

- 18.6 Single Nucleotide Polymorphism, Sequence and Omic Data 547  
References 548

## Volume 2

- List of Contributors xxiii  
Editors' Preface to the Fourth Edition xxvii  
Glossary xxix  
Abbreviations and Acronyms xxxix

- 19 Ethical Issues in Statistical Genetics 551  
*Susan E. Wallace and Richard Ashcroft*
- 19.1 Introduction 551
- 19.1.1 What Is Ethics? 552
- 19.1.2 Models for Analysing the Ethics of Population Genetic Research 553
- 19.2 Ethics and Governance in Population Genetics Research: Two Case Studies 554
- 19.2.1 'Healthy Volunteer' Longitudinal Cohort Studies: UK Biobank 555
- 19.2.2 Precision Medicine Approaches: 100,000 Genomes Project 556
- 19.2.3 The Scientific and Clinical Value of the Research 556
- 19.2.4 Recruitment of Participants 558
- 19.2.5 Consent 559
- 19.2.6 Returning Individual Genetic Research Results 563
- 19.2.7 Confidentiality and Security 564
- 19.3 Stewardship and Wider Social Issues 565
- 19.3.1 Benefit Sharing 566
- 19.3.2 Community Involvement and Public Engagement\* 567
- 19.3.3 Race, Ethnicity and Genetics 567
- 19.4 Conclusion 568  
Acknowledgements 568  
References 568
- 20 Descent-Based Gene Mapping in Pedigrees and Populations 573  
*E.A. Thompson*
- 20.1 Introduction to Genetic Mapping and Genome Descent 573
- 20.1.1 Genetic Mapping: The Goal and the Data 573
- 20.1.2 The Process of Meiosis and the Descent of DNA 574
- 20.1.3 Genetic Linkage Mapping: Association or Descent? 576
- 20.2 Inference of Local IBD Sharing from Genetic Marker Data 577
- 20.2.1 Identity by Descent at a Locus 577
- 20.2.2 Probabilities of Marker Data Given IBD Pattern 579
- 20.2.3 Modeling the Probabilities of Patterns of IBD 580
- 20.2.4 Inferring Local IBD from Marker Data 581
- 20.3 IBD-Based Detection of Associations between Markers and Traits 583
- 20.3.1 Trait Data Probabilities for Major Gene Models 583
- 20.3.2 Quantitative Trait Data Probabilities under Random Effects Models 584
- 20.3.3 IBD-Based Linkage Likelihoods for Major Gene Models 585
- 20.3.4 IBD-Based Linkage Likelihoods for Random-Effects Models 587
- 20.4 Other Forms of IBD-Based Genetic Mapping 589

|        |                                                                                               |     |
|--------|-----------------------------------------------------------------------------------------------|-----|
| 20.4.1 | IBD-Based Case-Control Studies                                                                | 589 |
| 20.4.2 | Patterns of IBD in Affected Relatives                                                         | 590 |
| 20.5   | Summary                                                                                       | 592 |
|        | Acknowledgements                                                                              | 592 |
|        | References                                                                                    | 593 |
| 21     | Genome-Wide Association Studies                                                               | 597 |
|        | <i>Andrew P. Morris and Lon R. Cardon</i>                                                     |     |
| 21.1   | Introduction                                                                                  | 597 |
| 21.2   | GWAS Design Concepts                                                                          | 599 |
| 21.2.1 | Phenotype Definition                                                                          | 599 |
| 21.2.2 | Structure of Common Genetic Variation and Design of GWAS Genotyping Technology                | 599 |
| 21.2.3 | Sample Size Considerations                                                                    | 601 |
| 21.2.4 | Genome-Wide Significance and Correction for Multiple Testing                                  | 601 |
| 21.2.5 | Replication                                                                                   | 602 |
| 21.3   | GWAS Quality Control                                                                          | 602 |
| 21.3.1 | SNP Quality Control Procedures                                                                | 603 |
| 21.3.2 | Sample Quality Control Procedures                                                             | 604 |
| 21.3.3 | Software                                                                                      | 606 |
| 21.4   | Single SNP Association Analysis                                                               | 606 |
| 21.4.1 | Generalised Linear Modelling Framework                                                        | 606 |
| 21.4.2 | Accounting for Confounding Factors as Covariates                                              | 606 |
| 21.4.3 | Coding of SNP Genotypes                                                                       | 607 |
| 21.4.4 | Imputed Genotypes                                                                             | 609 |
| 21.4.5 | Visualisation of Results of Single SNP GWAS Analyses                                          | 609 |
| 21.4.6 | Interactions with Non-Genetic Risk Factors                                                    | 609 |
| 21.4.7 | Bayesian Methods                                                                              | 611 |
| 21.4.8 | Software                                                                                      | 611 |
| 21.5   | Detecting and Accounting for Genetic Structure in GWASs                                       | 611 |
| 21.5.1 | Identification of Related Individuals                                                         | 612 |
| 21.5.2 | Multivariate Approaches to Identify Ethnic Outliers and Account for Population Stratification | 613 |
| 21.5.3 | Mixed Modelling Approaches to Account for Genetic Structure                                   | 614 |
| 21.5.4 | Software                                                                                      | 615 |
| 21.6   | Multiple SNP Association Analysis                                                             | 616 |
| 21.6.1 | Haplotype-Based Analyses                                                                      | 616 |
| 21.6.2 | SNP-SNP Interaction Analyses                                                                  | 617 |
| 21.6.3 | Gene-Based Analyses                                                                           | 619 |
| 21.6.4 | Software                                                                                      | 619 |
| 21.7   | Discussion                                                                                    | 620 |
|        | References                                                                                    | 623 |
| 22     | Replication and Meta-analysis of Genome-Wide Association Studies                              | 631 |
|        | <i>Frank Dudbridge and Paul Newcombe</i>                                                      |     |
| 22.1   | Introduction                                                                                  | 631 |
| 22.2   | Replication                                                                                   | 632 |
| 22.2.1 | Motivation                                                                                    | 632 |
| 22.2.2 | Different Forms of Replication                                                                | 632 |

|        |                                                                                          |     |
|--------|------------------------------------------------------------------------------------------|-----|
| 22.2.3 | Two-Stage Genome-Wide Association Studies                                                | 634 |
| 22.2.4 | Significance Thresholds for Replication                                                  | 634 |
| 22.2.5 | A Key Challenge: Heterogeneity                                                           | 635 |
| 22.3   | Winner's Curse                                                                           | 635 |
| 22.3.1 | Description of the Problem                                                               | 635 |
| 22.3.2 | Methods for Correcting for Winner's Curse                                                | 637 |
| 22.3.3 | Applicability of These Methods                                                           | 639 |
| 22.4   | Meta-analysis                                                                            | 640 |
| 22.4.1 | Motivation                                                                               | 640 |
| 22.4.2 | An Illustrative Example                                                                  | 640 |
| 22.4.3 | Fixed Effect Meta-analysis                                                               | 641 |
| 22.4.4 | Chi-Square Test for Heterogeneity in Effect                                              | 641 |
| 22.4.5 | Random Effects Meta-analysis                                                             | 642 |
| 22.4.6 | Interpretation and Significance Testing of Meta-analysis Estimates                       | 643 |
| 22.4.7 | Using Funnel Plots to Investigate Small Study Bias in Meta-analysis                      | 644 |
| 22.4.8 | Improving Analyses via Meta-analysis Consortia and Publicly Available Data               | 645 |
| 22.5   | Summary                                                                                  | 647 |
|        | References                                                                               | 647 |
| 23     | Inferring Causal Relationships between Risk Factors and Outcomes Using Genetic Variation | 651 |
|        | <i>Stephen Burgess, Christopher N. Foley, and Verena Zuber</i>                           |     |
| 23.1   | Background                                                                               | 651 |
| 23.1.1 | Correlation and Causation                                                                | 651 |
| 23.1.2 | Chapter Outline                                                                          | 652 |
| 23.2   | Introduction to Mendelian Randomization and Motivating Example                           | 652 |
| 23.2.1 | Instrumental Variable Assumptions                                                        | 653 |
| 23.2.2 | Assessing the Instrumental Variable Assumptions                                          | 654 |
| 23.2.3 | Two-Sample Mendelian Randomization and Summarized Data                                   | 655 |
| 23.3   | Monogenic Mendelian Randomization Analyses: The Easy Case                                | 655 |
| 23.4   | Polygenic Mendelian Randomization Analyses: The Difficult Case                           | 656 |
| 23.4.1 | Example: Low-Density Lipoprotein Cholesterol and Coronary Heart Disease Risk             | 656 |
| 23.4.2 | More Complex Examples                                                                    | 657 |
| 23.4.3 | Two-Stage Least Squares and Inverse-Variance Weighted Methods                            | 659 |
| 23.5   | Robust Approaches for Polygenic Mendelian Randomization Analyses                         | 660 |
| 23.5.1 | Median Estimation Methods                                                                | 660 |
| 23.5.2 | Modal Estimation Methods                                                                 | 660 |
| 23.5.3 | Regularization Methods                                                                   | 660 |
| 23.5.4 | Other Outlier-Robust Methods                                                             | 661 |
| 23.5.5 | MR-Egger Method                                                                          | 661 |
| 23.5.6 | Multivariable Methods                                                                    | 663 |
| 23.5.7 | Interactions and Subsetting                                                              | 663 |
| 23.5.8 | Practical Advice                                                                         | 664 |
| 23.6   | Alternative Approaches for Causal Inference with Genetic Data                            | 665 |
| 23.6.1 | Fine-Mapping and Colocalization                                                          | 665 |
| 23.6.2 | LD Score Regression                                                                      | 666 |
| 23.7   | Causal Estimation in Mendelian Randomization                                             | 667 |

|        |                                                                                                               |     |
|--------|---------------------------------------------------------------------------------------------------------------|-----|
| 23.7.1 | Relevance of Causal Estimate                                                                                  | 667 |
| 23.7.2 | Heterogeneity and Pleiotropy                                                                                  | 668 |
| 23.7.3 | Weak Instrument Bias and Sample Overlap                                                                       | 668 |
| 23.7.4 | Time-Dependent Causal Effects                                                                                 | 669 |
| 23.7.5 | Collider Bias                                                                                                 | 670 |
| 23.8   | Conclusion                                                                                                    | 670 |
|        | References                                                                                                    | 671 |
| 24     | Improving Genetic Association Analysis through Integration of Functional Annotations of the Human Genome      | 679 |
|        | <i>Qiongshi Lu and Hongyu Zhao</i>                                                                            |     |
| 24.1   | Introduction                                                                                                  | 679 |
| 24.2   | Types of Functional Annotation Data in GWAS Applications                                                      | 680 |
| 24.2.1 | Transcriptomic Annotation Data                                                                                | 680 |
| 24.2.2 | Epigenetic Annotation Data                                                                                    | 681 |
| 24.2.3 | DNA Conservation                                                                                              | 681 |
| 24.3   | Methods to Synthesize Annotation Data                                                                         | 682 |
| 24.3.1 | Genome Browsers and Annotator Software                                                                        | 682 |
| 24.3.2 | Supervised Learning Methods                                                                                   | 682 |
| 24.3.3 | Unsupervised Learning Methods                                                                                 | 684 |
| 24.3.4 | Improving Specificity of Computational Annotations                                                            | 685 |
| 24.4   | Methods to Integrate Functional Annotations in Genetic Association Analysis                                   | 685 |
| 24.4.1 | Partitioning Heritability and Genetic Covariance                                                              | 685 |
| 24.4.2 | Imputation-Based Gene-Level Association Analysis                                                              | 688 |
| 24.4.3 | Other Applications and Future Directions                                                                      | 690 |
|        | Acknowledgements                                                                                              | 690 |
|        | References                                                                                                    | 691 |
| 25     | Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci | 697 |
|        | <i>Solveig K. Sieberts and Eric E. Schadt</i>                                                                 |     |
| 25.1   | Introduction                                                                                                  | 697 |
| 25.1.1 | An Overview of Transcription as a Complex Process                                                             | 700 |
| 25.1.2 | Modeling Approaches for Biological Processes                                                                  | 702 |
| 25.1.3 | Human versus Experimental Models                                                                              | 704 |
| 25.2   | Modeling for eQTL Detection and Causal Inference                                                              | 705 |
| 25.2.1 | Heritability of Expression Traits                                                                             | 705 |
| 25.2.2 | Single-Trait eQTL Mapping                                                                                     | 706 |
| 25.2.3 | Joint eQTL Mapping                                                                                            | 706 |
| 25.2.4 | eQTL and Clinical Trait Linkage Mapping to Infer Causal Associations                                          | 708 |
| 25.3   | Inferring Gene Regulatory Networks                                                                            | 714 |
| 25.3.1 | From Assessing Causal Relationships among Trait Pairs to Predictive Gene Networks                             | 714 |
| 25.3.2 | Building from the Bottom Up or Top Down?                                                                      | 714 |
| 25.3.3 | Using eQTL Data to Reconstruct Coexpression Networks                                                          | 715 |
| 25.3.4 | An Integrative Genomics Approach to Constructing Predictive Network Models                                    | 718 |

|        |                                                                                                                                 |     |
|--------|---------------------------------------------------------------------------------------------------------------------------------|-----|
| 25.3.5 | Integrating Genetic Data as a <i>Structure</i> Prior to Enhance Causal Inference in the Bayesian Network Reconstruction Process | 720 |
| 25.3.6 | Incorporating Other Omics Data as Network Priors in the Bayesian Network Reconstruction Process                                 | 721 |
| 25.3.7 | Illustrating the Construction of Predictive Bayesian Networks with an Example                                                   | 722 |
| 25.4   | Conclusions                                                                                                                     | 723 |
| 25.5   | Software                                                                                                                        | 724 |
|        | References                                                                                                                      | 725 |
| 26     | Statistical Methods for Single-Cell RNA-Sequencing                                                                              | 735 |
|        | <i>Tallulah S. Andrews, Vladimir Yu. Kiselev, and Martin Hemberg</i>                                                            |     |
| 26.1   | Introduction                                                                                                                    | 735 |
| 26.2   | Overview of scRNA-Seq Experimental Platforms and Low-Level Analysis                                                             | 736 |
| 26.2.1 | Low-Throughput Methods                                                                                                          | 736 |
| 26.2.2 | High-Throughput Methods                                                                                                         | 737 |
| 26.2.3 | Computational Analysis                                                                                                          | 739 |
| 26.3   | Novel Statistical Challenges Posed by scRNA-Seq                                                                                 | 739 |
| 26.3.1 | Estimating Transcript Levels                                                                                                    | 739 |
| 26.3.2 | Analysis of the Expression Matrix                                                                                               | 747 |
|        | References                                                                                                                      | 753 |
| 27     | Variant Interpretation and Genomic Medicine                                                                                     | 761 |
|        | <i>K. Carss, D. Goldstein, V. Aggarwal, and S. Petrovski</i>                                                                    |     |
| 27.1   | Introduction and Current Challenges                                                                                             | 761 |
| 27.2   | Understanding the Effect of a Variant                                                                                           | 765 |
| 27.3   | Understanding Genomic Variation Context through Large Human Reference Cohorts                                                   | 771 |
| 27.4   | Functional Assays of Genetic Variation                                                                                          | 777 |
| 27.5   | Leveraging Existing Information about Gene Function Including Human and Model Phenotype Resources                               | 779 |
| 27.6   | Holistic Variant Interpretation                                                                                                 | 782 |
| 27.7   | Future Challenges and Closing Remarks                                                                                           | 783 |
| 27.8   | Web Resources                                                                                                                   | 786 |
|        | References                                                                                                                      | 788 |
| 28     | Prediction of Phenotype from DNA Variants                                                                                       | 799 |
|        | <i>M.E. Goddard, T.H.E. Meuwissen, and H.D. Daetwyler</i>                                                                       |     |
| 28.1   | Introduction                                                                                                                    | 799 |
| 28.2   | Genetic Variation Affecting Phenotype                                                                                           | 800 |
| 28.3   | Data on DNA Polymorphisms Used for Prediction of Genetic Effects                                                                | 801 |
| 28.4   | Prediction of Additive Genetic Values                                                                                           | 802 |
| 28.4.1 | An Equivalent Model                                                                                                             | 804 |
| 28.4.2 | Single-Step BLUP                                                                                                                | 804 |
| 28.4.3 | Multiple Traits                                                                                                                 | 805 |
| 28.4.4 | Gene Expression                                                                                                                 | 806 |
| 28.4.5 | Using External Information                                                                                                      | 806 |
| 28.5   | Factors Affecting Accuracy of Prediction                                                                                        | 806 |
| 28.6   | Other Uses of the Bayesian Genomic Selection Models                                                                             | 808 |

|        |                                                                                  |     |
|--------|----------------------------------------------------------------------------------|-----|
| 28.7   | Examples of Genomic Prediction                                                   | 809 |
| 28.7.1 | Cattle                                                                           | 809 |
| 28.7.2 | Humans                                                                           | 809 |
| 28.8   | Conclusions                                                                      | 810 |
|        | References                                                                       | 810 |
| 29     | Disease Risk Models                                                              | 815 |
|        | <i>Allison Meisner and Nilanjan Chatterjee</i>                                   |     |
| 29.1   | Introduction and Background                                                      | 815 |
| 29.1.1 | Disease Risk Models and Their Applications                                       | 815 |
| 29.1.2 | Examples of Available Disease Risk Models                                        | 817 |
| 29.1.3 | Incorporating Genetic Factors                                                    | 817 |
| 29.2   | Absolute Risk Model                                                              | 818 |
| 29.2.1 | General Software for Building Absolute Risk Models                               | 820 |
| 29.3   | Building a Polygenic Risk Score                                                  | 821 |
| 29.3.1 | Expected Performance                                                             | 821 |
| 29.3.2 | Standard Approach to Constructing a PRS: LD Clumping and $p$ -Value Thresholding | 823 |
| 29.3.3 | Advanced Approaches to Constructing a PRS                                        | 825 |
| 29.4   | Combining PRS and Epidemiologic Factors                                          | 826 |
| 29.5   | Model Validation                                                                 | 827 |
| 29.6   | Evaluating Clinical Utility                                                      | 828 |
| 29.7   | Example: Breast Cancer                                                           | 829 |
| 29.8   | Discussion                                                                       | 832 |
| 29.8.1 | Future Directions                                                                | 832 |
| 29.8.2 | Challenges                                                                       | 832 |
|        | References                                                                       | 833 |
| 30     | Bayesian Methods for Gene Expression Analysis                                    | 843 |
|        | <i>Alex Lewin, Leonardo Bottolo, and Sylvia Richardson</i>                       |     |
| 30.1   | Introduction                                                                     | 843 |
| 30.2   | Modelling Microarray Data                                                        | 845 |
| 30.2.1 | Modelling Intensities                                                            | 845 |
| 30.2.2 | Gene Variability                                                                 | 845 |
| 30.2.3 | Normalization                                                                    | 846 |
| 30.3   | Modelling RNA-Sequencing Reads                                                   | 846 |
| 30.3.1 | Alignments for RNA-Sequencing Data                                               | 846 |
| 30.3.2 | Likelihood for Read-Level Data                                                   | 847 |
| 30.3.3 | Likelihood for Transcript-Level Read Counts                                      | 848 |
| 30.3.4 | Likelihood for Gene-Level Read Counts                                            | 850 |
| 30.4   | Priors for Differential Expression Analysis                                      | 851 |
| 30.4.1 | Differential Expression from Microarray Data                                     | 851 |
| 30.4.2 | Differential Expression from RNA-Sequencing Data                                 | 856 |
| 30.5   | Multivariate Gene Selection Models                                               | 857 |
| 30.5.1 | Variable Selection Approach                                                      | 857 |
| 30.5.2 | Bayesian Shrinkage with Sparsity Priors                                          | 861 |
| 30.6   | Quantitative Trait Loci                                                          | 863 |
| 30.6.1 | Single-Response Models                                                           | 863 |
| 30.6.2 | Multiple-Response Models                                                         | 864 |

|        |                                                                                                    |     |
|--------|----------------------------------------------------------------------------------------------------|-----|
|        | Acknowledgements                                                                                   | 868 |
|        | References                                                                                         | 869 |
| 31     | <b>Modelling Gene Expression Dynamics with Gaussian Process Inference</b>                          | 879 |
|        | <i>Magnus Rattray, Jing Yang, Sumon Ahmed, and Alexis Boukouvalas</i>                              |     |
| 31.1   | Introduction                                                                                       | 879 |
| 31.1.1 | Covariance Function                                                                                | 880 |
| 31.1.2 | Inference                                                                                          | 882 |
| 31.2   | Applications to Bulk Time Series Expression Data                                                   | 883 |
| 31.2.1 | Identifying Differential Expression in Time                                                        | 884 |
| 31.2.2 | Identifying Changes between Two Time Course Experiments                                            | 885 |
| 31.2.3 | Hierarchical Models of Replicates and Clusters                                                     | 887 |
| 31.2.4 | Differential Equation Models of Production and Degradation                                         | 888 |
| 31.3   | Modelling Single-Cell Data                                                                         | 889 |
| 31.3.1 | Modelling Single-Cell Trajectory Data                                                              | 889 |
| 31.3.2 | Dimensionality Reduction and Pseudotime Inference                                                  | 890 |
| 31.3.3 | Modelling Branching Dynamics with Single-Cell RNA-Sequencing Data                                  | 892 |
| 31.4   | Conclusion                                                                                         | 893 |
|        | Acknowledgements                                                                                   | 894 |
|        | References                                                                                         | 894 |
| 32     | <b>Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models</b> | 899 |
|        | <i>Marco Grzegorzczuk and Dirk Husmeier</i>                                                        |     |
| 32.1   | Introduction                                                                                       | 899 |
| 32.2   | Methodology                                                                                        | 901 |
| 32.2.1 | Dynamic Bayesian Networks (DBN)                                                                    | 901 |
| 32.2.2 | Bayesian Linear Regression                                                                         | 902 |
| 32.2.3 | Bayesian Piecewise Linear Regression (NH-DBN)                                                      | 905 |
| 32.2.4 | Bayesian Piecewise Linear Regression with Coupled Regression Coefficients (Coupled NH-DBNs)        | 908 |
| 32.2.5 | NH-DBNs with More Flexible Allocation Schemes                                                      | 915 |
| 32.2.6 | NH-DBNs with Time-Varying Network Structures                                                       | 916 |
| 32.2.7 | Dynamic Bayesian Network Modelling                                                                 | 918 |
| 32.2.8 | Computational Complexity                                                                           | 920 |
| 32.3   | Application Examples                                                                               | 921 |
| 32.3.1 | Morphogenesis in <i>Drosophila</i>                                                                 | 921 |
| 32.3.2 | Synthetic Biology in Yeast                                                                         | 922 |
| 32.4   | Summary                                                                                            | 927 |
|        | Appendix A: Coupling Schemes                                                                       | 927 |
| A.1    | Hard Information Coupling Based on an Exponential Prior                                            | 928 |
| A.2    | Hard Information Coupling Based on a Binomial Prior                                                | 928 |
| A.3    | Soft Information Coupling Based on a Binomial Prior                                                | 929 |
|        | References                                                                                         | 929 |
| 33     | <b>DNA Methylation</b>                                                                             | 933 |
|        | <i>Kasper D. Hansen, Kimberly D. Siegmund, and Shili Lin</i>                                       |     |
| 33.1   | A Brief Introduction                                                                               | 933 |
| 33.2   | Measuring DNA Methylation                                                                          | 934 |



|        |                                                                                         |     |
|--------|-----------------------------------------------------------------------------------------|-----|
| 33.3   | Differential DNA Methylation                                                            | 936 |
| 33.3.1 | Differential Methylation with Bisulfite-Sequencing Data                                 | 936 |
| 33.3.2 | Differential Methylation with Capture-Sequence Data                                     | 939 |
| 33.3.3 | Differential Methylation with HumanMethylation Array Data                               | 940 |
| 33.4   | Other Topics of Interest                                                                | 941 |
|        | References                                                                              | 942 |
| 34     | Statistical Methods in Metabolomics                                                     | 949 |
|        | <i>Timothy M.D. Ebbels, Maria De Iorio, and David A. Stephens</i>                       |     |
| 34.1   | Introduction                                                                            | 949 |
| 34.2   | Preprocessing and Deconvolution                                                         | 950 |
| 34.2.1 | Nuclear Magnetic Resonance Spectroscopy                                                 | 950 |
| 34.2.2 | Liquid Chromatography – Mass Spectrometry                                               | 952 |
| 34.3   | Univariate Methods                                                                      | 954 |
| 34.3.1 | Metabolome-Wide Significance Levels                                                     | 956 |
| 34.3.2 | Sample Size and Power                                                                   | 957 |
| 34.4   | Multivariate Methods and Chemometrics Techniques                                        | 958 |
| 34.4.1 | Linear Regression Methods                                                               | 959 |
| 34.4.2 | Shrinkage Methods                                                                       | 960 |
| 34.5   | Orthogonal Projection Methods                                                           | 961 |
| 34.5.1 | Principal Components Analysis                                                           | 962 |
| 34.5.2 | Partial Least Squares                                                                   | 964 |
| 34.5.3 | Orthogonal Projection onto Latent Structures                                            | 965 |
| 34.6   | Network Analysis                                                                        | 966 |
| 34.7   | Metabolite Identification and Pathway Analysis                                          | 969 |
| 34.7.1 | Statistical Correlation Spectroscopy                                                    | 969 |
| 34.7.2 | Pathway and Metabolite Set Analysis                                                     | 971 |
| 34.8   | Conclusion                                                                              | 972 |
|        | References                                                                              | 972 |
| 35     | Statistical and Computational Methods in Microbiome and Metagenomics                    | 977 |
|        | <i>Hongzhe Li</i>                                                                       |     |
| 35.1   | Microbiome in Human Health and Disease                                                  | 977 |
| 35.2   | Estimation of Microbiome Features from 16S rRNA and Shotgun Metagenomic Sequencing Data | 980 |
| 35.2.1 | Estimation of Microbiome Features in 16S rRNA Data                                      | 980 |
| 35.2.2 | Estimation of Microbial Composition in Shotgun Metagenomic Data                         | 981 |
| 35.2.3 | Estimation of Microbial Gene/Pathway Abundance in Shotgun Metagenomic Data              | 982 |
| 35.2.4 | Quantification of Bacterial Growth Dynamics                                             | 982 |
| 35.2.5 | Microbial Diversity Index                                                               | 983 |
| 35.3   | Methods for Analysis of Microbiome as an Outcome of an Intervention or Exposure         | 983 |
| 35.3.1 | Modeling Multivariate Sparse Count Data as the Response Variable                        | 984 |
| 35.3.2 | Modeling High-Dimensional Compositional Response Data in Microbiome Studies             | 984 |
| 35.4   | Methods for Analysis of Microbiome as a Covariate                                       | 985 |
| 35.4.1 | Regression Analysis with Compositional Covariates                                       | 985 |
| 35.4.2 | Kernel-Based Regression in Microbiome Studies                                           | 986 |



|        |                                                                                                   |      |
|--------|---------------------------------------------------------------------------------------------------|------|
| 35.5   | Methods for Analysis of Microbiome as a Mediator                                                  | 987  |
| 35.6   | Integrative Analysis of Microbiome, Small Molecules and Metabolomics Data                         | 989  |
| 35.6.1 | Computational Analysis of Small Molecules from the Human Microbiota                               | 989  |
| 35.6.2 | Metabolic Modeling in Microbiome                                                                  | 990  |
| 35.7   | Discussion and Future Directions                                                                  | 991  |
|        | Acknowledgements                                                                                  | 991  |
|        | References                                                                                        | 992  |
| 36     | Bacterial Population Genomics                                                                     | 997  |
|        | <i>Jukka Corander, Nicholas J. Croucher, Simon R. Harris, John A. Lees, and Gerry Tonkin-Hill</i> |      |
| 36.1   | Introduction                                                                                      | 997  |
| 36.2   | Genetic Population Structure and Clustering of Genotypes                                          | 998  |
| 36.2.1 | Background                                                                                        | 998  |
| 36.2.2 | Model-Based Clustering                                                                            | 998  |
| 36.2.3 | Linkage Disequilibrium                                                                            | 1000 |
| 36.2.4 | Distance-Based Methods                                                                            | 1000 |
| 36.3   | Phylogenetics and Dating Analysis                                                                 | 1001 |
| 36.4   | Transmission Modeling                                                                             | 1004 |
| 36.4.1 | Challenges                                                                                        | 1004 |
| 36.5   | Genome-Wide Association Studies in Bacteria                                                       | 1008 |
| 36.5.1 | Background                                                                                        | 1008 |
| 36.5.2 | Phylogenetic Methods                                                                              | 1009 |
| 36.5.3 | Regression-Based Methods                                                                          | 1011 |
| 36.6   | Genome-Wide Epistasis Analysis                                                                    | 1012 |
| 36.7   | Gene Content Analysis                                                                             | 1013 |
|        | References                                                                                        | 1014 |
|        | Reference Author Index                                                                            | 1021 |
|        | Subject Index                                                                                     | 1109 |