

# 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, 2<sup>nd</sup> 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

Affymetrix is a registered trademark of Affymetrix, Inc.  
GeneChip is a registered trademark of Affymetrix, Inc.  
GenePix is a registered trademark of Molecular Devices Inc.  
Agilent is a trademark of Agilent Technologies

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

© 2008 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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

<b>Preface</b>	<b>v</b>
<b>List of Contributors</b>	<b>xi</b>
<b>1 The ALL Dataset</b>	<b>1</b>
F. Hahne and R. Gentleman	
1.1 Introduction . . . . .	1
1.2 The ALL data . . . . .	1
1.3 Data subsetting . . . . .	2
1.4 Nonspecific filtering . . . . .	3
1.5 BCR/ABL ALL1/AF4 subset . . . . .	4
<b>2 R and Bioconductor Introduction</b>	<b>5</b>
R. Gentleman, F. Hahne, S. Falcon, and M. Morgan	
2.1 Finding help in R . . . . .	5
2.2 Working with packages . . . . .	7
2.3 Some basic R . . . . .	8
2.4 Structures for genomic data . . . . .	11
2.5 Graphics . . . . .	20
<b>3 Processing Affymetrix Expression Data</b>	<b>25</b>
R. Gentleman and W. Huber	
3.1 The input data: CEL files . . . . .	25
3.2 Quality assessment . . . . .	28
3.3 Preprocessing . . . . .	32
3.4 Ranking and filtering probe sets . . . . .	33
3.5 Advanced preprocessing . . . . .	40
<b>4 Two-Color Arrays</b>	<b>47</b>
Florian Hahne and Wolfgang Huber	
4.1 Introduction . . . . .	47
4.2 Data import . . . . .	48
4.3 Image plots . . . . .	50

4.4	Normalization . . . . .	50
4.5	Differential expression . . . . .	57
<b>5</b>	<b>Fold-Changes, Log-Ratios, Background Correction, Shrinkage Estimation, and Variance Stabilization</b>	<b>63</b>
	W. Huber	
5.1	Fold-changes and (log-)ratios . . . . .	63
5.2	Background-correction and generalized logarithm . . . . .	65
5.3	Calling VSN . . . . .	70
5.4	How does VSN work? . . . . .	72
5.5	Robust fitting and the “most genes not differentially expressed” assumption . . . . .	74
5.6	Single-color normalization . . . . .	78
5.7	The interpretation of log-ratios . . . . .	79
5.8	Reference normalization . . . . .	81
<b>6</b>	<b>Easy Differential Expression</b>	<b>83</b>
	F. Hahne and W. Huber	
6.1	Example data . . . . .	83
6.2	Nonspecific filtering . . . . .	84
6.3	Differential expression . . . . .	85
6.4	Multiple testing correction . . . . .	87
<b>7</b>	<b>Differential Expression</b>	<b>89</b>
	W. Huber, D. Scholtens, F. Hahne, and A. von Heydebreck	
7.1	Motivation . . . . .	89
7.2	Nonspecific filtering . . . . .	90
7.3	Differential expression . . . . .	92
7.4	Multiple testing . . . . .	94
7.5	Moderated test statistics and the <b>limma</b> package . . . . .	95
7.6	Gene selection by Receiver Operator Characteristic (ROC) . . . . .	99
7.7	When power increases . . . . .	101
<b>8</b>	<b>Annotation and Metadata</b>	<b>103</b>
	W. Huber and F. Hahne	
8.1	Our data . . . . .	103
8.2	Multiple probe sets per gene . . . . .	106
8.3	Categories and overrepresentation . . . . .	107
8.4	Working with GO . . . . .	109
8.5	Other annotations available . . . . .	112
8.6	<b>biomaRt</b> . . . . .	113
8.7	Database versions of annotation packages . . . . .	115



<b>9</b>	<b>Supervised Machine Learning</b>	<b>121</b>
	R. Gentleman, W. Huber, and V. J. Carey	
9.1	Introduction . . . . .	121
9.2	The example dataset . . . . .	123
9.3	Feature selection and standardization . . . . .	124
9.4	Selecting a distance . . . . .	124
9.5	Machine learning . . . . .	126
9.6	Cross-validation . . . . .	129
9.7	Random forests . . . . .	132
9.8	Multigroup classification . . . . .	135
<b>10</b>	<b>Unsupervised Machine Learning</b>	<b>137</b>
	R. Gentleman and V. J. Carey	
10.1	Preliminaries . . . . .	137
10.2	Distances . . . . .	139
10.3	How many clusters? . . . . .	142
10.4	Hierarchical clustering . . . . .	144
10.5	Partitioning methods . . . . .	146
10.6	Self-organizing maps . . . . .	148
10.7	Hopach . . . . .	151
10.8	Silhouette plots . . . . .	152
10.9	Exploring transformations . . . . .	154
10.10	Remarks . . . . .	157
<b>11</b>	<b>Using Graphs for Interactome Data</b>	<b>159</b>
	T. Chiang, S. Falcon, F. Hahne, and W. Huber	
11.1	Introduction . . . . .	159
11.2	Exploring the protein interaction graph . . . . .	160
11.3	The co-expression graph . . . . .	162
11.4	Testing the association between physical interaction and coexpression . . . . .	164
11.5	Some harder problems . . . . .	165
11.6	Reading PSI-25 XML files from <i>IntAct</i> with the <b>Rintact</b> package . . . . .	165
<b>12</b>	<b>Graph Layout</b>	<b>173</b>
	F. Hahne, W. Huber, and R. Gentleman	
12.1	Introduction . . . . .	173
12.2	Layout and rendering using <b>Rgraphviz</b> . . . . .	175
12.3	Directed graphs . . . . .	180
12.4	Subgraphs . . . . .	185
12.5	Tooltips and hyperlinks on graphs . . . . .	187

<b>13 Gene Set Enrichment Analysis</b>	<b>193</b>
R. Gentleman, M. Morgan, and W. Huber	
13.1 Introduction . . . . .	193
13.2 Data analysis . . . . .	196
13.3 Identifying and assessing the effects of overlapping gene sets . . . . .	203
<b>14 Hypergeometric Testing Used for Gene Set Enrichment Analysis</b>	<b>207</b>
S. Falcon and R. Gentleman	
14.1 Introduction . . . . .	207
14.2 The basic problem . . . . .	208
14.3 Preprocessing and inputs . . . . .	209
14.4 Outputs and result summarization . . . . .	215
14.5 The conditional hypergeometric test . . . . .	218
14.6 Other collections of gene sets . . . . .	219
<b>15 Solutions to Exercises</b>	<b>221</b>
2 R and Bioconductor Introduction . . . . .	221
3 Processing Affymetrix Expression Data . . . . .	226
4 Two-Color Arrays . . . . .	230
5 Fold-Changes, Log-Ratios, Background Correction, Shrinkage Estimation, and Variance Stabilization . . . . .	231
6 Easy Differential Expression . . . . .	233
7 Differential Expression . . . . .	233
8 Annotation and Metadata . . . . .	234
9 Supervised Machine Learning . . . . .	241
10 Unsupervised Machine Learning . . . . .	249
11 Using Graphs for Interactome Data . . . . .	256
12 Graph Layout . . . . .	259
13 Gene Set Enrichment Analysis . . . . .	261
14 Hypergeometric Testing Used for Gene Set Enrichment Analysis . . . . .	265
<b>References</b>	<b>271</b>
<b>Index</b>	<b>277</b>

# 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