
Contents

Preface	xv
About the Authors	xxi
1 Introduction to Genomics	1
1.1 Genes, DNA and central dogma	1
1.1.1 What is a genome?	1
1.1.2 What is a gene?	2
1.1.3 How are genes controlled? Transcriptional and post-transcriptional regulation	4
1.1.4 What does a gene look like?	5
1.2 Elements of gene regulation	6
1.2.1 Transcriptional regulation	6
1.2.2 Post-transcriptional regulation	11
1.3 Shaping the genome: DNA mutation	13
1.4 High-throughput experimental methods in genomics	15
1.4.1 The general idea behind high-throughput techniques	15
1.4.2 High-throughput sequencing	16
1.5 Visualization and data repositories for genomics	19
2 Introduction to R for Genomic Data Analysis	23
2.1 Steps of (genomic) data analysis	23
2.1.1 Data collection	24
2.1.2 Data quality check and cleaning	24
2.1.3 Data processing	24
2.1.4 Exploratory data analysis and modeling	24
2.1.5 Visualization and reporting	25
2.1.6 Why use R for genomics ?	25
2.2 Getting started with R	26
2.2.1 Installing packages	27
2.2.2 Installing packages in custom locations	28
2.2.3 Getting help on functions and packages	28
2.3 Computations in R	29
2.4 Data structures	30
2.4.1 Vectors	30

2.4.2	Matrices	31
2.4.3	Data frames	33
2.4.4	Lists	34
2.4.5	Factors	35
	Data types	36
	Reading and writing data	37
2.6.1	Reading large files	38
	Plotting in R with base graphics	39
2.7.1	Combining multiple plots	42
2.7.2	Saving plots	43
	Plotting in R with ggplot2	44
2.8.1	Combining multiple plots	46
2.8.2	ggplot2 and tidyverse	48
	Functions and control structures (for, if/else, etc.)	49
2.9.1	User-defined functions	49
2.9.2	Loops and looping structures in R	50
	Exercises	56
2.10.1	Computations in R	56
2.10.2	Data structures in R	56
2.10.3	Reading in and writing data out in R	59
2.10.4	Plotting in R	60
2.10.5	Functions and control structures (for, if/else, etc.)	63

Statistics for Genomics

67

	How to summarize collection of data points: The idea behind statistical distributions	67
3.1.1	Describing the central tendency: Mean and median	67
3.1.2	Describing the spread: Measurements of variation	69
3.1.3	Precision of estimates: Confidence intervals	74
	How to test for differences between samples	79
3.2.1	Randomization-based testing for difference of the means	79
3.2.2	Using t-test for difference of the means between two samples	80
3.2.3	Multiple testing correction	83
3.2.4	Moderated t-tests: Using information from multiple comparisons	86
	Relationship between variables: Linear models and correlation	89
3.3.1	How to fit a line	92
3.3.2	How to estimate the error of the coefficients	96
3.3.3	Accuracy of the model	99
3.3.4	Regression with categorical variables	102
3.3.5	Regression pitfalls	104

3.4	Exercises	106
3.4.1	How to summarize collection of data points: The idea behind statistical distributions	106
3.4.2	How to test for differences in samples	107
3.4.3	Relationship between variables: Linear models and correlation	108

4 Exploratory Data Analysis with Unsupervised Machine Learning

111

4.1	Clustering: Grouping samples based on their similarity	111
4.1.1	Distance metrics	111
4.1.2	Hierarchical clustering	115
4.1.3	K-means clustering	119
4.1.4	How to choose “k”, the number of clusters	122
4.2	Dimensionality reduction techniques: Visualizing complex data sets in 2D	127
4.2.1	Principal component analysis	127
4.2.2	Other matrix factorization methods for dimensionality reduction	134
4.2.3	Multi-dimensional scaling	139
4.2.4	t-Distributed Stochastic Neighbor Embedding (t-SNE)	140
4.3	Exercises	144
4.3.1	Clustering	144
4.3.2	Dimension reduction	145

5 Predictive Modeling with Supervised Machine Learning

147

5.1	How are machine learning models fit?	148
5.1.1	Machine learning vs. statistics	149
5.2	Steps in supervised machine learning	149
5.3	Use case: Disease subtype from genomics data	150
5.4	Data preprocessing	151
5.4.1	Data transformation	152
5.4.2	Filtering data and scaling	154
5.4.3	Dealing with missing values	155
5.5	Splitting the data	156
5.5.1	Holdout test dataset	156
5.5.2	Cross-validation	157
5.5.3	Bootstrap resampling	158
5.6	Predicting the subtype with k-nearest neighbors	158
5.7	Assessing the performance of our model	159
5.7.1	Receiver Operating Characteristic (ROC) curves	162
5.8	Model tuning and avoiding overfitting	164
5.8.1	Model complexity and bias variance trade-off	167

5.8.2	Data split strategies for model tuning and testing	170
5.9	Variable importance	172
5.10	How to deal with class imbalance	174
5.10.1	Sampling for class balance	174
5.10.2	Altering case weights	175
5.10.3	Selecting different classification score cutoffs	175
5.11	Dealing with correlated predictors	175
5.12	Trees and forests: Random forests in action	176
5.12.1	Decision trees	176
5.12.2	Trees to forests	177
5.12.3	Variable importance	179
5.13	Logistic regression and regularization	180
5.13.1	Regularization in order to avoid overfitting	184
5.13.2	Variable importance	186
5.14	Other supervised algorithms	188
5.14.1	Gradient boosting	188
5.14.2	Support Vector Machines (SVM)	190
5.14.3	Neural networks and deep versions of it	193
5.14.4	Ensemble learning	195
5.15	Predicting continuous variables: Regression with machine learning	196
5.15.1	Use case: Predicting age from DNA methylation	196
5.15.2	Reading and processing the data	197
5.15.3	Running random forest regression	198
5.16	Exercises	200
5.16.1	Classification	200
5.16.2	Regression	201

Operations on Genomic Intervals and Genome Arithmetic 203

6.1	Operations on genomic intervals with GenomicRanges package	204
6.1.1	How to create and manipulate a GRanges object	204
6.1.2	Getting genomic regions into R as GRanges objects	207
6.1.3	Finding regions that do/do not overlap with another set of regions	210
6.2	Dealing with mapped high-throughput sequencing reads	214
6.2.1	Counting mapped reads for a set of regions	214
6.3	Dealing with continuous scores over the genome	215
6.3.1	Extracting subsections of Rle and RleList objects	218
6.4	Genomic intervals with more information: SummarizedExperiment class	221
6.4.1	Create a SummarizedExperiment object	221
6.4.2	Subset and manipulate the SummarizedExperiment object	222
6.5	Visualizing and summarizing genomic intervals	225

6.5.1	Visualizing intervals on a locus of interest	225
6.5.2	Summaries of genomic intervals on multiple loci	226
6.5.3	Making karyograms and circos plots	230
6.6	Exercises	234
6.6.1	Operations on genomic intervals with the GenomicRanges package	234
6.6.2	Dealing with mapped high-throughput sequencing reads	235
6.6.3	Dealing with contiguous scores over the genome	235
6.6.4	Visualizing and summarizing genomic intervals	235

7 Quality Check, Processing and Alignment of High-throughput Sequencing Reads 237

7.1	FASTA and FASTQ formats	237
7.2	Quality check on sequencing reads	239
7.2.1	Sequence quality per base/cycle	239
7.2.2	Sequence content per base/cycle	240
7.2.3	Read frequency plot	241
7.2.4	Other quality metrics and QC tools	241
7.3	Filtering and trimming reads	243
7.4	Mapping/aligning reads to the genome	246
7.5	Further processing of aligned reads	248
7.6	Exercises	248

8 RNA-seq Analysis 251

8.1	What is gene expression?	251
8.2	Methods to detect gene expression	252
8.3	Gene expression analysis using high-throughput sequencing technologies	252
8.3.1	Processing raw data	253
8.3.2	Alignment	253
8.3.3	Quantification	254
8.3.4	Within sample normalization of the read counts	255
8.3.5	Computing different normalization schemes in R	256
8.3.6	Exploratory analysis of the read count table	259
8.3.7	Differential expression analysis	264
8.3.8	Functional enrichment analysis	273
8.3.9	Accounting for additional sources of variation	277
8.4	Other applications of RNA-seq	290
8.5	Exercises	291
8.5.1	Exploring the count tables	291
8.5.2	Differential expression analysis	292
8.5.3	Functional enrichment analysis	292

8.5.4	Removing unwanted variation from the expression data	293
seq analysis		295
	Regulatory protein-DNA interactions	295
	Measuring protein-DNA interactions with ChIP-seq	296
	Factors that affect ChIP-seq experiment and analysis quality	298
9.3.1	Antibody specificity	298
9.3.2	Sequencing depth	298
9.3.3	PCR duplication	299
9.3.4	Biological replicates	299
9.3.5	Control experiments	299
9.3.6	Using tagged proteins	301
	Pre-processing ChIP data	301
9.4.1	Mapping of ChIP-seq data	301
	ChIP quality control	303
9.5.1	The data	303
9.5.2	Sample clustering	304
9.5.3	Visualization in the genome browser	308
9.5.4	Plus and minus strand cross-correlation	312
9.5.5	GC bias quantification	316
9.5.6	Sequence read genomic distribution	320
	Peak calling	327
9.6.1	Types of ChIP-seq experiments	327
9.6.2	Peak calling: Sharp peaks	332
9.6.3	Peak calling: Broad regions	341
9.6.4	Peak quality control	344
9.6.5	Peak annotation	356
	Motif discovery	359
9.7.1	Motif comparison	361
	What to do next?	363
	Exercises	364
9.9.1	Quality control	364
methylation analysis using bisulfite sequencing data		367
	What is DNA methylation?	367
10.1.1	How DNA methylation is set?	368
10.1.2	How to measure DNA methylation with bisulfite sequencing	368
	Analyzing DNA methylation data	368
	Processing raw data and getting data into R	369
	Data filtering and exploratory analysis	370
10.4.1	Reading methylation call files	370

10.4.2	Further quality check	372
10.4.3	Merging samples into a single table	373
10.4.4	Filtering CpGs	374
10.4.5	Clustering samples	376
10.4.6	Principal component analysis	378
10.5	Extracting interesting regions: Differential methylation and segmentation	379
10.5.1	Differential methylation	379
10.5.2	Methylation segmentation	384
10.5.3	Working with large files	387
10.6	Annotation of DMRs/DMCs and segments	388
10.6.1	Further annotation with genes or gene sets	390
10.7	Other R packages that can be used for methylation analysis	390
10.8	Exercises	390
10.8.1	Differential methylation	390
10.8.2	Methylome segmentation	391
11	Multi-omics Analysis	393
11.1	Use case: Multi-omics data from colorectal cancer	393
11.2	Latent variable models for multi-omics integration	399
11.3	Matrix factorization methods for unsupervised multi-omics data integration	400
11.3.1	Multiple factor analysis	401
11.3.2	Joint non-negative matrix factorization	404
11.3.3	iCluster	409
11.4	Clustering using latent factors	413
11.4.1	One-hot clustering	414
11.4.2	K-means clustering	415
11.5	Biological interpretation of latent factors	416
11.5.1	Inspection of feature weights in loading vectors	416
11.5.2	Making sense of factors using enrichment analysis	418
11.5.3	Interpretation using additional covariates	420
11.6	Exercises	422
11.6.1	Matrix factorization methods	422
11.6.2	Clustering using latent factors	423
11.6.3	Biological interpretation of latent factors	423
Bibliography		425
Index		437