

---

*Handbook of Statistical Genetics*

---

---

# *Handbook of Statistical Genetics*

---

*Edited by*

**D. J. Balding**

*University of Reading, UK*

**M. Bishop**

*UK HGMP Resource Centre, Cambridge, UK*

**C. Cannings**

*University of Sheffield, UK*

JOHN WILEY & SONS, LTD

Chichester • New York • Weinheim • Brisbane • Singapore • Toronto

Copyright © 2001 by John Wiley & Sons, Ltd.  
Baffins Lane, Chichester,  
West Sussex PO19 1UD, England

National 01243 779777  
International (+44) 1243 779777  
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk  
Visit our Home Page on <http://www.wiley.co.uk> or <http://www.wiley.com>

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1P 0LP, UK, without the permission in writing of the Publisher and the copyright owner, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for the exclusive use by the purchaser of the publication.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

*Other Wiley Editorial Offices*

John Wiley & Sons, Inc., 605 Third Avenue,  
New York, NY 10158-0012, USA

Wiley-VCH Verlag GmbH  
Pappelallee 3, D-69469 Weinheim, Germany

Jacaranda Wiley Ltd, 33 Park Road, Milton,  
Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,  
Jin Xing Distripark, Singapore 129809

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,  
Rexdale, Ontario, M9W 1L1, Canada

*Library of Congress Cataloging-in-Publication Data*

Handbook of statistical genetics/edited by D.J. Balding, M. Bishop, C. Cannings.  
p. cm. – (Wiley series in probability and statistics)  
Includes bibliographical references and index.  
ISBN 0-471-86094-8 (alk.paper)  
1. Genetics – Statistical methods – Handbooks, manuals, etc. I. Balding, D.J. II.  
Bishop, M. III. Cannings, C. (Christopher), 1942-IV. Series.

QH438.4.S73 H36 2000  
576.5'07'27 – dc21

00-043551

*British Library Cataloguing in Publication Data*

A catalogue record for this book is available from the British Library

ISBN 0-471-86094-8

Typeset in 10/12pt Times by Laser Words, Madras, India.

Printed and bound in Great Britain by Bookcraft, Midsomer Norton.

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

---

# Contents

---

<b>Editors' Preface</b>	<b>xxi</b>
<b>List of Contributors</b>	<b>xxiii</b>
<b>Part 1 BIOINFORMATICS</b>	<b>1</b>
<b>1 Chromosome Maps</b>	<b>3</b>
<i>T.P. Speed and H. Zhao</i>	
1.1 Introduction	3
1.2 Genetic Maps	5
1.2.1 Mendel's Two Laws	5
1.2.2 Basic Principles in Genetic Mapping	6
1.2.3 Meiosis, Chromatid Interference, Chiasma Interference, and Crossover Interference	8
1.2.4 Genetic Map Functions	9
1.2.5 Genetic Mapping for Three Markers	9
1.2.6 Genetic Mapping for Multiple Markers	11
1.2.7 Tetrads	14
1.2.8 Half-tetrads	16
1.2.9 Other Types of Data	17
1.2.10 Current State of Genetic Maps	17
1.2.11 Programs for Genetic Mapping	18
1.3 Physical Maps	20
1.3.1 Polytene Chromosomes	20
1.3.2 Cytogenetic Maps	21
1.3.3 Restriction Maps	21
1.3.4 Restriction Mapping Via Optical Mapping	22
1.3.5 Ordered Clone Maps	23
1.3.6 Contig Mapping using Restriction Fragments	25
1.3.7 Sequence-Tagged Site Maps	25
1.4 Radiation Hybrid Mapping	27
1.4.1 Haploid Data	28
1.4.2 Diploid Data	29
1.5 Other Physical Mapping Approaches	31
1.6 Gene Maps	31

1.7	Programs for Physical Mapping	31
	Acknowledgments	32
	References	32
<b>2</b>	<b>Statistical Significance in Biological Sequence Comparison</b>	<b>39</b>
	<i>W.R. Pearson and T.C. Wood</i>	
2.1	Introduction	39
2.2	Statistical Significance and Biological Significance	40
	2.2.1 'Molecular' Homology	41
	2.2.2 Examples of Similarity in Proteins	41
	2.2.3 Inferences from Protein Homology	42
2.3	Estimating Statistical Significance for Local Similarity Searches	43
	2.3.1 Measuring Sequence Similarity	43
	2.3.2 Statistical Significance of Local Similarity Scores	46
	2.3.3 Evaluating Statistical Estimates	57
2.4	Summary: Exploiting Statistical Estimates	61
	Acknowledgments	62
	References	62
<b>3</b>	<b>Probabilistic Models for the Study of Protein Evolution</b>	<b>67</b>
	<i>J.L. Thorne and N. Goldman</i>	
3.1	Introduction	67
3.2	The Dayhoff Model	68
3.3	Amino Acid Composition	70
3.4	Heterogeneity of Replacement Rates Among Sites	71
3.5	Protein Structure	72
3.6	Variation of Preferred Residues Among Sites	74
3.7	Models with a Physicochemical Basis	75
3.8	Codon-Based Models	76
3.9	Covariation Among Sites	77
3.10	Conclusion	79
	Acknowledgments	79
	References	80
<b>4</b>	<b>Statistical Approaches in Eukaryotic Gene Prediction</b>	<b>83</b>
	<i>V. Solovyev</i>	
4.1	Structural Organization and Expression of Eukaryotic Genes	83
4.2	Methods of Functional Signal Recognition	86
	4.2.1 Position-Specific Measures	87
	4.2.2 Content-Specific Measures	88
	4.2.3 Frame-Specific Measures	89
4.3	Performance Measures	89
4.4	Linear Discriminant Analysis	90
4.5	Prediction of Donor and Acceptor Splice Junctions	91
	4.5.1 Splice-Site Characteristics	91
	4.5.2 Donor Splice-Site Characteristics	95
	4.5.3 Acceptor Splice-Site Recognition	98
4.6	Identification of Promoter Regions in Human DNA	99
4.7	Recognition of Poly A Signals	105
4.8	Characteristics for Recognition of 3'-Processing Sites	106
4.9	Identification of Multiple Genes in Genomic Sequences	107

4.10	Discriminative and Probabilistic Approaches to Multiple Gene Prediction	108
4.10.1	HMM-Based Multiple Gene Prediction	108
4.10.2	Pattern-Based Multiple Gene Prediction Approach	111
4.11	Internal Exon Recognition	111
4.12	Recognition of Flanking Exons	112
4.12.1	5'-Terminal Exon-Coding Region Recognition	112
4.12.2	3'-Exon-Coding Region Recognition	113
4.13	Performance of Gene Identification Programs	114
4.14	Using Protein Similarity Information to Improve Gene Prediction	116
4.15	Annotation of Sequences from Genome Sequencing Projects	117
4.16	Infogene: A Gene-Centered Database of Known and Predicted Genes	119
	Acknowledgments	121
	References	122

## **5 Protein Structure 129**

*W.R. Taylor*

5.1	Introduction	129
5.1.1	Historical Background	129
5.1.2	Future Importance	130
5.2	Basic Principles	130
5.2.1	Hydrophobic Core	130
5.2.2	Secondary Structure	131
5.2.3	Protein Architecture and Topology	131
5.2.4	Domains	133
5.3	Structure Comparison and Classification	133
5.3.1	Limits of Structural Similarity	133
5.3.2	Protein Structure Comparison	136
5.3.3	Assessing Significance	138
5.3.4	Classification	141
5.3.5	Towards Automatic Classification	142
5.4	Protein Structure Prediction	143
5.4.1	Homology Modelling	143
5.4.2	Threading	144
5.4.3	Secondary Structure Prediction	145
5.5	Conclusions	145
5.5.1	From Deduction to Induction	145
5.5.2	Finite Biology	146
	References	146

## **Part 2 POPULATION GENETICS 151**

### **6 Mathematical Models in Population Genetics 153**

*C. Neuhauser*

6.1	A Brief History of the Role of Selection	153
6.2	Mutation, Random Genetic Drift, and Selection	154
6.2.1	Mutation	155
6.2.2	Random Genetic Drift	155

6.2.3	Selection	157
6.2.4	The Wright–Fisher Model	157
6.3	The Diffusion Approximation	158
6.3.1	Fixation	161
6.3.2	The Kolmogorov Forward Equation	162
6.3.3	Random Genetic Drift Versus Mutation and Selection	162
6.4	The Infinite Allele Model	163
6.4.1	The Infinite Allele Model with Mutation	163
6.4.2	Ewens’s Sampling Formula	165
6.4.3	The Infinite Allele Model with Selection and Mutation	165
6.5	Other Models of Mutation and Selection	166
6.5.1	The Infinitely Many Sites Model	166
6.5.2	Frequency-Dependent Selection	166
6.5.3	Overlapping Generations	167
6.6	Coalescent Theory	167
6.6.1	The Neutral Coalescent	167
6.6.2	The Ancestral Selection Graph	169
6.6.3	Varying Population Size	172
6.7	Detecting Selection	173
	Acknowledgments	175
	References	175
<b>7</b>	<b>Coalescent Theory</b>	<b>179</b>
	<i>M. Nordborg</i>	
7.1	Introduction	179
7.2	The Coalescent	180
7.2.1	The Fundamental Insights	180
7.2.2	The Coalescent Approximation	183
7.3	Generalizing the Coalescent	186
7.3.1	Robustness and Scaling	186
7.3.2	Variable Population Size	187
7.3.3	Population Structure on Different Time Scales	189
7.4	Geographical Structure	190
7.4.1	The Structured Coalescent	190
7.4.2	The Strong-Migration Limit	191
7.5	Segregation	192
7.5.1	Hermaphrodites	193
7.5.2	Males and Females	194
7.6	Recombination	195
7.6.1	The Ancestral Recombination Graph	195
7.6.2	Properties and Effects of Recombination	199
7.7	Selection	200
7.7.1	Balancing Selection	201
7.7.2	Selective Sweeps	203
7.7.3	Background Selection	204
7.8	Neutral Mutations	205
7.9	Conclusion	205
7.9.1	The Coalescent and ‘Classical’ Population Genetics	205
7.9.2	The Coalescent and Phylogenetics	206

7.9.3	Prospects	208
	Acknowledgments	208
	References	208
<b>8</b>	<b>Inference Under the Coalescent</b>	<b>213</b>
	<i>M. Stephens</i>	
8.1	Introduction	213
8.1.1	Likelihood-based Inference	214
8.2	The Likelihood and the Coalescent	218
8.3	Importance Sampling	220
8.3.1	Likelihood Surfaces	222
8.3.2	Ancestral Inference	223
8.3.3	Application and Assessing Reliability	223
8.4	Markov Chain Monte Carlo	224
8.4.1	Introduction	224
8.4.2	Choosing a Good Proposal Distribution	226
8.4.3	Likelihood Surfaces	226
8.4.4	Ancestral Inference	228
8.4.5	Example Proposal Distributions	229
8.4.6	Application and Assessing Reliability	232
8.5	Conclusions	235
8.5.1	Extensions to More Complex Demographic and Genetic Models	235
8.5.2	Choice of Method	235
8.5.3	Software and Internet Resources	235
	Acknowledgments	236
	References	236
<b>9</b>	<b>Inferences from Spatial Population Genetics</b>	<b>239</b>
	<i>F. Rousset</i>	
9.1	Introduction	239
9.2	Models in Spatial Population Genetics	241
9.2.1	Neutral Models of Geographical Variation	241
9.2.2	Clines	242
9.3	Methods of Inference	243
9.3.1	$F$ -Statistics	243
9.3.2	Likelihood Methods	247
9.4	Inference Under the Different Models	249
9.4.1	Migration Matrix Models	249
9.4.2	Island Model	250
9.4.3	Isolation by Distance	251
9.4.4	Inferences from Clines	254
9.4.5	Other Methods	255
9.5	Spatial Analyses in Practice	256
9.5.1	Objections to the Models	256
9.5.2	How do these Methods Perform?	257
9.6	Software	258
	Acknowledgments	258

Appendix A	Analysis of Variance and Probabilities of Identity	259
Appendix B	Likelihood Analysis of the Island Model	263
	References	265
<b>10</b>	<b>Analysis of Population Subdivision</b>	<b>271</b>
	<i>L. Excoffier</i>	
10.1	Introduction	271
10.2	The Fixation Index $F$	272
10.3	Wright's $F$ -Statistics in Hierarchic Subdivisions	274
	10.3.1 Multiple Alleles	276
	10.3.2 Sample Estimation of $F$ -Statistics	277
	10.3.3 $G$ -Statistics	278
10.4	Analysis of Genetic Subdivision Under an Analysis of Variance Framework	279
	10.4.1 The Model	280
	10.4.2 Estimation Procedure	283
	10.4.3 Dealing with Mutation and Migration Using Identity Coefficients	287
10.5	Relationship Between Different Definitions of Fixation Indexes	287
10.6	$F$ -Statistics and Coalescence Times	290
10.7	Analysis of Molecular Data: The Amova Framework	291
	10.7.1 Haplotypic Diversity	291
	10.7.2 Genotypic Data	294
	10.7.3 Multi-Allelic Molecular Data	294
	10.7.4 Dominant Data	297
	10.7.5 Relation of AMOVA to Other Approaches	298
10.8	Significance Testing	299
	10.8.1 Resampling Techniques	299
	10.8.2 Exact Tests	300
10.9	Related and Remaining Problems	301
	10.9.1 Testing Departure from Hardy–Weinberg Equilibrium	301
	10.9.2 What is the Underlying Genetic Structure of Populations?	301
	Acknowledgments	302
	References	302
<b>11</b>	<b>Linkage Disequilibrium and Recombination</b>	<b>309</b>
	<i>R.R. Hudson</i>	
11.1	Introduction	309
11.2	Tests of Association	311
	11.2.1 Haploid Data, Two Loci	311
	11.2.2 Haploid Data, More than Two Loci	314
	11.2.3 Diploid Data	315
11.3	Properties of Linkage Disequilibrium Under Population Genetic Models	318
	References	323

<b>Part 3</b>	<b>EVOLUTIONARY GENETICS</b>	<b>325</b>
<b>12</b>	<b>Adaptive Molecular Evolution</b>	<b>327</b>
	<i>Z. Yang</i>	
12.1	Introduction	327
12.2	Markov Model of Codon Substitution	329
12.3	Estimation of Synonymous and Non-Synonymous Substitution Rates Between Two Sequences	331
	12.3.1 <i>Ad hoc</i> Methods	331
	12.3.2 Maximum Likelihood Estimation	332
	12.3.3 A Numerical Example and Evaluation of Methods	336
12.4	Likelihood Calculation on a Phylogeny	338
12.5	Detecting Adaptive Evolution Along Lineages	339
	12.5.1 Likelihood Calculation Under Models of Variable $\omega$ Ratios Among Lineages	339
	12.5.2 Adaptive Evolution in the Primate Lysozyme	340
	12.5.3 Comparison with Methods Based on Reconstructed Ancestral Sequences	342
12.6	Inferring Amino Acid Sites Under Diversifying Selection	343
	12.6.1 Likelihood Calculation Under Models of Variable $\omega$ Ratios Among Sites	343
	12.6.2 Positive Selection in the HIV-1 <i>vif</i> Genes	345
	12.6.3 Comparison with Methods Based on Reconstructed Ancestral Sequences	347
12.7	Limitations of Current Methods	347
12.8	Computer Software	348
	Acknowledgments	348
	References	348
<b>13</b>	<b>Genome Evolution</b>	<b>351</b>
	<i>J.F.Y. Brookfield</i>	
13.1	Introduction	351
13.2	The Structure and Function of Genomes	353
	13.2.1 Genome Sequencing Projects	353
	13.2.2 Post-genomics	354
	13.2.3 The Origins and Functions of Introns	356
13.3	The Organization of Genomes	359
	13.3.1 The Relative Positions of Genes: are they Adaptive?	359
	13.3.2 Functional Linkage among Prokaryotes	360
	13.3.3 Gene Clusters	361
	13.3.4 Gene Duplications and Gene Families	362
	13.3.5 Apparent Genetic Redundancy	363
13.4	Population Genetics and the Genome	364
	13.4.1 The Impact of Chromosomal Position on Population Genetic Variability	364
	13.4.2 Codon Usage Bias	365
13.5	Population Genetics of Mobile DNAs	366
	13.5.1 Repetitive Sequences	366
	13.5.2 Transposable Elements: Parasites or Symbionts?	367

13.5.3	Copy Number Control	367
13.5.4	Selfish Transposable Elements and Sex	369
13.5.5	Phylogenies of Transposable Elements	370
13.6	Conclusions	371
	References	372
<b>14</b>	<b>Virus Evolution</b>	<b>377</b>
	<i>Y. Suzuki, A. Wyndham and T. Gojobori</i>	
14.1	Introduction: HIV as a Model for Virus Evolution	377
14.2	Background	378
	14.2.1 HIV and AIDS	378
	14.2.2 HIV Molecular Biology	378
	14.2.3 HIV-1 Co-receptor Usage and Phenotype Switch	379
14.3	Evolutionary Rate	380
	14.3.1 Some Common Estimation Methods	380
	14.3.2 Nucleotide Substitution in HIV-1	381
	14.3.3 The Mutation Rate of HIV-1	386
14.4	Natural Selection	387
	14.4.1 Methods for Examining Natural Selection	387
	14.4.2 What Kind of Selective Mechanisms are at Work in HIV-1?	387
14.5	Phylogenetic Relationships Between HIV and SIV Members	390
	14.5.1 Virus Phylogenies	390
	14.5.2 HIV-1 and the Primate Lentiviruses	390
	14.5.3 HIV-1 Subtypes	391
14.6	Recombination	392
	14.6.1 Escape from Muller's Ratchet	392
	14.6.2 Methods for Detecting Recombination	392
	14.6.3 Recombination in HIV-1	393
14.7	The Molecular Clock and Divergence Dates	394
	14.7.1 Estimating Divergence Events	394
	14.7.2 The Molecular Clock of HIV-1	394
	14.7.3 Divergence Events Among HIV-1 Subtypes and the Primate Lentiviruses	395
14.8	Population Dynamics and Models	397
	14.8.1 Stochastic or Deterministic?	397
	14.8.2 HIV-1 Generation Time	398
	14.8.3 Drug Resistance	399
	14.8.4 Immune Response	399
14.9	Concluding Remarks	403
	References	404
<b>15</b>	<b>Application of the Likelihood Function in Phylogenetic Analysis</b>	<b>415</b>
	<i>J.P. Huelsenbeck and J.P. Bollback</i>	
15.1	Introduction	415
15.2	History	417
	15.2.1 A Brief History of Maximum Likelihood in Phylogenetics	417
	15.2.2 A Brief History of Bayesian Inference in Phylogenetics	418
15.3	Likelihood Function	418

15.4	Developing an Intuition of Likelihood	424
15.5	Method of Maximum Likelihood	426
15.6	Bayesian Inference	429
15.7	Markov Chain Monte Carlo	431
15.8	Assessing Uncertainty of Phylogenies	435
15.9	Hypothesis Testing and Model Choice	436
15.10	Comparative Analysis	437
15.11	Conclusions	438
	References	439
<b>16</b>	<b>Phylogenetics: Parsimony and Distance Methods</b>	<b>445</b>
	<i>D. Penny and M. Hendy</i>	
16.1	Introduction	445
16.2	Data	446
16.2.1	Character State Matrix	447
16.2.2	Genetic Distances	447
16.2.3	Splits (bipartitions)	452
16.2.4	Sampling Error	455
16.3	Theoretical Background	456
16.3.1	Terminology for Graphs and Trees	456
16.3.2	Computational Complexity, Numbers of Trees	458
16.3.3	Three Parts of an Evolutionary Model	461
16.3.4	Stochastic Mechanisms of Evolution	464
16.4	Methods for Inferring Evolutionary Trees	466
16.4.1	Five Desirable Properties for Methods	467
16.4.2	Optimality Criteria	470
16.5	Search Strategies	478
16.5.1	Complete or Exact Searches	478
16.5.2	Heuristic Searches I: Limited (local) Searches	479
16.5.3	Heuristic Searches II: Hill-climbing and Related Methods	481
16.5.4	Quartets and Supertrees	482
16.6	Overview and Conclusions	482
	References	483
<b>Part 4</b>	<b>GENETIC EPIDEMIOLOGY</b>	<b>485</b>
<b>17</b>	<b>Nonparametric Linkage</b>	<b>487</b>
	<i>P. Holmans</i>	
17.1	Introduction	487
17.2	Pros and Cons of Model-Free Methods	488
17.3	Model-Free Methods for Dichotomous Traits	489
17.3.1	Affected Sib-Pair Methods	489
17.3.2	Parameter Estimation and Power Calculation using Affected Sib Pairs	491
17.3.3	Typing Unaffected Relatives in Sib-Pair Analyses	492
17.3.4	Application of Sib-Pair Methods to Multiplex Sibships	493
17.3.5	Methods for Analysing Larger Pedigrees	494

17.3.6	Extensions to Multiple-Marker Loci	495
17.3.7	Inclusion of Covariates	495
17.3.8	Multiple Disease Loci	496
17.3.9	Strategies for Genome Scans	497
17.4	Model-Free Methods for Quantitative Traits	498
17.4.1	Sampling Considerations	499
17.5	Conclusions	499
	References	500
<b>18</b>	<b>The Transmission/Disequilibrium Test</b>	<b>507</b>
	<i>W.J. Ewens and R.S. Spielman</i>	
18.1	Introduction	507
18.2	The Case–Control Test	508
18.3	The Transmission/Disequilibrium Test	509
18.4	Statistical Properties of the TDT	510
18.4.1	Validity	510
18.4.2	Data	511
18.4.3	Form of the Test Statistic	511
18.4.4	Mode of Inheritance	511
18.4.5	Inferring Parental Genotypes	511
18.4.6	Continuous Traits	511
18.4.7	Power	512
18.5	The TDT as a Test of Association	512
18.6	Generalizations of the TDT: More than Two Marker Alleles	513
18.7	Generalizations of the TDT: Unaffected Sibs	514
18.8	The S-TDT Used as a Test of Association	517
	References	518
<b>19</b>	<b>Population Association</b>	<b>519</b>
	<i>D. Clayton</i>	
19.1	Introduction	519
19.2	Measures of Association	520
19.3	Case–Control Studies	522
19.4	Tests for Association	524
19.5	Logistic Regression and Log-Linear Models	527
19.6	Stratification and Matching	529
19.7	Unmeasured Confounding	532
19.8	Multiple Alleles	534
19.9	Haplotype Analysis	537
19.10	Discussion	538
	References	539
<b>20</b>	<b>Linkage Analysis</b>	<b>541</b>
	<i>E.A. Thompson</i>	
20.1	Introduction	541
20.2	The Early Years	542
20.3	The Development of Human Genetic Linkage Analysis	544
20.4	The Pedigree Years: Segregation and Linkage Analysis	546
20.5	Likelihood and Location Score Computation	548

20.6	Linkage Analysis of Complex Traits	553
20.7	Map Estimation, Map Uncertainty, and the Meiosis Model	556
20.8	The Future	559
	Acknowledgment	560
	References	560

## **Part 5 ANIMAL AND PLANT GENETICS** **565**

### **21 Quantitative Trait Loci in Inbred Lines** **567**

*R.C. Jansen*

21.1	Introduction	567
21.1.1	Mendelian Factors and Quantitative Traits	567
21.1.2	The Genetics of Inbred Lines	568
21.1.3	Phenotype, Genotype and Environment	569
21.2	Segregation Analysis	570
21.2.1	Visualization of Quantitative Variation in a Histogram	570
21.2.2	Plotting Mixture Distributions on Top of the Histogram	572
21.2.3	Fitting Mixture Distributions	573
21.2.4	Wanted: QTLs!	574
21.3	Dissecting Quantitative Variation with the Aid of Molecular Markers	575
21.3.1	Molecular Markers	575
21.3.2	Mixture Models	576
21.3.3	Alternative Regression Mapping	580
21.3.4	Highly Incomplete Marker Data	581
21.3.5	ANOVA and Regression Tests	581
21.3.6	Maximum Likelihood Tests	582
21.3.7	Analysis-of-deviance Tests	583
21.3.8	How Many Parameters Can we Fit Safely?	584
21.4	QTL Detection Strategies	585
21.4.1	Model Selection and Genome Scan	585
21.4.2	Single-marker Analysis and Interval Mapping	586
21.4.3	Composite Interval Mapping	588
21.4.4	Multiple-QTL Mapping	589
21.4.5	Uncritical use of Model Selection Procedures	592
21.4.6	Final Comments	592
21.5	Bibliographic Notes	593
	Acknowledgments	594
	References	594

### **22 Mapping Quantitative Trait Loci in Outbred Pedigrees** **599**

*I. Hoeschele*

22.1	Introduction	599
22.2	Linkage Mapping via Least Squares or Maximum Likelihood and Fixed Effects Models	601
22.2.1	Least-Squares	601
22.2.2	Maximum Likelihood	604

22.3	Linkage Mapping via Residual Maximum Likelihood and Random Effects Models	605
22.3.1	Identity-by-Descent Probabilities of Alleles	605
22.3.2	Mixed Linear Model with Random QTL Allelic Effects	609
22.3.3	Mixed Linear Model with Random QTL Genotypic Effects	610
22.3.4	Relationship with other Likelihood Methods	612
22.4	Linkage Mapping via Bayesian Methodology	614
22.4.1	General	614
22.4.2	Bayesian Mapping of a Monogenic Trait	615
22.4.3	Bayesian QTL Mapping	616
22.5	Genotype Sampling in Complex Pedigrees	625
22.6	Fine Mapping of Quantitative Trait Loci	636
22.6.1	Fine Mapping Using Current Recombinations	636
22.6.2	Fine Mapping Using Historical Recombinations	637
22.7	Concluding Remarks	639
	Acknowledgments	639
	References	639
<b>23</b>	<b>Inferences About Breeding Values</b>	<b>645</b>
	<i>D. Gianola</i>	
23.1	Introduction	645
23.2	Landmarks	646
23.2.1	Statistical Genetic Models	646
23.2.2	Best Linear Unbiased Prediction	648
23.2.3	Variance and Covariance Component Estimation	651
23.2.4	BLUP and Unknown Dispersion Parameters	654
23.2.5	Bayesian Procedures	654
23.2.6	Nonlinear, Generalized Linear Models, and Longitudinal Responses	657
23.2.7	Effects of Selection on Inferences	661
23.2.8	Computing Software	663
23.3	Future Developments	664
	Acknowledgments	666
	References	666
<b>24</b>	<b>Marker-Assisted Selection and Introgression</b>	<b>673</b>
	<i>J.C. Whittaker</i>	
24.1	Introduction	673
24.2	Marker-Assisted Selection: Inbred Line Crosses	674
24.2.1	Results	676
24.2.2	Refinements	679
24.3	Marker-Assisted Selection: Outbred Populations	681
24.3.1	MAS via BLUP	682
24.3.2	Comments	683
24.3.3	Within-family MAS	685
24.4	Marker-Assisted Introgression	686
24.4.1	Inbred Line Crosses	687
24.4.2	Outbred Populations	688

24.5	Discussion	689
	Acknowledgments	690
	References	690
<b>Part 6 APPLICATIONS</b>		<b>695</b>
<b>25</b>	<b>Ethics in the Use of Statistics in Genetics</b>	<b>697</b>
	<i>D. Beyleveld</i>	
25.1	Introduction	697
25.2	What is Ethics?	698
	25.2.1 Uses of the Term 'Ethics'	698
	25.2.2 Morality	699
25.3	Normative Moral Theories and Institutionalized Consensus	700
	25.3.1 Normative Moral Theories	700
	25.3.2 Adjudicating Between Normative Moral Theories	702
	25.3.3 Consensus and Legitimation	704
25.4	Uses of Statistics	709
	25.4.1 Use of DNA Analysis as Forensic Evidence	709
	25.4.2 Heritability Studies	715
25.5	Concluding Remarks	718
	References	719
<b>26</b>	<b>Forensics</b>	<b>721</b>
	<i>B.S. Weir</i>	
26.1	Introduction	721
26.2	Principles of Interpretation	722
26.3	Profile Probabilities	724
	26.3.1 Allelic Independence	724
	26.3.2 Allele Frequencies	726
	26.3.3 Joint Profile Probabilities	727
	26.3.4 Dirichlet Distribution	730
26.4	Mixtures	730
26.5	Sampling Issues	733
	26.5.1 Allele Probabilities	733
	26.5.2 Coancestry	734
26.6	Other Forensic Issues	735
	26.6.1 Common Fallacies	735
	26.6.2 Relevant Population	736
	26.6.3 Database Searches	736
	26.6.4 Uniqueness of Profiles	736
26.7	Conclusion	737
	References	738
<b>27</b>	<b>Pharmacogenetics</b>	<b>741</b>
	<i>N.J. Schork, D. Fallin, H.K. Tiwari and M.A. Schork</i>	
27.1	Introduction: The Scope of Pharmacogenetics	742
27.2	General Issues in the Pharmacogenetic Analysis of Clinical Trials	743

27.3	Phenotypic and Outcome Assessment via Mixture Distribution Analysis	746
27.3.1	The Basic Normal Mixture Model	746
27.3.2	Hypothesis Testing	748
27.4	Optimal Genotyping Protocols via Extreme Sampling	749
27.5	Multilocus and Haplotype Analysis	752
27.5.1	The Potential Advantages of Studying Haplotypes	752
27.5.2	Haplotypes: Estimation and Testing	753
27.6	Assessing Sample Homogeneity	755
27.7	Sequential Pharmacogenetic Designs	757
27.7.1	The Basic Model	757
27.7.2	Matched Pair Sequential Pharmacogenetic Trial	758
27.8	Conclusions	762
	Acknowledgments	762
	References	762
<b>28</b>	<b>Statistical Basis of Risk Calculations</b>	<b>765</b>
	<i>R. Chakraborty</i>	
28.1	Introduction	765
28.2	Concept of Risk in the Formulation of Bayesian Inference	766
28.3	Various Stages of Risk Estimation	767
28.3.1	Population-based Risk Estimate	767
28.3.2	Conditional Probability	768
28.3.3	Joint and Posterior Risk Probability	768
28.4	Major Advances in Estimating the Conditional Probability	768
28.4.1	Conditional Risks for Single or Major Gene Defects	768
28.4.2	Conditional/Recurrence Risk for Multifactorial Diseases	769
28.4.3	Factors Affecting the Information Content of Conditional Risk Evaluation	769
28.5	Use of Genetic Data in Other Types of Risk Estimation	772
28.5.1	Radiation-induced Risk	772
28.5.2	Pharmacogenetic and Ecogenetic Aspects of Risk Evaluation	773
28.6	Summary and Concluding Remarks	774
	Acknowledgments	774
	References	774
<b>29</b>	<b>Conservation Genetics</b>	<b>779</b>
	<i>M.A. Beaumont</i>	
29.1	Introduction	779
29.2	Estimating Effective Population Size	780
29.2.1	Estimating $N_e$ Using Two Samples from the Same Population: the Temporal Method	781
29.2.2	Estimating $N_e$ from Two Derived Populations	785
29.2.3	Estimating $N_e$ Using One Sample	790
29.2.4	Inferring Past Changes in Population Size: Population Bottlenecks	793
29.3	Hybridization	798
29.3.1	Admixture	798
29.3.2	Genetic Mixture Modelling	802
	References	808

---

<b>30 Genetic History of the Human Species</b>	<b>813</b>
<i>J.H. Relethford</i>	
30.1 Introduction	813
30.2 Models of Modern Human Origins	814
30.2.1 Replacement – The Recent African Origin Model	815
30.2.2 Continuity – The Multiregional Evolution Model	815
30.2.3 Is There a Middle Ground?	816
30.2.4 The Genetic Evidence	817
30.3 Gene Trees	818
30.3.1 Mitochondrial DNA	819
30.3.2 Other Gene Trees	821
30.3.3 Interpretations of Gene Trees	822
30.3.4 Neandertal DNA	822
30.4 African Genetic Diversity	824
30.4.1 Bottlenecks and Population Age	825
30.4.2 Variation in Effective Population Size	825
30.5 Genetic Distances Between Living Human Populations	828
30.5.1 Levels of Genetic Differentiation	828
30.5.2 Patterns of Genetic Distances	830
30.6 Genetic Demography of the Human Species	832
30.6.1 Species Effective Population Size	832
30.6.2 Temporal Changes in Species Effective Size	834
30.6.3 Species Effective Size-What Does it Mean?	835
30.7 Conclusion	838
Acknowledgments	840
References	840
<b>Index</b>	<b>847</b>

---

# *Editors' Preface*

---

As the Human Genome Project nears its climax, with the (almost) complete sequence of mankind, we are approaching a new stage in our ability to understand genetic forces in man. Within a few months we shall have that complete sequence (to some degree of approximation), at least for a single individual, and this information allied to the advances in technology opens the door to the examination of subtler questions than has been possible before.

With details of the human sequence available there will be an opportunity to understand more details of our evolution. Already the steps from RFLPs, to Microsatellites, to SNPs has seen increasing, though still sketchy, understanding of our heritage. Sequence data in abundance will represent a further level of detail, providing information about the amino acid sequences (via the code) not available in the SNPs.

We have already seen attempts to unravel the genetics of complex diseases where multiple loci interact, and the new data streams will make it easier, or more difficult, to approach issues of susceptibility, severity, drug efficacy and drug side-effects, as well as opening up the possibility of examining many non-disease characteristics to scrutiny. We say "easier, or more difficult" because abundant data allows the potential to examine more complex questions, but also exposes issues of multiple testing, data dredging and the like, to which many are prone, and some addicted.

In parallel with the Human Genome project there have been major sequencing efforts in several other organisms. Some, such as the worm *C. Elegans*, have been completed and many more are in the pipe line. This information will be immensely valuable, both in unpicking the genetic networks within the organisms, but also in understanding broader questions in evolution and classification. These problems again need sophisticated mathematical, statistical and computational tools.

The Handbook is intended to contribute to the development of methods for exploiting this new data explosion. It should be useful both to statisticians interested in genetical applications, and geneticists seeking a deeper knowledge of statistical methods in their field. In order to be able to treat more advanced topics, an elementary knowledge of both statistics and genetics is assumed, roughly up to a first undergraduate course in each area. However, each chapter is intended to provide an accessible introduction to newcomers to the field, with little or no specialist knowledge beyond this.

It is always important for statistical tools to be well understood by those who seek to use them and doubly so in this new context where the data available is so complex and so abundant. We hope therefore that this book, the aim of which is to present the current set of tools and their theoretical underpinning, is timely. We have set ourselves the ambitious goal of covering the full range of statistics in genetics, which we have

identified as falling into six sections, Bioinformatics, Population Genetics, Evolutionary Genetics, Genetic Epidemiology, Animal and Plant Genetics, and Applications. No doubt there are other ways to slice the subject, and topics which others might have included or excluded. No doubt areas we have not even mentioned will become important in the near future, but strong theoretical background always makes it easier to step forward to the next advance.

We have assembled an outstanding set of authors to write the chapters, who have produced work of quality, and hopefully have achieved our aims; only the readership can decide. We should like to thank the authors for their efforts, for meeting their deadlines (and sometimes ours as well), and making our jobs as editors, in the main, relatively straightforward. We would like to express our appreciation to the staff of John Wiley and Sons for running the project smoothly from their end. In particular Sharon Clutton (commissioning editor), Helen Ramsey (publishing editor) and Rob Calver (editorial assistant) have made our jobs much easier than would have otherwise been the case.

**DAVID BALDING**  
**MARTIN BISHOP**  
**CHRIS CANNINGS**  
January 2001

---

# *List of Contributors*

---

**Dr M. Beaumont**

School of Animal and Microbial Sciences  
University of Reading  
Whiteknights  
PO BOX 228  
Reading RG6 6AJ  
UK

Houston, TX 77030  
USA

**Professor D. Clayton**

MRC Biostatistics Unit  
Institute of Public Health  
University Forvie Site  
Robinson Way  
Cambridge CB2 2SR

**Professor D. Beyleveld**

Sheffield Institute of Biotechnological Law and  
Ethics (SIBLE)  
Department of Law  
University of Sheffield  
Crookesmoor Building  
Conduit Road  
Sheffield S10 1FL  
UK

UK

**Professor W.J. Ewens**

Department of Biology  
University of Pennsylvania  
Philadelphia, PA 19104-6018  
USA

**Dr J.P. Bollback**

Department of Biology  
University of Rochester  
Rochester, NY 14627-0211  
USA

**Dr L. Excoffier**

Genetics and Biometry Laboratory  
Department of Anthropology and Ecology  
University of Geneva  
12 rue G. Revilliod  
1227 Geneva  
Switzerland

**Dr J.F.Y. Brookfield**

Institute of Genetics  
Queen's Medical Centre  
University of Nottingham  
Nottingham NG7 2UH  
UK

**D. Fallin**

CWRU School of Medicine  
Department of Epidemiology and  
Biostatistics  
Case Western Reserve University  
Rammelkamp Bldg, 2nd fl  
MetroHealth Medical Center  
2500 MetroHealth Drive  
Cleveland, OH 44109-1998  
USA

**Professor R. Chakraborty**

Human Genetics Center  
The University of Texas School of Public Health  
6901 Bertner Avenue  
Room S-244

**Professor D. Gianola**  
Department of Dairy Science  
1675 Observatory Drive  
University of Wisconsin  
Madison, WI 53706-2054  
USA

**Professor T. Gojobori**  
Laboratory for DNA Analysis  
Center for Information Biology  
National Institute of Genetics  
1111 Yata, Mishima  
Shizuoka-ken 411-8540  
Japan

**Dr N. Goldman**  
Department of Zoology  
University of Cambridge  
Downing Street  
Cambridge CB2 3EJ  
UK

**Professor M.D. Hendy**  
Institute for Fundamental Sciences  
Massey University  
Private Bag 11222  
Palmerston North  
New Zealand

**Professor I. Hoeschele**  
Departments of Dairy Science and Statistics  
Virginia Polytechnic Institute  
2160 Litton Reaves  
Blacksburg, VA 24061-0315  
USA

**Professor P. Holmans**  
Department of Psychiatry  
Washington University  
660 S Euclid Avenue  
Box 8067  
St Louis, MO 63100-1010  
USA

**Professor R. Hudson**  
Department of Ecology and Evolution  
University of Chicago

1101 E. 57th St.  
Chicago, IL 60637  
USA

**Professor J.P. Huelsenbeck**  
Department of Biology  
University of Rochester  
Rochester, NY 14627 0211  
USA

**Dr R.C. Jansen**  
Wageningen UR Centre for Biometry  
Plant Research International  
PO Box 16  
NL-6700  
AA Wageningen  
The Netherlands

**Professor C. Neuhauser**  
School of Mathematics  
University of Minnesota  
251 Vin H  
206 Church Street SE  
Minneapolis, MN 55455  
USA

**Professor M. Nordborg**  
Program in Molecular Biology  
Department of Biological Sciences  
University of Southern California  
Los Angeles, CA 90089-1340  
USA

**Professor W. Pearson**  
Department of Biochemistry and Molecular Genetics  
University of Virginia  
Jordan Hall HS #800733, Room 6-044  
Charlottesville, VA 22908  
USA

**Professor D. Penny**  
Institute of Molecular BioSciences  
Massey University  
Private Bag 11222  
Palmerston North  
New Zealand

**Professor J.H. Relethford**

Department of Anthropology  
State University of New York College at Oneonta  
311 Fitzelle Hall  
Oneonta, NY 13820  
USA

**Dr F. Rousset**

Laboratoire Génétique et Environnement  
Institut des Sciences de l'Évolution  
CCO65, USTL, Place E. Bataillon  
34095 Montpellier Cedex 05  
France

**Professor M.A. Schork**

School of Public Health  
Department of Biostatistics  
University of Michigan  
1420 Washington Heights  
Ann Arbor, MI 48109-2029  
USA

**Dr N.J. Schork**

Department of Statistical Genomics  
The GENSET Corporation  
875 Prospect Street  
Suite 206  
La Jolla, CA 92037-4264  
USA

**Professor V. Solovyev**

Director of Bioinformatics  
EOS Biotechnology  
225A Gateway Boulevard  
South San Francisco, CA 94080  
USA

**Professor T.P. Speed**

Department of Statistics and Program in  
Biostatistics  
367 Evans Hall, #3860  
University of California  
Berkeley, CA 94720-3860  
USA

**Dr R.S. Spielman**

Department of Genetics  
University of Pennsylvania

Philadelphia, PA 19104-6145  
USA

**Dr M. Stephens**

Department of Statistics  
University of Oxford  
1 South Parks Road  
Oxford OX1 3TG  
UK

**Dr Y. Suzuki**

Laboratory for DNA Analysis  
Center for Information Biology  
National Institute of Genetics  
1111 Yata, Mishima  
Shizuoka-ken 411-8540  
Japan

**Professor W.R. Taylor**

Division of Mathematical Biology  
National Institute for Medical Research  
The Ridgeway  
Mill Hill  
London NW7 1AA  
UK

**Professor E. Thompson**

Department of Statistics  
University of Washington  
Box 354322  
Seattle, WA 98195-4322  
USA

**Dr J.L. Thorne**

Department of Statistics  
North Carolina State University  
Box 8203  
Raleigh, NC 27695-8203  
USA

**Dr H.K. Tiwari**

CWRU School of Medicine  
Department of Epidemiology and  
Biostatistics  
Case Western Reserve University  
Rammelkamp Bldg, 2nd fl  
MetroHealth Medical Center

2500 MetroHealth Drive  
Cleveland, OH 44109-1998  
USA

**Professor B.S. Weir**  
Department of Statistics  
North Carolina State University  
Box 8203  
Raleigh, NC 27695-8203  
USA

**Dr J.C. Whittaker**  
Department of Applied Statistics  
University of Reading  
PO Box 240  
Reading RG6 6FN  
UK

**Dr T.C. Wood**  
Clemson University Genomics Institute  
100 Jordan Hall  
Clemson University  
Clemson, SC 29634  
USA

**Dr A. Wyndham**  
Laboratory for DNA Analysis  
Center for Information Biology  
National Institute of Genetics  
1111 Yata, Mishima  
Shizuoka-ken 411-8540  
Japan

**Dr Z. Yang**  
Department of Biology  
Galton Laboratory  
University College London  
4 Stephenson Way  
London NW1 2HE  
UK

**Dr H. Zhao**  
Department of Epidemiology and Public  
Health  
Yale University School of Medicine  
60 College Street  
New Haven, CT 06520  
USA