Use R!

Advisors:
Robert Gentleman • Kurt Hornik • Giovanni Parmigiani

Use R!

Series Editors: Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani

Albert: Bayesian Computation with R
Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data Analysis with R
Claude: Morphometrics with R
Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis: With R and GGobi
Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies
Kleiber/Zeileis, Applied Econometrics with R
Nason: Wavelet Methods in Statistics with R
Paradis: Analysis of Phylogenetics and Evolution with R
Peng/Dominici: Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health
Pfaff: Analysis of Integrated and Cointegrated Time Series with R, 2nd edition
Sarkar: Lattice: Multivariate Data Visualization with R
Spector: Data Manipulation with R

Florian Hahne · Wolfgang Huber
Robert Gentleman · Seth Falcon

Bioconductor Case Studies

Springer

Florian Hahne
Fred Hutchinson Cancer Research Center
Division of Public Health Sciences
Program in Computational Biology
1100 Fairview Avenue, N., M2-B876,
PO Box 19024, Seattle, WA 98109-1024
USA
fhahne@fhcrc.org

Wolfgang Huber
Wellcome Trust Genome Campus
European Bioinformatics Institute
EMBL Outstation Hinxton
Hinxton
Cambridge CB10 1SD
UK
huber@ebi.ac.uk

Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue, N., M2-B876
PO Box 19024, Seattle, Washington 98102-1024
USA
rgentleman@fhcrc.org

Seth Falcon
seth@userprimary.net
http://userprimary.net/user

Series Editors
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue, N., M2-B876
PO Box 19024, Seattle, Washington 98102-1024
USA

Kurt Hornik
Department für Statistik und Mathematik
Wirtschaftsuniversität Wien Augasse 2-6
A-1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive Cancer Center
at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

ISBN: 978-0-387-77239-4 e-ISBN: 978-0-387-77240-0
DOI: 10.1007/978-0-387-77240-0

Library of Congress Control Number: 2008926129

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Recent developments in genomics and molecular biology finally carry the promise of understanding the functions of complex biological systems on a whole genome level. These developments have led to enormous amounts of data generated in highthroughput technologies, most prominently in gene expression microarrays. Within the Bioconductor project, an increasing number of researchers are trying to establish solutions for the analysis of such data, combining knowledge from such diverse disciplines as statistics, computer science, bioinformatics, and molecular biology.

With microarrays becoming a standard technology in many molecular biology labs, there is increased demand for comprehensive yet easy to follow instructions to the complex data analysis process. After many years of teaching introductory Bioconductor courses we can identify the main topics of interest, the common misunderstandings and pitfalls, and have learned to better understand key problems with which beginners to the analysis tasks are often challenged. In this book, we try to guide the readers through each step of the data analysis process, beginning from import and data processing to the generation of lists of differentially expressed genes and finally the modeling and interpretation of these lists in downstream analyses. Every chapter focuses on real data use cases that illustrate the problem, and we present both executable code and detailed background information for each step. A companion Webpage to this book can be found at `http://www.bioconductor.org/pub/docs/BioconductorCaseStudies`

Acknowledgments

We would like to thank Stefano Iacus for organizing the annual Bioconductor courses in Bressanone, Italy, which are the basis for this book and May Alipao, who has helped to organize the Bioconductor courses at the Hutchinson Center. We also thank the many students who attended these and other courses and whose countless questions, helpful remarks, and enthusiasm provided a valuable source of inspiration during the genesis of this book, immensely shaping the outcome you hold in your hands right now. We thank all the co-authors of the individual chapters. Their

expert knowledge was highly appreciated and helped to clarify the key concepts and to point out critical steps. We would also like to thank the following individuals who contributed in one or another way to this book: James W. MacDonald, Marc Carlson, Nolwenn LeMeur, Brig Mecham, Joern Toedling, Steffen Durinck, Anna Freni Sterrantino, Deepayan Sarkar, Rafael Irizarry, Jean (Zhijin) Wu, and Li Long.

RG expresses his thanks and appreciation to Tanja and Sophie, for their encouragement and understanding during the long hours spent working on this and other projects.

Florian Hahne
Wolfgang Huber
Robert Gentleman
Seth Falcon

Contents

List of Contributors

V.J. Carey, Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

T. Chiang, European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

S. Falcon, Scientific Software Engineer

F. Hahne, Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

W. Huber, European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

M. Morgan, Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

D. Scholtens, Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

A. von Heydebreck, Global Technologies, Merck KGaA, Darmstadt, FRG