

# Contents

<b>I</b>	<b>Preprocessing data from genomic experiments</b>	<b>1</b>
<b>1</b>	<b>Preprocessing Overview</b>	<b>3</b>
	W. Huber, R. A. Irizarry, and R. Gentleman	
1.1	Introduction . . . . .	3
1.2	Tasks . . . . .	4
	1.2.1 Prerequisites . . . . .	5
	1.2.2 Stepwise and integrated approaches . . . . .	5
1.3	Data structures . . . . .	6
	1.3.1 Data sources . . . . .	6
	1.3.2 Facilities in R and Bioconductor . . . . .	7
1.4	Statistical background . . . . .	8
	1.4.1 An error model . . . . .	9
	1.4.2 The variance-bias trade-off . . . . .	11
	1.4.3 Sensitivity and specificity of probes . . . . .	11
1.5	Conclusion . . . . .	12
<b>2</b>	<b>Preprocessing High-density Oligonucleotide Arrays</b>	<b>13</b>
	B. M. Bolstad, R. A. Irizarry, L. Gautier, and Z. Wu	
2.1	Introduction . . . . .	13
2.2	Importing and accessing probe-level data . . . . .	15
	2.2.1 Importing . . . . .	15
	2.2.2 Examining probe-level data . . . . .	15
2.3	Background adjustment and normalization . . . . .	18
	2.3.1 Background adjustment . . . . .	18
	2.3.2 Normalization . . . . .	20
	2.3.3 vsn . . . . .	24
2.4	Summarization . . . . .	25
	2.4.1 <code>expresso</code> . . . . .	25
	2.4.2 <code>threestep</code> . . . . .	26
	2.4.3 <code>RMA</code> . . . . .	27
	2.4.4 <code>GCRMA</code> . . . . .	27
	2.4.5 <code>affypdnn</code> . . . . .	28

2.5	Assessing preprocessing methods . . . . .	29
2.5.1	Carrying out the assessment . . . . .	30
2.6	Conclusion . . . . .	32
<b>3</b>	<b>Quality Assessment of Affymetrix GeneChip Data</b>	<b>33</b>
	B. M. Bolstad, F. Collin, J. Brettschneider, K. Simpson, L. Cope, R. A. Irizarry, and T. P. Speed	
3.1	Introduction . . . . .	33
3.2	Exploratory data analysis . . . . .	34
3.2.1	Multi-array approaches . . . . .	35
3.3	Affymetrix quality assessment metrics . . . . .	37
3.4	RNA degradation . . . . .	38
3.5	Probe level models . . . . .	41
3.5.1	Quality diagnostics using PLM . . . . .	42
3.6	Conclusion . . . . .	47
<b>4</b>	<b>Preprocessing Two-Color Spotted Arrays</b>	<b>49</b>
	Y. H. Yang and A. C. Paquet	
4.1	Introduction . . . . .	49
4.2	Two-color spotted microarrays . . . . .	50
4.2.1	Illustrative data . . . . .	50
4.3	Importing and accessing probe-level data . . . . .	51
4.3.1	Importing . . . . .	51
4.3.2	Reading target information . . . . .	52
4.3.3	Reading probe-related information . . . . .	53
4.3.4	Reading probe and background intensities . . . . .	54
4.3.5	Data structure: the <i>marrayRaw</i> class . . . . .	54
4.3.6	Accessing the data . . . . .	56
4.3.7	Subsetting . . . . .	56
4.4	Quality assessment . . . . .	57
4.4.1	Diagnostic plots . . . . .	57
4.4.2	Spatial plots of spot statistics - <i>image</i> . . . . .	59
4.4.3	Boxplots of spot statistics - <i>boxplot</i> . . . . .	60
4.4.4	Scatter-plots of spot statistics - <i>plot</i> . . . . .	61
4.5	Normalization . . . . .	62
4.5.1	Two-channel normalization . . . . .	63
4.5.2	Separate-channel normalization . . . . .	64
4.6	Case study . . . . .	67
<b>5</b>	<b>Cell-Based Assays</b>	<b>71</b>
	W. Huber and F. Hahne	
5.1	Scope . . . . .	71
5.2	Experimental technologies . . . . .	71
5.2.1	Expression assays . . . . .	72
5.2.2	Loss of function assays . . . . .	72

5.2.3	Monitoring the response . . . . .	72
5.3	Reading data . . . . .	73
5.3.1	Plate reader data . . . . .	74
5.3.2	Further directions in normalization . . . . .	76
5.3.3	FCS format . . . . .	77
5.4	Quality assessment and visualization . . . . .	79
5.4.1	Visualization at the level of individual cells . . . . .	79
5.4.2	Visualization at the level of microtiter plates . . . . .	82
5.4.3	Brushing with Rggobi . . . . .	83
5.5	Detection of effectors . . . . .	85
5.5.1	Discrete Response . . . . .	85
5.5.2	Continuous response . . . . .	88
5.5.3	Outlook . . . . .	90
<b>6</b>	<b>SELDI-TOF Mass Spectrometry Protein Data</b>	<b>91</b>
	X. Li, R. Gentleman, X. Lu, Q. Shi, J. D. Iglehart, L. Harris, and A. Miron	
6.1	Introduction . . . . .	91
6.2	Baseline subtraction . . . . .	93
6.3	Peak detection . . . . .	95
6.4	Processing a set of calibration spectra . . . . .	96
6.4.1	Apply baseline subtraction to a set of spectra . . . . .	98
6.4.2	Normalize spectra . . . . .	99
6.4.3	Cutoff selection . . . . .	100
6.4.4	Identify peaks . . . . .	101
6.4.5	Quality assessment . . . . .	101
6.4.6	Get proto-biomarkers . . . . .	102
6.5	An example . . . . .	105
6.6	Conclusion . . . . .	108
<b>II</b>	<b>Meta-data: biological annotation and visualization</b>	<b>111</b>
<b>7</b>	<b>Meta-data Resources and Tools in Bioconductor</b>	<b>113</b>
	R. Gentleman, V. J. Carey, and J. Zhang	
7.1	Introduction . . . . .	113
7.2	External annotation resources . . . . .	115
7.3	Bioconductor annotation concepts: curated persistent packages and Web services . . . . .	116
7.3.1	Annotating a platform: HG-U95Av2 . . . . .	117
7.3.2	An Example . . . . .	118
7.3.3	Annotating a genome . . . . .	119
7.4	The <code>annotate</code> package . . . . .	119
7.5	Software tools for working with Gene Ontology (GO) . . . . .	120

7.5.1	Basics of working with the GO package . . . . .	121
7.5.2	Navigating the hierarchy . . . . .	122
7.5.3	Searching for terms . . . . .	122
7.5.4	Annotation of GO terms to LocusLink sequences: evidence codes . . . . .	123
7.5.5	The GO graph associated with a term . . . . .	125
7.6	Pathway annotation packages: KEGG and cMAP . . . . .	125
7.6.1	KEGG . . . . .	126
7.6.2	cMAP . . . . .	127
7.6.3	A Case Study . . . . .	129
7.7	Cross-organism annotation: the homology packages . . . . .	130
7.8	Annotation from other sources . . . . .	132
7.9	Discussion . . . . .	133
<b>8</b>	<b>Querying On-line Resources</b>	<b>135</b>
	V. J. Carey, D. Temple Lang, J. Gentry, J. Zhang, and R. Gentleman	
8.1	The Tools . . . . .	135
8.1.1	Entrez . . . . .	137
8.1.2	Entrez examples . . . . .	137
8.2	PubMed . . . . .	138
8.2.1	Accessing PubMed information . . . . .	139
8.2.2	Generating HTML output for your abstracts . . . . .	141
8.3	KEGG via SOAP . . . . .	142
8.4	Getting gene sequence information . . . . .	144
8.5	Conclusion . . . . .	145
<b>9</b>	<b>Interactive Outputs</b>	<b>147</b>
	C. A. Smith, W. Huber, and R. Gentleman	
9.1	Introduction . . . . .	147
9.2	A simple approach . . . . .	148
9.3	Using the <code>annaffy</code> package . . . . .	149
9.4	Linking to On-line Databases . . . . .	152
9.5	Building HTML pages . . . . .	153
9.5.1	Limiting the results . . . . .	153
9.5.2	Annotating the probes . . . . .	154
9.5.3	Adding other data . . . . .	155
9.6	Graphical displays with drill-down functionality . . . . .	156
9.6.1	HTML image maps . . . . .	157
9.6.2	Scalable Vector Graphics (SVG) . . . . .	158
9.7	Searching Meta-data . . . . .	159
9.7.1	Text searching . . . . .	159
9.8	Concluding Remarks . . . . .	160
<b>10</b>	<b>Visualizing Data</b>	<b>161</b>

W. Huber, X. Li, and R. Gentleman

- 10.1 Introduction . . . . . 161
- 10.2 Practicalities . . . . . 162
- 10.3 High-volume scatterplots . . . . . 163
  - 10.3.1 A note on performance . . . . . 164
- 10.4 Heatmaps . . . . . 166
  - 10.4.1 Heatmaps of residuals . . . . . 168
- 10.5 Visualizing distances . . . . . 170
  - 10.5.1 Multidimensional scaling . . . . . 173
- 10.6 Plotting along genomic coordinates . . . . . 174
  - 10.6.1 Cumulative Expression . . . . . 178
- 10.7 Conclusion . . . . . 179

**III Statistical analysis for genomic experiments 181**

**11 Analysis Overview 183**

V. J. Carey and R. Gentleman

- 11.1 Introduction and road map . . . . . 183
  - 11.1.1 Distance concepts . . . . . 184
  - 11.1.2 Differential expression . . . . . 184
  - 11.1.3 Cluster analysis . . . . . 184
  - 11.1.4 Machine learning . . . . . 184
  - 11.1.5 Multiple comparisons . . . . . 185
  - 11.1.6 Workflow support . . . . . 185
- 11.2 Absolute and relative expression measures . . . . . 185

**12 Distance Measures in DNA Microarray Data Analysis. 189**

R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim

- 12.1 Introduction . . . . . 189
- 12.2 Distances . . . . . 191
  - 12.2.1 Definitions . . . . . 191
  - 12.2.2 Distances between points . . . . . 192
  - 12.2.3 Distances between distributions . . . . . 195
  - 12.2.4 Experiment-specific distances between genes . . . . . 198
- 12.3 Microarray data . . . . . 199
  - 12.3.1 Distances and standardization . . . . . 199
- 12.4 Examples . . . . . 201
  - 12.4.1 A co-citation example . . . . . 203
  - 12.4.2 Adjacency . . . . . 207
- 12.5 Discussion . . . . . 208

**13 Cluster Analysis of Genomic Data 209**

K. S. Pollard and M. J. van der Laan

- 13.1 Introduction . . . . . 209

13.2	Methods . . . . .	210
13.2.1	Overview of clustering algorithms . . . . .	210
13.2.2	Ingredients of a clustering algorithm . . . . .	211
13.2.3	Building sequences of clustering results . . . . .	211
13.2.4	Visualizing clustering results . . . . .	214
13.2.5	Statistical issues in clustering . . . . .	215
13.2.6	Bootstrapping a cluster analysis . . . . .	216
13.2.7	Number of clusters . . . . .	217
13.3	Application: renal cell cancer . . . . .	222
13.3.1	Gene selection . . . . .	222
13.3.2	HOPACH clustering of genes . . . . .	223
13.3.3	Comparison with PAM . . . . .	224
13.3.4	Bootstrap resampling . . . . .	224
13.3.5	HOPACH clustering of arrays . . . . .	224
13.3.6	Output files . . . . .	226
13.4	Conclusion . . . . .	228
<b>14</b>	<b>Analysis of Differential Gene Expression Studies</b>	<b>229</b>
	D. Scholtens and A. von Heydebreck	
14.1	Introduction . . . . .	229
14.2	Differential expression analysis . . . . .	230
14.2.1	Example: ALL data . . . . .	232
14.2.2	Example: Kidney cancer data . . . . .	236
14.3	Multifactor experiments . . . . .	239
14.3.1	Example: Estrogen data . . . . .	241
14.4	Conclusion . . . . .	248
<b>15</b>	<b>Multiple Testing Procedures: the multtest Package and Applications to Genomics</b>	<b>249</b>
	K. S. Pollard, S. Dudoit, and M. J. van der Laan	
15.1	Introduction . . . . .	249
15.2	Multiple hypothesis testing methodology . . . . .	250
15.2.1	Multiple hypothesis testing framework . . . . .	250
15.2.2	Test statistics null distribution . . . . .	255
15.2.3	Single-step procedures for controlling general Type I error rates $\theta(F_{V_n})$ . . . . .	256
15.2.4	Step-down procedures for controlling the family-wise error rate . . . . .	257
15.2.5	Augmentation multiple testing procedures for controlling tail probability error rates . . . . .	258
15.3	Software implementation: R <code>multtest</code> package . . . . .	259
15.3.1	Resampling-based multiple testing procedures: MTP function . . . . .	260
15.3.2	Numerical and graphical summaries . . . . .	262
15.4	Applications: ALL microarray data set . . . . .	262

15.4.1	ALL data package and initial gene filtering . . . . .	262
15.4.2	Association of expression measures and tumor cellular subtype: Two-sample $t$ -statistics . . . . .	263
15.4.3	Augmentation procedures . . . . .	265
15.4.4	Association of expression measures and tumor molecular subtype: Multi-sample $F$ -statistics . . . . .	266
15.4.5	Association of expression measures and time to relapse: Cox $t$ -statistics . . . . .	268
15.5	Discussion . . . . .	270
<b>16</b>	<b>Machine Learning Concepts and Tools for Statistical Genomics</b>	<b>273</b>
	V. J. Carey	
16.1	Introduction . . . . .	273
16.2	Illustration: Two continuous features; decision regions . . . . .	274
16.3	Methodological issues . . . . .	276
	16.3.1 Families of learning methods . . . . .	276
	16.3.2 Model assessment . . . . .	281
	16.3.3 Metatheorems on learner and feature selection . . . . .	283
	16.3.4 Computing interfaces . . . . .	284
16.4	Applications . . . . .	285
	16.4.1 Exploring and comparing classifiers with the ALL data . . . . .	285
	16.4.2 Neural net initialization, convergence, and tuning . . . . .	287
	16.4.3 Other methods . . . . .	287
	16.4.4 Structured cross-validation support . . . . .	288
	16.4.5 Assessing variable importance . . . . .	289
	16.4.6 Expression density diagnostics . . . . .	289
16.5	Conclusions . . . . .	291
<b>17</b>	<b>Ensemble Methods of Computational Inference</b>	<b>293</b>
	T. Hothorn, M. Dettling, and P. Bühlmann	
17.1	Introduction . . . . .	293
17.2	Bagging and random forests . . . . .	295
17.3	Boosting . . . . .	296
17.4	Multiclass problems . . . . .	298
17.5	Evaluation . . . . .	298
17.6	Applications: tumor prediction . . . . .	300
	17.6.1 Acute lymphoblastic leukemia . . . . .	300
	17.6.2 Renal cell cancer . . . . .	303
17.7	Applications: Survival analysis . . . . .	307
17.8	Conclusion . . . . .	310
<b>18</b>	<b>Browser-based Affymetrix Analysis and Annotation</b>	<b>313</b>

C. A. Smith	
18.1	Introduction . . . . . 313
18.1.1	Key user interface features . . . . . 314
18.2	Deploying webbioc . . . . . 315
18.2.1	System requirements . . . . . 315
18.2.2	Installation . . . . . 315
18.2.3	Configuration . . . . . 316
18.3	Using webbioc . . . . . 317
18.3.1	Data Preprocessing . . . . . 317
18.3.2	Differential expression multiple testing . . . . . 318
18.3.3	Linked annotation meta-data . . . . . 320
18.3.4	Retrieving results . . . . . 321
18.4	Extending webbioc . . . . . 322
18.4.1	Architectural overview . . . . . 322
18.4.2	Creating a new module . . . . . 324
18.5	Conclusion . . . . . 326
<b>IV</b>	<b>Graphs and networks 327</b>
<b>19</b>	<b>Introduction and Motivating Examples 329</b>
R. Gentleman, W. Huber, and V. J. Carey	
19.1	Introduction . . . . . 329
19.2	Practicalities . . . . . 330
19.2.1	Representation . . . . . 330
19.2.2	Algorithms . . . . . 330
19.2.3	Data Analysis . . . . . 331
19.3	Motivating examples . . . . . 331
19.3.1	Biomolecular Pathways . . . . . 331
19.3.2	Gene ontology: A graph of concept-terms . . . . . 333
19.3.3	Graphs induced by literature references and citations . . . . . 334
19.4	Discussion . . . . . 336
<b>20</b>	<b>Graphs 337</b>
W. Huber, R. Gentleman, and V. J. Carey	
20.1	Overview . . . . . 337
20.2	Definitions . . . . . 338
20.2.1	Special types of graphs . . . . . 341
20.2.2	Random graphs . . . . . 343
20.2.3	Node and edge labeling . . . . . 344
20.2.4	Searching and related algorithms . . . . . 344
20.3	Cohesive subgroups . . . . . 344
20.4	Distances . . . . . 346



<b>21 Bioconductor Software for Graphs</b>	<b>347</b>
V. J. Carey, R. Gentleman, W. Huber, and J. Gentry	
21.1 Introduction . . . . .	347
21.2 The graph package . . . . .	348
21.2.1 Getting started . . . . .	349
21.2.2 Random graphs . . . . .	352
21.3 The RBGL package . . . . .	352
21.3.1 Connected graphs . . . . .	355
21.3.2 Paths and related concepts . . . . .	357
21.3.3 RBGL summary . . . . .	360
21.4 Drawing graphs . . . . .	360
21.4.1 Global attributes . . . . .	363
21.4.2 Node and edge attributes . . . . .	363
21.4.3 The function agopen and the Ragraph class . . . . .	365
21.4.4 User-defined drawing functions . . . . .	366
21.4.5 Image maps on graphs . . . . .	368
<b>22 Case Studies Using Graphs on Biological Data</b>	<b>369</b>
R. Gentleman, D. Scholtens, B. Ding, V. J. Carey, and W. Huber	
22.1 Introduction . . . . .	369
22.2 Comparing the transcriptome and the interactome . . . . .	370
22.2.1 Testing associations . . . . .	371
22.2.2 Data analysis . . . . .	373
22.3 Using GO . . . . .	374
22.3.1 Finding interesting GO terms . . . . .	375
22.4 Literature co-citation . . . . .	378
22.4.1 Statistical development . . . . .	380
22.4.2 Comparisons of interest . . . . .	382
22.4.3 Examples . . . . .	382
22.5 Pathways . . . . .	387
22.5.1 The graph structure of pathways . . . . .	388
22.5.2 Relating expression data to pathways . . . . .	390
22.6 Concluding remarks . . . . .	393
<b>V Case studies</b>	<b>395</b>
<b>23 limma: Linear Models for Microarray Data</b>	<b>397</b>
G. K. Smyth	
23.1 Introduction . . . . .	397
23.2 Data representations . . . . .	398
23.3 Linear models . . . . .	399
23.4 Simple comparisons . . . . .	400
23.5 Technical Replication . . . . .	403
23.6 Within-array replicate spots . . . . .	406

23.7	Two groups . . . . .	407
23.8	Several groups . . . . .	409
23.9	Direct two-color designs . . . . .	411
23.10	Factorial designs . . . . .	412
23.11	Time course experiments . . . . .	414
23.12	Statistics for differential expression . . . . .	415
23.13	Fitted model objects . . . . .	417
23.14	Preprocessing considerations . . . . .	418
23.15	Conclusion . . . . .	420
<b>24</b>	<b>Classification with Gene Expression Data</b>	<b>421</b>
	M. Dettling	
24.1	Introduction . . . . .	421
24.2	Reading and customizing the data . . . . .	422
24.3	Training and validating classifiers . . . . .	423
24.4	Multiple random divisions . . . . .	426
24.5	Classification of test data . . . . .	428
24.6	Conclusion . . . . .	429
<b>25</b>	<b>From CEL Files to Annotated Lists of Interesting Genes</b>	<b>431</b>
	R. A. Irizarry	
25.1	Introduction . . . . .	431
25.2	Reading CEL files . . . . .	432
25.3	Preprocessing . . . . .	432
25.4	Ranking and filtering genes . . . . .	433
	25.4.1 Summary statistics and tests for ranking . . . . .	434
	25.4.2 Selecting cutoffs . . . . .	437
	25.4.3 Comparison . . . . .	437
25.5	Annotation . . . . .	438
	25.5.1 PubMed abstracts . . . . .	439
	25.5.2 Generating reports . . . . .	441
25.6	Conclusion . . . . .	442
<b>A</b>	<b>Details on selected resources</b>	<b>443</b>
A.1	Data sets . . . . .	443
	A.1.1 ALL . . . . .	443
	A.1.2 Renal cell cancer . . . . .	443
	A.1.3 Estrogen receptor stimulation . . . . .	443
A.2	URLs for projects mentioned . . . . .	444
	<b>References</b>	<b>445</b>
	<b>Index</b>	<b>465</b>



<http://www.springer.com/978-0-387-25146-2>

Bioinformatics and Computational Biology Solutions

Using R and Bioconductor

Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R.; Dudoit,

S. (Eds.)

2005, XX, 474 p. 128 illus. in color., Hardcover

ISBN: 978-0-387-25146-2