

Contents

Preface	v
List of Contributors	xi
1 The ALL Dataset	1
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
2 R and Bioconductor Introduction	5
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
3 Processing Affymetrix Expression Data	25
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
4 Two-Color Arrays	47
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
5	Fold-Changes, Log-Ratios, Background Correction, Shrinkage Estimation, and Variance Stabilization	63
	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
6	Easy Differential Expression	83
	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
7	Differential Expression	89
	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 limma package	95
7.6	Gene selection by Receiver Operator Characteristic (ROC)	99
7.7	When power increases	101
8	Annotation and Metadata	103
	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	biomaRt	113
8.7	Database versions of annotation packages	115

9	Supervised Machine Learning	121
	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
10	Unsupervised Machine Learning	137
	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
11	Using Graphs for Interactome Data	159
	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 Rintact package	165
12	Graph Layout	173
	F. Hahne, W. Huber, and R. Gentleman	
12.1	Introduction	173
12.2	Layout and rendering using Rgraphviz	175
12.3	Directed graphs	180
12.4	Subgraphs	185
12.5	Tooltips and hyperlinks on graphs	187

13 Gene Set Enrichment Analysis	193
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
14 Hypergeometric Testing Used for Gene Set Enrichment Analysis	207
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
15 Solutions to Exercises	221
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
References	271
Index	277