Contents

1 Introduction 3

1.1 What is transcriptomics about? 5

1.2 Measuring gene expression 8

1.2.1 Microarrays 8

1.2.2 RNA-seq 10

1.3 The role of statistics in transcriptomics 14

2 Motivation, Aims, and Contributions 19

2.1 Motivation 21

2.2 Aims 22

2.3 Main contributions 24

2.3.1 Journal papers 25

2.3.2 Conferences 27

2.3.3 Software 28

2.3.4 Courses 28

3 Variable selection for multifactorial genomic data 29

3.1 Introduction 31

3.2 Methods 33

3.2.1 The ASCA-genes framework 34

3.2.2 Variable selection strategies 37

3.2.3 Data simulation 44

3.2.4 Performance indicators 49

3.3 Results 50

3.3.1 Simulated data 50

3.3.2 Experimental data: Hypoxia 62

3.3.3 Other applications 68

3.4 Discussion 69

4 RNA-seq data quality control **73**

4.1 Introduction 75

4.2 Data 78

4.3 Quality control analysis 80

4.3.1 Biotype distribution 80

4.3.2 Sequencing depth and expression quantification 85

4.3.3 Sequencing biases 90

4.4 Normalization 93

4.5 Filtering out low-count features 95

4.5.1 CPM method 96

4.5.2 Wilcoxon test 97

4.5.3 Proportion test 97

4.5.4 Comparing filtering methods 98

4.6 Discussion 100

5 Differential expression in RNA-seq **105**

5.1 Introduction 107

5.2 Data 109

5.2.1 Experimental data 109

5.2.1.1 Experimental data with technical replicates . 109

5.2.1.2 Experimental data with biological replicates . 110

5.2.2 Simulated data 111

5.2.2.1 Simulated data with technical replication . . 111

5.2.2.2 Simulation algorithm for biological replication 111

5.3 Methods 117

5.3.1 NOISeq 117

5.3.1.1 NOISeq-real 119

5.3.1.2 NOISeq-sim 120
5.3.2 NOISeqBIO 121
5.3.3 Other differential expression methods 126
5.3.4 Data processing 128
5.3.5 Tools for performance assessment 129
5.3.5.1 Performance indicators 129
5.3.5.2 Precision-recall curves and false discovery rate plots 129
5.3.5.3 Box plots 129
5.4 Results 130
5.4.1 NOISeq performance 130
5.4.1.1 Comparison on simulated data 130
5.4.1.2 Comparison on experimental data 131
5.4.2 NOISeqBIO performance 136
5.4.2.1 Preliminary studies on simulated data 136
5.4.2.2 Comparison on simulated data 148
5.4.2.3 Results on experimental datasets 160
5.5 Discussion 164

6 General discussion and Conclusions 169
6.1 General discussion 171
6.2 Conclusions 175
6.3 Reach and relevance 177
6.4 Future research lines 178

Appendix: Quality Control Report 181

References 189