

---

# Contents

<b>Preface</b> .....	VII
<b>List of Tables</b> .....	XVII
<b>List of Figures</b> .....	XIX
<b>Acronyms</b> .....	XXI
<b>1 Genetic Association Studies</b> .....	1
1.1 Overview of population-based investigations .....	2
1.1.1 Types of investigations .....	2
1.1.2 Genotype versus gene expression .....	4
1.1.3 Population-versus family-based investigations .....	6
1.1.4 Association versus population genetics .....	7
1.2 Data components and terminology .....	7
1.2.1 Genetic information .....	8
1.2.2 Traits .....	11
1.2.3 Covariates .....	12
1.3 Data examples .....	12
1.3.1 Complex disease association studies .....	13
1.3.2 HIV genotype association studies .....	16
1.3.3 Publicly available data used throughout the text .....	18
Problems .....	27
<b>2 Elementary Statistical Principles</b> .....	29
2.1 Background .....	30
2.1.1 Notation and basic probability concepts .....	30
2.1.2 Important epidemiological concepts .....	33
2.2 Measures and tests of association .....	37
2.2.1 Contingency table analysis for a binary trait .....	38
2.2.2 M-sample tests for a quantitative trait .....	44

2.2.3	Generalized linear model	48
2.3	Analytic challenges	55
2.3.1	Multiplicity and high dimensionality	55
2.3.2	Missing and unobservable data considerations	58
2.3.3	Race and ethnicity	60
2.3.4	Genetic models and models of association	61
	Problems	62
<b>3</b>	<b>Genetic Data Concepts and Tests</b>	<b>65</b>
3.1	Linkage disequilibrium (LD)	65
3.1.1	Measures of LD: $D'$ and $r^2$	66
3.1.2	LD blocks and SNP tagging	74
3.1.3	LD and population stratification	76
3.2	Hardy-Weinberg equilibrium (HWE)	78
3.2.1	Pearson's $\chi^2$ -test and Fisher's exact test	78
3.2.2	HWE and population substructure	82
3.3	Quality control and preprocessing	86
3.3.1	SNP chips	86
3.3.2	Genotyping errors	88
3.3.3	Identifying population substructure	89
3.3.4	Relatedness	92
3.3.5	Accounting for unobservable substructure	94
	Problems	95
<b>4</b>	<b>Multiple Comparison Procedures</b>	<b>97</b>
4.1	Measures of error	97
4.1.1	Family-wise error rate	98
4.1.2	False discovery rate	100
4.2	Single-step and step-down adjustments	101
4.2.1	Bonferroni adjustment	102
4.2.2	Tukey and Scheffe tests	105
4.2.3	False discovery rate control	109
4.2.4	The $q$ -value	112
4.3	Resampling-based methods	114
4.3.1	Free step-down resampling	114
4.3.2	Null unrestricted bootstrap	120
4.4	Alternative paradigms	123
4.4.1	Effective number of tests	123
4.4.2	Global tests	125
	Problems	127

<b>5</b>	<b>Methods for Unobservable Phase</b> . . . . .	129
5.1	Haplotype estimation . . . . .	130
5.1.1	An expectation-maximization algorithm . . . . .	130
5.1.2	Bayesian haplotype reconstruction . . . . .	137
5.2	Estimating and testing for haplotype–trait association . . . . .	140
5.2.1	Two-stage approaches . . . . .	140
5.2.2	A fully likelihood-based approach . . . . .	145
	Problems . . . . .	149
	Supplemental notes . . . . .	150
	Supplemental R scripts . . . . .	155
<b>6</b>	<b>Classification and Regression Trees</b> . . . . .	157
6.1	Building a tree . . . . .	157
6.1.1	Recursive partitioning . . . . .	157
6.1.2	Splitting rules . . . . .	158
6.1.3	Defining inputs . . . . .	167
6.2	Optimal trees . . . . .	173
6.2.1	Honest estimates . . . . .	174
6.2.2	Cost-complexity pruning . . . . .	174
	Problems . . . . .	179
<b>7</b>	<b>Additional Topics in High-Dimensional Data Analysis</b> . . . . .	181
7.1	Random forests . . . . .	182
7.1.1	Variable importance . . . . .	183
7.1.2	Missing data methods . . . . .	187
7.1.3	Covariates . . . . .	198
7.2	Logic regression . . . . .	198
7.3	Multivariate adaptive regression splines . . . . .	205
7.4	Bayesian variable selection . . . . .	209
7.5	Further readings . . . . .	211
	Problems . . . . .	212
	<b>Appendix R Basics</b> . . . . .	213
A.1	Getting started . . . . .	213
A.2	Types of data objects . . . . .	216
A.3	Importing data . . . . .	220
A.4	Managing data . . . . .	221
A.5	Installing packages . . . . .	224
A.6	Additional help . . . . .	225
	<b>References</b> . . . . .	227
	<b>Glossary of Terms</b> . . . . .	237
	<b>Glossary of Select R Packages</b> . . . . .	243

<b>Subject Index</b> .....	247
<b>Index of R Functions and Packages</b> .....	251



<http://www.springer.com/978-0-387-89553-6>

Applied Statistical Genetics with R  
For Population-based Association Studies

Foulkes, A.S.

2009, XXIII, 252 p., Softcover

ISBN: 978-0-387-89553-6