

4	Detection of Imprinting and Maternal Effects	55
4.1	Imprinting and maternal genotype effects – Two epigenetic factors	55
4.1.1	Imprinting effects on complex diseases	56
4.1.2	Maternal genotype effects on complex diseases	57
4.2	Confounding between imprinting and maternal effects	58
4.3	Evolving study designs	60
4.4	Methods for detecting imprinting and maternal effects using data from prospective studies	63
4.5	Methods for detecting imprinting and maternal effects using data from retrospective studies	64
4.5.1	Joint detection of imprinting and maternal genotype effects	65
4.5.2	Detection of imprinting assuming no maternal effect	68
4.5.3	Detection of maternal-child genotype interacting effect assuming no imprinting	70
4.6	Case studies	71
4.6.1	Case study 1 – Framingham Heart Study	72
4.6.2	Case study 2 – UK rheumatoid arthritis data	74
4.7	Software	75
4.8	Concluding remarks	76
5	Modeling and Analysis of Next-Generation Sequencing Data	77
5.1	Isolation, quality control and library preparation	78
5.2	Validation, pooling and normalization	80
5.3	Sequencing	81
5.3.1	Single-end vs. paired-end	81
5.3.2	Generations of sequencing technology	81
5.3.3	Various next-generation sequencing platforms	82
5.3.3.1	Illumina	82
5.3.3.2	SOLiD	83
5.3.3.3	Ion Torrent semiconductor sequencing	84
5.3.3.4	Pacific biosciences single molecule real-time sequencing	85
5.3.3.5	Nanopore technologies	87
5.3.3.6	Choosing a platform	87
5.4	Factors affecting NGS data accuracy	89
5.4.1	At the library preparation stage	89
5.4.2	At the sequencing stage	90
5.5	Applications of RNA-Seq	90
5.6	RNA-Seq data preprocessing and analysis	93
5.6.1	Base calling	93
5.6.2	Quality control and preprocessing of reads	95
5.6.2.1	Quality control	95
5.6.2.2	Preprocessing	96

5.6.3	Read alignment	97
5.6.4	Genome-guided transcriptome assembly and isoform finding	100
5.6.5	Quantification and comparison of expression levels	101
5.6.6	Normalization methods	103
5.6.7	Differential expression analysis	105
5.6.7.1	Binomial and Poisson-based approaches	105
5.6.7.2	Empirical Bayes approaches	108
5.6.7.3	Negative binomial-based approaches	108
5.6.8	Classification	112
5.6.8.1	Linear discriminant analysis	112
5.6.8.2	Support vector machine classifier	114
5.6.9	Further downstream analysis	116
6	Sequencing-Based DNA Methylation Data	119
6.1	DNA methylation	120
6.2	Evolving technologies for measuring DNA methylation	121
6.3	Methods for Detection of DMCs using BS-seq data	123
6.3.1	BS-seq data	123
6.3.2	Fisher's exact test	124
6.3.3	Logistic regression	126
6.3.4	Beta-binomial formulations	127
6.3.4.1	Parameter estimation.	128
6.3.4.2	Statistical inference – hypothesis testing	130
6.3.5	Smoothing	131
6.3.5.1	Smoothing as part of the beta-binomial procedures	131
6.3.5.2	BSmooth	132
6.4	Methods for detection of DMRs using BS-seq data	134
6.4.1	Rule-based procedure – follow-up on DMCs	134
6.4.2	Credible band procedure – a single step approach	134
6.4.3	Summary of methods for BS-seq data – Which methods to choose?	136
6.5	Methods for detection of DMRs using Cap-seq data	136
6.5.1	Cap-seq data	136
6.5.2	Direct methods	138
6.5.2.1	Quantification of methylation signals	138
6.5.2.2	Detection of DMRs	139
6.5.3	Two-step methods	140
6.5.3.1	Derivation of nucleotide-level data	140
6.5.3.2	Detection of DMRs	141
6.6	Case studies	142
6.6.1	Case study 1 – Detection of DMRs using BS-seq data	143
6.6.2	Case study 2 – Detection of DMRs using Cap-seq data	143
6.7	Software	144
6.8	Concluding remarks and statistical challenges	144

7	Modeling and Analysis of Spatial Chromatin Interactions	147
7.1	3D chromosome organization and spatial regulation	148
7.2	Evolving technologies for measuring long-range interaction	149
7.3	Methods for recapitulating 3D structures using Hi-C type data	151
7.3.1	Hi-C data	151
7.3.2	Non-parametric optimization-based methods	153
7.3.3	Model-based methods	154
7.4	Methods for detecting long-range interactions using ChIA-PET type data	156
7.4.1	ChIA-PET data	156
7.4.2	Detections of true chromatin interactions – individual pair analysis	158
7.4.3	Detections of true chromatin interactions – joint analysis of all pairs	159
7.4.4	Detections of pairs with differential chromatin interactions	161
7.5	Case studies	163
7.5.1	Case study 1 – Reconstruction of 3D structure from Hi-C data	164
7.5.2	Case study 2 – Detection of true chromatin interactions from ChIA-PET data	164
7.5.3	Case study 3 – Detection of differential chromatin interaction intensities from ChIA-PET data	166
7.6	Software	167
7.7	Concluding remarks, and statistical and computational challenge	168
8	Digital Improvement of Single Cell Hi-C Data	169
8.1	Sparsity of single cell Hi-C data	170
8.1.1	Digital improvement of data quality	171
8.1.2	Separating structural zeros from dropouts	171
8.2	Adaptation of scRNA imputation methods for improving scHi-C data quality	172
8.2.1	Preparation of scHi-C data for using scRNA methods	172
8.2.2	Machine-learning algorithms for imputation	173
8.2.3	A block algorithm to account for spatial correlations	173
8.3	Methods designed for improving Hi-C data quality	175
8.4	Self representation smoothing for imputation and structural zeros inference	176
8.4.1	SRS model and estimation	176
8.4.2	Inference for structural zeros	178
8.5	Bayesian modeling for identifying structural zeros and imputing dropouts	180
8.5.1	The HiCImpute model	180
8.5.2	Statistical inference	182

8.6	Case studies	182
8.6.1	Case study 1 – improved cell clustering with enhanced scHi-C data from scRNA imputation method	182
8.6.2	Case study 2 – cell clustering with enhanced data from methods proposed specifically for scHi-C analysis	183
8.6.3	Case study 3 – discovery of subtypes based on improved data	187
8.7	Software	187
8.8	Concluding remarks, and statisitcal and computational challenges	189
9	Metabolomics Data Preprocessing	191
9.1	Introduction	191
9.2	Data sets	193
9.2.1	HAPO metabolomics study	193
9.3	Technology platforms	194
9.4	Data formats	196
9.5	Peak identification and metabolite quantification for non-targeted data	196
9.5.1	Peak detection	197
9.5.2	Peak alignment	197
9.5.3	Peak deconvolution	199
9.5.4	Metabolite identification	200
9.6	Normalization	202
9.7	Experimental design	208
10	Metabolomics Data Analysis	211
10.1	Per-metabolite analyses	212
10.1.1	Redundant metabolite measurements	212
10.1.2	Descriptive statistics	214
10.1.3	Individual tests of metabolite association	216
10.1.4	Missing data	217
10.1.5	Multiple comparisons adjustment	218
10.2	Multivariate approaches	219
10.3	Pathway enrichment analyses	220
10.4	Network analyses	222
10.4.1	Constructing and describing networks	223
10.4.2	Metabolic subnetwork optimization	223
10.4.3	Local community detection	225
10.4.4	Differential metabolic networks	227
10.5	Case studies on data integration	229
10.5.1	Metabolomics and genetic variants	229
10.5.2	Metabolomics and transcriptomics	230

11 Appendix	233
11.1 Basics of probability	233
11.2 Random variables and probability distributions	245
11.3 Basics of stochastic processes	260
11.4 Hidden Markov models	264
11.5 Frequentist statistical inference	268
11.6 Bayesian inference	274
 Bibliography	 283
 Index	 335