

Giovanni Parmigiani
Elizabeth S. Garrett
Rafael A. Irizarry
Scott L. Zeger
Editors

The Analysis of Gene Expression Data

Methods and Software

With 113 Figures, Including 37 Color Plates



Springer

Contents

Preface

Contributors **xvn**

Color Insert **(facing page 236)**

**The Analysis of Gene Expression Data: An Overview of
Methods and Software** **1**

*Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry,
and Scott L. Zeger*

1.1	Measuring Gene Expression Using Microarrays.	1
1.1.1	Microarray Technologies.	1
1.1.2	Sources of Variation in Gene Expression Measurements Using Microarrays.	4
1.1.3	Phases of Microarray Data Analysis.	5
1.2	Design of Microarray Experiments.	7
1.2.1	Replication and Sample Size Considerations	7
1.2.2	Design of Two-Channel Arrays.	9
1.3	Data Storage	9
1.3.1	Databases.	9
1.3.2	Standards.	10
1.3.3	Statistical Analysis Languages.	11
1.4	Preprocessing	12
1.4.1	Image Analysis.	12
1.4.2	Visualizations for Quality Control.	12
1.4.3	Background Subtraction.	13
1.4.4	Probe-level Analysis of Oligonucleotide Arrays	14
1.4.5	Within-Array Normalization of cDNA Arrays	15
1.4.6	Normalization Across Arrays.	15
1.5	Screening for Differentially Expressed Genes.	16
1.5.1	Estimation or Selection?.	16
1.5.2	One Problem or Many?.	17
1.5.3	Selection and False Discovery Rates.	18
1.5.4	Beyond Two Groups.	19
1.6	Challenges of Genome Biometry Analyses.	19

1.7	Visualization and Unsupervised Analyses	21
1.7.1	Profile Visualization	21
1.7.2	Why Clustering?	22
1.7.3	Hierarchical Clustering	23
1.7.4	/c-Means Clustering and Self-Organizing Maps	25
1.7.5	Model-Based Clustering	26
1.7.6	Principal Components Analysis.	26
1.7.7	Multidimensional Scaling	27
1.7.8	Identifying Novel Molecular Subclasses.	27
1.7.9	Time Series Analysis	28
1.8	Prediction	29
1.8.1	Prediction Tools	29
1.8.2	Dimension Reduction.	30
1.8.3	Evaluation of Classifiers.	30
1.8.4	Regression-Based Approaches.	31
1.8.5	Classification Trees.	31
1.8.6	Probabilistic Model-Based Classification	32
1.8.7	Discriminant Analysis.	33
1.8.8	Nearest-Neighbor Classifiers.	33
1.8.9	Support Vector Machines.	33
1.9	Free and Open-Source Software.	33
1.9.1	Whitehead Institute Tools.	34
1.9.2	Eisen Lab Tools.	34
1.9.3	TIGR Tools.	34
1.9.4	GeneX and CyberT.	35
1.9.5	Projects at NCBI.	35
1.9.6	BRB.	35
1.9.7	The OOML library.	36
1.9.8	MatArray.	36
1.9.9	BASE.	36
1.10	Conclusion.	36

2 Visualization and Annotation of Genomic Experiments 46

Robert Gentleman and Vincent Carey

2.1	Introduction.	46
2.2	Motivations for Component-Based Software.	47
2.3	Formalism.	49
2.4	Bioconductor Software for Filtering, Exploring, and Interpreting Microarray Experiments.	50
2.4.1	Formal Data Structures and Methods for Multiple Microarrays.	50
2.4.2	Tools for Filtering Gene Expression Data: The Closure Concept	54

2.4.3	Expression Density Diagnostics: High-Throughput Exploratory Data Analysis for Microarrays	55
2.4.4	Annotation	57
2.5	Visualization	58
2.5.1	Chromosomes.	59
2.6	Applications.	64
2.6.1	A Case Study of Gene Filtering	64
2.6.2	Application of Expression Density Diagnostics	67
2.7	Conclusions.	70

Bioconductor R Packages for Exploratory Analysis and Normalization of cDNA Microarray Data 73

Sandrine Dudoit and Jean Yee Hwa Yang

3.1	Introduction	73
3.1.1	Overview of Packages.	73
3.1.2	Two-Color cDNA Microarray Experiments	75
3.2	Methods	76
3.2.1	Standards for Microarray Data	76
3.2.2	Object-Oriented Programming: Microarray Classes and Methods.	77
3.2.3	Diagnostic Plots	78
3.2.4	Normalization Using Robust Local Regression	79
3.3	Application: Swirl Microarray Experiment	80
3.4	Software	81
3.4.1	Package marrayClasses—Classes and Methods for cDNA Microarray Data	81
3.4.2	Package marrayInput—Data Input for cDNA Microarrays.	89
3.4.3	Package marrayPlots—Diagnostic Plots for cDNA Microarray Data	91
3.4.4	Package marrayNorm—Location and Scale Normalization for cDNA Microarray Data	96
3.5	Discussion.	99

An R Package for Analyses of Affymetrix Oligonucleotide Arrays 102

Rafael A. Irizarry, Laurent Gautier, and Leslie M. Cope

4.1	Introduction	102
4.2	Methods.	103
4.2.1	Notation	103
4.2.2	The CEL/CDF Convention	104
4.2.3	Probe Pair Sets.	106
4.2.4	Probe-Level Objects	107

Contents

4.2.5	Normalization	108
4.2.6	Exploratory Data Analysis of Probe-Level Data	111
4.3	Application	113
4.3.1	Expression Measures	113
4.4	Software	115
4.4.1	A Case Study	115
4.4.2	Extending the Package	118
4.5	Conclusion	118

DNA-Chip Analyzer (dChip) 120

Cheng Li and Wing Hung Wong

5.1	Introduction	120
5.2	Methods	121
5.2.1	Normalization of Arrays Based on an "Invariant Set".	121
5.2.2	Model-Based Analysis of Oligonucleotide Arrays	122
5.2.3	Confidence Interval for Fold Change	122
5.2.4	Pooling Replicate Arrays Considering Measurement Accuracy	124
5.3	Software and Applications	125
5.3.1	Reading in Array Data Files	125
5.3.2	Viewing an Array Image	127
5.3.3	Normalizing Arrays	129
5.3.4	Viewing PM/MM Data	129
5.3.5	Calculating Model-Based Expression Indexes	131
5.3.6	Filter Genes	132
5.3.7	Hierarchical Clustering	133
5.3.8	Comparing Samples	135
5.3.9	Mapping Genes to Chromosomes	137
5.3.10	Sample Classification by Linear Discriminant Analysis	138
5.4	Discussion	139

Expression Profiler 142

*Jaak Vilo, Misha Kapushesky, Patrick Kemmeren, Ugis Sarkans,
and Alvis Brazma*

6.1	Introduction	142
6.2	EPCLUST	143
6.2.1	EPCLUST: Data Import	143
6.2.2	EPCLUST: Data Filtering	144
6.2.3	EPCLUST: Data Annotation	146
6.2.4	EPCLUST: Data Environment	147

6.2.5	EPCLUST: Data Analysis	148
6.3	URLMAP: Cross-Linking of the Analysis Results Between the Tools and Databases.	151
6.4	EP:GO GeneOntology Browser.	152
6.5	EP:PPI: Comparison of Protein Pairs and Expression . .	153
6.6	Pattern Discovery, Pattern Matching, and Visualization Tools.	154
6.7	An Example of the Data Analysis and Visualizations Performed by the Tools in Expression Profiler.	154
6.8	Integration of Expression Profiler with Public Microarray Databases.	159
6.9	Conclusions.	160

An S-PLUS Library for the Analysis and Visualization of Differential Expression **163**

Jae K. Lee and Michael O'Connell

7.1	Introduction	163
7.2	Assessment of Differential Expression.	164
7.2.1	Local Pooled Error.	165
7.2.2	Tests for Differential Expression.	169
7.2.3	Cluster Analysis and Visualization.	171
7.3	Analysis of Melanoma Expression	174
7.3.1	Tests for Differential Expression.	175
7.3.2	Cluster Analysis and Visualization.	178
7.3.3	Annotation	180
7.4	Discussion	181

8 DRAGON and DRAGON View: Methods for the Annotation, Analysis, and Visualization of Large-Scale Gene Expression Data **185**

*Christopher M.L.S. Bouton, George Henry, Carlo Colantuoni,
and Jonathan Pevsner*

8.1	Introduction	185
8.2	System and Methods.	189
8.2.1	Overview of DRAGON.	189
8.2.2	DRAGON'S Hardware, Software, and Database Architecture	190
8.2.3	Cross-Referencing Information in DRAGON . .	192
8.2.4	The DRAGON Search and Annotate Tools . . .	193
8.2.5	The DRAGON View Data Visualization Tools .	196
8.2.6	DRAGON Gram: A Novel Visualization Tool . .	198
8.3	Implementation	199
8.4	Discussion and Conclusion.	204

9 SNOMAD: Biologist-Friendly Web Tools for the Standardization and Normalization of Microarray Data 210

Carlo Colantuoni, George Henry, Christopher M.L.S. Bouton, Scott L. Zeger, and Jonathan Pevsner

9.1	Introduction	210
9.2	Methods and Application	212
9.2.1	Overview of Experimental and Data Analysis Procedures	212
9.2.2	Background Subtraction	214
9.2.3	Global Mean Normalization	214
9.2.4	Standard Data Transformation and Visualization Methods	215
9.2.5	Local Mean Normalization Across Element Signal Intensity	217
9.2.6	Local Variance Correction Across Element Signal Intensity	219
9.2.7	Local Mean Normalization Across the Microarray Surface	223
9.3	Software	225
9.4	Discussion	226

10 Microarray Analysis Using the MicroArray Explorer 229

Peter F. Lemkin, Gregory C. Thornwall, and Jai Evans

10.1	Introduction	229
10.1.1	Need for the Methodology	230
10.1.2	Basic Ideas Behind the Approach	231
10.2	Methods—Statistical and Informatics Basis	232
10.2.1	Analysis Paradigm	235
10.2.2	Particular Analysis Methods	238
10.2.3	Data Conversion	238
10.3	Software	239
10.3.1	System Design—Software Implementation	244
10.3.2	How to Download the Software	247
10.3.3	Strengths and Weaknesses of the Approach	248
10.4	Applications	249
10.5	Discussion	251

11 Parametric Empirical Bayes Methods for Microarrays 254

Michael A. Newton and Christina Kendziorski

11.1	Introduction	254
11.2	EB Methods	256
11.2.1	Canonical EB Example	256
11.2.2	General Model Structure: Two Conditions	256

11.2.3	Multiple Conditions	258
11.2.4	The Gamma-Gamma and Lognormal-Normal Models	259
11.2.5	Model Fitting	260
11.3	Software	261
11.4	Application	263
11.5	Discussion	269

12 SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays **272**

John D. Storey and Robert Tibshirani

12.1	Introduction	272
12.2	Methods and Applications	273
12.2.1	Multiple Hypothesis Testing	273
12.2.2	An Application	275
12.2.3	Forming the Test Statistics	276
12.2.4	Calculating the Null Distribution	277
12.2.5	The SAM Thresholding Procedure	278
12.2.6	Estimating False-Discovery Rates	280
12.3	Software	283
12.3.1	Obtaining the Software	283
12.3.2	Data Formats	283
12.3.3	Response Format	284
12.3.4	Example Input Data File for an Unpaired Problem	285
12.3.5	Block Permutations	285
12.3.6	Normalization of Experiments	285
12.3.7	Handling Missing Data	287
12.3.8	Running SAM	287
12.3.9	Format of the Significant Gene List	288
12.4	Discussion	289

13 Adaptive Gene Picking with Microarray Data: Detecting Important Low Abundance Signals **291**

*Yi Lin, Samuel T. Nadler, Hong Lan, Alan D. Attie,
and Brian S. Yandell*

13.1	Introduction	291
13.2	Methods	292
13.2.1	Background Subtraction	292
13.2.2	Transformation to Approximate Normality	293
13.2.3	Differential Expression Across Conditions	295
13.2.4	Robust Center and Spread	297

13.2.5	Formal Evaluation of Significant Differential Expression	299
13.2.6	Simulation Studies	301
13.2.7	Comparison of Methods with <i>E. coli</i> Data	304
13.3	Software	304
13.4	Application	306
13.4.1	Diabetes and Obesity Studies	306
13.4.2	Software Example	308
14	MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments	313
	<i>Hao Wu, M. Kathleen Kerr, Xiangqin Cui, and Gary A. Churchill</i>	
14.1	Introduction	313
14.2	Methods	314
14.2.1	Data Acquisition	315
14.2.2	ANOVA Models for Microarray Data	315
14.2.3	Experimental Design for Microarrays	317
14.2.4	Data Transformations	321
14.2.5	Algorithms for Computing ANOVA Estimates	322
14.2.6	Statistical Inference	323
14.2.7	Cluster Analysis	327
14.3	Software	328
14.3.1	Availability	328
14.3.2	Functionality	329
14.4	Data Analysis with MAANOVA	334
14.5	Discussion	339
15	GeneClust	342
	<i>Kim-Anh Do, Bradley Broom, and Sijin Wen</i>	
15.1	Introduction	342
15.2	Methods	343
15.2.1	Algorithm	343
15.2.2	Choice of Cluster Size via the Gap Statistic	344
15.2.3	Supervised Gene Shaving for Class Discrimination	346
15.3	Software	347
15.3.1	The GeneShaving Package	347
15.3.2	GeneClust: A Faster Implementation of Gene Shaving	352
15.4	Applications	354
15.4.1	The Alon Colon Dataset	354
15.4.2	The NCI60 Dataset	356
15.5	Discussion	358

16 POE: Statistical Methods for Qualitative Analysis of Gene Expression 362

Elizabeth S. Garrett and Giovanni Parmigiani

16.1	Introduction	362
16.2	Methodology.	364
16.2.1	Mixture Model for Gene Expression.	364
16.2.2	Useful Representations of the Results.	366
16.2.3	Bayesian Hierarchical Model Formulation	367
16.2.4	Restrictions to Remove Ambiguity in the Case of Only Two Components.	368
16.2.5	Mining for Subsets of Genes.	368
16.2.6	Creating Molecular Profiles.	370
16.3	R Software Extension: POE.	371
16.3.1	An Example of Using POE on Simulated Data.	371
16.3.2	Estimating Posterior Probability of Expression Using <code>poe.fit</code>	372
16.3.3	Visualization Tools.	374
16.3.4	Gene-Mining Functions.	377
16.3.5	Molecular Profiling Tool.	379
16.4	Results of POE Applied to Lung Cancer Data	381
16.5	Discussion and Future Work	384

17 Bayesian Decomposition 388

Michael F. Ochs

17.1	Introduction	388
17.1.1	Role of Signaling and Metabolic Pathways	388
17.1.2	Gene Expression Microarrays.	389
17.2	Methods.	390
17.2.1	Matrix Decomposition.	390
17.2.2	Markov Chain Monte Carlo.	391
17.2.3	Bayesian Framework.	392
17.2.4	The Prior Distribution.	393
17.2.5	Summary Statistics.	395
17.3	Software.	396
17.3.1	Implementation.	396
17.3.2	Files and Installation.	396
17.3.3	Issues in the Application of Bayesian Decomposition	397
17.4	Application of Bayesian Decomposition to Yeast Cell Cycle Data.	398
17.4.1	Preparation of the Data	398
17.4.2	Running the Program	399
17.4.3	Visualizing the Output	400

17.4.4	Interpretation	402
17.4.5	Advantages of Bayesian Decomposition.	403
17.5	Discussion	403
18	Bayesian Clustering of Gene Expression Dynamics	409
	<i>Paola Sebastiani, Marco Ramoni, and Isaac S. Kohane</i>	
18.1	Introduction	409
18.2	Methods	411
18.2.1	Modeling Time	412
18.2.2	Probabilistic Scoring Metric	413
18.2.3	Heuristic Search	415
18.2.4	Statistical Diagnostics	416
18.3	Software	417
18.3.1	Screen 0: Welcome Screen	417
18.3.2	Screen 1: Getting Started	418
18.3.3	Screen 2: Analysis	418
18.3.4	Screen 3: Cluster Model	419
18.3.5	Screen 4: Pack and Go!	419
18.4	Application	420
18.4.1	Analysis	420
18.4.2	Statistical Diagnostics	421
18.4.3	Understanding the Model	421
18.5	Conclusions	424
19	Relevance Networks: A First Step Toward Finding Genetic Regulatory Networks Within Microarray Data	428
	<i>Atul J. Butte and Isaac S. Kohane</i>	
19.1	Introduction	428
19.1.1	Advantages of Relevance Networks.	429
19.2	Methodology.	431
19.2.1	Formal Definition of Relevance Networks	431
19.2.2	Finding Regulatory Networks in Phenotypic Data	432
19.2.3	Using Entropy and Mutual Information to Evaluate Gene-Gene Associations.	434
19.3	Applications	437
19.3.1	Finding Pharmacogenomic Regulatory Networks	437
19.3.2	Setting the Threshold	439
19.4	Software	440
	Index	447