduced expression of the genes, and not influence the no differentially expressed genes in the core network biomarkers of bladder cancer cells compared with normal bladder cells. The binding protein of the designed multiple drug combination can also be obtained using the DGIdb. The strategy leads to improved drug safety and efficacy in the treatment of bladder cancer.

RESULTS AND DISCUSSION

CONSTRUCTION OF IGEN

We first used NGS expression data and methylation profiles in normal bladder cells and stage 1 and 4 bladder cancer cells to identify a real IGEN for normal bladder cells and a general real IGEN for bladder cancer cells (see [Materials and Methods](#) section). By applying PGNP to the real IGEN of the normal bladder cells and the general real IGEN of the bladder cancer cells, we then obtained 115 core proteins/TFs for the core IGEN of the normal bladder cells and 138 core proteins/TFs for the core IGEN of the bladder cancer cells. To determine how the signaling cascades from the core receptor proteins to the core TFs participate in bladder carcinogenesis, the core proteins, which mediate the signal transductions from core receptor proteins to core TFs, and their corresponding genes were considered the core network biomarkers. The miRNAs with a high number of different connections regulating the genes of the core network biomarkers between normal and cancerous cells were also involved in the core network biomarkers. Moreover, to identify the mechanism of carcinogenesis from stage 1 to stage 4 bladder cancer, we used the identified parameters of models [\(17.1\)–\(17.3\)](#) in the general IGEN of bladder cancer cells as the initial condition of the constrained least square parameter estimation. We then applied the data for stage 1 and 4 bladder cancer cells to obtain the two real IGENs for stage 1 and 4 bladder cancer, respectively. Furthermore, we analyzed the connection changes of the core network biomarkers between normal bladder cells and stage 1 bladder cancer cells ([Fig. 17.3](#)) and between stage 1 and 4 bladder cancer cells ([Fig. 17.4](#)) to determine the mechanisms of bladder

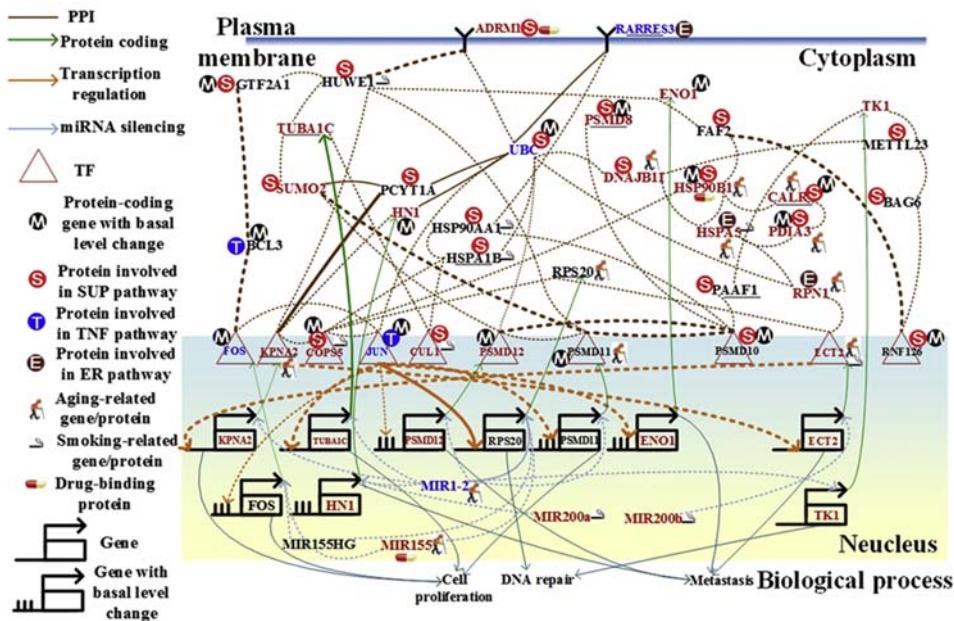


FIGURE 17.3

Comparison of genetic and epigenetic alterations and connection changes in the core network biomarkers of bladder carcinogenesis between normal bladder cells and stage 1 bladder cancer cells. Red, blue, and black gene/miRNA symbols represent the highly expressed genes, the suppressed genes, and the no differentially expressed genes in stage 1 bladder cancer cells, respectively, compared with normal bladder cells. *Dashed and solid lines* denote the identified connections in normal and cancerous cells, respectively. The identified connections of the core network biomarkers do not exist in normal bladder cells only. *Bold lines* indicate the high regulatory or interaction parameters, that is, a_{ij} , c_{li} , and d_{jk} , identified in the stochastic regression models (17.1)–(17.3) of the IGEN. The bold proteins, including RARRES3, TUBA1C, PSMD8, HSPA1B, RPS20, CALR, PAAF1, and KPNA2, were the identified core network biomarkers. The major factors, including downregulated miR1-2, the aging-related proteins, HSP90B1, CALR, HSPA5, PDIA3, RPN1, and ECT2, the smoking-related proteins, HUWE1, HSPA5, and ECT2, and the epigenetic regulation of ENO1, HSP90B1, CALR, and PDIA3, lead to the progression from normal bladder cells to stage 1 bladder cancer cells through the SUP and ER signaling pathways.

carcinogenesis and accordingly design multiple drug combinations for treating bladder cancer with minimal side effects.

To investigate the impact of the major risk factors, aging and smoking, on the core network biomarkers of bladder carcinogenesis, we highlighted the significantly expressed genes between young and old people and between nonsmokers and smokers in the core network biomarkers (P value $< .05$). Additionally, the genes with changes in the basal level of (17.1) between normal bladder cells and stage 1 bladder cancer cells and between stage 1 and 4 bladder cancer cells were also highlighted in the core network biomarkers of Figs. 17.3 and 17.4, respectively. The basal level change of a gene between

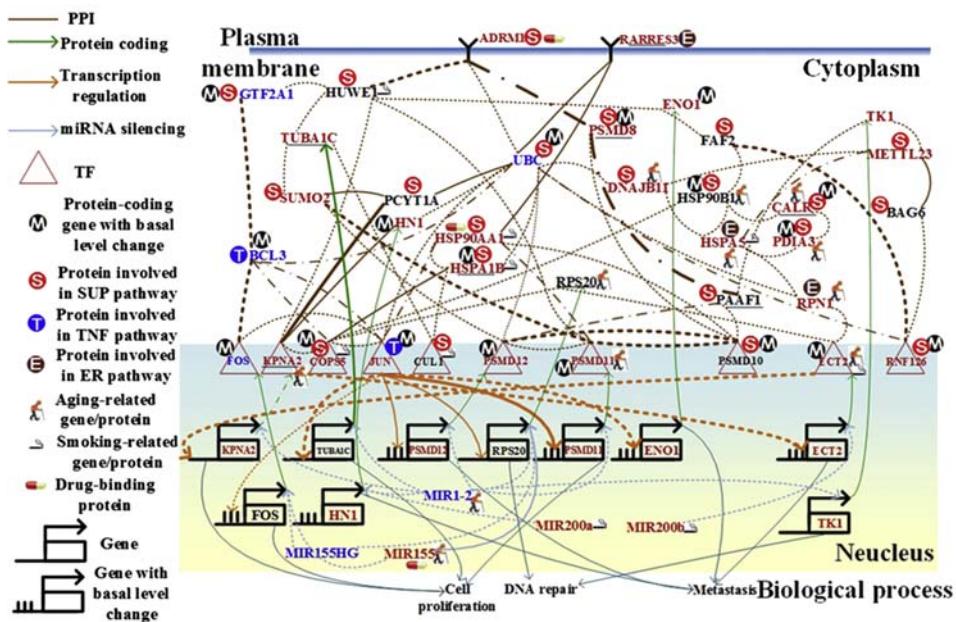


FIGURE 17.4

Comparison of genetic and epigenetic alterations and connection changes in the core network biomarkers of bladder carcinogenesis between stage 1 and stage 4 bladder cancer cells. Red, blue, and black gene/miRNA symbols represent the highly expressed genes, the suppressed genes, and the no differentially expressed genes in stage 4 bladder cancer cells, respectively, compared with stage 1 bladder cancer cells. *Dashed, dash-dot, and solid lines* denote the identified connections in stage 1 cancer cells, stage 4 cancer cells, and both stage 1 and 4 cancer cells, respectively. *Bold lines* indicate the high regulatory or interaction parameters, that is, a_{ij} , c_{li} , and d_{jk} , identified in the stochastic regression models (17.1)–(17.3) of the IGEN. The bold proteins RARRES3, TUBA1C, PSMD8, HSPA1B, RPS20, CALR, PAAF1, and KPNA2 were the identified core network biomarkers. The smoking-related protein HSP90AA1 and DNA methylation of ECT2 mediate metastasis of bladder cancer.

two cell types has been implicated in the epigenetic regulation of gene expression. The expression of a gene that exhibits a basal level change and a significant change (P value $< .05$) of its methylation profile between the two-bladder cell types is probably regulated by DNA methylation in bladder carcinogenesis.

PROJECTION OF THE CORE NETWORK BIOMARKERS INTO BIOLOGICAL PROCESSES AND SIGNALING PATHWAYS TO INVESTIGATE CARCINOGENIC MECHANISMS OF BLADDER CANCER

According to the information of the biological processes and signaling pathways in the GO and KEGG pathway databases, the roles of the genes in the core network biomarkers (Figs. 17.3 and 17.4) are

projected into three pathways: the SUP, TNF signaling, and ER signaling pathways and three biological processes: cell proliferation, DNA repair, and metastasis.

It has been reported that the SUP pathway is associated with increased proliferation in urinary bladder carcinogenesis [28]. HuaChanSu (HCS), a class of toxic steroids, has been used to show that the TNF pathway mediates the inhibition of cell proliferation in bladder cancer [29]. The viability of human bladder cancer cells is reduced by cantharidin through the ER pathway; the latter is a natural toxin [30]. Therefore, the proteins of the core network biomarkers participating in the SUP, TNF, and ER signaling pathways play an important role in bladder carcinogenesis. We then determined how the core network biomarkers mediate bladder carcinogenesis through the influences of aging, smoking, epigenetic regulation, and miRNA regulation.

The role of the SUP pathway is to degrade misfolded proteins, influence PPIs, translocate proteins, and stabilize protein structure. Owing to the accumulation of genetic mutations and epigenetic alterations in cancer cells, the SUP pathway plays a crucial role in the maintenance of many important cellular processes in cancer cells. The repressed activity of ubiquitin C (UBC), encodes for polyubiquitin precursor, is influencing degradation and translation of several proteins in stage 1 and stage 4 bladder cancer cells. For example, the repression of UBC affects the signal transduction of RARRES3, a tumor suppressor, in bladder carcinogenesis. To maintain the cellular functions of cancer cells, the regulation of the SUP pathway adapts to the accumulated genetic mutations and epigenetic alterations.

In normal cells, the TNF pathway is critical for inducing inflammation, which can cause cell death. Accumulated DNA damage, epigenetic alterations, or stresses can induce the TNF pathway, and the pathway then triggers cell death. JUN, one of the TFs in the TNF pathway, plays an important role in promoting the invasion and migration of bladder cancer cells [31]. We determined that the repressed expression of JUN in stage 1 bladder cancer cells leads to cancer cell immortality and causes accumulated genetic mutations and epigenetic alterations. Additionally, the results revealed that JUN was activated in stage 4 bladder cancer cells to mediate metastasis. The role of JUN in the metastasis of bladder cancer cells can also be supported [31]. It has also been reported that the TNF pathway acts as a switch between inflammation and cancer [32]. Moreover, downregulated BCL3, which participates in the TNF pathway in the adipose tissue of the bladder wall, leads to reduce inflammation in bladder carcinogenesis [33].

The ER pathway participates in the regulation of protein folding, protein synthesis, and post-translational modifications [34]. Misfolded proteins, arising from genetic mutations, epigenetic alterations, or stresses, induce the ER pathway to restore cellular homeostasis in normal cells. Owing to the immortal nature of cancer cells, the accumulated genetic mutations and epigenetic alterations in bladder cancer cells can activate most of the genes that contribute to the ER pathway in bladder carcinogenesis (Figs. 17.3 and 17.4). In the ER pathway, only RARRES3, a tumor suppressor gene, was downregulated in stage 1 bladder cancer cells.

THE IMPACT OF AGING, SMOKING, AND miRNA AND EPIGENETIC REGULATION ON BLADDER CARCINOGENESIS THROUGH THE CORE NETWORK BIOMARKERS

Major factors, including downregulated miR1-2 and aging- and smoking-related proteins, may lead to the progression from normal bladder cells to stage 1 bladder cancer cells through the SUP and ER signaling pathways. It has been reported that aging and smoking are the major factors that accumulate

genetic and epigenetic alterations and ultimately induce bladder carcinogenesis. In Fig. 17.3, our results reveal that ADRM1 regulates KPNA2, which promotes proliferation, and is mediated by the aging-related proteins, HSP90B1, CALR, HSPA5, PDIA3, RPN1, and ECT2, the smoking-related proteins, HUWE1, HSPA5, and ECT2, and the epigenetic regulation of ENO1, HSP90B1, CALR, and PDIA3, through the SUP and ER signaling pathways. ADRM1 knockdown leads to a reduction of cancer cell proliferation and has been found in gastric [35], ovarian [36], liver [37], and colorectal cancers [38] and acute leukemia [39]. Therefore, the results support the hypothesis that aging is the most important factor in inducing bladder carcinogenesis through the SUP pathway. Additionally, our results (Fig. 17.3) show that the inhibited aging-related miRNA miR1-2 in stage 1 bladder cancer cells leads to miR1-2 deregulation of genes including *KPNA2*, *TUBA1C*, *HN1*, *PSMD11*, *PSMD12*, and *TK1*, which influence cell proliferation, DNA repair, and metastasis. The miR1-2 has also been identified as a tumor suppressor in bladder cancer cells [40].

MiR1-2 AND MiR200B MEDIATE THE REDUCTION OF CELL PROLIFERATION AND METASTASIS THROUGH KPNA2 AND ECT2, RESPECTIVELY

The receptor ADRM1 signal triggers the signaling cascade from the smoking-related protein HUWE1 to the aging-related proteins HSP90B1 and RPS20 and the smoking-related TF COPS5. The TF COPS5 upregulates the metastasis-associated gene *ECT2*, which is suppressed by miR200b in stage 1 bladder cancer cells. The results show the cross-regulation between the transcription of the smoking-related protein COPS5 and the aging-related protein *ECT2*. The aging-related miRNA miR1-2 and the smoking-related miRNA miR200b act as a switch to depress the proliferation-associated protein KPNA2 in stage 1 and stage 4 bladder cancer cells (Figs. 17.3 and 17.4) and the metastasis-associated gene *ECT2* in stage 4 bladder cancer cells (Fig. 17.4), respectively.

THE SMOKING-RELATED PROTEIN HSP90AA1 AND DNA METHYLATION OF ECT2 MEDIATE THE METASTASIS OF BLADDER CANCER

Our results reveal that receptor RARRES3 signaling triggers the activated TF JUN mediated by the smoking-related protein HSP90AA1, and JUN then activates the metastasis-associated gene *PSMD12* in stage 4 bladder cancer cells (Fig. 17.4). Receptor ADRM1 signaling also triggers the metastasis-associated protein PSMD12 through the proteins PSMD8 and PAAF1 and epigenetic regulation in stage 4 bladder cancer cells. This shows that metastasis-associated *ECT2* is activated by epigenetic regulation in stage 4 bladder cancer cells. Receptor RARRES3 signaling also triggers the aging-related and proliferation-associated TF PSMD11 through the smoking-related protein HSP90AA1 in stage 4 bladder cancer cells. The activated TF JUN also regulates the proliferation-associated gene *PSMD11* and the DNA repair-associated gene *RPS20*. There is evidence that curcumin (diferuloylmethane) can suppress tumor initiation, promotion, and metastasis. Curcumin can also inhibit the expression of JUN [41]. Additionally, the RNAi-induced induction of *ECT2* suppresses cell migration, invasion, and metastasis [42]. Our results indicate that the upregulation of *ECT2* in stage 4 bladder cancer cells is regulated by epigenetic regulation of *ECT2* expression. This is also supported by the significant change in the DNA methylation profiles in *ECT2* between normal bladder cells and stage 4 bladder cancer cells (*P* value < .007).

FUNCTIONAL MODULE NETWORK ANALYSIS IN BLADDER CARCINOGENESIS

The activated DNA repair of bladder cancer cells leads to metastasis owing to the immortality of cancer cells. According to the modular information in the GO database and the KEGG pathway database, the genes/proteins in the core network biomarkers (Figs. 17.5 and 17.6) are projected into three pathways, the SUP pathway, the TNF signaling pathway, and the ER signaling pathway, and

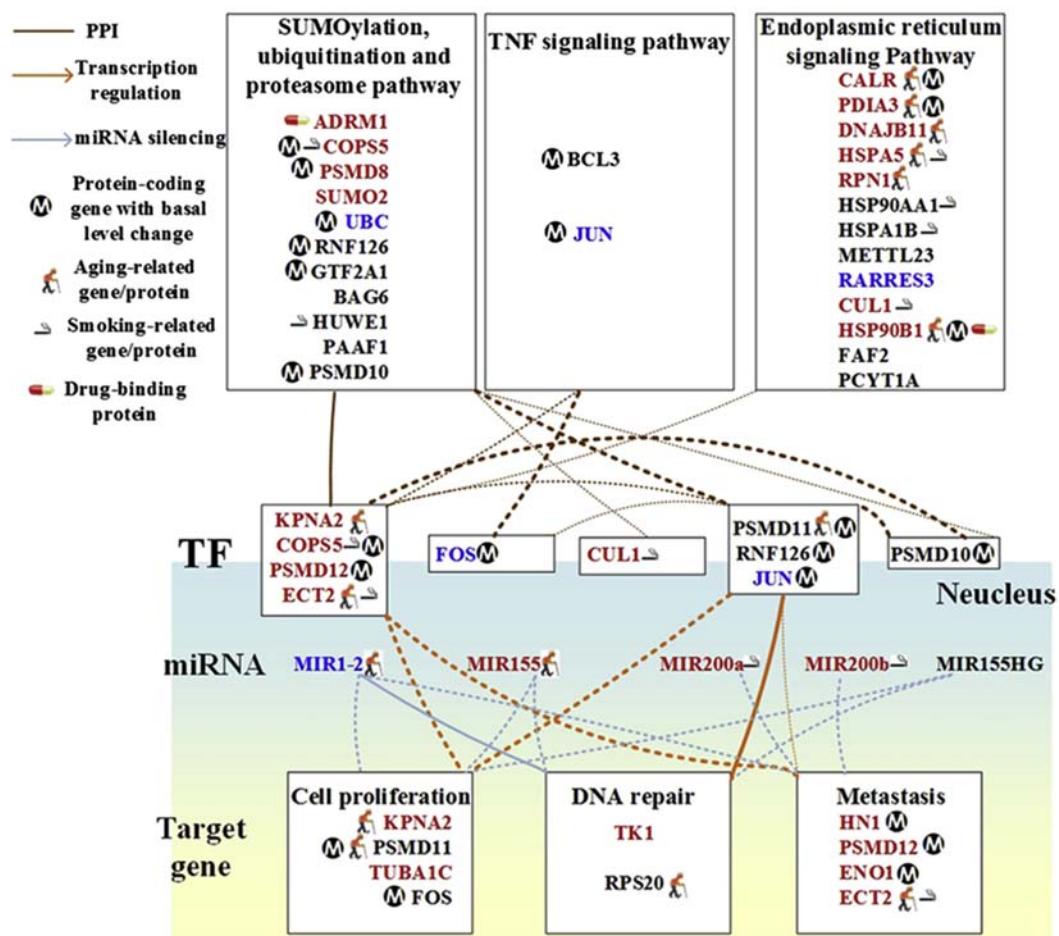


FIGURE 17.5

Module network of the core network biomarkers in Fig. 17.3 for investigating the bladder carcinogenic mechanisms from normal bladder cells to stage 1 bladder cancer cells. The notations of gene/miRNA symbols and line styles are the same as those in Fig. 17.3. The activated TFs KPNA2, COPSS5, PSMD12, and ECT2 play an important role in mediating the signal transduction of the SUP and ER pathways to activate cell proliferation and metastasis in stage 1 bladder cancer. The metastasis of the stage 1 bladder cancer is repressed by the activated miRNAs miR200a and miR200b.

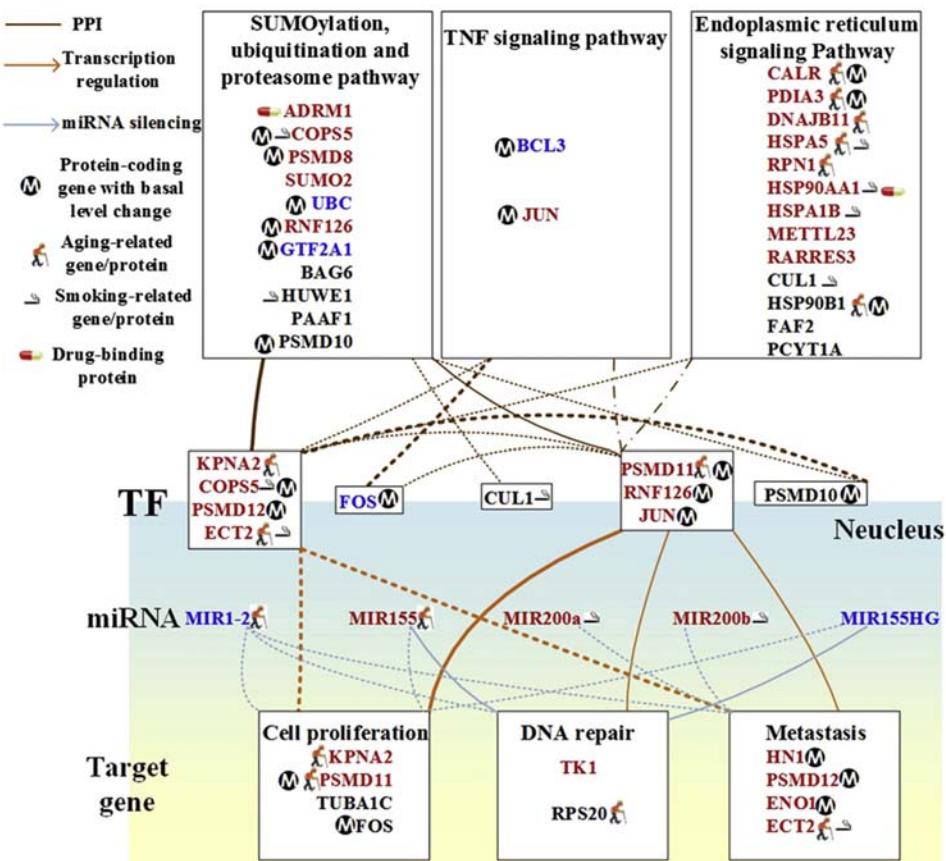


FIGURE 17.6

Module network of the core network biomarkers in Fig. 17.4 to investigate the bladder carcinogenic mechanisms from stage 1 to stage 4 bladder cancer cells. The notations of gene/miRNA symbols and line styles are the same as those in Fig. 17.4. The activated DNA repair of bladder cancer cells leads to metastasis owing to the immortality of cancer cells. The activated JUN in the TNF pathway induces cell proliferation, DNA repair, and metastasis in stage 4 bladder cancer cells.

three biological processes: cell proliferation, DNA repair, and metastasis. The module networks in Figs. 17.5 and 17.6 show that the activated TFs KPNA2, COPSS5, PSMD12, and ECT2 play an important role in mediating the signal transduction of the SUP and ER pathways to activate cell proliferation and metastasis in stage 1 bladder cancer. The metastasis of the stage 1 bladder cancer is repressed by the activated miRNAs miR200a and miR200b, as shown in Fig. 17.5. The activated signal transduction from SUP and ER pathways also triggers DNA repair through the epigenetically regulated TFs PSMD11 and RNF126.

Additionally, the activated TFs PSMD11, RNF126, and JUN mediate the signal transduction from SUP, TNF, and ER pathways to trigger cell proliferation, DNA repair, and metastasis in stage 4 bladder

cancer, as shown in Fig. 17.6. Although miR155 is activated in stage 1/4 bladder cancer, miR155 suppresses FOS and RPS20 in stage 1 bladder cancer, and miR155 only suppresses the DNA repair-associated gene *RPS20* in stage 4 bladder cancer. Furthermore, we suggest that DNA repair may play a critical role in repairing DNA damage, which results from genetic and epigenetic alterations, leading to phenotypic change of the bladder cells from normal cells to stage 1 cancer cells, and from stage 1 cancer cells to metastatic cancer cells.

In summary, aging and epigenetic regulation dominate bladder carcinogenesis through CALR, PDIA3, DNAJB11, HSPA5, RPN1, HSP90B1, KPNA2, ECT2, and PSMD11 and through COPSS5, PSMD8, RNF126, CALR, PDIA3, HSP90B1, PSMD12, PSMD11, JUN, HN1, and ENO1, respectively. Smoking promotes bladder carcinogenesis especially in metastasis. Finally, the cellular mechanisms from normal to stage 1 bladder cancer cells and from stage 1 to stage 4 bladder cancer cells are summarized in Fig. 17.7A and B, respectively. When the accumulated genetic mutations and epigenetic alterations lead to the deregulation of the TNF pathway in inflammation, the accumulated misfolded proteins in the ER pathway induce cell proliferation in stage 1 bladder cancer (Fig. 17.7A). Regulation of the ER and TNF pathways adapts to the accumulated genetic mutations and epigenetic alterations through the SUP pathway. The progression of DNA repair and cell proliferation in stage 1 bladder cancer ultimately results not only in the repression of miR200a and miR200b during metastasis but also in the regulation of the TNF pathway to metastasis, cell proliferation, and DNA repair in stage 4 bladder cancer (Fig. 17.7B).

TWO SEPARATE DRUG COMBINATIONS FOR TREATING STAGE 1 AND STAGE 4 BLADDER CANCER CELLS WITH MINIMAL SIDE EFFECTS

The design of a multiple drug combination for treating stage 1 bladder cancer depends on a strategy of inhibiting the highly expressed genes *ADRM1*, *COPS5*, *PSMD8*, *SUMO2*, *CALR*, *PDIA3*, *DNAJB11*, *HSPA5*, *RPN1*, *CUL1*, *HSP90B1*, *KPNA2*, *PSMD12*, *ECT2*, *TK1*, *TUBA1C*, *HN1*, and *ENO1*; activating the suppressed genes *UBC*, *JUN*, *RARRES3*, and *FOS*; and suppressing the drug's effect on the no differentially expressed genes *BAG6*, *HUWE1*, *PAAF1*, *PSMD10*, *FAF2*, *PCYT1A*, and *PSMD10*. According to the drug design strategy (see Materials and Methods section), a multiple drug combination comprising gefitinib, estradiol, yohimbine, and fulvestrant was obtained for treating stage 1 bladder cancer.

The design of a multiple drug combination for treating stage 4 bladder cancer depends on a strategy of inhibiting the highly expressed genes *ADRM1*, *COPS5*, *PSMD8*, *SUMO2*, *RNF126*, *CALR*, *PDIA3*, *DNAJB11*, *HSPA5*, *RPN1*, *HSP90AA1*, *HSPA1B*, *METTL23*, *RARRES3*, *KPNA2*, *PSMD12*, *ECT2*, *JUN*, *TK1*, *TUBA1C*, *HN1*, and *ENO1*; activating the suppressed genes *BCL3*, *FOS*, *UBC*, and *GTF2AI*; and suppressing the drug's effect on the no differentially expressed genes, which are the same as those in stage 1 bladder cancer. We obtained a multiple drug combination comprising gefitinib, estradiol, chlorpromazine, and LY294002 for treating stage 4 bladder cancer. According to the information in DGIdb, miR-155, the HSP90 protein family, ADRM1, and estrogen receptor are the direct targets of the multiple drug combination comprising gefitinib, estradiol, yohimbine, and fulvestrant in stage 1 bladder cancer, respectively (Figs. 17.3 and 17.5), whereas the same proteins are also the direct targets of the multiple drug combination comprising gefitinib, estradiol, chlorpromazine, and LY294002 in stage 4 bladder cancer, respectively (Figs. 17.4 and 17.6). Moreover, the analysis of drug response genome-wide microarray data reveals that high doses of yohimbine can

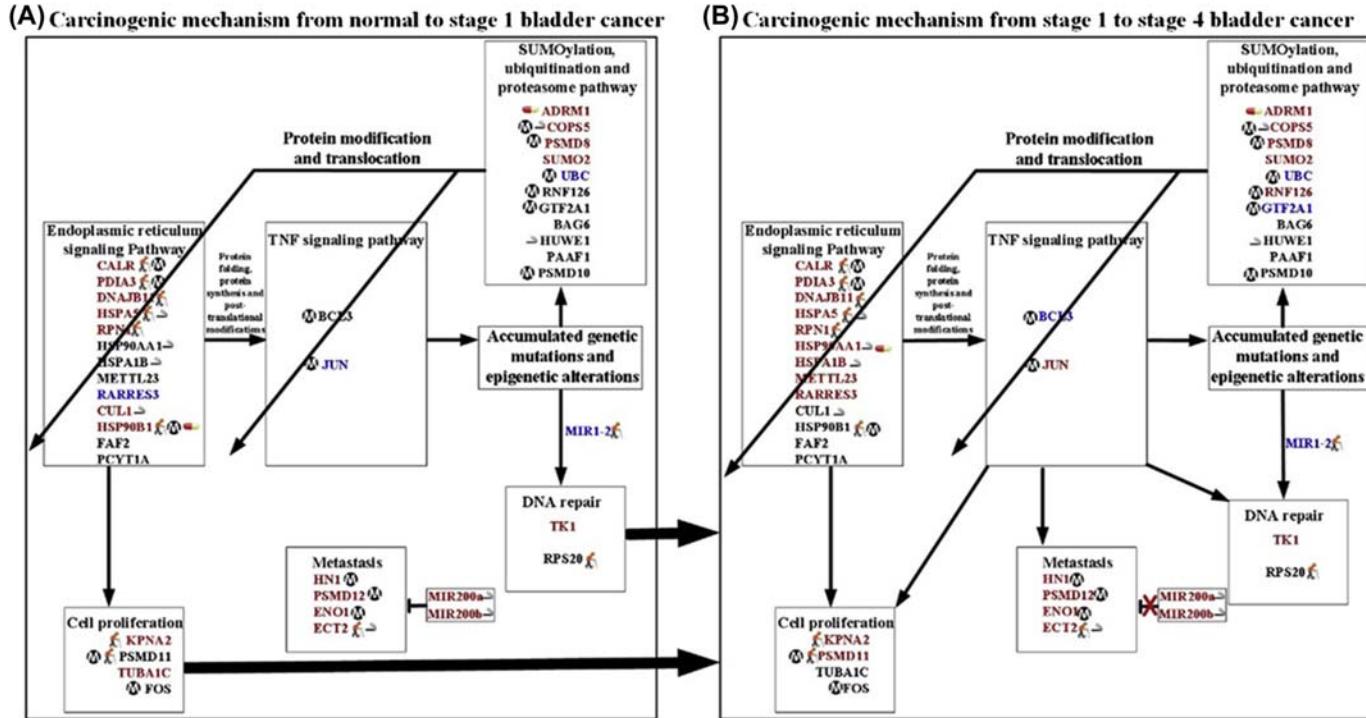
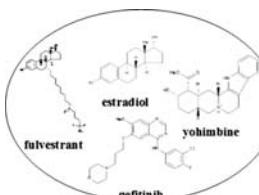
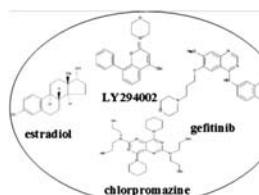


FIGURE 17.7

The carcinogenic mechanisms from normal to stage 1 bladder cancer cells (A), and from stage 1 to stage 4 bladder cancer cells (B). When the accumulated genetic mutations and epigenetic alterations lead to the deregulation of the TNF pathway in inflammation, the accumulated misfolded proteins in the ER pathway induce cell proliferation in stage 1 bladder cancer (A). The regulations of ER and TNF pathways are adaptive to the accumulated genetic mutations and epigenetic alterations through the SUP pathway. The progression of DNA repair and cell proliferation in stage 1 bladder cancer ultimately results not only in the repression of miR200a and miR200b during metastasis but also in the regulation of the TNF pathway to metastasis, cell proliferation, and DNA repair in stage 4 bladder cancer (B).

Table 17.1 The Multiple Drug Design Strategy and Potential Multiple Drug Combination for Stage 1 and 4 Cancers

	Stage 1 Bladder Cancer	Stage 4 Bladder Cancer
The highly expressed genes for potential inhibition strategy of multiple drug design	<i>ADRM1, COPSS5, PSMD8, SUMO2, CALR, PDIA3, DNAJB11, HSPA5, RPN1, CUL1, HSP90B1, KPNA2, PSMD12, ECT2, TK1, TUBA1C, HN1, and ENO1</i>	<i>ADRM1, COPSS5, PSMD8, SUMO2, RNF126, CALR, PDIA3, DNAJB11, HSPA5, RPN1, HSP90AA1, HSPA1B, METTL23, RARRES3, KPNA2, PSMD12, ECT2, JUN, TK1, TUBA1C, HN1, and ENO1</i>
The suppressed genes for potential activation strategy of multiple drug design	<i>UBC, JUN, RARRES3, and FOS</i>	<i>BCL3, FOS, UBC, and GTF2A1</i>
The nondifferentially expressed genes to avoid side effect of multiple drug design	<i>BAG6, HUWE1, PAAF1, PSMD10, FAF2, PCYTIA, and PSMD10</i>	<i>BAG6, HUWE1, PAAF1, PSMD10, FAF2, PCYTIA, and PSMD10</i>
The potential multiple drug combination		

activate BAG6 in stage 1 bladder cancer, whereas high doses of chlorpromazine can activate HSPA5 and JUN in stage 4 bladder cancer. Therefore, low-dose yohimbine and low-dose chlorpromazine could avoid side effects in the treatment of stage 1 and stage 4 bladder cancer cells, respectively.

Ultimately, we designed one specific drug combination for treating stage 1 bladder cancer and another specific drug combination for treating stage 4 bladder cancer with minimal side effects (Table 17.1).

CONCLUSION

In this chapter, we proposed a new method for constructing an IGEN for characterizing cellular mechanisms in bladder carcinogenesis by using system regression modeling and large-scale database mining. We then applied PGP to obtain the core network biomarkers of the IGEN. By comparing the connection changes of the core network biomarkers between normal bladder cells and stage 1 bladder cancer cells and between stage 1 and stage 4 bladder cancer cells, we investigated the progression mechanisms of bladder carcinogenesis. Database mining provided all possible candidates for genetic and miRNA regulations and protein interactions in IGEN. We used AIC and statistical assessment to prune the false-positive regulations and interactions by applying the regression-coupling model to NGS data and methylation profiles. We compared the connection differences in the core network biomarkers between different cellular types to explore bladder carcinogenic mechanisms.

According to the comparison of the connection changes in the core network biomarkers between normal cells and stage 1 cancer cells and between stage 1 and stage 4 cancer cells, we investigated how the genetic and epigenetic regulations, miRNA regulations, and aging-related and smoking-related genes affect the biological functions that lead to bladder carcinogenesis. According to gene expression changes in the core network biomarkers between normal bladder cells and stage 1 bladder cancer cells and between stage 1 and stage 4 bladder cancer cells, we then identified two separate drug combinations for treating stage 1 and 4 bladder cancer cells. Therefore, the proposed IGEN construction method and PGP provide potential network biomarkers for bladder cancer diagnosis and treatment.

REFERENCES

- [1] Schroeder GL, Lorenzo-Gomez M-F, Hautmann SH, et al. A side by side comparison of cytology and biomarkers for bladder cancer detection. *J Urol* 2004;172(3):1123–6.
- [2] Li C-W, Chen B-S. Identifying functional mechanisms of gene and protein regulatory networks in response to a broader range of environmental stresses. *Comp Funct Genom* 2010;2010:20. Article ID 408705.
- [3] Chen BS, Li CW. Measuring information flow in cellular networks by the systems biology method through microarray data. *Front Plant Sci* 2015;6. article 390.
- [4] Chernov AV, Reyes L, Xu Z, et al. Mycoplasma CG- and GATC-specific DNA methyltransferases selectively and efficiently methylate the host genome and alter the epigenetic landscape in human cells. *Epigenetics* 2015;10(4):303–18.
- [5] Bandres E, Agirre X, Bitarte N, et al. Epigenetic regulation of microRNA expression in colorectal cancer. *Int J Cancer* 2009;125(11):2737–43.
- [6] Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 2007;39(4):457–66.
- [7] Valinluck V, Tsai H-H, Rogstad DK, Burdzy A, Bird A, Sowers LC. “Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* 2004;32(14):4100–8.
- [8] Ghosh S, Yates AJ, Frühwald MC, Miecznikowski JC, Plass C, Smiraglia DJ. Tissue specific DNA methylation of CpG islands in normal human adult somatic tissues distinguishes neural from non-neural tissues. *Epigenetics* 2010;5(6):527–38.
- [9] Chen B-S, Li C-W. On the interplay between entropy and robustness of gene regulatory networks. *Entropy* 2010;12(5):1071–101.
- [10] Chen BS, Tsai KW, Li CW. Using nonlinear stochastic evolutionary game strategy to model an evolutionary biological network of organ carcinogenesis under a natural selection scheme. *Evol Bioinf Online* 2015;11:155–78.
- [11] Chen B, Wong S, Li C. On the calculation of system entropy in nonlinear stochastic biological networks. *Entropy* 2015;17(10):6801–33.
- [12] Wang Y-C, Chen B-S. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC Med Genom* 2011;4. article 2.
- [13] Tu C-T, Chen B-S. On the increase in network robustness and decrease in network response ability during the aging process: a systems biology approach via microarray data. *IEEE ACM Trans Comput Biol Bioinf* 2013;10(2):468–80.
- [14] Blagosklonny MV. Prevention of cancer by inhibiting aging. *Canc Biol Ther* 2008;7(10):1520–4.
- [15] Lamb J, Crawford ED, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313(5795):1929–35.

- [16] Fink SEK, Pahernik S, Hallscheidt P, Zeier M. Uro- thelial carcinoma. *Onkologie* 2015;21(8):739–45.
- [17] Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res* 2016;44(D1):D1036–44.
- [18] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;136(2):215–33.
- [19] Bovolenta LA, Acencio ML, Lemke N. HTRIDb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genom* 2012;13. article 405.
- [20] Zheng G, Tu K, Yang Q, et al. ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics* 2008;24(20):2416–7.
- [21] Matys V, Fricke E, Geffers R, et al. TRANSFAC@: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;31(1):374–8.
- [22] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27–30.
- [23] Johansson R. System modeling and identification. Englewood Cliffs (NJ, USA): Prentice Hall; 1993.
- [24] Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 2007;25(1):117–24.
- [25] Coleman TF, Hulbert LA. A direct active set algorithm for large sparse quadratic programs with simple bounds. *Math Program* 1989;45(3):373–406.
- [26] Gill PE, Murray W, Wright MH. Practical optimization. New York (NY, USA): Academic Press; 1981.
- [27] Seber GAF, Lee AJ. Linear regression analysis. 2nd ed. Hoboken (NJ, USA): Wiley- Interscience; 2003.
- [28] Romanenko A, Kakehashi A, Morimura K, et al. Urinary bladder carcinogenesis induced by chronic exposure to persistent low-dose ionizing radiation after Chernobyl accident. *Carcinogenesis* 2009;30(11):1821–31.
- [29] Yang T, Shi RL, hang L, et al. Huachansu suppresses human bladder cancer cell growth through the Fas/Fasl and TNF- alpha/TNFR1 pathway in vitro and in vivo. *J Exp Clin Canc Res* 2015;34. article 21.
- [30] Su C, Liu S, Lee K, et al. Cantharidin induces apoptosis through the calcium/PKC-regulated endoplasmic reticulum stress pathway in human bladder cancer cells. *Am J Chin Med* 2015;43(3):581–600.
- [31] Lee E-J, Lee S-J, Kim S, et al. Interleukin-5 enhances the migration and invasion of bladder cancer cells via ERK1/2- mediated MMP-9/NF- κ B/AP-1 pathway: involvement of the p21WAF1 expression. *Cell Signal* 2013;25(10):2025–38.
- [32] Sethi G, Sung B, Aggarwal BB. TNF: a master switch for inflammation to cancer. *Front Biosci* 2008;13(13):5094–107.
- [33] O'Rourke RW, Metcalf MD, White AE, et al. Depot- specific differences in inflammatory mediators and a role for NK cells and IFN- γ in inflammation in human adipose tissue. *Int J Obes* 2009;33(9):978–90.
- [34] Martinon F. Targeting endoplasmic reticulum signaling pathways in cancer. *Acta Oncol* 2012;51(7):822–30.
- [35] Jang SH, Park JW, Kim HR, Seong JK, Kim HK. ADRM1 gene amplification is a candidate driver for metastatic gastric cancers. *Clin Exp Metastasis* 2014;31(6):727–33.
- [36] Fejzo MS, Anderson L, von Euw EM, et al. Amplification target ADRM1: role as an oncogene and therapeutic target for ovarian cancer. *Int J Mol Sci* 2013;14(2):3094–109.
- [37] Yang X, Miao XY, Wen Y, Hu JX, Dai WD, Yin BL. A possible connection between adhesion regulating molecule 1 overexpression and nuclear factor kappa B activity in hepatocarcinogenesis. *Oncol Rep* 2012;28(1):283–90.
- [38] Chen W, Hu X-T, Shi Q-L, Zhang F-B, He C. Knock-down of the novel proteasome subunit Adrm1 located on the 20q13 amplicon inhibits colorectal cancer cell migration, survival and tumorigenicity. *Oncol Rep* 2009;21(2):531–7.
- [39] Zheng X, Guo Y, Chen Y, et al. Knockdown of adhesion- regulating molecule 1 inhibits proliferation in HL60 cells. *Acta Haematol* 2015;134(2):88–100.

- [40] Yamasaki T, Yoshino H, Enokida H, et al. Novel molecular targets regulated by tumor suppressors microRNA-1 and microRNA-133a in bladder cancer. *Int J Oncol* 2012;40(6):1821–30.
- [41] Aggarwal BB, Kumar A, Bharti AC. Anticancer potential of curcumin: preclinical and clinical studies. *Anticancer Res* 2003;23(1):363–98.
- [42] Xie J, Lei P, Hu Y. Small interfering RNA-induced inhibition of epithelial cell transforming sequence 2 suppresses the proliferation, migration and invasion of osteosarcoma cells. *Exp Ther Med* 2015;9(5):1881–6.

FURTHER READING

- Ferreira AEN, Ponces Freire AMJ, Voit EO. A quantitative model of the generation of N-epsilon-(carboxymethyl) lysine in the Maillard reaction between collagen and glucose. *Biochem J* 2003;376(1):109–21.
- Voit EO, Ferreira AEN. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge (UK): Cambridge University Press; 2000.

EPIGENOME-WIDE ANALYSIS OF DNA METHYLATION IN COLORECTAL CANCER

18

Nurul-Syakima Ab Mutalib, Rashidah Baharuddin, Rahman Jamal

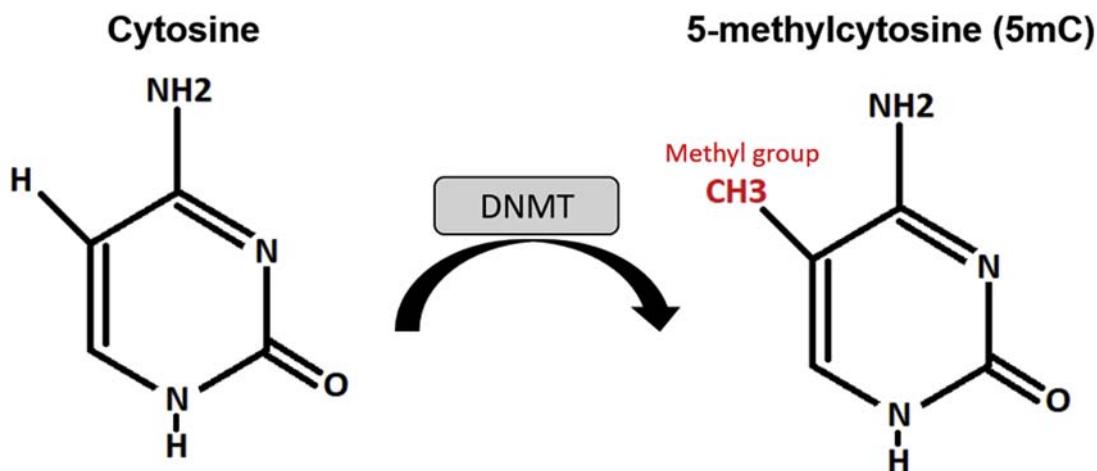
UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

INTRODUCTION

Epigenetics is the study of heritable changes in gene function that is independent from the alterations in the DNA sequence and is crucial for the preservation of tissue-specific gene expression patterns in humans [1]. Perturbation of epigenetic processes can cause alterations of gene function which subsequently could lead to cell transformation, which is the hallmark of cancer [2]. Conventionally seen as a genetic disease, the initiation and progression of cancer is now accepted to involve epigenetic aberrations together with genetic alterations [1]. Nevertheless, epigenetic abnormalities are reversible and this notion has led to the development of epigenetic therapy, which is making promising progress with four epigenetic-based drugs that have already been approved by the United States Food and Drug Administration (US FDA) for cancer treatment [3].

Epigenetic markers can be broadly divided into histone modification and DNA methylation. DNA methylation, one of the well-characterized epigenetic markers, is the process when a methyl group (CH_3) is added to the fifth carbon of cytosine (C) by DNA methyltransferases (DNMTs), producing 5-methylcytosine (5mC) (Fig. 18.1) [4]. Even though most of DNA methylation takes place in CpG dinucleotides (CpGs), it is also found in non-CpG sites (CpA, CpC, and CpT) that distinctively influence gene function and structure [5]. DNA methylation has been implicated in various cellular processes, including transcription factor repression [6], X-chromosome inactivation [7], genomic imprinting [8], modification of chromatin structure [9], and carcinogenesis [10].

The mechanisms of DNA methylation in regulating cellular processes are mainly described on the CpG islands and the gene bodies. CpG (cytosine–phosphate–guanine) islands refer to the ~ 500 base pair stretches of DNA with at least 55% G + C content, often clustered at the 5'-ends of genes (promoter region) and have observed:expected frequency of ≥ 0.6 [11]. About 70% of the CpG loci in the mammalian genome are methylated [12], however, in normal cells CpG islands are protected from DNA methylation via unknown mechanisms [13]. Around half of CpG islands are situated in gene promoters and when these CpGs are methylated, gene expression will be potentially downregulated by changing the chromatin structure and this in turn will inhibit the initiation of transcription processes [14]. On the contrary, gene body methylation does not block the aforementioned processes and might even trigger elongation of transcription, which correlates with active transcription and therefore increases

**FIGURE 18.1**

DNA methylation takes place at cytosine bases when a methyl group (CH_3) is added at the fifth carbon on the cytosine ring by the DNA methyltransferase (DNMT).

gene expression [15]. In addition, there is also evidence showing that gene body methylation affects gene splicing [16].

Gene promoters are located immediately adjacent to a gene, demonstrate directionality, and are inclined to have a higher degree of redundant activity across various cell types [17]. Outside of promoter regions, distal enhancers play a key role in the cell type-specific regulation of gene expression [18]. Enhancers are 10 times more abundant than promoters and genes in humans [19] and their methylation state could subsequently result in different gene expression patterns, permitting the formation of hundreds of different cell functions and identities [20]. It has been shown that the disruption of enhancer activity through epigenetic alteration can influence cell type-specific roles, causing a wide range of abnormalities including cancer [20]. In cancers, these modifications can stimulate an “identity crisis” whereby enhancers associated with oncogenes are triggered whereas those associated with cell fate are deactivated. In addition, chromatin also plays key functions in transcriptional regulation and maintaining genomic stability [21]. Investigation of the epigenetic mechanisms controlling the activities of all of these components can uncover therapeutic opportunities in managing cancers, including CRC.

The important DNA methylation findings aforementioned would not have been discovered without the development of profiling approaches, both experimental and computational. The advancement of microarray and next-generation sequencing (NGS) technologies has significantly improved DNA methylation profiling, providing an unparalleled comprehensive and unbiased view of the DNA methylation landscape in the human cancer genome. However, the unraveling of DNA methylation landscape would not have been possible without the parallel development of epigenetics computational tools. This chapter will briefly summarize the available epigenome-wide approach for DNA methylation investigation and subsequently focus on computational epigenetics and epigenomics analysis in CRC. Finally, current updates on epigenetic biomarkers in CRC will also be covered.

APPROACHES TO ANALYZE DNA METHYLATION IN COLORECTAL CANCER

Generally, techniques for epigenome-wide DNA methylation comprise bisulfite conversion of unmethylated cytosine to uracil, immunoprecipitation with antibodies targeting the methylated DNA, digestion of restriction sites containing the CpG sites by methylation-sensitive restriction enzymes (MREs), and then hybridization to a microarray chip or the use of NGS [22]. Microarray-based platforms are commonly used to profile DNA methylation due to their relatively cheaper cost and simpler analysis pipeline compared to sequencing-based platforms. The most recent approach that combines bisulfite conversion with hybridization-based arrays is the use of the Infinium MethylationEPIC BeadChip by Illumina [23,24]. This microarray beadchip contains more than 850,000 methylation sites, which include over 90% of content from the previous version Human-Methylation450K BeadChip, CpG sites outside of CpG islands, non-CpG methylated sites identified in human stem cells (CHH sites), differentially methylated sites identified in tumor versus normal, FANTOM5 enhancers, ENCODE open chromatin and enhancers, DNase hypersensitive sites, and microRNA promoter regions [23]. It has been perhaps the most comprehensive methylation microarray platform to date.

Other available alternatives for microarray-based profiling are Human CpG Microarrays [25] and Human DNA Methylation Microarray [26] from Agilent Technologies. The Human CpG Microarrays cover 27,800 CpG islands spanning 21 megabases (Mb), targeted by 237,220 probes in or within 95 base pairs (bp) of CpG islands [25]. The Human DNA Methylation Microarray contains a slightly higher content, comprising promoter regions and CpG islands. It contains 27,627 expanded CpG islands and 5081 unmethylated regions targeted by 237,227 probes [26]. In contrast to the MethylationEPIC BeadChip which utilizes bisulfite conversion, Human CpG Microarray and Human DNA Methylation Microarray are based on methylated DNA immunoprecipitation (MeDIP). Although the described microarrays offer single base pair resolution, the knowledge obtained will be confined to the previously identified regions; therefore novel discovery is not possible with this approach.

The most comprehensive method for methylation profiling thus far is the bisulfite conversion followed by whole-genome bisulfite sequencing (WGBS), which includes MethylC-seq [27] and BS-seq [28,29]. These approaches quantify epigenome-wide single-cytosine methylation levels and directly evaluate the ratio of methylated to unmethylated bases. However, to obtain a complete DNA methylome, 90 gigabases (Gb) of data are required which will proportionally increase the cost and even with this 50% of the reads may even fail to cover the CpG sites [30]. In addition, due to the bisulfite conversion which changes the cytosine residues to uracil and leaving 5mC unaffected, the resulting genomes will have reduced GC content and, therefore, possess lower base complexity which will subsequently affect sequencing quality [30]. A relatively cheaper approach would be the reduced representation bisulfite sequencing (RRBS) which permits measurement of single-CpG methylation in CpG-dense regions only [31,32]. Around 3 Gb of sequencing is needed to achieve the similar level of sequencing depth for most regions of interest [30], as compared to WGBS which requires 90 Gb. In RRBS, genomic DNA will be first digested with the methylation-insensitive restriction enzyme *MspI* followed by sequencing. However, RRBS only enables interrogation of 8%–15% of the CpGs in the human genome.

MRE-seq or methylation-sensitive restriction enzymes sequencing [33], as the name implies, includes parallel restriction enzyme digestions with several MREs including *HpaII*, *Hin6I*, and *AciI*.

These MREs only recognize the restriction sites with unmethylated CpGs, whereas the methylated CpGs are predicted by the lack of sequencing reads at recognition sequences. By incorporating multiple restriction recognition sites, MRE-seq is able to interrogate up to 30% of the genome while only requiring approximately 3 Gb of sequencing data [34]. On the other hand, MeDIP-seq (methylated DNA immunoprecipitation sequencing) detects methylated CpGs and involves immunoprecipitation of 5 mC using antibody followed by sequencing [33,35]. Combination of MRE-seq and MeDIP-seq has been proposed as a cost-effective alternative for WGBS [36]. Another approach, MBD-seq (methyl-CpG binding domain sequencing), uses the MBD2 protein methyl-CpG binding domain for enrichment of methylated double-stranded DNA fragments [37]. The comparisons of several sequencing-based approaches for investigating DNA methylation are comprehensively reviewed by others [34,38]. All of the DNA methylation approaches described in this chapter are summarized in [Table 18.1](#).

EPIGENOME-WIDE ANALYSIS OF DNA METHYLATION IN COLORECTAL CANCER

Epigenome-wide analysis offers possibilities for the discovery of novel regulatory mechanisms that are susceptible to environmental and lifestyle modification, thus influencing predisposition to diseases, including CRC. Epigenome-wide screening approaches as discussed previously include an increasing number of not only CpG sites, but also other distal regulatory elements. [Table 18.2](#) illustrates the selected epigenome-wide analysis of DNA methylation in this cancer.

DNA METHYLATION BIOMARKERS IN COLORECTAL CANCER

CRC is curable in most of the cases if it is detected at an early stage. The detection of this cancer can be achieved through noninvasive and invasive procedures. The noninvasive procedures for CRC include the fecal occult blood test (FOBT). Despite the improvement of CRC detection rates using this procedure, FOBT has a low sensitivity [58]. Hence, invasive methods such as colonoscopy are more effective and have become gold standard in CRC detection. However, this method requires extensive preparation of bowel and some complications may occur during the procedure such as bleeding and bowel injury [58]. Therefore, due to the limited compliance in colonoscopy and FOBT approaches, much research has been devoted to investigate the methylation biomarkers that control the transformation of normal colonic epithelial cells to adenocarcinomas [59–61]. Potential biomarkers can be identified through various biospecimens such as blood, stool, and tissue-based [62]. It is necessary to provide promising biomarkers with better sensitivity and specificity in order to improve diagnosis, prognostication, and surveillance as well as predicting response towards the therapy.

BLOOD-BASED DNA METHYLATION BIOMARKERS

Several blood screening biomarker tests have been shown to facilitate the noninvasive detection at early stages of tumor development and perhaps may be useful for the diagnosis of CRC. Various genes including protein phosphatase 1 regulatory subunit 3C (*PPP1R3C*), EF-hand domain family member D1 (*EFHD1*) [63], syndecan-2 (*SDC2*) [64], neurogenin 1 (*NEUROG1*) [65], nerve growth factor

Table 18.1 Epigenome-Wide Approaches to Study DNA methylation

Methods	Feature	References
Illumina Infinium MethylationEPIC BeadChip	<p>Covers more than 850,000 methylation sites which consist of:</p> <ul style="list-style-type: none"> >90% of content contained on the Illumina HumanMethylation450K BeadChip CpG sites outside of CpG islands Non-CpG methylated sites identified in human stem cells (CHH sites) Differentially methylated sites identified in tumor versus normal FANTOM5 enhancers ENCODE open chromatin and enhancers DNase hypersensitive sites MicroRNA promoter regions 	[23,24]
Agilent Human CpG Microarrays	<ul style="list-style-type: none"> Contains 27,800 CpG islands spanning 21 Mb regions with 237,220 probes (\pm 95 bp of CpG islands) 	[25]
Agilent Human DNA Methylation Microarray	<ul style="list-style-type: none"> Enables interrogation of 27,627 expanded CpG islands (including promoters) and 5081 UMR regions with 237,227 probes 	[26]
MethylC-seq	<ul style="list-style-type: none"> Massively parallel sequencing of bisulfite-converted genomic DNA Provides the highest CpG coverage at 95% 	[27] [38]
Bisulfite sequencing (BS-seq)	<ul style="list-style-type: none"> Uses unconventional sequencing adapter strategy to produce four possible distinct DNA sequences generated through copying a bisulfite-converted duplex DNA strand The bioinformatics strategy is different and more complex than MethylC-seq Provides the highest CpG coverage at 95% 	[28,29] [38]
Reduced representation bisulfite sequencing (RRBS)	<ul style="list-style-type: none"> Genomic DNA is digested with the methylation-insensitive restriction enzyme MspI followed by massively parallel sequencing Only enables interrogation of small portion of the CpGs in human genome (8%–15%) 	[31,32] [34]
MRE-seq (methylation-sensitive restriction enzymes sequencing)	<ul style="list-style-type: none"> Interrogation of unmethylated CpG sites at single-CpG site resolution via sequencing of the size-selected fragments from genomic DNA digested with the methylation-sensitive restriction enzymes (MREs) HpaII, Hin6I, and AciI CpG coverage at 30% 	[33] [34]
MeDIP-seq (methylated DNA immunoprecipitation sequencing)	<ul style="list-style-type: none"> Antibody-based immunoprecipitation of 5mC and massively parallel sequencing to detect methylated CpGs CpG coverage at 67% 	[33,35] [38]
MBD-seq (methyl-CpG binding domain sequencing)	<ul style="list-style-type: none"> The methylated DNA is enriched by methyl-CpG binding domain of MBD2 protein followed by massively parallel sequencing CpG coverage at 61% 	[37] [38]

The \pm sign indicates 95 basepair upstream and downstream of the CpG islands.

Table 18.2 Selected Epigenome-Wide Analysis of DNA methylation in CRC and the Related Analysis Tool

Techniques	Main Findings	References	Analysis Tool and references
WGBS	Discovery of aberrant superenhancer methylation in CRC	[39]	Bismark V0.7.4 software [40]
RRBS	Identified hypermethylated CpG sites in <i>L3MBTL1</i> , <i>NKX6-2</i> , <i>PREX1</i> , <i>TRAF7</i> , <i>PRDM14</i> , and <i>NEFM</i> genes which were enriched in Wnt/β-catenin, PI3k/AKT, VEGF, and JAK/STAT3 signaling pathways	[41]	Zymo Research proprietary analysis pipeline
RRBS	<i>ATXN7L1</i> , <i>BMP3</i> , <i>EID3</i> , <i>GAS7</i> , <i>GPR75</i> , and <i>TNFAIP2</i> genes were significantly hypermethylated in CRC versus normal tissues	[42]	Zymo Research proprietary analysis pipeline
BS-seq	Regions of focal hypermethylation in the tumors were located primarily at CpG islands and were concentrated within regions of long-range (>100 kilobases) hypomethylation. These hypomethylated domains covered nearly half of the genome and coincided with late replication and attachment to the nuclear lamina in human cell lines	[43]	Customized pipeline
Microarray (Illumina Infinium array 450K)	Aberrant methylation of <i>CCNE1</i> , <i>CCNDBP1</i> , <i>PON3</i> , <i>DDX43</i> , and <i>CHL1</i> genes might be associated with the CRC recurrence	[44]	ChAMP [45,46]
Microarray (Illumina Infinium array 450K)	Hypomethylation of <i>BPIL3</i> and <i>HBBP1</i> genes and hypermethylation of <i>TIFPI2</i> , <i>ADHFE1</i> , <i>FLII</i> , and <i>TLX1</i> genes in rectal cancer	[47]	Customized pipeline
Microarray (Illumina Infinium array 27K)	The cell adhesion molecule pathway is the most enriched and includes <i>JAM2</i> , <i>NCAM1</i> , <i>ITGA8</i> , and <i>CNTN1</i> genes	[48]	Partek Genomic Suite 6.6
Microarray (Illumina Infinium array 450K)	Identified three classes of CRC and two classes of adenomas based on DNA methylation patterns	[49]	minfi [50]
Microarray (Illumina Infinium array 450K)	Methylation in <i>SND1</i> , <i>ADHFE1</i> , <i>OPLAH</i> , <i>TLX2</i> , <i>Clorf70</i> , <i>ZFP64</i> , <i>NR5A2</i> , and <i>COL4A</i> genes differentiates CRC tissues from adjacent normal colonic mucosa	[51]	IMA [52] and Partek Genomic Suite 6.6
Microarray (Illumina Infinium array 27K)	Identified four DNA methylation-based subgroups of CRC	[53]	RPMM [54,55]
Microarray (Illumina Infinium array 450K)	Hypermethylation was more common in the right colon	[56]	RPMM [54,55]
Microarray (Illumina Infinium array 27K)	Hypermethylation of <i>ADHFE1</i> , <i>BOLL</i> , <i>SLC6A15</i> , <i>ADAMTS5</i> , <i>TFPI2</i> , <i>EYA4</i> , <i>NPY</i> ,	[56]	Customized pipeline

Table 18.2 Selected Epigenome-Wide Analysis of DNA methylation in CRC and the Related Analysis Tool—cont'd

Techniques	Main Findings	References	Analysis Tool and references
Microarray (comprehensive high-throughput array-based relative methylation; CHARM)	<i>TWIST1</i> , <i>LAMA1</i> , and <i>GAS7</i> genes and hypomethylation of <i>MAEL</i> and <i>SFT2D3</i> genes in CRC tissues compared to the normal mucosa Most methylation alterations in CRC occur in CpG island shores	[57]	Customized pipeline

receptor (*NGFR*), and tomoregulin-2 (*TMEFF2*) [66] have emerged as potential blood-based methylation markers for the early detection of CRC with sensitivities ranging from 51% to 96%.

Interestingly, a blood-based assay that detects methylated septin 9 gene (*SEPT9*) has been commercialized by Epigenomics AG (Berlin, Germany) for use as a diagnostic test. This assay is currently being marketed in several countries and it provides overall sensitivity of 95% and specificity of 85% for the detection of methylated *SEPT9* in CRC patients. The *SEPT9* gene belongs to a class of GTPase and methylation at the promoter region of this gene is known to be associated with CRC, thus warranting its use for blood-based CRC screening [67]. In addition, Toth et al. [68] have investigated the sensitivity and specificity of this gene in blood as a potential biomarker for CRC. They validated on 34 CRC patients and 24 healthy controls. Out of 34 CRC patients, 33 had positive methylation and no methylation was found in healthy controls [68]. A more recent systematic review and meta-analysis assessed the accuracy of methylated *SEPT9* in detecting CRC, and the pooled sensitivity, specificity, and AUC of methylated *SEPT9* were remarkable, of 0.71, 0.92, and 0.88, respectively [69]. Methylated *SEPT9* could be meaningful for clinical use.

STOOL-BASED DNA METHYLATION BIOMARKERS

A stool test is painless and quite acceptable in general population for CRC screening. This test has been recommended for screening of average risk asymptomatic individuals aged 50 years and above by American College of Gastroenterology (ACG). Numerous hypermethylated genes have been detected in the fecal samples. For instance, fibrillin-1 gene (*FBN-1*) was found to have higher methylation levels in CRC stool samples. In 2013, Guo et al. performed methylation analysis of this gene across 75 CRC patients and 30 healthy controls [70]. Fifty four patients exhibited hypermethylated *FBN-1* in their stools. Their findings revealed that this test has 72% sensitivity and 93.3% specificity for CRC detection in stools.

Another potential fecal methylation marker is secreted frizzled-related protein gene 2 (*SFRP2*). Inactivation of this gene via promoter methylation causes activation of Wnt signaling pathway [71], which is well known in CRC. Previous studies indicated that increased methylation of *SFRP2* is a potential biomarker for early detection of CRC in stools [72,73]. Huang and colleagues [72] reported that *SFRP2* methylation is detectable in 94.2% of CRC patients, which enables fecal DNA testing as a

noninvasive screening tool. This finding is supported by Lu et al. [74] and Yang et al. [75] where they found that this gene is hypermethylated in most of the CRC cases. More examples of potential biomarkers discovered in the stools are vimentin (*VIM*), tissue factor pathway inhibitor 2 (*TFPI2*), NDRG family member 4 protein (*NDRG4*), and bone morphogenetic protein 3 (*BMP3*) [76]. These genes have the potential as biomarkers for early screening of CRC.

PROGNOSTIC BIOMARKERS

DNA methylation profiles could be useful as biomarkers in predicting patient response towards therapy and risk of tumor recurrence, which are important for patient outcome. For example, promoter methylation of mutL homolog 1 (*MLH1*), transcription factor AP-2 epsilon (*TFAP2E*), and BCL2 interacting protein 3 (*BNIP3*) was associated with chemoresistance towards 5-fluorouracil (5-FU) through inhibition of the apoptosis pathway [77–79]. Adding to that, T-box 5 (*TBX5*), disheveled binding antagonist of beta catenin 2 (*DACT2*), Ras association domain family member 1 (*RASSF1A*), and adenomatous polyposis coli 1A (*APC1A*) genes were discovered as potential biomarkers in CRC. Methylation of these genes in CRC correlated with a worse prognosis [80–82]. Therefore, these genes may serve as epigenetic biomarkers to predict the outcome of CRC patients.

Methylation analysis has successfully identified a few predictive markers that act as indicators of disease recurrence in CRC including Wnt family member 5A (*WNT5A*), branched chain amino acid transaminase 1 (*BCAT1*), and IKAROS family zinc finger 1 (*IKZF1*) [83,84]. In these two studies, the authors observed that these genes could possibly be involved in tumor growth and invasiveness which lead to cancer recurrence. However, the underlying mechanisms of recurrence are being intensively studied now. Examples of the promising biomarkers are shown in Table 18.3.

COMPUTATIONAL TOOLS FOR DNA METHYLATION

High-throughput DNA methylation data require sophisticated computational tools to facilitate the analyses and to assist in data interpretation in order for the researchers to comprehend the biological meaning. Depending on the platforms used and users' skills in computational analysis, the selection of epigenetics computational tools will be different. Therefore it is important to understand the principles behind the tools and their applications in answering the research questions posed by the respective researchers. There are many epigenome-wide DNA methylation analysis tools available, however, this chapter will only focus on the selected tools which are widely used.

The first review on computational epigenetics tools was provided by Bock and Lengauer almost a decade ago [98] and briefly focused on the ground-breaking computational studies that have revolutionized epigenetic research. Two years later, Lim et al. [99] reviewed the epigenetics computational strategies available in 2010 and covered several databases and bioinformatics tools. Robinson and Pelizzola [100] provided a short review on computational analyses which addressed the vital experimental issues closely linked with the epigenetics profiling approaches (such as bias and overrepresentation issues), the available computational tools to overcome those issues, and tools to enable the integration of multiple epigenetic data. Since the explosion of high-throughput methylation data, many computational tools have been developed thus far. However, most of the tools are poorly organized and are scattered, making the search for the right tools a time-consuming effort. For this

Table 18.3 DNA methylation Biomarkers in CRC

Gene(s)	Methylation Status	Potential Application	References
Axin 2 (<i>AXIN2</i>) and dickkopf WNT signaling pathway inhibitor 1 (<i>DKK1</i>)	Hypermethylation	Predictor of recurrence in stage II CRC	[85]
Long interspersed element-1 (<i>LINE-1</i>)	Hypomethylation	<i>LINE-1</i> hypomethylation is associated with worse prognosis of patient with CRC	[86]
O-6-methylguanine-DNA-methyltransferase (<i>MGMT</i>)	Hypermethylation	Epigenetic silencing of <i>MGMT</i> increases response of CRC cells to alkylating agents such as temozolomide (reduces chemoresistance)	[87]
Suppressor of cytokine signaling 1 (<i>SOCS1</i>)	Hypermethylation	Promoter methylation of <i>SOCS1</i> is associated with lymph node metastasis, TNM stage, and poor overall survival	[88]
Calcium voltage-gated channel subunit alpha1 G (<i>CACNA1G</i>)	Hypermethylation	Hypermethylation of <i>CACNA1G</i> acts as a potential biomarker for poor survival in CRC	[89]
Integrin subunit alpha 4 (<i>ITGA4</i>) and tissue factor pathway inhibitor 2 (<i>TFPI2</i>)	Hypermethylation	Possible risk biomarkers for the development of CRC	[90]
B-cell CLL/lymphoma 6B (<i>BCL6B</i>)	Hypermethylation	Epigenetic silencing of <i>BCL6B</i> induced CRC proliferation and metastasis	[91]
Insulin-like growth factor binding protein 3 (<i>IGFBP3</i>)	Hypermethylation	Stage II and III CRC patients with <i>IGFBP3</i> hypermethylation do not benefit from adjuvant 5FU-based chemotherapy	[92]
B-cell lymphoma 2 (<i>BCL2</i>), apoptotic peptidase activating factor 1 (<i>APAF1</i>), and tumor protein 53 (<i>p53</i>)	Hypermethylation	Methylation of apoptosis pathway markers leads to poor prognosis of CRC	[93]
Transcription factor 3 (<i>TCF3</i>)	Hypomethylation	Predictor of recurrence in stage II and III CRC	[94]
Chromosome 9 open reading frame 50 (<i>C9orf50</i>) and thrombomodulin (<i>THBD</i>)	Hypermethylation	Potential biomarkers for noninvasive early CRC detection	[95]
ALX homeobox 4 (<i>ALX4</i>)	Hypermethylation	Promoter methylation for detection of precancerous lesions	[96]
Ubiquitin C-terminal hydrolase L1 (<i>UCHL1</i> ; also known as <i>PGP9.5</i>)	Hypomethylation	<i>PGP9.5</i> hypomethylation was observed in patient with early-stage CRC	[97]

reason, a web-based curation called OMICtools was created in 2013 and it provides manually curated list of tools for epigenome-wide analysis [101].

DNA methylation profiling using microarray platform is preferable to many biologists owing to its relative simplicity. Because of the wide usage, the computational tools for this platform are widely available. One of the most frequently used tools is Chip Analysis Methylation Pipeline (ChAMP) [45,46], which was originally developed for Illumina Infinium array 450K. Based on R, ChAMP uses the IDAT files as input and utilizes the data import, quality control, and normalization options offered by minfi [50]. The data are filtered for probe detection $P > .01$ by default. When the IDAT files are not available, users are offered the options to upload a matrix of M, beta (β), or raw intensity values, thus offering flexibility. Depending on the research objectives, users can opt to filter out individual probes or probe sets and single nucleotide polymorphisms (SNPs) based on a user-selected minor allele frequency (MAF). These filtering features will prevent biases due to genetic variation in downstream statistical analyses for detecting the differentially methylated CpGs. After preprocessing, the later steps include normalization, calling the differentially methylated region (DMR), and detection of copy number aberrations (CNAs) [45,46]. The normalization step included in ChAMP is of particular importance because Illumina Infinium array 450K is based on two different assays, that is, Infinium I and Infinium II [24,102]. The available options for normalization of the two assays are Beta-Mixture Quantile Normalization (BMQ) [103], Peak Based Correction (PBC) [104], Subset-Quantile Normalization (SQN) [105], and Subset-Quantile Within Array Normalization (SWAN) [106]. In addition to basic DMR identification analysis, ChAMP enables batch effect correction using ComBat [107] and CNA detection [108]. Recently, an updated version of ChAMP has been released [109] to accommodate the Infinium HumanMethylationEPIC BeadChip which has double the array contents compared to the 450K array [110]. This updated version adds more features, including identification of differentially methylated genomic blocks (DMB) using DMRcate [111] and Bumphunter [112], gene set enrichment analysis (GSEA) [113], cell type heterogeneity correction using RefbaseEWAS [114], and web-based graphical interfaces by integrating Shiny and Plotly for enhanced result visualization [109].

The high-throughput sequencing methods to study DNA methylation comprise direct sequencing of bisulfite-treated DNA (BS-seq) or enrichment-based methods such as MeDIP-seq or MBD-seq. In contrast to methylation microarray, analysis of BS-seq data represents higher level of challenge to both the non-bioinformaticians and bioinformaticians. The most frequently used computational tool for BS-seq data analysis is Bismark [40], a package for mapping and determination of methylation state of BS-seq reads. Mapping of bisulfite-treated sequences to a reference genome is known to pose a substantial challenge due to low complexity of the DNA code after bisulfite conversion, libraries that consist of four DNA strands to be analyzed, and the fact that each read can hypothetically exist in all probable methylation states. Despite the number of short read aligners which are available, none are specifically designed to perform bisulfite mapping [40]. Therefore, Bismark is thought to be the most reliable computational tool to analyze BS-seq data. Bismark will search for a unique alignment by performing four alignment processes concurrently. The reads will be transformed into a G-to-A and C-to-T version, followed by parallel alignment of each of them to the preconverted version of the reference genome using short read aligner Bowtie [40], enabling Bismark to differentiate the strand origin of the reads. This alignment approach will also precisely handle incomplete methylation in an unbiased manner because the residual cytosines in the reads are first converted into a complete bisulfite-converted form in silico prior to the alignment. Additionally, Bismark is written in Perl

language and executed from the command line, and therefore the users need background knowledge in bioinformatics and programming.

A more recently developed tool P3BSseq (Parallel Processing Pipeline software for automatic analysis of Bisulfite Sequencing) enables analysis of both WGBS and RRBS [115]. As reviewed by Krueger et al. in 2012, sequencing-based DNA methylation data processing is a complicated process and requires multistep computational analysis: quality control of the raw reads, removal of low quality bases and adapters, mapping of trimmed reads to the reference genome, deduplication of aligned reads, and, lastly, methylation calling [116]. Each step requires different tools, and they are scattered and dependent on the operating system; the input/output formats are incompatible between steps, and need specific but poorly documented parameters [115]. Certain tools are only available for single-end reads or only for directional libraries [115]. P3BSseq was developed to tackle these issues. Operating under Linux/Unix, P3BSseq performs trimming, alignment, annotation, bisulfite conversion quality controls, BED methylome generation and produces NIH-compliance report files [115]. The list of selected computational tools for DNA methylation study is provided in Table 18.4.

WORKFLOW FOR DNA METHYLATION ANALYSIS IN CRC

Global DNA methylation analysis can be overwhelming for beginners because the available tools are scattered; different packages need to be combined for different purposes. A review by Morris and Beck [55] introduced the most prominent pipelines intended for the new array-based DNA methylation researchers. However, it has been 3 years since the review was published, therefore this chapter will propose a more comprehensive and updated workflow. DNA methylation analysis starts with raw data upload, and usually it is in IDAT format which is acceptable for the majority of analysis tools. However, a flexible tool that can accept the raw intensity, M values, or beta values in text format is highly preferable. Following data upload are several quality control steps to remove bias, filtering of probes with low detection *P* values, batch and background correction, correction of heterogeneity in cell population, and normalization. Following quality control, users have the option to filter SNPs according to predetermined MAF.

Subsequently, the DMRs will be determined and statistical analyses shall be performed to determine whether the differences in methylation level are significant or not. Usually the final output from this step will only be a text file with probe or gene identifier with *P* values. Users can use the simple function in Microsoft Excel to calculate delta beta values, which normally are the values included in publications. In addition, the output from a typical DNA methylation analysis pipeline might not include “proper” graphical visualizations. Even if included, the figures are not customizable and not in a “publication-ready” format. Therefore, the scatter plots or pie chart can be generated using the Microsoft Excel or GraphPad Prism (GraphPad Software, Inc., San Diego, CA). The text file containing the methylation loci and beta value can also be uploaded into a program that permits graphical visualization of the methylation data. The web-based tool Morpheus (Broad Institute) can be used to generate heat maps and also perform hierarchical or k-means clustering. A typical DNA methylation analysis workflow is illustrated in Fig. 18.2. To date, ChAMP is the most comprehensive DNA methylation analysis tool available and is recently updated in December 2017 [109].

On the other hand, epigenome-wide sequencing-based DNA methylation analysis is relatively in infancy and the tools are still being developed. Moreover, only a handful of sequencing-based DNA

Table 18.4 Data Analysis Tools for DNA methylation Study

Tools	Feature	Microarray/sequencing	Availability	References
P3BSseq	A processing pipeline optimized for directional and nondirectional libraries, single-end and paired-end reads of WGBS and RRBS	WGBS RRBS	http://sourceforge.net/p/p3bsseq/wiki/Home/	[115]
epiGbs	A reference genome-free RRBS method that enables cost-effective analysis of DNA methylation and genetic variation in hundreds of samples	RRBS	https://github.com/thomasvangurp/epiGBS	[117]
metilene	A program to identify DMRs within whole-genome and targeted data. Able to analyze sample pairs without replicates and can estimate missing methylation data	WGBS Targeted BS-seq	http://www.bioinf.uni-leipzig.de/Software/metilene/	[118]
DMRcate	A Bioconductor (R) package for de novo identification and extraction of DMRs from the human genome using WGBS and Illumina Infinium Array (450K and EPIC) data. It permits filtering of probes possibly affected by SNPs and cross-hybridization	Microarray (Illumina Infinium array 450K and EPIC) WGBS	http://bioconductor.org/packages/release/bioc/html/DMRcate.html	[111]
M3D	A Bioconductor (R) package that uses a kernel method to identify DMRs	RRBS	https://bioconductor.org/packages/release/bioc/html/M3D.html	[119]
ChAMP	A Bioconductor (R) package that offers quality control metrics, various normalization methods, identification of DMRs, and copy number alterations	Microarray (Illumina Infinium array 450K and EPIC)	https://bioconductor.org/packages/release/bioc/html/ChAMP.html	[45,46]
SMAP	A streamlined methylation analysis pipeline for BS-seq	RRBS	https://github.com/gaosj lucky/SMAPdigger	[120]
RnBeads	A flexible tool for comprehensive analysis of DNA methylation data from any platforms	WGBS RRBS Microarray (Illumina Infinium array 450K and EPIC) Any protocols producing single base pair resolution DNA methylation information	http://rnbeads.mpi-inf.mpg.de/	[121]
BEAT	A Bioconductor (R) package that provides quantitative high-resolution analysis of DNA methylation patterns from BS-seq data, including the detection of regional epimutation events. Also supports analysis of single-cell BS-seq data	BS-seq	www.bioconductor.org/packages/devel/bioc/html/BEAT.html	[122]

DMAP	A C-based tool for RRBS and WGBS data, which includes a suite of statistical tools and a different investigating approach for analyzing DNA methylation data	RRBS WGBS	http://www.otago.ac.nz/biochemistry/research/otago652955.html	[123]	
BSmooth	An analysis pipeline which includes alignment, quality control, and is able to handle low-coverage experimental designs	WGBS	http://rafalab.jhsph.edu/bsmooth	[124]	
Bismark	A tool for fast analysis of BS-seq data which performs read mapping and methylation calling in a single step. The resulting output discriminates between cytosines in CpG, CHG, and CHH context	WGBS	www.bioinformatics.bbsrc.ac.uk/projects/bismark/	[40]	
Quantitative differentially methylated regions (QDMRs)	A tool to identify DMRs from genome-wide methylation profiles by adapting Shannon entropy	Microarray RRBS	http://bioinfo.hrbmu.edu.cn/qdmr/	[125]	
MethPipe	A computational pipeline to detect DMRs, allele-specific methylation, and partially methylated domains	WGBS RRBS	http://smithlabresearch.org/software/methpipe/	[126]	
methylPipe	A Bioconductor (R) package for the analysis of CpG and non-CpG methylation from WGBS data that also enables integration with other epigenomic data sets	WGBS	https://bioconductor.org/packages/release/bioc/html/methylPipe.html	[127]	
minfi	A Bioconductor (R) package that provides comprehensive analysis and takes cellular heterogeneity into account	Microarray (Illumina Infinium array 450K and EPIC) BS-seq RRBS	http://bioconductor.org/packages/release/bioc/html/minfi.html http://code.google.com/p/moabs/	[50]	
Model-based analysis of bisulfite sequencing (MOABS)	A C-based tool for aligning WGBS data and detecting DMRs. Need replicates	Microarray (Illumina Infinium array 450K)	http://www.rforge.net/IMA/	[128]	
Illumina Methylation Analyzer (IMA)	A package for automated analysis pipeline for determining region-level and site-level methylation changes	Microarray (Illumina Infinium array 450K)	https://cran.r-project.org/web/packages/RPMM/	[52]	
Recursively partitioned mixture model (RPMM)	A model-based unsupervised clustering method developed for DNA methylation data in beta distribution	Microarray (Illumina Infinium array 27K and 450K)		[54,55]	

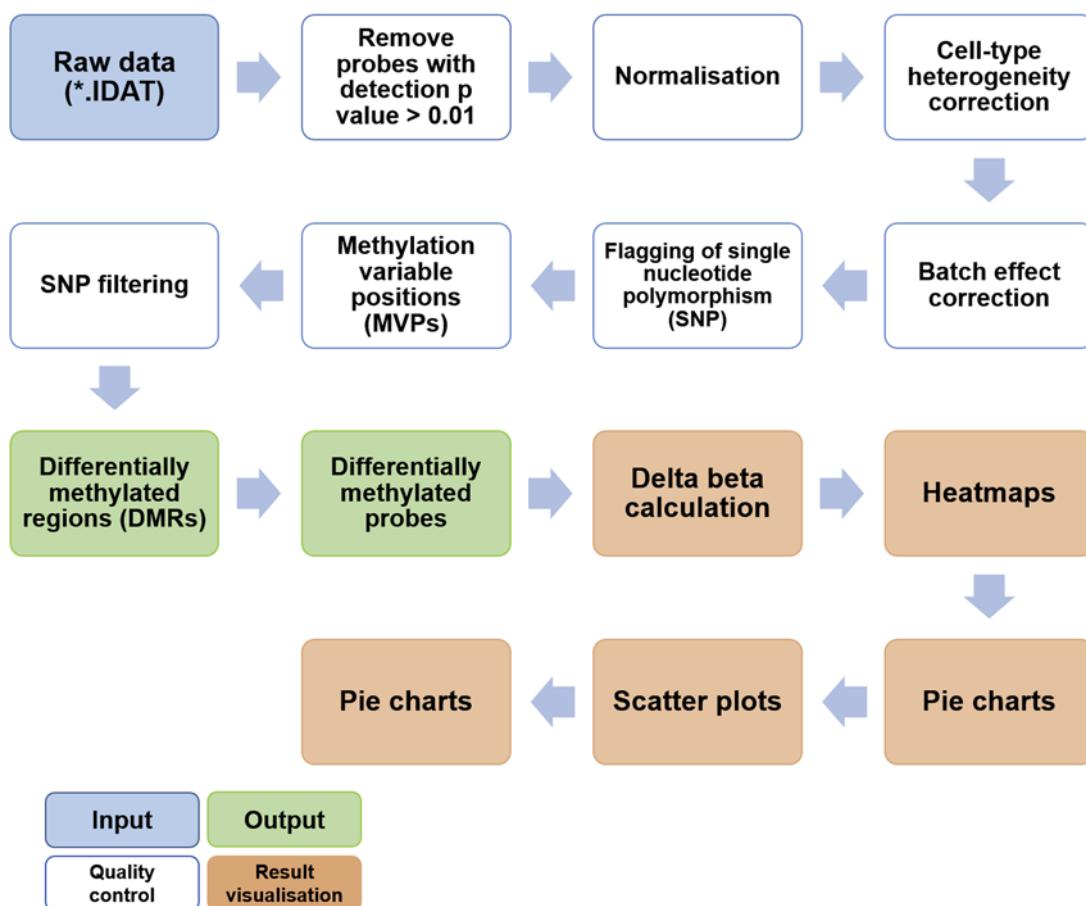
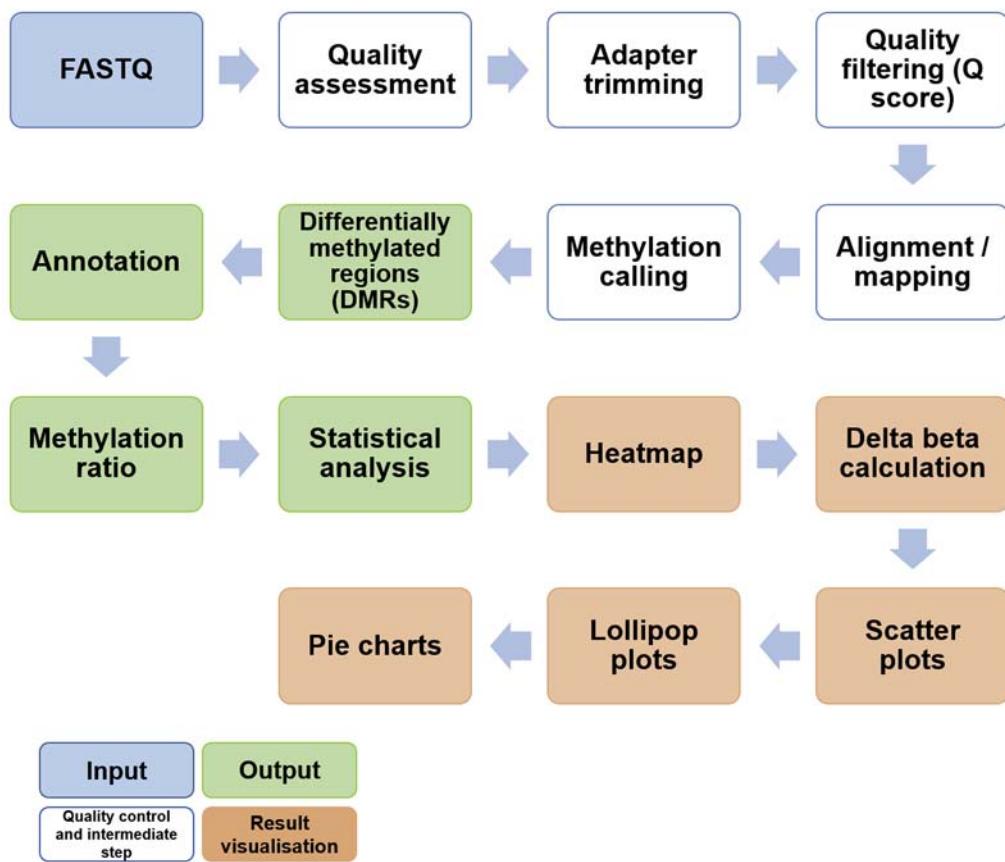


FIGURE 18.2

Typical workflow for array-based DNA methylation analysis.

methylation analyses performed in CRC to date and majority of the researchers used in-house or proprietary pipelines for the complete analysis (refer to [Table 18.2](#)). This is still a big challenge which needs to be overcome by both the bioinformaticians and biologists alike. SMAP is one of the comprehensive pipelines for analyzing sequence-based data [120], but it has never been used in CRC DNA methylation analysis. Besides, it operates in a UNIX/Linux shell which is the main obstacle for most conventional biologists. Nevertheless, an example of analysis workflow for epigenome-wide sequencing-based DNA methylation is illustrated in [Fig. 18.3](#). With the output from the analysis, users can calculate the delta beta values using Microsoft Excel, generate scatter plots or pie chart using the Microsoft Excel or GraphPad Prism, and create heat maps as well as hierarchical clustering using Morpheus. In addition, lollipop plots can also be generated using the web-based Methylation plotter [129] or R-based MethVisual [130].

**FIGURE 18.3**

Typical workflow for epigenome-wide sequencing-based DNA methylation analysis.

CONCLUSION

The advancement in techniques for interrogating the DNA methylation signature has dramatically increased the interest in epigenetics research. Improvement of existing approaches and development of novel methodologies have triggered many active areas of research, especially when the needs for single-cell resolution methylation analysis are emerging. Moving forward, epigenome-wide single-cell DNA methylation analysis will be the new frontier, especially in cancer research in order to understand tumor heterogeneity and address the current gaps in knowledge. In addition, the

computational analysis remains the bottleneck in epigenome-wide sequencing-based DNA methylation research, and therefore there is an urgent need for better, more organized, integrated, comprehensive, user-friendly tools that also offer graphical visualizations.

ACKNOWLEDGMENT

The authors would like to acknowledge Long Research Grant Scheme (LRGS/2014/UKM-UKM/K/01) from Ministry of Higher Education Malaysia for supporting our colorectal cancer research.

REFERENCES

- [1] Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 2010;31(1):27–36.
- [2] Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. *Genome Med* 2014;6(8):66.
- [3] Jones PA, Issa JP, Baylin S. Targeting the cancer epigenome for therapy. *Nat Rev Genet* 2016;17(10):630–41.
- [4] Zakhari S. Alcohol metabolism and epigenetics changes. *Alcohol Research* 2013;35(1):6–16.
- [5] Kim KD, El Baidouri M, Jackson SA. Accessing epigenetic variation in the plant methylome. *Brief Funct Genom* 2014;13(4):318–27.
- [6] Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 2017;356(6337).
- [7] Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum Mol Genet* 2015;24(6):1528–39.
- [8] Barlow DP, Bartolomei MS. Genomic imprinting in mammals. *Cold Spring Harbor Perspect Biol* 2014;6(2).
- [9] Lorincz MC, Dickerson DR, Schmitt M, Groudine M. Infragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* 2004;11(11):1068–75.
- [10] Klutstein M, Nejman D, Greenfield R, Cedar H. DNA methylation in cancer and aging. *Canc Res* 2016;76(12):3446–50.
- [11] Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. *Cell* 1985;40(1):91–9.
- [12] Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* 1989;83(2):181–8.
- [13] Robertson KD, Jones PA. DNA methylation: past, present and future directions. *Carcinogenesis* 2000;21(3):461–7.
- [14] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;13(7):484–92.
- [15] Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Canc Cell* 2014;26(4):577–90.
- [16] Maunakea AK, Chepelev I, Cui K, Zhao K. Infragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res* 2013;23(11):1256–69.
- [17] Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. *Trends Genet* 2015;31(5):274–80.
- [18] Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev* 2009;19(6):541–9.

- [19] Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, et al. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol* 2014;15(9):469.
- [20] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature* 2012;489(7414):75–82.
- [21] Kron KJ, Bailey SD, Lupien M. Enhancer alterations in cancer: a source for a cell identity crisis. *Genome Med* 2014;6(9):77.
- [22] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;13(10):705–19.
- [23] Infinium methylationEPIC kit. November 24, 2017. <https://www.illumina.com>.
- [24] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98(4):288–95.
- [25] Human CpG island microarray. November 24, 2017. <http://www.genomics.agilent.com>.
- [26] Human DNA methylation microarrays. November 24, 2017. <https://www.genomics.agilent.com>.
- [27] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462(7271):315–22.
- [28] Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 2008;452(7184):215–9.
- [29] Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. *Genome Res* 2010;20(3):320–31.
- [30] Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res* 2013;23(9):1541–53.
- [31] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005;33(18):5868–77.
- [32] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;454(7205):766–70.
- [33] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D’Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 2010;466(7303):253–7.
- [34] Nair SS, Coolen MW, Stirzaker C, Song JZ, Statham AL, Strbenac D, et al. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* 2011;6(1):34–44.
- [35] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 2005;37(8):853–62.
- [36] Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* 2015;72:29–40.
- [37] Serre D, Lee BH, Ting AH. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* 2010;38(2):391–9.
- [38] Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 2010;28(10):1097–105.
- [39] Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* 2016;17:11.
- [40] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 2011;27(11):1571–2.

- [41] Ashktorab H, Shakoori A, Zarnogi S, Sun X, Varma S, Lee E, et al. Reduced representation bisulfite sequencing determination of distinctive DNA hypermethylated genes in the progression to colon cancer in African Americans. *Gastroenterol Res Pract* 2016;2016:2102674.
- [42] Ashktorab H, Daremipouran M, Goel A, Varma S, Leavitt R, Sun X, et al. DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia. *Epigenetics* 2014; 9(4):503–12.
- [43] Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* 2011;44(1):40–6.
- [44] Baharudin R, Ab Mutalib NS, Othman SN, Sagap I, Rose IM, Mohd Mokhtar N, Jamal R. Identification of predictive DNA methylation biomarkers for chemotherapy response in colorectal cancer. *Front Pharmacol* 2017;8:47.
- [45] Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* 2014;30(3):428–30.
- [46] Butcher LM, Beck S. Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods* 2015;72:21–8.
- [47] Vymetalkova V, Vodicka P, Pardini B, Rosa F, Levy M, Schneiderova M, et al. Epigenome-wide analysis of DNA methylation reveals a rectal cancer-specific epigenomic signature. *Epigenomics* 2016;8(9): 1193–207.
- [48] Kok-Sin T, Mokhtar NM, Ali Hassan NZ, Sagap I, Mohamed Rose I, Harun R, Jamal R. Identification of diagnostic markers in colorectal cancer via integrative epigenomics and genomics data. *Oncol Rep* 2015; 34(1):22–32.
- [49] Luo Y, Wong CJ, Kaz AM, Dzieciatkowski S, Carter KT, Morris SM, et al. Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* 2014;147(2):418–29.
- [50] Aryee MJ, Jaffe AE, Corradia-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;30(10):1363–9.
- [51] Naumov VA, Generozov EV, Zaharjevskaya NB, Matushkina DS, Larin AK, Chernyshov SV, et al. Genome-scale analysis of DNA methylation in colorectal cancer using infinium human methylation450 beadchips. *Epigenetics* 2013;8(9):921–34.
- [52] Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu SIMA. An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* 2012;28(5):729–30. <https://doi.org/10.1093/bioinformatics/bts013>.
- [53] Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012;22(2):271–82.
- [54] Houseman EA, Christensen BC, Yeh RF, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, Kelsey KT. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinf* 2008;9:365. <https://doi.org/10.1186/1471-2105-9-365>.
- [55] Koestler DC, Christensen BC, Marsit CJ, Kelsey KT, Houseman EA. Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. *Stat Appl Genet Mol Biol* 2013;12(2):225–40. <https://doi.org/10.1515/sagmb-2012-0068>.
- [56] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487(7407):330–7.

- [57] Kim YH, Lee HC, Kim SY, Yeom YI, Ryu KJ, Min BH, et al. Epigenomic analysis of aberrantly methylated genes in colorectal cancer identifies genes commonly affected by epigenetic alterations. *Ann Surg Oncol* 2011;18(8):2338–47.
- [58] Young PE, Womeldorf CM. Colonoscopy for colorectal cancer screening. *J Canc* 2013;4(3):217–26.
- [59] Lech G, Slotwiński R, Słodkowski M, Krasnodebski IW. Colorectal cancer tumour markers and biomarkers: recent therapeutic advances. *World J Gastroenterol* 2016;22(5):1745–55.
- [60] Gonzalez-Pons M, Cruz-Correia M. Colorectal cancer biomarkers: where are we now? *BioMed Res Int* 2015;2015:149014.
- [61] Iannone A, Losurdo G, Pricci M, Girardi B, Massaro A, Principi M, et al. Stool investigations for colorectal cancer screening: from occult blood test to DNA analysis. *J Gastrointest Canc* 2016;47(2):143–51.
- [62] Solé X, Crous-Bou M, Cordero D, Olivares D, Guinó E, Sanz-Pamplona R, et al. Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS One* 2014;9(9):e106748.
- [63] Takane K, Midorikawa Y, Yagi K, Sakai A, Aburatani H, Takayama T, et al. Aberrant promoter methylation of PPP1R3C and EFHD1 in plasma of colorectal cancer patients. *Cancer Med* 2014;3(5):1235–45.
- [64] Oh T, Kim N, Moon Y, Kim MS, Hoehn BD, Park CH, et al. Genome-wide identification and validation of a novel methylation biomarker, SDC2, for blood-based detection of colorectal cancer. *J Mol Diagn* 2013;15(4):498–507.
- [65] Herbst A, Rahmig K, Stieber P, Philipp A, Jung A, Ofner A, et al. Methylation of NEUROG1 in serum is a sensitive marker for the detection of early colorectal cancer. *Am J Gastroenterol* 2011;106(6):1110–8.
- [66] Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, et al. DNA methylation biomarkers for blood-based colorectal cancer screening. *Clin Chem* 2008;54(2):414–23.
- [67] Behrouz Sharif S, Hashemzadeh S, Mousavi Ardehaie R, Eftekharasadat A, Ghojazadeh M, Mehrtash AH, et al. Detection of aberrant methylated SEPT9 and NTRK3 genes in sporadic colorectal cancer patients as a potential diagnostic biomarker. *Oncol Lett* 2016;12(6):5335–43.
- [68] Tóth K, Sipos F, Kalmár A, Patai AV, Wichmann B, Stoehr R, et al. Detection of methylated SEPT9 in plasma is a reliable screening method for both left- and right-sided colon cancers. *PLoS One* 2012;7(9):e46000.
- [69] Nian J, Sun X, Ming S, Yan C, Ma Y, Feng Y, et al. Diagnostic accuracy of methylated SEPT9 for blood-based colorectal cancer detection: a systematic review and meta-analysis. *Clin Transl Gastroenterol* 2017;8(1):e216.
- [70] Guo Q, Song Y, Zhang H, Wu X, Xia P, Dang C. Detection of hypermethylated fibrillin-1 in the stool samples of colorectal cancer patients. *Med Oncol* 2013;30(4):695.
- [71] Suzuki H, Watkins DN, Jair KW, Schuebel KE, Markowitz SD, Chen WD, et al. Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nat Genet* 2004;36(4):417–22.
- [72] Huang Z, Li L, Wang J. Hypermethylation of SFRP2 as a potential marker for stool-based detection of colorectal cancer and precancerous lesions. *Dig Dis Sci* 2007;52(9):2287–91.
- [73] Zhang H, Qi J, Wu YQ, Zhang P, Jiang J, Wang QX, et al. Accuracy of early detection of colorectal tumours by stool methylation markers: a meta-analysis. *World J Gastroenterol* 2014;20(38):14040–50.
- [74] Lu H, Huang S, Zhang X, Wang D, Zhang X, Yuan X, et al. DNA methylation analysis of SFRP2, GATA4/5, NDRG4 and VIM for the detection of colorectal cancer in fecal DNA. *Oncol Lett* 2014;8(4):1751–6.
- [75] Yang Q, Huang T, Ye G, Wang B, Zhang X. Methylation of SFRP2 gene as a promising noninvasive biomarker using feces in colorectal cancer diagnosis: a systematic meta-analysis. *Sci Rep* 2016;6:33339.
- [76] Ahlquist DA, Taylor WR, Mahoney DW, Zou H, Domanico M, Thibodeau SN, et al. The stool DNA test is more accurate than the plasma septicin 9 test in detecting colorectal neoplasia. *Clin Gastroenterol Hepatol* 2012;10(3):272–277.e1.
- [77] Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology* 2010;138(6):2073–2087.e3.

- [78] Ebert MP, Tänzer M, Balluff B, Burgermeister E, Kretzschmar AK, Hughes DJ, et al. TFAP2E-DKK4 and chemoresistance in colorectal cancer. *N Engl J Med* 2012;366(1):44–53.
- [79] He J, Pei L, Jiang H, Yang W, Chen J, Liang H. Chemoresistance of colorectal cancer to 5-fluorouracil is associated with silencing of the BNIP3 gene through aberrant methylation. *J Canc* 2017;8(7):1187–96.
- [80] Yu J, Ma X, Cheung KF, Li X, Tian L, Wang S, et al. Epigenetic inactivation of T-box transcription factor 5, a novel tumor suppressor gene, is associated with colon cancer. *Oncogene* 2010;29(49):6464–74.
- [81] Wang S, Dong Y, Zhang Y, Wang X, Xu L, Yang S, et al. DACT2 is a functional tumor suppressor through inhibiting Wnt/β-catenin pathway and associated with poor survival in colon cancer. *Oncogene* 2015;34(20):2575–85.
- [82] Nilsson TK, Löf-Öhlin ZM, Sun XF. DNA methylation of the p14ARF, RASSF1A and APC1A genes as an independent prognostic factor in colorectal cancer patients. *Int J Oncol* 2013;42(1):127–33.
- [83] Ahn JB, Chung WB, Maeda O, Shin SJ, Kim HS, Chung HC, et al. DNA methylation predicts recurrence from resected stage III proximal colon cancer. *Cancer* 2011;117(9):1847–54.
- [84] Young GP, Pedersen SK, Mansfield S, Murray DH, Baker RT, Rabbitt P, et al. A cross-sectional study comparing a blood test for methylated BCAT1 and IKZF1 tumor-derived DNA with CEA for detection of recurrent colorectal cancer. *Cancer Med* 2016;5(10):2763–72.
- [85] Kandimalla R, Linnekamp JF, van Hooff S, Castells A, Llor X, Andreu M, et al. Methylation of WNT target genes AXIN2 and DKK1 as robust biomarkers for recurrence prediction in stage II colon cancer. *Oncogenesis* 2017;6(4):e308.
- [86] Swets M, Zaalberg A, Boot A, van Wezel T, Frouws MA, Bastiaannet E, Gelderblom H, van de Velde CJ, Kuppen PJ. Tumor LINE-1 methylation level in association with survival of patients with stage II colon cancer. *Int J Mol Sci* 2016;18(1):E36.
- [87] Calegari MA, Inno A, Monterisi S, Orlandi A, Santini D, Basso M, et al. A phase 2 study of temozolomide in pretreated metastatic colorectal cancer with MGMT promoter methylation. *Br J Canc* 2017;116(10):1279–86.
- [88] Kang XC, Chen ML, Yang F, Gao BQ, Yang QH, Zheng WW, et al. Promoter methylation and expression of SOCS-1 affect clinical outcome and epithelial-mesenchymal transition in colorectal cancer. *Biomed Pharmacother* 2016;80:23–9.
- [89] Cha Y, Kim KJ, Han SW, Rhee YY, Bae JM, Wen X, et al. Adverse prognostic impact of the CpG island methylator phenotype in metastatic colorectal cancer. *Br J Canc* 2016;115(2):164–71.
- [90] Gerecke C, Scholtka B, Löwenstein Y, Fait I, Gottschalk U, Rogoll D, et al. Hypermethylation of ITGA4, TFPI2 and VIMENTIN promoters is increased in inflamed colon tissue: putative risk markers for colitis-associated cancer. *J Canc Res Clin Oncol* 2015;141(12):2097–107.
- [91] Hu S, Cao B, Zhang M, Linghu E, Zhan Q, Brock M, et al. Epigenetic silencing BCL6B induced colorectal cancer proliferation and metastasis by inhibiting P53 signaling. *Am J Cancer Res* 2015;5(2):651–62.
- [92] Perez-Carbonell L, Balaguer F, Toiyama Y, Egoavil C, Rojas E, Guarinos C, et al. IGFBP3 methylation is a novel diagnostic and predictive biomarker in colorectal cancer. *PLoS One* 2014;9(8):e104285.
- [93] Benard A, Zeestraten EC, Goossens-Beumer IJ, Putter H, van de Velde CJ, Hoon DS, et al. DNA methylation of apoptosis genes in rectal cancer predicts patient survival and tumor recurrence. *Apoptosis* 2014;19(11):1581–93.
- [94] Li C, Cai S, Wang X, Jiang Z. Hypomethylation-associated up-regulation of TCF3 expression and recurrence in stage II and III colorectal cancer. *PLoS One* 2014;9(11):e112005.
- [95] Lange CP, Campan M, Hinoue T, Schmitz RF, van der Meulen-de Jong AE, Slingerland H, et al. Genome-scale discovery of DNA-methylation biomarkers for blood-based detection of colorectal cancer. *PLoS One* 2012;7(11):e50266.
- [96] Tänzer M, Balluff B, Distler J, Hale K, Leodolter A, Röcken C, et al. Performance of epigenetic markers SEPT9 and ALX4 in plasma for detection of colorectal precancerous lesions. *PLoS One* 2010;5(2):e9061.

- [97] Mizukami H, Shirahata A, Goto T, Sakata M, Saito M, Ishibashi K, et al. PGP9.5 methylation as a marker for metastatic colorectal cancer. *Anticancer Res* 2008;28(5A):2697–700.
- [98] Bock C, Lengauer T. Computational epigenetics. *Bioinformatics* 2008;24(1):1–10.
- [99] Lim SJ, Tan TW, Tong JC. Computational Epigenetics: the new scientific paradigm. *Bioinformation* 2010; 4(7):331–7.
- [100] Robinson MD, Pelizzola M. Computational epigenomics: challenges and opportunities. *Front Genet* 2015; 6:88.
- [101] OmicTools. November 24, 2017. <https://omictools.com>.
- [102] Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;6(6):692–702.
- [103] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013;29(2):189–96.
- [104] Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium methylation 450K technology. *Epigenomics* 2011;3(6):771–84.
- [105] Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 2012; 4(3):325–41.
- [106] Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* 2012;13(6):R44.
- [107] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8(1):118–27.
- [108] Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, et al. Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol* 2014;15(2):R30.
- [109] Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for illumina BeadChips. *Bioinformatics* 2017.
- [110] Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016;8(3):389–99.
- [111] Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. *Epigenet Chromatin* 2015;8:6.
- [112] Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* 2012;13(1):166–78.
- [113] Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11(2):R14.
- [114] Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinf* 2012;13:86.
- [115] Luu PL, Gerovska D, Arrospide-Elgarresta M, Retegi-Carrión S, Schöler HR, Araúzo-Bravo MJ. P3BSseq: parallel processing pipeline software for automatic analysis of bisulfite sequencing data. *Bioinformatics* 2017;33(3):428–31.
- [116] Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* 2012;9(2):145–51.
- [117] Van Gurp TP, Wagemaker NC, Wouters B, Vergeer P, Ouborg JN, Verhoeven KJ. epiGBS: reference-free reduced representation bisulfite sequencing. *Nat Methods* 2016;13(4):322–4.
- [118] Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* 2016;26(2):256–62.
- [119] Mayo TR, Schweikert G, Sanguinetti G. M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics* 2015;31(6):809–16.

- [120] Gao S, Zou D, Mao L, Zhou Q, Jia W, Huang Y, et al. SMAP: a streamlined methylation analysis pipeline for bisulfite sequencing. *GigaScience* 2015;4:29.
- [121] Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* 2014;11(11):1138–40.
- [122] Akman K, Haaf T, Gravina S, Vijg J, Tresch A. Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data. *Bioinformatics* 2014;30(13):1933–4.
- [123] Stockwell PA, Chatterjee A, Rodger EJ, Morison IM. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* 2014;30(13):1814–22.
- [124] Hansen KD, Langmead B, Irizarry RA. BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;13(10):R83.
- [125] Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res* 2011;39(9):e58.
- [126] Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 2013;8(12):e81148.
- [127] Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, et al. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinf* 2015;16:313.
- [128] Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, et al. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* February 24, 2014;15(2):R38.
- [129] Methylation plotter. A web tool for dynamic visualization of DNA methylation data. *Source Code Biol Med* 2014;9:11. <https://doi.org/10.1186/1751-0473-9-11>.
- [130] Zackay A, Steinhoff C. MethVisual – visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing. *BMC Res Notes* 2010;3:337. <https://doi.org/10.1186/1756-0500-3-337>.

FURTHER READING

- Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009; 41(2):178–86.
- Morris TJ, Beck S. Analysis pipelines and packages for infinium HumanMethylation450 BeadChip (450k) data. *Methods* 2015;72:3–8. <https://doi.org/10.1016/j.ymeth.2014.08.011>.
- Stracci F, Zorzi M, Grazzini G. Colorectal cancer screening: tests, strategies, and perspectives. *Front Public Health* 2014;2:210.

INTEGRATIVE OMIC ANALYSIS OF NEUROBLASTOMA 19

Kamalakannan Palanichamy

*Department of Radiation Oncology, The Ohio State University College of Medicine and Comprehensive Cancer Center,
Columbus, OH, United States*

INTRODUCTION

Initially, I have provided a brief overview of neuroblastoma and its diagnosis and prognosis. Subsequently, I have provided a snapshot of next-generation sequencing and various “omic” analyses and ways to integrate them for advancing the cure for these devastating high-risk neuroblastomas. Finally, I have summarized the sequencing studies that have been published so far, and how these studies have helped us understand the biology of neuroblastoma.

NEUROBLASTOMA

Neuro has a Greek origin, meaning nerve and *blastoma* has a Latin origin meaning cancer in precursor cells. Neuroblastoma is a pediatric tumor composed of embryonic tissue in the developmental stage. The neuroblastoma cells are immature, yet to develop specific functions and are undifferentiated. More specifically, neuroblastomas are cancers of the sympathetic nervous system. About 98% of neuroblastomas are not inherited. There are no known risk factors for neuroblastoma since they are developed at a very young age and very rare in children over 10 years. Similar to other cancers, there are four stages: stage I—IV; stage II and stage IV have two subcategories; three risk groups, low risk, intermediate risk, and high risk with a 5-year survival of 95%, 90%, and 50% respectively. The characteristics of each stage are provided in **Table 19.1**.

Neuroblastoma is unique when compared to other cancers due to its tendency for spontaneous regression of tumors in infancy even though the original diagnosis was a metastatic disease. Patients with stage IV high-risk neuroblastoma (HRN) often undergo intensive multidisciplinary therapy such as surgery, radiotherapy, chemotherapy, and autologous hematopoietic stem cell transplantation. The current standard of care treatment for neuroblastoma is high-dose chemotherapy followed by administering patients with previously collected blood-forming cells (autologous stem cell transplant) to restore immune function. Patients who showed response undergo adjuvant treatment with a vitamin A derivative (isotretinoin), or immunotherapy, or differentiation therapy. However, despite of this treatment regimen, more than 50% of HRN cases succumb to the disease. More effective therapeutic regimens are required to improve the treatment outcome of this devastating disease. The prognostic

Table 19.1 Neuroblastoma Staging System

Stage	Subtype	Diagnosis
Stage I		Localized without lymph node involvement and may be surgically removed.
Stage II	IIA	Localized but surgical resection may not be possible.
	IIB	Localized with lymph node involvement and surgical resection may not be possible.
Stage III		Advanced, may be a large tumor with or without lymph node involvement.
Stage IV	IV	Advanced with metastatic spread.
	IVS	A special category applies to children less than 1-year old.

factors of neuroblastoma include v-myc myelocytomatosis viral related oncogene, neuroblastoma (*MYCN*) gene amplification, 11q aberration, and DNA ploidy. Research in the past has contributed to a marginal increase in overall survival rates, and about 25% of patients do not respond well to current therapies. Current treatment protocols are primarily derived from adult cancer therapeutic regimen and have undesirable short- and long-term side effects. Recent advances in genomics have revealed that pediatric cancers are genetically distinct from adult counterparts and require alternative treatment approaches.

OMICS: GENOMICS, TRANSCRIPTOMICS, PROTEOMICS, EPIGENOMICS, AND METABOLOMICS

The evolution of next-generation sequencing (NGS) technology has led to our enhanced comprehension of disease etiology through various interrelated and interdependent segments of the genome. The biological matter within the cell is composed of DNA, RNA, protein, and metabolites. Genomics is the study of genome by utilizing DNA which includes both the coding region (exons) and the noncoding region (introns). Transcriptomics is the study of transcriptome comprising the complete set of RNA transcripts that are produced by the genome. Proteomics is the study of proteins and their abundance, variations, modifications, and interactions. Epigenomics is the study of epigenome, a congregation of chemical compounds that can instruct genome. The chemical modifications include DNA methylation, histone modification, chromatin accessibility, etc. Metabolomics is the study of metabolome, which refers to the complete set of small molecules or metabolites, which primarily depends on the state of the system.

Next-generation sequencing (NGS) is high-throughput sequencing that has revitalized cancer medicine. The advances in NGS have led to the generation of large omic data sets and pose a big challenge from bioinformatics perspective to exploit the full potential of omic data. Currently emerging bioinformatics tools with novel algorithms for DNA-Seq, RNA-Seq, Chip-Seq, Methyl-Seq, etc. focus on de novo assembly or alignment of sequence reads, quality control, quantification, annotation, visualization, integration with other omes, etc., to retrieve and resolve complex biological information. Genome-Seq or DNA-Seq or Whole Exome-Seq (WES) enables us to analyze the genome by decrypting the multiple coding and noncoding DNA sequences and map the genome to decipher relationships. Some of the commonly used tools involve genome annotation, genome editing, genome variant, DNA structure analysis, and comparative genomics. Epigenome modification includes DNA methylation, histone modification, and nucleosome positioning. Analysis of DNA-Seq data allows us

to map the genes in the genome, exons—introns, regulatory elements, indels, repeats, and mutations. Genomic variants are differences in the genome that make everyone unique. Some of the single nucleotide polymorphisms (SNPs), insertions, deletions, substitutions, and structural variations are linked to specific phenotypes or diseases, but most of them have still unknown effects. Bioinformatics tools have been developed to study the different types of genomic variations, rearrangements, and modifications. Several programs proposed here can aid in predicting their effects on gene regulation and miRNA expression and identify cancer and disease-associated mutational signatures.

INTEGRATIVE OMICS

Integrative omics data analysis is challenging due to the complexity of individual omic data. Recent advances in informatics are helpful in analyzing multiomic data efficiently. There are two approaches to resolve bioinformatics of omics landscape. The first one is an individual omics informatics tool with a focus on the specific task in both single- and multiomic settings. The second one is the integrated software platform, which can assemble multiple tools and perform complicated tasks. Due to the complexity, a workflow management system (WMS) is required for effective and efficient processing of omic data from a simple software toolkit to the integrated multiomic platforms. The following sections will briefly discuss strategies of successful omics informatics. Several omics informatics tools have been developed for a specific purpose, but they have limited comprehensive analysis capabilities. Benchmarking of bioinformatics tools are lacking and requires development. There are several tools to analyze the NGS data for sequence alignment and assembly [1–5]. The following section will provide an overview of the NGS tools.

TOOLS FOR NGS DATA ANALYSIS AND INTEGRATIVE OMICS

The first task in the NGS data analysis is the sequence alignment and assembly. Millions of short read sequences are mapped to the reference genome, which occurs with a computational time. Some of the recent commonly used computational tools for sequence alignment, genome assembly, and transcriptome assembly are provided in Table 19.2.

WORKFLOW

The general workflow includes generating FASTQ files from raw reads, alignment of the reads to generate a SAM (sequence alignment map) or BAM (binary version of SAM) file, followed by variant calling and filtering to obtain analysis ready variant VCF (variant call format) files. After extracting

Table 19.2 Sequence Alignment and Assembly tools

Sequence Alignment	Genome Assembly	Transcriptome Assembly
MOSAIK [6]	Cotex [7]	Bowtie-TopHat-Cufflinks pipeline [8]
Bowtie 2 [9]	SOAP denovo [10]	Oases [11]
Stampy [12]		Trinity [13] Cufflinks [14]

genetic variants with relative abundance from individual omes, integrative omics occurs, which consists of multiple sequential processing steps. Refinement of data at this stage is done by using statistical and machine learning techniques and incorporating external databases. This crucial step increases the feasibility of experimental validation of the hypothesis generated from the analysis. Machine learning frameworks, such as Bayesian integration and probabilistic graphical models, can incorporate data from multiple omes [15,16]. CNAmet is an R-package for integrative analysis of copy number, DNA methylation, and gene expression data. The package is available at <http://csbi.ltdk.helsinki.fi/CNAmet> [17]. Other approaches such as multiple concerted disruption (MCD) analysis are also possible [18].

One of the major goals of integrative omics is to identify molecular subtypes, and determine the association of clinical parameters and outcomes with molecular subtypes known as medical genomics. Personalized medicine based on medical genomics for neuroblastoma is emerging and we can anticipate a more matured database will emerge soon. This is possible because of the advances in sequencing technologies accompanied with high speed computing capabilities. Integrating all the omics (genome, transcriptome, epigenome, proteome, and metabolome) data followed by clustering analysis with clinical data (health history, radiology report, pathology report, treatment, outcome, etc.) will provide us with the medical genome. The medical genomics data from clinical cohorts can be used to make appropriate clinical decisions for treating neuroblastomas. Schematic representation of the informatics from omes and clinical parameters leading to genomic medicine is shown in Fig. 19.1. The current exciting advances in this area will allow oncologists to employ personalized medicine to treat neuroblastoma.

The following are the widely used software tools to interrogate NGS data. The statistical tool ‘Picard’ is a variant calling program consisting of a set of Java tools to interact SAM/VCF files and is available at <http://broadinstitute.github.io/picard>. Samtools is another statistical tool for variant calling available at <http://www.htslib.org> [19]. Crossbow is a variant discovery toolkit which combines aligner Bowtie and the SNP caller SOAPsnp [20]. HugeSeq integrates several tools such as BWA, GATK, SAMtools, VCFtools, BEDtools, etc. into an efficient pipeline [21]. Churchill is another fast, scalable variant discovery platform available at <http://churchill.nchri.org> [22]. Software tools such as SAMtools and GATK can process data from multiple omes [10,21]. The Python toolkit includes Bcbio-nextgen, Ruffus, Bpipe, etc. Bcbio-nextgen provides scalable and reproducible pipelines, with distributed computation, and also supports configurable pipelines and automatic validation (<https://github.com/chapmanb/bcbio-nextgen>). Ruffus is another lightweight scalable library for creating and executing pipelines [23]. Bpipe is a simple, dedicated programming language for defining and executing bioinformatics pipelines with easy parallelism and restarting jobs [24]. Cpipe is a variant detection pipeline designed for diagnostic settings and is available at <http://cpipeline.org> [25]. COSMOS is another python library for NGS analysis and workflow management with a user interface to monitor job status and distribute resources based on demand [26]. NGSANE is a lightweight production informatics for analyzing high-throughput data available at <https://github.com/BauerLab/ngsane> [27].

Most of the tools discussed so far are adequate to conduct computational intensive omic data analysis, but most of them are not sufficient for integrative omic analyses. Open source Omics Pipe is a cloud-based modular tool that streamlines multiomics data analysis. Genomic and transcriptomic data analysis including functionality to interact with TCGA (The Cancer Genome Atlas) data sets can be conducted at <http://sulab.scripps.edu/omicspipe> and the source code is available at

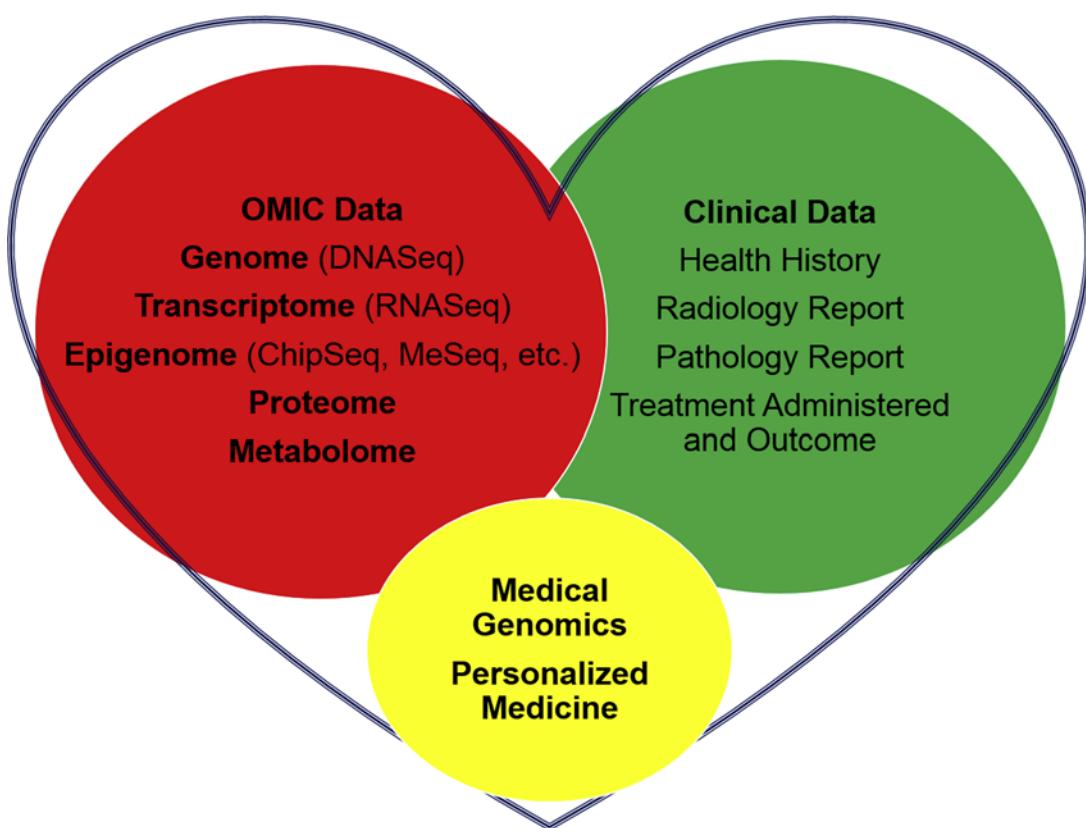


FIGURE 19.1

Schematic representation of deciphering medical omics from integrative omics

https://bitbucket.org/sulab/omics_pipe [28]. Taverna workflow suite is an open source suite to execute multiple workflows typically used in high-throughput omics analyses (<http://www.taverna.org.uk>). The workflows are reusable, executable, and can be shared for repurpose [29]. GenePattern developed by Broad Institute provides practice pipelines and a simple user interface for genomic data analysis [30]. Recently they have released a GenePattern notebook environment with an easy-to-use interface that provides access to hundreds of genomic tools without the need to write a code. The package is available at <http://www.genepattern-notebook.org> [31]. GenomeSpace is a cloud-based tool for integrative genomics that facilitates integrative analysis by nonprogrammers, and is available at <http://www.genomespace.org> [32]. Galaxy is another web-based portal which facilitates multiomic analyses with searchable remote resources, combining data and visualizing results. Galaxy can be accessed at <http://g2.bx.psu.edu> [33,34]. Illumina is the widely used platform for NGS and it has tools such as BaseSpace which is an Illumina genomic cloud computing platform which integrates a broad range of library tools and is available at <https://basespace.illumina.com>. NextBio Research is a data analysis platform owned by Illumina that focuses on gene function, drug, and disease mechanisms and is available at <https://www.nextbio.com>.

NEUROBLASTOMA OMICS

Genome

Whole-genome sequencing comprising 87 cases of stage I–IV neuroblastomas identified structural defects and local shedding of chromosomes (chromothripsis) in 18% of HRN. Genomic landscape of neuroblastoma revealed two molecular defects, chromothripsis and neuritogenesis gene alterations in HRN [35]. Neuroblastoma with chromothripsis are associated with poor prognosis. Structural alterations affected odd oz/ten-m homolog 3 (*ODZ3*), protein tyrosine phosphatase, receptor type D (*PTPRD*), and CUB and sushi domain-containing protein 1 (*CSMD1*), essential for neuronal growth cone stabilization [36–38]. The regulators of Rac/Rho pathway such as alpha thalassemia/mental retardation syndrome X-linked (*ATRX*), T-lymphoma invasion and metastasis-inducing protein 1 (*TIAM1*), and others were mutated leading to defects in neuritogenesis. Interestingly, most of these poor outcome neuroblastomas did not carry *MYCN* amplifications. Within the HRN, *MYCN* amplification status did not correlate to the mutation frequency. Lack of recurrent mutations indicates that neuroblastoma carry a few early somatic tumor driver mutations.

The National Cancer Institute (NCI) led initiative, “Therapeutically Applicable Research to Generate Effective Treatments” (TARGET), utilizes comprehensive molecular profiling to determine the genetic drivers responsible for initiation and progression of therapeutic resistant pediatric cancers. The goal of this consortium is to identify prognostic markers and therapeutic targets in order to develop new effective treatment strategies. The TARGET repository has data sets comprising gene expression (Affymetrix Human Exon ST Array), Copy Number & LOH (Affymetrix SNP 6.0 Array), DNA methylation (Illumina 450K Chip), WGS (DNA-Seq), WES (RNA-Seq). Summary of the patient cohorts with stage, *MYCN* status, and mutational count for both European and American neuroblastoma cohorts are provided in Fig. 19.2.

Genetic landscape of high-risk neuroblastoma (HRN): This is the most aggressive neuroblastoma with survival rates less than 50%. HRN originates from the sympathetic nervous system and is

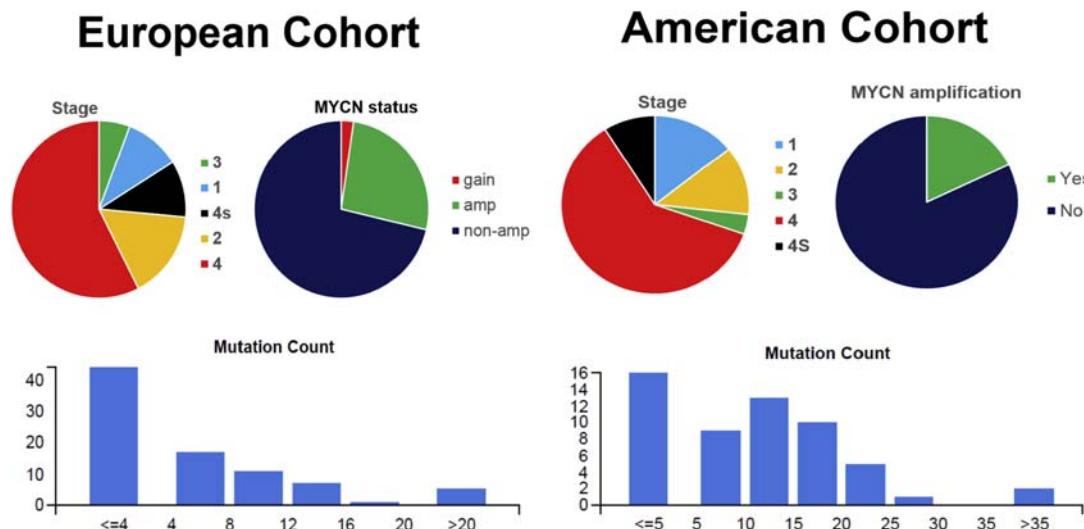


FIGURE 19.2

Mutational spectrum of European and American neuroblastoma cohorts

associated with widespread metastasis. The somatic mutational spectrum of 240 HRN cases from whole exome, genome, and transcriptome sequencing explored through TARGET initiative identified a low median exonic frequency of 0.6 per Mb. About 80% were nonsilent, with few recurrently mutated genes in these tumors. Somatic mutation frequencies were found in anaplastic lymphoma receptor tyrosine kinase (*ALK*) (9.2% of cases), protein tyrosine phosphatase nonreceptor type 11 (*PTPN11*) (2.9% of cases), alpha thalassemia/mental retardation syndrome X-linked (*ATRX*) (2.5% of cases), and *MYCN* (1.7% of cases) [39]. The relative scarcity of recurrent somatic mutations in these cases raises concerns on the current therapeutic regimens since most of them rely on frequently mutated oncogenic drivers. Across the coding regions of all the cases, 5291 candidate somatic mutations in 3960 genes were identified. Genes with a significant frequency of somatic mutation across 240 HRN are provided in Table 19.3 and pathogenic germline variants are provided in Table 19.4.

Table 19.3 Mutational frequency of genes in HRN

Gene	Description	Mutations
<i>ALK</i>	Anaplastic lymphoma receptor tyrosine kinase	22
<i>PTPN11</i>	Protein tyrosine phosphatase nonreceptor type 11	7
<i>ATRX</i>	Alpha thalassemia/mental retardation syndrome X-linked	6
<i>OR5T1</i>	Olfactory receptor family 5, subfamily T, member 1	3
<i>PDE6G</i>	Phosphodiesterase 6G, cGMP-specific	2
<i>MYCN</i>	v-myc myelocytomatosis viral related oncogene, neuroblastoma	4
<i>NRAS</i>	Neuroblastoma RAS viral (v-ras) oncogene homolog	2

Table 19.4 Pathogenic germline variants in HRN

Gene	Genome Position (hg19)	cDNA Change	Protein Change
<i>ALK</i>	Chr2:29,432,664	c.3824G > A	p.Arg1275Gln
<i>CHEK2</i>	Chr22:29,121,242	c.433C > T	p.Arg145Trp
<i>CHEK2</i>	Chr22:29,121,015	c.542G > A	p.Arg181His
<i>CHEK2</i>	Chr22:29,121,018	c.539G > A	p.Arg180His
<i>PINK1</i>	Chr1:20,972,133	c.1040T > C	p.Leu437Pro
<i>PINK1</i>	Chr1:20,971,042	c.836G > A	p.Arg279His
<i>BARD1</i>	Chr2:215,657,051	c.334C > T	p.Arg112 ^a
<i>BARD1</i>	Chr2:215,595,215	c.1921C > T	p.Arg641 ^a
<i>TP53</i>	Chr17:7,578,194	c.655C > T	p.Pro219Ser
<i>PALB2</i>	Chr16:23,646,182	c.1684+1C > A	Splice at Gly562

BARD1: *BRCA1 associated RING domain protein*.

CHEK2: *Serine/threonine protein kinase chk2*.

PALB2: *Partner and localizer of BRCA2*.

PINK1: *Serine/threonine protein kinase, mitochondrial*.

TP53: *Tumor protein P53*.

^aStop codon.

Whole-genome sequencing of 56 neuroblastomas comprising 39 HRN and 17 low-risk neuroblastomas (LRN) discovered rearrangements at Chr 5p15.33. This is contiguous to telomerase reverse transcriptase (*TERT*) in HRN in 12 out of 39 samples and is mutually exclusive with *MYCN* amplification and *ATRX* mutation. This rearrangement juxtaposes the *TERT* coding sequence to strong enhancer elements leading to a massive chromatin remodeling and DNA methylation. Most of the HRN are associated with *TERT* rearrangements, *MYCN* amplification, or *ATRX* mutations, all of which facilitates telomere lengthening. In contrast, LRN is characterized by the absence of such alterations, and lacks the ability to immortalize and proliferate [40].

Neuroblastomas are rare in adolescents and highly associated with age. Patients less than 18 months of age are more likely have a tumor that spontaneously regresses than in older children [41]. Individuals with African ancestry are more likely to have a more malignant phenotype than people of European descent [42]. Neuroblastoma are more common in boys than in girls. HRN occurs in older children with either *MYCN* amplification and 1p deletion or 11q deletions without *MYCN* amplification. Both high-risk groups exhibit 17q gain [43–45]. The large size of chromosomal aberrations on 11q and 17q have decreased the ability to identify causative genes, whereas genes such as cadherin 5 (*CDH5*), calmodulin binding transcription activator 1 (*CAMTA1*), and castor zinc finger 1 (*CASZ1*) are considered to be involved for 1p alteration [46–48].

In addition to *MYCN*, the two other oncogenes *ALK* and lin-28 homolog B (*LIN28B*) were found to be amplified [49,50]. *MYCN* activates certain genes by promoter binding and represses certain genes by the association of repressor protein complexes. Binding of *MYCN* to enhancer of zeste homolog 2 (*EZH2*), which is highly expressed in neuroblastoma and a core component of the polycomb repressive complex 2 (*PRC2*), may play a key role in activating or repressing genes. Oncogenic functions of *MYCN* include regulation of DNA replication, transcription, splicing, miRNA regulation, etc. [51–53]. *MYCN* and *LIN28B* work cooperatively in regulating each other. *LIN28B* regulates *MYCN* through *let-7* binding and *MYCN* directly regulates *LIN28B* through promoter binding [54]. *ALK* can increase *MYCN* expression by promoter activation [55].

Analysis of the genome and exome showed genomic alterations associated with the molecular pathogenesis of neuroblastoma [35,39,40,56–58]. Specifically, somatic point mutations and somatic structural variants in the *PTPRD*, *ODZ3*, *CSMD1*, and AT rich interaction domain 1A (*ARID1A*) genes [35,57], a few high-frequency recurrent somatic mutations in the *ALK*, chromodomain helicase DNA binding protein 9 (*CHD9*), protein tyrosine kinase 2 (*PTK2*), neuron navigator 3 (*NAV3*), neuron navigator 1 (*NAV1*), frizzled class receptor 1 (*FZDI*), *ATRX*, *ARID1B*, *TIAM1*, *ALK*, *PTPN11*, *OR5T1*, *PDE6G*, *MYCN*, and *NRAS* genes [35,39,56,57], and rearrangements in *TERT* gene superenhancer region are discovered in neuroblastoma patients with worst survival [40,58]. At the ends of chromosomes, telomeres are DNA–protein structures that protect the genome from insults or damage and shorten progressively over time. It is well known that telomere lengthening is associated with higher risk of cancer. A study investigated the causal relevance of telomere length for the risk of cancer and nonneoplastic diseases identified longer telomeres were associated with higher odds of neoplastic disease. The GWAS studies published up to 2015 were used in this study to measure odds ratio (OR) and 95% confidence intervals (CI) for disease per standard deviation (SD, Study-specific relative risks for disease per unit change or quantile comparison of telomere length were transformed to an SD scale) higher telomere length due to germline genetic variation. They found a stronger association with higher OR (95% CI) for neuroblastoma, 2.98 (1.92–4.62) [59].

TRANSCRIPTOME AND EPIGENOME

In general, DNA methylation is detected using the following approaches: utilizing methylation-sensitive restriction enzymes recognizing “GC” rich sequences, bisulfite treatment to convert unmethylated cytosine to thymine, affinity-based enrichment using 5-methyl cytosine antibody, etc. NGS technology uses the following methods: whole-genome bisulfite sequencing (WGBS-Seq), methylation array, reduced representation bisulfite sequencing (RRBS-Seq), methylated DNA immunoprecipitation (MeDIP), Tet-assisted bisulfite sequencing (Tab-Seq), and hmC methylation arrays. Initial epigenetic alteration studies have identified caspase-8 (*CASP8*) and RAS-association domain family 1 isoform A (*RASSF1A*) DNA methylations as key events for the development and progression of neuroblastoma [60–62]. The methylation status of these genes is significantly associated with survival [63,64]. A study reporting the integrative approach to analyze the methylomes, transcriptomes, and copy number variations in about 100 cases of neuroblastomas show that the DNA methylation patterns were clustered based on patient subgroups. Transcriptome integration revealed intragenic enhancer methylation as a mechanism for HRN associated transcriptional deregulation. This integrative analysis shows the cooperation between *PRC2* activity and DNA methylation resulting in the inhibition of tumor-suppressive differentiation program contributing to neuroblastoma pathogenesis [65]. Recurrent somatic mutations are rare in neuroblastomas, suggesting that epigenetic mechanisms may drive neuroblastoma development and progression.

Hierarchical clustering of neuroblastoma methyl-Seq data identified two major. All *MYCN* amplified tumors mapped to cluster 1, showing the impact of *MYCN* on DNA methylation patterns. In cluster 2, there are subgroups, low-risk 4S patients (age <1.5 years) were enriched in subgroup 2s with no *MYCN* amplification. Segmental chromosomal aberrations were rare in this group. The remaining patients of cluster 2 were *MYCN*-nonamplified tumors, but were HRN due to the enrichment of other variables associated with poor outcomes such as age >1.5 years, 11q deletion, chromosomal alterations. Cluster 1 tumors had higher *MYC(N)* target gene activity than cluster 2 tumors, which might be due to *c-MYC* activity as indicated by elevated *MYC* mRNA levels. Cluster 1 patients have a poor survival outcome, cluster 2 patients have an intermediate survival outcome, and subgroup 2s patients have an excellent survival outcome.

The protocadherin beta family (*PCDHB*) gene body CpG methylation is associated with HRN CpG island methylator phenotype (CIMP) and CIMP-positive samples are *PCDHB* hypermethylated. Four CpG categories were associated with HRNs:

1. 860 CpGs representing 341 genes exhibited hypermethylation with downregulation of the corresponding gene (HyperDownHR).
2. 396 CpGs representing 167 genes exhibited hypomethylation with upregulation of the corresponding gene (HypoUpHR).
3. 602 CpGs representing 178 genes exhibited hypermethylation with upregulation of the corresponding gene (HyperUpHR).
4. 887 CpGs representing 2273 genes exhibited hypomethylation with downregulation of the corresponding gene (HypoDownHR).

Intragenic CpGs whose methylation was negatively associated with gene expression (HyperDownHR and HypoUpHR) were enriched with enhancer elements, highlighting the role of nonpromoter DNA

methylation in neuroblastoma pathogenesis. Genes such as C-X-C-motif chemokine receptor 4 (*CXCR4*), galanin and GMAP prepropeptide (*GAL*), leucine rich repeat neuronal 1 (*LRRN1*), ornithine decarboxylase 1 (*ODC1*), *TWIST1*, and Wolf-Hirschhorn syndrome candidate 1 protein (*WHSC1*) were highly expressed in HRNs with HypoUpHR. In the HyperDownHR cluster, genes such as ATP binding cassette subfamily B member 1 (*ABCB1*), calcium voltage-gated channel subunit alpha 1G (*CACNA1G*), *CD44*, dual specificity phosphatase 23 (*DUSP23*), PR/SET domain 2 (*PRDM2*), retinol binding protein 1 (*RBPI*), secreted frizzled related protein 1 (*SFRP1*), chromodomain helicase DNA binding protein 5 (*CHD5*), and neurotrophic receptor tyrosine kinase 1 (*NTRK1*). All these genes have been previously described to be involved in the biology of HRNs [66–74].

INTEGRATIVE OMICS

The neuroblastoma sequencing studies have shown that the cellular and molecular basis vastly differs from adult cancers and current treatment approaches for neuroblastomas raise concerns. Therefore, current approaches require critical evaluation and new targeted therapies based on etiology, genomics, transcriptomics, epigenomics, and vulnerabilities are required to cater this unmet need in HRN. Equipping with next-generation sequencing technologies and other platforms to interrogate integrative (GenOMICS, TranscriptOMICS, EpigenOMICS, ProteOMICS, MetabolOMICS) omics could address the current challenges, leading to clinically relevant discoveries. However, integrating different omics data is convoluted and the approaches to integrate them are getting better and better. Integrative omics offers many advantages in understanding the underlying genetic and molecular factors that play critical roles in disease initiation and progression. Integrative omics increases the confidence in findings, simultaneously decreasing false positives due to the assimilation of data obtained at genome, transcriptome, epigenome, and proteome levels. Multiple sources of evidence from different data types point to a similar observation, resulting in increased accuracy. This comprehensive integrative approach allows a better understanding of biological process reliance and regulation at different levels. The two best approaches, meta-dimensional and multistage analyses, are useful for conducting integrative omic analyses [75]. Meta-dimensional analysis integrates multiple data sets in a single study through three approaches such as concatenation-, transformation-, and model-based: integrating multiple data matrices from different omics data into one large input matrix before model construction (concatenation-based); transforming each omics data type into an intermediate form, such as a graph matrix that represents a network before multiple omics data are combined (transformation-based); or generating different models by using different types of omic data as training sets before integrating into a final model, thus allowing independent analysis for each omics data type (model-based) [76–78]. Multistage analysis, as the name implies, divides data analysis into multiple steps while capturing associations between enriched signals among different omic data types at each step [79]. The goal of the integrative omic analysis is to identify effective models that predict phenotypic traits and outcomes, elucidate biomarkers, and generate insights into the genetic underpinnings of the heritability of complex traits. Emerging machine learning and informatics advances will soon develop an integrative analysis platform where one can perform a descriptive analysis to find underlying relationship between different omes, and to predict a response using one or more omes.

NETWORK MODELING, REVERSE ENGINEERING MODELING, AND DYNAMIC MODELING

Network-based modeling uses the pre-constructed network. The two types of data (i) contact- or phenotype-specific biological information such as genomic, transcriptomic, epigenomic, proteomic, or metabolomics data, and (ii) interaction information such as protein–protein/gene–gene–protein interactions or pathways, are used. Interaction study at genome and proteome scale facilitates the discovery of novel functional cross talk between them. This enables us to overlay biological data onto network scaffold and investigate the function of a given network. Specific examples of such database can be found in resources such as KEGG, BioGRID, STRING, etc. Additional resources include IPA, NeAT, etc. [80–88]. The reverse engineering modeling approach uses data from reverse-engineering networks. One of the major advantages of reverse engineering methods is that no prior knowledge is required to infer a cause–effect relationship [89–91]. NetDecoder is a program that utilizes reverse engineering based approach [86]. Dynamic modeling utilizes mathematical approaches that facilitates the modeling of altered signaling pathways due to cancer progression and altered signaling pathways. It is capable of uncovering the phosphorylation events due to gain or loss of protein–protein interactions in establishing novel pathway cross talk [92–94].

MACHINE LEARNING-BASED MODELING

Machine learning algorithms are mathematical model mapping methods used to learn or uncover underlying patterns embedded in the data. Machine learning comprises a group of computational algorithms that can perform pattern recognition, classification, and prediction on data by learning from existing data (training set). The most common machine learning approaches in biology are support vector machines (SVM) and artificial neural networks (ANN) [95–97]. ANN model with deep neuron layers can be used to predict sequence specificities of DNA- and RNA-binding proteins, noncoding variants, alternative splicing, and quantitative structure–activity relationship (QSAR) of drugs [98–101]. Deep learning models have outperformed other machine learning methods in identifying more complex features from data [102]. To achieve complex results, deep learning techniques require a higher volume of data and computational time, compared to other machine learning algorithms. Omic data such as genome, transcriptome, epigenome, proteome, and metabolome may be integrated into a single model, which has large dimensions, and requires extensive time to build an appropriate model. Data collection can be minimized by reducing the dimension of input data, which can be done before or after data integration with principal component analysis (PCA), or after data integration with feature selection algorithms [103]. Mode of action by network identification (MNI) combines reverse engineering network modeling with machine learning to decipher regulatory interactions. MNI uses a training set of multidimensional omic data to identify genetic components and network that correspond to a specific state. MNI, using a set of ordinary differential equations, directed graph relating the amounts of biomolecules to each other can be generated. For example, when transcriptomic data are used as training data, regulatory influences between genes can be inferred. In addition to MNI, another network-based system CellNet classifies cellular states based on the status of gene regulatory network [104,105]. Both MNI and CellNet utilize machine learning integrated reverse engineering methods.

SUMMARY AND FUTURE DIRECTIONS

Integrative genomic, transcriptomic, epigenomic data of neuroblastomas show sparse mutations when compared to other cancers. The lack of information from the mutational landscape complicates the understanding of the initiation, progression, and metastasis of neuroblastomas. Currently, neuroblastomas are treated with chemotherapeutics widely used in adult cancers. Although mutational spectrum in neuroblastoma is not incredibly complex, it poses an enormous challenge. In stage IV and IVS, the age is the deciding criterion for a complete relapse or regression. Future research in this area from the immune perspective is essential and taking an integrative omic approach can equip us for a successful targeting of this devastating disease.

REFERENCES

- [1] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011; 12(5):363–76.
- [2] Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet* 2013;14(3):157–67.
- [3] Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet* 2011;12(10):671–82.
- [4] Engstrom PG, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 2013;10(12):1185–91.
- [5] Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet* 2013;14(5):333–46.
- [6] Lee WP, et al. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* 2014;9(3):e90581.
- [7] Iqbal Z, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;44(2):226–32.
- [8] Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;7(3):562–78.
- [9] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.
- [10] Li R, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010;20(2):265–72.
- [11] Schulz MH, et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;28(8):1086–92.
- [12] Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011;21(6):936–9.
- [13] Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;29(7):644–52.
- [14] Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28(5):511–5.
- [15] Hu P, et al. Integrating multiple resources to identify specific transcriptional cooperativity with a Bayesian approach. *Bioinformatics* 2014;30(6):823–30.
- [16] Wellcome Trust Case Control C, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 2012;44(12):1294–301.
- [17] Louhimo R, Hautaniemi S. CNAMet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* 2011;27(6):887–8.
- [18] Chari R, et al. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst Biol* 2010;4:67.

- [19] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27(21):2987–93.
- [20] Langmead B, et al. Searching for SNPs with cloud computing. *Genome Biol* 2009;10(11):R134.
- [21] Lam HY, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol* 2012;30(3):226–9.
- [22] Kelly BJ, et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol* 2015; 16:6.
- [23] Goodstadt L. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics* 2010;26(21): 2778–9.
- [24] Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* 2012;28(11):1525–6.
- [25] Sadedin SP, et al. Cpipe: a shared variant detection pipeline designed for diagnostic settings. *Genome Med* 2015;7(1):68.
- [26] Gafni E, et al. COSMOS: Python library for massively parallel workflows. *Bioinformatics* 2014;30(20): 2956–8.
- [27] Buske FA, et al. NGSANE: a lightweight production informatics framework for high-throughput data analysis. *Bioinformatics* 2014;30(10):1471–2.
- [28] Fisch KM, et al. Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics* 2015;31(11):1724–8.
- [29] Wolstencroft K, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* 2013;41(Web Server issue):W557–61.
- [30] Reich M, et al. GenePattern 2.0. *Nat Genet* 2006;38(5):500–1.
- [31] Reich M, et al. The GenePattern notebook environment. *Cell Syst* 2017;5(2):149–51. e1.
- [32] Qu K, et al. Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nat Methods* 2016;13(3):245–7.
- [33] Giardine B, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;15(10): 1451–5.
- [34] Boekel J, et al. Multi-omic data analysis using Galaxy. *Nat Biotechnol* 2015;33(2):137–9.
- [35] Molenaar JJ, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* 2012;483(7391):589–93.
- [36] Kraus DM, et al. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J Immunol* 2006;176(7):4419–30.
- [37] Sun QL, et al. Growth cone steering by receptor tyrosine phosphatase delta defines a distinct class of guidance cue. *Mol Cell Neurosci* 2000;16(5):686–95.
- [38] Zheng L, et al. Drosophila Ten-m and filamin affect motor neuron growth cone guidance. *PLoS One* 2011; 6(8):e22956.
- [39] Pugh TJ, et al. The genetic landscape of high-risk neuroblastoma. *Nat Genet* 2013;45(3):279–84.
- [40] Peifer M, et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* 2015; 526(7575):700–4.
- [41] Maris JM. Recent advances in neuroblastoma. *N Engl J Med* 2010;362(23):2202–11.
- [42] Gamazon ER, et al. Trans-population analysis of genetic mechanisms of ethnic disparities in neuroblastoma survival. *J Natl Cancer Inst* 2013;105(4):302–9.
- [43] Speleman F, De Preter K, Vandesompele J. Neuroblastoma genetics and phenotype: a tale of heterogeneity. *Semin Canc Biol* 2011;21(4):238–44.
- [44] Schleiermacher G, Janoueix-Lerosey I, Delattre O. Recent insights into the biology of neuroblastoma. *Int J Canc* 2014;135(10):2249–61.

- [45] Domingo-Fernandez R, et al. The role of genetic and epigenetic alterations in neuroblastoma disease pathogenesis. *Pediatr Surg Int* 2013;29(2):101–19.
- [46] Henrich KO, et al. CAMTA1, a 1p36 tumor suppressor candidate, inhibits growth and activates differentiation programs in neuroblastoma cells. *Canc Res* 2011;71(8):3142–51.
- [47] Liu Z, et al. CASZ1 inhibits cell cycle progression in neuroblastoma by restoring pRb activity. *Cell Cycle* 2013;12(14):2210–8.
- [48] Fujita T, et al. CHD5, a tumor suppressor gene deleted from 1p36.31 in neuroblastomas. *J Natl Cancer Inst* 2008;100(13):940–9.
- [49] Mosse YP, et al. Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* 2008;455(7215):930–5.
- [50] Molenaar JJ, et al. LIN28B induces neuroblastoma and enhances MYCN levels via let-7 suppression. *Nat Genet* 2012;44(11):1199–206.
- [51] Conacci-Sorrell M, McDowell L, Eisenman RN. An overview of MYC and its interactome. *Cold Spring Harb Perspect Med* 2014;4(1):a014357.
- [52] Wang C, et al. EZH2 Mediates epigenetic silencing of neuroblastoma suppressor genes CASZ1, CLU, RUNX3, and NGFR. *Canc Res* 2012;72(1):315–24.
- [53] Schulte JH, et al. MYCN regulates oncogenic MicroRNAs in neuroblastoma. *Int J Canc* 2008;122(3):699–704.
- [54] Beckers A, et al. MYCN-targeting miRNAs are predominantly downregulated during MYCN-driven neuroblastoma tumor formation. *Oncotarget* 2015;6(7):5204–16.
- [55] Umapathy G, et al. The kinase ALK stimulates the kinase ERK5 to promote the expression of the oncogene MYCN in neuroblastoma. *Sci Signal* 2014;7(349):ra102.
- [56] Lasorsa VA, et al. Exome and deep sequencing of clinically aggressive neuroblastoma reveal somatic mutations that affect key pathways involved in cancer progression. *Oncotarget* 2016;7(16):21840–52.
- [57] Sausen M, et al. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet* 2013;45(1):12–7.
- [58] Valentijn LJ, et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat Genet* 2015;47(12):1411–4.
- [59] Telomeres Mendelian Randomization C, et al. Association between telomere length and risk of cancer and non-neoplastic diseases: a mendelian randomization study. *JAMA Oncol* 2017;3(5):636–51.
- [60] Teitz T, et al. Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. *Nat Med* 2000;6(5):529–35.
- [61] Astuti D, et al. RASSF1A promoter region CpG island hypermethylation in phaeochromocytomas and neuroblastoma tumours. *Oncogene* 2001;20(51):7573–7.
- [62] Lazcoz P, et al. Frequent promoter hypermethylation of RASSF1A and CASP8 in neuroblastoma. *BMC Canc* 2006;6:254.
- [63] Yang Q, et al. Association of epigenetic inactivation of RASSF1A with poor outcome in human neuroblastoma. *Clin Canc Res* 2004;10(24):8493–500.
- [64] Abe M, et al. CpG island methylator phenotype is a strong determinant of poor prognosis in neuroblastomas. *Canc Res* 2005;65(3):828–34.
- [65] Henrich KO, et al. Integrative genome-scale analysis identifies epigenetic mechanisms of transcriptional deregulation in unfavorable neuroblastomas. *Canc Res* 2016;76(18):5523–37.
- [66] Liberman J, et al. Involvement of the CXCR7/CXCR4/CXCL12 axis in the malignant progression of human neuroblastoma. *PLoS One* 2012;7(8):e43665.
- [67] Cimmino F, et al. Galectin-1 is a major effector of TrkB-mediated neuroblastoma aggressiveness. *Oncogene* 2009;28(19):2015–23.

- [68] Hossain S, et al. NLRR1 enhances EGF-mediated MYCN induction in neuroblastoma and accelerates tumor growth in vivo. *Canc Res* 2012;72(17):4587–96.
- [69] Hogarty MD, et al. ODC1 is a critical determinant of MYCN oncogenesis and a therapeutic target in neuroblastoma. *Canc Res* 2008;68(23):9735–45.
- [70] Valsesia-Wittmann S, et al. Oncogenic cooperation between H-Twist and N-Myc overrides failsafe programs in cancer cells. *Canc Cell* 2004;6(6):625–30.
- [71] Hudlebusch HR, et al. MMSET is highly expressed and associated with aggressiveness in neuroblastoma. *Canc Res* 2011;71(12):4226–35.
- [72] Decock A, et al. Neuroblastoma epigenetics: from candidate gene approaches to genome-wide screenings. *Epigenetics* 2011;6(8):962–70.
- [73] Koyama H, et al. Mechanisms of CHD5 inactivation in neuroblastomas. *Clin Canc Res* 2012;18(6):1588–97.
- [74] Lau DT, et al. Prognostic significance of promoter DNA methylation in patients with childhood neuroblastoma. *Clin Canc Res* 2012;18(20):5690–700.
- [75] Ritchie MD, et al. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 2015;16(2):85–97.
- [76] Fridley BL, et al. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet Epidemiol* 2012;36(4):352–9.
- [77] Kim D, et al. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inf* 2012;45(6):1191–8.
- [78] Holzinger ER, et al. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* 2014;30(5):698–705.
- [79] Salazar BM, et al. Neuroblastoma, a paradigm for big data science in pediatric oncology. *Int J Mol Sci* 2016;18(1).
- [80] Kanehisa M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;44(D1):D457–62.
- [81] Chatr-Aryamontri A, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2015;43(Database issue):D470–8.
- [82] Szklarczyk D, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;43(Database issue):D447–52.
- [83] Szklarczyk D, Jensen LJ. Protein-protein interaction databases. *Meth Mol Biol* 2015;1278:39–56.
- [84] Brohee S, et al. Network Analysis Tools: from biological networks to clusters and pathways. *Nat Protoc* 2008;3(10):1616–29.
- [85] Kramer A, et al. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 2014;30(4):523–30.
- [86] da Rocha EL, et al. NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Res* 2016;44(10):e100.
- [87] Garcia-Alcalde F, et al. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* 2011;27(1):137–9.
- [88] Wang LY, et al. EpiRegNet: constructing epigenetic regulatory network from high throughput gene expression data for humans. *Epigenetics* 2011;6(12):1505–12.
- [89] Marbach D, et al. Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;9(8):796–804.
- [90] Margolin AA, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf* 2006;7(Suppl. 1):S7.
- [91] McKinney-Freeman S, et al. The transcriptional landscape of hematopoietic stem cell ontogeny. *Cell Stem Cell* 2012;11(5):701–14.

- [92] Altrock PM, Liu LL, Michor F. The mathematics of cancer: integrating quantitative models. *Nat Rev Canc* 2015;15(12):730–45.
- [93] Kolch W, et al. The dynamic control of signal transduction networks in cancer cells. *Nat Rev Canc* 2015;15(9):515–27.
- [94] Li H, et al. Pathway sensitivity analysis for detecting pro-proliferation activities of oncogenes and tumor suppressors of epidermal growth factor receptor-extracellular signal-regulated protein kinase pathway at altered protein levels. *Cancer* 2009;115(18):4246–63.
- [95] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;16(6):321–32.
- [96] Tarca AL, et al. Machine learning and its applications to biology. *PLoS Comput Biol* 2007;3(6):e116.
- [97] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–44.
- [98] Alipanahi B, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33(8):831–8.
- [99] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12(10):931–4.
- [100] Xiong HY, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 2015;347(6218):1254806.
- [101] Ma J, et al. Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model* 2015;55(2):263–74.
- [102] Mnih V, et al. Human-level control through deep reinforcement learning. *Nature* 2015;518(7540):529–33.
- [103] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;23(19):2507–17.
- [104] Cahan P, et al. CellNet: network biology applied to stem cell engineering. *Cell* 2014;158(4):903–15.
- [105] Morris SA, et al. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* 2014;158(4):889–902.

COMPUTATIONAL ANALYSIS OF EPIGENETIC MODIFICATIONS IN MELANOMA

20

Ming Tang, Kunal Rai

Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, United States

INTRODUCTION

Epigenetic information is coded in the form of modifications on our genetic material. DNA bases and histone proteins are heavily modified with various chemical moieties (such as methylation, acetylation, ubiquitination, SUMOylation, and phosphorylation). Some of these modifications are laid in specific manner on certain kinds of epigenetic elements and therefore can act as markers for identification of such elements in a genome-wide manner. For example, acetylation on histone H3 at lysine 27 marks active enhancers or promoters [1,2]. Since the advent of genomic technologies in the past decade, such as DNA microarrays and next-generation sequencing, a number of approaches have been developed to map these modified bases or histones at the genome-wide scale. These studies have yielded wealth of epigenomic data in various biological contexts including various solid and hematological malignancies. This chapter focuses on epigenomic alterations in cutaneous melanoma, the most aggressive form of skin cancer. We briefly summarize the current knowledge on alterations in various epigenetic processes, mention the methods for genome-wide analyses of these processes, and provide a brief description of computational analyses of most popular methods (Table 20.1).

Melanoma is a highly aggressive disease, which primarily arises from melanocytes present in the skin. Rate of melanoma occurrence is increasing every decade with an alarming rate [3]. As such, melanoma represents 5% and 4% of cancer cases in men and women, respectively. In 2017, an estimated 87,100 new cases of melanoma will be diagnosed and approximately 9730 people will die of this disease [4]. Treatment strategies for melanoma patients have been rapidly improved with the advent of immune checkpoint blockade agents and adaptive T-cell therapies, which have been approved by FDA in the past 2–3 years [5,6]. These therapies have provided remarkably durable responses in melanoma. However, response rates remain low [6–11]. Therefore, a major unmet need in melanoma therapy is to identify biomarkers of responses to immunotherapies. Epigenetic marks could be potentially used as biomarkers. Hence, there is a need for comprehensive understanding of epigenomic aberrations in melanoma, especially those associated with response to immunotherapies.

Melanoma arises from melanocytes that produce melanin for absorption of reactive oxygen species (ROS) derived from cellular response to ultraviolet (UV) radiation [12]. Due to its ability to induce cyclobutane pyrimidine dimers, UV irradiation causes a large number of somatic mutations in the

Table 20.1 Available Computational Tools for Analysis of Epigenomic Data Sets

Epigenomic Elements	Methods	Tools	References
DNA methylation (5mC)	Infinium HM450K beadchip array	Minfi, ChAMP (R)	[117,118]
DNA methylation (5mC)	RRBS	Methylkit, BiSeq (R)	[36,38]
DNA methylation (5mC)	WGBS	Bsseq (R)	[119]
DNA methylation (5mC)	MeDIP-Seq	MEDIPS (R)	[120]
DNA methylation (5hmC)	OxBS-Seq	oxBS-MLE, MLML	[121,122]
DNA methylation (5hmC)	TAB-Seq	MLML	[122]
Histone modification	ChIP-Seq	MACS, HOMER	[75,123]
Transcription factor binding	ChIP-Seq, ChIP-exo	MACS, MACE	[75,124]
Chromatin accessibility	DNaseI-Seq	F-Seq, HotSpot	[125,126]
Chromatin accessibility	MNase-Seq	DANPOS	[127]
Chromatin accessibility	FAIRE-Seq	HOMER, ZINBA, F-Seq	[123,125,128]
Chromatin accessibility	ATAC-Seq	MACS, nucleoATAC	[75,110]
High-order chromatin structure	Hi-C	HiC-pro, HiCCUPS, HOMER, diffHic	[97,123,129,130]
High-order chromatin structure	Capture Hi-C	CHiCAGO	[131]
High-order chromatin structure	4C	w4CSeq, FourCSeq	[132,133]
High-order chromatin structure	5C	HiFive	[134]
Long-range chromatin interaction	ChIA-PET	Mango, CHIA-PET2	[135,136]
Long-range chromatin interaction	HiChIP	HiC-pro	[130]

DNA [13]. Hence, melanoma is typified by a large number of somatic mutations that harbor UV signature [14]. This poses challenges in interpretation of epigenomic aberrations observed in melanoma as well as in determining their functional roles.

DNA MODIFICATIONS

5-Methylcytosine

5-Methylcytosine is the most abundant DNA base modification present in the eukaryotic cells that is associated with gene repression. The distribution and aberrations in 5-methylcytosine have been studied extensively during normal development and in cancer including melanoma (reviewed in Refs. [15,16]). Historically, it was found that cancers display hypomethylation at the global level and hypermethylation on promoters of specific tumor suppressor genes (leading to their silencing) [17]. In melanoma, pregenomic era studies identified several important genes' promoters to be hypermethylated including *CDKN2A/CDKN2B* (cyclin dependent kinase inhibitor 2A/2B), *PTEN* (phosphatase and tensin homolog), *MGMT* (O-6-methylguanine-DNA methyltransferase), *RASSF1A* (Ras association domain family member 1), *RAR-β2* (retinoic acid receptor beta 2), *TBC1D16* (TBC1 domain family member 16), and *FES* (tyrosine protein kinase FES) [18–24]. The largest study

defining 5-methylcytosine patterns in melanoma is the TCGA (The Cancer Genome Atlas Group) study, which profiled 333 melanoma samples using Infinium Human Methylation 450K beadchip array [14]. Clustering analysis of differentially methylated probes in the TCGA study identified four subgroups of melanoma samples: “normal” like, hypomethylated, hypermethylated, and super-hypermethylated CIMP (CpG-island methylator phenotype) clusters [14]. Here, CIMP cluster showed marginally significant enrichment of *IDH1/2* (isocitrate dehydrogenase 1/2) and *ARID2* (AT-rich interaction domain 2) mutations. Another recent study compared DNA methylation profiles derived from MIRA-Seq of 27 melanoma tumors with normal melanocytes and identified *KIT* (KIT proto-oncogene tyrosine kinase receptor), *PAX3* (paired box 3), and *SOX10* (SRY-box 10) as hypermethylated and downregulated genes [25]. Finally, another study followed methylation patterns in premalignant nevi ($n = 14$), primary tumors ($n = 33$), and metastatic tumors ($n = 28$) using Human-Methylation450 BeadChip array, leading to identification of some other developmental genes such as *HoxA9* (homeobox-A9) as well as potential biomarkers in methylation of *PON3* (paraoxonase 3) and *OVOL1* (ovo-like transcriptional repressor 1) [26]. For detailed information on aberrant methylation in melanoma, please refer to the review article by Micevic et al. [15].

Most popular methods to generate DNA methylation profiles remain Infinium HM450K beadchip array (replaced with Infinium Methylation EPIC 850K beadchip array), MeDIP-Seq (methylated DNA immunoprecipitation followed by sequencing), RRBS (reduced representation bisulfite sequencing), and WGBS (whole genome bisulfite sequencing) depending on the need of depth and economic considerations. Infinium HM450K array is a probe-based technology that covers about 450,000 CpGs that are present in the promoter region 5'UTR, CpG island, CpG shores, CpG shelves, first exon, gene body, and 3'UTR, providing a comprehensive view of methylation on 99% of RefSeq genes. MeDIP-Seq is dependent on immunoprecipitation of methylated DNA using 5-methylcytosine specific antibody followed by next-generation sequencing [27]. The gold standard however, for determining the DNA methylome remains bisulfite-based. In essence, WGBS is whole-genome resequencing, preceded by treatment of genomic DNA with sodium bisulfite [28,29]. Excluding repetitive regions, this technique is capable of determining the state of virtually all cytosines in the genome. But given the costs and bioinformatic challenges, WGBS is used less than other methods. Therefore, we focus here on RRBS, the “reduced” form of WGBS, which is low cost, yet single-base resolution, at the expense of genome coverage.

Reduced Representation Bisulfite Sequencing (RRBS)

RRBS is a cost-efficient method for genome-wide DNA methylation profiling [30,31]. Genomic DNA is first digested by a methylation-insensitive restriction enzyme (e.g., *BglII*, *MspI*) and size selected to produce a small subset of the genomic DNA enriched for CpG sites in most of the promoters and CpG islands. Bisulfite conversion is performed and sequencing library is constructed subsequently.

To process RRBS data, the raw sequencing fastq files undergo quality control first. Low quality bases and adapters are trimmed off by tools such as *Trim Galore* [32]. Lambda spike-in DNA is used as a control for bisulfite conversion rate. The lambda DNA used in the spike-in is unmethylated. If the bisulfite conversion is efficient, high percentage of cytosines (C) in the lambda DNA should be converted to thymine (T). Then, quality controlled fastq reads are aligned to reference genome using aligners such as *Bismark* [33], *BSMAP* [34], and *BWA-meth* [35]. *Bismark* or *MethylDackel* can be used to extract the methylation calls from the aligned bam files.

For each CpG site, the number of reads supporting the C (methylated) and the number of reads supporting the T are extracted from the bam files. Methylation level of each CpG is calculated as the number of C divided by the total number of reads covering that site. This ratio is also known as the beta value. Differentially methylated loci are then identified by statistical tests assuming the methylation level of a CpG site follows a beta distribution [36,37]. Downstream differential methylation analysis can be carried out by packages such as *methylKit* [38], *Biseq* [36], and *DMAP* [39]. For a full list of packages, one can refer to this review [40].

So far, not many studies have used RRBS in studying global DNA methylation changes in melanoma. A recent study used RRBS to profile paired primary and metastatic melanoma cell lines identified *EBF3* (early B cell factor 3) to be hypermethylated in the metastatic cell lines [41]. Unexpectedly, hypermethylation of *EBF3* promoter associated with increased gene expression, which is somewhat contradictory to what is known for promoter methylation. Given that RRBS is cost effective, more studies can be carried out using it to study global DNA methylation changes in melanoma. Moreover, RRBS profiling on the tumor samples rather than the cell lines will give more insight on how DNA methylation contributes to melanoma tumorigenesis. However, the computational challenge of analyzing tumor RRBS data is that tumor samples contain a lot of microenvironment cells. Teasing apart the normal stromal cell contribution in DNA methylation is critical. One algorithm, *MethylPurify* [42], has been developed to deconvolute the normal cell composite and analyze differential methylation for RRBS data and WGBS data.

5-Hydroxymethylcytosine

5-Hydroxymethylcytosine is a recently discovered (in 2009) modification of DNA that is derived from oxidation of 5-methylcytosine by TET (Tet methylcytosine dioxygenase) enzymes [43]. It is subsequently converted to 5-formylcytosine and 5-carboxylcytosine during the consecutive oxidation processes [44,45]. We have limited information on distribution and aberrations in these marks in melanoma. A prominent study in melanoma showed that 5hmC levels are reduced during transition to metastasis, consistent with downregulation of TET2 enzyme [46]. They found strong correlation between loss of 5hmC and poor prognosis of patients, thereby suggesting 5hmC levels as a potential biomarker in melanoma. Further studies are required to determine the patterns of 5hmC, 5fmc, and 5camC in melanoma tumors to determine heterogeneity among patients and global patterns for their downstream regulatory pathways. In this space, multiple methodologies have been recently developed including OxBS-Seq (oxidative bisulfite sequencing), TAB-Seq, and hMeDIP-Seq (5-hydroxymethylated DNA immunoprecipitation followed by sequencing) (reviewed in Ref. [47]).

HISTONE MODIFICATIONS AND CHROMATIN STATES

DNA wraps on histone proteins to form nucleosomes, which are the fundamental units of chromatin [48,49]. Nucleosome is composed of an octamer of the four core histones (H3, H4, H2A, H2B) [50,51]. The N-terminals of the histones are subject to various posttranslational modifications including acetylation, methylation, and phosphorylation [52,53]. A specific combination of histone modifications is termed “histone code”, which correlates with the gene expression pattern. For example, histone H3 lysine 27 acetylation (H3K27ac) is usually found at active promoters or enhancers while H3K27me3 is usually located at the repressive promoters [54,55].

Recently, alterations in enhancers have been shown in multiple malignancies by H3K27ac profiles [56]. However, more than 100 epigenetic modifications have been identified [57,58] without clear understanding of their biological roles and interdependence. Furthermore, there are an even larger number of possible combinatorial patterns of these histone and DNA modifications, and it is these combinatorial patterns—not individual modifications—that dictate epigenetic states [53]. Hence there is tremendous need to identify alterations in these chromatin states during cancer progression. Comprehensive knowledge of epigenome alterations in cancers has been lagging in part due to technical (e.g., generation of large-scale data), analytical (e.g., algorithms to define combinatorial states), and biological (e.g., lack of “germline normal” equivalence) challenges. However, with the recent development of high-throughput ChIP-sequencing methods, computational approaches to predict combinatorial patterns, and a surge in epigenome profiling studies [59–63], it is now possible to determine epigenetic states associated with different stages of tumorigenesis and therapeutic resistance.

Indeed, we have recently exploited these advances to define chromatin state changes associated with nontumorigenic to tumorigenic transition in melanoma using an optimized high-throughput ChIP-Seq protocol [64,65] for 35 epigenomic marks [66] and cutting-edge computational algorithms such as *ChromHMM* [67]. ChIP-Seq is short for chromatin-immunoprecipitation followed by sequencing, which is a gold standard approach to study protein and DNA interaction *in vivo*. The basic steps are as follows: First, the DNA-binding protein *in vivo* is crosslinked with formaldehyde and the chromatin is sheared into 200–600 bp short fragments. Then, the DNA–protein complex is immunoprecipitated with an antibody specific to the protein of interest. Finally, the DNA is purified and made to a library for sequencing [68,69].

Chromatin-Immunoprecipitation Followed by Sequencing (ChIP-Seq)

Analysis of ChIP-Seq data involves a series of steps of quality control and preprocessing. Briefly, the quality of the sequencing fastq reads is assessed by *fastqc*, and then the reads are aligned to the reference genome using aligners such as *bowtie* and *bwa*. *Bowtie* has two versions: *bowtie1* and *bowtie2*. *Bowtie2* is better for read length greater than 50 bp and it can tolerate indels. It is still very common to use 36 bp single-end sequencing library for ChIP-Seq, so *bowtie1* is preferred for ChIP-Seq with short reads and if one does not care about the indels. The aligned bam files can be used to call peaks. Finally, the resulting peak files and raw signal (bigwig) files are visualized in a genome browser such as IGV.

To accommodate great need for defining chromatin states in melanoma tumor samples, we have established an integrated platform for high-throughput ChIP-Seq which constitutes a wet-lab module and a computational module [70]. The computational module utilizes a processing pipeline [71] based on *snakemake* [72]. *Snakemake* is a workflow management system, an extension of python language by adding declarative code to define rules. Rules describe how to create output files from input files, which is very similar to GNU Make. *Snakemake* greatly reduces the complexity of creating workflows by providing a fast and comfortable execution environment, together with a clean and modern specification language in python style. We incorporated all the preprocessing steps in this pipeline including quality control of raw fastq reads, aligning to genome by *bowtie1* [73], assessing ChIP quality by *phantompeakqual* following ENCODE standard [74], down-sampling all samples, making RPKM normalized bigwig tracks by *deepTools*, calling peaks using both *macs1* and *macs2* [75],

generating superenhancer calls by *Rose* [76], and running *ChromHMM* [67] models. We added a *multiQC* [77] module in the pipeline, which generates an HTML output aggregating all the quality control results: *fastqc* [78], *samtools* [79] *flagstat*, and *bowtie* alignment report.

The pipeline was implemented in a way that all jobs are submitted to the computing cluster; independent jobs are run in a parallelized manner; dependent jobs start after the upstream jobs finish. In this way, computing capacity can be maximized by taking advantage of a multicore computing cluster. Moreover, there is a configuration file that can be used to fine-tune the running parameters. One can use a different genome (e.g., mouse) to align reads; set peak calling *P* value cutoffs; change the number of reads to one that all samples are downsampled to; choose different number of chromatin state for *ChromHMM*; etc. If certain files failed, rerunning the whole pipeline will only rerun the files that failed. Of course, one can force rerun any desired step for any files. Our pipeline ensures computation reproducibility as *snakemake* records exact commands that are used to generate the outputs and the versions of the tools that are used to run the commands.

Various findings using ChIP-Seq have been reported in studying melanoma progression. ChIP-Seq against two important histone modifications representing activated (H3K27ac) and repressed (H3K27me3) chromatin marks was performed, which identified *TEADs* (TEA domain transcription factor) as the regulators of the invasive state of melanoma [80]. Another study used ChIP-Seq to identify MITF (melanogenesis associated transcription factor) and SMARCA4 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 4) binding sites and showed that MITF interacts with chromatin remodeling complex comprising SMARCA4 and CHD7 (chromodomain helicase DNA binding protein 7). Laurette et al. further showed that *SMARCA4* is essential for melanoma cell proliferation in vitro and for normal melanocyte development in vivo [81]. More recently, using H3K27ac as a surrogate for superenhancers, a study showed activation of superenhancers at neural crest progenitor (NCP) genes in both zebrafish and human melanomas, identifying an epigenetic mechanism for control of this NCP signature leading to melanoma [82]. Using ChIP-Seq, *CTCF* (CCCTC-binding factor) was shown to have the reduced binding at the mutant alleles in melanoma. Topologically associating domains with mutated *CTCF* anchors contained differentially expressed cancer-related genes [83].

HIGHER-ORDER CHROMATIN STRUCTURE

Human genome DNA is ~ 2 m long if it is stretched. How a ~ 2 μm nucleus harbors this much longer DNA? Eukaryotic DNAs are not linear inside of the nucleus. The conventional model is that DNA wraps on the histone proteins to form nucleosomes with 11 nanometer (nm) in diameter [48,50]. These “beads on a string” structures are then further folded into several different scales of high-order structures until they form chromosomes. Recent studies found that chromosomes are spatially segregated into sub-megabase scale domains: topologically associating domains (TADs) [84,85]. TADs are quite conserved across species [84] and tissue types [86]. Importantly, the high-order chromatin structure has been shown to be critical in regulating gene expression and determining cell identity. Misregulation of high-order chromatin structure can cause diseases [87,88] including cancer [89]. However, this view is challenged by a recent electron microscopy tomography (EMT) with a labeling method (ChromEM) study in living cells [90], in which the authors showed that no such high-order structure was observed. Instead, DNA and nucleosomes assemble into disordered chains that have diameters between 5 and 24 nm, with different particle arrangements, densities, and structural conformations.

Chromosome Conformation Capture Based Methods

There are two ways to study higher-order chromatin structures: microscope-based [91] and genomic sequencing-based assay [92]. Here, we will focus on the genomic approaches. To study long-range chromatin interaction, several methods are developed including Hi-C [92], capture Hi-C [93], ChIA-PET [94], and HiChIP [93]. The resolution of the interaction map can range from megabase [92] to kilobase [95,96]. Architectural proteins such as CTCF and cohesin are shown to be enriched in the TAD boundaries [84,97]. Interestingly, different studies show discrepant observations on whether CTCF and cohesin are required for the TAD formation [98–102].

All these methods have a similar experimental procedure. First, chromatin is crosslinked by formaldehyde, a restriction enzyme is used to digest the chromatin, and then free-ends of the DNA are ligated in a diluted volume and the ligated DNA fragment pairs are subject to high-throughput sequencing. Interaction of the genomic sites is then identified by computational methods. The difference between ChIA-PET, HiChIP, and Hi-C is that the interaction pairs are enriched by applying an antibody to the protein of interest, which mediates the looping confirmation, while Hi-C is capturing all the interactions in the genome. Paired-end sequencing of the Hi-C library results in fastq reads files similar to other high-throughput sequencing assays. Raw reads are quality controlled and then aligned to the reference genome. Note that the paired-end reads are aligned separately because the insert size of the Hi-C ligation product can vary drastically. The data are then binned into fixed genomic interval sizes, to aggregate data and remove noise. Data are further normalized by mappability, GC content, and fragment length [103]. Lastly, the interactions and TADs are predicted by various tools [104].

So far, no studies have used Hi-C to investigate higher-order chromatin structure changes during melanoma progression. Integrating Hi-C data in a different cell type and mutation data from melanoma, a recent study showed that genomic regions display similar mutation profile if they are in close spatial proximity to late-replicating domains [105]. Hi-C and other types of long-range interaction assays in melanoma are urgently needed to investigate how high-order chromatin changes contribute to melanomagenesis.

NUCLEOSOME POSITIONING

Nucleosome packaging has a significant effect on the availability of DNA sequences to proteins such as transcription factors, which are central players in regulating gene expression. Thus, open or accessible regions of the genome are considered as the places where regulatory elements reside. Characterizing the accessible regions of the genome is critical in studying how the transcription is regulated. There are several methods that are developed to locate the accessible regions in a genome-wide scale utilizing the high-throughput sequencing techniques [106]: DNase I digestion (DNaseI-Seq), formaldehyde-assisted isolation of regulatory elements (FAIRE-Seq), and the more recent assay of transposase-accessible chromatin (ATAC-Seq) [107]. All of the assays probe the open chromatin regions harboring regulatory elements such as enhancers and promoters. For example, ENCODE project has used DNaseI-Seq to profile the accessible chromatin landscape of the human genome [108]. A complementary method micrococcal nuclease digestion (MNase-Seq) can be used to identify nucleosome positioning.

ATAC-Seq

Of all the methods, ATAC-Seq is gaining popularity due to its less laborious steps and less number of cells as starting materials. This method uses hyperactive Tn5 transposase, which inserts sequencing

adapters into accessible regions of chromatin, to detect accessible regions of the genome [107]. Sequencing reads are mapped to reference genome to infer open regions, to footprint the transcription factor motifs and to infer nucleosome positioning.

ATAC-Seq analysis is thoroughly described in Ref. [107]. Briefly, the sequencing reads are quality controlled with *fastqc* and the adapters are trimmed. Then the reads are aligned back to reference genome using *bowtie2* [109]. It is quite common to have mitochondrial sequences in the reads, so the mitochondrial reads are removed from the aligned bam files. Peaks are called using *MACS2* [75]. Nucleosome positioning is inferred by *nucleoATAC* [110]. We have developed a *snakemake* [72] based pipeline [111] to preprocess ATAC-Seq data.

Surprisingly, few studies have assayed chromatin accessibility during melanoma progression. Using DNase-Seq data from melanocyte generated by ENCODE, it is reported that DNA mutation rate in melanomas is highly increased at active transcription factor binding sites and nucleosome embedded DNA, compared to their flanking regions [112]. Genome-wide DNA accessibility data are scarce in melanoma. A recent study in zebrafish model of melanoma identified superenhancers (using ChIP-Seq) and open chromatin regions (using ATAC-Seq) near neural-crest identity defining *crestin* and *sox10* genes in melanoma tumors. It will be quite interesting to investigate the chromatin accessibility change during melanoma progression. Assaying on the primary tumors is possible because only a small number of cells are required for ATAC-Seq.

FUTURE PERSPECTIVE

Although we have begun to understand the epigenetic aberrations in the cancer, we are still in the early stages of our understanding of cancer epigenome. We need better methods to define epigenome from tumors rather than cell lines as cell lines have been cultured for a long time under artificial conditions. Further, given the heterogeneity within patient samples, we need to profile a large number of patient samples to determine the subsets where specific epigenetic therapies may be useful. Hence high-throughput versions of the existing methodologies are needed. On the other hand, to get a better understanding of intratumor heterogeneity, we need to define epigenome in tumor cells as well as those in the microenvironment. This is especially important in melanoma given the durable effects of immunotherapy. Importantly, in the light of precision medicine efforts, this needs to be done in a large set of patient samples. Here, methodologies that profile epigenome at the single cell level will be highly useful. Indeed, significant advances have been made in determining DNA methylation, higher-order chromatin structure, and ATAC-Seq profiles from single cells [113–116]. These need to be applied in melanoma. Finally, it needs to be determined if specific genetic mutations have differential effects on the epigenome as it could identify specific patient populations that may be benefited from epigenetic therapy.

REFERENCES

- [1] Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* February 2011;470(7333):279–83.
- [2] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* December 2010;107(50):21931–6.

- [3] Caini S, Gandini S, Sera F, Raimondi S, Farnol MC, Boniol M, et al. Meta-analysis of risk factors for cutaneous melanoma according to anatomical site and clinico-pathological variant. *Eur J Cancer* 2009; 45(17):3054–63.
- [4] Tas F. Metastatic behavior in melanoma: timing, pattern, survival, and influencing factors. *J Oncol* 2012; 2012:647684.
- [5] Reiss KA, Forde PM, Brahmer JR. Harnessing the power of the immune system via blockade of PD-1 and PD-L1: a promising new anticancer strategy. *Immunotherapy* 2014;6(4):459–75.
- [6] Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med* 2012;366(26):2443–54.
- [7] Brahmer JR, Tykodi SS, Chow LQ, Hwu WJ, Topalian SL, Hwu P, et al. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N Engl J Med* 2012;366(26):2455–65.
- [8] Hodi FS, O’Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* 2010;363(8):711–23.
- [9] Postow MA, Chesney J, Pavlick AC, Robert C, Grossmann K, McDermott D, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *N Engl J Med* May 21, 2015;372(21):2006–17.
- [10] Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, et al. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in metastatic or locally advanced, unresectable melanoma. *J Clin Oncol* June 10, 2015;33(17):1889–94.
- [11] Topalian SL, Sznol M, McDermott DF, Kluger HM, Carvajal RD, Sharfman WH, et al. Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J Clin Oncol* 2014;32(10):1020–30.
- [12] Lin JY, Fisher DE. Melanocyte biology and skin pigmentation. *Nature* 2007;445(7130):843–50.
- [13] Bertolotto C. Melanoma: from melanocyte to genetic alterations and clinical options. *Science* 2013;2013: 635203.
- [14] Cancer Genome Atlas N. Genomic classification of cutaneous melanoma. *Cell* 2015;161(7):1681–96.
- [15] Micevic G, Theodosakis N, Bosenberg M. Aberrant DNA methylation in melanoma: biomarker and therapeutic opportunities. *Clin Epigenet* 2017;9:34.
- [16] Baylin SB, Jones PA. A decade of exploring the cancer epigenome – biological and translational implications. *Nat Rev Cancer* 2011;11(10):726–34.
- [17] Feinberg AP. Alterations in DNA methylation in colorectal polyps and cancer. *Prog Clin Biol Res* 1988;279: 309–17.
- [18] Olvedy M, Tisserand JC, Luciani F, Boeckx B, Wouters J, Lopez S, et al. Comparative oncogenomics identifies tyrosine kinase FES as a tumor suppressor in melanoma. *J Clin Invest* 2017;127(6):2310–25.
- [19] Kohonen-Corish MR, Cooper WA, Saab J, Thompson JF, Trent RJ, Millward MJ. Promoter hypermethylation of the O(6)-methylguanine DNA methyltransferase gene and microsatellite instability in metastatic melanoma. *J Invest Dermatol* 2006;126(1):167–71.
- [20] Venza M, Visalli M, Biondo C, Lentini M, Catalano T, Teti D, et al. Epigenetic regulation of p14ARF and p16INK4A expression in cutaneous and uveal melanoma. *Biochim Biophys Acta* 2015;1849(3):247–56.
- [21] Mirmohammadsadegh A, Marini A, Nambiar S, Hassan M, Tannapfel A, Ruzicka T, et al. Epigenetic silencing of the PTEN gene in melanoma. *Cancer Res* 2006;66(13):6546–52.
- [22] Hoon DS, Spugnardi M, Kuo C, Huang SK, Morton DL, Taback B. Profiling epigenetic inactivation of tumor suppressor genes in tumors and plasma from cutaneous melanoma patients. *Oncogene* 2004;23(22): 4014–22.
- [23] Fan J, Eastham L, Varney ME, Hall A, Adkins NL, Chetel L, et al. Silencing and re-expression of retinoic acid receptor beta2 in human melanoma. *Pigment Cell Melanoma Res* 2010;23(3):419–29.
- [24] Tellez CS, Shen L, Estecio MR, Jelinek J, Gershenwald JE, Issa JP. CpG island methylation profiling in human melanoma cell lines. *Melanoma Res* 2009;19(3):146–55.

- [25] Jin SG, Xiong W, Wu X, Yang L, Pfeifer GP. The DNA methylation landscape of human melanoma. *Genomics* 2015;106(6):322–30.
- [26] Wouters J, Vizoso M, Martinez-Cardus A, Carmona FJ, Govaere O, Laguna T, et al. Comprehensive DNA methylation study identifies novel progression-related and prognostic markers for cutaneous melanoma. *BMC Med* 2017;15(1):101.
- [27] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 2008;26(7):779–85.
- [28] Lister R, O’Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 2008;133(3):523–36.
- [29] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462(7271):315–22.
- [30] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* October 12, 2005;33(18):5868–77. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki901>.
- [31] Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods* February 10, 2010;7(2):133–6. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.1414>. Nature Publishing Group.
- [32] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* May 2, 2011;17(1):10. Available from: <http://journal.embnet.org/index.php/embnetjournal/article/view/200>.
- [33] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* June 1, 2011;27(11):1571–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21493656>.
- [34] Xi Y, Li WBSMAP. Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* July 27, 2009;10(1):232. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19635165>.
- [35] Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. January 6, 2014. Available from: <http://arxiv.org/abs/1401.1129>.
- [36] Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* July 1, 2013;29(13):1647–53. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23658421>.
- [37] Feng H, Conneely KN, Wu HA. Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* April 2014;42(8):e69. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24561809>.
- [38] Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* October 3, 2012;13(10):R87. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23034086>.
- [39] Stockwell PA, Chatterjee A, Rodger EJ, Morison IM. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* July 1, 2014;30(13):1814–22. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu126>. Oxford University Press.
- [40] Wang F, Zhang N, Wang J, Wu H, Zheng X. Tumor purity and differential methylation in cancer epigenomics. *Brief Funct Genomics* May 19, 2016;15(6):elw016. Available from: <https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elw016>. Oxford University Press.
- [41] Chatterjee A, Stockwell PA, Ahn A, Rodger EJ, Leichter AL, Eccles MR. Genome-wide methylation sequencing of paired primary and metastatic cell lines identifies common DNA methylation changes and a role for EBF3 as a candidate epigenetic driver of melanoma metastasis. *Oncotarget* January 24, 2017;8(4):6085–101. Available from: <http://www.oncotarget.com/abstract/14042>.

- [42] Zheng X, Zhao Q, Wu H-J, Li W, Wang H, Meyer CA, et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol* August 7, 2014; 15(7):419. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0419-x>. BioMed Central.
- [43] Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* (80-) 2009;324(5929): 930–5.
- [44] Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* (80-) 2011;333(6047):1300–3.
- [45] He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* (80-) 2011;333(6047):1303–7.
- [46] Lian CG, Xu Y, Ceol C, Wu F, Larson A, Dresser K, et al. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* 2012;150(6):1135–46.
- [47] Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;15(10):647–61.
- [48] Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science* (80-) 1974;184(4139): 868–71.
- [49] Olins AL, Olins DE. Spheroid chromatin units (v bodies). *Science* (80-) 1974;183(4122):330–2.
- [50] Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 1997;389:251–60.
- [51] Kornberg RD, Thomas JO. Chromatin structure; oligomers of the histones. *Science* (80-) 1974;184(4139): 865–8.
- [52] Kouzarides T. Chromatin modifications and their function. *Cell* 2007;128(4):693–705.
- [53] Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* January 6, 2000;403(6765): 41–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10638745>.
- [54] Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell* May 18, 2007;129(4):823–37. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17512414>.
- [55] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* August 2, 2007;448(7153):553–60. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17603471>.
- [56] Herz H-M, Hu D, Shilatifard A. Enhancer Malfunction in cancer. *Mol Cell* March 2014;53(6):859–66. Available from: <http://www.sciencedirect.com/science/article/pii/S1097276514002056>.
- [57] Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* August 2011;21(8):1273–83. Available from: <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=3149494&tool=pmcentrez&rendertype=abstract>.
- [58] Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* September 16, 2011;146(6):1016–28. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21925322>.
- [59] Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* May 5, 2011;473(7345):43–9. Available from: <http://www.nature.com/doifinder/10.1038/nature09906>. Nature Research.
- [60] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature* September 6, 2012;489(7414):57–74. Available from: <https://doi.org/10.1038/nature11247>. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

- [61] Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* August 15, 2013;154(4):888–903. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23953118>.
- [62] Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* May 23, 2013;153(5):1134–48. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23664764>.
- [63] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. *Nature* February 18, 2015;518(7539):317–30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25693563>.
- [64] Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat Protoc* 2013;8(3):539–54.
- [65] Cheng CS, Rai K, Garber M, Hollinger A, Robbins D, Anderson S, et al. Semiconductor-based DNA sequencing of histone modification states. *Nat Commun* 2013;4:2672. Available from: <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=3917140&tool=pmcentrez&rendertype=abstract>.
- [66] Fiziev P, Akdemir KC, Miller JP, Keung EZ, Samant NS, Sharma S, et al. Systematic epigenomic analysis reveals chromatin states associated with melanoma progression. *Cell Rep* April 25, 2017;19(4):875–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28445736>. Elsevier.
- [67] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* February 28, 2012;9(3):215–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22373907>. NIH Public Access.
- [68] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10:669–80.
- [69] Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* December 2012;13(12):840–52. Available from: <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=3591838&tool=pmcentrez&rendertype=abstract>. Nature Publishing Group.
- [70] Terranova C, Tang M, Orouji E, Maitituoheti M, Raman AT, Amin SB, et al. An integrated platform for genome-wide mapping of chromatin states using high-throughput ChIP-sequencing in tumor tissues. *J Vis Exp* April 5, 2018;(134). Available from: <https://www.jove.com/video/56972/an-integrated-platform-for-genome-wide-mapping-chromatin-states-using>.
- [71] Tang M. pyflow-ChIPSeq: a snakemake based ChIP-seq pipeline. Zenodo; 2017. <http://doi.org/10.5281/zenodo.819971>.
- [72] Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* October 1, 2012;28(19):2520–2. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480>. Oxford University Press.
- [73] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
- [74] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* September 1, 2012;22(9):1813–31. Available from: <http://genome.cshlp.org/content/22/9/1813.full>.
- [75] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol* 2008;9:R137.
- [76] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* April 11, 2013;153(2):307–19. Available from: <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=3653129&tool=pmcentrez&rendertype=abstract>.

- [77] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* October 1, 2016;32(19):3047–8. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354>. Oxford University Press.
- [78] Babraham Bioinformatics – FastQC. A Quality control tool for high throughput sequence data [Internet]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [79] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
- [80] Verfaillie A, Imrichova H, Atak ZK, Dewaele M, Rambow F, Hulselmans G, et al. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* April 9, 2015; 6:6683. Available from: <http://www.nature.com/doifinder/10.1038/ncomms7683>. Nature Publishing Group.
- [81] Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, et al. Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells. *eLife Sciences Publications, Ltd.*; March 24, 2015. p. 4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25803486>.
- [82] Kaufman CK, Mosimann C, Fan ZP, Yang S, Thomas AJ, Ablain J, et al. A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* January 29, 2016; 351(6272):aad2197. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26823433>. American Association for the Advancement of Science.
- [83] Poulos RC, Thoms JAI, Guan YF, Unnikrishnan A, Pimanda JE, Wong JWH. Functional mutations form at CTCF-cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif. *Cell Rep* December 13, 2016;17(11):2865–72. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27974201>.
- [84] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* April 11, 2012;485(7398):376–80. Available from: <http://www.nature.com/doifinder/10.1038/nature11082>. Nature Research.
- [85] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* February 3, 2012;148(3):458–72. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22265598>. Elsevier.
- [86] Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* November 2016;17(8):2042–59. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S2211124716314814>.
- [87] Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* May 21, 2015;161(5): 1012–25. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25959774>. Elsevier.
- [88] Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* October 5, 2016;538(7624): 265–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27706140>.
- [89] Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* December 23, 2015;529(7584): 110–4. Available from: <http://www.nature.com/doifinder/10.1038/nature16490>. Nature Research.
- [90] Ou HD, Phan S, Deerinck TJ, Thor A, Ellisman MH, O’Shea CC. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* July 28, 2017;357(6349):eaag0025. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28751582>. American Association for the Advancement of Science.

- [91] Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 2017;543(7646):519–24.
- [92] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* October 9, 2009;326(5950):289–93. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2858594&tool=pmcentrez&rendertype=abstract>.
- [93] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* May 4, 2015;47(6):598–606. Available from: <http://www.nature.com/doifinder/10.1038/ng.3286>. Nature Publishing Group.
- [94] Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Bin MY, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009;462:58–64.
- [95] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* October 20, 2013;503(7475):290. Available from: <http://www.nature.com/doifinder/10.1038/nature12644>. Nature Research.
- [96] Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* September 19, 2016;13(11):919–22. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.3999>. Nature Research.
- [97] Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* December 18, 2014;159(7):1665–80. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25497547>.
- [98] Heidari N, Phanstiel DH, He C, Grubert F, Jahanbanian F, Kasowski M, et al. Genome-wide map of regulatory interactions in the human genome. *Genome Res* September 16, 2014;24(12):1905–17. Available from: <http://genome.cshlp.org/content/early/2014/09/15/gr.176586.114?top=1>.
- [99] Kubo N, Ishii H, Gorkin D, Meitinger F, Xiong X, Fang R, et al. Preservation of chromatin organization after acute loss of CTCF in mouse embryonic stem cells. *Org* March 20, 2017:118737. Available from: <https://www.biorxiv.org/content/early/2017/03/20/118737>. Cold Spring Harbor Laboratory.
- [100] Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* May 18, 2017;169(5):930–44. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28525758>. Elsevier.
- [101] Rodriguez-Carballo E, Lopez-Delisle L, Zhan Y, Fabre P, Beccari L, El-Idrissi I, et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Org* September 28, 2017:193706. Available from: <https://www.biorxiv.org/content/early/2017/09/28/193706>. Cold Spring Harbor Laboratory.
- [102] Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* September 27, 2017;551(7678):51–6. Available from: <http://www.nature.com/doifinder/10.1038/nature24281>. Nature Research.
- [103] Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* January 15, 2015;72:65–75. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25448293>. NIH Public Access.
- [104] Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat Methods* June 12, 2017;14(7):679–85. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.4325>. Nature Research.
- [105] Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* February 19, 2013;4:1502. Available from: <http://www.nature.com/doifinder/10.1038/ncomms2502>. Nature Publishing Group.
- [106] Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. *Nat Rev Genet* July 12, 2011;12(8):554–64. Available from: <http://www.nature.com/doifinder/10.1038/nrg3017>. Nature Publishing Group.

- [107] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* January 5, 2015;109:21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25559105>. NIH Public Access.
- [108] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature* September 6, 2012;489(7414):75–82. Available from: <http://www.ncbi.nlm.nih.gov/article/3721348&tool=pmcentrez&rendertype=abstract>.
- [109] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;357–9.
- [110] Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* November 1, 2015;25(11):1757–70. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26314830>. Cold Spring Harbor Laboratory Press.
- [111] Tang M. pyflow-ATACseq: a snakemake based ATAC-seq pipeline. Zenodo 2017. <http://doi.org/10.5281/zenodo.1043588>.
- [112] Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* April 13, 2016;532(7598):264–7. Available from: <http://www.nature.com/doifinder/10.1038/nature17661>. Nature Publishing Group.
- [113] Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, Tanay A, et al. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat Protoc* 2015;10(12):1986–2003.
- [114] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;523(7561):486–90.
- [115] Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014;11(8):817–20.
- [116] Kelsey G, Stegle O, Reik W. Single-cell epigenomics: Recording the past and predicting the future. *Science* (80-) 2017;358(6359):69–75.
- [117] Morris TJ, Beck S. Analysis pipelines and packages for infinium HumanMethylation450 BeadChip (450k) data. *Methods* January 15, 2015;72:3–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25233806>. Elsevier.
- [118] Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k Chip analysis methylation pipeline. *Bioinformatics* February 1, 2014;30(3):428–30. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt684>. Oxford University Press.
- [119] Hansen KD, Langmead B, Irizarry RA. BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* October 3, 2012;13(10):R83. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-10-r83>. BioMed Central.
- [120] Lienhard M, Grimm C, Morkel M, Herwig R, Chavez LMEDIPS. genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* January 15, 2014;30(2):284–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24227674>. Oxford University Press.
- [121] Xu Z, Taylor JA, Leung Y-K, Ho S-M, Niu L. oxBS-MLE: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. *Bioinformatics* August 13, 2016;32(23):btw527. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27522082>.
- [122] Qu J, Zhou M, Song Q, Hong EE, Smith AD. MMLM: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* October 15, 2013;29(20):2645–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23969133>.
- [123] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors Prime cis-regulatory elements required for Macrophage and B Cell Identities. *Mol Cell* May 28, 2010;38(4):576–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20513432>.

- [124] Wang L, Chen J, Wang C, Uusküla-Reimand L, Chen K, Medina-Rivera A, et al. MACE: model based analysis of ChIP-exo. *Nucleic Acids Res* November 10, 2014;42(20):e156. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25249628>.
- [125] Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* November 1, 2008;24(21):2537–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18784119>.
- [126] John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* March 23, 2011;43(3):264–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21258342>.
- [127] Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, et al. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* February 1, 2013;23(2):341–51. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23193179>. Cold Spring Harbor Laboratory Press.
- [128] Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* July 25, 2011;12(7):R67. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-7-r67>. BioMed Central.
- [129] Lun ATL, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* December 19, 2015;16(1):258. Available from: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0683-0>. BioMed Central.
- [130] Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* December 1, 2015;16(1):259. Available from: <http://genomebiology.com/2015/16/1/259>. BioMed Central.
- [131] Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* December 15, 2016;17(1):127. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0992-2>. BioMed Central.
- [132] Cai M, Gao F, Lu W, Wang K. w4CSeq: software and web application to analyze 4C-seq data. *Bioinformatics* November 1, 2016;32(21):3333–5. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw408>. Oxford University Press.
- [133] Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EEM, Huber W. FourCSeq: analysis of 4C sequencing data. *Bioinformatics* October 1, 2015;31(19):3085–91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26034064>. Oxford University Press.
- [134] Sauria ME, Phillips-Cremins JE, Corces VG, Taylor J. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol* December 24, 2015;16(1):237. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0806-y>. BioMed Central.
- [135] Phanstiel DH, Boyle AP, Heidari N, Snyder MP. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* October 1, 2015;31(19):3092–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26034063>.
- [136] Li G, Chen Y, Snyder MP, Zhang MQ. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res* January 9, 2017;45(1):e4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27625391>. Oxford University Press.

DNA METHYLOME OF ENDOMETRIAL CANCER

21

Golnaz Asaadi Tehrani

Molecular Genetics, Department of Genetics, Zanjan Branch, Islamic Azad University, Zanjan, Iran

INTRODUCTION

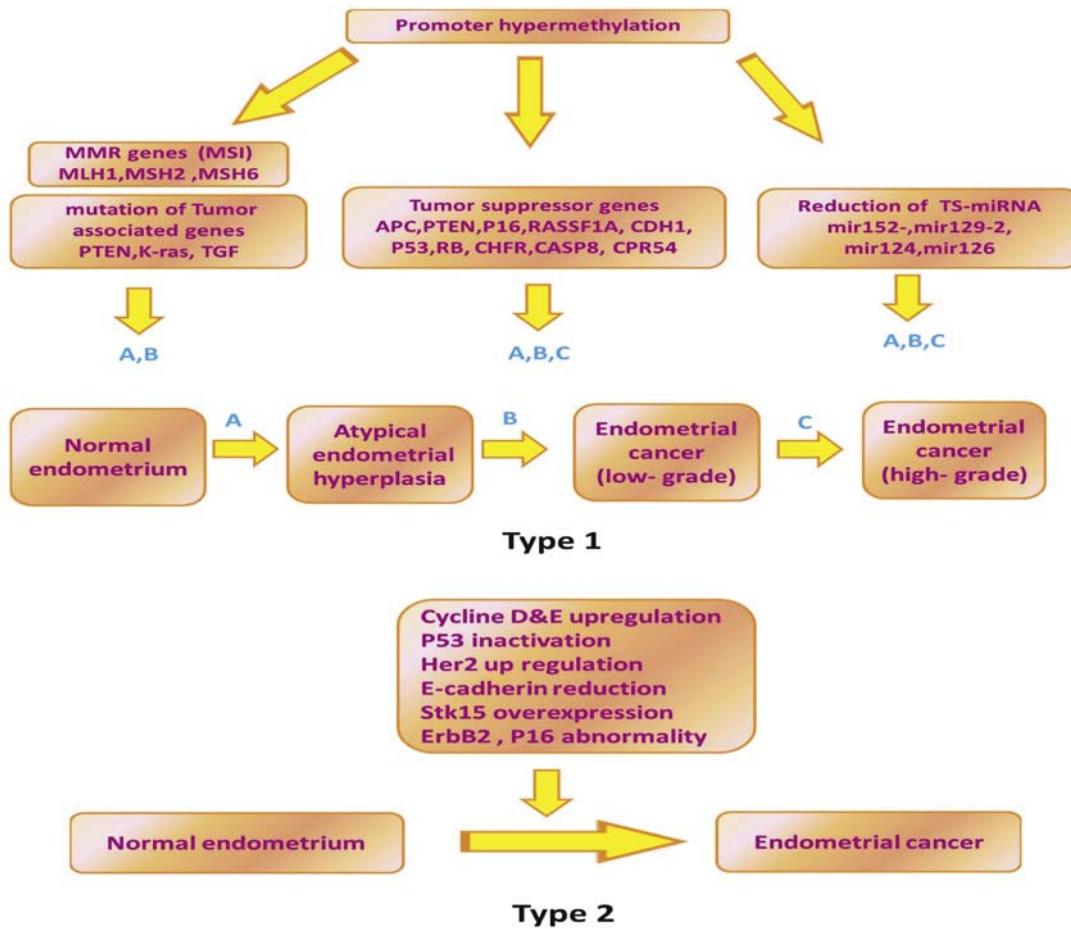
Endometrial Cancer (EC) is the most common gynecological tumor in developed countries and the fourth most prevalent cancer among women after breast, colon, and lung cancers, with 42,160 new cases and 7780 deaths occurring in the United States in 2009, with an approximate of 7400 die from the disease. It is the second most common malignancy in female with hereditary nonpolyposis colorectal cancers and fourth most common cause of death among cancer patients in Europe and North America. However, in some parts of the world, it is less common; lowest rates are reported from India and Southeast Asia as a whole [1,2].

Despite recent advances made in the treatment and diagnosis of EC, about 81,500 women are affected every year in the European Union, and the incidence is still increasing. Overall morbidity and mortality of this disease is low because most of the patients are diagnosed at early stages by abnormal bleeding; however, about 30% are identified in advanced stages. The frequency and mortality rate of EC has greatly increased in the past few years. It is associated with an approximately fourfold increase in mortality with ovarian cancer and twofold increase in mortality with breast cancer. More than 90% of cases occur in women older than 50 years, with an average age of 63 years [3,4].

The risk of EC increases with many factors, including early onset of menstruation, nulliparity, obesity, diabetes, infertility, late menopause, and menopausal estrogen therapy, all of which are associated with estrogen exposure. Most endometrial malignancies develop in endometrial gland cells and are referred to as endometrioid adenocarcinomas (61%–71%), some adenosquamous (14%–24%), and rest are papillary serous tumor and clear cell carcinoma.

Based on clinicopathological properties and molecular characteristics, endometrial carcinomas have been classified into two groups (Fig. 21.1):

- Type I (endometrioid endometrial carcinoma, EEC) occurs most often in obese postmenopausal women and occasionally in anovulatory premenopausal women. It is typically low-grade (I-II) adenocarcinomas, classically estrogen-related, and is characterized by a favorable prognosis. The lesions are commonly well differentiated, proceeded by endometrial hyperplasia, and comprise approximately 80% of sporadic tumors. In terms of molecular level, type I cancers have near-diploid karyotypes, microsatellite instability and are linked to mutations or downregulation of *PTEN*, *K-ras*, and *CTNNB1* genes [5,6].

**FIGURE 21.1**

Genetic and epigenetic alterations of endometrial cancer types 1 and 2. A, B, and C indicated progression of endometrial carcinogenesis from normal tissue, atypical hyperplasia and low-grade endometrial cancer, for each step altered genes are specified. Mismatch repair (MMR) and tumor-associated genes are involved only in step A and B alterations, but TSGs and TS-miRNAs are effective in all A, B, and C steps.

- Type II (nonendometrioid endometrial carcinoma, NEEC) tumors comprise a heterogeneous, poorly differentiated group of tumors of high-grade endometriosis; serous papillary or clear cell morphology that primarily occurs in older postmenopausal women. It may be estrogen independent and often accompanied by surrounding endometrial atrophy. Type II tumors are characterized often by *P53* mutations, *Her-2/neu* overexpression, *ErbB2*, and *P16* abnormalities, loss of heterozygosity (LOH) at several chromosomal loci and an aneuploid karyotype. These tumors are often characterized by an aggressive clinical course and poor prognosis [5–7].

MOLECULAR SIGNALING PATHWAYS OF ENDOMETRIAL CARCINOMA

Endometrial carcinogenesis is a complex process requiring the acquisition of genetic abnormalities in oncogenes, tumor suppressor genes, and DNA repair genes. Basically, abnormal cell proliferation in EC is controlled by steroid hormones and signaling pathways downstream of growth factors and their tyrosine kinase receptors. Cross-talk occurrence between them is critical for molecular and cellular functions. The most important signaling pathways have been identified to be involved in the multiple-step development of EC, including phosphatidylinositol 3-kinase/AKT serine/threonine kinase1 provided/mammalian target of rapamycin (PI3K/AKT/mTOR), mitogen-activated protein kinase (MAPK/ERK), wingless-type MMTV integration site family member (WNT/β-catenin), vascular endothelial growth factor/vascular endothelial growth factor receptor (VEGF/VEGFR), tumor protein p53/cyclin-dependent kinase inhibitor 1A (P53/P21), and cyclin-dependent kinase inhibitor 2A/RB transcriptional corepressor 1 provided (P16INK4a/pRB) [8] (Fig. 21.2):

PI3/AKT/mTOR

Predominance of PI3K/Akt/mTOR signaling in EC results from the fact that 26%–80% of sporadic endometrial carcinomas carry somatically acquired inactivating mutations and/or deletions of the *PTEN*

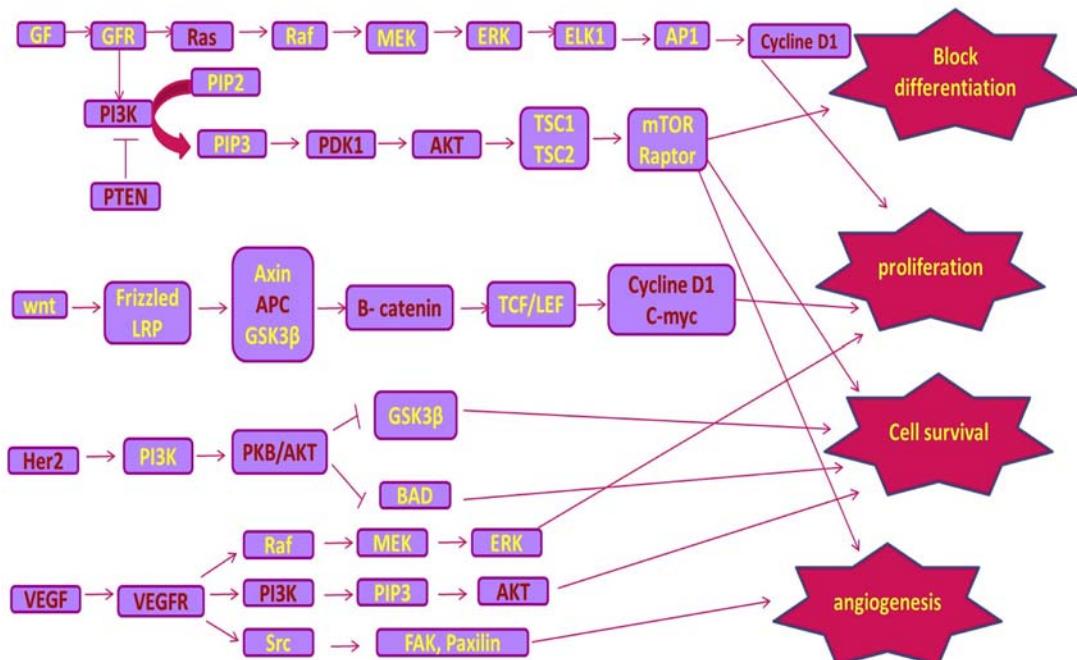


FIGURE 21.2

Signaling pathways contribute to the carcinogenesis of endometrial cancer. Altered genes are labeled in red and wild type genes are labeled in yellow.

gene. PTEN is a phosphatase that negatively regulates the PI3K/Akt signaling pathway and dephosphorylates both tyrosine phosphate and serine/threonine phosphate residues. High rate of mutations in the *PIK3CA* gene, which encodes the PI3K catalytic subunit alpha, has been reported in both types of EC. *PIK3CA* mutations were detected in 20%–30% of ECs. Amplification of the *PIK3CA* gene, however, is much more common in type II EC, with a prevalence of about 46%. Activation of AKT frequently occurs in EC but is independent to *PTEN* or *PIK3CA* activation status [9,10].

MAPK/ERK

K-ras as an oncogene has a signaling function from activated membrane receptors in MAPK pathway. Activated Ras is as a result of mutations in this gene, which cause continuous and excessive signaling, induces proliferation and carcinogenesis. K-ras mutations have been identified in the 10%–30% of endometrial carcinomas and 6%–16% of endometrial hyperplasia cases, although some investigators have reported an almost complete absence of K-ras mutations in serous and clear cell carcinomas of endometrium. Incidence of K-ras mutations has been showed to be significantly higher in tumor invasion proliferation, and incidence of mutation in grade 1 tumors was higher than that in grade 2 and 3 tumors. K-ras is involved in two stages of carcinogenesis: a shift from endometrial hyperplasia to EC and invasive proliferation of well-differentiated tumor cells. Other researchers found a higher frequency of K-ras mutations in microsatellite instability (MSI)-positive carcinomas in comparison with MSI-negative tumors [11,12].

WNT/β-CATENIN

CTNNB1 gene encodes β-catenin, a component of E-cadherins with an important role in cell adhesions. It is involved in regulation of cell proliferation and differentiation through wnt signaling pathway. β-catenin degradation is prevented by mutations, so the transcription levels of target genes will increase. These mutations are also detected in atypical endometrial hyperplasia; therefore, β-catenin mutations are implicated in the early stages of carcinogenesis. *CTNNB1* mutations have been found in 20%–40% of cases of type I EC.

During the menstrual cycle, estradiol enhances WNT/β-catenin signaling, and constitutive activation of WNT/β-catenin signaling will cause endometrial hyperplasia, which may develop further into EC. Analysis of endometrial carcinoma, nuclear β-catenin expression has been found, which is complained by low-grade EC and associated with a second carcinoma. Results of a research confirmed 31% increased expression of *CTNNB1* in EECs and 3% in NEECs. In addition, 25% of EC cases had β-catenin nuclear accumulation without mutations for this gene [13,14].

VEGF/VEGFR

VEGF and angiopoietins are key factors in angiogenesis, which are critical in spreading and metastasis of the tumors. Increased expression of VEGF-A and VEGF-B in the beginning of post-menopausal endometrium and endometrioid EC had been reported; also high expression of VEGF-D in carcinoma cells and stroma cells, are closely related with high level of VEGFR-3 in carcinoma cells and endothelial cells, suggested their significant role in myometrial invasion and lymph node metastasis [15,16].

HER-2/NEU

HER-2/neu is a tyrosine kinase membrane receptor in the epidermal growth factor (EGF) receptor family. Mutations of this gene are also found in breast and ovarian cancers. HER-2/neu expression in EC has a strong inverse correlation with differentiation. However, the incidence of gene amplification differs from 14% to 63% in all cancers, and overexpression of the protein ranges from 9% to 74% [17,18].

EPIGENETIC ALTERNATIONS IN ENDOMETRIAL CARCINOMA

One of the main problems in cancer research is apoptosis resistance of cancer cells, which is the major goal of cancer therapy. In fact, tumor cells developed numerous strategies such as genetic mutations and epigenetic modifications, which made them resistant to apoptosis. Studying of epigenetic mechanisms is a popular and expanding field in biomedical cancer researches. Epigenetic changes, including DNA methylation, covalent histone modifications, nucleosome positioning, and noncoding RNA molecules (miRNAs), are increasingly regarded as key events in the development of EC [18,19].

DNA methylation is one of the most common and best studied epigenetic modifications in mammals. In this process, methyl groups (-CH₃) from S-adenylmethionine are added to cytosine bases at CpG sites, which is performed by DNA methyltransferase (DNMT) enzymes, including Dnmt1, Dnmt2, Dnmt3a, DnMT3b, and Dnmt3L. DNMT1 maintains attachment of methyl groups to hemimethylated DNA during replication, whereas DNMT3A and DNMT3B catalyze de novo methylation of DNA. This modified base is sometimes referred to as the fifth nucleotide of the human genome because about 4% of the cytosine residues in the genome are methylated. CpG island regions usually in promoter upstream sites of the genes, rich of CpG sites (60%–70%), are the goal of methylation process and have an important role in gene expression. Many tumor suppressor genes in cancer cells are inactivated by aberrant DNA methylation in promoter CpG islands, which suggests that aberrant DNA methylation may cause carcinogenesis similar to gene mutations [20,21].

Early DNA methylation researches focused on detection and quantification of methylation status, recently using microarray hybridization technologies and next-generation sequencing platforms, allow construction of genomic maps of DNA methylation. Experimental approaches consist of the following:

ENZYME DIGESTION-BASED METHODS

Restriction enzyme: Methylation-sensitive restriction enzymes (MREs) such as BstU1, Hpa II, Not1, and SmaI cleavage only unmethylated target sequences, and methylated DNA remains intact; DNA fragments are size selected, then sequencing technologies predict genome-wide DNA methylation level.

Comparative high-throughput arrays of relative methylation (CHARM) uses McrBC enzyme, in which it recognizes RmC(N)55–103RmC sites, which cleaves half of the methylated DNAs and all the methylated CGIs methylated DNA, subsequently uses array hybridization. This method is able to detect differentially methylated regions (DMRs) at CGI shores, which are otherwise not detectable with CpG-directed enrichment methods, such as methylated DNA immunoprecipitation (MeDIP).

AFFINITY ENRICHMENT-BASED METHODS

MBD (Methyl-CpG-binding domain) proteins: This method is based on the capacity of MBD proteins that specifically bind to methylated DNA sequences and could be profiled by MBD chips or MBDCap-seq.

MeDIP: Using an antimethylcytosine, antibody would immunoprecipitate DNA with methylated CpG sites. Subsequently, DNA fractions enriched by MeDIP will be evaluated by MeDIP-chip or MeDIP-seq.

BISULFITE CONVERSION-BASED METHODS

Microarray, whole-genome bisulfite sequencing (WGBS): In this method, after DNA extraction and fragmentation, adenine (A) tails will add to the 3' end; methylated adaptors are ligated to the DNA fragments, bisulfite treatment, PCR amplification, and sequencing resulted library. WGBS has the ability to assess the methylation status of low CpG-density regions, such as intergenic and distal regulatory elements.

Reduced-representation bisulfite sequencing (RRBS): These are cost-effective protocols, widely used in profiling large-scale samples. After MspI restriction digestion, bisulfite conversion and next-generation sequencing will be performed. A size selection of MspI-digested fragments between 40 and 220 bps was found to cover 85% of CGIs, mostly in promoters, which compose only 1%–3% of the mammalian genome, thereby significantly decreasing the amount of sequencing.

Bioinformatics analysis of WGBS and RRBS methods includes data processing, DNA methylation quantification profiling, identification of the DMRs, and visualization of the methylome. Aligning bisulfite-converted reads are based on two algorithms: wild-card aligners that substitute Cs with Ys in the reference genome, and reads with both Cs and Ts can then be aligned. This method results in higher genomic coverage. However, the three-letter aligners convert all Cs in the reference genome and the read into Ts, and thus, standard aligners with lower mappability can be adopted because of reduced sequence complexity. The alignment profile can be visualized with tools such as the UCSC genome browser, WBSA, IGV, and Methylation plotter, which results in greater clarity at a single-base resolution across the genome.

Recently, The Cancer Genome Atlas Consortium (TCGA) profiled DNA methylation of more than 300 EC samples using array-based DNA methylation platforms (HumanMethylation27 BeadChip and Human Methylation 450 BeadChip), which interrogate 27,578 CpG sites and 482,421 CpG sites, respectively.

In one study, whole genome DNA methylation changes in two classical types of EC by applying MeDIP-seq and MRE-seq methods discovered 27,009 and 15,676 recurrent DMRs for EAC and uterine papillary serous carcinoma (UPSC). More than 80% of DMRs were in intergenic and intronic regions. In addition, large-scale demethylation of X chromosome was detected in UPSC, accompanied by decreased XIST expression. Majority of DMRs harbored promoter or enhancer functions and associated with uterine development and disease [22] (Table 21.1).

Methylation panel for EC using epigenomic analysis filtered out 180 methylated genes and detected 14 hypermethylated genes in EC. Results suggested the potential use of methylated BHLHFE22, CDO1, and CELF4 panel for EC screening. Based on the analysis of DNA methylation of 1135 DMC and 1488 DECs obtained between tumor and normal groups, PCDHs clusters, DDP6, and TNXB were found to be associated with tumorigenesis and may be novel candidate biomarkers for EC [23].

Table 21.1 Computational DNA Methylome Analysis of Endometrioid Adenocarcinoma (EAC) or Type I and Uterine Papillary Serous Carcinoma (UPSC) or type 2

Resource	Purpose	URL, Software, Equation
DMR identification	Identification of DMRs in the R 2.15 environment between the DNA methylome of normal endometrium and each cancer sample	http://epigenome.wustl.edu/Mnm
Genomic features	CpG islands and RefSeq gene coding loci features microRNA gene cluster Human lincRNA Catalog	UCSC genome browser: http://mirstart.mbc.nctu.edu.tw/ http://www.broadinstitute.org/genome_bio/human_lincrnas/ http://cancergenome.nih.gov/
TCGA DNA methylation data	Downloading processed DNA methylation data of uterine corpus endometrial carcinomas (Infinium Human Methylation 450 BeadChip platform) from TCGA	
TCGA RNA expression data	Downloading processed mRNA-seq and miRNA-seq data of uterine corpus endometrial carcinomas (Illumina GA and Illumina HiSeq platform) from TCGA	http://cancergenome.nih.gov/
Validation of DMRs using TCGA Infinium 450K data	Computation of the average methylation level (aML) in each cancer type group. n : the number of samples in a specific cancer type group; w : the number of available Infinium CpG probes within the corresponding DMR; $BVij$: beta value of the j^{th} Infinium CpG probe within that DMR in the i^{th} sample. DNA methylation change (DMC) aML_C : the averaged methylation level of a DMR in a specific cancer type group; aML_n : the averaged methylation level of a DMR in the normal control group EAC tpDMRs and UPSC tpDMRs validation to examine the significance of DNA methylation difference between cancer group and normal controls.	$aML = \sum_i^n \left(\sum_j^w BVij/w \right) / n$ $DMC = aML_C - aML_n$ Mann-Whitney U test

Continued

Table 21.1 Computational DNA Methylome Analysis of Endometrioid Adenocarcinoma (EAC) or Type I and Uterine Papillary Serous Carcinoma (UPSC) or type 2—cont'd

Resource	Purpose	URL, Software, Equation
Enrichment calculation	Calculation of the binding site ES for each genomic feature, DHS, and transcription factor with respect to DMRs: $n - hit$: the number of DMRs that contain specific a genomic feature, experimentally annotated DHS, or TFBS; $n - DMR$: the total number of DMRs $N - hit$: the number of genomic windows with a specific genomic feature, annotated DHS, or TFBS; $N - all$: the number of 500 bp windows in the human genome	$ES = \frac{n - hit}{n - DMR}$ $N - hit/N - all$

DHS, *DNase I hypersensitive site*; DMR, *Differentially methylated region*; EC, *Enrichment score*; TCGA, *The Cancer Genome Atlas*; TFBS, *Transcription factor binding site*.

Studies of aberrant DNA hypermethylation associated with EC have been found that hypermethylation of gene promoters is linked to reduced expression of several genes in this cancer. Table 21.2 listed a number of genes frequently hypermethylated in EC.

DNA MISMATCH REPAIR GENES

MSI is a characteristic of tumor cells in which the MMR system has failed. This change is found in certain types of cancers and especially prevalent in EC. *hMLH1* and *hMSH2* are MMR genes that have a strong association with EC. Approximately, 20%–30% of cases of sporadic endometrial carcinomas, showed MSI. While *hMLH1* and *hMSH2* mutations are rare (less than 10%) in sporadic ECs with the MSI + phenotype. However reduced expression of *hMLH1* and other MMR genes is common in ECs lacking detectable mutations. A strong association between *hMLH1* promoter methylation, transcriptional silencing and MSI+ phenotype was observed in sporadic endometrial cancer, particularly in the endometrioid type. It has been shown that *hMLH1* promoter methylation as an early event in the procession from normal endometrium to carcinoma, and as a feature of a subset of precursor lesion. However, *hMLH2* methylation is very rare (1.4%) in endometrial cancer [24].

Accurate identification of MMR deficiency in EC may be important to identify patients with a higher risk of recurrence. In screening for Lynch syndrome, use of five microsatellite markers, two mononucleotide repeats (BAT26 and BAT25), and three dinucleotide repeats (D5S346, D2S123, and D17S250) is recommended. MSI is observed in certain types of cancer, including 20%–30% of cases of EC. Research results suggest that MMR gene abnormalities occur frequently in EC. Women with

Table 21.2 Genes Evaluated to Be as a Biomarker for Endometrial Adenocarcinomas, Their Function, and Genetic or Epigenetic Alterations

Symbol	Gene	Function	Genetic/Epi genetic Alternations
<i>APC</i>	Adenomatous polyposis coli	Tumor suppressor gene, cell adhesion, signal transduction, stabilization of the cytoskeleton, regulation of cell cycle, and apoptosis	Hypermethylation
<i>PTEN</i>	Phosphatase and tensin homolog	Tumor suppressor gene, controls proliferation and apoptosis	Mutation/deletion/hypermethylation
<i>P16</i>	Cyclin-dependent kinase inhibitor 2A (CDKN2K)	Tumor suppressor gene, cell cycle regulation, involved in senescence	Mutation/hypermethylation
<i>hMLH1</i>	Human Mut-L homolog 1	DNA mismatch repair gene	Hypermethylation
<i>RSK4</i>	X-linked Ribosomal S6 Kinase RPS6KA6	Serine—threonine protein kinases, which are regulated by growth factors, putative tumor suppressor gene, and is a target of the ERK signaling pathway	Hypermethylation
<i>SPRY2</i>	Sprouty RTK signaling antagonist 2	An antagonist of fibroblast growth factor pathways and may negatively modulate respiratory organogenesis	Hypermethylation
<i>CDH1</i>	E-cadherin	Epithelial cell—cell adhesion, suppresses invasion and metastasis	Hypermethylation
<i>RASSF1A</i>	Ras association domain family protein 1	Tumor suppressor gene, cell cycle regulation, microtubule stabilization, cellular adhesion and inhibits tumor formation, apoptosis	Hypermethylation
<i>HOXA11</i>	Homeobox A11	Expressed in endometrial epithelium and is known to play a role in uterine embryogenesis	Hypermethylation
<i>COMT</i>	Catechol-O-methyltransferase	Important in the metabolism of catechol drugs	Hypermethylation

Continued

Table 21.2 Genes Evaluated to Be as a Biomarker for Endometrial Adenocarcinomas, Their Function, and Genetic or Epigenetic Alterations—cont'd

Symbol	Gene	Function	Genetic/Epigenetic Alterations
<i>RARβ2</i>	Retinoic acid receptor	Apoptosis, involved in senescence, inhibition of proliferation	Hypermethylation
<i>ERα</i>	Estrogen receptor α	Steroid receptor, regulation of cell proliferation	Mutation
<i>PRβ</i>	Progesterone receptor β	Steroid receptor, growth regulation	Mutation
<i>GPR54</i>	G-protein-coupled receptor 54	Regulation of endocrine function, play a role in the onset of puberty	Hypermethylation, mutation
<i>TGFβRII</i>	Transforming growth factor beta receptor II	Cell proliferation, cell cycle arrest, wound healing, immunosuppression, and tumorigenesis	Hypermethylation
<i>TP53</i>	Tumor protein p53	Tumor suppressor gene, involved in apoptosis, cell cycle arrest, and senescence	Mutation
<i>TP73</i>	Tumor protein p73	Involved in cellular responses to stress and development	Mutation
<i>CHFR</i>	checkpoint with fork head and ring finger domains	Cell cycle progression and tumorigenesis	Hypermethylation
<i>THBS1</i>	Thrombospondin-1	Regulated by the TP53, and Rb can act to promote or suppress angiogenesis and fibrinolysis	Hypermethylation
<i>THBS2</i>	Thrombospondin-2	Similar structure to THBS1, function unknown	Hypermethylation
<i>CTNNB1</i>	β -catenin	Participates in the tissue adherens complex	Mutation/ hypermethylation
<i>VDR</i>	Vitamin D receptor (25-hydroxyvitamin D3-1 α hydroxylase)	Involved with antiproliferation and prodifferentiation	Hypermethylation
<i>ERBB2</i>	erb-b2 receptor tyrosine kinase	Amplification or overexpression of this oncogene plays an important role in the development and progression of certain aggressive types of cancers	Amplification

Table 21.2 Genes Evaluated to Be as a Biomarker for Endometrial Adenocarcinomas, Their Function, and Genetic or Epigenetic Alterations—cont'd

Symbol	Gene	Function	Genetic/Epigenetic Alterations
<i>K-ras</i>	KRAS protooncogene, GTPase	Protooncogene, regulating cell division	Mutation
<i>MGMT</i>		DNA repair gene	Hypermethylation
<i>PIK3CA</i>	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	Regulator of cellular growth, transformation, adhesion, apoptosis, survival, and motility	Mutation
<i>AKT1</i>	AKT serine/threonine kinase 1	Phosphorylates and inactivates components of the apoptotic machinery	Mutation
<i>RUNX3</i>	Runt-related transcription factor 3	Transcription factor that is active or suppress transcription, tumor suppressor gene	Deletion, hypermethylation
<i>HAND2</i>	Heart and neural crest derivatives expressed 2	Transcriptional factor expressed in endometrial stroma	Hypermethylation
<i>AURKA</i>	Aurora kinase A	Putative oncogene-accurate segregation of chromosomes during mitosis, tumor development, and progression	Mutation, hypermethylation
<i>CCND1</i>	Cyclin D1	Regulators of CDK	Mutation

Lynch syndrome have a high risk for EC, with a life-long incidence of 40%–60%, which is similar to or greater than that of colon cancer. Potential screening methods include transvaginal ultrasound and endometrial biopsy. Transvaginal or transabdominal sialography is used to evaluate endometrial conditions and thickness [25].

MMR gene methylations are particularly important and contribute to carcinogenesis. Strong association between *hMLH1* promoter methylation, transcriptional silencing, and MSI + phenotype was reported in sporadic EC. This is an early event in the procession from normal endometrium to carcinoma. Furthermore, as a feature of a subset of precursor lesion, inactivation of MMR genes induces MSI in many tumor suppressor genes, including *PTEN*, *TGF-βR2*, *IGF2R*, and *BAX*. Aberrant methylation of *hMLH1* gene has been detected in 40.4% of patients with EC, which is an important step in the early stages of carcinogenesis, with the loss of DNA MMR function, proposed to lead to mutation of genes such as *PTEN*. Epigenetic inhibition of *hMLH1* expression is more frequent than that of *hMSH2* (1.4%) in EC. In patients with EC, aberrant hypermethylation of *hMLH1*, *APC*, *E-cadherin*, and *CHFR* in 40.4%, 22.0%, 14.0%, and 13.3% of cases has been reported, respectively. In addition, patients with MMR gene defects had significantly younger age (≤ 60 years) and better prognosis in terms of early stage, negative nodal status, and longer survivals [26,27].

The *EPCAM* gene encodes epithelial cell adhesion molecules and is overexpressed in most of the cancers. It is a homophilic intracellular adhesion molecule that may promote metastasis of cancer cells by inhibiting intracellular adhesion due to E-cadherin. There are various opinions on the role of *EPCAM* in carcinogenesis. Researches have showed that epigenetic mutation in the 3'-untranslated region (UTR) of *EPCAM* inactivated *hMSH2* and was involved in carcinogenesis of EC [28].

MGMT is a silenced DNA repair gene that is present in 48% of EECs. Loss of *MGMT* function leads to recognition of O(6)-methylguanine as adenine by DNA polymerases. O(6)-methylguanine is a promutagenic form that leads to G-to-A mutations [29].

STEROID RECEPTOR GENES

The cyclic production of estrogen and progesterone during the menstrual cycle and declining sex steroid hormone levels after menopause are directly correlated with endometrial proliferation and/or atrophic morphological changes. Specific role of the three *Er α* (A,B,C) isoforms in EC is still not clear. In an investigation the expression level of three isoforms of *ER α* in endometrial cancer cell lines was investigated and promoter methylation of *ER α -C* isoform found in 94% of EC tissues; However they found no *ER α -C* expression, as well as restoration of the expression with 5-aza-2' deoxycytidine treatment. However, in another study reported only 24% of 25 endometrial cancer cases; methylation was correlated with ER-negative status of the tumors. Other researches did not find any association between loss of ER expression and de novo methylation of the ER gene. Analysis of PR gene indicated a positive association between promoter hypermethylation of PR-B and reduced mRNA expression, in one study more than 70% of EC samples were promoter methylated for PR-B, but in all cancer and normal samples, PR-A promoter was unmethylated [30,31].

TUMOR SUPPRESSOR GENES

Promoter hypermethylation of tumor suppressor genes is one of the most important events in many cancers, including endometrial carcinoma.

PTEN is also known as MMAC1/TEP1 gene located on chromosome 10q23.3 and is responsible for Cowden syndrome and Bannayan-Zonana syndromes. A number of studies have shown that *PTEN* promoter methylation in about 20% is of sporadic type I endometrial carcinoma, and this methylation was significantly associated with metastatic disease and MSI phenotype. But results of some researchers suggested that *PTEN* pseudogene is predominantly methylated, not *PTEN*. More studies with the ability to make distinction between promoter methylation of *PTEN* and its pseudogene are needed for the better understanding of the mechanisms of *PTEN* inactivation in endometrial carcinogenesis [32].

P16INK4a encodes a cyclin-dependent kinase, which belongs to a family of inhibitor-dependent kinase INK4, with direct connection to CDK4 and CDK6, arrest cell cycle in G1 phase. Reduced expression of *P16* detected in 16% of endometrial tumors. Promoter methylation of *p16* gene has been observed between 11% and 75% of sporadic ECs; however, other studies have reported much lower frequencies of *p16* methylation, in which variability may be due to assay sensitivity, primer design, or study sample size. In addition, abnormal epigenetic modifications of *p16* play an important role during EC carcinogenesis. High methylation of P16INK4a promoter was observed up to 75% of sporadic ECs. The molecular mechanisms of *p16* inactivation in EC remain unclear. In a study, it has been reported that *P16* promoter methylation is associated with advanced stage and poorer survival of EC;

another research found that rare *P16* methylation occur in a subgroup of aggressive endometrial carcinomas with poor prognosis [33,34].

RASSF1A induces cell cycle arrest through the Rb-mediated checkpoint by inhibiting the accumulation of cyclin D1. *RASSF1A* promoter methylation is an early event in type I endometrial carcinogenesis. It has been reported that methylation occurs in 33%–85% of ECs and is associated with reduced expression of *RASSF1A*, but normal endometrium adjacent to endometrioid showed 36% methylation. Hypermethylation of this gene has been found to be correlated with LOH in which reflecting the importance of *RASSF1A* in endometrial carcinoma. Furthermore, a number of researches confirmed that hypermethylation of *RASSF1A* gene is associated with advanced stage, recurrence, and survival of the patients [35–37].

APC is a tumor suppressor gene and a regulator of wnt signaling pathway through β-catenin stability and/or degradation. *APC* alternations and its effect in wnt signaling pathway are thought to be associated with endometrial carcinogenesis. In several studies, it has been reported that *APC* promoter methylation, approximately 20%–45% and more frequent in tumors with MSI than those without MSI. But no significant associations of *APC* promoter methylation with the clinicopathological factors or recurrence and distant metastases have been observed in ECs.

CDH1 encodes E-cadherin, which is the major cadherin molecule expressed in epithelial cells. The cadherin mediated cell adhesion system, and it is known to act as an invasion suppressor system in cancer cells. Hypermethylation of E-cadherin promoter region is a frequent event in endometrial carcinoma, which may play an important role in the progression of carcinogenesis. *CDH1* promoter hypermethylation has been reported in 40% of endometrial carcinomas and 10% of endometrial hyperplasias. Decreased expression of E-cadherin was detected in 43% of endometrial carcinomas and 5% of endometrial hyperplasias. Promoter hypermethylation of *CDH1* has been results it's down regulation and effects on both cancer progression in clinical pathology and 5-year survival rates. Finding of researches suggest that aberrant methylation of *CDH1*, promotes invasion and metastasis of cancer cells and worsens the prognosis of endometrial cancer [39].

P53 is defined as a typical tumor suppressor gene, crucial cell cycle regulator, apoptosis, and DNA repair induction. Its mutation association with endometrial carcinogenesis has been reported in 90% of type II endometrial cases and 10%–20% of grade 3, type I EC. Results of a research showed that p53 mutations occurred only at sites with positive p53 protein expression in endometrioid adenocarcinoma, which were poorly differentiated regions of cancer tissues. *P53* is also implicated in the early stage of carcinogenesis of serous adenocarcinoma. In addition, based on the results of the other research, the *p53* signature is found frequently in normal endometria adjacent to serous adenocarcinoma but rarely detected in other normal endometria or tissues adjacent to endometrioid adenocarcinoma. Atypical epithelium develops features similar to serous adenocarcinoma and covers endometrial cortical layers but is not invasive. This state is referred to as serous endometrial intraepithelial carcinoma and finally progresses to serous adenocarcinoma [38,39].

pRB is an important tumor suppressor gene in which cell cycle progression is inhibited in G1 phase. Indeed, phosphorylation of the Rb protein by cyclineD/cdk4,6 complex cause E2F release, increase DNA polyactivity, and cell cycle proliferation. RB mutations were first identified in retinoblastoma and then in other cancers such as bladder, esophageal, small cell lung as well as EC carcinogenesis. In EC, LOH was found in 18% of RB genes, and *pRB* downregulation was consistent with LOH. The incidence of mutations increased with the advancement of the clinical stage [40].

OTHER RELATED GENES

HOXA11, which encodes a transcription factor, is involved in proliferation, differentiation, and embryologic development of the endometrium. *HOXA11* is expressed in both epithelium and stroma in the adult uterus and appears to be regulated by ovarian steroids. Observation of some researchers suggested that *HOXA11* methylation may contribute to endometrial tumorigenesis. It has been showed that methylation of the HOXA11 promoter was more frequent (90%) in recurrent EC than in primary tumors, that later recurred demonstrated, methylation of this locus, whereas only 51% of nonrecurrent primary tumors were methylated [41].

CHFR is an M-phase checkpoint gene that regulates progression of the cell cycle. *CHFR* down-regulation by aberrant hypermethylation increases the paclitaxel sensitivity of gastric cancer and ECs. Findings suggest that examination of *CHFR* expression could form the basis of personalized cancer treatment [42].

COMT (catechol-O-methyltransferase) is an enzyme that plays an important role in estrogen-induced cancers because *COMT* inactivates catechol estrogens that have cancer-promoting activities. It has been found that methylation of the *COMT* promoter (69%–94%) selectively inactivated membrane-bound COMT (MB-COMT) and may contribute to endometrial carcinogenesis via estrogen [43].

SPRY2 is an antagonist of the fibroblast growth factor (FGF) receptor and inhibits cell proliferation and angiogenesis by inhibiting the RAS-MAPK pathway, downstreaming the FGF receptor. It has been found that *SPRY2* expression depended on the menstrual cycle in normal endometria and *SPRY2* involved for the development of glandular structures. Its expression is extremely low in highly invasive cancer other than endometrioid adenocarcinoma [44].

CASP8 is an apoptosis-related gene involved in cell death via Fas ligands. *CASP8* has been found to be methylated in EC [45].

MICRORNA ABERRANT METHYLATION IN ENDOMETRIAL CARCINOMA

Currently the noninvasive diagnosis of EEC mainly relies on the combination of ultrasound, MRI, and serological markers, but none of them are completely satisfactory. CA125 has been used for many years as a serum marker for EC diagnosis and screening. However, it has been recognized that CA125 levels have poor specificity in the detection of EC. A large number of miRNAs are proved to be present in circulation. By examining TCGA miRNA-seq data, it has been shown that expression levels of certain miRNAs correlated with DNA methylation changes in their promoters.

miRNAs are short, noncoding RNAs of about 18–25 bases, which may cause mRNA degradation or the inhibition of protein translation by directly binding to the 3'-untranslated region of their target mRNAs and that regulate expression of genes. Through modulation of the protein expression of their target genes, which act as oncogenes or tumor suppressors, miRNAs are closely associated with the development and progression of human cancer. miRNAs have been found to be downregulated by methylation of DNA in various cancers, and these miRNAs are referred to as tumor suppressor miRNAs (TS-miRNAs) [46].

Significant demethylation in some of the miRNAs increases their expression level in EC. miR-205, miR-200 family (miR-200a, 200b, 200c, 141, and 429) this family together with miR-205 regulates the expression of target genes *ZEB1* and *ZEB2*, which have been implicated to be involved in epithelial-to-mesenchymal transition and tumor progression. miR-96 cluster (hsa-miR-96, hsa-miR-182, and hsa-miR-183) and three miRNAs (miR-499, miR-135b and miR-205) are upregulated in EC.

In contrast, even with strongly hypermethylated promoters, some miRNAs exhibited upregulated expression level, including miR-25, miR-93, miR-99B, miR-324, miR106B, miR-3074, miR-199a-3p (inhibit tumor proliferation by suppression of mTOR) and five miRNAs (miR-10b, miR-195, miR-30a-5p, miR-30a-3p, and miR-21). The regulatory relationship between promoter hypermethylation and miRNA expression is likely complex and needs further investigation [47].

TS-miRNAs INVOLVED IN ENDOMETRIAL CANCER WITH THEIR FUNCTION INCLUDING miR-129-2, miR-152, miR-124, miR-126, miR-137, AND miR-491

Microarray assays showed that *SOX4* gene (SRY-related high-mobility group box 4), is highly expressed in EC cells, and miR-129-2 was a negative regulator of *SOX4*. It was reported that in 68% of patients, miR-129-2 expression was silenced by methylation; in contrast demethylation can increase the miR-129-2 expression, reduce *SOX4* expression, and suppress proliferation of cancer cells. Furthermore, hypermethylation of miR-129-2 was statistically related to MSI and methylation status of *hMLH1* [48].

miR-152 is identified as a TS-miRNA candidate in EC; its methylation is reported to be altered in acute lymphocytic leukemia. Methylation of miR-152 is completely consistent with expression, and hypermethylation in the promoter region reduces the expression. DNMT1 is a well-known target of miR-152. E2F3, MET, and Rictor have been identified as additional targets. Both E2F3 and MET genes are identified as cancer genes; E2F3 is a transcription inhibitor; and MET encodes a cell surface receptor for hepatocyte growth factor. Activation of mTORC2-Akt signaling contributes to canceration of the endometrium, therefore silencing of miR-152 appears to lead to activation of multiple targets and may be suggested as a new goal for EC treatment [49].

miR-124 exerts tumor suppressor effects and is frequently methylated in multiple cancer types. Results of a research demonstrated that miR-124 is downregulated in endometrial carcinoma, and the loss of its expression is mediated by DNA methylation. The methylation-mediated repression of miR-124 leads to the overexpression of IQGAP1 (IQ motif containing GTPase-activating protein 1), which in turn accelerates cancer cell proliferation, EMT, and invasion [50].

miR126 typically uses a suppressive role in numerous types of cancer, including non—small cell lung cancer, cervical cancer, gastric cancer, colon cancer, clear cell renal cell carcinoma, and chronic myelogenous leukemia. Recently, miR-126 was suggested to function as a tumor suppressor in EC. It was observed that miR-126 was notably downregulated in EC tissues, compared with matched adjacent tissues. Furthermore, it was demonstrated that miR-126 served a suppressive role in mediating EC cell migration and invasion; in addition, it was confirmed that IRS1 was a direct target gene of miR-126 in the EC cells and insulin-like growth factor 1 receptor signaling contributes to the development of endometrial hyperplasia [51].

Hundreds of lncRNA genes also underwent DNA methylation changes in EC. In a research, the number of hypomethylated lncRNA genes was 207 for EAC and 245 for UPSC, with 68 in common (exception of classic examples such as XIST, H19, and HOTAIR). The roles of lncRNAs and consequences of their abnormal DNA methylation in endometrial carcinogenesis remain to be elucidated. However, interesting candidate genes emerged, for example, promoter region of a tumor suppressor noncoding RNA, MEG3 (maternally expressed gene 3), was highly methylated in EAC, which was significantly associated with downregulation [47].

DNA METHYLATION MACHINERY IN ENDOMETRIUM

DNA methylation mechanism is based on three types of enzymes: writer, eraser, and reader.

Writers are comprised of DNMT enzymes that modify five positions of the cytosine residues and establish or maintain DNA methylation pattern. There are at least five known members of the DNMT group (DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3L) included in three families (1, 2, and 3). DNMT1 is targeted to replication foci, where it has a very high preference for hemimethylated DNA strands, and it is associated with DNA replication machinery. During replication, DNMT1 reproduces the parental DNA methylation pattern in the daughter DNA strands. DNMT2 is an RNA methyltransferase, which can methylate cytosine in the anticodon loop of tRNA^{ASP} despite its high sequence and structural similarity to other DNMT enzymes. DNMT3A and DNMT3B enzymes predominantly catalyze de novo methylation, which occurs in both unmethylated DNA strands. DNMT3L as a regulatory factor interacts with DNMT3A and DNMT3B enzymes and enhances de novo methylation activity of both enzymes, by stabilizing the active catalytic sites in maintenance enzymes.

Erasers: DNA methylation can be passively erased because of inhibition of maintenance of DNMT, which is involved in sustaining the DNA methylation pattern. Modified cytosine residues (5-mC) can chemically react at two sites: the amino group and the methyl group. The amino group can be deaminated to a carbonyl group by the protein AID/APOBEC (activation-induced cytidine deaminase/apolipoprotein B mRNA-editing enzyme complex), which converts 5-mC into thymine and results in a GC-to-TA transition mutation.

Readers: DNA methylation can be recognized by a free family of proteins with various DNA-binding domains, including MBD (methyl-CpG-binding domain) proteins, UHRF (ubiquitin-like, containing PHD and RING finger domain) proteins, and zinc finger proteins. The MBD family of proteins plays a key role in gene silencing and includes several members, namely MeCP2, MBD1, MBD2, MBD3, and MBD4 [21].

The human endometrium is a hormone-sensitive tissue, which undergoes cyclic morphological and biochemical changes on an average of every 28 days. The menstrual cycle is divided into three phases: proliferative, secretory, and menstrual phases, which are under control of the ovarian steroid hormones oestradiol (E2) and progesterone (P4). They interact with their receptors (oestrogen and progesterone receptors), also known as nuclear transcription factors.

Currently, scientific studies confirmed different expression of DNMTs during the menstrual cycle, DNMT expression, and functional regulation, by steroid hormones (Fig. 21.3), but molecular mechanisms are still poorly understood. Some studies obtained results that DNMT1, DNMT3A, and DNMT3B expressions, correlated with oestrogen levels, were significantly downregulated in the midsecretory phase compared with the proliferative phase. In contrast, other studies reported that DNMT1 expression level was higher in the secretory phase, than in the proliferative phase, whereas DNMT3B expression levels were not changed in any phase. Another researcher confirmed that global DNA methylation status and progesterone receptor level were significantly higher during the proliferative phase with a decrease toward the end of the secretory phase [52,53].

Another important component of DNA methylation machinery is the MBD family of proteins, which establish cross talk between DNA methylation and histone modifications in gene silencing. It has been demonstrated that MBD2 expression is higher in the secretory phase of the endometrium,

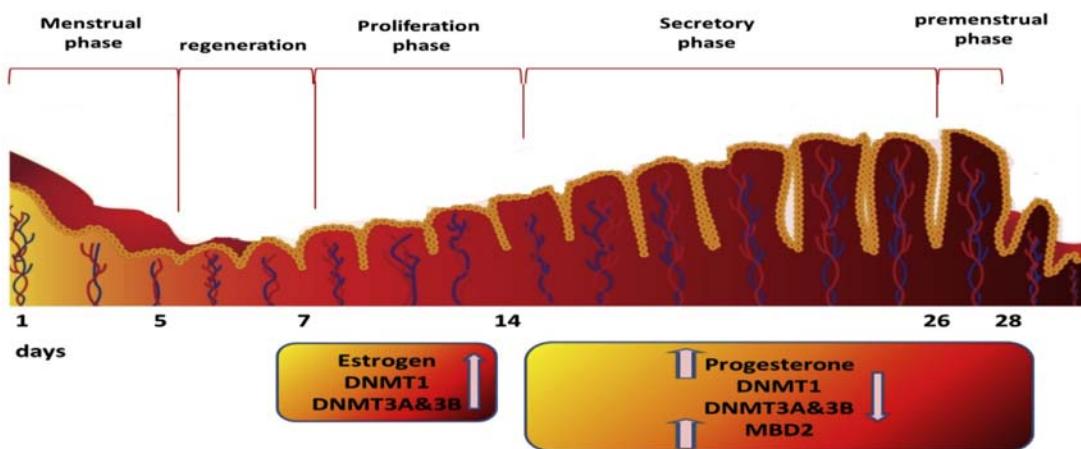


FIGURE 21.3

Menstrual cycle phases associated with gene expression changes of DNA methylation machinery proteins. *DNMTs*, DNA methyltransferases; *MBD2*, DNA methyl-binding protein 2.

compared with the proliferative and menstrual phases, but there were no significant differences in MBD1 and MeCP2 expression throughout the normal menstrual cycle [54].

APPLICATION OF DNA HYPERMETHYLATION FOR TREATMENT

Epigenetic patterns that commonly observe in cancer cells inconstant to irreversible genetic changes can be partly or fully reversed. Indeed, this is the main reason for the development of anticancer drugs for epigenetic modification. Treatment of cancer using demethylating agents to restore expression of cancer suppressor genes, silenced through methylation, has been attempted for some time, and use of methylation inhibitors to treat cancer has a long history. However, toxicity caused by administration of a high dose of these drugs was problematic and prevented practical use. More recently, antitumor effects have been obtained at lower doses with fewer side effects, and better efficacy may be achievable through combination with other chemotherapeutic agents [55,56].

Administration of low doses of 5-aza-2'-deoxycytidine as a methylation inhibitor obtained a response rate of 60% in patients with acute myelocytic leukemia; however, the disadvantage of this approach is that demethylation is not sequence specific and removal of methyl groups with biological importance or activation of oncogenes that are silenced by methylation may occur. Therefore, development of sequence-specific demethylating agents based on binding sequence of transcription factors is a current area of research [57,58].

To detect sensitivity of anticancer drugs, researchers examined utilization of epigenetic aberrations as genetic markers. Some researchers suggested that treatment may be selected based on the specific methylation properties of tumor cells. For example, methylation of *CHFR*, M-phase checkpoint gene plays an important role in the development of EC. Aberrant methylation of this gene is correlated with the sensitivity of tumors to microtubule inhibitors therefore may be a marker of sensitivity of EC to

anticancer drugs. In the mitotic stress condition during M phase, *CHFR* activated decomposes Aurora and PLK1 and delay early prophase to late prophase progression and results cell cycle arrest. If CFTR gene promoter was unmethylated, cell cycle arrest in G2/M phase would occur and mitotic index would be low, but hyper methylation of CFTR promoter region would reduce its expression and result in observation of high mitotic index. On the other hand treatment with 5-aza-2'-deoxycytidine, restore the mitotic index reduction, so methylation index detection of *CHFR* can be suggested as an effective marker for prediction of the effects of anticancer drugs [55,56,59].

Various molecular pathways and genes are associated with carcinogenesis of EC, and consideration of these genes may be important for cancer treatment. Targeting agents include bevacizumab, afilbercept, and thalidomide (monoclonal antibodies against VEGF-A, antiangiogenic); gefitinib and erlotinib (EGFR inhibitors); trastuzumab (monoclonal antibody against extracellular domain of HER2); and temsirolimus and ridaforolimus (mTOR inhibitors). It has been reported that HDACIs, such as vorinostat and valproic acid, were effective in six EC cell lines. In EC, expression of miR-152 has been shown to be inhibited by DNA hypermethylation, so miR-152 may potentiate the use of it in the treatment of EC. The expression levels of miR-200c and miR-205 were significantly higher in EC; identification of such biomarkers may permit diagnosis and prediction of prognosis, which currently depends mainly on histological techniques, allowing early detection of EC and selection of the appropriate treatment regimen. DNMT1, MET, E2E3, and Rictor, all of which are associated with DNA methylation and cell proliferation, which have been identified as improved targeting of these drugs are needed for practical clinical use [60].

FUTURE DIRECTIONS AND CONCLUSION

Recent studies showed that epigenetic alterations in normal cellular processes and abnormal changes leading to endometrial carcinogenesis and aberrant DNA methylation profile result in various pathological forms of the endometrium and EC development. It is, therefore, necessary to recognize the differences between DNA methylation changes that occur naturally in the menstrual cycle phases and those in cancer cells.

In type I EC development, promoter hypermethylation of MMR genes, *MLH1* and *hMSH2* genes play a significant role, and reduced expression of other genes such as *APC*, *CHFR*, *Sprouty 2*, *RASSF1A*, *GPR54*, *CDH1*, and *RSK4* by hypermethylation has been found. Suppression of gene expression by miRNAs also occurs in EC, and expression of the miRNA itself may be increased or decreased by promoter methylation, based on the differences between normal and cancerous endometrial tissue, and may contribute to cancerization of the endometrium. These kinds of epigenetic data have potential for discovery of biomarkers for prevention, diagnosis, risk assessment, and treatment of EC, with considerable potential for clinical application. However, the carcinogenic mechanisms remain largely unknown, particularly with regard to de novo carcinogenesis of type II EC.

Cancer-specific DNA methylation may be useful for diagnosis using methods such as Methylation Specific PCR (MSP) for detection of several cancers, including EC. Aberrant DNA hypermethylation can be detected with a high level of sensitivity, and cancer cells can be detected in minute quantities of endometrial samples. Currently, the main goals of epigenetic studies in cancer research are to identify gene hypermethylation that is directly related to canceration and to use these findings in diagnosis and treatment. Agents that regulate methylation of specific genes may ultimately be particularly effective as anticancer therapy.

Complete DNA methylome of EC investigation provides the whole-genome DNA methylation map for this common and deadly disease. From these data sets, tens of thousands of DMRs specific to this cancer have been identified. The majority of DMRs harbor regulatory functions including promoters and enhancers that are important to developmental and pathological changes of the uterus. Many methylation changes were found in CpG island shores and were associated with expression changes of nearby genes. In addition, large-scale of demethylation of chromosome X in UPSC accompanied by decreased *XIST* expression has observed. Remethylation of transposable elements in cancers might provide a novel mechanism to deregulate normal endometrium-specific enhancers derived from specific transposable elements. Results of researches demonstrate that DNA methylation changes are an important signature of EC and regulate gene expression by affecting not only proximal promoters but also distal enhancers, including those derived from transposable elements.

REFERENCES

- [1] Le T, Bentley J, Farrell S, Fortier MP, Giede C, Kupets R, et al. Epidemiology and investigations for suspected endometrial cancer. *J Obstet Gynaecol Can* 2013;35(4):380–1.
- [2] Leslie KK, Thiel KW, Goodheart MJ, Koen De Geest MD, Jiae Y, Yang S. Endometrial cancer. *Obstet Gynecol Clin* 2012;39(2):255–68.
- [3] Burke WM, Mario Leitao JO, Paola Gehrig ES, Olawaiye AB, et al. Endometrial cancer: a review and current management strategies: Part I. *Gynecol Oncol* 2014;134:385–92.
- [4] Plataniotis G, Castiglione M. Endometrial cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2010;21(5):41–5.
- [5] Armstrong AJ, Hurd WW, Elguero S, Barker NM, Zanotti KM. Diagnosis and management of endometrial hyperplasia. *J Minim Invasive Gynecol* 2012;19:562–71.
- [6] Lax S. Precursor lesions of endometrial carcinoma, diagnostic approach and molecular pathology. *Pathologie* 2011;32(2):255–64.
- [7] Sorosky JI. Endometrial cancer. *Obstet Gynecol* 2012;120:383–97.
- [8] Ma X, Ma CX, Jianghui W. Endometrial carcinogenesis and molecular signaling pathways. *Am J Mol Biol* 2014;4:134–49.
- [9] Oda K. Targeting ras-PI3K/mTOR pathway and the predictive biomarkers in endometrial cancer. *Gan to Kagakuryoho. Cancer Chemother* 2011;38:1084–7.
- [10] Mori N, Kyo S, Sakaguchi J, Mizumoto Y, Ohno S, Maida Y, et al. Concomitant activation of AKT with extracellular-regulated kinase 1/2 occurs independently of PTEN or PIK3CA mutations in endometrial cancer and may be associated with favorable prognosis. *Cancer Sci* 2007;98:1881–8.
- [11] Dobrzycka B, Terlikowski SJ, Mazurek A, Kowalcuk O, Niklińska Wa CL, et al. Mutations of the KRAS Oncogene in Endometrial Hyperplasia and Carcinoma. *Folia Histochemica et Cytobiologica/Polish Academy of Sciences. Pol Histochem Cytochem Soc* 2009;47:65–8.
- [12] Alexander-Sefre F, Salvesen HB, Ryan A, Singh N, Akslen LA, MacDonald N, et al. Molecular assessment of depth of myometrial invasion in stage I endometrial cancer: a model based on K-ras mutation analysis. *Gynecol Oncol* 2003;91:218–25.
- [13] Wang Y, van der Zee M, Fodde R, Blok LJ. Wnt/β-Catenin and sex hormone signaling in endometrial homeostasis and cancer. *Oncotarget* 2010;1:674–84.
- [14] Kiewisz J, Wasniewski T, Kmiec Z. Participation of WNT and β-catenin in physiological and pathological endometrial changes: association with angiogenesis. *BioMed Res Int* 2015;11. <https://doi.org/10.1155/2015/854056>.

- [15] Holland CM, Day K, Evans A, Smith SK. Expression of the VEGF and angiopoietin genes in endometrial atypical hyperplasia and endometrial cancer. *Br J Cancer* 2003;89:891–8.
- [16] Yokoyama Y, Charnock-Jones DS, Licence D, Yanaihara A, Hastings JM, Holland CM, et al. Expression of vascular endothelial growth factor (VEGF)-D and its receptor, VEGF receptor 3, as a prognostic factor in endometrial carcinoma. *Clin Canc Res* 2003;9:1361–9.
- [17] Growdon WB, Groeneweg J, Byron V, DiGloria C, Borger DR, Tambouret R, et al. HER2 over-expressing high grade endometrial cancer expresses high levels of p95HER2 variant. *Gynocol Oncol* 2015;137(1):160–6. <https://www.ncbi.nlm.nih.gov/pubmed/25602714-comments>.
- [18] Buza N, Roque DM, Santin AD. HER2/neu in endometrial cancer: a promising therapeutic target with diagnostic challenges. *Arch Pathol Lab Med* 2014;138(3):343–50.
- [19] Chaudhry P, Asselin E. Resistance to chemotherapy and hormone therapy in endometrial cancer. *Endocr Relat Cancer* 2009;16(2):363–80.
- [20] Kanwal R, Gupta S. Epigenetic modifications in cancer. *Clin Genet* 2012;81(4):303–11.
- [21] Caplakova V, Babusikova E, Blahovocova E, Balharek T, Zeliskova M, Hatok J, et al. DNA methylation machinery in the endometrium and endometrial cancer. *Anticancer Res* 2016;36:4407–20.
- [22] Zhang BO, Xing XY, Li J, Lowdon RF, Zhou Y, Lin N, et al. Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers. *BMC Genom* 2014;15:868.
- [23] Huang RL, Su PH, Liao YP, Wu TI, Hsu YT, Lin WY, et al. Integrated epigenomics analysis reveals a DNA methylation panel for endometrial cancer detection using cervical scrapings. *Clin Canc Res* 2016. <https://doi.org/10.1158/1078-0432.CCR-16-0863>.
- [24] Masuda K, Banno K, Yanokura M, Kobayashi Y, Iori Kisu I, Ueki A, et al. Relationship between DNA mismatch repair deficiency and endometrial cancer. *Mol Biol Int* 2011. <https://doi.org/10.4061/2011/256063>.
- [25] Stelloo E, Jansen AML, Osse1 EM, Nout RA, Creutzberg CL, Ruano D, et al. Practical guidance for mismatch repair-deficiency testing in endometrial cancer. *Ann Oncol* 2017;28:96–102.
- [26] Zighelboim I, Goodfellow PJ, Gao F, Gibb RK, Powell MA, Rader JS, et al. Microsatellite instability and epigenetic inactivation of MLH1 and outcome of patients with endometrial carcinomas of the endometrioid type. *J Clin Oncol* 2007;25:2042–8.
- [27] Horowitz N, Pinto K, Mutch DG, Herzog TJ, Rader JS, Gibb R, et al. Microsatellite instability, MLH1 promoter methylation, and loss of mismatch repair in endometrial cancer and concomitant atypical hyperplasia. *Gynecol Oncol* 2002;86:62–8.
- [28] van der Gun BTF, Melchers LJ, Ruiters MH, de Leij LFMH, McLaughlin PM, Rots MG. EpCAM in carcinogenesis: the good, the bad or the ugly. *Carcinogenesis* 2010;31(11):1913–21.
- [29] Meng H, Tao M, Jo L. Freudenheim.DNA methylation in endometrial cancer. *Epigenetics* 2010;5(6):491–8.
- [30] Hori M, Iwasaki M, Shimazaki J, Inagawa S, Itabashi M. Assessment of hypermethylated DNA in two promoter regions of the estrogen receptor alpha gene in human endometrial diseases. *Gynecol Oncol* 2000;76:89–96.
- [31] Navari JR, Roland PY, Keh P, Salvesen HB, Akslen LA, Iversen OE, et al. Loss of estrogen receptor (ER) expression in endometrial tumors is not associated with de novo methylation of the 5' end of the ER gene. *Clin Cancer Res* 2000;6:4026–32.
- [32] B HBT, Li J, Barkoh BA, Luthra R, Mills GB, et al. Clinical assessment of PTEN loss in endometrial carcinoma: Immunohistochemistry out-performs gene sequencing. *Mod Pathol* 2012;25(5):699–708.
- [33] Azizi M, Asaadi Tehrani G. Evaluation of promoter hypermethylation of tumor-suppressor genes p14 and p16 in Iranian endometrial carcinoma patients. *Middle East J Cancer* October 2017;8(4):179–86.
- [34] Hu ZY, Ld T, Zhou Q, Xiao L, Cao Y. Aberrant promoter hypermethylation of p16 gene in endometrial carcinoma. *Tumor Biol* 2015;36(3):1487–91.

- [35] Arafa M, Kridelka F, Mathias V, Vanbellinghen JF, Renard I, Foidart JM, et al. High frequency of RASSF1A and RAR β 2 gene promoter methylation in morphologically normal endometrium adjacent to endometrioid adenocarcinoma. *Histopathology* 2008;53:525–32.
- [36] Pallarés J, Velasco A, Eritja N, Santacana M, Dolcet X, Cuatrecasas M, et al. Promoter hypermethylation and reduced expression of RASSF1A are frequent molecular alterations of endometrial carcinoma. *Mod Pathol* 2008;21:691–9.
- [37] Jabbara N, Asaadi Tehrani G, Lalooha F, Farzam SA, Elmizadeh KH. Promoter hypermethylation analysis of the tumor suppressor genes RASSF1A and RASSF2A in Iranian endometrial carcinoma patients. *Int J Cancer Manag* 2017;10(4):e8629.
- [38] Feng YZ, Shiozawa T, Horiuchi A, Shih HC, Miyamoto T, Kashima H, et al. Intratumoral heterogeneous expression of p53 correlates with p53 mutation, Ki-67, and cyclin A expression in endometrioid-type endometrial adenocarcinomas. *Virchows Arch* 2005;447:816–22.
- [39] Zheng W, Xiang L, Fadare O, Kong BA. Proposed model for endometrial serous carcinogenesis. *Am J Surg Pathol* 2011;35:e1–14.
- [40] Semczuk A, Marzec B, Roessner A, Jakowicki JA, Wojcierowski J, Schneider-Stock R, et al. Loss of heterozygosity of the retinoblastoma gene is correlated with the altered pRb expression in human endometrial cancer. *Virchows Arch* 2002;441:577–83.
- [41] Whitcomb BP, Mutch DG, Herzog TJ, Rader JS, Gibb RK, Goodfellow PJ, et al. Frequent HOXA11 and THBS2 promoter methylation, and a methylator phenotype in endometrial adenocarcinoma. *Clin Cancer Res* 2003;9:2277–87.
- [42] Wang X, Yang Y, Xu C, Xiao L, Shen H, Zhang X, et al. CHFR suppression by hypermethylation sensitizes endometrial cancer cells to paclitaxel. *Int J Gynecol Cancer* 2011;21:996–1003.
- [43] Sasaki M, Kaneuchi M, Sakuragi N, Dahiya R. Multiple promoters of catechol-O-methyltransferase gene are selectively inactivated by CpG hypermethylation in endometrial cancer. *Cancer Res*. 2003;63(12):3101–6.
- [44] Velasco A, Pallares J, Santacana M, Gatius S, Fernandez M, Domingo M, et al. Promoter hypermethylation and expression of sprouty 2 in endometrial carcinoma. *Hum Pathol* 2011;42:185–93.
- [45] Banno K, Yanokura M, Iida M, Masuda K, Aoki D. Carcinogenic mechanisms of endometrial cancer: involvement of genetics and epigenetics. *J Obstet Gynaecol Res* 2014;40(8):1957–67.
- [46] Banno K, Kisu I, Yanokura M, Masuda K, Kobayashi Y, Ueki A, et al. Endometrial cancer and hypermethylation: regulation of DNA and MicroRNA by epigenetics. *Biochem Res Int* 2012. <https://doi.org/10.1155/2012/738274>.
- [47] Ulfenborg B, Jurcevic S, Lindlöf A, Karin Klinga-Levan K, Olsson B. miREC: a database of miRNAs involved in the development of endometrial cancer. *BMC Res Notes* 2015;8:104.
- [48] Huang YW, Liu JC, Deatherage DE, Luo J, Mutch DG, Goodfellow PJ, et al. Epigenetic repression of microRNA-129-2 leads to overexpression of SOX4 oncogene in endometrial cancer. *Canc Res* 2009;69(23):9038–46.
- [49] Tsuruta T, Kozaki K, Uesugi A, Furuta M, Hirasawa A, Imoto I, et al. miR-152 is a tumor suppressor microRNA that is silenced by DNA hypermethylation in endometrial cancer. *Canc Res* 2011;71(20):6450–62.
- [50] Dong P, Ihira k XY, Watari H, Sharon JB, Hanley SH JB, Yamad Y, et al. Reactivation of epigenetically silenced miR-124 reverses the epithelial-to-mesenchymal transition and inhibits invasion in endometrial cancer cells via the direct repression of IQGAP1 expression. *Oncotarget* 2016;7(15):20260–70.
- [51] Zhao X, Zhu D, Lu C, Yan D, Li L, Chen Z. MicroRNA-126 inhibits the migration and invasion of endometrial cancer cells by targeting insulin receptor substrate 1. *Oncol Lett* 2016;11:1207–12.
- [52] Yamagata Y, Asada H, Tamura I, Lee L, Maekawa R, TaniguchiK, et al. Matsuoka A, Tamura H and Sugino N: DNA methyltransferase expression in the human endometrium: downregulation by progesterone and estrogen. *Hum Reprod* 2009;24:1126–32.

- [53] Vincent ZL, Farguar CM, Mitchell MD, Ponnampalam AP. Expression and regulation of DNA methyltransferases in human endometrium. *Fertil Steril* 2011;95(4):1522–5.
- [54] van Kaam KJ, Delvoux B, Romano A, D’Hooghe T, Dunselman GA, Groothuis PG. Deoxyribonucleic acid methyltransferase and methyl-CpG-binding domain proteins in human endometrium and endometriosis. *Fertil Steril* 2011;95(4):1421–7.
- [55] Muraki Y, Banno K, Yanokura M, Kobayashi Y, Kawaguchi M, Nomura H, et al. Epigenetic DNA hypermethylation: clinical applications in endometrial cancer (Review). *Oncol Rep* 2009;22:967–72.
- [56] Ono A, Kisu I, Banno K, Yanokura M, Masuda K, Kobayashi Y, et al. Epigenetic aberrant hypermethylation of DNA in endometrial cancer: application as a biomarker. *J Cancer Ther* 2011;2:610–5.
- [57] Issa JPJ, Garcia-Manero G, Giles FJ, Mannari R, Thomas D, Faderl S, et al. Phase 1 study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2'-Deoxycytidine [Decit-abine] in hematopoietic malignancies. *Blood* 2003;103(5):1635–40.
- [58] Yang AS, Doshi KD, Choi SW, Joel B, Mason JB, Mannari RK, et al. DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia. *Cancer Res* 2006;66(10):5495–503.
- [59] Koga K, Kitajima Y, Miyoshi A, Sato K, Sato S, Miyazaki K. The significance of aberrant CHFR methylation for clinical response to microtubule inhibitors in gastric cancer. *J Gastroenterol* 2006;41(2):133–9.
- [60] Nogami Y, Banno K, Kisu I, Yankura M, Umen K, Masuda K, et al. Current status of molecular-targeted drugs for endometrial cancer (Review). *Mol Clin Oncol* 2013;1:799–804.

FURTHER READING

Ignatov A, Bischoff J, Ignatov T, Schwarzenau C, Krebs T, Kuester D, et al. APC promoter hypermethylation is an early event in endometrial tumorigenesis. *Cancer Sci* 2010;101:321–7.

EPIGENETICS AND EPIGENOMICS ANALYSIS FOR AUTOIMMUNE DISEASES

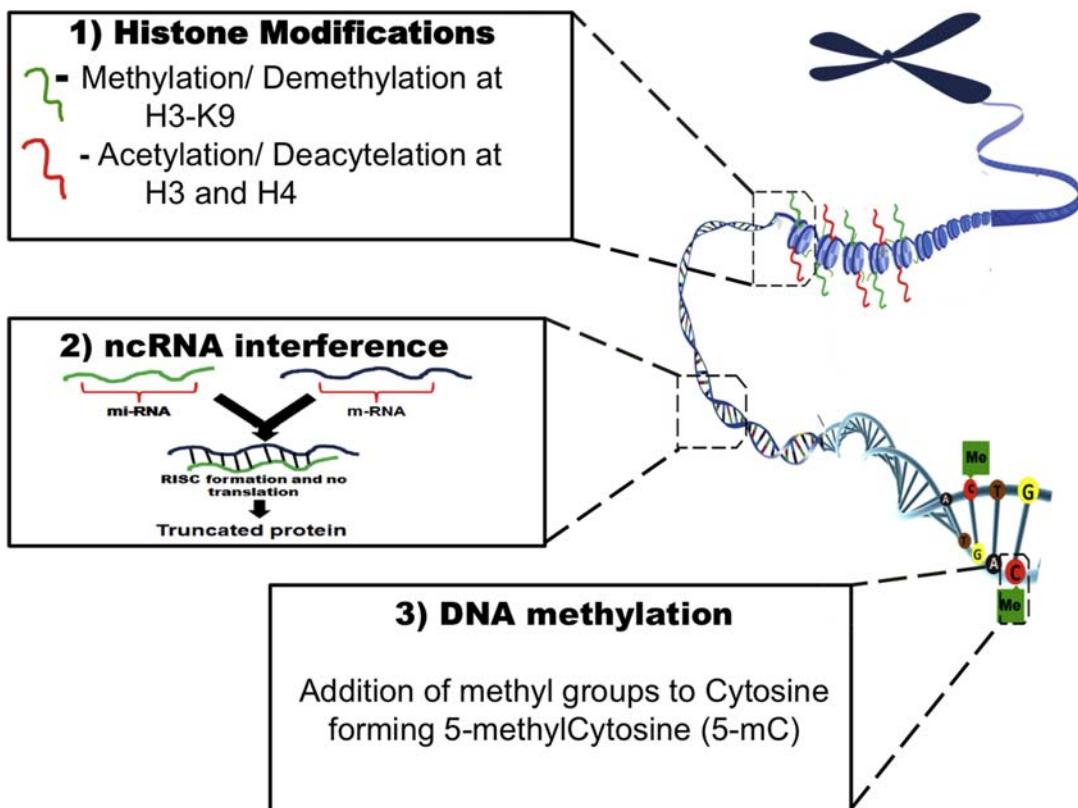
22

Bhawna Gupta¹, Kumar Sagar Jaiswal¹, Arup Ghosh², Sunil Kumar Raghav²

¹School of Biotechnology, Kalinga Institute of Industrial Technology, Bhubaneswar, India; ²Institute of Life Sciences, Bhubaneswar, India

The primary activity of the body's immune system is self-recognition; however, any disturbance in the inherent redundancy of the autoimmune network leads to the formation of a self-centered immune system thereby resulting in target tissue destruction and thus a multiorgan ramification [1]. More than 80 autoimmune diseases have been identified involving nearly any body part. Despite the heterogeneity of the organs affected, autoimmune diseases share epidemiological and clinical features [2]. Altered combinations of gene sequences are known in patients with autoimmune diseases implicating an important role of genomics in understanding the pathogenesis and progression of autoimmunity [3–6]. However, perturbations of the autoimmune network have been seen following changes in both the internal and external environments. Presence of environmental triggers such as pollutants, smoking tobacco, radiations, infectious agents, alcohol, and others play a significant role in generating autoimmune responses [6–8]. A frequent discordance of autoimmune diseases in monozygotic twins also supports the persistence of epigenetics in establishing autoimmune responses [9,10].

The epigenetic research aims to understand the mitotic and/or meiotic alterations in the DNA that are external and heritable. These factors bring stochastic changes in the DNA, changing gene expression patterns thereby manipulating the protein profile and thus may immortalize the autoimmune network. Epigenetic changes have been well reported in autoimmune diseases [11,12] such as rheumatoid arthritis (RA) [13–15], systemic lupus Erythematosus (SLE) [16–18], multiple sclerosis (MS) [19–21], type 2 diabetes (T2D) [22,23], and various environmental factors have been held responsible for creating differential epigenetic marks thereby leading to disease onset and progression [24]. There are a variety of epigenetic modifications (Fig. 22.1) such as DNA methylation, histone modifications, and nonhistone proteins influencing chromatin structure on interaction with histones and DNA in autoimmune diseases [29–31]. These modifications can be readily altered and thus epigenetic information can be reprogrammed dynamically and can also be propagated with substantially lower fidelity [31]. Thus understanding the mechanisms of epigenetic regulation will be imperative not only in identifying the pathogenesis and progression of an autoimmune disease but also in building therapeutic interventions for the betterment of the patients. Moreover, for autoimmune diseases, there is a need to contemplate the extent and mechanism of genetic and epigenetic variations interacting to produce a clinical symptom. It is also important to understand if genetic variations render the DNA sequences unstable and disposed to mutation or is it that the genetic variations drive epigenetic changes by altering the chromatin. A study by Trynka et al. reported colocalization of

**FIGURE 22.1**

Epigenetic regulation works by three different mechanisms. (1) **Histone modifications**—posttranslational modifications of histone proteins such as methylation, acetylation, sumoylation, and ubiquitylation. Such modifications affect gene expression by bringing conformational changes in histone proteins [25]. Histone acetyltransferases (HATs) and histone deacetylases (HDACs) lead to acetylation and deacetylations of histones, respectively [25]. (2) Transcriptional **interferences by ncRNA**—noncoding RNAs such as miRNA, siRNA, and piRNA play important role in gene expressions. miRNAs target specific transcripts or mRNA with complementary sequence and inhibit the complete translation of a functional protein, whereas siRNAs lead to mRNA degradation through RNA-induced transcriptional silencing [26]. (3) **DNA methylation**—addition of methyl groups to the DNA base cytosine leads to the formation of 5-methylcytosine (5-mC), which can modulate transcription by serving as an interaction site for specific chromatin binder or reader proteins [27,28].

H3K4me3 marks in CD4⁺ T cells with 31 SNPs in RA. CD4⁺ T cells [32] thus referring to a potential interactome between the gene modifications and epigenetic variations as well as identifying tissue-specific epigenetic dysregulation.

Recent technical advances have made it possible to measure epigenetic changes in any case-control cohort using state-of-the-art high-throughput techniques such as ChIP-on-chip, bisulfite sequencing, and ChIP-seq. These techniques, however, are dependent a lot on the quality of the tissue

samples collected, the antibodies for a significant and specific chromatin immunoprecipitation, the quality and the sensitivity of the reagents and analytical instruments used and finally the bioinformatics is expected to make significant contributions in data analysis and presentation. The bioinformatics tools will be further used to disentangle the highly interwoven epigenetic mechanisms regulating their target genes and to delineate the complex networks of synergistic and antagonistic interactions bringing about these regulations. The data then need to be statistically powerful and biologically meaningful to understand the development of the autoimmune disease. With the revolutionary changes in bioinformatics abilities we are able to analyze, in parallel, a large number of high-throughput epigenetic data sets within a span of few days.

STUDY DESIGN AND DATA ACQUISITION METHODS

Ascertaining the right time of tissue collection, identification of the admissible tissue samples that will accurately represent the epigenetic manipulations in the disease and understanding the need of either cross-sectional or case-control studies for proper biomarker identification has been always intriguing for any research. It is imperative to understand epigenetic regulation of a particular cell type of the target tissue; however, methodological inability of cell separation techniques produce nonspecific signals and thus may mitigate the biological significance of the analysis. However, recent advances in the analytical tools for epigenome data has made cautious bioinformatics adjustments of mixed cellular populations [33]. Numerous methods and study designs have been used to accomplish epigenome-wide association scan (EWAS) for different autoimmune diseases; however, each of these studies has their own strengths and weaknesses [9]. Studies involving monozygotic twins have contributed a lot toward understanding epigenetic regulations in diseases such as SLE and RA, as they abrogate the effects of genetic variations; however, cross-sectional and case-control cohorts are perchance most frequently included for the ease of tissue collection from a large population set [10]. A major point of concern with the study designs using such cohorts is the appropriate selection of tissues that may be regarded as "controls." A number of EWAS reports in RA used joint tissues from patients with osteoarthritis as controls; however, osteoarthritis is a nonautoimmune inflammatory condition with actively migrating cells in the joints and hence may not be regarded as an ideal control. The longitudinal studies followed over the course of time for a disease with spiked-in internal controls are another good option; however, because of technical reasons comprising continuous patient follow-ups, and invasive procedures of specimen collection, reports of large longitudinal studies are unavailable till date for analysis of autoimmune diseases.

It is although imperative to establish epigenetic changes in specific autoimmune diseases, however, before designing an epigenetic study; we need to explore the already existing information and data sets available. A wealth of epigenetics data has been deposited in public resource databases such as DataBase of human Transcriptional Start Sites (DBTSS), European Promoter Database (EPD), ChromDB, and CHREMOFAC for different autoimmune diseases spanning different tissues and cell types and is accessible for analyses that may help us to identify potential target genes as well as to design further protocols for more robust conclusions about epigenetic associations.

To identify and measure the epigenetic modifications in a cell type, several high-throughput techniques have been developed and are often used. Broadly, these can be array based or rely on the next-generation sequencing (NGS).

MICROARRAY-BASED DETECTION

Microarrays have emerged as an important tool to examine changes involving all aspects of the epigenetic interactions. Microarrays can be customized based on their specific use for identification of regulatory epigenetic marks (Fig. 22.2). Recently, methylation-specific oligonucleotide (MSO) microarrays have been developed for analyzing DNA methylation patterns. These use PCR products of bisulfite-modified DNA as targets followed by synthetic oligonucleotides with the capability of discriminating between methylated and unmethylated cytosine as probes [11]. MSO microarray is capable enough of identifying methylation signals of DNA isolated from tissues and fibroblasts from cancer patients. This approach has also been successful in quantifying DNA methylation levels in autoimmune diseases such as SLE [12], where DNA hypomethylation in CD4+ T cells is associated with disease development [13]. Both of the above studies followed the protocol wherein they used genomic DNA followed by bisulfite processing and sequencing, preparation of methylation-specific probes on the printing arrays followed by subsequent hybridization, scanning, and data processing [11,12]. For the array-based detection, the genomic DNA from cell lines, tissue samples, and peripheral blood mononuclear cells (PBMCs) are used as test samples. Isolation of genomic DNA is followed by PCR to amplify CpG regions of any gene to be tested. The amplified product then undergoes bisulfite treatment to produce uracil from the unmethylated cytosine. This bisulfite-treated DNA can be labeled either at the end terminus with Cy5-dCTP or Cy3-dCTP. Oligonucleotides are designed to incorporate CpG sites of test DNA sample. These synthesized oligonucleotides have an amino-linker C6 [$\text{NH}_2(\text{CH}_2)_6$] attached to its 5' end (IDT). Probes can be printed on slides with the help of any microarrayer. Any microarray scanner such as GenePix 4000A scanner (Axon Instruments) or LuxScan-10 K/A microarray scanner (CapitalBio Corp, Beijing, China) can be used to screen the slides and capture the images. Genepix pro 3.0 is the most preferred program for analysis of captured images or spots. Once base-line correction is done, mean pixel intensities of spots can be statistically processed with any statistical program. MSO microarray was used to identify hypomethylated promoter regions of CD70 and CD11a in SLE and determine the exact methylation status in regulators of genes [12].

Microarrays can also be used for genome-wide histone modification analysis. Acetylation microarrays have been developed to determine acetylation levels in certain genes. This is a combined approach involving chromatin immunoprecipitation (ChIP) and hybridization of DNA to specific probes on microarray glass slides [14–16]. More often, isolated chromatin is immunoprecipitated (ChIP) using antibodies against modified histones such as H3ac, H3K4me2 or H3K27me3. The immunoprecipitated DNA samples are amplified, labeled with Cy5-dUTP and Cy3-dUTP, and hybridized on a microarray chip with unique probes for human genes. This two-color fluorescence labeling technique helps in simultaneous hybridization of samples but separate detection of signals thereby providing a comparative analysis as well as the relative expression of genes. The techniques have been often named as ChIP-on-chip. Hybridization images thus obtained are scanned using a DNA microarray scanner, the color intensities are extracted followed by identification of enriched genomic regions with a peak detection algorithm. The immunoprecipitated chromatin region can be identified, and the gene expression patterns can thus be correlated by gene expression profiling. Using a similar approach, the potential role of miRNAs in the epigenetic control can also be determined by global miRNA profiling [14].

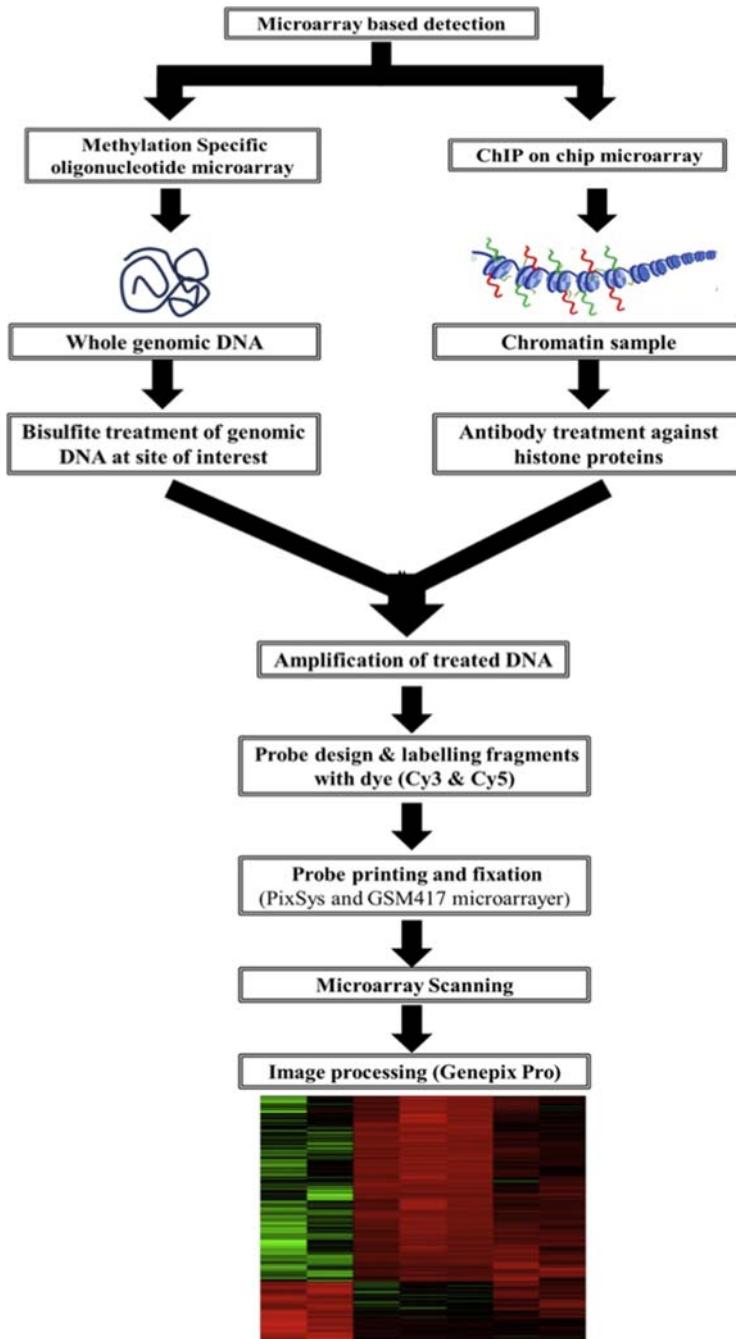


FIGURE 22.2

To decipher epigenetic regulation in autoimmune disease two different approaches in microarray can be used. Methylation specific oligonucleotide microarray is the composition of bisulfite treatment of DNA to convert unmethylated cytosine to uracil and then hybridization with specific probes. ChIP on chip microarray combines chromatin immunoprecipitation with the help of histone protein antibodies and then hybridization with specific probes. In the end both of these method generate fluorescence and which can be analyzed with the help of different softwares thereby generating mean pixel intensities, which can be used to generate heat-map for gene association studies.

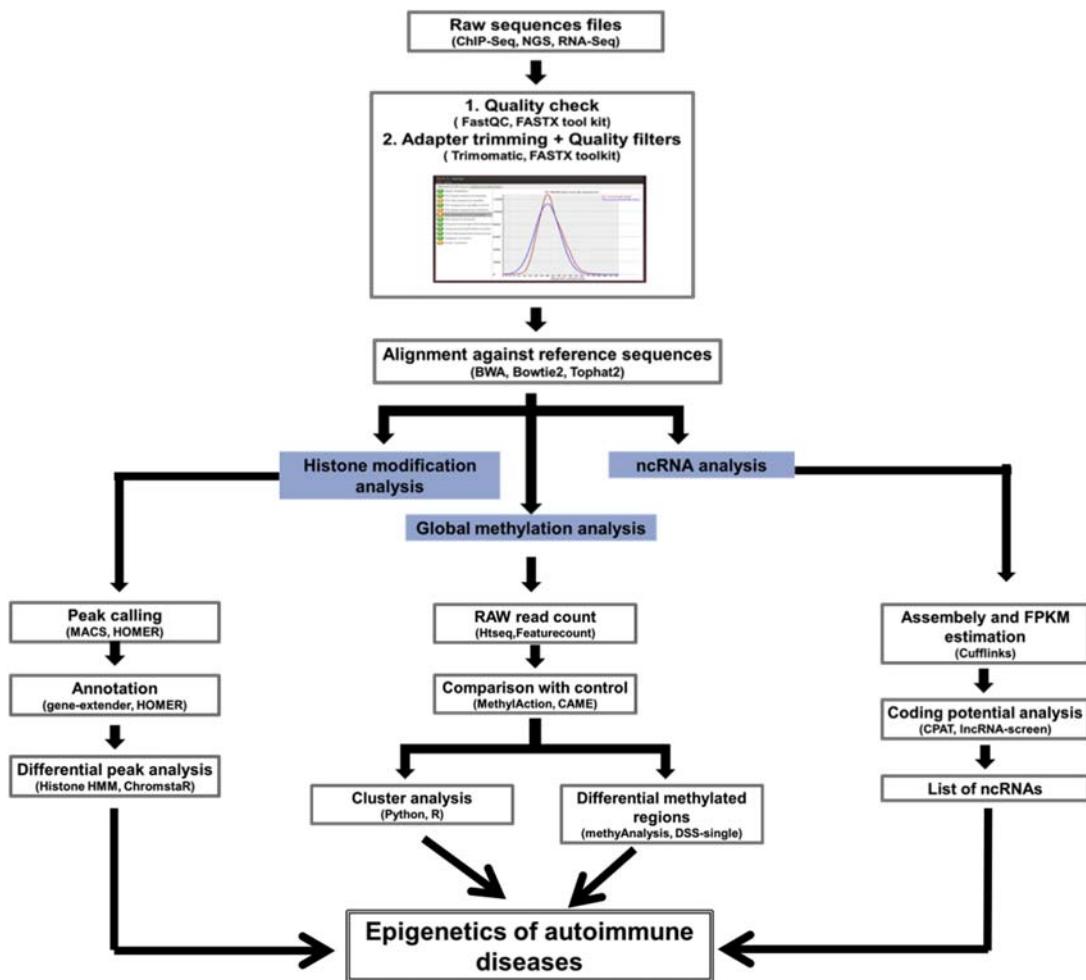
In autoimmune diseases, nucleic acids from PBMCs and local tissue samples are more preferred for assessing epigenomic regulation. Synovial fibroblasts and inflamed lower intestinal mucosa constitute tissue samples for RA and inflammatory bowel disease, respectively [17]. In autoimmune diseases, inflammatory responses are generated through collective action of genetic factors from macrophages, T cells, synovial fibroblasts, growth factors, inflammatory cytokines [18], and chemokines [19]. All of these biological entities can be found in human blood, hence blood impose as one of the best samples for microarrays.

NEXT-GENERATION SEQUENCING

NGS has paved the way for sequence analysis on every region of the genome as well as transcriptome. Analyzed regions include whole genome classified as whole genome sequencing, whole exons as whole exome sequencing, RNA sequencing (RNA-seq), chromatin immunoprecipitation followed by NGS (ChIP-seq), and DNA methylation sequencing. NGS has allowed us to identify DNA sequences from the rarest of the samples with ultimate accuracy using customized approaches (Fig. 22.3) thereby delineating the exact gene loci involved in pathogenesis as well as the progression of the diseases.

ChIP-Seq is one of the important applications of NGS technology [20–22]. This differs from microarray as precipitated DNA fragments are sequenced rather than probed and visualized through fluorescent tagging [23]. ChIP-seq is more preferable for identification of genomic regions bound by transcription factors to identify the enhancers/repressor elements and the underlying DNA sequence motifs [34]. This technology has excelled in epigenomics including posttranslational profiling of histone modification marks, to identify chromatin structure and nucleosome positioning [35]. Like ChIP-on-chip, in ChIP-seq antibodies against specific protein of interest either transcription factors or specific to histone tail modifications are used to immunoprecipitate the protein-bound chromatin complexes followed by high-throughput sequencing of pulled down and purified DNA fragments. For sequencing purposes, several NGS sequencers are available with different vendors and well-established NGS analysis pipelines are widely reported in literature. For example, if sequencing has to be done in Illumina Genome Analyzer; whole process starting from sample preparation up to obtaining data is well described [29]. The raw sequences obtained could be further analyzed for motif discovery. Sequences are purged to remove any kind of overlapping or duplications of peak using software (Table 22.1) such as PeakSplitter [30] can be used. Then heat maps and positional profiles of nucleotides and dinucleotide compositions are generated. Motifs discovered through sequence analysis are then compared using annotated motif databases such as JASPER to find out associating transcription factors [31].

RNA sequencing is the newly developed method for analyzing mRNA expression wherein the sequencing pipeline is similar to that of ChIP-seq except that the starting material is RNA instead of fragmented DNA [44]. Careful evaluation of RNA including small RNAs (sRNAs) [45] have revealed the active participation of RNA in disease development, especially microRNAs (miRNAs) [46]. Although not translated, miRNAs regulate the process of protein-coding transcripts and can control synthesis of proteins important for pathogenesis and or progression of diseases. As miRNAs are short sequences, NGS has played an important role in analyzing their presence and association to different autoimmune diseases. Although fragility of RNA limits the use of RNA-seq [45,46], careful handling and in-depth investigation can help to analyze the entire transcriptome, including differential splicing and allelic expression [47,48]. miRNA transcriptome analysis of resting and cytokine-activated mouse

**FIGURE 22.3**

Overview of sequencing-based epigenetics analysis for autoimmune diseases. Different DNA modifications can be analyzed following specific pipelines and software.

NK (natural killer) cells was done through the use of Illumina GA and ABI SOLiD platforms. Unique bioinformatics pipeline was used to analyze 302 known and 21 novel mature miRNAs from small RNA libraries of NK cell [49]. Illumina platform is widely acceptable for analyses of genes with lesser expression, alternative splice variants, and novel transcripts [50]. High level of correspondence was observed between SOLiD platform and Affymetrix Exon 1.0ST arrays on comparison of RNA-seq data in terms of exon-level fold changes and detection. Extremely low background error rate helped in obtaining the greatest detection correspondence in RNA-seq analysis. The discrepancies between

Table 22.1 Tools for MeDIP Microarray Data Analysis

Tool	Use	Source	References
MEDME	Enhanced estimate of DNA methylation	http://bioconductor.org/packages/release/bioc/html/MEDME.html	[36]
Batman	A Bayesian deconvolution strategy for DRM identification	https://github.com/dasmoth/batman	[37]
BayMeth	Empirical Bayes model based on the Poisson distribution	https://bioconductor.org/packages/release/bioc/html/RepiTools.html	[38]
ChIPOTle	An excel macros for identification of sites of protein–DNA interaction	https://doi.org/10.1186/gb-2005-6-11-r97	[39]
PAVIS	Offering two main functions: annotation (relative information between quay peaks and other peaks in genome) and visualization (simultaneous view of multiple peaks in the context of genomic features)	https://manticore.niehs.nih.gov/pavis2	[40]
EaSeq	GUI controlled peak-finding, quantitation, normalization, clustering, distance analysis, randomization, scoring of ChIP-seq data	http://easeq.net/	[41]
MMDiff	Kernel-based method for detection of statistically significant difference between read enrichment profiles in different ChIP-Seq samples.	http://www.bioconductor.org/packages/3.2/bioc/html/MMDiff.html	[42]
ChIP-Array 2	Web-based tool integrating transcription factor (TF) binding and transcriptome data to construct a gene regulatory network	http://jjwanglab.org/chip-array-v2/	[43]

RNA-seq and transcriptomics analysis on SOLiD platform are not crisp and clear as observed in Illumina [51].

NGS technologies have overcome the major problems of the microarrays by eliminating the need of species-specific probes and the need of long probes. In addition, ChIP-seq can generate data with higher accuracy, which is optimal for chromatin protein footprinting [52]. There are several commercially available sequencers with different coverage and base calling accuracy rates, such as Illumina Genome Analyzer Platform can provide 28–100 bases per sequencing read, Roche 454 generates 250–400 bases, and SOLiD from Applied Biosystems (Thermofisher) can generate 4–30 base pair reads. As the histone protein-binding regions have ~20-bp footprint, researchers usually opt for higher coverage rather than long reads. In addition, paired end reads increase the accuracy during peak calling.

Although widely acceptable for research in autoimmune diseases, NGS has its own limitations. NGS produces large amount of data, which poses challenge for data analysts. Sometimes, this also brings errors because of its low detection capability on low-frequency mutations because of shorter and repeated reads. Available bioinformatics tools for NGS data analysis are time-consuming and require professionals to run it, which may also limit the use of NGS.

EPIGENETIC CHANGES IN AUTOIMMUNE DISEASES

Autoimmune diseases are a result of immune system failure to differentiate between self- and nonself, thereby generating an immune response against its own cells and tissues. A large number of researches have been performed using blood from patients with autoimmune diseases because of easy availability of blood samples for analysis. With already established techniques for separation of different blood cells, it has become possible to study and analyze perturbed epigenetic landscapes in autoimmune diseases [53] and generate cell-specific signatures [54–58]. However, target tissue–specific cells have also been frequently isolated and studied for their epigenetic modifications correlating to autoimmune responses [59].

RHEUMATOID ARTHRITIS

RA is a systemic autoimmune disease characterized by the progressive destruction of joints specially small joints with activated synovial fibroblasts [60,61], accumulated synovial fluid [61], selective migration of immune cells at the site of inflammation [62,63] leading to severe pain, joint stiffness, distorted joints, and thereby reduced mobility. Almost 1% of the world population suffers from RA with the disease-affecting females more than the male population [64]. A number of nongenetic factors have been involved in pathogenesis of RA including smoking [65], alcohol [66], inappropriate lifestyle [67,68] as well as environmental pollutants [69–71]. An altered gut microbiota is proposed to be the cause of RA [72,73], whereas a number of antiinflammatory food have been seen to relieve disease symptoms [71,74]. All these reports eventually point toward an epigenetic modulation of the disease. These studies have used state-of-the-art techniques including ChIP-Seq, rtPCR, bisulphite sequencing followed by curated computational tools to identify epigenetic signatures for RA.

RA synovial fibroblasts (RASFs) are reported to be the key players in onset and progression of the disease [75]. Hypomethylation of DNA has been frequently observed as an effect of deficiency of DNA methyltransferases (DNMTs) in proliferating RASF [76,77] that leads to increased translation of matrix-degrading enzymes, adhesion molecules, and receptors involved in the pathogenesis of RA [76–78]. Studies designed to target the deregulated epigenome of RASF show a better protective strategies for prolonged remission of RA [79,80]. These reports bring a new dimension to the current therapeutic interventions where prolonged remission is difficult and a frequent relapse appears because of stable epigenetic modifications continuously changing DNA configuration. Moreover, using techniques such as pyrosequencing and ChIP-Seq, analysis of PBMCs has confirmed hypomethylation at CpG sites of several inflammatory cytokines leading to their increased expression [81,82] with further deleterious effects in RA patients. Hypomethylation in CpG sequences presents upstream of L1 ORF and IL6 promoter in monocytes are generally due to reduced expression of DNMTs, which in turn leads to increased expression of certain growth factors, receptors, adhesion molecules, and cytokines causing irreversible phenotypic changes in different cells, for example, in RASF [83,84]. Death receptor-3 (DR-3) is a proapoptotic protein activating NF- κ B. Changes in promoter methylation of CpG islands within DR-3 gene and on IL6 promoter, analyzed using bisulfite sequencing and qRT-PCR, leads to a decrease in DR-3 protein expression resulting in apoptosis-resistant RA synovial cells as well as reducing IL6 promoter transcriptional activity and thereby gene expression of IL6 gene [85–88].

Altered patterns of histone modifications are also frequently observed at the promoters of crucial genes in RASF [60,61,89] with an imbalance in the expression of histone acetyltransferases (HATs) and histone deacetylases (HDACs). Matrix metalloproteinases (MMPs) and enzyme from ADAMTS family (a disintegrin and metalloproteinase domain with thrombospondin motifs) cause heavy cartilage destruction [90,91], and their genes are reported to be controlled by chromatin modifications and histone acetylations [92–94]. Hence HDAC inhibitors are now considered as new chondroprotective therapeutic agents to suppress destructive MMPs and ADAMTs in synovial tissues [95–97]. Recent genome-wide study by de la Rica et al. compared RASF DNA methylation with miRNA expression and RASF transcriptome data from the Gene Expression Omnibus (GEO) in an integrated analysis. Several target genes including IL6R, CAPN8, and DPP4 had differential methylation status. Around 200 out of the 714 genes identified had difference in methylation and had decreased gene expression. This led to identification of several novel and known-associated miRNAs. According to their report, many of the CpG sites were hypermethylated leading to decreased expression of miRNA, but four CpG sites with hypomethylation led to increase in miRNA expression [98].

Unique H2A histone variants such as macroH2A are known to interfere with the interaction of transcription factor NF- κ B in RA leading to prevent the function of some nucleosome-restructuring proteins [99]. Moreover, microarray analyses of TNF- α -stimulated RASFs and PBMCs from RA patients show constitutive overexpression of microRNAs such as miR-146, miR-203, and miR-155 [100]. TNF- α and IL-1 β enhance the expression of miR-155, which inhibits the regular function of MMPs [100]. Using microarray and rtPCR analyzes, researchers observed that the CD4+ T cells from RA patients when treated with proinflammatory cytokines upregulate miR-146, which in turn downregulates NF- κ B pathway [101–103]. miR-203 also causes repression of MMPs and inhibition of IL-6 [104]. miR-124 targets cyclin-dependent kinase 2 (CDK-2) and monocyte chemoattractant protein 1 (MCP-1) leading to increased cell proliferation [105,106].

More recently, researchers are able to comprehend that some of these epigenetic differences can be attributed to the altered frequency as well as the heterogeneity of the cell populations. We thus need new and better technologies to effectively measure and filter out methylation changes due to altered cell frequencies and those that are likely to cause the diseases and also those that are a consequence of the disease. Recently, Liu and colleagues performed a genome-wide epigenome analysis on PBMCs using the Illumina 450K methylation bead array followed by a series of statistical algorithms to reduce the confounding factors that have hampered the previous analyses [107]. Using this approach, they could identify two clusters within the MHC region with differential methylation patterns that potentiate pathogenesis of RA. Moreover, cellular heterogeneity may mislead and dilute the significance of epigenetic landmarks for a disease hence epigenetic analysis at the single cell level will enable detailed investigations of epigenetic states associated with each clonotype of a specific cell population. Although methods for parallel single-cell genome-wide methylome and transcriptome sequencing have been developed [108–111], still technical advances are urgent.

Most of these studies have analyzed total or mixed cell populations of PBMCs; however, owing to the cell specificity of methylation marks, these reports may thus skew the results and interpretations [112,113].

SYSTEMIC LUPUS ERYTHEMATOSUS

SLE is one of the most explored autoimmune diseases with respect to epigenetic modifications. SLE is a chronic autoimmune condition, with a characteristic feature of producing autoantibodies against a variety of nuclear antigens, which affects almost any organ of the body [114–117]. Methylation status of certain genes in T cells of SLE patients has been measured using techniques including DNA microarrays, bisulfite sequencing, and qPCR. A reduced methylation of SLE accordant genes including ITGAL (integrin subunit alpha L) (CD11a), CD40LG (CD40 ligand), PRF1 (perforin 1), CD70, IFGNR2, MMP14, LCN2 (lipocalin-2), and rRNA (18S and 28S) gene promoters has been frequently reported [10,89,118–120]. DNA hypomethylation affects the T-cell chromatin structure leading to overexpression of these genes resulting in cell hyperactivity and generation of immune and inflammatory response [121–123]. Hypomethylation in E1B promoter region of CD5 in B cells lead to autoimmunity. This is because of reduced expression of DNMT1 ultimately leading to defective cell signaling [124].

Studies on murine and human models of SLE have shown that during apoptosis histone gets modified to generate immunogenicity. These antibodies are directed toward released nuclear components postapoptosis leading to pathogenesis of SLE [59,125]. Apoptosis leads to changes in nucleosome structure generating immunogenic epitopes, which further form autoantibodies against some chromatin components [60,61]. Histone modifications such as histone-3-lysine4 trimethylation (H3K4me3), histone-3-lysine8 (H4K8) triacetylation, histone-3-lysine27 trimethylation (H3K27me3), and histone-2B-lysine12 acetylation (H2BK12ac) are reported to facilitate increased apoptosis in nucleosomes thereby generating autoimmunogenicity with activation of antigen-presenting cells, production of autoantibody, and a subsequent inflammatory response [62,63]. Certain studies have looked on global acetylation and have found an altered pattern in histone H3 and H4 of active SLE CD4+ T cells [64]. ChIP-chip analysis of monocytes from SLE patients has shown changes in acetylation status of histone H4 with an increase in expression of interferon family genes playing a key role in progression of SLE [65–67].

MULTIPLE SCLEROSIS

MS is a chronic inflammatory disease and leads to neurodegeneration due to destruction in myelin sheaths [40,41]. According to recent reports, hypomethylation in promoter region of peptidyl arginine deiminase type II (PAD2) have been held responsible for the progression of this disease [68]. PAD2 is one of the important enzymes involved in citrullination of myelin basic protein (MBP). This whole process of citrullination is associated with important biological effects including promoting protein autocleavage, increasing the chances of generating and creating new epitopes, and also modulating the immune system [43,126,127]. In MS, activity of demethylase gets enhanced leading to hypomethylation in promoter region of the PAD2 [69]. This hypomethylation results in an increased expression of PAD2 that in turn increases the citrullination of MBP leading to a hike in the production of immunodominant peptides. These immunodominant peptides are involved in autocleavage of MBP with subsequent irreversible changes in its biological properties leading to proteolytic digestion, myelin instability, and a chronic inflammation response [70–72].

An altered pattern in histone acetylation of white matter is also been observed in patients with MS. The level of oligodendrocyte differentiation inhibitors such as ID2 (inhibitor of DNA binding 2), TCF7L2 (transcription factor 7-Like 2 T-Cell-Specific, HMG-box), and SOX2 SRY (sex-determining region Y-box 2) increases because of hyperacetylation of H3 in their promoter regions [73].

TYPE 1 DIABETES

Type 1 diabetes (T1D) is mostly attributed to T cell wherein it is known to be a T cell–mediated autoimmune disease involving the destruction of insulin-producing pancreatic β cells [74]. GWAS studies and metaanalyses have shown people harboring genetic polymorphisms in MHC class II especially at DR, DQ loci, PTPN22, CTLA4, and IL-rRA [75,76]. ChIP-seq analysis of T cell has assessed damage in pancreas, which causes generation of autoimmune response leading to cellular proliferation, all of which can change the status of epigenomics [77,78].

During new onset of T1D, methylation level of insulin gene increases several times in patients compared with healthy persons [79]. Other factors such as lymphocyte maturation and expression level of cytokines are hugely affected due to epigenetic modifications [80]. Identification of the epigenetic modification behind this phenomena was done through ChIP-chip analysis of blood samples from patients, and it was found to be an altered H3K9 demethylation leading to overexpression of CTLA4 [128]. Overexpression of miR-236, a specific microRNA, has correlated itself with disease severity [129]. Assessment of miRNA levels in PBMC of T1D patients has shown upregulation of miR-146a and downregulation of miR-20b, miR-31, miR-99a, miR-151, miR-125b, miR-335, miR-365, and miR-100 [130]. miR-21a and miR-93 expression level goes down in PBMCs of T1D patients when compared with healthy persons [85].

ANALYZING EPIGENETIC CHANGES IN AUTOIMMUNE DISEASES

The immune system is the primary interface between humans and the environment and thus both genetic and epigenetic modifications have a great impact on the development of autoimmune diseases. The study of epigenetics was limited to DNA methylation, different types of histone tail modifications, and changes in chromatin remodeling machinery [86,87]. However, with the recent technical advancements, other important epigenetic aspects such as transcription regulation by noncoding RNAs also have become an integral part of these studies [88]. In this chapter, we will thus discuss the epigenetic basis of autoimmune diseases in context of computational methodologies used for the detection of DNA methylation, histone tail modifications, and regulation of noncoding RNA.

DNA METHYLATION

One of the most studied epigenetic modifications is the methylation patterns of cytosine in CpG dinucleotide-rich regions. After the discovery of methylation-sensitive restriction enzymes, it was observed that CpG sites were either completely unmethylated or completely methylated [90]. Although southern blot analysis using methylation-sensitive restriction endonucleases provides a basic idea about methylated DNA, it requires a large amount of quality DNA and incomplete digestion incorporates sequence biases. There are several conventional techniques such as PCR and restriction digestion; however, for genome-wide detection of methylation sites and methylation patterns

methylation-specific restriction digestion (HELP-seq, MSCC), bisulfite treatment (WGBS), immunoprecipitation (MeDIP, MIRA-seq)-based techniques coupled with microarray or massive parallel sequencing [91] as well using double-stranded DNA-binding antimethylcytosine proteins (MBDs) such as MBD1, MBD3L1 [131] are frequently used. Except of the strand sensitivity both MeDIP and MBD data analysis strategies are similar in nature and are grouped as affinity-based enrichment methods. Although the biochemical methods for detecting DNA methylation are different yet the data analysis strategy is similar in case of both methylation-specific restriction digestion and 5'mC (5' methyl cytosine) immunoprecipitation, we may thus categorically group the computational tools needed for methylation data analysis.

Data Analysis Post Immunoprecipitation Studies

In enzyme-based genome-wide methylation discovery, antimethylcytosine-based immunoprecipitation provides more global and less biased information than methylation site-specific restriction digestion [132]. The basic principle of methylated DNA immunoprecipitation or MeDIP-chip is similar to that of the transcription factor–binding ChIP-chip experiments. Genomic DNA is sheared down to a size of about 400 bp, then target sequences are enriched using specific antibodies and finally input DNA (experimental control) and methylated DNA sequences labeled with Cy3 (green) and Cy5 (red) are hybridized to microarray slides containing probes for annotated human CpG islands [133].

After signal intensities from the chip are reordered, they go through several quality control and normalization steps to reduce technical and dye-related biases before driving any inference from the data. Methylation log ratio known as M is calculated using signal intensity in input sample as a background for comparison with the methylated DNA sample. However, the M value cannot define the exact methylated regions, the nonlinear relationship between enrichment, and expected methylation value is captured by multiparameter logistic regression model or Bayesian deconvolution strategy [37,134]. Table 22.1 shows different analytical tools that may be considered for analysis of MeDIP data.

Data Analysis Post Bisulfite Treatment

The bisulfite treatment of DNA using commercially available kits such as EZ-96 Methylation MagPrep (Zymo) converts the unmethylated cytosine into uracil, whereas the methylated cytosines remain unchanged [98,135]. The DNA is fragmented as discussed above, and the methylation levels of the fragmented DNA are determined by hybridizing to CpG-specific microarray chips such as Human-Methylation450 BeadChips (Illumina) [98,136,137]. The signal intensities acquired from microarray chip imaging are converted to β scores ranging from 0 to 1 using computational tools such as BMIQ wherein 0 denotes no methylation and 1 represents completely methylated DNA [136,137] that uses the following equation for β -score calculations:

$$\beta = \frac{M}{M + U + 100} - 1$$

where, M = methylated DNA intensity, U = unmethylated DNA intensity.

The β scores are then normalized for correcting probe design bias (that may result in type-1 and type-2 error for Illumina Human Methylation 450 platform) [138]. In clinical samples, high variance is an expected phenomenon and as a quality control measurement, principal component analysis is performed to filter out the outliers. Once the preprocessing and normalization is done, genome-wide

Table 22.2 Tools for Bisulfite Microarray Data Analysis

Tool	Use	Source	References
ComBat	Batch effects removal	http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html	[126]
BMIQ	Probe bias correction	https://code.google.com/archive/p/bmiq/	[138]
GenomeStudio	Genotyping analysis of microarray data	https://sapac.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html?langsel=/in/	NA
iEVORA	Differentially methylated CpG identification	https://github.com/aet21/iEVORA	[127]
Lumi	Quality control, variance normalization, etc.	https://doi.org/10.18129/B9.bioc.lumi	[140]
BSmooth	Alignment, quality control, and analysis pipeline; precise data generation with low coverage of input	http://rafalab.jhsph.edu/bsmooth	[141]
MethPipe	Pipeline for analysis of whole genome bisulfite sequencing and reduced representation bisulfite sequencing (RRBS) data	http://smithlabresearch.org/software/methpipe/	[142]
MOABS	Analysis of large-scale DNA methylation data from single cytosine level to region level	https://code.google.com/archive/p/moabs	[143]

methylation pattern is analyzed using standard SNP-calling protocols to get genotypic data for quantitative trait locus association [139]. Table 22.2 shows different computational tools that may be used to analyze the bisulfite microarray data.

Although the microarray-based detection is often used as a high-throughput method for detecting DNA methylation patterns, however, even with better and newer platforms such as Illumina Infinium HumanMethylation450 BeadChip only about 450,000 out of nearly 55 million CpG present per diploid cell can be analyzed.

Data Analysis Post Next-Generation Sequencing

NGS or massive parallel sequencing has changed the definition of modern-day high-throughput studies by providing true single-nucleotide resolution. This new technology removed the biases and limitations that microarray chip-based system had [144]. As the analysis costs came down, the usage and acceptability of this platform increased tremendously to understand disease pathology as well as for diagnostics and prognostic purposes.

The DNA sample (library) preparation method is almost the same as that followed for microarray. Different platforms for sequencing provided by companies such as Helicos Bioscience, Illumina, ABI

Biosciences use different approaches such as sequencing by synthesis, clonal cluster, or sequencing by ligation and thus generate several hundred gigabases of DNA sequences [145]. However, the analysis of any NGS data starts with primary quality control measures such as checking the quality of raw reads by Phred score, the adapter and low-quality base trimming, GC bias and duplicate sequence removal (e.g., using PICARD, DEDUPE tools). The next major challenge is to map the randomly assigned reads to respective regions in the genome and various sequence alignment tools are then referred. For the bisulfite sequence alignment files wherein the uracil (U) in bisulfite-treated DNA is converted to thymine (T) resulting in four different strands of DNA from a single loci amplification [146], computational tools such as Bismark are frequently used that convert C to T and G to A in directional or nondirectional sequence data and alignment is carried out with Bowtie with methylation-specific index files [146,147]. Another tool named BS Seeker uses a similar approach but is limited only to single-end read alignment that outperforms most of the bisulfite sequence alignment files [107]. Further analysis and visualization of methylated loci is done using a set of tools such as MethTools, Bis-SNP, QUMA, CpG PatternFinder ([Table 22.3](#)).

While bisulfite sequence analysis is used for quantification of DNA methylation, the immune precipitation sequencing (MeDIP) or methylcytosine binding (MethylCAP) based technique is used for detecting methylated regions for understanding differentially methylated regions (DMRs) between case and control groups; however, after target enrichment and barcoding, the library is sequenced using any of the tools mentioned earlier in [Table 22.3](#) [108]. After alignment using any short-read aligners such as BWA-MEM and Bowtie2, duplicate and low-quality reads are filtered out. Then detection of methylated regions is done using ChIP-seq peak callers such as MACS2 and FindPeaks [109,110]. From the peak calling, information of genomic regions with methylation enrichment is obtained, which is used for DMR detection before a sample-wide normalization. There are several R packages and stand-alone tools available for DMR detection; some of them are listed in [Table 22.4](#).

Table 22.3 Tools for Bisulfite Sequencing Data Analysis

Tool	Use	Source	References
BSMAP	DNA methylation sequence aligner	https://github.com/zyndagj/BSMAP	[148]
Bismark	DNA methylation sequence aligner	https://www.bioinformatics.babraham.ac.uk/projects/bismark/	[146]
BS Seeker	Single-end BS-seq aligner	NA	[107]
MethTools	Analysis and Visualization	https://genome.imb-jena.de/methtools/	[149]
QUMA	Methylation quantification and visualization	http://quma.cdb.riken.jp/	[150]
Bis-SNP	DNA methylation calling	http://people.csail.mit.edu/dnaase/bissnp2011/	[151]
Minfi	Tools to analyze and visualize Illumina Infinium methylation arrays	http://www.bioconductor.org/packages/release/bioc/html/minfi.html	[152]
coMET	Visualisation of regional epigenome-wide association scan (EWAS) results and DNA comethylation patterns	http://www.bioconductor.org/packages/release/bioc/html/coMET.html	[153]

Table 22.4 Tools for MeDIP Data Analysis

Tool	Use	Source	References
MEDIPS	DRM finding, visualization	https://bioconductor.org/packages/release/bioc/html/MEDIPS.html	[112]
csaw	Differential binding	https://bioconductor.org/packages/release/bioc/html/csaw.html	[154]
DISMISS	Strand-specific methylation detection	https://github.com/uhkniazi/dismiss	[155]
ChIPpeakAnno	Annotation	https://bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html	[156]
MeQA	Quality control and analysis pipeline	http://life.tongji.edu.cn/meqa/	[157]

Although for both bisulfite sequencing and MeDIP, MethylCAP uses two different strategies for genome-wide methylation site discovery, often they are used as complementary methods to validate one another.

Most of the studies done to date regarding methylation patterns in autoimmune disorders are mostly array based. DNA methylation studies in some well-known autoimmune disorders such as Sjögren's syndrome, T1D, and SLE suggest a direct association of methylation pattern and immune response-related gene expression [98,111,112].

Monozygotic twin pair is a great example for identification of genetic and epigenetic components in pathogenesis. Although a lot of studies have confirmed some the genetic components of T1D, the epigenetics part is still poorly understood. The above heatmap is the visual representation of differential methylation pattern in monozygotic twin pairs. The study by Rakyan et al. [113] recruited 15 monozygotic twin pairs, one with diagnosed with diabetes below the age of 20 years and other with a low diabetic risk. The methylation levels were determined using Illumina HumanMethylation27 BeadChip, and the raw signal is plotted using R package gplots (<https://cran.r-project.org/package=gplots>) after scaling showing a pattern in DNA methylation between the two groups, represented as two primary branches of the tree. A visual representation of obtained results can be found in Fig. 22.4.

HISTONE MODIFICATION ANALYSIS

Another major component of epigenetic modifications leading to gene expression disruption depends on the modifications of the histones. There are several known histone tail modifications, and the field is expanding further with the help of NGS. The most frequently studied histone modifications include methylation and acetylation at the lysine residue. The protocols and strategies used for DNA library preparation to analyze histone modifications for the chip-based assay or for NGS is very similar to those of MedDIP; however, the antibodies used for fragment enrichment is specific to the histone mark and the needs are to be studied. In addition, the tools and algorithms used in bioinformatics analysis of histone ChIP-seq data are same as the ones discussed above for the immunoprecipitation sequencing files.

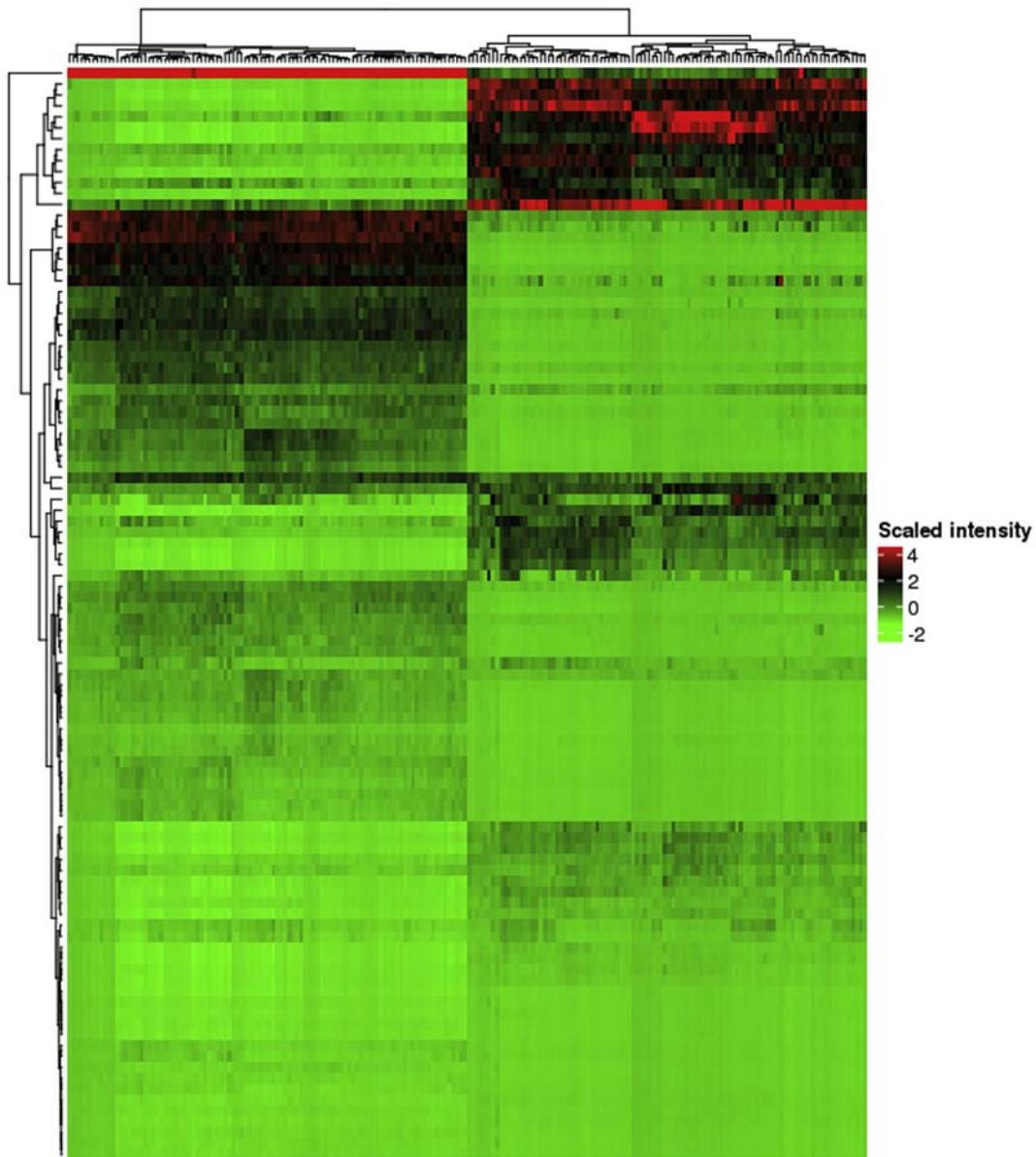


FIGURE 22.4

Clustered heatmap of DNA methylation pattern in CD14 + Monocyte cells from discordant monozygotic twins with and without childhood-onset of type 1 diabetes. (Microarray data GSE56606).

Raw reads after quality control are aligned to reference genome and then are randomly subsampled to a significant size for peak calling. The input DNA is also maintained as the negative control for true positive peak calling, and further analysis is done using differential binding detection tools to get comprehensive modification marks important in the disease pathophysiology. The Encode project has a detailed guideline of histone ChIP-seq data processing and analysis for experiments with or without any replicates [114].

Several histone modifications are known today with well-studied roles in cancer biology or development of an organism, but unlike transcription factor–binding events, the histone modification has a unique challenge in determining well-defined binding peaklike feature detection using conventional tools [115]. To address these challenges in histone modification data analysis, there are several specialized tools available, some of them are mentioned in Tables 22.5 and 22.6.

One of the preliminary ways to check patterns of NGS data after alignment is to use a genome browser. Integrative Genomics Viewer (IGV) [116] by Broad Institute is one of the most popular tools for visualization and exploration of sequencing data. To showcase the different methylation patterns in human primary macrophages and monocytes, the Wiggle (Wig) formatted data loaded with associated BED files containing detected using MACS2. The peaks in data represent corresponding tag density of associated region marked by a bar below the region. An example of obtained results has been depicted in Fig. 22.5.

MiRNA AND TARGETS PREDICTION

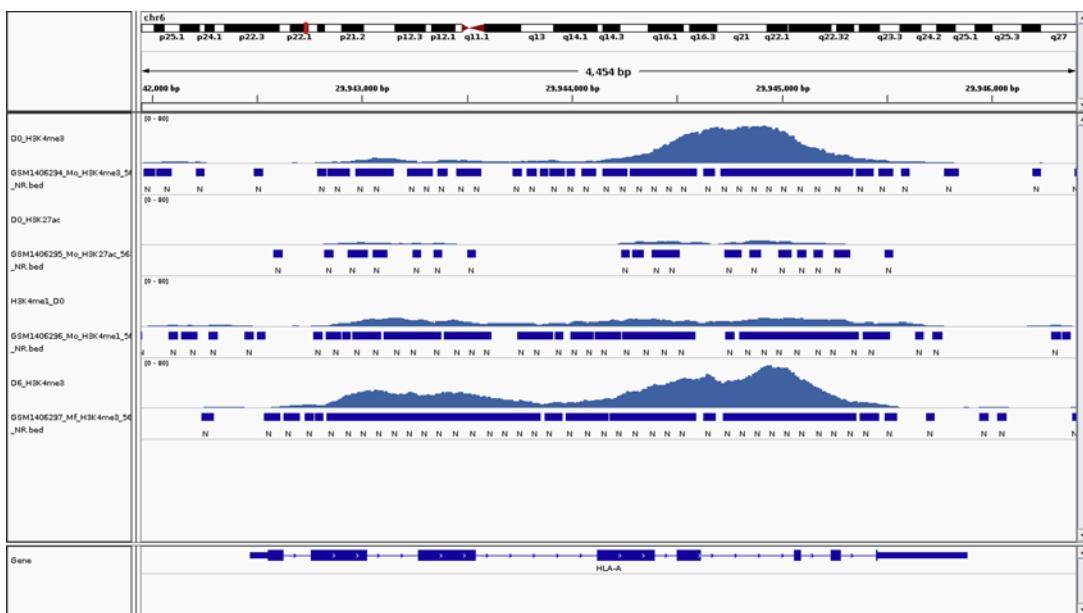
A microRNA is a group of small noncoding RNAs that after maturation bind to target mRNAs and form a multiprotein RNA-induced silencing complex (RISC). In immature form, they are about 77 nucleotides long but after maturation, the length is reduced to 18–22 nucleotides [117]. For the detection of miRNA in pathological conditions, both RNA-seq and miRNA-seq can be used; however, with higher resolution, cellular abundance of miRNA is very low. The sequence aligners such as Bowtie2, BWA-MEM are again used with several references such as ribosomal RNA, tRNA, and the unaligned reads are aligned to miRBase in a sequential manner [160]. In another approach based on reading length distribution from the aligned file the miRNA reads can be identified in the range of

Table 22.5 Some of the Popular Histone Modification Detection and Peak-Finding Tools

Tool	Use	Algorithm	Source
histoneHMM	Differential analysis of histone modification	Hidden Markov model	http://histonehmm.molgen.mpg.de/
HOMER	Peak calling	Hypergeometric optimization	http://homer.ucsd.edu/
Epigramp	Epigenome status form DNA motifs	Fast edge preserving Bayesian reconstruction	http://wanglab.ucsd.edu/star/epigramp/
Segway	Annotation and signal detection	Gaussian mixture models	https://hoffmanlab.org/proj/segway/
MACS 1.4	Peak calling	Model-based analysis	http://liulab.dfci.harvard.edu/MACS/
ChromHMM	Chromatin state detection	Hidden Markov model	http://compbio.mit.edu/ChromHMM/

Table 22.6 Tools for Analysis of Histone Modification ChIP-Seq Data

Tool	Use	Source	References
MACS	Peak calling from ChIP-seq short reads	http://liulab.dfci.harvard.edu/MACS/	[158]
histoneHMM	Differential histone modification analysis	http://histonehmm.molgen.mpg.de/	[115]
HOMER	Peak calling, motif search	http://homer.ucsd.edu/homer/	[159]
Pavis	Annotation	https://manticore.niehs.nih.gov/pavis2	[40]
DiffBind	Differential binding analysis using ChIP-seq data	http://bioconductor.org/packages/release/bioc/html/DiffBind.html	NA

**FIGURE 22.5**

Enrichment of different histone modification patterns on HLA-A gene (MHC-1) in human primary monocytes (Data source GSE58310), visualized using IGV genome browser.

18–24 nucleotides. There are several pipelines and Web server–based tools available for association with gene expression patterns. Broadly, they can be grouped into two categories as tools to analyze differential expression (Table 22.6) and tools for miRNA target prediction (Tables 22.7 and 22.8).

The role of miRNA in gene expression regulation and disease pathology is still not well understood for autoimmune disorders. Despite the limitations candidate miRNA molecules such as miR-146a, upregulation in case of RA has provided a hope to use them for diagnostic purpose [165].

Table 22.7 Tools to Analyze Differential Expression of miRNA

Tool	Algorithm Used	Source	References
DESeq2	Negative binomial distribution	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	[161]
EdgeR	Empirical Bayes estimation	https://bioconductor.org/packages/release/bioc/html/edgeR.html	[162]
miARma-Seq	Analysis pipeline	http://miarmaseq.idoproteins.com/	[163]
CAP-miRSeq	Analysis pipeline	https://bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html	[164]

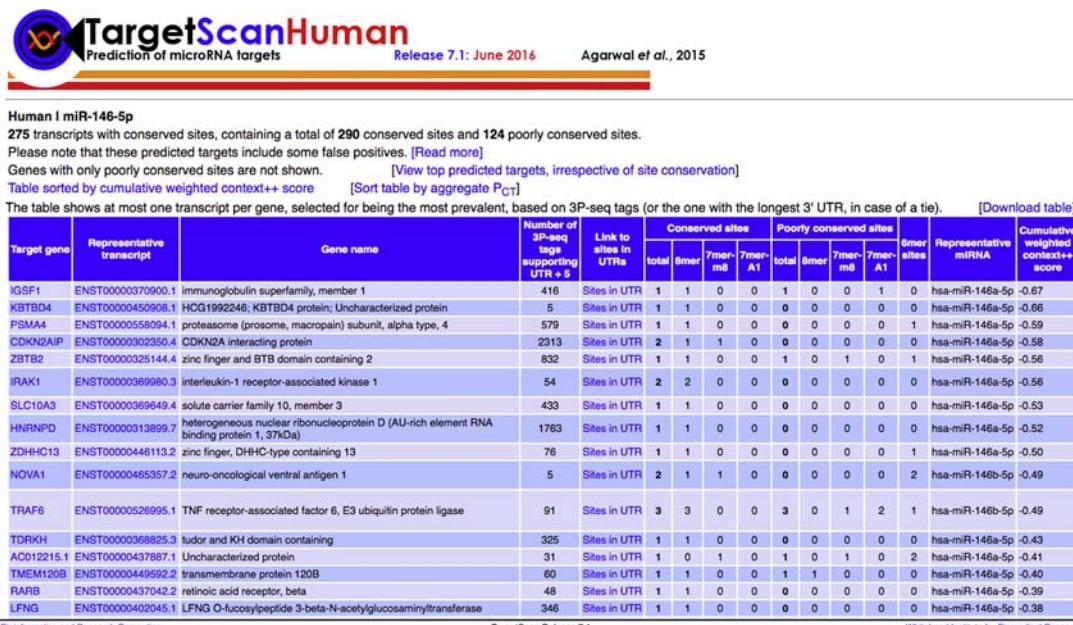
Table 22.8 Tools for miRNA Target Prediction

Tool	Source
TargetScan	http://www.targetscan.org/
Diana Tools	http://diana.imis.athena-innovation.gr/DianaTools/index.php
miRanda	http://www.microrna.org/microrna/getGeneForm.do
PITA	http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html
PicTar	http://pictar.mdc-berlin.de/

To get the functional impact of microRNA or other noncoding RNA, one of the important steps is searching target genes of miRNA molecules having differential expression. In this example, the target genes of miRNA-146, one of the upregulated [166] miRNA molecules in case of RA is searched using TargetScanHuman, one of the well-known miRNA target prediction tools (Fig. 22.6). Several tools used to analyze differential expression and target prediction of miRNA can be found in Tables 22.6 and 22.7, respectively.

EPIGENETIC DATABASES

Deciphering gene regulation with respect to epigenetic mechanisms has been in the prime focus of researchers since last decade, which has resulted in plenty of research articles on molecules and regulatory factors involved in epigenetic regulation [167–169]. Most of these regulatory factors are enzymes involved in histone modification, methylation and demethylation of CpG islands, and differential expression and regulation by miRNA. After experimental data analyses, validation and submission are important part of research, hence there is always a need of data repository systems or databases. Databases such as NCBI, contain huge information and come with associated tools to visualize and use those data as per own convenience. Databases also serve as important go-to points to retrieve any information about any topic. Some of the databases related to epigenomics such as HIstome, MethylomeDB are frequently referred while searching for epigenetic signatures for various diseases.

**FIGURE 22.6**

miRNA-146 target prediction in human using target scan Web server.

HISTOME

It is a freely accessible electronic database with stored information of about five types of histones, eight classes of posttranslational modifications, and various histone-modifying enzymes. Updated version of this database includes complete information of 50 histone proteins along with 150 histone-modifying enzymes. Many of the data fields are linked and cross-referenced with other databases (e.g., UniprotKB/Swiss-Prot, HGNC, OMIM, Unigene, etc.). Complete nucleotide sequences of promoter regions (-700 TSS + 300) for all gene entries can also be found on this database [170]. This database can be easily accessed through www.histome.net, <http://www.iiserpune.ac.in/~coee/histome/index.php>, or <http://www.actrec.gov.in/histome/index.php>.

METHYLOMEDB

This contains genome-wide DNA methylation data of brain tissues from human and mouse (<http://www.neuroepigenomics.org/methylomedb/index.html>). Methylation Mapping Analysis by Paired-end Sequencing (Methyl-MAPS) has been used to generate methylation profiles and has been analyzed by Methyl-Analyzer software. The methylation profiles include complete information of 80% CpG dinucleotides in human and mouse brains. This database includes an integrated genome

browser (modified from UCSC Genome Browser), which helps users to browse difference in DNA methylation profiles of a specific genomic loci, to look for certain changes in pattern of methylation, and to compare methylation patterns between individual samples [171].

METHBASE

MethBase is a central reference methylome database created from public bisulfite sequencing data sets. It contains hundreds of methylomes from well-studied organisms. For each methylome, Methbase provides methylation level at individual sites, regions of allele-specific methylation, hypo- or hypermethylated regions, partially methylated regions, and detailed metadata and summary statistics. These results are generated with the MethPipe software package, a stand-alone, comprehensive pipeline for analyzing bisulfite sequencing data, both WGBS and RRBS [142].

miRWALK2.0

miRWalk2.0 is a freely accessible, comprehensive archive, supplying the biggest available collection of predicted and experimentally verified miRNA–target interactions. It houses possible binding site interaction information (including “central pairing sites”) between genes (encompassing the complete sequence as well as mitochondrial genomes) and miRNAs resulting from the miRWalk algorithm by walking with a heptamer (7 nts) seed of miRNA from positions 1 to 6. These different starting positions are considered because it has recently been identified that miRNAs also regulate the expression of their target genes by annealing from nucleotides 4 to 15 [172]. This database also includes a comparative platform of miRNA-binding sites on mitochondrial genomes [173]. This database is available through <http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/>.

ROADMAP EPIGENOMICS

The NIH Roadmap Epigenomics Mapping Consortium was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. The project has generated high-quality, genome-wide maps of several key histone modifications, chromatin accessibility, DNA methylation, and mRNA expression across hundreds of human cell types and tissues. This Web portal serves as a supplementary data repository system [174]. This Web-interface database can be used by visiting http://egg2.wustl.edu/roadmap/web_portal/.

CONCLUSION

Treatments based on epigenetic modulators are now frequently explored for their effective use in treating and managing autoimmune diseases. Although a lot has been reported regarding epigenetic modifications for autoimmune diseases, there is still an important scope to identify specific epigenetic signature and landmarks for each autoimmune disorder. It is thus imperative to identify a suitable study design that can specifically answer the questions, identify the best tissue types, and the corresponding controls, protocols of collection and processing followed by a comprehensive analysis of the data using the well-established computational tools. The tools should provide ample information regarding

baseline noise, extent of outliers, and true positive detections, should be helpful for the biologist to segregate information based on cell type as well as disease progression. A lot of EWAS has already generated resources, which can be explored effectively.

REFERENCES

- [1] Theofilopoulos AN. The basis of autoimmunity: Part I Mechanisms of aberrant self-recognition. *Immunol Today* 1995;16(2):90–8.
- [2] Jacobson DL, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin Immunol Immunopathol* 1997;84(3):223–43.
- [3] Holoshitz J. The rheumatoid arthritis HLA-DRB1 shared epitope. *Curr Opin Rheumatol* 2010;22(3):293.
- [4] Rioux JD, Abbas AK. Paths to understanding the genetic basis of autoimmune disease. *Nature* 2005; 435(7042):584–9.
- [5] Raghav SK, Gupta B, Agrawal C, Chaturvedi VP, Das HR. Expression of TNF- α and related signaling molecules in the peripheral blood mononuclear cells of rheumatoid arthritis patients. *Mediat Inflamm* 2006; 2006.
- [6] Khanna S, Jaiswal KS, Gupta B. Managing rheumatoid arthritis with dietary interventions. *Front Nutr* 2017;4.
- [7] Nielen MM, van Schaardenburg D, Reesink HW, Van de Stadt RJ, van der Horst-Bruinsma IE, de Koning MH, Habibuw MR, Vandebroucke JP, Dijkmans BA. Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis Rheum* 2004;50(2): 380–6.
- [8] Edwards C, Cooper C. Early environmental factors and rheumatoid arthritis. *Clin Exp Immunol* 2006; 143(1):1–5.
- [9] Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;12(8):529–41.
- [10] Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, Berdasco M, Fraga MF, O'Hanlon TP, Rider LG. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res* 2010;20(2):170–9.
- [11] Gitan RS, Shi H, Chen C-M, Yan PS, Huang TH-M. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res* 2002;12(1):158–64.
- [12] Zhang X, Zhou D, Zhao M, Luo Y, Zhang P, Lu Z, Lu Q. A proof-of-principle demonstration of a novel microarray-based method for quantifying DNA methylation levels. *Mol Biotechnol* 2010;46(3):243–9.
- [13] Richardson B. DNA methylation and autoimmune disease. *Clin Immunol* 2003;109(1):72–9.
- [14] Robyr D, Grunstein M. Genomewide histone acetylation microarrays. *Methods* 2003;31(1):83–9.
- [15] Hecht A, Strahl-Bolsinger S, Grunstein M. Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature* 1996;383(6595):92.
- [16] Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001;409(6819):533–8.
- [17] Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci* 1997;94(6): 2150–5.
- [18] Feldmann M, Brennan FM, Maini RN. Rheumatoid arthritis. *Cell* 1996;85:1277–89.
- [19] Thomson A. The cytokine handbook. 2nd. London: Academic Press; 1994.
- [20] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;316(5830):1497–502.

- [21] Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;129(4):823–37.
- [22] Robertson G, Hirst M, Bainbridge M, Biletsky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;4(8):651–7.
- [23] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;448(7153):553–60.
- [24] Hedrich C. Genetic variation and epigenetic patterns in autoimmunity. *J Genet Syndr Gene Ther* 2011;2:2.
- [25] Greer JM, McCombe PA. The role of epigenetic mechanisms and processes in autoimmune disorders. *Biol Targets Ther* 2012;6:307.
- [26] Sampath D, Liu C, Vasan K, Sulda M, Puduvali VK, Wierda WG, Keating MJ. Histone deacetylases mediate the silencing of miR-15a, miR-16, and miR-29b in chronic lymphocytic leukemia. *Blood* 2012;119(5):1162–72.
- [27] Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;14(3):204–20.
- [28] Baubec T, Schübeler D. Genomic patterns and context specific interpretation of DNA methylation. *Curr Opin Genet Dev* 2014;25:85–92.
- [29] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456(7218):53–9.
- [30] Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinf* 2010;11(1):415.
- [31] Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, Van Helden J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc* 2012;7(8):1551–68.
- [32] Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 2013;45(2):124–30.
- [33] Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 2014;15(2):R31.
- [34] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;457(7231):854–8.
- [35] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10(10):669–80.
- [36] Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res* 2008;18(10):1652–9.
- [37] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graef S, Johnson N, Herrero J, Tomazou EM. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 2008;26(7):779.
- [38] Riebler A, Menigatti M, Song JZ, Statham AL, Stirzaker C, Mahmud N, Mein CA, Clark SJ, Robinson MD. BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach. *Genome Biol* 2014;15(2):R35.
- [39] Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* 2005;6(11):R97.
- [40] Huang W, Loganathanraj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics* 2013;29(23):3097–9.

- [41] Lerdrup M, Johansen JV, Agrawal-Singh S, Hansen K. An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nat Struct Mol Biol* 2016;23(4):349.
- [42] Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC Genom* 2013;14(1):826.
- [43] Wang P, Qin J, Qin Y, Zhu Y, Wang LY, Li MJ, Zhang MQ, Wang J. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Res* 2015;43(W1):W264–9.
- [44] Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *BioMed Res Int* 2012;2012.
- [45] Friedländer MR, Adamidi C, Han T, Lebedeva S, Isenbarger TA, Hirst M, Marra M, Nusbaum C, Lee WL, Jenkin JC. High-resolution profiling and discovery of planarian small RNAs. *Proc Natl Acad Sci* 2009;106(28):11546–51.
- [46] Kang W, Friedländer MR. Computational prediction of miRNA genes from small RNA sequencing data. *Front Bioeng Biotechnol* 2015;3.
- [47] Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013;45(10):1238–43.
- [48] Nolte-'t Hoen EN, Buermans HP, Waasdorp M, Stoorvogel W, Wauben MH, t Hoen PA. Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions. *Nucleic Acids Res* 2012;40(18):9272–85.
- [49] Fehniger TA, Wylie T, Germino E, Leong JW, Magrini VJ, Koul S, Keppel CR, Schneider SE, Koboldt DC, Sullivan RP. Next-generation sequencing identifies the natural killer cell microRNA transcriptome. *Genome Res* 2010;20(11):1590–604.
- [50] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;18(9):1509–17.
- [51] Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genom* 2010;11(1):282.
- [52] Bergman CM, Carlson JW, Celniker SE. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* 2004;21(8):1747–9.
- [53] Gupta B, Hawkins RD. Epigenomics of autoimmune diseases. *Immunol Cell Biol* 2015;93(3):271–6.
- [54] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009;459(7243):108–12.
- [55] Cui K, Zang C, Roh T-Y, Schones DE, Childs RW, Peng W, Zhao K. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* 2009;4(1):80–93.
- [56] Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 2010;6(5):479–91.
- [57] Hawkins RD, Hon GC, Yang C, Antosiewicz-Bourget JE, Lee LK, Ngo Q-M, Klugman S, Ching KA, Edsall LE, Ye Z. Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell Res* 2011;21(10):1393–409.
- [58] Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 2013;153(5):1134–48.
- [59] Bruns A, Blass S, Hausdorf G, Burmester GR, Hiepe F. Nucleosomes are major T and B cell autoantigens in systemic lupus erythematosus. *Arthritis Rheum* 2000;43(10):2307–15.

- [60] Koutouzov S, Jeronimo AL, Campos H, Amoura Z. Nucleosomes in the pathogenesis of systemic lupus erythematosus. *Rheum Dis Clin N Am* 2004;30(3):529–58.
- [61] van Bavel CC, Dieker JW, Tamboer WP, van der Vlag J, Berden JH. Lupus-derived monoclonal autoantibodies against apoptotic chromatin recognize acetylated conformational epitopes. *Mol Immunol* 2010; 48(1):248–56.
- [62] Hu N, Qiu X, Luo Y, Yuan J, Li Y, Lei W, Zhang G, Zhou Y, Su Y, Lu Q. Abnormal histone modification patterns in lupus CD4+ T cells. *J Rheumatol* 2008;35(5):804–10.
- [63] Amoura Z, Koutouzov S, Piette J-C. The role of nucleosomes in lupus. *Curr Opin Rheumatol* 2000;12(5): 369–73.
- [64] Van Bavel CC, Dieker JW, Kroese Y, Tamboer WP, Voll R, Muller S, Berden JH, Van Der Vlag J. Apoptosis-induced histone H3 methylation is targeted by autoantibodies in systemic lupus erythematosus. *Ann Rheum Dis* 2011;70(1):201–7.
- [65] Zhang Z, Song L, Maurer K, Petri MA, Sullivan KE. Global H4 acetylation analysis by ChIP-chip in systemic lupus erythematosus monocytes. *Gene Immun* 2010;11(2):124–33.
- [66] Dai Y, Zhang L, Hu C, Zhang Y. Genome-wide analysis of histone H3 lysine 4 trimethylation by ChIP-chip in peripheral blood mononuclear cells of systemic lupus erythematosus patients. *Clin Exp Rheumatol* 2010; 28(2):158.
- [67] Zhang Z, Song L, Maurer K, Bagashev A, Sullivan K. Monocyte polarization: the relationship of genome-wide changes in H4 acetylation with polarization. *Gene Immun* 2011;12(6):445–56.
- [68] Mastronardi FG, Noor A, Wood DD, Paton T, Moscarello MA. Peptidyl argininedeiminase 2 CpG island in multiple sclerosis white matter is hypomethylated. *J Neurosci Res* 2007;85(9):2006–16.
- [69] Moscarello MA, Mastronardi FG, Wood DD. The role of citrullinated proteins suggests a novel mechanism in the pathogenesis of multiple sclerosis. *Neurochem Res* 2007;32(2):251–6.
- [70] Musse AA, Boggs JM, Harauz G. Deimination of membrane-bound myelin basic protein in multiple sclerosis exposes an immunodominant epitope. *Proc Natl Acad Sci U S A* 2006;103(12):4422–7.
- [71] Tranquill LR, Cao L, Ling NC, Kalbacher H, Martin RM, Whitaker JN. Enhanced T cell responsiveness to citrulline-containing myelin basic protein in multiple sclerosis patients. *Mult Scler J* 2000;6(4):220–5.
- [72] Calabrese R, Zampieri M, Mechelli R, Annibali V, Guastafierro T, Ciccarone F, Coarelli G, Umeton R, Salvetti M, Caiafa P. Methylation-dependent PAD2 upregulation in multiple sclerosis peripheral blood. *Mult Scler J* 2012;18(3):299–304.
- [73] Pedre X, Mastronardi F, Bruck W, López-Rodas G, Kuhlmann T, Casaccia P. Changed histone acetylation patterns in normal-appearing white matter and early multiple sclerosis lesions. *J Neurosci* 2011;31(9): 3435–45.
- [74] Eizirik DL, Colli ML, Ortis F. The role of inflammation in insulitis and β-cell loss in type 1 diabetes. *Nat Rev Endocrinol* 2009;5(4):219–26.
- [75] Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 2009;41(6):703–7.
- [76] Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 2008;40(12):1399–401.
- [77] Kauri LM, Wang G-S, Patrick C, Bareggli M, Hill DJ, Scott FW. Increased islet neogenesis without increased islet mass precedes autoimmune attack in diabetes-prone rats. *Lab Invest* 2007;87(12):1240–51.
- [78] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008; 40(7):897–903.

- [79] Akirav EM, Lebastchi J, Galvan EM, Henegariu O, Akirav M, Ablamunits V, Lizardi PM, Herold KC. Detection of β cell death in diabetes using differentially methylated circulating DNA. *Proc Natl Acad Sci* 2011;108(47):19018–23.
- [80] Pfleger C, Meierhoff G, Kolb H, Schloot NC, Group pS. Association of T-cell reactivity with β -cell function in recent onset type 1 diabetes patients. *J Autoimmun* 2010;34(2):127–35.
- [81] Fu L-h, Ma C-l, Cong B, Li S-j, Chen H-y, Zhang J-g. Hypomethylation of proximal CpG motif of interleukin-10 promoter regulates its expression in human rheumatoid arthritis. *Acta Pharmacol Sin* 2011;32(11):1373–80.
- [82] Kim Y-I, Logan JW, Mason JB, Roubenoff R. DNA hypomethylation in inflammatory arthritis: reversal with methotrexate. *J Lab Clin Med* 1996;128(2):165–72.
- [83] Neidhart M, Rethage J, Kuchen S, Kunzler P, Crowl RM, Billingham ME, Gay RE, Gay S. Retrotransposable L1 elements expressed in rheumatoid arthritis synovial tissue. *Arthritis Rheum* 2000;43:2634–47.
- [84] Kuchen S, Seemayer CA, Rethage J, Rv K, Kuenzler P, Michel BA, Gay RE, Gay S, Neidhart M. The L1 retroelement-related p40 protein induces p38 δ MAP kinase. *Autoimmunity* 2004;37(1):57–65.
- [85] Salas-Pérez F, Codner E, Valencia E, Pizarro C, Carrasco E, Pérez-Bravo F. MicroRNAs miR-21a and miR-93 are down regulated in peripheral blood mononuclear cells (PBMCs) from patients with type 1 diabetes. *Immunobiology* 2013;218(5):733–7.
- [86] Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. *Cell* 2007;128(4):635–8.
- [87] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* 2016;17(8):487.
- [88] Bernstein E, Allis CD. RNA meets chromatin. *Genes Dev* 2005;19(14):1635–55.
- [89] Oelke K, Lu Q, Richardson D, Wu A, Deng C, Hanash S, Richardson B. Overexpression of CD70 and overstimulation of IgG synthesis by lupus T cells and T cells treated with DNA methylation inhibitors. *Arthritis Rheum* 2004;50(6):1850–60.
- [90] Felsenfeld G. A brief history of epigenetics. *Cold Spring Harb Perspect Biol* 2014;6(1):a018200.
- [91] Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* 2016;5(1):3.
- [92] Buckland J. Rheumatoid arthritis: HDAC and HDACi: pathogenetic and mechanistic insights. *Nat Rev Rheumatol* 2011;7(12):682.
- [93] Grabiec AM, Reedquist KA. Histone deacetylases in RA: epigenetics and epiphenomena. *Arthritis Res Ther* 2010;12(5):142.
- [94] Huber LC, Brock M, Hemmatazad H, Giger OT, Moritz F, Trenkmann M, Distler JH, Gay RE, Kolling C, Moch H. Histone deacetylase/acetylase activity in total synovial tissue derived from rheumatoid arthritis and osteoarthritis patients. *Arthritis Rheum* 2007;56(4):1087–93.
- [95] Grabiec AM, Korchynskyi O, Tak PP, Reedquist KA. Histone deacetylase inhibitors suppress rheumatoid arthritis fibroblast-like synoviocyte and macrophage IL-6 production by accelerating mRNA decay. *Ann Rheum Dis* 2012;71(3):424–31.
- [96] Choo Q-Y, Ho PC, Tanaka Y, Lin H-S. Histone deacetylase inhibitors MS-275 and SAHA induced growth arrest and suppressed lipopolysaccharide-stimulated NF- κ B p65 nuclear accumulation in human rheumatoid arthritis synovial fibroblastic E11 cells. *Rheumatology* 2010. keq108.
- [97] Nasu Y, Nishida K, Miyazawa S, Komiyama T, Kadota Y, Abe N, Yoshida A, Hirohata S, Ohtsuka A, Ozaki T. Trichostatin A, a histone deacetylase inhibitor, suppresses synovial inflammation and subsequent cartilage destruction in a collagen antibody-induced arthritis mouse model. *Osteoarthritis Cartilage* 2008;16(6):723–32.
- [98] Paul DS, Teschendorff AE, Dang MA, Lowe R, Hawa MI, Ecker S, Beyan H, Cunningham S, Fouts AR, Ramelius A. Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat Commun* 2016;7:13555.

- [99] Susser E, Neugebauer R, Hoek HW, Brown AS, Lin S, Labovitz D, Gorman JM. Schizophrenia after prenatal famine: further evidence. *Arch Gen Psychiatr* 1996;53(1):25–31.
- [100] Stanczyk J, Pedrioli DML, Brentano F, Sanchez-Pernaute O, Kolling C, Gay RE, Detmar M, Gay S, Kyburz D. Altered expression of MicroRNA in synovial fibroblasts and synovial tissue in rheumatoid arthritis. *Arthritis Rheum* 2008;58(4):1001–9.
- [101] Li J, Wan Y, Guo Q, Zou L, Zhang J, Fang Y, Zhang J, Zhang J, Fu X, Liu H. Altered microRNA expression profile with miR-146a upregulation in CD4+ T cells from patients with rheumatoid arthritis. *Arthritis Res Ther* 2010;12(3):R81.
- [102] Nakasa T, Miyaki S, Okubo A, Hashimoto M, Nishida K, Ochi M, Asahara H. Expression of microRNA-146 in rheumatoid arthritis synovial tissue. *Arthritis Rheum* 2008;58(5):1284–92.
- [103] Niimoto T, Nakasa T, Ishikawa M, Okuhara A, Izumi B, Deie M, Suzuki O, Adachi N, Ochi M. MicroRNA-146a expresses in interleukin-17 producing T cells in rheumatoid arthritis patients. *BMC Musculoskel Disord* 2010;11(1):209.
- [104] Stanczyk J, Ospelt C, Karouzakis E, Filer A, Raza K, Kolling C, Gay R, Buckley CD, Tak PP, Gay S. Altered expression of microRNA-203 in rheumatoid arthritis synovial fibroblasts and its role in fibroblast activation. *Arthritis Rheum* 2011;63(2):373–81.
- [105] Nakamachi Y, Kawano S, Takenokuchi M, Nishimura K, Sakai Y, Chin T, Saura R, Kurosaka M, Kumagai S. MicroRNA-124a is a key regulator of proliferation and monocyte chemoattractant protein 1 secretion in fibroblast-like synoviocytes from patients with rheumatoid arthritis. *Arthritis Rheum* 2009;60(5):1294–304.
- [106] Kawano S, Nakamachi Y. miR-124a as a key regulator of proliferation and MCP-1 secretion in synoviocytes from patients with rheumatoid arthritis. *Ann Rheum Dis* 2011;70(Suppl 1):i88–91.
- [107] Chen P-Y, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinf* 2010;11(1):203.
- [108] Keravnou A, Ioannides M, Tsangaras K, Loizides C, Hadjidianni MD, Papageorgiou EA, Kyriakou S, Antoniou P, Mina P, Achilleos A. Whole-genome fetal and maternal DNA methylation analysis using MeDIP-NGS for the identification of differentially methylated regions. *Genetics Res* 2016;98.
- [109] Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 2008;24(15):1729–30.
- [110] Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* 2012;7(9):1728.
- [111] Sekigawa I, Okada M, Ogasawara H, Kaneko H, Hishikawa T, Hashimoto H. DNA methylation in systemic lupus erythematosus. *Lupus* 2003;12(2):79–85.
- [112] Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* 2013;30(2):284–6.
- [113] Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, Daunay A, Busato F, Mein CA, Manfras B. Identification of type 1 diabetes–associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genetics* 2011;7(9):e1002300.
- [114] Consortium EP. A user’s guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology* 2011;9(4):e1001046.
- [115] Heinig M, Colomé-Tatché M, Taudt A, Rintisch C, Schafer S, Pravenec M, Hubner N, Vingron M, Johannes F. histoneHMM: differential analysis of histone modifications with broad genomic footprints. *BMC Bioinf* 2015;16(1):60.
- [116] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* 2011;29(1):24.

- [117] Van Wynsberghe PM, Chan S-P, Slack FJ, Pasquinelli AE. Analysis of microRNA expression and function. In: *Methods in cell biology*, vol. 106. Elsevier; 2011. p. 219–52.
- [118] Lu Q, Kaplan M, Ray D, Ray D, Zacharek S, Gutsch D, Richardson B. Demethylation of ITGAL (CD11a) regulatory sequences in systemic lupus erythematosus. *Arthritis Rheum* 2002;46(5):1282–91.
- [119] Lu Q, Wu A, Tesmer L, Ray D, Yousif N, Richardson B. Demethylation of CD40LG on the inactive X in T cells from women with lupus. *J Immunol* 2007;179(9):6352–8.
- [120] Kaplan MJ, Lu Q, Wu A, Attwood J, Richardson B. Demethylation of promoter regulatory elements contributes to perforin overexpression in CD4+ lupus T cells. *J Immunol* 2004;172(6):3652–61.
- [121] Lei W, Luo Y, Lei W, Luo Y, Yan K, Zhao S, Li Y, Qiu X, Zhou Y, Long H. Abnormal DNA methylation in CD4+ T cells from patients with systemic lupus erythematosus, systemic sclerosis, and dermatomyositis. *Scand J Rheumatol* 2009;38(5):369–74.
- [122] Lu Q, Wu A, Ray D, Deng C, Attwood J, Hanash S, Pipkin M, Lichtenheld M, Richardson B. DNA methylation and chromatin structure regulate T cell perforin gene expression. *J Immunol* 2003;170(10):5124–32.
- [123] Lu Q, Ray D, Gutsch D, Richardson B. Effect of DNA methylation and chromatin structure onITGAL expression. *Blood* 2002;99(12):4503–8.
- [124] Garaud S, Le Dantec C, Jousse-Joulin S, Hanrotel-Saliou C, Saraux A, Mageed RA, Youinou P, Renaudineau Y. IL-6 modulates CD5 expression in B cells from patients with lupus by regulating DNA methylation. *J Immunol* 2009;182(9):5623–32.
- [125] Schett G, Smolen J, Zimmermann C, Hiesberger H, Hoefler E, Fournel S, Muller S, Rubin R, Steiner G. The autoimmune response to chromatin antigens in systemic lupus erythematosus: autoantibodies against histone H1 are a highly specific marker for SLE associated with increased disease activity. *Lupus* 2002;11(11):704–15.
- [126] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;28(6):882–3.
- [127] Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, Fasching PA, Widschwendter M. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* 2016;7:10478.
- [128] Miao F, Smith DD, Zhang L, Min A, Feng W, Natarajan R. Lymphocytes from patients with type 1 diabetes display a distinct profile of chromatin histone H3 lysine 9 dimethylation. *Diabetes* 2008;57(12):3189–98.
- [129] Sebastiani G, Grieco FA, Spagnuolo I, Galleri L, Cataldo D, Dotta F. Increased expression of microRNA miR-326 in type 1 diabetic patients with ongoing islet autoimmunity. *Diabetes Metab Res Rev* 2011;27(8):862–6.
- [130] Hezova R, Slaby O, Faltejskova P, Mikulkova Z, Buresova I, Raja KM, Hodek J, Ovesna J, Michalek J. microRNA-342, microRNA-191 and microRNA-510 are differentially expressed in T regulatory cells of type 1 diabetic patients. *Cell Immunol* 2010;260(2):70–4.
- [131] Bogdanović O, Veenstra GJC. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* 2009;118(5):549–65.
- [132] Rauch T, Wang Z, Zhang X, Zhong X, Wu X, Lau SK, Kernstine KH, Riggs AD, Pfeifer GP. Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. *Proc Natl Acad Sci* 2007;104(13):5527–32.
- [133] Weng Y-I, Huang TH-M, Yan PS. Methylated DNA immunoprecipitation and microarray-based analysis: detection of DNA methylation in breast cancer cell lines. In: *Molecular endocrinology*. Springer; 2009. p. 165–76.
- [134] Siegmund KD. Statistical approaches for the analysis of DNA methylation microarray data. *Hum Genet* 2011;129(6):585–95.

- [135] Li Y, Tollefson TO. DNA methylation detection: bisulfite genomic sequencing analysis. In: Epigenetics protocols. Springer; 2011. p. 11–21.
- [136] Guo S, Zhu Q, Jiang T, Wang R, Shen Y, Zhu X, Wang Y, Bai F, Ding Q, Zhou X. Genome-wide DNA methylation patterns in CD4⁺ T cells from Chinese Han patients with rheumatoid arthritis. *Mod Rheumatol* 2017;27(3):441–7.
- [137] Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, Chatham WW, Kimberly RP. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4⁺ T-cell populations. *PLoS Genetics* 2013;9(8):e1003678.
- [138] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2012;29(2):189–96.
- [139] Chen Y-a, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;8(2):203–9.
- [140] Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008;24(13):1547–8.
- [141] Hansen KD, Langmead B, Irizarry RA. BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;13(10):R83.
- [142] Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* 2013;8(12):e81148.
- [143] Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* 2014;15(2):R38.
- [144] Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings Funct Genomics Proteomics* 2009;8(3):174–83.
- [145] Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;11(1):31.
- [146] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;27(11):1571–2.
- [147] Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010. 11.17. 11-11.17. 14.
- [148] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf* 2009;10(1):232.
- [149] Grunau C, Schattevoy R, Mache N, Rosenthal A. MethTools—a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res* 2000;28(5):1053–8.
- [150] Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. *Nucleic Acids Res* 2008;36(suppl_2):W170–5.
- [151] Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* 2012;13(7):R61.
- [152] Aryee MJ, Jaffe AE, Corradia-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;30(10):1363–9.
- [153] Martin TC, Yet I, Tsai P-C, Bell JT. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinf* 2015;16(1):131.
- [154] Lun AT, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res* 2015;44(5):e45.
- [155] Niazi U, Geyer KK, Vickers MJ, Hoffmann KF, Swain MT. DISMISS: detection of stranded methylation in MeDIP-Seq data. *BMC Bioinf* 2016;17(1):295.

- [156] Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinf* 2010;11(1):237.
- [157] Huang J, Renault V, Sengenes J, Touleimat N, Michel S, Lathrop M, Tost J. MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. *Bioinformatics* 2011;28(4):587–8.
- [158] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. Model-based analysis of chip-seq (MACS). *Genome Biol* 2008;9(9):R137.
- [159] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 2010;38(4):576–89.
- [160] Tam S, Tsao M-S, McPherson JD. Optimization of miRNA-seq data preprocessing. *Briefings Bioinf* 2015; 16(6):950–63.
- [161] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
- [162] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139–40.
- [163] Andrés-León E, Núñez-Torres R, Rojas AM. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 2016;6:25749.
- [164] Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, Kocher J-P. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genom* 2014;15(1):423.
- [165] Ceribelli A, Nahid MA, Satoh M, Chan EK. MicroRNAs in rheumatoid arthritis. *FEBS Lett* 2011;585(23): 3667–74.
- [166] Chan EK, Ceribelli A, Satoh M. MicroRNA-146a in autoimmunity and innate immune responses. *Ann Rheum Dis* 2012. annrheumdis-2012–202203.
- [167] Van Attikum H, Gasser SM. Crosstalk between histone modifications during the DNA damage response. *Trends Cell Biol* 2009;19(5):207–17.
- [168] Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. *Science* 2010;330(6004):612–6.
- [169] Chow J, Heard E. X inactivation and the complexities of silencing a sex chromosome. *Curr Opin Cell Biol* 2009;21(3):359–66.
- [170] Khare SP, Habib F, Sharma R, Gadewal N, Gupta S, Galande S. HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res* 2011;40(D1):D337–42.
- [171] Xin Y, Chanrion B, O'donnell AH, Milekic M, Costa R, Ge Y, Haghghi FG. MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res* 2011;40(D1):D1245–9.
- [172] Shin C, Nam J-W, Farh KK-H, Chiang HR, Shkumatava A, Bartel DP. Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell* 2010;38(6):789–802.
- [173] Dweep H, Gretz N. miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. *Nat Methods* 2015;12(8):697.
- [174] Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, Madden PA, Smirnov I, Costello JF, Wang T. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat Biotechnol* 2015;33(4):345.

This page intentionally left blank

COMPUTATIONAL EPIGENETICS IN LUNG CANCER

23

S. Babichev^{1,2}, V. Lytvynenko², M. Korobchynskyi³, I. Sokur⁴

¹*Jan Evangelista Purkyně University in Usti nad Labem, Usti nad Labem, Czech Republic;* ²*Kherson National Technical University, Kherson, Ukraine;* ³*Military-Diplomatic Academy named Eugene Bereznyak, Kyiv, Ukraine;*
⁴*Kherson Regional Oncology Dispancer, Kherson, Ukraine*

INTRODUCTION

Reconstruction of gene regulatory network based on gene expression profiles is one of the current problems of modern bioinformatics. Gene regulatory network is a set of genes, which interact with each other to control the specific cell functions. Qualitatively reconstructed gene regulatory network allows us to increase the quality of epigenetics investigation in lung cancer. Gene expression profiles obtained from DNA microarray experiments or RNA sequences technology serve as the basis for reconstructing gene regulatory networks [1]. High dimension of feature space is one of the distinctive peculiarities of the studied profiles. About tens of thousands of genes are contained in the studied data. The reconstruction of gene regulatory network based on the whole data set of gene expression profiles is very complicated task because of the following reasons: it requests large computer resources; it needs a lot of time expenses to process the information; and complexity of the obtained network complicates the interpretation of results. Therefore, it is necessary firstly to reduce uninformativity of the gene expression profiles in terms of statistical and entropy criteria and to divide gene expression profiles into subsets, each of which will include a group of genes that performs similar functions in the studied biological object. Biclustering technology is relevant to solve this problem nowadays. Implementation of this technology allows objects and features to be grouped according to their mutual correlations. So, in papers [2,3] the authors provided a review of a large quantity of biclustering approaches existing in literature with the analysis of their advantages and disadvantages. In Ref. [4] the authors proposed and implemented a convex biclustering method using gene expression profiles of lung cancer patients. The authors showed the efficiency of the proposed method during the simulation process. However, it should be noted that one of the significant problems of this technology about qualitative implementation is the selection of biclustering level during objects and features grouping. The qualitative validation of the obtained model is another task, which has no solution nowadays. A high dimension of feature space promotes to large quantity of the obtained biclusters. The limitation of their quantity by removing small biclusters leads to the loss of some useful information. To solve this problem, we propose cluster–bicluster technology, the implementation of which involves two stages:

clustering of gene expression profiles at the first stage and biclustering of the obtained clusters at the second stage of the performed experiment.

There are a lot of clustering algorithms nowadays. Each of them has its advantages and disadvantages and is focused on a specific type of data [5–8]. One of the essential disadvantages of the existing clustering algorithms is reproducibility error, in other words, successful clustering results obtained on one data set do not repeat while using another similar data set. This error reduction can be achieved by careful verification of the obtained model using “fresh information,” which has not been used during the simulation process. A higher degree of coincidence between clustering results on similar data corresponds to a higher degree of the obtained model objectivity. This idea is the basis of objective clustering inductive technology, the main conception of which was presented in Ref. [9]. Practical implementation of the objective clustering inductive technology is possible based on various clustering algorithms. The choice of clustering algorithm is determined by the structure and character of the studied data. The practical implementation of this technology based on agglomerative hierarchical, self-organizing SOTA (self-organizing tree algorithm) [10,11], and DBSCAN (density-based spatial clustering of application with noise) [12] clustering algorithms were presented in Ref. [13–16]. One of the key conditions of a successful implementation of this technology is about the careful determination of internal, external, and complex balance clustering quality criteria, which should be taken into account both the character of the objects of grouping within clusters and character of the cluster distribution in the feature space.

This chapter presents the conceptual basis of the objective clustering inductive technology based on DBSCAN and SOTA clustering algorithms and its implementation within the framework of hybrid model of cluster–biclusler analysis. The lung cancer patients’ gene expression profiles [17] were used during simulation process. This data set includes the gene expression profiles of 96 patients. Ten of them were healthy and 86 patients were divided by the degree of the disease severity into three groups. Size of the initial data matrix was (96×7129) . To our knowledge, the implementation of this technology for gene expression profiles grouping allows us to gather more useful information for the following gene regulatory network reconstruction. Hence, this gene expression profiles can promote better understanding on the epigenetics in lung cancer.

CONCEPTUAL BASIS OF THE OBJECTIVE CLUSTERING INDUCTIVE TECHNOLOGY

Initial data set of gene expression profiles is presented as a matrix: $A = \{x_{ij}\}$, $i = 1, \dots, n; j = 1, \dots, m$, where n is the quantity of studied objects or conditions of the performed experiment, m is the quantity of studied genes. The aim of clustering process is a partition of genes expression profiles into nonempty subsets of pairwise nonintersecting clusters in accordance with clustering quality criteria taking into account the properties of the studied profiles:

$$K = \{K_s\}, s = 1, \dots, k; \bigcup_{s=1}^k K_s = A; \bigcap_{s=1}^k K_s = \emptyset$$

where k is the quantity of clusters. Objective clustering technology is based on the inductive methods of complex systems analysis, which involves sequential enumeration of clustering within a given range of the number of clusters change to select from them the best variants [9]. Let W – is a set of all

admissible clustering for given data set A . Clustering is the best (an optimal) in terms of clustering quality criteria $QC(K)$ if the following condition is performed:

$$K_{opt} = \arg \min_{K \subseteq W} QC(K) \text{ or } K_{opt} = \arg \max_{K \subseteq W} QC(K)$$

where $QC(K)$ – is the quality criterion for K clustering. Clustering $K_{opt} \subseteq W$ is the objective, if difference of objects and clusters distribution in different clustering for equal power subsets A and B (including the same quantity of pairwise similar objects) is minimal:

$$QC(K_{obj}) = \arg \min_{K \subseteq W} (QC(K)^A - QC(K)^B)$$

Architecture of the objective clustering inductive technology is presented in Fig. 23.1.

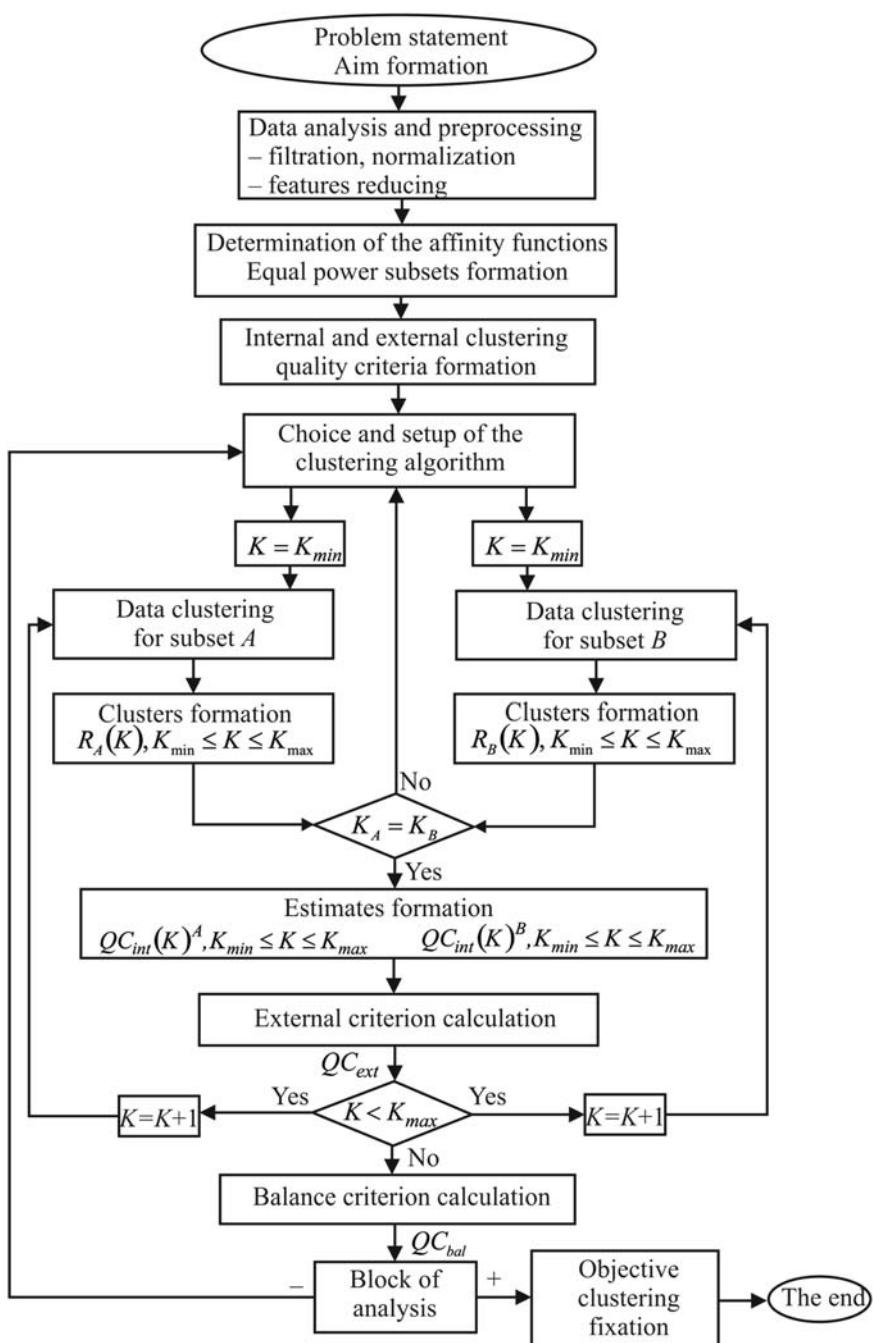
The technology implementation involves the following steps:

1. The studied data affinity function determination. Division of the initial data set into two equal power subsets A and B using chosen affinity function. The equal power subsets include the same quantity of the pairwise similar objects.
2. The internal, external, and complex balance clustering quality criteria formation.
3. Choice of clustering algorithm and setup of its initial parameters. These parameters are changed during the algorithm operation to obtain different variants of the grouped studied data.
4. Data clustering on the subsets A and B concurrently and clusters formation within the range of admissible clustering ($K_{min} \leq K \leq K_{max}$). If the clusters quantity in various clustering differs, it is necessary to change the setup of the algorithm or to use another admissible clustering algorithm and to repeat step 3 of this procedure.
5. Calculation of the internal QC_{int} and external QC_{ext} clustering quality criteria for current clustering on the equal power subsets A and B .
6. Calculation and analysis of the complex balance clustering quality criterion values. Fixation of the objective clustering in accordance with global maximum values of the complex balance clustering quality criterion. If the maximum value of the complex balance criterion is less than admissible (sign “–”), it is necessary to choose another clustering algorithm.

The implementation of the proposed technology allows us to increase the objectivity of determination of the used clustering algorithms optimal parameters at the early stage of gene expression profiles grouping. In the case of computational epigenetics in lung cancer based on the use of gene regulatory network, qualitative and objective grouping of genes before biclustering process allows us to save more useful information for the following interpretation of the character of corresponding genes interaction for simulation of the obtained gene network.

AFFINITY METRIC AND CLUSTERING QUALITY CRITERIA TO ESTIMATE THE PROXIMITY OF GENE EXPRESSION PROFILES

It is obvious that qualitative clustering corresponds to a higher division ability of different clusters and a higher density of objects concentration inside clusters. Thus, it is important to determine the proximity metric of gene expression profiles. In Ref. [18] the authors presented the results of their research on the comparison of three well-known dissimilarity metrics to estimate the proximity level of

**FIGURE 23.1**

Architecture of the objective clustering inductive technology.

numeric vectors: Manhattan, Euclidean, and correlation distances. The evaluation of the effectiveness of the metrics under investigation was performed using the model data representing the lung cancer patients' gene expression profiles in two different clusters. The centers of the corresponding clusters are calculated by the formula:

$$C_S = \frac{1}{N_S} \sum_{i=1}^{N_S} x_i^S$$

where N_S is the quantity of the gene expression profiles in cluster S , x_i^S is i -th gene expression profile in cluster S . The simulation process included the following steps:

- calculation of an average distance d_{int} from gene expression profiles to centres of clusters, where these profiles are

$$d_{int}(X^{S,P}, C_{S,P}) = \frac{1}{N} \left(\sum_{i=1}^{N_S} d(x_i^S, C_S) + \sum_{j=1}^{N_P} d(x_j^P, C_P) \right)$$

- calculation of an average distance d_{ext} from profiles to centers of neighboring clusters:

$$d_{ext}(X^{S,P}, C_{S,P}) = \frac{1}{N} \left(\sum_{i=1}^{N_S} d(x_i^S, C_P) + \sum_{j=1}^{N_P} d(x_j^P, C_S) \right)$$

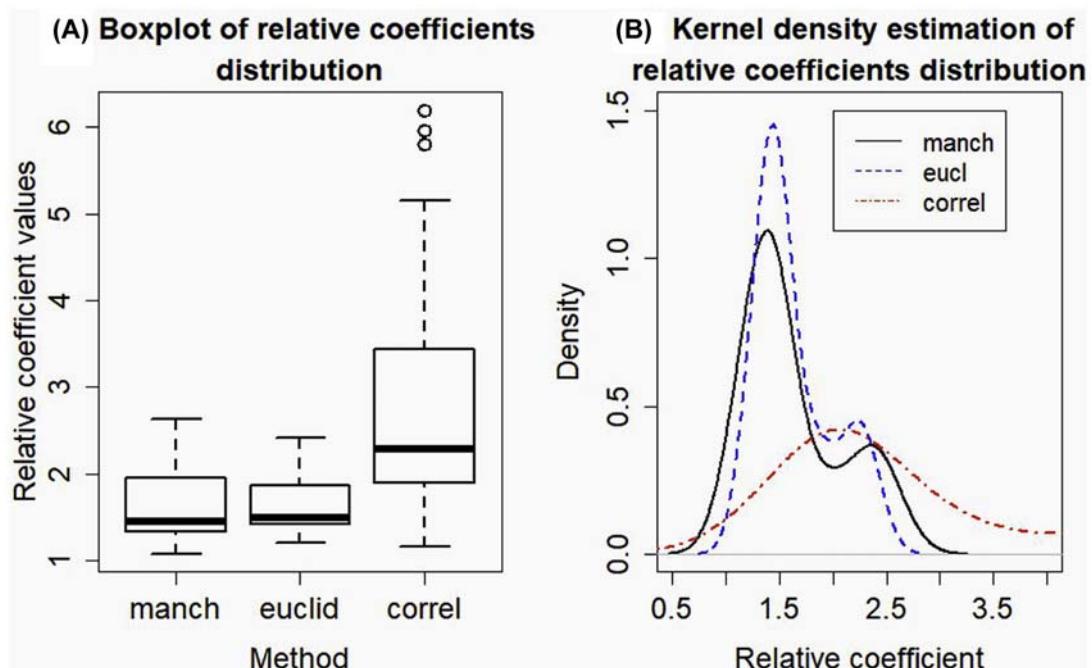
- calculation of relative distance:

$$d_{rel}(X^{S,P}, C_{S,P}) = \frac{d_{ext}(X^{S,P}, C_{S,P})}{d_{int}(X^{S,P}, C_{S,P})},$$

where $X^{S,P}$ and $C_{S,P}$ are the sets of the gene expression profiles in clusters S and P and the centers of these clusters, respectively, x_i^S and x_i^P are i -th gene expression profiles in clusters S and P , N_S and N_P are the number of gene expression profiles in clusters S and P , N is the total number of gene expression profiles.

It is obvious that the higher value of the relative distance corresponds to the higher separating ability of the used affinity metric. To estimate the effectiveness of the metrics, we used the lung cancer patients' data of Array Express Database [17], which includes the gene expression profiles of 96 patients, 10 of them were healthy (these patients do not have lung cancer disease) and 86 patients were divided by the severity of the disease into three groups (well, moderate, and poor). Each of the gene expression profiles included 7129 gene expressions. The class of healthy patients (10 profiles) and the class of patients with poor state of health (21 profiles) were used during the simulation process. The results of the relative distance values distribution while using Manhattan, Euclidean, and correlation distances are shown in Fig. 23.2.

The analysis of Fig. 23.2 allows us to conclude that in the case of gene expression profiles the correlation metric has higher separating ability in comparison with Euclidean and Manhattan metrics

**FIGURE 23.2**

Visualization of the relative distance values distribution: (A) boxplot; (B) kernel density estimation.

because the values of the relative criterion, which were calculated based on the correlation distance, are higher in comparison with the use of Euclidean and Manhattan distances.

It is obvious that a qualitative clustering corresponds to a higher division ability of different clusters and a higher density of objects concentration inside the clusters. Thus, the internal clustering quality criteria should be complex and taking into account both the character of the objects distribution within different clusters and character of the clusters distribution in features space. The first component of the complex internal criterion is calculated as an average distance from the objects to the cluster centers, where these objects are

$$QCW = \frac{1}{N} \sum_{S=1}^K \sum_{i=1}^{N_S} d(x_i^S, C_S)$$

The second component of this criterion is calculated as an average distance between the centers of the clusters. This component takes into account the singularity of the clusters distribution in the feature space:

$$QCB = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K d(C_i, C_j)$$

where K is the quantity of clusters; N is the total quantity of objects; N_S is the quantity of objects in cluster S ; x_i^S is i -th vector in cluster S ; C_i , C_j , and C_S are the centers of clusters i , j , and S , respectively; $d(\cdot)$ is the metric used to estimate the proximity level of the studied vectors. Various combinations of these components allow us to obtain the clustering quality criteria for the studied data. During the simulation process, the following internal criteria to estimate the quality of the data grouping were used:

- Calinski-Harabasz criterion (CH) [19]:

$$QC_{CH} = \frac{QCB \cdot (N - K)}{QCW \cdot (K - 1)} \rightarrow \max,$$

where N is the number of gene expression profiles, K is the number of clusters.

- WB index (WB) [20]:

$$QC_{WB} = \frac{K \cdot QCW}{QCB} \rightarrow \min$$

- Hartigan index (H) [21]:

$$QC_H = \left| \log_2 \frac{QCB}{QCW} \right| \rightarrow \min$$

- C-index (C) [22]:

$$QC_C = \frac{QCW - QCW_{\min}}{QCW_{\max} - QCW_{\min}} \rightarrow \min,$$

where QCW_{\min} and QCW_{\max} are the average of the minimum and maximum distances between gene expression profiles inside the clusters, respectively.

- Ball and Hall index (BH) [23]:

$$QC_{BH} = \frac{QCW}{K} \rightarrow \min$$

- Xie-Beni index (XB) [24]:

$$QC_{XB} = \frac{1}{N} \frac{QCW}{\min \delta(C_S, C_P)} \rightarrow \min,$$

where

$$\delta(C_S, C_P) = \min_{i \in S, j \in P} d(x_i^S, x_j^P)$$

and C_S , C_P are the centers of clusters S and P , respectively; x_i^S , x_j^P are the i -th and j -th gene expression profiles in clusters S and P .

The external clustering quality criterion is calculated as the normalized difference of the internal clustering quality criteria for the equal power subsets A and B :

$$QC_{ext}(A, B) = \frac{|QC_{int}(A) - QC_{int}(B)|}{QC_{int}(A) + QC_{int}(B)}$$

It is obvious that the objective clustering corresponds to extremum values of both the internal and external clustering quality criteria. However, it is possible that the extrema of these criteria correspond to different clustering. Thus, it is necessary to determine a complex balance clustering quality criterion, which takes into account both the character of the objects and clusters distribution in various clustering and difference between clustering results, which are obtained based on the two equal power subsets. Harrington desirability function [25] was used to calculate the complex balance clustering quality criterion. Implementation of this function involves transformation of the spacing the internal and external criteria into reaction scale Y , the values of which are changed linearly within the range from -2 to 5 . Then, the private desirabilities of the appropriate criteria are calculated by the formula:

$$d = \exp(-\exp(-Y))$$

The chart of Harrington desirability function versus the reaction index Y and criteria values is shown in Fig. 23.3.

Transformation of the criteria spacing into reaction scales are performed by linear equation:

$$Y = a - b \cdot QC$$

The parameters a and b are determined empirically. The general Harrington desirability index value is calculated as a geometric average of private desirabilities index:

$$D = \sqrt[n]{\prod_{i=1}^n d_i}$$

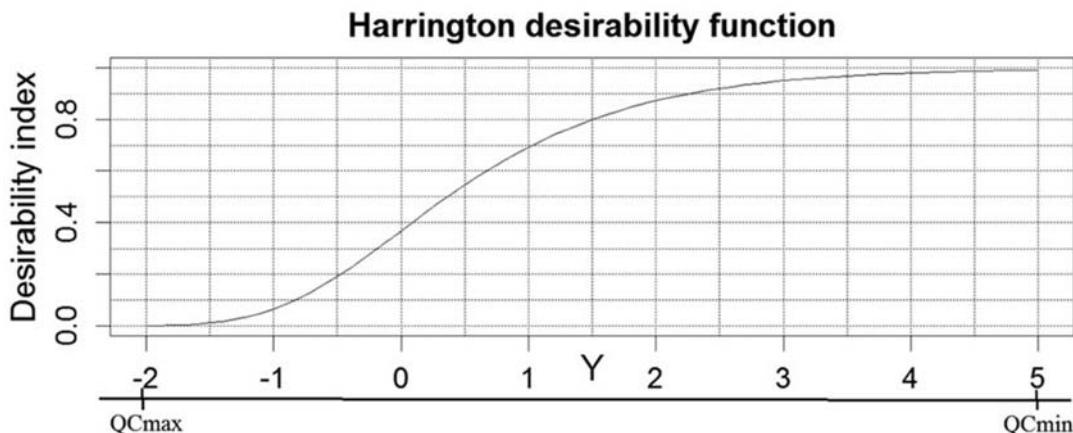


FIGURE 23.3

Harrington desirability function.

In the case of the objective clustering inductive technology, the general Harrington desirability index was used as a complex balance criterion. It is obvious that the largest value of the complex balance criterion corresponds to the best parameters of the clustering algorithm operation.

SIMULATION OF THE OBJECTIVE CLUSTERING PROCESS USING LUNG CANCER PATIENTS' GENE EXPRESSION PROFILES

The lung cancer patients' gene expression profiles were used to estimate the effectiveness of the internal, external, and complex balance clustering quality criteria within the framework of the objective clustering inductive technology [17]. Firstly, the data were divided into two equal power subsets. These subsets contain the same quantity of pairwise similar gene expression profiles. Then, each of these subsets was sequentially divided into clusters from $K_{min} = 2$ to $K_{max} = 5$. In the case of the use of a two-cluster structure, the first cluster were the gene expression profiles of healthy patients (NORM) and the gene expression profiles of the patients with a good state of health (WELL), the second cluster included the gene expression profiles of the patients with poor (POOR) and moderate (MODERATE) health status. In the case of a three-cluster structure, the first cluster contains data of the healthy patients, the second cluster contains the data of the patients with good state of health, and the third cluster includes the gene expression profiles of the patients with poor and moderate health status. In the case of a four-cluster structure the first cluster contains the data of healthy patients, the second cluster contains the data of patients with a good state of health, the third cluster includes the gene expression profiles of the patients with a poor state of health, and the fourth cluster contains the gene expression of the patients with a moderate health status. To obtain a five-cluster structure, the gene expression profiles of the patients with a moderate state of health were divided into two random groups. The best clustering corresponds to the four-cluster structure because all the gene expression profiles in this case are distributed into a corresponding cluster by patients' health status. The correlation distance was used to estimate the proximity level of the appropriate gene expression profiles. Fig. 23.4 shows the charts of the internal clustering quality criteria for equal power of subsets A and B versus clusters quantity.

The analysis of the obtained charts allows us to conclude that Xie-Beni (Fig. 23.4D), Ball and Hall (Fig. 23.4E), and C (Fig. 23.4F) indexes are not effective to estimate gene expression profiles clustering objectivity because the changes of character for these criteria do not allow us to determine the optimal clustering. The values of these criteria do not have extrema, which correspond to the optimal four-cluster structure. Hartigan index (H) shows better results in comparison with Xie-Beni (XB), Ball and Hall (BH) and C indexes (Fig. 23.4C). The value of this criterion for subset A has a minimum for the optimal four-cluster structure, but it is problematic to distinguish the differences between four- and five-cluster structures for subset B. Calinski-Harabasz and WB index criteria show the best results for being selected as an optimal clustering. The values of these criteria have local extrema corresponding to the optimal four-cluster structure. Fig. 23.5 shows the charts of the external clustering quality criteria versus the clusters quantity, which were calculated based on Calinski-Harabasz criteria, Hartigan and WB indexes.

Analysis of the obtained charts has shown the external criterion, which is calculated based on Hartigan index. Hartigan index does not allow us to allocate the optimal four-cluster structure in the

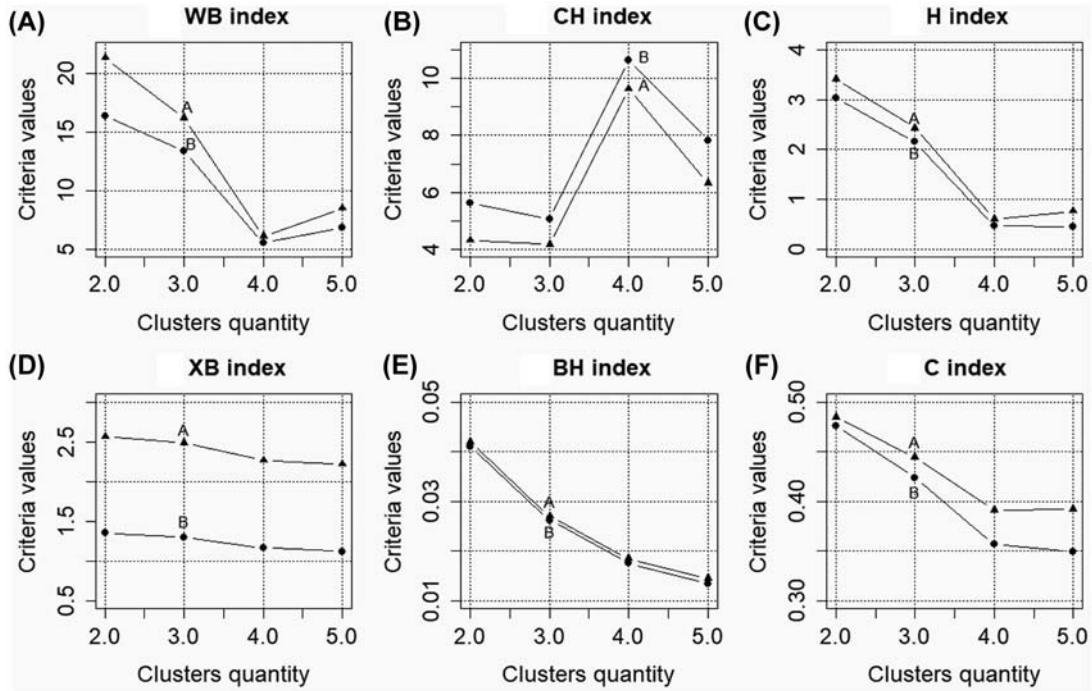


FIGURE 23.4

Charts of the internal clustering quality criteria versus the clusters quantity: (A) WB index; (B) Calinski-Harabasz criterion; (C) Hartigan index; (D) Xie-Beni index; (E) Ball and Hall index; (F) C index.

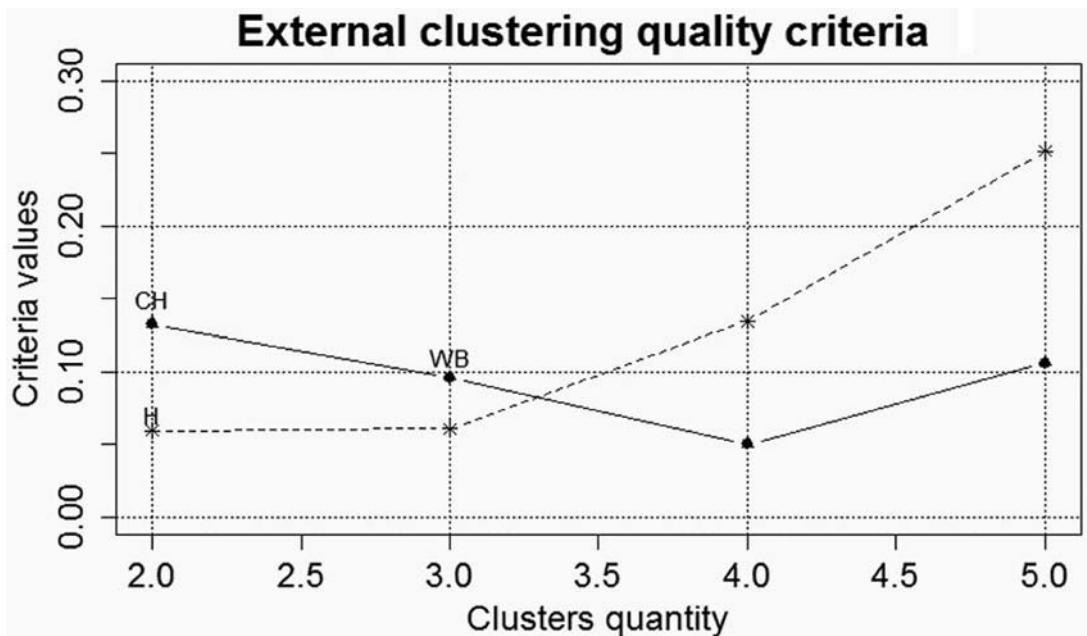


FIGURE 23.5

Charts of the external clustering quality criteria versus the clusters quantity.

case of gene expression profiles analysis. The value of this criterion monotonically increases with the increasing cluster quantity. The external criteria, which are calculated based on *WB* index, and Calinski-Harabasz criterion are effective to solve this task. These criteria showed similar results and with the minimum values corresponded to the optimal four-cluster structure.

To increase the sensitivity of the internal clustering quality criterion and corresponding external criterion, we propose a complex internal criterion, which is calculated as a multiplicative combination of *WB* index and Calinski-Harabasz criteria:

$$QC_{CX} = \frac{QC_{WB}}{QC_{CH}} = \frac{K(K-1)QCW^2}{(N-K)QCB^2} \rightarrow \min$$

Fig. 23.6 shows the charts of the complex internal, external, and balance clustering quality criteria versus the clusters quantity. The complex balance criterion is calculated with the use of the complex internal and external criteria.

The analysis of the obtained charts shows a high efficiency of the proposed criteria to determine an optimal clustering in the case of gene expression profiles grouping. The value of Harrington desirability index, which is used as a complex balance clustering quality criterion and contains both the internal and external criteria, reaches its maximum value in the case of the optimal four-cluster structure. In this case, the complex balance criterion takes into account both the character of objects and cluster grouping in various clustering and difference between the clustering results, obtained on equal power data subsets.

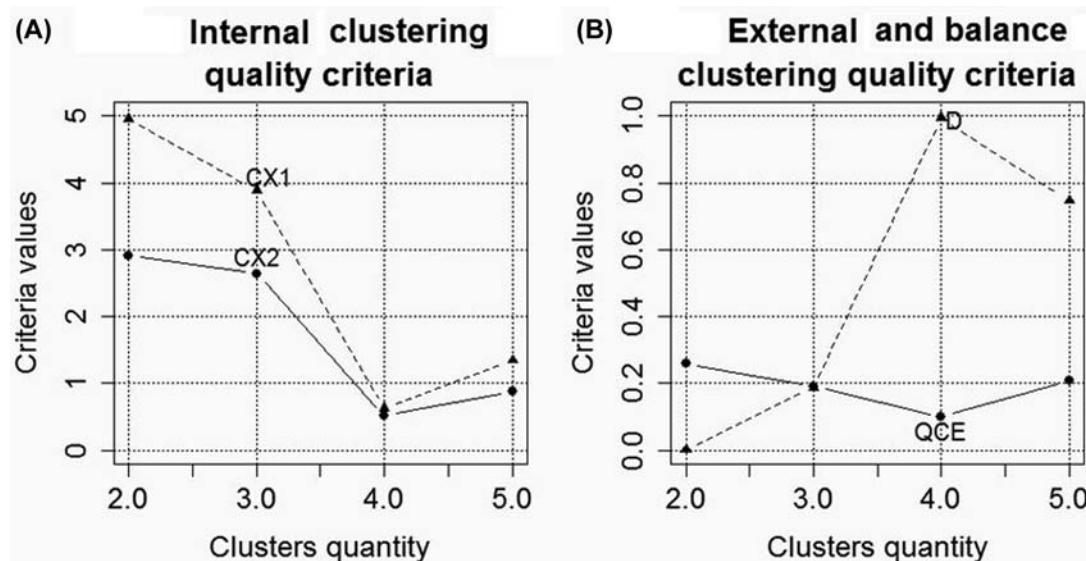


FIGURE 23.6

Charts of the internal, external, and complex balance clustering quality criteria.

PRACTICAL IMPLEMENTATION OF SOTA AND DBSCAN CLUSTERING ALGORITHMS WITHIN THE FRAMEWORK OF THE OBJECTIVE CLUSTERING INDUCTIVE TECHNOLOGY

SOTA clustering algorithm [10] represents a type of self-organizing neural networks based on Kohonen maps and Fritzke algorithm of a spatial cell structure growing [11]. Opposed to Kohonen maps that reflect a set of high-dimensional input data on the elements of two-dimensional array of small dimension, SOTA algorithm generates a binary topological tree. Fritzke algorithm performs self-organization output nodes of the network in such a way that the quantity of the nodes increases in the field of higher variability of the distances between gene expression profiles in the node. Two parameters are used to determine the effectiveness of SOTA clustering algorithm operation: a weight coefficient of the sister's cell (*s*cell) and a maximum divergence coefficient value. Weight coefficients of the parent's and winner's cells are determined automatically: $p_{cell} = s_{cell} \times 5$; $w_{cell} = p_{cell} \times 2$. This ratio is recommended by the authors of the algorithm [10]. The block scheme of the objective clustering model based on SOTA clustering algorithm is shown in Fig. 23.7.

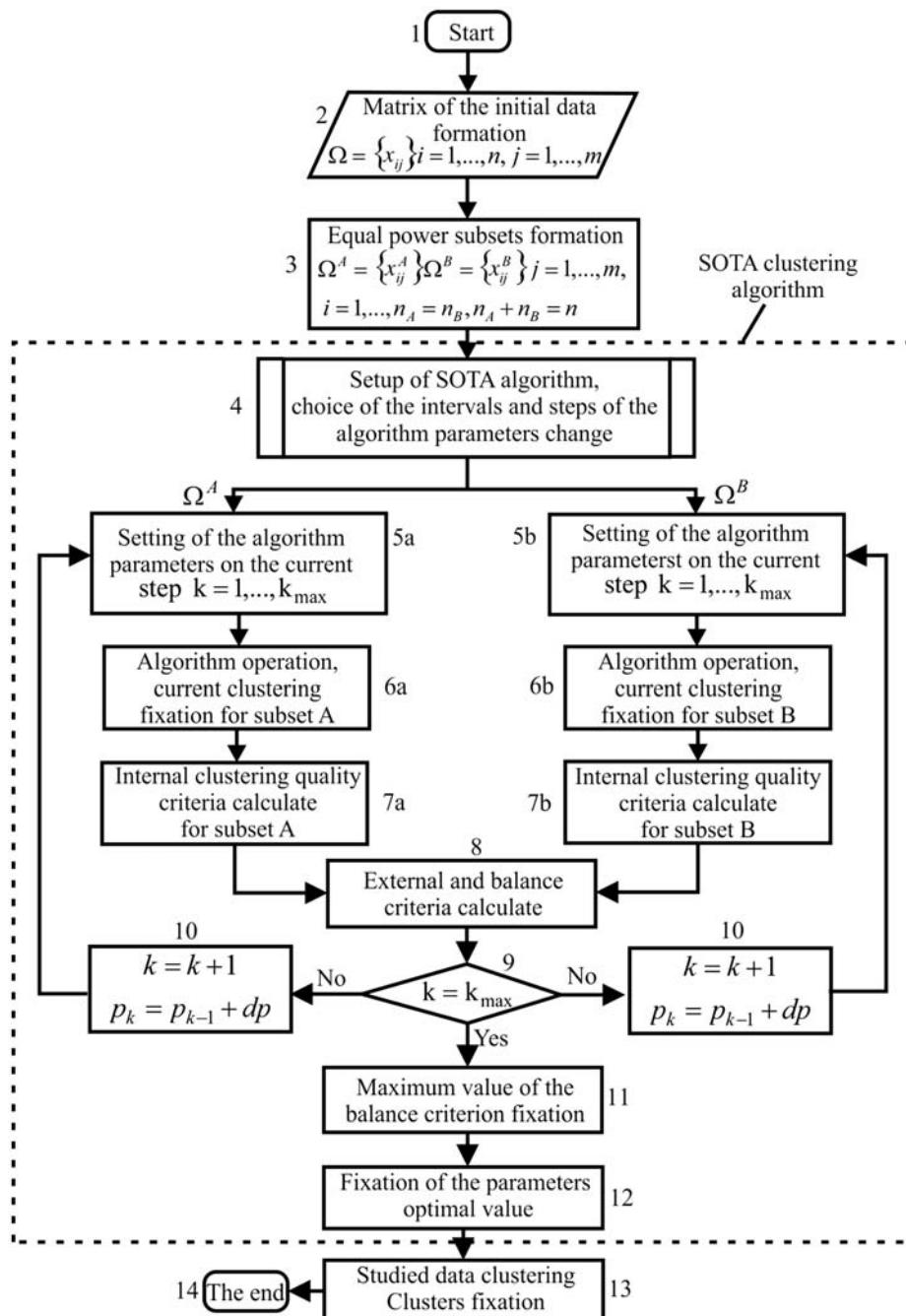
The implementation of this model involves the following steps:

1. Presentation of the studied data as a matrix $n \times m$, where n is the quantity of the studied objects or rows and m is the quantity of genes or columns.
2. Division of the initial data set into two equal power subsets.
3. Setup of SOTA clustering algorithm. Setting of the *s*cell weight parameter initial value, the interval and step of its change.
4. Data clustering on the equal power subsets *A* and *B* concurrently. The clusters formation and the internal, external, and balance clustering quality criteria calculation within a range of the algorithm parameters change.
5. Fixation of the optimal *s*cell parameter corresponding to the maximum value of the balance criterion.
6. Setting of the initial value of the maximum divergence parameter, range, and step of its change. Repeating step 4 of this procedure. Fixation of the optimal maximum divergence parameter.
7. Full data clustering by SOTA clustering algorithm using the optimal parameters of the algorithm operation.

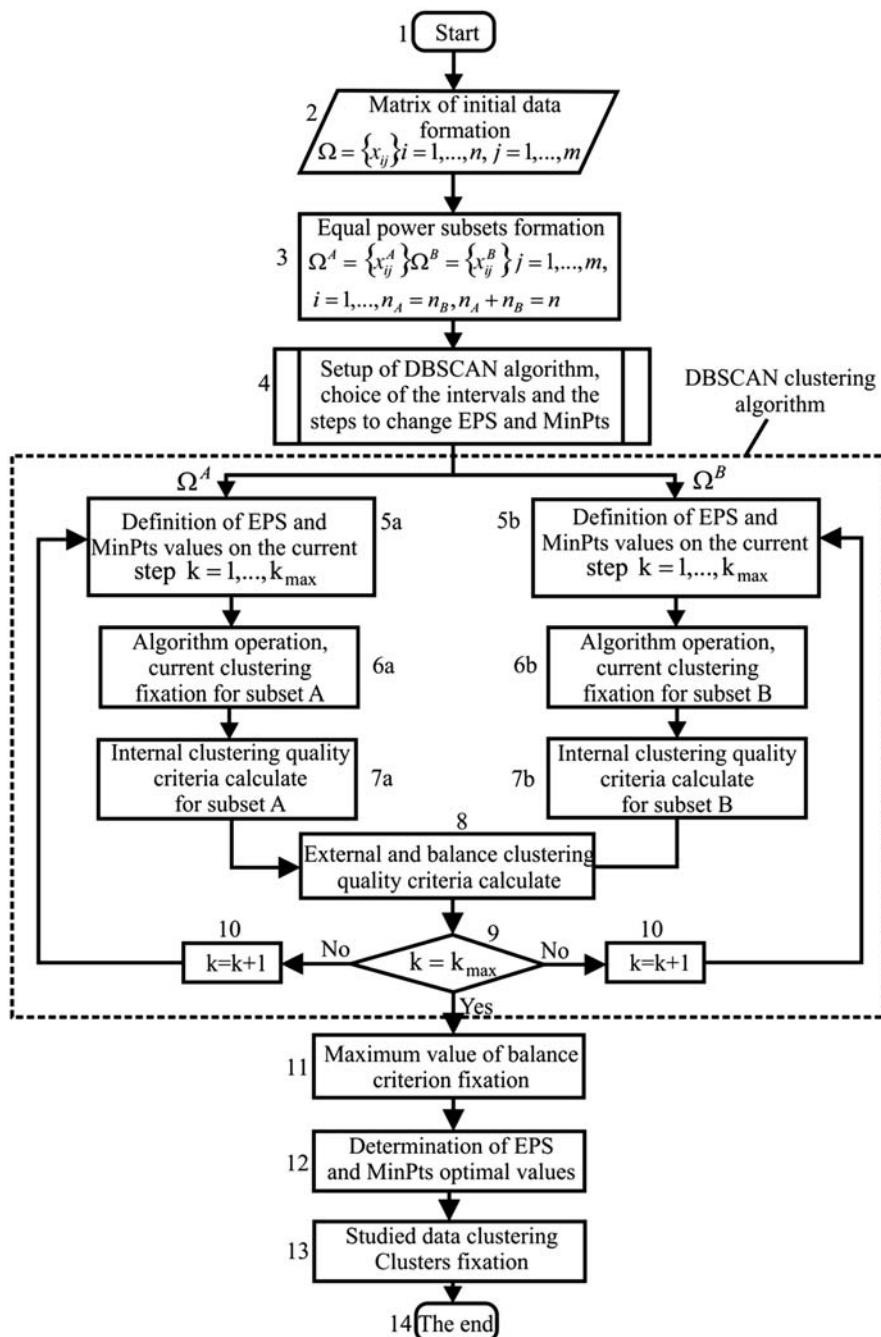
DBSCAN clustering algorithm [12] initially needs two parameters: EPS-neighborhood of points (*EPS*) and the least quantity of the points within EPS-neighborhood (*MinPts*). The choice of these parameters determines the character of the studied objects grouping during the algorithm operation. The technology based on the sorted *k*-dist graph is proposed by the authors to determine the optimal parameters of the algorithm operation [12]. However, the implementation of this technology does not allow us to determine the *EPS* and *MinPts* values exactly, and this fact influences the quality of the algorithm operation. To determine the *EPS* and *MinPts* optimal values, we propose to use the objective clustering inductive technology. The structural scheme of the objective clustering model based on DBSCAN clustering algorithm is presented in Fig. 23.8.

The implementation of this model involves the following steps:

1. Matrix of the studied data formation. Division of the initial data set into two equal power subsets.

**FIGURE 23.7**

Block scheme of the objective clustering model based on SOTA clustering algorithm.

**FIGURE 23.8**

Block scheme of the objective clustering model based on DBSCAN clustering algorithm.

2. The distance matrix between the studied gene expression profiles for both subsets is calculated using correlation distance. This distance matrix is used as the input data for the following processing.
3. Setup of DBSCAN clustering algorithm, choice of the range and steps of the EPS and MinPts values change.
4. Fixation of the MinPts value ($\text{MinPts} = 3$). Initialization of the $\text{EPS} = \text{EPS}_{\min}$.
5. Data clustering on the two subsets A and B using DBSCAN algorithm in the range from EPS_{\min} to EPS_{\max} . Clustering fixation at each step.
6. Calculation of the internal, external, and complex balance clustering quality criteria at each step of the algorithm operation.
7. Analysis of the balance criterion values. Fixation of the optimal value EPS , which corresponds to the maximum value of the balance clustering quality criterion.
8. Data clustering on the two equal power subsets A and B in the range from MinPts_{\min} to MinPts_{\max} . Clustering fixation at each step.
9. Repeating steps 6 and 7 of this algorithm for the MinPts values. Fixation of the EPS and MinPts optimal values, which correspond to the maximum of the complex balance clustering quality criterion.
10. Clustering of the studied data using the obtained parameters of DBSCAN algorithm operation.

RESULTS OF THE SIMULATION AND DISCUSSION

To evaluate the effectiveness of the algorithm operation within the framework of the proposed technology, we used the gene expression profiles obtained from lung cancer patients [17]. The simulation was performed using R software (version 3.4.2, packages “clValid,” “dbSCAN”) [26]. Fig. 23.9 shows the charts of the internal (Fig. 23.9A), external, and complex balance criteria (Fig. 23.9B) versus the EPS-neighborhood values. Two thousand gene expression profiles were investigated during the simulation process. Firstly, these profiles were divided into two equal power subsets using the correlation metric. Then, the dissimilarity matrices for all pairs of the studied objects of both subsets using the correlation distance were calculated. These dissimilarity matrixes have been used as the input data for the next steps of DBSCAN clustering algorithm operation. Three values of EPS-neighborhood have been selected based on the maximum values of the complex balance criterion, which is shown in Fig. 23.9B: $\text{EPS}_1 = 0.13$; $\text{EPS}_2 = 0.17$; $\text{EPS}_3 = 0.44$. Fig. 23.9C shows the charts of the complex balance criterion for the selected EPS versus the MinPts values. The analysis of the charts allows concluding that the best clustering in terms of maximum value of the complex balance clustering quality criterion is achieved using the following parameters of DBSCAN algorithm operation: (1) $\text{EPS} = 0.13$, $\text{MinPts} = 3$; (2) $\text{EPS} = 0.17$, $\text{MinPts} = 8$; (3) $\text{EPS} = 0.44$, $\text{MinPts} = 6$. However, the detailed analysis of the obtained results has shown what in the cases a and b the number of the clusters in the obtained clustering on the equal power subsets A and B differs. Only in the case of $\text{EPS} = 0.44$ and $\text{MinPts} = 6$ both clustering contained the same quantity of clusters.

In this case, the studied data were divided as follows: the first cluster contained 1663 profiles, the second cluster contained 16 profiles, and there were 321 profiles in the third cluster. The objects in the third cluster were identified as the noise component. Results of the simulation have also shown that the largest quantity of the gene expression profiles were concentrated in the first cluster. This fact can

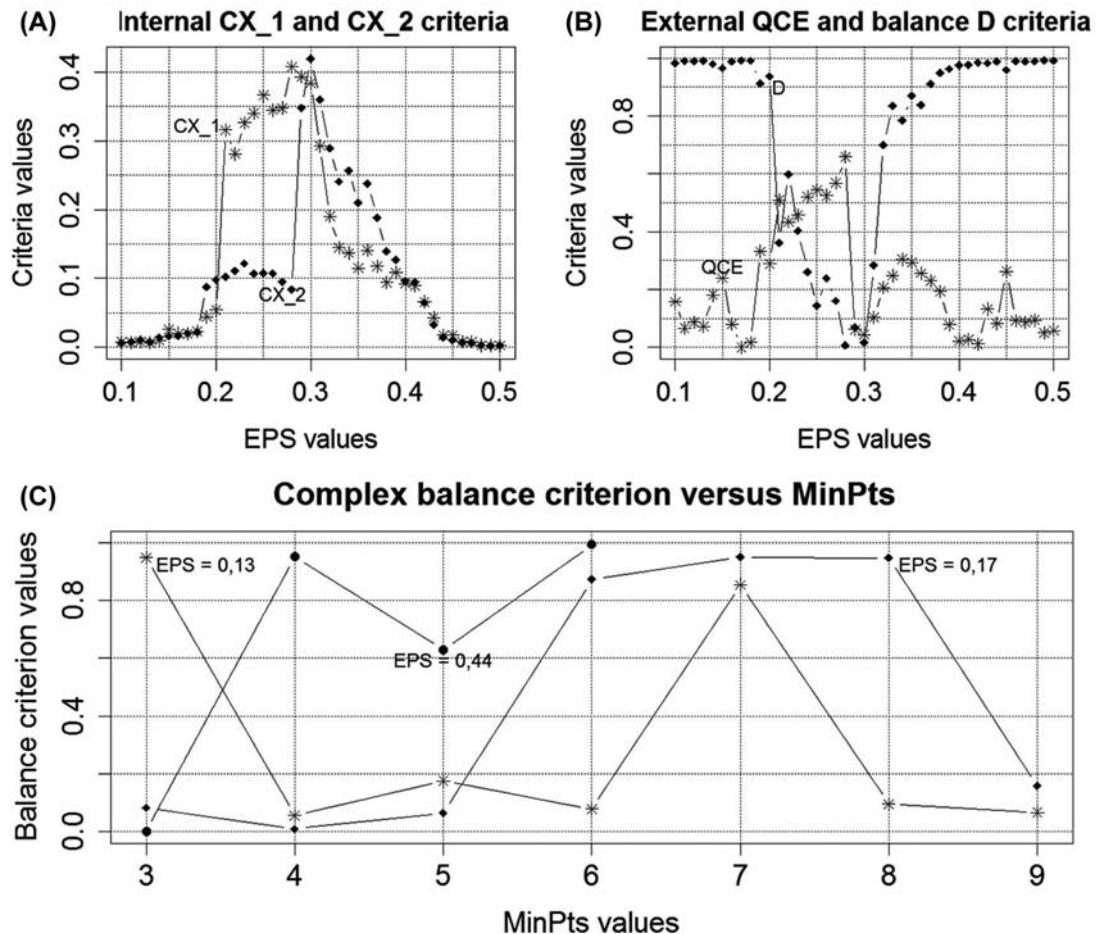


FIGURE 23.9

Results of the simulation of the lung cancer patients' gene expression profiles grouping with the use of DBSCAN clustering algorithm: (A) the internal clustering quality criteria for equal power subsets A and B versus the *EPS* values; (B) the external and complex balance clustering quality criteria versus the *EPS* values; (C) the complex balance criterion versus the *MinPts* values.

be explained in the following way: these genes define the main processes, which are carried out in biological organisms. Therefore, they have more correlation between each other in comparison with the genes in other clusters or the genes, which are identified as the noise.

The results of the simulation in the case of SOTA clustering algorithm application are presented in Fig. 23.10.

The maximum divergence value $E = 0.001$ was used during the simulation process. As it can be seen, the internal clustering quality criteria CX_1 and CX_2 are not effective to determine the optimal

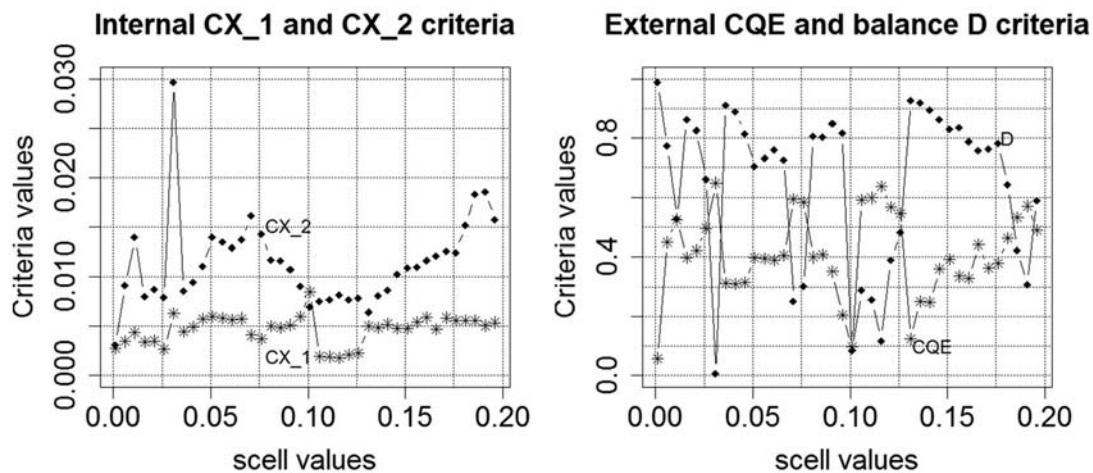


FIGURE 23.10

Results of the simulation of the lung cancer patients' gene expression profiles grouping in the case of SOTA clustering algorithm application.

parameters of the algorithm operation because the minimum values of these criteria do not correlate with each other. The external clustering quality criterion *CQE* has several local minimums corresponding to the successful grouping of the studied vectors. However, the analysis of the complex balance criterion values, which takes into account both the internal and external criteria, allows us to conclude that the best clustering was corresponded to 0.001 *scell* value. In this case, 6659 gene expression profiles were divided into two clusters. The first cluster contained 4276 gene expression profiles and the second was 2383. Variation of the maximum divergence value in the range from 0.001 to 1 has not changed the clustering results. Obtained results create the conditions to develop a step-by-step technology of gene expression profiles grouping at the early stage of the gene regulatory network reconstruction. Objective clustering based on DBSCAN algorithm allows us to select the genes with higher level of their mutual correlation. The noise component is also removed at this step. Then at the second step of data grouping, the selected profiles are divided into two groups using SOTA clustering algorithm. At the third step of the gene expression profiles grouping, biclustering technology is implemented on the obtained clusters.

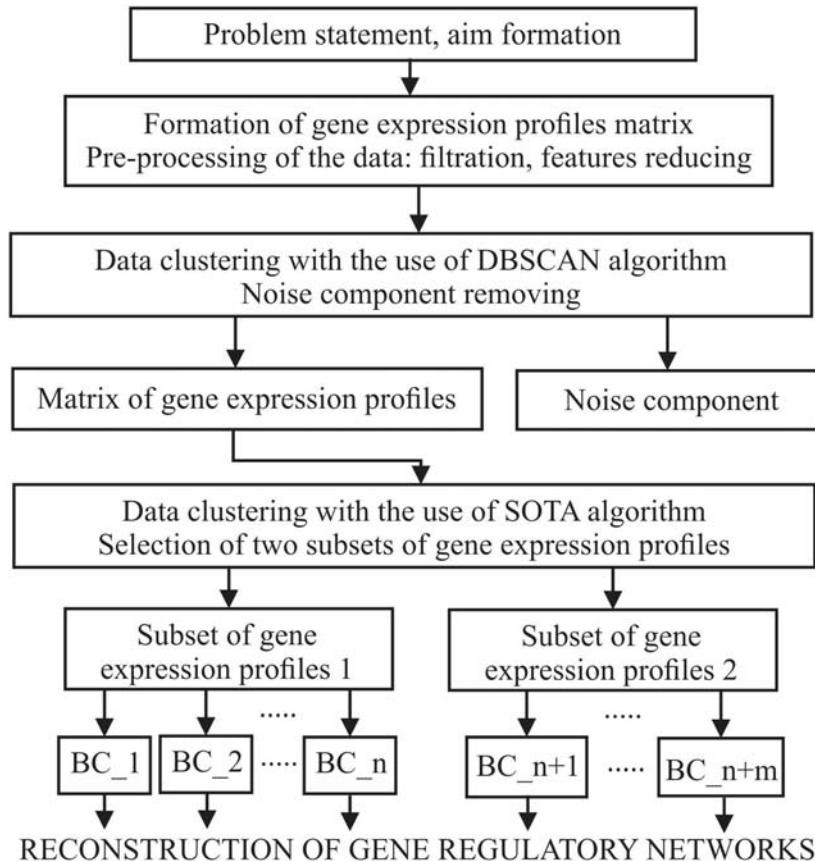
HYBRID MODEL OF CLUSTER–BICLUSTER ANALYSIS OF GENE EXPRESSION PROFILES

Structural block scheme of the cluster–bicluster analysis hybrid model is shown in Fig. 23.11.

Practical implementation of this model involves the following stages:

Stage I. Formation of data and their preprocessing.

1. Formation of matrix of the gene expression profiles, data filtration, and noninformative genes reducing.

**FIGURE 23.11**

Structural block scheme of the cluster–bicluster analysis hybrid model.

Stage II. Gene expression profiles clustering.

- Division of the initial gene expression profiles data set into two equal power subsets, which contain the same quantity of pairwise similar gene expression profiles.
- Determination of the optimal parameters for DBSCAN and SOTA clustering algorithms according to the technology presented in [15,16].
- Data clustering with the use of DBSCAN algorithm. Selection of a noise component. Formation of a new matrix of gene expression profiles for the following processing.
- Clusterization of the obtained gene expression profiles with the use of SOTA clustering algorithm. Formation of two subsets of gene expression profiles for the following bicluster analysis.

Stage III. Gene expression profiles biclustering on the obtained clusters.

- Choice of the biclustering algorithm, setup of its parameters.

- 7. Biclustering process.** Formation of the biclusters, which contain mutually correlated genes and conditions of experiment performing.

Stage IV. Reconstruction of the gene regulatory networks based on the data of the obtained biclusters.

It should be noted that practical implementation of hereinbefore-described step-by-step technology allows us to save more useful information at the stage of gene expression profiles preprocessing. Initial data set of the gene expression profiles, which can be obtained by DNA microarray experiments or by RNA sequencing methods, contains tens of thousands of genes. Bicluster analysis with the use of the initial data allows us to receive the biclusters of mutually correlated genes and conditions of experiment performing. However, the parallel gene expression profiles processing with the use of step-by-step data clustering technology promote to higher level of concentration of mutually correlated genes and conditions. This fact influences the following process of the gene regulatory networks reconstruction and simulation of their operation.

In the case of investigation of the epigenetics in lung cancer, a correct reconstruction of gene networks based on the gene expression profiles of lung cancer patients allow us to study the mechanism of interaction of appropriate groups of genes for better understanding of the particularities of these interactions on genetic level.

CONCLUSIONS

This chapter presents the hybrid model of cluster–bicluster technology based on the complex use of the objective clustering inductive technology of the gene expression profiles based on DBSCAN and SOTA clustering algorithms and bicluster analysis methods. Practical implementation of this technology at the early stage of gene regulatory network reconstruction allows us to increase the quality of the reconstructed gene network by more careful grouping of the mutually correlated genes and the conditions of experiment performing. This fact promotes to better understanding of the epigenetics in lung cancer during the simulation of the reconstructed gene regulatory network based on the lung cancer patient' gene expression profiles.

The implementation of the objective clustering inductive technology involves the concurrent data clustering on the two equal power subsets, which include the same quantity of pairwise similar objects. The correlation metric has been used as the proximity metric of lung cancer patients' gene expression profiles. The internal clustering quality criteria take into account both the character of objects distribution within the clusters relative to the centers of the appropriate clusters and character of the clusters distribution in features space. The external clustering quality criterion has been calculated as a normalized difference of the internal clustering quality criteria. Simulation process involved sequential evaluation of the internal and external criteria during the increase of the clusters quantity from K_{min} to K_{max} . The objective clustering is corresponded to the global minimum of the external clustering quality criterion. The gene expression profiles of Array Express Database, which were investigated on lung cancer, have been used as the experimental data. The number of clusters was changed from two to five during the simulation process. The objective clustering corresponded to the four-cluster structure. The results of the simulation have shown that WB index and Calinski-Harabasz criteria are the most effective in determining the objective clustering. These criteria have clearly expressed extrema, which corresponded to the optimal four-cluster structure both in the case of estimation of character of

objects and clusters distribution within the equal power subsets and in the case of estimation of the result of clustering difference on these subsets. To increase the sensitivity of the internal clustering quality criterion and corresponding external criterion the complex internal criterion has been proposed. It was calculated as a multiplicative combination of *WB* index and Calinski-Harabasz criteria. This criterion has been used as internal clustering quality criteria during the simulation process. The external criteria were calculated as a normalized difference of the internal clustering quality criteria, which were calculated based on the two equal power subsets. General Harrington desirability index, which was calculated on the basis of the internal and external criteria, has been used as a complex balance clustering quality criteria. The determination of the optimal parameters of the used algorithm operation was performed on the basis of the maximum value of the complex balance clustering quality criterion during algorithm operation. The results of the simulation have shown high efficiency of the proposed technology. In the case of DBSCAN clustering algorithm application, the noise component has been selected during the algorithm operation. Implementation of the proposed technology also allows us to group the gene expression profiles based on their similarity. The gene expression profiles with high correlation coefficient have been distributed into one cluster. This fact allows us to select the groups of gene expression profiles, which determine the main processes in the biological organisms. In the case of SOTA clustering algorithm application, the studied gene expression profiles were divided into two groups. Then, bicluster analysis is performed on the obtained clusters. Reconstruction of the gene regulatory networks involves the use of the genes and conditions of the obtained biclusters. To our knowledge, a correct reconstruction of gene networks based on gene expression profiles of lung cancer patients can promote to a better understanding of epigenetics regulations in lung cancer.

REFERENCES

- [1] Wei LK, Au A. Computational epigenetics. In: Handbook of epigenetics. 2nd ed. 2017. p. 167–90.
- [2] Pontes B, Giraldez R, Aguilar-Ruiz JS. Biclustering on expression data: a review. *J Biomed Inf* 2015;57: 163–80.
- [3] Kaiser S. Biclustering: methods, software and application. Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften am Institut Fur Statistik an der Fakultat Fur Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universitat Munchen; 2011.
- [4] Chi E, Allen G, Baraniuk R. Convex biclustering. *Biometrics* 2016;73:10–9.
- [5] Kriegel H, Kröger P, Sander J, Zimek A. Density-based clustering. *Wiley Interdiscip Rev Data Min Knowl Discov* 2011;1(3):231–40.
- [6] Berkhin PA. Survey of clustering data mining techniques. Grouping multidimensional data. Recent advances in clustering. Berlin, Heidelberg: Springer-Verlag; 2006. p. 25–72.
- [7] Bodyanskiy Y, Dolotov A, Vynokurova O. Self-learning cascade spiking neural network for fuzzy clustering based on group method of data handling. *J Autom Inf Sci* 2013;45(3):23–33.
- [8] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. 1998.
- [9] Madala H, Ivakhnenko A. Inductive learning algorithms for complex systems modeling. CRC Press; 1994. p. 26–51.
- [10] Dorazo J, Corazo J. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J Mol Evol* 1997;44(2):226–59.
- [11] Fritzke B. Growing cell structures a self-organizing network for unsupervised and supervised learning. *Neural Network* 1994;7(9):1441–60.

- [12] Ester M, Kriegel H, Sander J, Xu J. A density-based algorithm for discovering clusters in large spatial datasets with noise. In: Proceedings of the second international conference on knowledge discovery and data mining. Portland; 1996. p. 226–31.
- [13] Babichev S, Taif MA, Lytvynenko V. Inductive model of data clustering based on the agglomerative hierarchical algorithm. In: Proceeding of the 2016 IEEE first international conference on data stream mining and processing (DSMP); 2016. p. 19–22.
- [14] Babichev S, Taif MA, Lytvynenko V, Korobchynskyi M. Objective clustering inductive technology of gene expression sequences features. Communication in computer and information science. In: Proceeding of the 13th international conference beyond databases, architectures and structures. Ustron, Poland; 2017. p. 359–72.
- [15] Babichev S, Lytvynenko V, Osypenko V. Implementation of the objective clustering inductive technology based on the DBSCAN clustering algorithm. In: Proceeding of the XIIth IEEE international scientific and technical conference; 2017. p. 479–84.
- [16] Babichev S, Gozhyu A, Kornelyuk A, Lytvynenko V. Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm. Biopolym Cell 2017;33(5):379–92.
- [17] Beer D, Kardia S, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8(8):216–24.
- [18] Babichev S, Taif MA, Lytvynenko V, Osypenko V. Criterial analysis of gene expression sequences to create the objective clustering inductive technology. In: Proceeding of the 2017 IEEE 37th international conference on electronics and nanotechnology (ELNANO); 2017. p. 244–9.
- [19] Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat 1974;3:1–27.
- [20] Zhao Q, Xu M, Fräntti P. Sum-of-squares based cluster validity index and significance analysis. In: Proceeding of international conference on adaptive and natural computing algorithms; 2009. p. 313–22.
- [21] Hartigan J. Clustering algorithms. New York (NY): Wiley; 1975.
- [22] Hubert L, Schultz J. Quadratic assignment as a general data-analysis strategy. Br J Math Stat Psychol 1976; 29:190–241.
- [23] Ball JH, Hall GJ. Isodata: a novel method of data analysis and pattern classification. Menlo Park: Stanford Research Institute; 1965.
- [24] Xie X, Beni G. A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell (TPAMI) 1991; 13(8):841–7.
- [25] Harrington J. The desirability function. Ind Qual Control 1965;21(10):494–8.
- [26] Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comput Graph Stat 1996;5(3): 299–314.

This page intentionally left blank

Index

'Note: Page numbers followed by "f" indicate figures and "t" indicate tables.'

A

- Acetylation microarrays, 330
- Adenomatous polyposis coli 1A (*APC1A*), 296
- Affinity enrichment-based methods, 376–377
- Aflibercept, 360
- Akaike information criterion (AIC), 266–268
- ALlelic-Imbalance (ALI) detection, 73–74
- ALI detection. *See ALlelic-Imbalance (ALI) detection*
- Allele-specific DNA methylation, 165–167
- ALS-SMA, 138
- Alzheimer's disease (AD), 5
 - DNA methylation, 133, 135
 - DNA repair, 135
 - epigenetic changes, 132–133
 - epigenetic mechanisms, 132
 - gene-wise DNA methylation changes, 134–135
 - gene-wise histone alterations, 136–137
 - genome-wide DNA methylation alternations, 135
 - genomic risk factors, 137
 - histone acetylation changes, 136
 - hydroxymethylation, 134
 - hypomethylation, 134
 - polymorphisms, 137–138
 - systems level modules, 138, 139t–140t
- American College of Gastroenterology (ACG), 295
- Amyotrophic lateral sclerosis (ALS), 138
- Angiotensin II receptor type 1 (*AGTR1*), 227
- Antimethylcytosine, 374
- Antimethylcytosine-based immunoprecipitation, 377
- APC, 380
 - Apoptosis, 333–334
 - ArrayExpress, 168–169, 401
 - Artificial neural networks (ANN), 321
 - Assay of transposase-accessible chromatin (ATAC-Seq), 348–350
- Autoimmune diseases
 - data acquisition methods, 327–334
 - data analysis
 - post bisulfite treatment, 334
 - post immunoprecipitation studies, 333–334
 - post next-generation sequencing, 378–380
 - DNA methylation, 333–334
 - epigenetic changes in, 329–330, 333
 - multiple sclerosis, 331–332
 - rheumatoid arthritis, 330
 - systemic lupus erythematosus, 330–332

- type 1 diabetes, 332–333
- epigenetic databases, 384
- Histome, 385
- histone modification analysis, 380–382, 382t–383t
- MethBase, 386
- MethylomeDB, 385–386
- microarray-based detection, 328–330
- miRNA and targets prediction, 382–384, 384t
- miRWalk2.0, 386
- next-generation sequencing, 328–330
- ROADMAP Epigenomics, 386
- study designs, 327–334
- 5-aza-2'-deoxycytidine, endometrial carcinoma, 359

B

- Ball and Hall index (*BH*), 403, 405
- BAM, 313–314
- Bayesian deconvolution strategy, 377
- Bayesian integration, 313–314
- BCL2 interacting protein 3 (*BNIP3*), 296
- BeadStudio Software, 227
- Beta-Binomial Hierarchical model, 35
- Beta distribution
 - difference of two binomial proportions, 57–58
 - measurements with replicates distribution, 58–59
 - methylation ratio, 61–62
- Beta-Mixture Quantile Normalization (BMQN), 298
- BET proteins, 152
- Bevacizumab, 360
- Binomial proportion, 61
 - single, 62–63
 - two, 57–58
- BioGRID, 266–268
- Bioinformatics methods, 209–210
- Bioinformatics techniques, 172–173
- Bioinformatics tools, 312–313
- Biological significance, 33–34
- Biological variation, 33–34
- Biomarkers, 249
 - colorectal cancer
 - blood-based, 292–295
 - prognostic, 296
 - stool-based, 295–296
 - Bismark, 122, 298
 - Bisulfite conversion-based methods, 377
 - Bisulfite microarray data, 212–213, 223, 334

- Bisulfite-sequencing (BS-seq), 2–3, 122, 211–212, 221–222, 327–328, 379, 379t
- Bladder cancer
aging, 266, 268–269, 279–280
integrated genetic and epigenetic genome-wide network construction of, 276–278
core network biomarkers, 278–280
ECT2, DNA methylation of, 280
functional module network analysis, 281–283
genome-wide network projection, 273–276
miR1-2, 280
miR200b, 280
miRNA repression ability, 271–273
multiple drug combinations, 276, 283–285, 285t
omics data, data preprocessing of, 269
protein interaction ability, 271–273
smoking-related protein HSP90AA1, 280
statistical significance testing, 271–273
stochastic regression models for, 270–271
TF regulatory ability, 271–273
principal genome-wide network projection, 266
smoking, 266, 268–269, 279–280
- Blood, 172
- Bone morphogenetic protein 3 (*BMP3*), 295–296
- Bowtie, 69–70, 213, 347
- Bpipe, 314
- Branched chain amino acid transaminase 1 (*BCAT1*), 296
- Breast cancer
computational tools, 240, 243t
databases, 240, 242t
DNA methylation, 233–235
epigenetic modifications, 6
epigenetics, 233
histone modification, 235–238
noncoding RNAs, 238–240
“Broad” enrichment regime, 12–13, 13f
- Bromodomain and extraterminal (BET) proteins, 250
- BS-Seeker2, 122
- BS-seq, 298
- BWA, 69–70, 213
- Bwa*, 347
- C**
- CA125, 384
- Cadherin 5 (CDH5), 318
- Caenorhabditis elegans*, 79
- Calinski-Harabasz criterion (*CH*), 403, 407
- Calmodulin binding transcription activator 1 (*CAMTA1*), 318
- Cancer development, 188–190
- Cancer-specific differentially DNA-methylated regions (cDMRs), 210–211
- Cancer-specific DNA methylation, 360
- Capture Hi-C, 348
- Carcinogen-metabolizing genes, 228
- Cardiac reprogramming process, 156
- Cardiac specific chromatin signature, 151–152
- Cardiomyocytes (CMs), 149
chromatin conformation, 155–156
transcription
chromatin modification landscapes identification, 151–152
dynamics during heart development, 151–152
eukaryotic nucleus, 150
heart disease, dynamics of regulatory cis-elements, 152
- Cardiovascular diseases (CVDs)
cardiomyocytes, 149
induced cardiomyocytes, 149
iPSC reprogramming, 149
large-scale mapping, 150
- CASP8, 384
- Caspase-8 (*CASP8*), 319
- Castor zinc finger 1 (*CASZ1*), 318
- Catechol-O-methyltransferase (COMT), 382–383
- CDH1, 380
- Cell cycle inhibitor genes, 233
- CellNet, 321
- Cell type proportions, 172–173
- CHFR, 382
- ChIA-PET, 348
- Chip Analysis Methylation Pipeline (ChAMP), 298–299
- ChIP-enrichment
“broad” enrichment regime, 12–13, 13f
calling problem, 16
chromatin states segmentation, 16–19
genomic features, 16
H3K27me3, 14
normR, 14
“peak calling”, 12
“peaky” enrichment regime, 12–13, 13f
realistic background model, 14
sequencing depth normalization, 12–13
systematic biases, 12
- ChIP-on-chip, 327–328, 330
- ChIP-seq. *See* Chromatin-immunoprecipitation followed by sequencing (ChIP-seq)
- CHREMOFAC, 329
- Chromatin, 290
formaldehyde, 348
structure, higher-order, 347–356
- Chromatin conformation (CTCF), 155–156
- Chromatin immunoprecipitation (ChIP), 330
- Chromatin-immunoprecipitation followed by sequencing (ChIP-seq), 3–4, 110, 327–328, 330, 347
- Allelic-Imbalance detection, 73–74

- all-in-one data analysis pipelines, 72–73
 bioinformatics analysis pipeline, 67
 design, 68–69
 differential enrichment detection, 71–72
 mapping, 69–70
 melanoma, 347
 peak calling, 70–71
 quality, 69
 web lab experiment protocol, 67
- Chromatin states segmentation, 16
- Chromatin structure, 3
- ChromDB, 329
- ChromHMM, 16–17, 114, 347
- Chromosomal rearrangements, 247–248
- Chromosome conformation capture based methods, 347
- Chromothripsis, 316
- Cistrome, 72–73
- Cluster-bicluster analysis, 413–415
- Clustering algorithm
- density-based spatial clustering of application with noise, 408–411
 - self-organizing tree algorithm, 398, 408–411
- Clustering quality criteria, 399–405
- Clusters of genes, 111
- CNAmet, 313–314
- Coding sequence (CDS) regions, 84
- Colorectal cancer, 6
- biomarkers, 292–296, 297t
 - computational tools, 296–299, 300t–301t
 - epigenome-wide analysis, 291–292, 293t–295t
 - workflow, 299–302
- Comparative high-throughput arrays of relative methylation (CHARM), 374
- Competitive endogenous RNA (ceRNA), 6, 238–240
- Computational network-based model, 6
- Confidence interval (CI)
- difference of two binomial proportions, 63–64
 - single binomial proportion, 62–63
- Connectivity Map (CMAP), 266–268
- Convolutional neural network (CNN), 116
- Copy number alterations (CNAs), 247–248, 298
- Copy number variation (CNV), 74
- Core network biomarkers
- in bladder carcinogenesis, 273, 278–280
 - downstream genes in, 268
 - ER pathway, 268, 279
 - module network of, 268
 - SUP pathway, 268, 279
 - TFs/proteins in, 268, 276–277
 - TNF pathway, 268, 279
- Core proteins, 273–275
- COSMOS, 314
- COX7A1*, 186
- CpG island. *See* Cytosine-phosphate-guanine (CGIs) islands
- CpG-island methylator phenotype (CIMP) clusters, 343–344
- Cpipe, 314
- Credible methylation difference (CDIF), 34–37
- Crossbow, 314
- Cross-referencing data, 169–170
- CTNNB1* gene, 370
- Cumulative distribution function (cdf), 25
- Cyclin dependent kinase inhibitor 2A/RB transcriptional corepressor 1 provided (P16INK4a/pRB), 368
- Cytochrome c oxidase polypeptide 7A1 (COX7A1), 186
- Cytosine-phosphate-guanine (CpG), 182
- Cytosine-phosphate-guanine (CGIs) islands, 2–3, 219–220, 234–235, 289–290

D

- Data analysis tools, for DNA methylation, 300t–301t
- DataBase of human Transcriptional Start Sites (DBTSS), 329
- Data integration and visualization, an online platform (DaVIE), 213–214
- DaVIE. *See* Data integration and visualization, an online platform (DaVIE)
- Death receptor-3 (DR-3), 332
- Deep networks, 116
- DeepSEA, 116
- Dementia, 132–133
- Density-based spatial clustering of application with noise (DBSCAN) clustering algorithm, 408–411
- Diabetes mellitus, 181
- DIANA-microT-CDS, 85
- Differential enrichment (DE) analysis tool, 71–72
- Differentially methylated CpGs (DMCs), 44–45, 64
- Differentially methylated genomic blocks (DMB), 298
- Differentially methylated region (DMR), 33–34, 44–45, 298, 374
- hidden Markov model, 65
 - obesity, 169–170
 - simply group DMCs, 65
- DiseaseMeth version 2.0, 121
- Disheveled binding antagonist of beta catenin 2 (*DACT2*), 296
- DMRs. *See* Differentially methylated region (DMR)
- DNA damage, 251
- DNA fraction, 189
- DNA hydroxymethylation, 249–250
- DNA hypermethylation, 202–203, 386
- DNA hypomethylation, 202–203
- in autoimmune diseases, 329
 - RA synovial fibroblasts, 332
 - systemic lupus erythematosus, 333

- DNA methylation, 1–3, 289. *See also* Model-based analysis of bisulfite sequencing data (MOABS)
- Alzheimer’s disease (AD), 133, 135
 - gene-wise changes, 134–135
 - genome-wide alternations, 135
 - assessment technique, 165–167
 - autoimmune diseases, 333–334
 - in autoimmune diseases, 329
 - breast cancer, 233–235
 - cancer-specific, 360
 - in colorectal cancer
 - biomarkers, 292–296, 297t
 - computational tools, 296–299, 300t–301t
 - epigenome-wide analysis, 291–292, 293t–295t
 - workflow, 299–302
 - cytosine-phosphate-guanine islands, 289–290
 - detection of allele-Specific, 44–46
 - of ECT2, 280
 - in endometrium, 385–386
 - erasers, 358
 - readers, 358
 - writers, 358
 - head and neck cancer, 204–205
 - bisulfite microarray data, 212–213
 - bisulfite-sequencing data, 211–212
 - data, 210–214
 - data visualization and statistical analysis, 213–214
 - enrichment-based data, 213
 - during heart development and disease
 - hydroxymethylation, 155
 - mice lacking DNMT, 154
 - promoter regions, 154–155
 - whole-genome bisulfite sequencing, 153
 - in heterogeneous tissues, 167
 - oral cancer
 - advancement, 227–228
 - biomarker, 224–227, 225t
 - data visualization, 224
 - in prostate cancer, 249
 - schizophrenia (SCZ)
 - antipsychotic drugs, 127–128
 - methodology, 121–122
 - network, 123
 - prediction applications, 123, 124t
 - software and data sets, 167–169
 - type 2 diabetes mellitus, 182–183
- DNA methyltransferase (DNMT), 1–2, 202–203, 289, 332, 373
- DNA mismatch repair genes, 377–378
- DNA modifications
 - 5-hydroxymethylcytosine, 346
 - 5-methylcytosine, 345–346
- DNA repair, 135
 - genes, 233
- DNase I digestion (DNaseI-Seq), 348
- DNMT1, 153
- Dnmt3a, 154
- Drug Gene Interaction Database (DGIdb), 266–268
- Dynamic modeling, 321
- ## E
- Early-life adversity, 127–128
- ECT2, DNA methylation of, 280
- Effect size, 170
- EF-hand domain family member D1 (*EFHD1*), 292–295
- Embryonic stem cell (ESC), 47–48
- Encode project, 380–382
- Encyclopedia of DNA Elements (ENCODE), 253–255
- The Encyclopedia of DNA Elements (ENCODE), 107–108, 169–170
- Endometrial carcinoma (EC)
 - classification, 367
 - DNA hypermethylation for treatment, 386
 - DNA methylation machinery in, 385–386
 - epigenetic alterations
 - affinity enrichment-based methods, 376–377
 - bisulfite conversion-based methods, 377
 - DNA mismatch repair genes, 377–378
 - enzyme digestion-based methods, 376–384
 - steroid receptor genes, 378–380
 - tumor suppressor genes, 380–382
 - hypermethylated genes in, 376, 378t
 - microRNA aberrant methylation in, 384–386
 - molecular signaling pathways, 368–370
 - HER-2/neu, 375–376
 - MAPK/ERK, 373–376
 - PI3/AKT/mTOR, 370–372
 - VEGF, 375
 - VEGFR, 375
 - WNT/β-catenin, 373–374
 - mortality rate, 365–366
 - risk factors, 366–367
 - tumor suppressor miRNAs in, 385- Endometrioid adenocarcinoma (EAC), 372t, 376
- Endometrioid endometrial carcinoma (EEC), 343
- Endometrium, DNA methylation in, 385–386
- Endoplasmic reticulum (ER) signaling pathway, 268, 279, 281–282
- Enhancer, 290, 346
- Enhancer of zeste homolog 2 (*EZH2*), 318
- Enrichment-based method, 223–224
- Enzyme digestion-based methods, 376–384
- EPCAM* gene, 377
- EpiCSeg, 17

Epidaurus, 6–7, 253
 Epidermal growth factor (EGF), 372
 Epigenetic databases, 384
 Epigenetic instability, 248
 Epigenetic markers, 289
 Epigenetic modifications, 107–108
 Epigenetic regulatory network, 5–6
 Epigenetics, 131, 150, 182, 289
 and cancer, 6–7
 computational approaches, 1
 DNA methylation, 1–3
 histone modifications, 3–4
 mechanisms, 1
 metabolic and cardiac disorders, 4–5
 miRNAs, 4
 neurological disorders, 5–6
 Epigenomics, 150, 312
 covalent modifications of DNA, 11
 differences, 19–21
 functional elements, 11
 histone proteins, 11
 Erasers, 358
 ERG, 250
 Erlotinib, 360
 Estrogen, 377–378
 Estrogen response genes, 233
 Euclidean metrics, 401–402
 European Promoter Database (EPD), 329
 Exons, 312

F

False discovery rate (FDR), 30–31
 FASTQ files, 313–314
 Fecal occult blood test (FOBT), 292
 Fibrillin-1 gene (*FBN-1*), 295
 Fisher's exact test *P*-value (FETP) method, 33–34, 39–44
 5-fluorouracil (5-FU), 296
 5-Hydroxymethylcytosine (5hmC), 249
 5-methylcytosine (5mC), 289
 Forkhead Box I2 (*FOXI2*), 227
 Formaldehyde-assisted isolation of regulatory elements (FAIRE-Seq), 348
 Free energy, 82
 Fritzke algorithm, 408
 The Functional Annotation of the Mammalian Genome 5 (FANTOM5) Project, 169–170
 Functional module network analysis, 281–283

G

Galaxy, 72–73, 314–315
 Gefitinib, 360
 Gene-coding region, 126

Gene Expression Omnibus (GEO), 168–169, 332
 Gene expression profiles
 affinity metrics, 399–405
 cluster-bicluster analysis, 413–415
 clustering quality criteria, 399–405
 objective clustering process, 405–407
 Gene fusions, 247–248
 GenePattern, 314–315
 Gene promoters, 290
 Gene regulatory network (GRN), 265, 270, 397–398
 Gene set enrichment analysis (GSEA), 48, 170, 298
 Gene signature association analysis (GS2A), 253
 GenomeSpace, 314–315
 Genome-wide association studies (GWAS), 127
 Genome-wide profiling efforts, 128
 Genomics, 312
 Glucokinase gene (*GCK*), 185
 G–U wobble, 82

H

Harrington desirability function, 404
 Hartigan index, 405–407
 Head and neck cancer (HNC)
 computational analysis
 bioinformatics methods, 209–210
 DNA methylation data, 210–214
 development, 201–202
 DNA methylation, 204–205
 bisulfite sequencing, 206–208
 combined bisulfite restriction analysis assay, 206
 enrichment-based methods, 209
 methylated DNA immunoprecipitation, 209
 methylation specific PCR, 206
 microarray, 208–209
 pyrosequencing, 208
 whole genome bisulfite sequencing, 208
 environmental risk factors, 201–202
 epigenetic alteration, 202–203, 202t
 molecular changes, 202t
 Hematopoietic stem cell (HSC), 47–48
 HER-2/neu, 375–376
 Heterogeneity, 109
 Hi-C, 348
 HICHiP, 72, 348
 Hidden Markov model (HMM), 2–3, 65, 114
 High performance liquid chromatography ultraviolet (HPLC-UV), 165–167
 High-risk neuroblastoma (HRNB), 311–312
 CpGs, 319
 genetic landscape of, 316–317
 MYCN amplification, 318
 somatic mutations, 316–317

- High-throughput technologies, 1
 Histone, 385
 Histone acetylation, 3
 Histone acetyltransferase (HAT), 3, 250, 332
 Histone code, 150–152, 203
 Histone deacetylase (HDAC), 136, 250, 332
 Histone modification, 3–4, 152
 Alzheimer's disease (AD), 136–137
 autoimmune diseases, 380–382, 382t–383t
 breast cancer, 235–238
 chromatin structure alter, 203
 ChromHMM segmentation for, 17
 gene expression, 111
 melanoma, 346
 in prostate cancer, 250
 RA synovial fibroblasts, 332
 systemic lupus erythematosus, 333–334
 type 2 diabetes mellitus, 183
 Histone octamer, 150
 H3K4me3, 14–16
 H3K27me3, 14
 hMLH1 expression, 377
HNC. *See* Head and neck cancer (HNC)
HOXA9, 227
HOXA11, 382
HuaChanSu (HCS), 279
HugeSeq, 314
 Human body, 172
 Human CpG Microarrays, 291
 Human DNA Methylation Microarray, 291
 Human genome DNA, 347
 Human Genome Project (HGP), 266
 HumanMethylation450 BeadChips, 377
 Human Transcriptional Regulation Interactions database (HTRIdb), 266–268
 Hydroxymethylation, 134, 155
 5-hydroxymethylcytosine (5hmC), 33, 52–57
 in melanoma, 346
 methylation ratio, 58
 Hypermethylation, 249
 Hypomethylated regions, 66
 Hypomethylation, 134, 249, 334
- I**
- IKAROS family zinc finger 1 (*IKZF1*), 296
 Illumina, 314–315
 Illumina BeadScan, 212–213
 Illumina Genome Analyzer Platform, 331
 Illumina GenomeStudio software, 212–213
 Illumina HiSeq2500 NGS technology, 5
- Immunoprecipitation
 antimethylcytosine-based, 377
 methylated DNA, 377
 Induced cardiomyocytes (iCMs), 149
 Induced pluripotent stem cell (iPSC) reprogramming, 149
 Infinium beadarrays, 170
 Infinium 450K array, 249
 Infinium MethylationEPIC BeadChip, 291
 Ingenuity pathway, 228
 Insulin resistance, 185
 Integrated genetic and epigenetic genome-wide network (IGEN)
 construction of, 276–278
 core network biomarkers, 278–280
 ECT2, DNA methylation of, 280
 functional module network analysis, 281–283
 genome-wide network projection, 273–276
 miR1-2, 280
 miR200b, 280
 miRNA repression ability, 271–273
 multiple drug combinations, 276, 283–285, 285t
 omics data, data preprocessing of, 269
 protein interaction ability, 271–273
 smoking-related protein HSP90AA1, 280
 statistical significance testing, 271–273
 stochastic regression models for, 270–271
 TF regulatory ability, 271–273
 Integrated Genetic and Epigenetic Network (IGEN) system, 7
 Integrated Transcription Factor Platform (ITFP), 266–268
 Integrative analysis, 251–252
 deep learning, 116–117
 functional regulatory regions identifications, 114–115
 quality control and data preprocessing, 109–110
 self-organizing map (SOM), 115
 TF bind and histone modifications relationship, 111
 mixture model, 112–114
 regression analysis, 111–112
 Integrative epigenomic approaches, 6–7
 Integrative Genomics Viewer (IGV), 382
 Integrative omics, 313, 320
 International Diabetes Federation (IDF) data, 181
 International Human Epigenome Consortium (IHEC), 221
 Intestinal mucosa, 330
 Introns, 312
 Irreproducible discovery rate (IDR), 71
- J**
- JASPER, 330
 Journal of Clinical Investigation, 187–188

K

KEGG, 122
 Kohonen maps, 408
 K-ras, 370
 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, 266–268

L

Lambda spike-in DNA, 345
 Library complexity, 69
 Linear discriminant analysis (LDA), 115
 lncRNA genes, 357
 Local AU flanking content, 83
 Logistic model, 111–112
 Long interspersed nuclear elements (LINE), 249, 252–255
 Long noncoding RNA (lncRNAs), 250–251
 Low-risk neuroblastomas (LRN), 318
 Lung cancer
 DBSCAN clustering algorithm, 408–411
 gene expression profiles
 affinity metric and clustering quality criteria, 399–405
 cluster-bicluster analysis, 413–415
 objective clustering process, 398–399, 405–407
 SOTA clustering algorithm, 408–411
 Lynch syndrome, 376–377

M

Machine learning-based modeling, 321
 Machine learning-based target prediction tools, 84
 MacroH2A, 333
 MACS2, 14, 15f
 Manhattan metrics, 401–402
 Mann–Whitney U-statistic test, 26–27
 Markov Chain Monte Carlo (MCMC), 113–114
 Matrix metalloproteinases (MMPs), 332
 MeDIP, 298, 328t, 374, 380t
 MeInfoText, 210
 Melanin, 343
 Melanocytes, 343
 Melanogenesis associated transcription factor (MITF), 347
 Melanoma
 ATAC-Seq analysis, 348–350
 chromatin states, 346
 computational tools for epigenomic data, 349t–350t
 DNA modifications
 5-hydroxymethylcytosine, 346
 5-methylcytosine, 345–346
 higher-order chromatin structure, 347–356
 histone modifications, 346
 from melanocytes, 343
 nucleosome positioning, 348

treatment for, 343

Mendelian randomization, 171–172
 Metabolic and cardiac disorders, 4–5
 Metabolomics, 312
 Meta-dimensional analysis, 320
 Metastatic castration-resistant prostate cancer (mCRPC), 247
 MethBank, 121
 MethBase, 386
 MethDB, 210
 Methyl-Analyzer software, 385–386
 Methylated DNA immunoprecipitation (MeDIP), 209, 291, 319
 Methylation array, 319
 MethylatIOn INTegration (Mint), 256
 Methylation Mapping Analysis, 385–386
 Methylation quantitative trait loci (meQTLs), 127
 Methylation-sensitive restriction enzymes (MREs), 291–292, 373
 Methylation-specific oligonucleotide (MSO) microarrays, 329
 Methyl-CpG-binding domain (MBD) proteins, 374
 5-Methylcytosine, 345–346
 MethylomeDB, 121, 385–386
 Methylomics
 bisulfite-microarray method, 223
 bisulfite-sequencing method, 221–222
 DNA methylation data visualization, 224
 enrichment-based method, 223–224
 MethylPurify, 346
 Methylumi, 212–213
 miARma-Seq, 86
 Microarray, 329–330
 acetylation, 330
 autoimmune diseases, 328–330
 scanner, 329
 microRNAs (miRNAs), 4, 203, 250–251, 331
 applications, 80
 biogenesis, 79–80
 cell differentiation, 191
 endometrial carcinoma, aberrant methylation in, 384–386
 inhibition of β-cell proliferation, 191
 insulin sensitivity, 191
 negative regulation of β-cell survival, 191
 next-generation sequencing, 79
 target prediction, 382–384, 384t
 DIANA-microT-CDS, 85
 evolutionary conservation status, 82
 experimental validations, 84
 free energy, 82
 G–U wobble, 82
 high-throughput RNA-seq experiments, 86
 human, for computational tools, 86, 87t–94t
 local AU flanking content, 83

- microRNAs (miRNAs) (*Continued*)
- machine learning, 84
 - miRanda, 84
 - miRDB, 85
 - pattern-based approach, 84
 - PITA, 85
 - plant and animal, 80
 - prediction efficiency, 81
 - principles, 80–84
 - regulation, 80
 - seed sequence complementarity, 81
 - selected tools principle, 86, 95t–98t
 - STarMir webserver, 85
 - target-site abundance, 83
 - target-site accessibility, 83, 85
 - 3' UTR compensatory binding, 82–83
- Microsatellite instability (MSI), 376
- MicroSNiPer, 256
- Minfi, 212–213
- Minimum free energy (MFE), 82
- miR1-2, 280
- miR-124, 386–387
- miR126, 357
- miR-152, 386
- miR155, 282–283
- miR-203, 333
- miRanda, 84
- miR200b, 280
- miRDB, 85
- miRNA regulatory network (MRN), 270–271
- miRNAs. *See* microRNAs (miRNAs)
- miRNA-target gene association data, 266–268
- miRNomics, 86
- miRWalk2.0, 386
- Mismatch repair (MMR), 377
- Mitogen-activated protein kinase (MAPK/ERK), 368, 373–376
- Mixture model, 112–114
- MIxture Quantile (BMIQ) normalization, 168
- Model-based analysis of bisulfite sequencing data (MOABS)
- implementation, 37–39
 - novel and significant analysis, 57
 - performance
 - allele-specific DNA methylation detection, 44–46
 - vs. FETP, 39–44
 - 5hmC detection, 52–57
 - TFBSs, 47–52
 - software pipeline, 37–39
- Model-based analysis of regulation of gene expression (MARGE), 6–7, 253
- Mode of action by network identification (MNI), 321
- Modern bioinformatics tools, 1
- Molecular interaction network, 128
- Monozygotic twin pair, 380
- MOSAiCS, 70–71
- mRNA. *See* microRNAs (miRNAs)
- Multiparameter logistic regression model, 377
- Multiple concerted disruption (MCD) analysis, 313–314
- Multiple drug combination, 276, 283–285, 285t
- Multiple omics data, 109
- Multiple sclerosis (MS), 331–332
- Multiplicity-adjusted P value, 29–30, 30t
- MutL homolog 1 (*MLH1*), 296
- MYCN* amplification, 318–319
- Myelin basic protein (MBP), 334
- N**
- National Cancer Institute (NCI), 316
- National Center for Biotechnology Information (NCBI)
- Entrez Gene database, 266–268
- NDEx, 122
- NDRG family member 4 protein (*NDRG4*), 295–296
- Nebula, 72–73
- Nerve growth factor receptor (*NGFR*), 292–295
- Nervous system, 131–132
- NetDecoder, 321
- Network-based modeling, 321
- Neural crest progenitor (NCP) genes, 347
- Neuritogenesis, 316
- Neuroblastomas
- cells, 311
 - embryonic tissue, 311
 - high-risk, 311–312
 - low-risk, 318
 - omics
 - dynamic modeling, 321
 - epigenomics, 312
 - genome, 316–318
 - integrative, 313, 320
 - machine learning-based modeling, 321
 - network-based modeling, 321
 - next-generation sequencing, 312–313
 - proteomics, 312
 - reverse engineering, 321
 - transcriptome, 319–320
 - transcriptomics, 312
- personalized medicine for, 314
- prognostic factors for, 311–312
- staging system, 311, 312t
- treatment for, 311–312
- Neurodiseases (NDs), 138
- Neurogenin 1 (*NEUROG1*), 292–295
- Neurological disorders, 5–6
- NextBio Research, 314–315

Next-generation sequencing (NGS), 107–108, 249, 312–313
 autoimmune diseases, 328–330
 microRNAs, 79
 sequence alignment and assembly tools, 313, 313t
 of transcriptome, 124t
 NGSANE, 314
 NIH Roadmap Epigenomics Mapping Consortium, 107–108
 Noncoding RNAs (ncRNA), 203, 238–240
 Nonendometrioid endometrial carcinoma (NEEC), 344
 normR, 14, 15f
 Novel miRNAs, 4
 NPBin, 74
 Nuclear factor-B (NF-KB), 188
 Nucleosome, 3, 346, 348

O

Obesity, 167–169
 consequence, 171–172
 differentially methylated regions, 169–170
 Objective clustering inductive technology
 architecture, 399
 conceptual basis of, 398–399
 implementation, 399
 O-6-methylguanine-DNA-methyltransferase (MGMT), 377
 Omics data, 269
 Omics, neuroblastoma
 dynamic modeling, 321
 epigenomics, 312
 genome, 316–318
 integrative, 313, 320
 machine learning-based modeling, 321
 network-based modeling, 321
 next-generation sequencing, 312–313
 proteomics, 312
 reverse engineering, 321
 transcriptome, 319–320
 transcriptomics, 312
 Omics Pipe, 314–315
 OMICtools, 296–298
 Oncogenes, 219
 One-way analysis of variance (ANOVA), 269
 Oral cancer
 development, 219
 DNA hypomethylation, 220
 DNA methylation
 advancement, 227–228
 biomarker, 224–227, 225t
 epigenetic regulation, 220
 epigenetic silencing, 220
 epigenetics study, 221
 genotoxic agents, 220
 methylomics

bisulfite-microarray method, 223
 bisulfite-sequencing method, 221–222
 DNA methylation data visualization, 224
 enrichment-based method, 223–224
 Oral premalignant lesions (OPLs), 227
 Oral squamous cell carcinoma (OSCC), 219. *See also* Oral cancer
 oxBS-seq, 52–57

P

P53, 380
 PANTHER version 11, 122
 Parallel Processing Pipeline software for automatic analysis of Bisulfite Sequencing (P3BSseq), 299
 Parkinson's disease, 5–6
 Pathway Commons, 122
 PD_NGSAtlas, 122
 PDX1, 187
 Peak Based Correction (PBC), 298
 Peak calling, 12, 70–71, 107–108
 “Peaky” enrichment regime, 12–13, 13f
 Peptidyl arginine deiminase type II (PAD2), 334
 Phantompkqual, 347
 Phosphatidylinositol 3-kinase/AKT serine/threonine kinase1 provided/mammalian target of rapamycin (PI3K/AKT/mTOR), 368, 370–372
 Phred score, 378–379
 Picard, 314
 P16INK4a, 378–379
 PITA, 85
 Platelet-derived growth factor receptors (PDGF-R), 125–126
 Polycomb repression complex (PRC) marks, 250
 Polycomb repressive complex 1 (PRC1), 156
 Polycomb repressive complex 2 (PRC2), 318
 Position weight matrixes (PWMs), 116–117
 Position-wise content, 69
 Posttranslational histone modifications, 114
PPARGC1A, 185
 pRB, 380–382
 Precursor miRNA (pre-miRNA), 4
 Primary miRNA (pri-miRNA), 4
 Principal component analysis (PCA), 321
 Principal genome-wide network projection (PGNP), 266, 273–276
 Probabilistic graphical models, 313–314
 Probabilistic identification of combinations of target sites (PicTar), 85–86
 Probability density function, 57–58
 Proenkephalin (*PENK*), 227
 Progesterone, 377–378
 Prognostic biomarkers, 296

Prostate cancer (PCa)

- genomic alterations in
 - histone modifications, 250
 - long noncoding RNA, 250–251
 - microRNA, 250–251
- incidence, 247
- integrative analysis, 251–252
 - tools, 252–255
- potential applications for, 255–256
- prostate-specific antigen, 247
- single-nucleotide polymorphisms, 251–252
- tumor studies, 251–252, 252t

Prostate-specific antigen (PSA), 247

Protein interaction databases, 123

Protein phosphatase 1 regulatory subunit 3C (*PPP1R3C*), 292–295

Protein-protein interaction networks (PPINs), 265, 271

PTEN, 378

PubMeth, 210

Python toolkit, 314

Q

Quality control metrics, 109

Quality of reads, 69

R

Random Forest based Enhancer identification from Chromatin States (RFECS), 115

RAS-association domain family 1 isoform A (*RASSF1A*), 319Ras association domain family member 1 (*RASSF1A*), 296

RA synovial fibroblasts (RASFs), 332

R/Bioconductor, 212–213

Reactome pathway database, 122

Read alignment model, 37–39

Reduced representation bisulfite sequencing (RRBS), 291, 319, 346, 375

RefSeq genes, 170

RegNetDriver, 6–7, 253–255

Regression on Correlated Probes (RCP), 168

Reverse engineering modeling, 321

RFECS. *See* Random Forest based Enhancer identification from Chromatin States (RFECS)

Rheumatoid arthritis (RA), 330

Ridaforolimus, 360

RNA-induced silencing complex (RISC), 382–383

RNA sequencing (RNA seq), 110, 331

ROADMAP Epigenomics, 386

Roadmap Epigenomics Consortium (REMC), 253–255

Roadmap Epigenomics Project, 169–170

RRBS, 52–57

R, statistics software, 73

Ruffus, 314

S

Samtools, 314

Satellite cell, 172

SchizConnect, 122

Schizophrenia (SCZ)

candidate genes, 123

copy number variants, 121

disease mechanism, 123–125

DNA methylation

antipsychotic drugs, 127–128

methodology, 121–122

network, 123

prediction applications, 123, 124t

and epigenetic review, 126

pathway enrichment analysis, 125–126

polymorphisms, 121

SDMN, 123–125

Schizophrenia differential methylation genes (SDMG), 123–125

Schizophrenic differential methylation network (SDMN), 123–125

SCZ. *See* Schizophrenia (SCZ)Secreted frizzled-related protein gene 2 (*SFRP2*), 295–296

Seed region, 79

Seed sequence complementarity, 81

Self-organizing map (SOM), 115

Self-organizing tree algorithm (SOTA), 398, 408–411

Septin 9 gene (*SEPT9*), 295

Sequence alignment map (SAM), 313–314

Sequence-based imperfections, 82

Serpin family E member 1 (*SERPINE1*), 220Sigma², 256

Signal-to-noise ratio, 69

Signal transductions, 276–277

Significance-based Modules Integrating the Transcriptome and Epigenome (SMITE), 256

Single binomial proportion, 62–63

Single-cell DNA methylation, 167

Single-nucleotide variants (SNVs), 247–248

Skeletal muscle, 172, 192

Small RNAs (sRNAs), 331

SMAP, 299–302

Smoking-related protein HSP90AA1, 280

Snakemake, 347

Somatic reprogramming process, 156

SOX4 gene, 386

Spinal muscular atrophy (SMA), 138

SPRY2, 383

Squamous cell carcinomas (SCCs), 219

STarMir webserver, 85

Statistical methods, 107–108

Statistical models, 23

- closure principle, 27–28
 data analysis methods, 23
 dependencies, 23
 detect, 25
 finite sample modification, 28–29
 limiting null distributions, 26–27
 multiple test problems, 25–26
 multivariate distributions, 24
 real data analysis, 29–30
 two-group models, 24
 Statistical power, 33–34
 Steroid receptor genes, 378–380
 Stochastic regression models, 270–271
 of gene regulatory networks, 270, 272
 of miRNA regulatory network, 270–273
 of protein-protein interaction networks, 271
 Stool test, 295
 Structural variations (SV), 70, 247–248
 Studentized permutation approach, 28–29
 Subset-Quantile Normalization (SQN), 298
 Subset-Quantile Within Array Normalization (SWAN), 298
 SUMOylation, ubiquitination, and proteasome (SUP)
 pathway, 268, 279, 281–282
 Support vector machines (SVM), 321
 Suv39h1, 188
 Syndecan-2 (*SDC2*), 292–295
 Synovial fibroblasts, 330
 Systemic lupus erythematosus (SLE), 330–332
 SZDB, 122
 SZGR 2.0, 122
- T**
- TargetScan, 266–268
 Target-site abundance, 83
 Target-site accessibility, 83, 85
 Taverna workflow suite, 314–315
 T-box 5 (*TBX5*), 296
 T cell, ChIP-seq analysis of, 334
 T2DM. *See* Type 2 diabetes mellitus (T2DM)
 TEA domain transcription factor (TEADs), 347
 Telomerase reverse transcriptase (*TERT*), 318
 Temsirolimus, 360
 Ten-eleven translocase (*TET*) enzymes, 249–250
 Tet-assisted bisulfite sequencing (Tab-Seq), 319
 Tet methylcytosine dioxygenase (TET), 346
 Thalidomide, 360
 The Cancer Genome Atlas (TCGA), 107–108, 240, 314–315
 Consortium, 375
 network, 247
 Therapeutically Applicable Research to Generate Effective Treatments (TARGET), 316
 3D genome analysis techniques, 253–255
- Tissue factor pathway inhibitor 2 (*TFPI2*), 295–296
TMPRSS2-ERG fusion, 248
 Tn5 transposase, 348
 Tomoregulin-2 (*TMEFF2*), 292–295
 Topologically associated domains (TADs), 255
 Top-ranked pathways, 125–126
 Transcriptional start sites (TSS), 253
 Transcription factor (TF), 126
 gene expression
 mixture model, 112–114
 regression analysis, 111–112
 self-organizing map (SOM), 115
 Transcription factor AP-2 epsilon (*TFAP2E*), 296
 Transcription factor binding sites (TFBSs), 47–52
 TRANSFAC, 266–268
 Trastuzumab, 360
 Trim Galore, 345
 Tumor necrosis factor (TNF) signaling pathway, 268, 279, 281–282
 Tumor protein p53/cyclin-dependent kinase inhibitor 1A (P53/P21), 368
 Tumor suppressor genes, 189, 380–382
 Tumor suppressor miRNAs (TS-miRNAs), 385
 Two binomial proportions, 57–58
 Type 1 diabetes (T1D), 332–333
 Type 2 diabetes mellitus (T2DM), 5
 cancer development, 188–190
 characteristics, 181
 clinical presentation, 181
 concordance, 182
 drugs, 192–193
 epigenetic regulation, 185–187
 insulin resistance, 181
 miRNAs functional role. *See* microRNAs (miRNAs)
 pathogenesis, 187
 vascular complications, 187–188
- U**
- Ubiquitin C (UBC), 279
 UK Ovarian Cancer Population Study, 29–30
 United States Food and Drug Administration (US FDA), 289
 Uterine papillary serous carcinoma (UPSC), 372t, 376
 3' UTR compensatory binding, 82–83
- V**
- Valproic acid, 360
 Variant call format (VCF), 313–314
 Vascular endothelial growth factor (VEGF), 368, 375
 Vascular endothelial growth factor receptor (VEGFR), 368, 375
 Vimentin (*VIM*), 295–296
 Vorinostat, 360

W

- Web-based bioinformatics analysis pipelines, 72–73
Whole genome bisulfite sequencing (WGBS), 47–48,
165–167, 208, 291, 319, 374
Wingless-type MMTV integration site family member (WNT/
β-catenin), 368, 373–374
Wnt family member 5A (*WNT5A*), 296
Workflow management system (WMS), 313
Writers, 358

X

- Xie-Beni index (*XB*), 403, 405

Z

- Zic family member 1 (*ZIC1*), 227

Computational Epigenetics and Diseases

Volume 9

Edited by
Loo Keat Wei

Computational Epigenetics and Diseases, written by leading scientists in this evolving field, provides a comprehensive and cutting-edge knowledge of computational epigenetics in human diseases. In particular, the major computational tools, databases, and strategies for computational epigenetics analysis, for example, DNA methylation, histone modifications, microRNA, noncoding RNA, and ceRNA, are summarized, in the context of human diseases.

This book discusses bioinformatics methods for epigenetic analysis specifically applied to human conditions such as obesity, diabetes mellitus, schizophrenia, Alzheimer disease and autoimmune disorders. Additionally, different organ cancers, such as breast, lung, and colon, are discussed.

This book is a valuable source for graduate students and researchers in genetics and bioinformatics, and several biomedical field members interested in applying computational epigenetics in their research.

Key Features

- Provides a comprehensive and cutting-edge knowledge of computational epigenetics in human diseases.
- Summarizes the major computational tools, databases, and strategies for computational epigenetics analysis, such as DNA methylation, histone modifications, microRNA, noncoding RNA, and ceRNA.
- Covers the major milestones and future directions of computational epigenetics in various kinds of human diseases such as obesity, diabetes, heart disease, neurological disorders, cancers, blood disorders and autoimmune diseases.

Science/Life Sciences/Genetics
and Genomics



ACADEMIC PRESS

An imprint of Elsevier

elsevier.com/books-and-journals

ISBN 978-0-12-814513-5



9 780128 145135