

HANDBOOK OF STATISTICAL GENOMICS

FOURTH EDITION

Editors
DAVID J. BALDING
IDA MOLTKE
JOHN MARIONI

VOLUME 1

WILEY

BALDING
MOLTKE
MARIONI

BALDING
MOLTKE
MARIONI

HANDBOOK OF
STATISTICAL GENOMICS

HANDBOOK OF
STATISTICAL GENOMICS

FOURTH
EDITION

VOLUME 2

WILEY

WILEY

FOURTH
EDITION

VOLUME 1

WILEY

WILEY

Handbook of Statistical Genomics

Handbook of Statistical Genomics

Volume 1

Edited by

David J. Balding

University of Melbourne, Australia

Ida Moltke

University of Copenhagen, Denmark

John Marioni

University of Cambridge, United Kingdom

Fourth Edition

Founding Editors

Chris Cannings

Martin Bishop

WILEY

This fourth edition first published 2019
© 2019 John Wiley & Sons Ltd

Edition History

John Wiley & Sons, Ltd (1e, 2001), John Wiley & Sons, Ltd (2e, 2003), John Wiley & Sons, Ltd (3e, 2007)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of David J. Balding, Ida Moltke and John Marioni to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Balding, D. J., editor. | Moltke, Ida, editor. | Marioni, John, editor.

Title: Handbook of statistical genomics / edited by David J. Balding (University of Melbourne, Australia),

Ida Moltke (University of Copenhagen, Denmark), and John Marioni (University of Cambridge, United Kingdom).

Other titles: Handbook of statistical genetics. | Handbook of statistical genetics.

Description: Fourth edition. | Hoboken, NJ : Wiley, 2019– | Previous title: Handbook of statistical genetics. |

Includes bibliographical references and indexes. |

Identifiers: LCCN 2018060346 (print) | LCCN 2019003813 (ebook) | ISBN 9781119429227 (Adobe PDF) |

ISBN 9781119429258 (ePub) | ISBN 9781119429142 (hardcover)

Subjects: LCSH: Genetics—Statistical methods—Handbooks, manuals, etc.

Classification: LCC QH438.4.S73 (ebook) | LCC QH438.4.S73 H36 2019 (print) | DDC 572.8/60727—dc23

LC record available at <https://lcn.loc.gov/2018060346>

Cover Design: Wiley

Cover Image: © Zita / Shutterstock

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

10 9 8 7 6 5 4 3 2 1

Contents

Volume 1

List of Contributors	<i>xxiii</i>
Editors' Preface to the Fourth Edition	<i>xxvii</i>
Glossary	<i>xxix</i>
Abbreviations and Acronyms	<i>xxxix</i>

1	Statistical Modeling and Inference in Genetics	1
	<i>Daniel Wegmann and Christoph Leuenberger</i>	
1.1	Statistical Models and Inference	1
1.1.1	Statistical Models	2
1.1.2	Inference Methods and Algorithms	4
1.2	Maximum Likelihood Inference	4
1.2.1	Properties of Maximum Likelihood Estimators	6
1.2.2	Quantifying Confidence: the Fisher Information Matrix	8
1.2.3	Newton's Method	9
1.2.4	Latent Variable Problems: the EM Algorithm	11
1.2.5	Approximate Techniques	16
1.3	Bayesian Inference	20
1.3.1	Choice of Prior Distributions	21
1.3.2	Bayesian Point Estimates and Confidence Intervals	22
1.3.3	Markov Chain Monte Carlo	23
1.3.4	Empirical Bayes for Latent Variable Problems	30
1.3.5	Approximate Bayesian Computation	31
1.4	Model Selection	37
1.4.1	Likelihood Ratio Statistic	37
1.4.2	Bayesian Model Choice	38
1.5	Hidden Markov Models	40
1.5.1	Bayesian Inference of Hidden States Using Forward-Backward Algorithm	42
1.5.2	Finding the Most Likely Hidden Path (Viterbi Algorithm)	43
1.5.3	MLE Inference of Hierarchical Parameters (Baum–Welch Algorithm)	44
	Acknowledgements	46
	References	47

2	Linkage Disequilibrium, Recombination and Haplotype Structure	51
	<i>Gil McVean and Jerome Kelleher</i>	
2.1	What Is Linkage Disequilibrium?	51
2.2	Measuring Linkage Disequilibrium	53
	2.2.1 Single-Number Summaries of LD	54
	2.2.2 The Spatial Distribution of LD	56
	2.2.3 Various Extensions of Two-Locus LD Measures	60
2.3	Modelling Linkage Disequilibrium and Genealogical History	60
	2.3.1 A Historical Perspective	60
	2.3.2 Coalescent Modelling	62
	2.3.3 Relating Genealogical History to LD	67
2.4	Data Analysis	69
	2.4.1 Estimating Recombination Rates	69
	2.4.2 Methods Exploiting Haplotype Structure	72
2.5	Prospects	75
	Acknowledgements	75
	References	76
3	Haplotype Estimation and Genotype Imputation	87
	<i>Jonathan Marchini</i>	
3.1	Haplotype Estimation	87
	3.1.1 A Simple Haplotype Frequency Model	88
	3.1.2 Hidden Markov Models for Phasing	89
	3.1.3 Phasing in Related Samples	93
	3.1.4 Phasing Using Sequencing Data	94
	3.1.5 Phasing from a Reference Panel	95
	3.1.6 Measuring Phasing Performance	96
3.2	Genotype Imputation	97
	3.2.1 Uses of Imputation in GWASs	98
	3.2.2 Haploid Imputation	99
	3.2.3 Imputation Methods	100
	3.2.4 Testing Imputed Genotypes for Association	103
	3.2.5 Summary Statistic Imputation	104
	3.2.6 Factors Affecting Accuracy	104
	3.2.7 Quality Control for Imputed Data	107
3.3	Future Directions	109
	References	109
4	Mathematical Models in Population Genetics	115
	<i>Nick Barton and Alison Etheridge</i>	
4.1	Introduction	115
4.2	Single-Locus Models	116
	4.2.1 Random Drift and the Kingman Coalescent	117
	4.2.2 Diffusion Approximations	120
	4.2.3 Spatially Structured Populations	126
4.3	Multiple Loci	130
	4.3.1 Linkage Equilibrium	131
	4.3.2 Beyond Linkage Equilibrium	134
4.4	Outlook	140
	References	140

5	Coalescent Theory	145
	<i>Magnus Nordborg</i>	
5.1	Introduction	145
5.2	The Coalescent	146
5.2.1	The Fundamental Insights	146
5.2.2	The Coalescent Approximation	148
5.3	Generalizing the Coalescent	151
5.3.1	Robustness and Scaling	151
5.3.2	Variable Population Size	152
5.3.3	Population Structure on Different Time-Scales	153
5.4	Geographical Structure	155
5.4.1	The Structured Coalescent	155
5.4.2	The Strong-Migration Limit	156
5.5	Diploidy and Segregation	157
5.5.1	Hermaphrodites	157
5.5.2	Males and Females	159
5.6	Recombination	159
5.6.1	The Ancestral Recombination Graph	160
5.6.2	Properties and Effects of Recombination	163
5.7	Selection	164
5.7.1	Balancing Selection	165
5.7.2	Selective Sweeps	166
5.7.3	Background Selection	168
5.8	Neutral Mutations	168
5.9	Concluding Remarks	169
5.9.1	The Coalescent and 'Classical' Population Genetics	169
5.9.2	The Coalescent and Phylogenetics	169
5.9.3	Prospects	171
	Acknowledgements	171
	References	171
6	Phylogeny Estimation Using Likelihood-Based Methods	177
	<i>John P. Huelsenbeck</i>	
6.1	Introduction	177
6.1.1	Statistical Phylogenetics	178
6.1.2	Chapter Outline	178
6.2	Maximum Likelihood and Bayesian Estimation	179
6.2.1	Maximum Likelihood	179
6.2.2	Bayesian Inference	180
6.3	Choosing among Models Using Likelihood Ratio Tests and Bayes Factors	184
6.4	Calculating the Likelihood for a Phylogenetic Model	186
6.4.1	Character Matrices and Alignments	186
6.4.2	The Phylogenetic Model	186
6.4.3	Calculating the Probability of a Character History	187
6.4.4	Continuous-Time Markov Model	188
6.4.5	Marginalizing over Character Histories	189
6.5	The Mechanics of Maximum Likelihood and Bayesian Inference	192
6.5.1	Maximum Likelihood	192
6.5.2	Bayesian Inference and Markov Chain Monte Carlo	193

6.6	Applications of Likelihood-Based Methods in Molecular Evolution	199
6.6.1	A Taxonomy of Commonly Used Substitution Models	199
6.6.2	Expanding the Model around Groups of Sites	202
6.6.3	Rate Variation across Sites	204
6.6.4	Divergence Time Estimation	206
6.7	Conclusions	212
	References	213
7	The Multispecies Coalescent	219
	<i>Laura Kubatko</i>	
7.1	Introduction	219
7.2	Probability Distributions under the Multispecies Coalescent	221
7.2.1	Gene Tree Probabilities	221
7.2.2	Site Pattern Probabilities	227
7.2.3	Species Tree Likelihoods under the Multispecies Coalescent	229
7.2.4	Model Assumptions and Violations	230
7.3	Species Tree Inference under the Multispecies Coalescent	231
7.3.1	Summary Statistics Methods	231
7.3.2	Bayesian Full-Data Methods	234
7.3.3	Site Pattern-Based Methods	235
7.3.4	Multilocus versus SNP Data	236
7.3.5	Empirical Examples	237
7.4	Coalescent-Based Estimation of Parameters at the Population and Species Levels	239
7.4.1	Speciation Times and Population Sizes	239
7.4.2	Hybridization and Gene Flow	240
7.4.3	Species Delimitation	241
7.4.4	Future Prospects	242
	Acknowledgements	242
	References	242
8	Population Structure, Demography and Recent Admixture	247
	<i>G. Hellenthal</i>	
8.1	Introduction	247
8.1.1	'Admixture' versus 'Background' Linkage Disequilibrium	248
8.2	Spatial Summaries of Genetic Variation Using Principal Components Analysis	249
8.3	Clustering Algorithms	251
8.3.1	Defining 'Populations'	251
8.3.2	Clustering Based on Allele Frequency Patterns	252
8.3.3	Incorporating Admixture	253
8.3.4	Incorporating Admixture Linkage Disequilibrium	254
8.3.5	Incorporating Background Linkage Disequilibrium: Using Haplotypes to Improve Inference	255
8.3.6	Interpreting Genetic Clusters	258
8.4	Inferring Population Size Changes and Split Times	259
8.4.1	Allele Frequency Spectrum Approaches	260
8.4.2	Approaches Using Whole-Genome Sequencing	261
8.5	Identifying/Dating Admixture Events	262
8.5.1	Inferring DNA Segments Inherited from Different Sources	263
8.5.2	Measuring Decay of Linkage Disequilibrium	265

8.6	Conclusion	267
	Acknowledgements	268
	References	268
9	Statistical Methods to Detect Archaic Admixture and Identify Introgressed Sequences	275
	<i>Liming Li and Joshua M. Akey</i>	
9.1	Introduction	275
9.2	Methods to Test Hypotheses of Archaic Admixture and Infer Admixture Proportions	277
9.2.1	Genetic Drift and Allele Frequency Divergence in Genetically Structured Populations	277
9.2.2	Three-Population Test	277
9.2.3	D -Statistic	279
9.2.4	F_4 -Statistic	282
9.3	Methods to Identify Introgressed Sequences	283
9.3.1	S^* -Statistic	284
9.3.2	Hidden Markov and Conditional Random Field Models	287
9.3.3	Relative Advantages and Disadvantages of Approaches to Detect Introgressed Sequences	289
9.4	Summary and Perspective	289
	References	290
10	Population Genomic Analyses of DNA from Ancient Remains	295
	<i>Torsten Günther and Mattias Jakobsson</i>	
10.1	Introduction	295
10.2	Challenges of Working with and Analyzing Ancient DNA Data	296
10.2.1	Sequence Degradation	296
10.2.2	Contamination	297
10.2.3	Handling Sequence Data from Ancient Material	300
10.2.4	Different Sequencing Approaches and the Limitations in their Resulting Data	301
10.2.5	Effects of Limited Amounts of Data on Downstream Analysis	301
10.3	Opportunities of Ancient DNA	302
10.3.1	Population Differentiation in Time and Space	303
10.3.2	Continuity	306
10.3.3	Migration and Admixture over Time	307
10.3.4	Demographic Inference Based on High-Coverage Ancient Genomes	308
10.3.5	Allele Frequency Trajectories	308
10.4	Some Examples of How Genetic Studies of Ancient Remains Have Contributed to a New Understanding of the Human Past	310
10.4.1	Archaic Genomes and the Admixture with Modern Humans	310
10.4.2	Neolithic Revolution in Europe and the Bronze Age Migrations	311
10.5	Summary and Perspective	313
	Acknowledgements	313
	References	314
11	Sequence Covariation Analysis in Biological Polymers	325
	<i>William R. Taylor, Shaun Kandathil, and David T. Jones</i>	
11.1	Introduction	325

11.2	Methods	326
	11.2.1	DCA Method 326
	11.2.2	PSICOV 327
	11.2.3	plmDCA, GREMLIN and CCMpred 327
11.3	Applications	328
	11.3.1	Globular Protein Fold Prediction 328
	11.3.2	Transmembrane Protein Prediction 328
	11.3.3	RNA Structure Prediction 328
	11.3.4	Protein Disordered Regions 329
	11.3.5	Protein–Protein Interactions 329
	11.3.6	Allostery and Dynamics 330
	11.3.7	CASP 330
11.4	New Developments	332
	11.4.1	Sequence Alignment 332
	11.4.2	Comparison to Known Structures 333
	11.4.3	Segment Parsing 334
	11.4.4	Machine Learning 335
	11.4.5	Deep Learning Methods 336
	11.4.6	Sequence Pairing 338
	11.4.7	Phylogeny Constraints 339
11.5	Outlook	340
	Acknowledgements	341
	References	342
12	Probabilistic Models for the Study of Protein Evolution	347
	<i>Umberto Perron, Iain H. Moal, Jeffrey L. Thorne, and Nick Goldman</i>	
12.1	Introduction	347
12.2	Empirically Derived Models of Amino Acid Replacement	348
	12.2.1	The Dayhoff and Eck Model 348
	12.2.2	Descendants of the Dayhoff Model 350
12.3	Heterogeneity of Replacement Rates among Sites	351
12.4	Protein Structural Environments	351
12.5	Variation of Preferred Residues among Sites	353
12.6	Models with a Physicochemical Basis	355
12.7	Codon-Based Models	355
12.8	Dependence among Positions	357
12.9	Stochastic Models of Structural Evolution	359
12.10	Conclusion	360
	Acknowledgements	361
	References	361
13	Adaptive Molecular Evolution	369
	<i>Ziheng Yang</i>	
13.1	Introduction	369
13.2	Markov Model of Codon Substitution	371
13.3	Estimation of Synonymous and Non-synonymous Substitution Rates between Two Sequences and Test of Selection on the Protein	372
	13.3.1	Heuristic Estimation Methods 372
	13.3.2	Maximum Likelihood Estimation 374

13.3.3	Bayesian Estimation	377
13.3.4	A Numerical Example	377
13.4	Likelihood Calculation on a Phylogeny	379
13.5	Detecting Adaptive Evolution along Lineages	380
13.5.1	Likelihood Calculation under Models of Variable ω Ratios among Lineages	380
13.5.2	Adaptive Evolution in the Primate Lysozyme	381
13.5.3	Comparison with Methods Based on Reconstructed Ancestral Sequences	382
13.6	Inferring Amino Acid Sites under Positive Selection	384
13.6.1	Likelihood Ratio Test under Models of Variable ω Ratios among Sites	384
13.6.2	Methods that Test One Site at a Time	386
13.6.3	Positive Selection in the HIV-1 <i>vif</i> Genes	386
13.7	Testing Positive Selection Affecting Particular Sites and Lineages	388
13.7.1	Branch-Site Test of Positive Selection	388
13.7.2	Clade Models and Other Variants	389
13.8	Limitations of Current Methods	390
13.9	Computer Software	391
	References	391
14	Detecting Natural Selection	397
	<i>Aaron J. Stern and Rasmus Nielsen</i>	
14.1	Introduction	397
14.2	Types of Selection	398
14.2.1	Directional Selection	398
14.2.2	Balancing Selection	399
14.2.3	Polygenic Selection	399
14.3	The Signature of Selection in the Genome	399
14.3.1	The Signature of Positive Directional Selection	400
14.3.2	Balancing Selection	403
14.3.3	Polygenic Selection	403
14.3.4	Confounders	404
14.4	Methods for Detecting Selection	405
14.4.1	Substitution-Based Methods	405
14.4.2	Methods Comparing Substitutions and Diversity	406
14.4.3	Methods Using the Frequency Spectrum	407
14.4.4	Methods Using Genetic Differentiation	408
14.4.5	Methods Using Haplotype Structure	410
14.4.6	Why Full-Likelihood Methods Are Intractable for Population Samples	412
14.4.7	Composite Likelihood Methods	412
14.4.8	Approximate Bayesian Computation	413
14.4.9	Machine Learning Methods	413
14.5	Discussion	414
	References	415
15	Evolutionary Quantitative Genetics	421
	<i>Bruce Walsh and Michael B. Morrissey</i>	
15.1	Introduction	421

15.2	Resemblances, Variances, and Additive Genetic Values	422
15.2.1	Fisher's Genetic Decomposition	422
15.2.2	Additive Genetic Variances and Covariances	423
15.3	Parent–Offspring Regressions and the Response to Selection	423
15.3.1	Single-Trait Parent–Offspring Regressions	424
15.3.2	Selection Differentials and the Breeder's Equation	424
15.3.3	Multiple-Trait Parent–Offspring Regressions	425
15.3.4	The Genetic and Phenotypic Covariance Matrices	425
15.3.5	The Multivariate Breeder's Equation	425
15.4	The Infinitesimal Model	426
15.4.1	Linearity of Parent–Offspring Regressions under the Infinitesimal Model	426
15.4.2	Allele Frequency Changes under the Infinitesimal Model	426
15.4.3	Changes in Variances	427
15.4.4	The Equilibrium Additive Genetic Variance	429
15.5	Inference of σ_A^2 and \mathbf{G}	430
15.6	Fitness	432
15.6.1	Individual Fitness	432
15.6.2	Episodes of Selection	433
15.7	The Robertson–Price Identity, and Theorems of Selection	434
15.7.1	Description of the Theorems	435
15.7.2	Empirical Operationalization of the Theorems	436
15.8	The Opportunity for Selection	437
15.9	Selection Coefficients	438
15.9.1	Measures of Selection on the Mean	439
15.9.2	Measures of Selection on the Variance	439
15.10	Fitness Functions and the Characterization of Selection	441
15.10.1	Individual and Mean Fitness Functions	441
15.10.2	Gradients and the Local Geometry of Fitness Surfaces	442
15.11	Multivariate Selection	444
15.11.1	Short-Term Changes in Means: The Multivariate Breeder's Equation	444
15.11.2	The Effects of Genetic Correlations: Direct and Correlated Responses	444
15.11.3	Selection Gradients and Understanding which Traits Affect Fitness	446
15.12	Inference of Selection Gradients	447
15.12.1	Ordinary Least Squares Analysis	448
15.12.2	Flexible Inference of Fitness Functions with Associated Selection Gradient Estimates	449
15.12.3	Normality and Selection Gradients	450
15.13	Summary	451
	References	452
16	Conservation Genetics	457
	<i>Mark Beaumont and Jinliang Wang</i>	
16.1	Introduction	457
16.2	Estimating Effective Population Size	458
16.2.1	Methods Based on Heterozygosity Excess	459
16.2.2	Methods Based on Linkage Disequilibrium	460
16.2.3	Methods Based on Relatedness	463
16.2.4	Methods Based on Temporal Changes in Allele Frequency	466

16.3	Estimating Census Size by the Genotype Capture–Recapture Approach	470
16.3.1	Methods Based on Multilocus Genotype Mismatches	472
16.3.2	Methods Based on Pairwise Relatedness	472
16.3.3	Methods Based on Pairwise Relationships	473
16.3.4	Methods Based on Pedigree Reconstruction	474
16.4	Inferring Genetic Structure	475
16.4.1	Measuring Genetic Differentiation	475
16.4.2	Population Assignment	477
16.4.3	Population Clustering and Inference of Ancestry Proportions	479
16.4.4	Inferring Levels of Recent Gene Flow	483
16.4.5	Landscape Genetics	486
16.5	Deintrogression Strategies	487
16.6	Genetic Species Delimitation	489
16.7	Conclusions and Outlook	491
	Acknowledgements	492
	References	492
17	Statistical Methods for Plant Breeding	501
	<i>Ian Mackay, Hans-Peter Piepho, and Antonio Augusto Franco Garcia</i>	
17.1	Introduction	501
17.2	Heritability and the Breeder’s Equation in Plant Breeding	502
17.3	The Breeding System of Plants	504
17.4	Polyploidy in Plants and Its Genetic Consequences	505
17.5	Genomic Rearrangements in Plants	509
17.6	Genetic Architecture of Traits in Plants	510
17.7	Response to the Environment and Plasticity	511
17.8	Genomic Selection	514
17.8.1	Genotype–Environment Interaction	514
17.8.2	Quantitative Trait Loci and Major Genes	516
17.8.3	Genomic Selection and Cross Prediction	517
17.8.4	Genomic Selection and Phenotyping Cost	517
17.8.5	Mate Selection	517
17.8.6	Sequential Selection	518
17.8.7	Genomic Prediction of Hybrid Performance and Heterosis	518
17.8.8	Marker Imputation	519
17.9	Experimental Design and Analysis	519
17.10	Conclusions	521
	References	521
18	Forensic Genetics	531
	<i>B.S. Weir</i>	
18.1	Introduction	531
18.2	Principles of Interpretation	532
18.3	Profile Probabilities	534
18.3.1	Genetic Models for Allele Frequencies	535
18.3.2	Y-STR Profiles	539
18.4	Mixtures	542
18.4.1	Combined Probabilities of Inclusion and Exclusion	542
18.4.2	Likelihood Ratios	542
18.5	Behavior of Likelihood Ratio	546

- 18.6 Single Nucleotide Polymorphism, Sequence and Omic Data 547
References 548

Volume 2

- List of Contributors** xxiii
Editors' Preface to the Fourth Edition xxvii
Glossary xxix
Abbreviations and Acronyms xxxix

- 19 Ethical Issues in Statistical Genetics** 551
Susan E. Wallace and Richard Ashcroft
- 19.1 Introduction 551
- 19.1.1 What Is Ethics? 552
- 19.1.2 Models for Analysing the Ethics of Population Genetic Research 553
- 19.2 Ethics and Governance in Population Genetics Research: Two Case Studies 554
- 19.2.1 'Healthy Volunteer' Longitudinal Cohort Studies: UK Biobank 555
- 19.2.2 Precision Medicine Approaches: 100,000 Genomes Project 556
- 19.2.3 The Scientific and Clinical Value of the Research 556
- 19.2.4 Recruitment of Participants 558
- 19.2.5 Consent 559
- 19.2.6 Returning Individual Genetic Research Results 563
- 19.2.7 Confidentiality and Security 564
- 19.3 Stewardship and Wider Social Issues 565
- 19.3.1 Benefit Sharing 566
- 19.3.2 Community Involvement and Public Engagement 567
- 19.3.3 Race, Ethnicity and Genetics 567
- 19.4 Conclusion 568
Acknowledgements 568
References 568
- 20 Descent-Based Gene Mapping in Pedigrees and Populations** 573
E.A. Thompson
- 20.1 Introduction to Genetic Mapping and Genome Descent 573
- 20.1.1 Genetic Mapping: The Goal and the Data 573
- 20.1.2 The Process of Meiosis and the Descent of DNA 574
- 20.1.3 Genetic Linkage Mapping: Association or Descent? 576
- 20.2 Inference of Local IBD Sharing from Genetic Marker Data 577
- 20.2.1 Identity by Descent at a Locus 577
- 20.2.2 Probabilities of Marker Data Given IBD Pattern 579
- 20.2.3 Modeling the Probabilities of Patterns of IBD 580
- 20.2.4 Inferring Local IBD from Marker Data 581
- 20.3 IBD-Based Detection of Associations between Markers and Traits 583
- 20.3.1 Trait Data Probabilities for Major Gene Models 583
- 20.3.2 Quantitative Trait Data Probabilities under Random Effects Models 584
- 20.3.3 IBD-Based Linkage Likelihoods for Major Gene Models 585
- 20.3.4 IBD-Based Linkage Likelihoods for Random-Effects Models 587
- 20.4 Other Forms of IBD-Based Genetic Mapping 589

20.4.1	IBD-Based Case–Control Studies	589
20.4.2	Patterns of IBD in Affected Relatives	590
20.5	Summary	592
	Acknowledgements	592
	References	593
21	Genome-Wide Association Studies	597
	<i>Andrew P. Morris and Lon R. Cardon</i>	
21.1	Introduction	597
21.2	GWAS Design Concepts	599
21.2.1	Phenotype Definition	599
21.2.2	Structure of Common Genetic Variation and Design of GWAS Genotyping Technology	599
21.2.3	Sample Size Considerations	601
21.2.4	Genome-Wide Significance and Correction for Multiple Testing	601
21.2.5	Replication	602
21.3	GWAS Quality Control	602
21.3.1	SNP Quality Control Procedures	603
21.3.2	Sample Quality Control Procedures	604
21.3.3	Software	606
21.4	Single SNP Association Analysis	606
21.4.1	Generalised Linear Modelling Framework	606
21.4.2	Accounting for Confounding Factors as Covariates	606
21.4.3	Coding of SNP Genotypes	607
21.4.4	Imputed Genotypes	609
21.4.5	Visualisation of Results of Single SNP GWAS Analyses	609
21.4.6	Interactions with Non-Genetic Risk Factors	609
21.4.7	Bayesian Methods	611
21.4.8	Software	611
21.5	Detecting and Accounting for Genetic Structure in GWASs	611
21.5.1	Identification of Related Individuals	612
21.5.2	Multivariate Approaches to Identify Ethnic Outliers and Account for Population Stratification	613
21.5.3	Mixed Modelling Approaches to Account for Genetic Structure	614
21.5.4	Software	615
21.6	Multiple SNP Association Analysis	616
21.6.1	Haplotype-Based Analyses	616
21.6.2	SNP–SNP Interaction Analyses	617
21.6.3	Gene-Based Analyses	619
21.6.4	Software	619
21.7	Discussion	620
	References	623
22	Replication and Meta-analysis of Genome-Wide Association Studies	631
	<i>Frank Dudbridge and Paul Newcombe</i>	
22.1	Introduction	631
22.2	Replication	632
22.2.1	Motivation	632
22.2.2	Different Forms of Replication	632

22.2.3	Two-Stage Genome-Wide Association Studies	634
22.2.4	Significance Thresholds for Replication	634
22.2.5	A Key Challenge: Heterogeneity	635
22.3	Winner's Curse	635
22.3.1	Description of the Problem	635
22.3.2	Methods for Correcting for Winner's Curse	637
22.3.3	Applicability of These Methods	639
22.4	Meta-analysis	640
22.4.1	Motivation	640
22.4.2	An Illustrative Example	640
22.4.3	Fixed Effect Meta-analysis	641
22.4.4	Chi-Square Test for Heterogeneity in Effect	641
22.4.5	Random Effects Meta-analysis	642
22.4.6	Interpretation and Significance Testing of Meta-analysis Estimates	643
22.4.7	Using Funnel Plots to Investigate Small Study Bias in Meta-analysis	644
22.4.8	Improving Analyses via Meta-analysis Consortia and Publicly Available Data	645
22.5	Summary	647
	References	647
23	Inferring Causal Relationships between Risk Factors and Outcomes Using Genetic Variation	651
	<i>Stephen Burgess, Christopher N. Foley, and Verena Zuber</i>	
23.1	Background	651
23.1.1	Correlation and Causation	651
23.1.2	Chapter Outline	652
23.2	Introduction to Mendelian Randomization and Motivating Example	652
23.2.1	Instrumental Variable Assumptions	653
23.2.2	Assessing the Instrumental Variable Assumptions	654
23.2.3	Two-Sample Mendelian Randomization and Summarized Data	655
23.3	Monogenic Mendelian Randomization Analyses: The Easy Case	655
23.4	Polygenic Mendelian Randomization Analyses: The Difficult Case	656
23.4.1	Example: Low-Density Lipoprotein Cholesterol and Coronary Heart Disease Risk	656
23.4.2	More Complex Examples	657
23.4.3	Two-Stage Least Squares and Inverse-Variance Weighted Methods	659
23.5	Robust Approaches for Polygenic Mendelian Randomization Analyses	660
23.5.1	Median Estimation Methods	660
23.5.2	Modal Estimation Methods	660
23.5.3	Regularization Methods	660
23.5.4	Other Outlier-Robust Methods	661
23.5.5	MR-Egger Method	661
23.5.6	Multivariable Methods	663
23.5.7	Interactions and Subsetting	663
23.5.8	Practical Advice	664
23.6	Alternative Approaches for Causal Inference with Genetic Data	665
23.6.1	Fine-Mapping and Colocalization	665
23.6.2	LD Score Regression	666
23.7	Causal Estimation in Mendelian Randomization	667

23.7.1	Relevance of Causal Estimate	667
23.7.2	Heterogeneity and Pleiotropy	668
23.7.3	Weak Instrument Bias and Sample Overlap	668
23.7.4	Time-Dependent Causal Effects	669
23.7.5	Collider Bias	670
23.8	Conclusion	670
	References	671
24	Improving Genetic Association Analysis through Integration of Functional Annotations of the Human Genome	679
	<i>Qiongshi Lu and Hongyu Zhao</i>	
24.1	Introduction	679
24.2	Types of Functional Annotation Data in GWAS Applications	680
24.2.1	Transcriptomic Annotation Data	680
24.2.2	Epigenetic Annotation Data	681
24.2.3	DNA Conservation	681
24.3	Methods to Synthesize Annotation Data	682
24.3.1	Genome Browsers and Annotator Software	682
24.3.2	Supervised Learning Methods	682
24.3.3	Unsupervised Learning Methods	684
24.3.4	Improving Specificity of Computational Annotations	685
24.4	Methods to Integrate Functional Annotations in Genetic Association Analysis	685
24.4.1	Partitioning Heritability and Genetic Covariance	685
24.4.2	Imputation-Based Gene-Level Association Analysis	688
24.4.3	Other Applications and Future Directions	690
	Acknowledgements	690
	References	691
25	Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci	697
	<i>Solveig K. Sieberts and Eric E. Schadt</i>	
25.1	Introduction	697
25.1.1	An Overview of Transcription as a Complex Process	700
25.1.2	Modeling Approaches for Biological Processes	702
25.1.3	Human versus Experimental Models	704
25.2	Modeling for eQTL Detection and Causal Inference	705
25.2.1	Heritability of Expression Traits	705
25.2.2	Single-Trait eQTL Mapping	706
25.2.3	Joint eQTL Mapping	706
25.2.4	eQTL and Clinical Trait Linkage Mapping to Infer Causal Associations	708
25.3	Inferring Gene Regulatory Networks	714
25.3.1	From Assessing Causal Relationships among Trait Pairs to Predictive Gene Networks	714
25.3.2	Building from the Bottom Up or Top Down?	714
25.3.3	Using eQTL Data to Reconstruct Coexpression Networks	715
25.3.4	An Integrative Genomics Approach to Constructing Predictive Network Models	718

25.3.5	Integrating Genetic Data as a Structure Prior to Enhance Causal Inference in the Bayesian Network Reconstruction Process	720
25.3.6	Incorporating Other Omics Data as Network Priors in the Bayesian Network Reconstruction Process	721
25.3.7	Illustrating the Construction of Predictive Bayesian Networks with an Example	722
25.4	Conclusions	723
25.5	Software	724
	References	725
26	Statistical Methods for Single-Cell RNA-Sequencing	735
	<i>Tallulah S. Andrews, Vladimir Yu. Kiselev, and Martin Hemberg</i>	
26.1	Introduction	735
26.2	Overview of scRNA-Seq Experimental Platforms and Low-Level Analysis	736
26.2.1	Low-Throughput Methods	736
26.2.2	High-Throughput Methods	737
26.2.3	Computational Analysis	739
26.3	Novel Statistical Challenges Posed by scRNA-Seq	739
26.3.1	Estimating Transcript Levels	739
26.3.2	Analysis of the Expression Matrix	747
	References	753
27	Variant Interpretation and Genomic Medicine	761
	<i>K. Carss, D. Goldstein, V. Aggarwal, and S. Petrovski</i>	
27.1	Introduction and Current Challenges	761
27.2	Understanding the Effect of a Variant	765
27.3	Understanding Genomic Variation Context through Large Human Reference Cohorts	771
27.4	Functional Assays of Genetic Variation	777
27.5	Leveraging Existing Information about Gene Function Including Human and Model Phenotype Resources	779
27.6	Holistic Variant Interpretation	782
27.7	Future Challenges and Closing Remarks	783
27.8	Web Resources	786
	References	788
28	Prediction of Phenotype from DNA Variants	799
	<i>M.E. Goddard, T.H.E. Meuwissen, and H.D. Daetwyler</i>	
28.1	Introduction	799
28.2	Genetic Variation Affecting Phenotype	800
28.3	Data on DNA Polymorphisms Used for Prediction of Genetic Effects	801
28.4	Prediction of Additive Genetic Values	802
28.4.1	An Equivalent Model	804
28.4.2	Single-Step BLUP	804
28.4.3	Multiple Traits	805
28.4.4	Gene Expression	806
28.4.5	Using External Information	806
28.5	Factors Affecting Accuracy of Prediction	806
28.6	Other Uses of the Bayesian Genomic Selection Models	808

28.7	Examples of Genomic Prediction	809
28.7.1	Cattle	809
28.7.2	Humans	809
28.8	Conclusions	810
	References	810
29	Disease Risk Models	815
	<i>Allison Meisner and Nilanjan Chatterjee</i>	
29.1	Introduction and Background	815
29.1.1	Disease Risk Models and Their Applications	815
29.1.2	Examples of Available Disease Risk Models	817
29.1.3	Incorporating Genetic Factors	817
29.2	Absolute Risk Model	818
29.2.1	General Software for Building Absolute Risk Models	820
29.3	Building a Polygenic Risk Score	821
29.3.1	Expected Performance	821
29.3.2	Standard Approach to Constructing a PRS: LD Clumping and p -Value Thresholding	823
29.3.3	Advanced Approaches to Constructing a PRS	825
29.4	Combining PRS and Epidemiologic Factors	826
29.5	Model Validation	827
29.6	Evaluating Clinical Utility	828
29.7	Example: Breast Cancer	829
29.8	Discussion	832
29.8.1	Future Directions	832
29.8.2	Challenges	832
	References	833
30	Bayesian Methods for Gene Expression Analysis	843
	<i>Alex Lewin, Leonardo Bottolo, and Sylvia Richardson</i>	
30.1	Introduction	843
30.2	Modelling Microarray Data	845
30.2.1	Modelling Intensities	845
30.2.2	Gene Variability	845
30.2.3	Normalization	846
30.3	Modelling RNA-Sequencing Reads	846
30.3.1	Alignments for RNA-Sequencing Data	846
30.3.2	Likelihood for Read-Level Data	847
30.3.3	Likelihood for Transcript-Level Read Counts	848
30.3.4	Likelihood for Gene-Level Read Counts	850
30.4	Priors for Differential Expression Analysis	851
30.4.1	Differential Expression from Microarray Data	851
30.4.2	Differential Expression from RNA-Sequencing Data	856
30.5	Multivariate Gene Selection Models	857
30.5.1	Variable Selection Approach	857
30.5.2	Bayesian Shrinkage with Sparsity Priors	861
30.6	Quantitative Trait Loci	863
30.6.1	Single-Response Models	863
30.6.2	Multiple-Response Models	864

Acknowledgements 868
References 869

- 31 Modelling Gene Expression Dynamics with Gaussian Process Inference 879**
Magnus Rattray, Jing Yang, Sumon Ahmed, and Alexis Boukouvalas
- 31.1 Introduction 879
- 31.1.1 Covariance Function 880
- 31.1.2 Inference 882
- 31.2 Applications to Bulk Time Series Expression Data 883
- 31.2.1 Identifying Differential Expression in Time 884
- 31.2.2 Identifying Changes between Two Time Course Experiments 885
- 31.2.3 Hierarchical Models of Replicates and Clusters 887
- 31.2.4 Differential Equation Models of Production and Degradation 888
- 31.3 Modelling Single-Cell Data 889
- 31.3.1 Modelling Single-Cell Trajectory Data 889
- 31.3.2 Dimensionality Reduction and Pseudotime Inference 890
- 31.3.3 Modelling Branching Dynamics with Single-Cell RNA-Sequencing Data 892
- 31.4 Conclusion 893
- Acknowledgements 894
References 894
- 32 Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models 899**
Marco Grzegorzczuk and Dirk Husmeier
- 32.1 Introduction 899
- 32.2 Methodology 901
- 32.2.1 Dynamic Bayesian Networks (DBN) 901
- 32.2.2 Bayesian Linear Regression 902
- 32.2.3 Bayesian Piecewise Linear Regression (NH-DBN) 905
- 32.2.4 Bayesian Piecewise Linear Regression with Coupled Regression Coefficients (Coupled NH-DBNs) 908
- 32.2.5 NH-DBNs with More Flexible Allocation Schemes 915
- 32.2.6 NH-DBNs with Time-Varying Network Structures 916
- 32.2.7 Dynamic Bayesian Network Modelling 918
- 32.2.8 Computational Complexity 920
- 32.3 Application Examples 921
- 32.3.1 Morphogenesis in *Drosophila* 921
- 32.3.2 Synthetic Biology in Yeast 922
- 32.4 Summary 927
- Appendix A: Coupling Schemes 927
- A.1 Hard Information Coupling Based on an Exponential Prior 928
- A.2 Hard Information Coupling Based on a Binomial Prior 928
- A.3 Soft Information Coupling Based on a Binomial Prior 929
- References 929
- 33 DNA Methylation 933**
Kasper D. Hansen, Kimberly D. Siegmund, and Shili Lin
- 33.1 A Brief Introduction 933
- 33.2 Measuring DNA Methylation 934

33.3	Differential DNA Methylation	936
33.3.1	Differential Methylation with Bisulfite-Sequencing Data	936
33.3.2	Differential Methylation with Capture-Sequence Data	939
33.3.3	Differential Methylation with HumanMethylation Array Data	940
33.4	Other Topics of Interest	941
	References	942
34	Statistical Methods in Metabolomics	949
	<i>Timothy M.D. Ebbels, Maria De Iorio, and David A. Stephens</i>	
34.1	Introduction	949
34.2	Preprocessing and Deconvolution	950
34.2.1	Nuclear Magnetic Resonance Spectroscopy	950
34.2.2	Liquid Chromatography – Mass Spectrometry	952
34.3	Univariate Methods	954
34.3.1	Metabolome-Wide Significance Levels	956
34.3.2	Sample Size and Power	957
34.4	Multivariate Methods and Chemometrics Techniques	958
34.4.1	Linear Regression Methods	959
34.4.2	Shrinkage Methods	960
34.5	Orthogonal Projection Methods	961
34.5.1	Principal Components Analysis	962
34.5.2	Partial Least Squares	964
34.5.3	Orthogonal Projection onto Latent Structures	965
34.6	Network Analysis	966
34.7	Metabolite Identification and Pathway Analysis	969
34.7.1	Statistical Correlation Spectroscopy	969
34.7.2	Pathway and Metabolite Set Analysis	971
34.8	Conclusion	972
	References	972
35	Statistical and Computational Methods in Microbiome and Metagenomics	977
	<i>Hongzhe Li</i>	
35.1	Microbiome in Human Health and Disease	977
35.2	Estimation of Microbiome Features from 16S rRNA and Shotgun Metagenomic Sequencing Data	980
35.2.1	Estimation of Microbiome Features in 16S rRNA Data	980
35.2.2	Estimation of Microbial Composition in Shotgun Metagenomic Data	981
35.2.3	Estimation of Microbial Gene/Pathway Abundance in Shotgun Metagenomic Data	982
35.2.4	Quantification of Bacterial Growth Dynamics	982
35.2.5	Microbial Diversity Index	983
35.3	Methods for Analysis of Microbiome as an Outcome of an Intervention or Exposure	983
35.3.1	Modeling Multivariate Sparse Count Data as the Response Variable	984
35.3.2	Modeling High-Dimensional Compositional Response Data in Microbiome Studies	984
35.4	Methods for Analysis of Microbiome as a Covariate	985
35.4.1	Regression Analysis with Compositional Covariates	985
35.4.2	Kernel-Based Regression in Microbiome Studies	986

35.5	Methods for Analysis of Microbiome as a Mediator	987
35.6	Integrative Analysis of Microbiome, Small Molecules and Metabolomics Data	989
35.6.1	Computational Analysis of Small Molecules from the Human Microbiota	989
35.6.2	Metabolic Modeling in Microbiome	990
35.7	Discussion and Future Directions	991
	Acknowledgements	991
	References	992
36	Bacterial Population Genomics	997
	<i>Jukka Corander, Nicholas J. Croucher, Simon R. Harris, John A. Lees, and Gerry Tonkin-Hill</i>	
36.1	Introduction	997
36.2	Genetic Population Structure and Clustering of Genotypes	998
36.2.1	Background	998
36.2.2	Model-Based Clustering	998
36.2.3	Linkage Disequilibrium	1000
36.2.4	Distance-Based Methods	1000
36.3	Phylogenetics and Dating Analysis	1001
36.4	Transmission Modeling	1004
36.4.1	Challenges	1004
36.5	Genome-Wide Association Studies in Bacteria	1008
36.5.1	Background	1008
36.5.2	Phylogenetic Methods	1009
36.5.3	Regression-Based Methods	1011
36.6	Genome-Wide Epistasis Analysis	1012
36.7	Gene Content Analysis	1013
	References	1014
	Reference Author Index	1021
	Subject Index	1109

List of Contributors

V. Aggarwal

Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Sumon Ahmed

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Joshua M. Akey

Department of Ecology and Evolutionary Biology and Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA

Tallulah S. Andrews

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Richard Ashcroft

School of Law, Queen Mary University of London, London, UK

Nick Barton

Institute of Science and Technology Austria, Klosterneuburg, Austria

Mark Beaumont

School of Biological Sciences, Bristol University, Bristol, UK

Leonardo Bottolo

Department of Medical Genetics, University of Cambridge, Cambridge, UK, The Alan Turing Institute, London, UK, and MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Alexis Boukouvalas

Prowler.io, Cambridge, UK

Stephen Burgess

MRC Biostatistics Unit, University of Cambridge and Cardiovascular Epidemiology Unit, University of Cambridge, UK

Lon R. Cardon

BioMarin Pharmaceutical, Novato, CA, USA

K. Carss

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Nilanjan Chatterjee

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Jukka Corander

Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, Department of Biostatistics, University of Oslo, Oslo, Norway, and Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Nicholas J. Croucher

Department of Infectious Disease
Epidemiology, Imperial College London,
London, UK

H.D. Daetwyler

Agriculture Victoria, AgriBio, Bundoora,
Victoria, Australia, and School of Applied
Systems Biology, La Trobe University,
Bundoora, Victoria, Australia

Maria De Iorio

Department of Statistical Science,
University College London, London, UK

Frank Dudbridge

Department of Health Sciences, University
of Leicester, Leicester, UK

Timothy M.D. Ebbels

Computational and Systems Medicine,
Department of Surgery and Cancer,
Imperial College London, London, UK

Alison Etheridge

University of Oxford, UK

Christopher N. Foley

MRC Biostatistics Unit, University of
Cambridge, UK

Antonio Augusto Franco Garcia

University of São Paulo, Piracicaba, Brazil

M.E. Goddard

Faculty of Veterinary and Agricultural
Sciences, University of Melbourne,
Parkville, Victoria, Australia, and
Agriculture Victoria, AgriBio, Bundoora,
Victoria, Australia

Nick Goldman

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

D. Goldstein

Institute for Genomic Medicine, Columbia
University Medical Center, New York, USA

Marco Grzegorzcyk

Bernoulli Institute (BI), Faculty of Science
and Engineering, Rijksuniversiteit
Groningen, Groningen, Netherlands

Torsten Günther

Human Evolution, Department of
Organismal Biology, Uppsala University,
Sweden

Kasper D. Hansen

Department of Biostatistics, Johns Hopkins
Bloomberg School of Public Health, and
McKusick-Nathans Institute of Genetic
Medicine, Johns Hopkins School of
Medicine, Johns Hopkins University,
Baltimore, MD, USA

Simon R. Harris

Infection Genomics, Wellcome Sanger
Institute, Hinxton, Cambridgeshire, UK

Garrett Hellenthal

University College London Genetics
Institute (UGI), Department of Genetics,
Evolution and Environment, University
College London, London, UK

Martin Hemberg

Wellcome Sanger Institute, Hinxton,
Cambridgeshire, UK

John P. Huelsenbeck

Department of Integrative Biology,
University of California, Berkeley, CA, USA

Dirk Husmeier

School of Mathematics & Statistics,
University of Glasgow, Glasgow, UK

Mattias Jakobsson

Human Evolution, Department of
Organismal Biology, Uppsala University,
Sweden

David T. Jones

University College London, London, UK

Shaun Kandathil

University College London, London, UK

Jerome Kelleher

University of Oxford, UK

Vladimir Yu. Kiselev

Wellcome Sanger Institute, Hinxton,
Cambridgeshire, UK

Laura Kubatko

Ohio State University, Columbus, OH, USA

John A. Lees

Department of Microbiology, School of
Medicine, New York University, New York,
USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Christoph Leuenberger

University of Fribourg, Switzerland

Hongzhe Li

Department of Biostatistics, Epidemiology
and Informatics, Perelman School of
Medicine, University of Pennsylvania,
Philadelphia, USA

Liming Li

Department of Ecology and Evolutionary
Biology, Princeton University, Princeton,
NJ, USA

Shili Lin

Department of Statistics, Ohio State
University, Columbus, OH, USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Qiongshi Lu

Department of Biostatistics and Medical
Informatics, University of Madison-
Wisconsin, Madison, WI, USA

Jonathan Marchini

Regeneron Genetics Center, Tarrytown,
NY, USA

Gil McVean

University of Oxford, UK

Allison Meisner

Department of Biostatistics, Johns Hopkins
Bloomberg School of Public Health,
Baltimore, MD, USA

Ian Mackay

IMplant Consultancy Ltd, Chelmsford, UK

T.H.E. Meuwissen

Norwegian University of Life Sciences,
Ås, Norway

Iain H. Moal

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

Andrew P. Morris

Department of Biostatistics, University of
Liverpool, Liverpool, UK

Michael B. Morrissey

School of Biology, University of St Andrews,
St Andrews, UK

Paul Newcombe

MRC Biostatistics Unit, University of
Cambridge, Cambridge, UK

Rasmus Nielsen

Department of Integrative Biology and
Department of Statistics, University of
California, Berkeley, CA, USA

Magnus Nordborg

Gregor Mendel Institute, Austrian
Academy of Sciences, Vienna BioCenter,
Vienna, Austria

Umberto Perron

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

S. Petrovski

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Hans-Peter Piepho

University of Hohenheim, Stuttgart, Germany

Magnus Rattray

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Sylvia Richardson

MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, and The Alan Turing Institute, London, UK

Eric E. Schadt

Sema4, Stamford, CT, USA, and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Solveig K. Sieberts

Sage Bionetworks, Seattle, WA, USA

Kimberly D. Siegmund

Department of Preventive Medicine, Keck School of Medicine of USC, Los Angeles, USA

David A. Stephens

Department of Mathematics and Statistics, McGill University, Montreal, Canada

Aaron J. Stern

Graduate Group in Computational Biology, University of California, Berkeley, CA

William R. Taylor

The Francis Crick Institute, London, UK

E.A. Thompson

Department of Statistics, University of Washington, Seattle, WA, USA

Jeffrey L. Thorne

Department of Statistics & Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Gerry Tonkin-Hill

Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Susan E. Wallace

Department of Health Sciences, University of Leicester, Leicester, UK

Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Jinliang Wang

Institute of Zoology, Zoological Society of London, UK

Daniel Wegmann

University of Fribourg, Switzerland

B.S. Weir

Department of Biostatistics, University of Washington, Seattle, WA, USA

Jing Yang

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Ziheng Yang

Department of Genetics, Evolution and Environment, University College London, London, UK

Hongyu Zhao

Department of Biostatistics, Yale University, New Haven, CT, USA

Verena Zuber

MRC Biostatistics Unit, University of Cambridge, UK and School of Public Health, Imperial College London

Editors' Preface to the Fourth Edition

After a break of more than 10 years since the third edition, we are pleased to present the fourth edition of the *Handbook of Statistical Genomics*. Genomics has moved on enormously during this period, and so has the *Handbook*: almost everything is new or much revised, with only a small amount of material carried forward from previous editions. Two new editors have joined, Ida Moltke from Copenhagen and John Marioni from Cambridge. With sadness we note the death of founding editor Professor Chris Cannings during 2018. He first saw the need and had the vision for the *Handbook*, one of his many contributions to mathematical and statistical genetics. We also acknowledge the fundamental contribution of the other founding editor, Professor Martin Bishop.

While the content has changed, the mission has not: the *Handbook* is intended as an introduction suitable for advanced graduate students and early-career researchers who have achieved at least a good first-year undergraduate level in both statistics and genetics, and preferably more in at least one of those fields. The chapters are not thorough literature reviews, but focus on explaining the key ideas, methods and algorithms, citing key recent and historic literature for further details and references.

The change of title (from *Genetics* to *Genomics*) is not intended to indicate a substantial change of focus, but reflects both changes in the field, with increased emphasis on transcriptomics and epigenetics for example, and changes in usage. We interpret 'genomics' broadly, to include studies of whole genomes and epigenomes, near-genome processes such as transcription and metabolomics, as well as genomic mechanisms underlying whole-organism outcomes related to selection, adaptation and disease. We also interpret 'statistics' broadly, to include for example relevant aspects of data science and bioinformatics.

The 36 chapters are intended to be largely independent, so that to benefit from the *Handbook* it is not necessary to read every chapter, or to read chapters in order. This structure necessitates some duplication of material, which we have tried to minimize but not always eliminate. Alternative approaches to the same topic by different authors can be beneficial. The extensive subject and author indexes allow easy reference to topics arising in different chapters.

For those with minimal genetics background the glossary has been newly updated. Thanks to Keren Carss for contributing many new terms, in particular those relevant to genomic medicine. Gerry Tonkin-Hill and John Lees also contributed some advanced statistical terminology, but the glossary is predominantly of genetic terms. For those with limited background in statistical modeling and inference we have added an initial chapter that covers these topics, ranging from basic concepts to state-of-art models and methods.

We thank the many commentators on previous editions who were generous in their praise and helpful feedback. No doubt many more improvements will be possible for future editions and we welcome comments e-mailed to any of the editors. We are grateful to all of our authors for taking the time to write and update their chapters with care, and we would like to express our appreciation to all the professional staff working with and for Wiley who helped us to bring this project to fruition.

