

Contents

I Bioinformatics for Life and Personal Genome Interpretation

1	Bioinformatics for Life	3
1.1	Introduction	4
1.2	Life and Information	5
1.3	The Human Genome Project	8
1.4	Development of Microarray Technology and its Medical Applications	9
1.5	Explosion of Data from Next-Generation Sequencing	12
1.6	The Era of Systems Biology and Biomedical Informatics	14
	Bibliography	15
2	Next-Generation Sequencing Technology and Personal Genome Data Analysis	17
2.1	Introduction	18
2.2	Data Format	20
2.2.1	FASTQ Format.....	20
2.2.2	CSFASTA Format	21
2.2.3	Concept of QV	21
2.3	Alignment of NGS Sequencing Reads	22
2.3.1	Sequence Alignment with a Hash Table.....	23
2.3.2	Sequence Alignment with a Suffix Tree	23
2.4	SNP and InDel Detection	24
2.5	Annotation of Sequence Variation and Function Prediction	25
2.5.1	Annotation and Clinical Interpretation of Common Variations	25
2.5.2	Annotation and Clinical Interpretation of Rare Variants	26
2.5.3	KEGG Disease Pathways Mapping	27
2.5.4	Pharmacogenomics	28
2.5.5	Distribution of Genetic Variants in the Population	28
	Bibliography	30
3	Personal Genome Data Analysis	33
3.1	Prerequisites	34
3.2	SNP Annotations and Interpretations Using SNPedia and Promethease	34
3.2.1	Create Input File	34
3.2.2	Running Promethease.....	35
3.2.3	Results Format and Interpretations of Promethease.....	35
3.3	Interpretation of the Correlation of Rare SNPs with Diseases Using SIFT and KEGG	37
3.3.1	Run the SIFT Web Tool.....	38
3.3.2	KEGG DISEASE Pathway Mapping	38
3.4	Pharmacogenomic Analysis	41
3.4.1	Pharmacogenomic Analysis Using GENotation.....	41
3.5	Calculations of Allele Frequencies in a Population	43
	Bibliography	45

4	Personal Genome Interpretation and Disease Risk Prediction.....	47
4.1	Introduction	48
4.2	SNP Prioritization	48
4.3	Prediction of Genetic Disease Risk	49
4.3.1	Algorithm for Prediction of Genetic Disease Risk.....	49
4.3.2	Practice Data	49
4.3.3	Risk Prediction Using GENOTation.....	50
4.3.4	Data Analysis Using Promethease	52
4.4	Resources for Analyzing an Individual Genome.....	55
4.4.1	dbGAP	55
4.4.2	SNPedia	57
4.4.3	PheGenl.....	58
4.5	Disease-Related Sequence Variation Analysis.....	62
4.5.1	Verification of File List and Moving to Practice Directory Using Terminal.....	62
4.5.2	Extracting Main Disease-Causing Variants from Sequence Data	63
4.5.3	Finding the Disease Gene of a Child Using Kinship Data.....	72
4.5.4	Analysis of Gene Lists Known to Affect Phenotype	72
	Bibliography	75

II Advanced Microarray Data Analysis

5	Advanced Microarray Data Analysis	79
5.1	Introduction	80
5.2	Microarray Experiment.....	81
5.3	Structure and Normalization of Microarray Data.....	82
5.4	Differentially Expressed Genes (DEGs).....	84
5.5	Cluster Analysis and Interpretation.....	85
5.6	Classification Analysis.....	86
5.6.1	Linear Discriminant Analysis (LDA)	87
5.6.2	Support Vector Machine (SVM)	87
5.6.3	K-Nearest Neighborhood (KNN).....	88
5.7	Gene Set Enrichment Analysis.....	88
5.7.1	Analysis Method and Feature	89
5.7.2	Gene Set Database	90
5.8	Survival Analysis and Prognostic Subgroup Prediction	90
	Bibliography	92
6	Gene Expression Data Analysis.....	95
6.1	Introduction	96
6.2	Prerequisites.....	96
6.3	Differential Gene Expression Analysis.....	103
6.3.1	Analysis of Significant Differential Expression in Two-Group Comparisons: T-Test and SAM.....	104
6.3.2	Analysis of Significant Differential Expression in Comparisons of More than Three Groups: ANOVA	109

6.4	Clustering Analysis	111
6.4.1	Hierarchical Clustering.....	111
6.4.2	K-Means Clustering.....	113
6.4.3	SOM Clustering.....	114
6.5	Classification Analysis	115
6.5.1	LDA (Linear Discriminant Analysis).....	116
6.5.2	KNN (K-Nearest Neighbor).....	116
6.5.3	SVM (Support Vector Machine).....	116
6.6	Basic Data Processing in R Program	117
6.6.1	Understanding the Data Structure.....	117
	Bibliography	120
7	Gene Ontology and Biological Pathway-Based Analysis	121
7.1	Introduction	122
7.2	Prerequisites	124
7.3	Dataset and Biological Interpretation Tools	124
7.3.1	Generate Example Data Set.....	125
7.3.2	Biological Interpretation Tools.....	128
7.4	DAVID	129
7.5	ArrayXPath	131
7.6	BioLattice	131
	Bibliography	134
8	Gene Set Approaches and Prognostic Subgroup Prediction	135
8.1	Introduction	137
8.2	Prerequisites	139
8.3	Input Files	139
8.3.1	Gene Expression Data File Format.....	140
8.3.2	Phenotype Label File Format.....	141
8.3.3	Gene Set Database File.....	141
8.3.4	Microarray Chip Annotation File.....	142
8.3.5	Data Input.....	142
8.4	GSEA Execution	143
8.4.1	Required Fields.....	143
8.4.2	Basic Fields.....	145
8.4.3	GSEA Execution.....	145
8.4.4	GSEA Analysis Results.....	145
8.5	Leading Edge Subset Analysis	146
8.6	GSEA Analysis Via R	147
8.6.1	Installation.....	148
8.6.2	Input File.....	149
8.6.3	GSEA Execution.....	149
8.7	Survival Analysis	152
8.7.1	Install CGDS-R Package.....	152
8.7.2	Get List of Cancer Studies from Server.....	153
8.7.3	Extract Samples and Features.....	153
8.7.4	Get Mutation Profiles for BRCA1 and BRCA2.....	154

8.7.5	Extraction Samples with BRCA1 and BRCA2 Methylation	154
8.7.6	Clinical Data Integration.....	155
8.7.7	Survival Analysis.....	156
	Bibliography	157
9	MicroRNA Data Analysis	159
9.1	Introduction	160
9.2	Prerequisites	160
9.3	Retrieving miRNA-mRNA Pair Expression Data	160
9.3.1	Input File.....	160
9.3.2	Missing Value Removal	161
9.4	miRNA-mRNA Expression Correlation Coefficient	162
9.4.1	Finding One miRNA and mRNA Correlation.....	162
9.4.2	Finding the Correlation Coefficient of All Pair-Wise mRNAs of Several Genes.....	163
9.4.3	Find Significant miRNA-mRNA Pairs	165
9.4.4	Draw Significant miRNA-mRNA Plot.....	166
9.5	Verifying a Significant miRNA-mRNA Correlations in Information Databases	167
9.5.1	Verify mRNA Target Against Verified Experiments	167
9.5.2	Confirm Predicted Target Gene.....	169
	Bibliography	171

**III Network Biology, Sequence, Pathway
and Ontology Informatics**

10	Network Biology, Sequence, Pathway and Ontology Informatics	175
10.1	Introduction	176
10.2	Sequence Data Analysis	177
10.2.1	Sequence Information.....	177
10.2.2	Pair-wise Alignment.....	178
10.2.3	Multiple Alignment.....	178
10.2.4	Global Alignment.....	178
10.2.5	Local Alignment	179
10.3	Phylogenetic Tree Analysis	179
10.4	Visualization for Sequence Analysis	180
10.5	Biological Pathway Analysis	181
10.6	Gene Ontology	182
10.7	Biomedical Text Mining	183
10.8	Biomedical Network Analysis	184
	Bibliography	186
11	Motif and Regulatory Sequence Analysis	189
11.1	Introduction	190
11.2	Sequence Data Structure	190
11.3	Sequence Alignment and Phylogenetic Tree Analysis	193
11.3.1	Sequence Alignment.....	193
11.3.2	Sequence Alignment Practice.....	196
11.3.3	Phylogenetic Tree Analysis	199

11.4	Transcription Regulatory Site Prediction Using Sequence Alignments	202
11.4.1	Transcription Factors and Sequence Motifs	202
11.4.2	Conserved Sequence Region Detection.....	202
11.5	UCSC Genome Browser	206
11.6	Prediction of Targeting microRNAs	209
	Bibliography	211
12	Molecular Pathways and Gene Ontology	213
12.1	Introduction	214
12.2	Prerequisites.....	214
12.3	Gene Ontology	214
12.3.1	Search GO Annotation for a Single Gene.....	214
12.3.2	Search GO Annotations for a Gene List.....	216
12.3.3	Calculation of the Semantic Distance Between Genes.....	217
12.3.4	GO Annotation Analysis Using R.....	218
12.4	Biological Pathway Analysis	224
12.4.1	Search a Biological Pathway of a Specific Gene	224
12.4.2	Search Significant Biological Pathways of a Gene List	224
12.4.3	Data Analysis Using Gene Expression Data	225
12.4.4	Biological Pathway Analysis Using R.....	226
12.5	Biological Text Mining – COREMINE	229
	Bibliography	231
13	Biological Network Analysis.....	233
13.1	Introduction	234
13.2	Preparations	234
13.3	The Network Analysis Tools	234
13.3.1	Major Network Analysis Tools	234
13.3.2	Introduction to igraph and an Example of Its Use	235
13.4	Introduction to Data and Publication Used for Analysis	240
13.4.1	Introduction to Publication and Dataset.....	240
13.4.2	Introduction to Data and Preprocessing	240
13.5	Analysis of Protein Interaction Networks.....	241
13.5.1	Visualization	241
13.5.2	Distribution of Connections	242
13.5.3	Evolutionary Analysis	244
	Bibliography	246

IV SNPS, GWAS and CNVs, Informatics for Genome Variants

14	SNPs, GWAS, CNVs: Informatics for Human Genome Variations	249
14.1	Introduction	250
14.2	dbSNP	252
14.3	International HapMap Project	253
14.4	PharmGKB	253

14.5	Genome-Wide Association Studies (GWAS)	254
14.5.1	Allelic Test.....	254
14.5.2	Genotypic Test	255
14.5.3	Multiple Testing Correction	256
14.6	Definition and Importance of Copy-Number Variation	256
14.7	Analysis Methods of CNV	257
14.7.1	Chip Method to Detect CNV.....	258
14.7.2	Sequencing Method to Detect CNV.....	258
14.8	Conclusion	258
	Bibliography	259
15	SNP Data Analysis	261
15.1	Introduction	262
15.1.1	dbSNP	262
15.1.2	International HapMap Project	262
15.1.3	PharmGKB	262
15.2	dbSNP	263
15.2.1	rsID, ssID Search	263
15.2.2	Entrez SNP Search.....	264
15.3	1000 Genomes Project	266
15.3.1	Data	266
15.3.2	Browser	268
15.4	PharmGKB	270
15.4.1	Search PharmGKB.....	270
15.4.2	Clinical Annotations.....	270
15.5	Variant Analysis Using NGS	271
15.5.1	Preparation	271
15.5.2	Extraction of Disease-Associated Variants from Variants Identified Through NGS	272
	Bibliography	279
16	GWAS Data Analysis	281
16.1	Introduction	282
16.2	Prerequisites	282
16.3	*.ped and *.map Files of the PLINK	283
16.3.1	The genotypes.ped File.....	283
16.3.2	The genotypes.map File	284
16.4	Configuration of gPLINK and Other Programs that Work with gPLINK	285
16.5	Validation of the GWAS Data Set and Summary of Statistics	285
16.6	Filtering Out Data with a Threshold	288
16.7	Basic Association Test	291
16.8	Additive Genotypic Test	292
16.9	Manhattan Plot	294
	Bibliography	297
17	CNV Analysis	299
17.1	Introduction	300
17.2	Prerequisites	300

17.3	Data	301
17.3.1	Normalization	301
17.4	Genomic Alteration Detection	302
17.4.1	Single & Multi-sample Segmentation and Allele-specific Segmentation.....	302
17.4.2	Identification of Gain and Loss at Genomic Position.....	304
17.5	Visualization	308
17.5.1	Analysis of the Differences Between Samples Using the Heatmap	308
17.6	Obtaining Genomic Regions	309
	Bibliography	312

V Metagenome and Epigenome, Basic Data Analysis

18	Metagenome and Epigenome Data Analysis	315
18.1	Metagenome	316
18.2	Epigenome	318
18.2.1	DNA Methylation	318
18.2.2	Histone Modification	319
18.2.3	Non-coding RNA (ncRNA)	320
18.3	Epigenome Databases and Analysis Tools	320
18.3.1	DNA Methylation Databases and Analysis Tools.....	320
18.3.2	Histone Modification Databases and Analysis Tools	321
18.3.3	Noncoding RNA Databases and Analysis Tools	321
18.4	Epigenome Analysis	322
	Bibliography	322
19	Metagenome Data Analysis	325
19.1	Introduction	326
19.2	Prerequisites	326
19.3	Metagenome Analysis Tool	326
19.3.1	Major Metagenome Analysis Tools.....	326
19.3.2	Introduction to metagenomeSeq	327
19.4	Metagenome Analysis	327
19.4.1	Dataset	327
19.4.2	Basic Example of metagenomeSeq.....	327
19.4.3	Statistical Testing	329
19.4.4	Aggregating Counts.....	329
19.5	Visualization	330
	Bibliography	337
20	Epigenome Database and Analysis Tools	339
20.1	Introduction	340
20.2	Prerequisites	340
20.3	Epigenome Database	340
20.3.1	dbEM: A Database of Epigenetic Modifiers	340
20.3.2	EpiFactors	341
20.3.3	MethylomeDB	343
20.3.4	NGSmethDB: High-Quality Methylomes and Differential Methylation	345

20.4	Epigenome Analysis Tool	346
20.4.1	Bsseq: Analyzing WGBS with the Bsseq Package.....	346
20.4.2	MethPipe: a Computational Pipeline for Analyzing Bisulfite Sequencing Data	351
20.4.3	PAVIS: Peak Annotation and Visualization.....	351
	Bibliography	352
21	Epigenome Data Analysis	353
21.1	Introduction	354
21.2	Preparations	354
21.3	Input Sequence Data	354
21.3.1	Download Sequence Data	354
21.3.2	Specify Input Sequence Data.....	355
21.4	Data Filtering	355
21.4.1	Alignment Control (AC).....	355
21.4.2	Quality Control (QC).....	356
21.5	Analysis of Methylation Status	356
21.5.1	Sequence Name	357
21.5.2	Alignment	357
21.5.3	Sequence Start and End Positions.....	358
21.5.4	Length of the Reference Sequence.....	358
21.5.5	Methylated Positions in the Sample Sequence.....	359
21.5.6	Methylated Positions in the Reference Sequence.....	360
21.6	Exploratory Analysis and Visualization	360
21.6.1	Plotting Methylation Status.....	360
21.6.2	Lollipop Figure: Methylation Status.....	360
21.6.3	Neighboring Co-occurrence Display	362
21.6.4	Distant Co-occurrence Display	364
21.7	Statistical Tests	364
21.7.1	Fisher’s Exact Test.....	364
21.7.2	Clustering Analysis	365
21.7.3	Correspondence Analysis.....	366
	Bibliography	367