

Contents

Volume 1

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

1	Statistical Modeling and Inference in Genetics	1
	<i>Daniel Wegmann and Christoph Leuenberger</i>	
1.1	Statistical Models and Inference	1
	1.1.1 Statistical Models	2
	1.1.2 Inference Methods and Algorithms	4
1.2	Maximum Likelihood Inference	4
	1.2.1 Properties of Maximum Likelihood Estimators	6
	1.2.2 Quantifying Confidence: the Fisher Information Matrix	8
	1.2.3 Newton's Method	9
	1.2.4 Latent Variable Problems: the EM Algorithm	11
	1.2.5 Approximate Techniques	16
1.3	Bayesian Inference	20
	1.3.1 Choice of Prior Distributions	21
	1.3.2 Bayesian Point Estimates and Confidence Intervals	22
	1.3.3 Markov Chain Monte Carlo	23
	1.3.4 Empirical Bayes for Latent Variable Problems	30
	1.3.5 Approximate Bayesian Computation	31
1.4	Model Selection	37
	1.4.1 Likelihood Ratio Statistic	37
	1.4.2 Bayesian Model Choice	38
1.5	Hidden Markov Models	40
	1.5.1 Bayesian Inference of Hidden States Using Forward-Backward Algorithm	42
	1.5.2 Finding the Most Likely Hidden Path (Viterbi Algorithm)	43
	1.5.3 MLE Inference of Hierarchical Parameters (Baum-Welch Algorithm)	44
	Acknowledgements	46
	References	47

2	Linkage Disequilibrium, Recombination and Haplotype Structure	51
	<i>Gil McVean and Jerome Kelleher</i>	
2.1	What Is Linkage Disequilibrium?	51
2.2	Measuring Linkage Disequilibrium	53
2.2.1	Single-Number Summaries of LD	54
2.2.2	The Spatial Distribution of LD	56
2.2.3	Various Extensions of Two-Locus LD Measures	60
2.3	Modelling Linkage Disequilibrium and Genealogical History	60
2.3.1	A Historical Perspective	60
2.3.2	Coalescent Modelling	62
2.3.3	Relating Genealogical History to LD	67
2.4	Data Analysis	69
2.4.1	Estimating Recombination Rates	69
2.4.2	Methods Exploiting Haplotype Structure	72
2.5	Prospects	75
	Acknowledgements	75
	References	76
3	Haplotype Estimation and Genotype Imputation	87
	<i>Jonathan Marchini</i>	
3.1	Haplotype Estimation	87
3.1.1	A Simple Haplotype Frequency Model	88
3.1.2	Hidden Markov Models for Phasing	89
3.1.3	Phasing in Related Samples	93
3.1.4	Phasing Using Sequencing Data	94
3.1.5	Phasing from a Reference Panel	95
3.1.6	Measuring Phasing Performance	96
3.2	Genotype Imputation	97
3.2.1	Uses of Imputation in GWASs	98
3.2.2	Haploid Imputation	99
3.2.3	Imputation Methods	100
3.2.4	Testing Imputed Genotypes for Association	103
3.2.5	Summary Statistic Imputation	104
3.2.6	Factors Affecting Accuracy	104
3.2.7	Quality Control for Imputed Data	107
3.3	Future Directions	109
	References	109
4	Mathematical Models in Population Genetics	115
	<i>Nick Barton and Alison Etheridge</i>	
4.1	Introduction	115
4.2	Single-Locus Models	116
4.2.1	Random Drift and the Kingman Coalescent	117
4.2.2	Diffusion Approximations	120
4.2.3	Spatially Structured Populations	126
4.3	Multiple Loci	130
4.3.1	Linkage Equilibrium	131
4.3.2	Beyond Linkage Equilibrium	134
4.4	Outlook	140
	References	140

5	Coalescent Theory	145
	<i>Magnus Nordborg</i>	
5.1	Introduction	145
5.2	The Coalescent	146
5.2.1	The Fundamental Insights	146
5.2.2	The Coalescent Approximation	148
5.3	Generalizing the Coalescent	151
5.3.1	Robustness and Scaling	151
5.3.2	Variable Population Size	152
5.3.3	Population Structure on Different Time-Scales	153
5.4	Geographical Structure	155
5.4.1	The Structured Coalescent	155
5.4.2	The Strong-Migration Limit	156
5.5	Diploidy and Segregation	157
5.5.1	Hermaphrodites	157
5.5.2	Males and Females	159
5.6	Recombination	159
5.6.1	The Ancestral Recombination Graph	160
5.6.2	Properties and Effects of Recombination	163
5.7	Selection	164
5.7.1	Balancing Selection	165
5.7.2	Selective Sweeps	166
5.7.3	Background Selection	168
5.8	Neutral Mutations	168
5.9	Concluding Remarks	169
5.9.1	The Coalescent and 'Classical' Population Genetics	169
5.9.2	The Coalescent and Phylogenetics	169
5.9.3	Prospects	171
	Acknowledgements	171
	References	171
6	Phylogeny Estimation Using Likelihood-Based Methods	177
	<i>John P. Huelsenbeck</i>	
6.1	Introduction	177
6.1.1	Statistical Phylogenetics	178
6.1.2	Chapter Outline	178
6.2	Maximum Likelihood and Bayesian Estimation	179
6.2.1	Maximum Likelihood	179
6.2.2	Bayesian Inference	180
6.3	Choosing among Models Using Likelihood Ratio Tests and Bayes Factors	184
6.4	Calculating the Likelihood for a Phylogenetic Model	186
6.4.1	Character Matrices and Alignments	186
6.4.2	The Phylogenetic Model	186
6.4.3	Calculating the Probability of a Character History	187
6.4.4	Continuous-Time Markov Model	188
6.4.5	Marginalizing over Character Histories	189
6.5	The Mechanics of Maximum Likelihood and Bayesian Inference	192
6.5.1	Maximum Likelihood	192
6.5.2	Bayesian Inference and Markov Chain Monte Carlo	193

6.6	Applications of Likelihood-Based Methods in Molecular Evolution	199
6.6.1	A Taxonomy of Commonly Used Substitution Models	199
6.6.2	Expanding the Model around Groups of Sites	202
6.6.3	Rate Variation across Sites	204
6.6.4	Divergence Time Estimation	206
6.7	Conclusions	212
	References	213
7	The Multispecies Coalescent	219
	<i>Laura Kubatko</i>	
7.1	Introduction	219
7.2	Probability Distributions under the Multispecies Coalescent	221
7.2.1	Gene Tree Probabilities	221
7.2.2	Site Pattern Probabilities	227
7.2.3	Species Tree Likelihoods under the Multispecies Coalescent	229
7.2.4	Model Assumptions and Violations	230
7.3	Species Tree Inference under the Multispecies Coalescent	231
7.3.1	Summary Statistics Methods	231
7.3.2	Bayesian Full-Data Methods	234
7.3.3	Site Pattern-Based Methods	235
7.3.4	Multilocus versus SNP Data	236
7.3.5	Empirical Examples	237
7.4	Coalescent-Based Estimation of Parameters at the Population and Species Levels	239
7.4.1	Speciation Times and Population Sizes	239
7.4.2	Hybridization and Gene Flow	240
7.4.3	Species Delimitation	241
7.4.4	Future Prospects	242
	Acknowledgements	242
	References	242
8	Population Structure, Demography and Recent Admixture	247
	<i>G. Hellenthal</i>	
8.1	Introduction	247
8.1.1	'Admixture' versus 'Background' Linkage Disequilibrium	248
8.2	Spatial Summaries of Genetic Variation Using Principal Components Analysis	249
8.3	Clustering Algorithms	251
8.3.1	Defining 'Populations'	251
8.3.2	Clustering Based on Allele Frequency Patterns	252
8.3.3	Incorporating Admixture	253
8.3.4	Incorporating Admixture Linkage Disequilibrium	254
8.3.5	Incorporating Background Linkage Disequilibrium: Using Haplotypes to Improve Inference	255
8.3.6	Interpreting Genetic Clusters	258
8.4	Inferring Population Size Changes and Split Times	259
8.4.1	Allele Frequency Spectrum Approaches	260
8.4.2	Approaches Using Whole-Genome Sequencing	261
8.5	Identifying/Dating Admixture Events	262
8.5.1	Inferring DNA Segments Inherited from Different Sources	263
8.5.2	Measuring Decay of Linkage Disequilibrium	265

8.6	Conclusion	267
	Acknowledgements	268
	References	268
9	Statistical Methods to Detect Archaic Admixture and Identify Introgressed Sequences	275
	<i>Liming Li and Joshua M. Akey</i>	
9.1	Introduction	275
9.2	Methods to Test Hypotheses of Archaic Admixture and Infer Admixture Proportions	277
9.2.1	Genetic Drift and Allele Frequency Divergence in Genetically Structured Populations	277
9.2.2	Three-Population Test	277
9.2.3	<i>D</i> -Statistic	279
9.2.4	<i>F</i> ₄ -Statistic	282
9.3	Methods to Identify Introgressed Sequences	283
9.3.1	<i>S</i> *-Statistic	284
9.3.2	Hidden Markov and Conditional Random Field Models	287
9.3.3	Relative Advantages and Disadvantages of Approaches to Detect Introgressed Sequences	289
9.4	Summary and Perspective	289
	References	290
10	Population Genomic Analyses of DNA from Ancient Remains	295
	<i>Torsten Günther and Mattias Jakobsson</i>	
10.1	Introduction	295
10.2	Challenges of Working with and Analyzing Ancient DNA Data	296
10.2.1	Sequence Degradation	296
10.2.2	Contamination	297
10.2.3	Handling Sequence Data from Ancient Material	300
10.2.4	Different Sequencing Approaches and the Limitations in their Resulting Data	301
10.2.5	Effects of Limited Amounts of Data on Downstream Analysis	301
10.3	Opportunities of Ancient DNA	302
10.3.1	Population Differentiation in Time and Space	303
10.3.2	Continuity	306
10.3.3	Migration and Admixture over Time	307
10.3.4	Demographic Inference Based on High-Coverage Ancient Genomes	308
10.3.5	Allele Frequency Trajectories	308
10.4	Some Examples of How Genetic Studies of Ancient Remains Have Contributed to a New Understanding of the Human Past	310
10.4.1	Archaic Genomes and the Admixture with Modern Humans	310
10.4.2	Neolithic Revolution in Europe and the Bronze Age Migrations	311
10.5	Summary and Perspective	313
	Acknowledgements	313
	References	314
11	Sequence Covariation Analysis in Biological Polymers	325
	<i>William R. Taylor, Shaun Kandathil, and David T. Jones</i>	
11.1	Introduction	325

11.2	Methods	326
11.2.1	DCA Method	326
11.2.2	PSICOV	327
11.2.3	plmDCA, GREMLIN and CCMpred	327
11.3	Applications	328
11.3.1	Globular Protein Fold Prediction	328
11.3.2	Transmembrane Protein Prediction	328
11.3.3	RNA Structure Prediction	328
11.3.4	Protein Disordered Regions	329
11.3.5	Protein-Protein Interactions	329
11.3.6	Allostery and Dynamics	330
11.3.7	CASP	330
11.4	New Developments	332
11.4.1	Sequence Alignment	332
11.4.2	Comparison to Known Structures	333
11.4.3	Segment Parsing	334
11.4.4	Machine Learning	335
11.4.5	Deep Learning Methods	336
11.4.6	Sequence Pairing	338
11.4.7	Phylogeny Constraints	339
11.5	Outlook	340
	Acknowledgements	341
	References	342
12	Probabilistic Models for the Study of Protein Evolution	347
	<i>Umberto Perron, Iain H. Moal, Jeffrey L. Thorne, and Nick Goldman</i>	
12.1	Introduction	347
12.2	Empirically Derived Models of Amino Acid Replacement	348
12.2.1	The Dayhoff and Eck Model	348
12.2.2	Descendants of the Dayhoff Model	350
12.3	Heterogeneity of Replacement Rates among Sites	351
12.4	Protein Structural Environments	351
12.5	Variation of Preferred Residues among Sites	353
12.6	Models with a Physicochemical Basis	355
12.7	Codon-Based Models	355
12.8	Dependence among Positions	357
12.9	Stochastic Models of Structural Evolution	359
12.10	Conclusion	360
	Acknowledgements	361
	References	361
13	Adaptive Molecular Evolution	369
	<i>Ziheng Yang</i>	
13.1	Introduction	369
13.2	Markov Model of Codon Substitution	371
13.3	Estimation of Synonymous and Non-synonymous Substitution Rates between Two Sequences and Test of Selection on the Protein	372
13.3.1	Heuristic Estimation Methods	372
13.3.2	Maximum Likelihood Estimation	374

13.3.3	Bayesian Estimation	377
13.3.4	A Numerical Example	377
13.4	Likelihood Calculation on a Phylogeny	379
13.5	Detecting Adaptive Evolution along Lineages	380
13.5.1	Likelihood Calculation under Models of Variable ω Ratios among Lineages	380
13.5.2	Adaptive Evolution in the Primate Lysozyme	381
13.5.3	Comparison with Methods Based on Reconstructed Ancestral Sequences	382
13.6	Inferring Amino Acid Sites under Positive Selection	384
13.6.1	Likelihood Ratio Test under Models of Variable ω Ratios among Sites	384
13.6.2	Methods that Test One Site at a Time	386
13.6.3	Positive Selection in the HIV-1 <i>vif</i> Genes	386
13.7	Testing Positive Selection Affecting Particular Sites and Lineages	388
13.7.1	Branch-Site Test of Positive Selection	388
13.7.2	Clade Models and Other Variants	389
13.8	Limitations of Current Methods	390
13.9	Computer Software	391
	References	391
14	Detecting Natural Selection	397
	<i>Aaron J. Stern and Rasmus Nielsen</i>	
14.1	Introduction	397
14.2	Types of Selection	398
14.2.1	Directional Selection	398
14.2.2	Balancing Selection	399
14.2.3	Polygenic Selection	399
14.3	The Signature of Selection in the Genome	399
14.3.1	The Signature of Positive Directional Selection	400
14.3.2	Balancing Selection	403
14.3.3	Polygenic Selection	403
14.3.4	Confounders	404
14.4	Methods for Detecting Selection	405
14.4.1	Substitution-Based Methods	405
14.4.2	Methods Comparing Substitutions and Diversity	406
14.4.3	Methods Using the Frequency Spectrum	407
14.4.4	Methods Using Genetic Differentiation	408
14.4.5	Methods Using Haplotype Structure	410
14.4.6	Why Full-Likelihood Methods Are Intractable for Population Samples	412
14.4.7	Composite Likelihood Methods	412
14.4.8	Approximate Bayesian Computation	413
14.4.9	Machine Learning Methods	413
14.5	Discussion	414
	References	415
15	Evolutionary Quantitative Genetics	421
	<i>Bruce Walsh and Michael B. Morrissey</i>	
15.1	Introduction	421

15.2	Resemblances, Variances, and Additive Genetic Values	422
15.2.1	Fisher's Genetic Decomposition	422
15.2.2	Additive Genetic Variances and Covariances	423
15.3	Parent-Offspring Regressions and the Response to Selection	423
15.3.1	Single-Trait Parent-Offspring Regressions	424
15.3.2	Selection Differentials and the Breeder's Equation	424
15.3.3	Multiple-Trait Parent-Offspring Regressions	425
15.3.4	The Genetic and Phenotypic Covariance Matrices	425
15.3.5	The Multivariate Breeder's Equation	425
15.4	The Infinitesimal Model	426
15.4.1	Linearity of Parent-Offspring Regressions under the Infinitesimal Model	426
15.4.2	Allele Frequency Changes under the Infinitesimal Model	426
15.4.3	Changes in Variances	427
15.4.4	The Equilibrium Additive Genetic Variance	429
15.5	Inference of σ_A^2 and G	430
15.6	Fitness	432
15.6.1	Individual Fitness	432
15.6.2	Episodes of Selection	433
15.7	The Robertson-Price Identity, and Theorems of Selection	434
15.7.1	Description of the Theorems	435
15.7.2	Empirical Operationalization of the Theorems	436
15.8	The Opportunity for Selection	437
15.9	Selection Coefficients	438
15.9.1	Measures of Selection on the Mean	439
15.9.2	Measures of Selection on the Variance	439
15.10	Fitness Functions and the Characterization of Selection	441
15.10.1	Individual and Mean Fitness Functions	441
15.10.2	Gradients and the Local Geometry of Fitness Surfaces	442
15.11	Multivariate Selection	444
15.11.1	Short-Term Changes in Means: The Multivariate Breeder's Equation	444
15.11.2	The Effects of Genetic Correlations: Direct and Correlated Responses	444
15.11.3	Selection Gradients and Understanding which Traits Affect Fitness	446
15.12	Inference of Selection Gradients	447
15.12.1	Ordinary Least Squares Analysis	448
15.12.2	Flexible Inference of Fitness Functions with Associated Selection Gradient Estimates	449
15.12.3	Normality and Selection Gradients	450
15.13	Summary	451
	References	452
16	Conservation Genetics	457
	<i>Mark Beaumont and Jinliang Wang</i>	
16.1	Introduction	457
16.2	Estimating Effective Population Size	458
16.2.1	Methods Based on Heterozygosity Excess	459
16.2.2	Methods Based on Linkage Disequilibrium	460
16.2.3	Methods Based on Relatedness	463
16.2.4	Methods Based on Temporal Changes in Allele Frequency	466

16.3	Estimating Census Size by the Genotype Capture–Recapture Approach	470
16.3.1	Methods Based on Multilocus Genotype Mismatches	472
16.3.2	Methods Based on Pairwise Relatedness	472
16.3.3	Methods Based on Pairwise Relationships	473
16.3.4	Methods Based on Pedigree Reconstruction	474
16.4	Inferring Genetic Structure	475
16.4.1	Measuring Genetic Differentiation	475
16.4.2	Population Assignment	477
16.4.3	Population Clustering and Inference of Ancestry Proportions	479
16.4.4	Inferring Levels of Recent Gene Flow	483
16.4.5	Landscape Genetics	486
16.5	Deintrogression Strategies	487
16.6	Genetic Species Delimitation	489
16.7	Conclusions and Outlook	491
	Acknowledgements	492
	References	492
17	Statistical Methods for Plant Breeding	501
	<i>Ian Mackay, Hans-Peter Piepho, and Antonio Augusto Franco Garcia</i>	
17.1	Introduction	501
17.2	Heritability and the Breeder's Equation in Plant Breeding	502
17.3	The Breeding System of Plants	504
17.4	Polyploidy in Plants and Its Genetic Consequences	505
17.5	Genomic Rearrangements in Plants	509
17.6	Genetic Architecture of Traits in Plants	510
17.7	Response to the Environment and Plasticity	511
17.8	Genomic Selection	514
17.8.1	Genotype–Environment Interaction	514
17.8.2	Quantitative Trait Loci and Major Genes	516
17.8.3	Genomic Selection and Cross Prediction	517
17.8.4	Genomic Selection and Phenotyping Cost	517
17.8.5	Mate Selection	517
17.8.6	Sequential Selection	518
17.8.7	Genomic Prediction of Hybrid Performance and Heterosis	518
17.8.8	Marker Imputation	519
17.9	Experimental Design and Analysis	519
17.10	Conclusions	521
	References	521
18	Forensic Genetics	531
	<i>B.S. Weir</i>	
18.1	Introduction	531
18.2	Principles of Interpretation	532
18.3	Profile Probabilities	534
18.3.1	Genetic Models for Allele Frequencies	535
18.3.2	Y-STR Profiles	539
18.4	Mixtures	542
18.4.1	Combined Probabilities of Inclusion and Exclusion	542
18.4.2	Likelihood Ratios	542
18.5	Behavior of Likelihood Ratio	546

- 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