

Molecular Data Analysis Using R

Csaba Ortutay

Zsuzsanna Ortutay

WILEY Blackwell

Contents

- Foreword, xiii
- Preface, xv
- Acknowledgements, xix
- About the Companion Website, xxi
- 1 Introduction to R statistical environment, 1**
 - Why R?, 1
 - Installing R, 2*
 - Interacting with R, 2
 - Graphical interfaces and integrated development environment (IDE) integration, 3*
 - Scripting and sourcing, 3*
 - The R history and the R environment file, 4*
 - Packages and package repositories, 4
 - Comprehensive R Archive Network, 5*
 - Bioconductor, 6*
 - Working with data, 7
 - Basic operations in R, 8
 - Some basics of graphics in R, 10
 - Getting help in R, 12
 - Files for practicing, 13
 - Study exercises and questions, 14
 - References, 14
 - Webliography, 15
- 2 Simple sequence analysis, 17**
 - Sequence files, 17
 - FASTA sequence format, 18*
 - GenBank flat file format, 19*
 - Reading sequence files into R, 20
 - Obtaining sequences from remote databases, 21
 - Seqinr package, 21*
 - Ape package, 22*
 - Descriptive statistics of nucleotide sequences, 24
 - Descriptive statistics of proteins, 28
 - Aligned sequences, 31
 - Visualization of genes and transcripts in a professional way, 34

- Files for practicing, 37
- Study exercises and questions, 38
- References, 38
- Webliography, 39
- Packages, 40
- 3** Annotating gene groups, 41
 - Enrichment analysis: an overview, 41
 - Overview of two different methods, 41*
 - Enrichment analysis results, 42*
 - Common aspects of the two different approaches, 43*
 - Overrepresentation analysis, 46
 - Hypergeometric test using GOstats, 47*
 - ORA analysis using topGO, 48*
 - Enrichment analysis of microarray sets with topGO, 51*
 - Gene set enrichment analysis, 52
 - GSEA with R, 56*
 - Files for practicing, 61
 - Study exercises and questions, 61
 - References, 62
 - Webliography, 62
 - Packages, 63
- 4** Next-generation sequencing: introduction and genomic applications, 65
 - High-throughput sequencing background, 65
 - Experimental background, 66*
 - Single-end and paired-end sequencing reads, 67*
 - Assemble reads, 69*
 - How many reads? Depth of coverage, 71*
 - Storing data in files, 72
 - FASTQ, 72*
 - SAM and BAM files, 76*
 - Variant call format files, 77*
 - General data analysis workflow, 77
 - Data processing considerations, 78*
 - Quality checking and screening read sequences, 80
 - Quality checking for one file, 82*
 - Quality inspection for multiple files in a project, 82*
 - Quality filtering of FASTQ files, 83*
 - Handling alignment files and genomic variants, 84
 - Alignment and variation visualization, 88*
 - Simple handling of VCF files, 89*
 - Genomic applications: low- and medium-depth sequencing, 91
 - Aneuploidy sequencing and copy number variation identification, 92*

- SNP identification and validation, 92*
- Exome sequencing, 93*
- Genomic region resequencing, 93*
- Full genome and metagenome sequencing, 94*
- Files for practicing, 94
- Study exercises and questions, 94
- References, 95
- Webliography, 97
- Packages, 97
- 5** Quantitative transcriptomics: qRT-PCR, 99
 - Transcriptome, 99
 - Polymerase chain reaction, 100*
 - Standards for qPCR, 102*
 - R packages, 104*
 - Understanding delta Ct, 104
 - Calculation of delta Ct, 105*
 - Requirements for real delta Ct calculations, 107*
 - Absolute quantification, 110
 - Value prediction, the professional way, 114*
 - Relative quantification using the ddCt method, 115
 - Comparison of two conditions, 116*
 - Comparison of multiple experimental conditions, 118*
 - Quality control with melting curve, 121
 - Files for practicing, 123
 - Study exercises and questions, 123
 - References, 123
 - Webliography, 124
 - Packages, 124
- 6** Advanced transcriptomics: gene expression microarrays, 125
 - Microarray analysis: probes and samples, 125
 - Experimental background, 126*
 - Archiving and publishing microarray data, 128
 - Minimum information standard, 128*
 - Data preprocessing, 128
 - Accessing data from CEL files, 129*
 - Quality control, 131*
 - Normalization, 132*
 - Differential gene expression, 133
 - Annotating results, 136*
 - Creating normalized expression set from Illumina data, 138
 - Automated data access from GEO, 140
 - Files for practicing, 142

- Study exercises and questions, 142
- References, 143
- Webliography, 144
- Packages, 144
- 7** Next-generation sequencing in transcriptomics: RNA-seq experiments, 145
 - High-throughput RNA sequencing background, 145
 - Experimental background, 145*
 - RNA-seq applications, 146*
 - Differential expression with different resolutions, 147*
 - Preparing count tables, 148
 - Alignment files to read counts, 148*
 - Differential expression in simple comparison, 151*
 - A naive t-test approach, 151*
 - Single factor analysis with edgeR, 153*
 - Differential expression with DESeq, 156*
 - Complex experimental arrangements, 159
 - Experimental factors and design matrix, 160*
 - GLM with edgeR, 161*
 - GLMs with DESeq, 162*
 - Heatmap visualization, 163*
 - Files for practicing, 164
 - Study exercises and questions, 164
 - References, 165
 - Webliography, 166
 - Packages, 166
- 8** Deciphering the regulome: from ChIP to ChIP-seq, 167
 - Chromatin immunoprecipitation, 167
 - Experimental background, 168*
 - Fragment analysis, 168*
 - ChIP data in ENCODE, 169*
 - ChIP with tiling microarrays, 169
 - High-throughput sequencing of ChIP fragments, 176
 - Connecting annotation to peaks, 181*
 - Analysis of binding site motifs, 182
 - Files for practicing, 186
 - Study exercises and questions, 187
 - References, 187
 - Webliography, 188
 - Packages, 189
- 9** Inferring regulatory and other networks from gene expression data, 191
 - Gene regulatory networks, 191
 - Data for gene network inference, 192*

- Reconstruction of co-expression networks, 193
- Gene regulatory network inference focusing of master regulators, 201
- Integrated interpretation of genes with GeneAnswers, 207
- Files for practicing, 211
- Study exercises and questions, 212
- References, 213
- Packages, 214
- 10** Analysis of biological networks, 215
 - A gentle introduction to networks, 215
 - Networks and their components and features, 215*
 - Random networks, 220*
 - Biological networks, 221*
 - Files for storing network information, 223
 - Important network metrics in biology, 227
 - Distance-based measures, 228*
 - Degree and related measures, 230*
 - Vulnerability, 231*
 - Community structure of a network, 234*
 - Graph visualization, 236
 - Cytoscape, 240*
 - Files for practicing, 241
 - Study exercises and questions, 241
 - References, 242
 - Webliography, 243
 - Packages, 243
- 11** Proteomics: mass spectrometry, 245
 - Mass spectrometry and proteomics: why and how?, 245
 - File formats for MS data, 246
 - Accessing the raw data of published studies, 247*
 - Identification of peptides in the samples, 249
 - Peptide mass fingerprinting, 249*
 - Peptide identification by using MS/MS spectra, 250*
 - Quantitative proteomics, 254
 - Getting protein-specific annotation, 258
 - Files for practicing, 259
 - Study exercises and questions, 259
 - References, 259
 - Webliography, 260
 - Packages, 260
- 12** Measuring protein abundance with ELISA, 261
 - Enzyme-linked immunosorbent assays, 261
 - Accessing ELISA data, 264*

Concentration calculation with a standard curve, 264

Preparing reference data, 267

Fitting linear model, 268

Fitting of a logistic model, 269

Concentration calculations by employing models, 270

Comparative calculations using concentrations, 271

Files for practicing, 277

Study exercises and questions, 277

References, 277

Packages, 278

13 Flow cytometry: counting and sorting stained cells, 279

Theoretical aspects of flow cytometry, 279

Experiment types: diagnosis versus discovery, 280

Measurement arrangements, 281

Fluorescent dyes, 281

Tubes versus plates, 285

Instruments, 285

What about data?, 287

Files, 287

Workflows, 288

Data preprocessing, 289

Handling all samples together, 290

Compensation, 292

Quality assurance, 292

Using workflow objects and transformation, 296

Normalization, 298

Cell population identification, 299

Manual gating, 300

Automatic gating, 304

Relating cell populations to external variables, 305

Reporting results, 307

MIFlowCyt, 307

FlowRepository.org, 308

Files for practicing, 308

Study exercises and questions, 309

References, 309

Webliography, 310

Packages, 310

Glossary, 311

Index, 323