

Cambridge University Press

978-1-107-02752-7 - *Advances in Statistical Bioinformatics Models and Integrative Inference for High-Throughput Data*

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

ADVANCES IN STATISTICAL BIOINFORMATICS

Providing genome-informed personalized treatment is a goal of modern medicine. Identifying new translational targets in nucleic acid characterizations is an important step toward that goal. The information tsunami produced by such genome-scale investigations is stimulating parallel developments in statistical methodology and inference, analytical frameworks, and computational tools.

Within the context of genomic medicine and with a strong focus on cancer research, this book describes the integration of high-throughput bioinformatics data from multiple platforms to inform our understanding of the functional consequences of genomic alterations. This includes rigorous and scalable methods for simultaneously handling diverse data types such as gene expression array, miRNA, copy number, methylation, and next-generation sequencing data.

This material is written for statisticians who are interested in modeling and analyzing high-throughput data. Chapters by experts in the field offer a thorough introduction to the biological and technical principles behind multiplatform high-throughput experimentation.

Dr. Kim-Anh Do is Professor and Chair of the Department of Biostatistics at The University of Texas MD Anderson Cancer Center.

Dr. Zhaohui Steve Qin is an Associate Professor in the Department of Biostatistics and Bioinformatics at the Rollins School of Public Health, Emory University.

Dr. Marina Vannucci is a Professor in the Department of Statistics at Rice University, Director of the Interinstitutional Graduate Program in Biostatistics at Rice University, and an adjunct faculty member of The University of Texas MD Anderson Cancer Center.

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

ADVANCES IN STATISTICAL BIOINFORMATICS

Models and Integrative Inference for
High-Throughput Data

Edited by

KIM-ANH DO

The University of Texas MD Anderson Cancer Center, Houston, TX

ZHAOHUI STEVE QIN

Emory University, Atlanta, GA

MARINA VANNUCCI

Rice University, Houston, TX



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107027527

© Cambridge University Press 2013

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2013

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

Library of Congress Cataloguing in Publication data

Advances in statistical bioinformatics : models and integrative inference
for high-throughput data / [edited by] Kim-Anh Do, University of Texas MD
Anderson Cancer Center, Zhaohui Steve Qin, Emory University, Atlanta GA,
Marina Vannucci, Rice University, Houston, TX.

pages cm

Includes bibliographical references and index.

ISBN 978-1-107-02752-7 (hardback)

1. Bioinformatics – Statistical methods. 2. Biometry. 3. Genetics – Technique.
I. Do, Kim-Anh, 1960 – II. Qin, Zhaohui Steve, 1972 – III. Vannucci, Marina, 1966–
QH324.2.A395 2013

572.80285–dc23 2012049273

ISBN 978-1-107-02752-7 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication,
and does not guarantee that any content on such websites is, or will remain, accurate
or appropriate.

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

Contents

<i>List of Contributors</i>	<i>page</i> vii
<i>Preface</i>	xi
1. An Introduction to Next-Generation Biological Platforms <i>Virginia Mohlere, Wenting Wang, and Ganiraju Manyam</i>	1
2. An Introduction to The Cancer Genome Atlas <i>Bradley M. Broom and Rehan Akbani</i>	31
3. DNA Variant Calling in Targeted Sequencing Data <i>Wenyi Wang, Yu Fan, and Terence P. Speed</i>	54
4. Statistical Analysis of Mapped Reads from mRNA-Seq Data <i>Ernest Turro and Alex Lewin</i>	77
5. Model-Based Methods for Transcript Expression-Level Quantification in RNA-Seq <i>Zhaonan Sun, Han Wu, Zhaohui Qin, and Yu Zhu</i>	105
6. Bayesian Model-Based Approaches for Solexa Sequencing Data <i>Riten Mitra, Peter Mueller, and Yuan Ji</i>	126
7. Statistical Aspects of ChIP-Seq Analysis <i>Jonathan Cairns, Andy G. Lynch, and Simon Tavaré</i>	138
8. Bayesian Modeling of ChIP-Seq Data from Transcription Factor to Nucleosome Positioning <i>Raphael Gottardo and Sangsoon Woo</i>	170
9. Multivariate Linear Models for GWAS <i>Chiara Sabatti</i>	188
10. Bayesian Model Averaging for Genetic Association Studies <i>Christine Peterson, Michael Swartz, Sanjay Shete, and Marina Vannucci</i>	208
11. Whole-Genome Multi-SNP-Phenotype Association Analysis <i>Yongtao Guan and Kai Wang</i>	224

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

vi	<i>Contents</i>	
12.	Methods for the Analysis of Copy Number Data in Cancer Research <i>Bradley M. Broom, Kim-Anh Do, Melissa Bondy, Patricia Thompson, and Kevin Coombes</i>	244
13.	Bayesian Models for Integrative Genomics <i>Francesco C. Stingo and Marina Vannucci</i>	272
14.	Bayesian Graphical Models for Integrating Multiplatform Genomics Data <i>Wenting Wang, Veerabhadran Baladandayuthapani, Chris C. Holmes, and Kim-Anh Do</i>	292
15.	Genetical Genomics Data: Some Statistical Problems and Solutions <i>Hongzhe Li</i>	312
16.	A Bayesian Framework for Integrating Copy Number and Gene Expression Data <i>Yuan Ji, Filippo Trentini, and Peter Mueller</i>	331
17.	Application of Bayesian Sparse Factor Analysis Models in Bioinformatics <i>Haisu Ma and Hongyu Zhao</i>	350
18.	Predicting Cancer Subtypes Using Survival-Supervised Latent Dirichlet Allocation Models <i>Keegan Korthauer, John Dawson, and Christina Kendzioriski</i>	366
19.	Regularization Techniques for Highly Correlated Gene Expression Data with Unknown Group Structure <i>Brent A. Johnson</i>	382
20.	Optimized Cross-Study Analysis of Microarray-Based Predictors <i>Xiaogang Zhong, Luigi Marchionni, Leslie Cope, Edwin S. Iversen, Elizabeth S. Garrett-Mayer, Edward Gabrielson, and Giovanni Parmigiani</i>	398
21.	Functional Enrichment Testing: A Survey of Statistical Methods <i>Laila M. Poisson and Debashis Ghosh</i>	423
22.	Discover Trend and Progression Underlying High-Dimensional Data <i>Peng Qiu</i>	445
23.	Bayesian Phylogenetics Adapts to Comprehensive Infectious Disease Sequence Data <i>Jennifer A. Tom, Janet S. Sinsheimer, and Marc A. Suchard</i>	460
	<i>Index</i>	477

Color plates follow page 104.

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

List of Contributors

Rehan Akbani

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Veerabhadran Baladandayuthapani

*Department of Biostatistics, The University of Texas MD Anderson Cancer
Center*

Melissa Bondy

Department of Pediatrics, Baylor College of Medicine

Bradley M. Broom

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Jonathan Cairns

Department of Oncology, University of Cambridge

Kevin Coombes

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Leslie Cope

Departments of Oncology and Biostatistics, Johns Hopkins University

John Dawson

*Department of Biostatistics and Medical Informatics, University of Wisconsin-
Madison*

Kim-Anh Do

*Department of Biostatistics, The University of Texas MD Anderson Cancer
Center*

Yu Fan

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

viii

List of Contributors

Edward Gabrielson

Departments of Oncology and Pathology, Johns Hopkins University

Elizabeth S. Garrett-Mayer

Department of Biostatistics, Medical University of South Carolina

Debashis Ghosh

Department of Statistics, Penn State University

Raphael Gottardo

Public Health Sciences Division, Fred Hutchinson Cancer Research Center

Yongtao Guan

Department of Molecular and Human Genetics, Baylor College of Medicine

Chris C. Holmes

Department of Statistics, University of Oxford

Edwin S. Iversen

Institute of Statistics and Decision Sciences, Duke University

Yuan Ji

Center for Clinical and Research Informatics, NorthShore University Health-System

Brent A. Johnson

Department of Biostatistics and Bioinformatics, Emory University

Christina Kendziorski

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

Keegan Korthauer

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

Alex Lewin

Department of Epidemiology and Biostatistics, Imperial College London

Hongzhe Li

Department of Biostatistics and Epidemiology, University of Pennsylvania

Andy G. Lynch

Department of Oncology, University of Cambridge

Haisu Ma

Program in Computational Biology and Bioinformatics, Yale University

Ganiraju Manyam

Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center

Luigi Marchionni

Department of Oncology, Johns Hopkins University

Riten Mitra

Department of Mathematics, University of Texas at Austin

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

List of Contributors

ix

Virginia Mohlere

*Department of Biostatistics, The University of Texas MD Anderson Cancer
Center*

Peter Mueller

Department of Mathematics, University of Texas at Austin

Giovanni Parmigiani

Department of Biostatistics, Harvard University

Christine Peterson

Department of Statistics, Rice University

Laila M. Poisson

Department of Public Health Sciences, Henry Ford Hospital

Zhaohui Qin

Department of Biostatistics and Bioinformatics, Emory University

Peng Qiu

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Chiara Sabatti

Department of Health Research and Policy, Stanford University

Sanjay Shete

*Department of Biostatistics, The University of Texas MD Anderson Cancer
Center*

Janet S. Sinsheimer

Department of Human Genetics, University of California

Terence P. Speed

Department of Statistics, University of California

Francesco C. Stingo

*Department of Biostatistics, The University of Texas MD Anderson Cancer
Center*

Marc A. Suchard

*Departments of Biomathematics, Biostatistics, and Human Genetics, University
of California*

Zhaonan Sun

Department of Statistics, Purdue University

Michael Swartz

Division of Biostatistics, The University of Texas Health Science Center

Simon Tavaré

Department of Oncology, University of Cambridge

Patricia Thompson

Department of Cellular and Molecular Medicine, University of Arizona

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

x

List of Contributors

Jennifer A. Tom

Department of Statistics, University of California

Filippo Trentini

Department of Decision Science, Bocconi University

Ernest Turro

Department of Oncology, University of Cambridge

Marina Vannucci

Department of Statistics, Rice University

Kai Wang

Department of Computer Science and Engineering, University of California

Wenting Wang

Translational and Clinical Science, OSI Pharmaceuticals

Wenyi Wang

*Department of Bioinformatics and Computational Biology, The University of
Texas MD Anderson Cancer Center*

Sangsoon Woo

Fred Hutchinson Cancer Research Center

Han Wu

Department of Statistics, Purdue University

Hongyu Zhao

Division of Biostatistics, School of Public Health, Yale University

Xiaogang Zhong

Department of Applied Mathematics and Statistics, Johns Hopkins University

Yu Zhu

Department of Statistics, Purdue University

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

Preface

Providing genome-informed personalized treatment is an important goal of modern medicine. Identifying new translational targets in nucleic acid characterizations is an important step toward that goal. The information tsunami produced by such genome-scale investigations is stimulating parallel developments in statistical methodology and inference, analytical frameworks, and computational tools. Within the context of genomic medicine and with a strong focus on cancer research, this book describes the integration of high-throughput bioinformatics data from multiple platforms to inform our understanding of the functional consequences of genomic alterations. This includes rigorous and scalable methods for simultaneously handling diverse data types such as gene expression array, miRNA, copy number, methylation, and next-generation sequencing data. This book is intended for statisticians who are interested in modeling and analyzing high-throughput data. It covers the development and application of rigorous statistical methods (Bayesian and non-Bayesian) in the analysis of high-throughput bioinformatics data that arise from problems in medical and cancer research and molecular and structural biology. The specific focus of the volume is to provide an overview of the current state of the art of methods to integrate novel high-throughput multiplatform bioinformatics data, for a better understanding of the functional consequences of genomic alterations. The introductory description of biological and technical principles behind multiplatform high-throughput experimentation may be helpful to statisticians who are new to this research area.

Chapter 1 provides a detailed introduction to the next-generation high-throughput technology platforms that are the main workhorses in today's biomedical research laboratories and sets the scene for the subsequent methodology chapters. This chapter is mainly aimed at nonbiologists and details the unique measurement technologies, including next-generation DNA sequencing, genome profiling, and gene silencing, with associated idiosyncrasies for

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

xii

Preface

the different platforms. It also generates an overall outline of issues that statistical methodologies can address. Chapter 2 briefly describes The Cancer Genome Atlas (TCGA) project, an ambitious undertaking of the National Institutes of Health to identify all key genomic changes in the major types and subtypes of cancer. The description includes the history and goals of the TCGA project; how samples are collected and analyzed on multiple platforms; how the resulting data are processed, stored, and made available to qualified researchers; and what tools can be used to analyze TCGA data.

Subsequent chapters focus on specific methodological developments and are grouped approximately by the data types, with several chapters discussing the integration of at least two different data types. The central statistical topics addressed include experimental design, model building, group comparisons, regulatory networks, Bayesian networks, and gene interactions. The general theme of each chapter is to review existing methods, followed by a specific novel method developed by the author(s). Results are often demonstrated on simulated data and/or a real application data set. Additionally, relevant software may be discussed.

Chapter 3 describes a novel statistical method for analyzing the new array-based sequencing data. The novel method named SRMA increases the accuracy of identifying rare variants and thereby reduces the costs of subsequent sequence verifications. Chapters 4 and 5 discuss statistical approaches for quantifying gene expression and differential expression using RNA-seq data. Chapter 4 covers a wide range of topics, from read mapping, transcriptome assembly, and normalization to Poisson models to measure gene expression levels, methods to detect differentially expressed transcripts, and transcripts showing allelic imbalance. Chapter 5 focuses on transcript-level expression quantification using model-based methods. The authors provide a detailed review of six major approaches and discuss the advantages and limitations of all the methods. The authors then conduct performance comparisons using a series of real data sets to help researchers gain in-depth understanding of RNA-seq data.

Chapter 6 reviews a Bayesian approach for base calling, which uses a hierarchical model to account for the different sources of noise in the Solexa sequencing data. Chapters 7 and 8 survey statistical methodologies and Bayesian modeling for the analysis of ChIP sequencing data. Chapter 7 offers a detailed overview of the ChIP-seq experiment and steps required in the data analysis part, including read mapping, peak-calling, validation, and motif analysis. All main algorithms designed for the analysis of ChIP-seq data are discussed. In Chapter 8, the authors present a detailed description of the PICS/PING framework they have developed to analyze transcription factor and nucleosome

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)*Preface*

xiii

positioning ChIP-seq data. Chapters 9 through 11 discuss advanced statistical approaches for conducting association tests, particularly under the setting of genome-wide association study (GWAS). Chapter 9 surveys the standard methods of analysis for GWAS data, compares them with the underlying genetic model, and describes statistical approaches, such as penalized methods, that have attempted to bridge the gap between the theoretical models and the methods of analysis, with particular emphasis on Bayesian methods. Chapter 10 describes Bayesian techniques that can improve the reliability of inference through the incorporation of prior biological knowledge in SNP association studies. These methods can be used to identify the subset of SNPs most relevant to the disease under study and construct effective estimates that reflect uncertainty over model choice. The authors conclude with a brief discussion of Bayesian modeling and variable selection approaches for genome-wide association studies. Chapter 11 reviews recent developments in multi-SNP analysis, focusing on Bayesian variable selection regression, and compares them with penalized regression approaches. The authors explain the advantage of multi-SNP analysis in quantifying the total heritable signal in the data, including an interesting approach that can achieve this goal without identifying individual SNPs. The authors also discuss machine learning methods approaches for binary phenotypes.

Chapter 12 describes the problem of interpreting copy number data in the context of cancer research, specifically the problems that arise because of tumor ploidies significantly different from normal and the impact of normal DNA contamination of tumor samples, especially those from solid tumors. The authors then review a model that enables recovery of the copy number alterations in the tumor DNA from estimates of the tumor DNA fraction and ploidy, along with several algorithms for estimating these model parameters. Chapters 13 through 16 deal with integrated data analysis. Chapter 13 describes Bayesian variable selection models for integrative genomics. The authors first look into models that incorporate external biological information into the analysis of experimental data, in particular gene expression data. The authors then focus on Bayesian models that achieve an even greater type of integration, by incorporating into the modeling experimental data from different platforms, together with prior knowledge. In particular, they apply graphical models to integrate gene expression data with microRNA expression data. In Chapter 14, the authors discuss the problem of modeling the fundamental biological relationships among different types of genomic alterations surveyed in the same set of patient samples. The authors illustrate how to solve the problem using an objective Bayesian model selection approach for Gaussian graphical models and use the glioblastoma study in The Cancer Genome Atlas as an example. Three data

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

xiv

Preface

types, microRNA, gene expression, and patient survival time, are used in this integration study. Chapter 15 presents several recent statistical formulations and analysis methods for differential co-expression analysis and for multi-tissue gene expression data analysis and methods for eQTL analysis based on RNA-seq data. Chapter 16 considers the joint modeling of microarray RNA expression and DNA copy number data. The authors propose Bayesian mixture models for the observed copy numbers and gene expression measurements that define latent Gaussian probit scores for DNA and RNA and integrate the two platforms via a regression of the RNA probit scores on the DNA probit scores.

Chapters 17 through 19 discuss emerging ideas in genomic data analysis. Chapter 17 reviews the basic framework of Bayesian sparse factor modeling, a highly flexible and versatile approach for multivariate analysis, and describes its applications in bioinformatics, such as in transcription regulatory network inference and biological pathway analysis. In Chapter 18, the authors discuss applying the survival-supervised latent Dirichlet allocation (survLDA) model to utilize rich, diverse data types, such as high-throughput genomic information from multiple platforms, to make informed decisions for a particular patient's well-being, for personalized genomic medicine. The authors use simulation studies to understand what conditions can lead to an increased predictive power of survLDA. In Chapter 19, the author discusses how to achieve reliable estimation and variable selection in the linear model in the presence of high collinearity. The author examines deficiencies of the elastic net and argues in favor of a little-known competitor, the "Berhu" penalized least squares estimator, for high-dimensional regression analyses of genomic data.

Chapter 20 provides a simple, practical, and comprehensive technique for measuring consistency of molecular classification results across microarray platforms, without requiring subjective judgments about membership of samples in putative clusters. This methodology will be of value in consistently typing breast and other cancers across different studies and platforms in the future. Chapter 21 surveys a variety of pathway analysis methodologies for functional enrichment testing and discusses their strengths and weaknesses. A study of the gene expression profile differences between metastatic and localized prostate cancer is used for illustration. Chapter 22 discusses the problem of recovering progression patterns from high-dimensional data. The author argues that if the ordering of the cancer samples can be recovered, such ordering layout trajectories may reflect certain aspects of cancer progression and therefore lead to a better understanding of the disease. The final chapter, Chapter 23, reviews the evolving aims of phylogenetic inference, with successful insights derived from modern viral surveillance, and the techniques that can help to overcome the computational limitations of Bayesian phylogenetic inference.

Cambridge University Press

978-1-107-02752-7 - Advances in Statistical Bioinformatics Models and Integrative Inference
for High-Throughput Data

Edited by Kim-Anh Do, Zhaohui Steve Qin and Marina Vannucci

Frontmatter

[More information](#)

Preface

xv

We thank our colleagues, friends, and collaborators for contributing their ideas and insights to this collection. We are excited by the continuing opportunities for statistical developments in the area of integrated high-throughput bioinformatics data. We hope our readers will enjoy reading about new technology advances and new trends in statistical development.

Kim-Anh Do
Zhaohui Steve Qin
Marina Vannucci