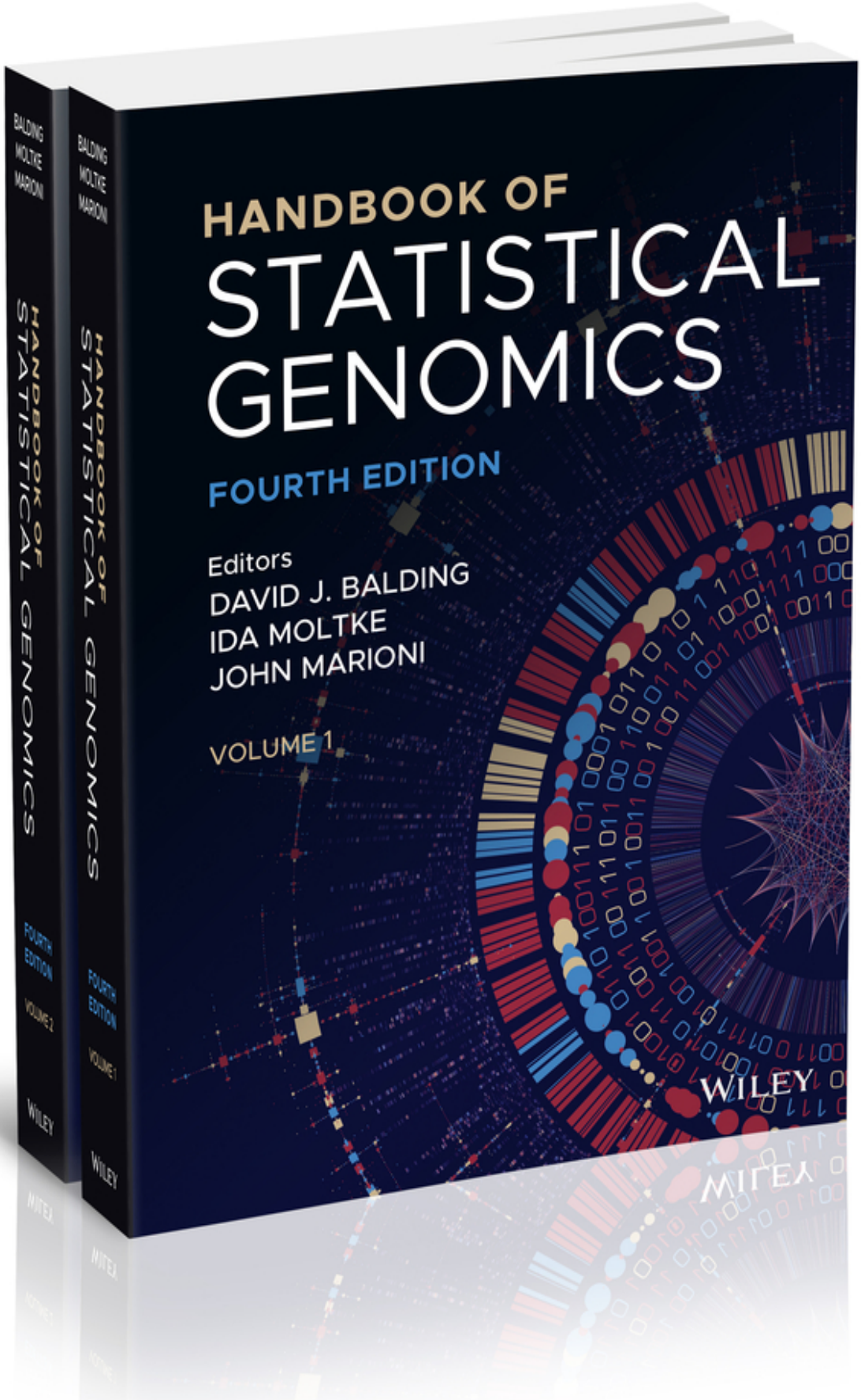
HANDBOOK OF
STATISTICAL
GENOMICS
FOURTH EDITION
Editors
DAVID J. BALDING
IDA MOLTKE
JOHN MARIONI
VOLUME 1
WILEY
BALDING
MOLTKE
MARIONI
HANDBOOK OF
STATISTICAL GENOMICS
FOURTH EDITION
VOLUME 2
WILEY

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

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

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.

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://lccn.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

Volume 2

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 Grzegorczyk
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.