

Contents

Preface	ix
Introduction: why studying transcriptomics?	xi
1 A primer on data distributions and their visualisation	1
1.1 Stochastic processes and Poisson distribution	1
1.1.1 Gaussian and t-Student distributions	3
1.1.2 Parameters of a distribution	3
1.2 Representation of quantitative biological data	6
1.2.1 Violin plot	7
1.2.2 Scatter plot	7
1.3 Lists of genes and Venn diagrams	8
2 Next-generation RNA sequencing	11
2.1 Introduction	11
2.2 Advantages of RNA-seq	11
2.3 RNA purification	12
2.3.1 RNA quality assessment	13
2.3.2 Abundant RNA species	14
2.3.3 Tissue-specific abundant RNA species	15
2.4 Library preparation	17
2.4.1 RNA fragmentation	17
2.4.2 Reverse transcription	18
2.4.3 Addition of the adapters	21
2.4.4 Quality control	21
2.5 Library sequencing	22
2.5.1 Bridge amplification and sequencing by synthesis	22
2.5.2 Single-end or paired-end sequencing?	23
2.5.3 Choosing the right read length	23
2.6 Other applications of next-generation RNA sequencing .	24

3	RNA-seq raw data processing	27
3.1	Introduction	27
3.2	General quality assessment	27
3.2.1	The analysis of Kmer levels permits estimation of the presence of artifact sequences	30
3.3	Removal of artefacts	31
3.4	Mapping the reads to the reference genome	32
3.4.1	The Burrows-Wheeler transform	32
3.4.2	Application of the Burrows-Wheeler transform to genome mapping	35
3.4.3	Optimal storage of the suffixes index	38
3.4.4	Structure of a SAM file	39
3.5	Complexity and depth of the sequencing	40
3.5.1	The Negative Binomial distribution is commonly used to estimate the complexity of a library	40
3.5.2	Marginal value of additional sequencing	42
	Appendix 3.1 – Negative binomial derivation	43
4	Differentially expressed gene detection and analysis	45
4.1	Introduction	45
4.2	Counting genes in a dataset	45
4.3	Detection of differentially expressed genes	46
4.3.1	When is a difference significant?	49
4.3.2	Modelling the data distribution	50
4.3.3	Testing the negative binomial model	53
4.4	Testing alternative splicing	55
	Appendix 4.1 – Properties of TPM	57
	Appendix 4.2 – On errors and statistical tests	58
5	Unbiased clustering methods	59
5.1	DEGs analysis: the issue of complexity	59
5.2	Clustering	60
5.2.1	Hierarchical clustering	61
5.2.2	K-means clustering	65
5.2.3	Fuzzy C-means	68
5.3	Principal component analysis	70
5.3.1	An intuitive view of PCA	70
5.3.2	A computational approach to PCA	74
5.4	Multi dimensional scaling	78
5.5	Nonlinear multidimensional mapping	80
5.6	Self-organising maps	83

6	Knowledge-based clustering methods	85
6.1	Introduction	85
6.2	Testing for gene set overrepresentation	85
6.3	KEGG pathways	87
6.4	Gene ontology	89
6.5	Gene set enrichment analysis	94
	6.5.1 Non-parametric GSEA for multiple samples	94
	6.5.2 Generally applicable gene enrichment	96
	Appendix 6.1 – GO redundancy reduction: clustering of annotations	97
7	Network analysis	99
7.1	Introduction	99
7.2	Biological networks	101
7.3	A primer on graph theory	101
	7.3.1 Algebraic graph theory offers some powerful tools to analyse graphs... and biological networks	102
	7.3.2 Topological properties of a graph	104
7.4	Weighted gene coexpression network analysis	108
	7.4.1 Module analysis	109
7.5	In conclusion: an explained example of the power of (neuro)genomics analysis	113
8	Mesoscale transcriptome analysis	121
8.1	Introduction	121
8.2	A comprehensive dataset of the human brain transcriptome	122
8.3	Evolutionary biology	126
8.4	Systems biology	131
8.5	Molecular neurobiology	134
8.6	Brain networks	136
9	Microscale transcriptome analysis	141
9.1	Introduction	141
9.2	It's a TRAP!	142
	9.2.1 RNA-seq can be used to study local translation in developing and mature visual circuits	145
9.3	Use of FACS to obtain the transcriptome at the cell-population scale	153
	9.3.1 Network analysis of cell-population transcriptomes and novel cytological properties of neurons	154

9.4	Single-cell RNA-seq strategies	157
9.4.1	A day in single-cell RNA sequencing	159
9.4.2	A closer view on the taxonomy of mouse primary visual cortex as revealed by single-cell RNA-seq	160
9.5	Other applications of single-cell RNA-seq to the nervous system	167
	Conclusion: new tools, new challenges	168
	References	169
	Index	175