
Contents

Preface	xiii
Symbol Description	xv
1 Introduction	1
1.1 Heredity, Genetics, and Genomics	1
1.2 Principles of Population Genomics	2
1.2.1 Units	2
1.2.2 Genome Structures	3
1.2.3 Mutations	9
1.2.4 Drift and Selection	11
1.3 R Packages and Conventions	12
1.4 Required Knowledge and Other Readings	16
2 Data Acquisition	17
2.1 Samples and Sampling Designs	17
2.1.1 How Much DNA in a Sample?	17
2.1.2 Degraded Samples	18
2.1.3 Sampling Designs	19
2.2 Low-Throughput Technologies	22
2.2.1 Genotypes From Phenotypes	22
2.2.2 DNA Cleavage Methods	23
2.2.3 Repeat Length Polymorphism	24
2.2.4 Sanger and Shotgun Sequencing	24
2.2.5 DNA Methylation and Bisulfite Sequencing	25
2.3 High-Throughput Technologies	27
2.3.1 DNA Microarrays	27
2.3.2 High-Throughput Sequencing	27
2.3.3 Restriction Site Associated DNA	28
2.3.4 RNA Sequencing	29
2.3.5 Exome Sequencing	29
2.3.6 Sequencing of Pooled Individuals	30
2.3.7 Designing a Study With HTS	30
2.3.8 The Future of DNA Sequencing	30
2.4 File Formats	31
2.4.1 Data Files	31

2.4.2	Archiving and Compression	32
2.5	Bioinformatics and Genomics	35
2.5.1	Processing Sanger Sequencing Data With <code>sangerseqR</code>	36
2.5.2	Read Mapping With <code>Rsubread</code>	36
2.5.3	Managing Read Alignments With <code>Rsamtools</code>	39
2.6	Simulation of High-Throughput Sequencing Data	41
2.7	Exercises	44
3	Genomic Data in R	47
3.1	What is an R Data Object?	47
3.2	Data Classes for Genomic Data	49
3.2.1	The Class "loci" (<code>pegas</code>)	49
3.2.2	The Class "genind" (<code>adegenet</code>)	50
3.2.3	The Classes "SNPbin" and "genlight" (<code>adegenet</code>)	51
3.2.4	The Class "SnpMatrix" (<code>snpStats</code>)	52
3.2.5	The Class "DNABin" (<code>ape</code>)	53
3.2.6	The Classes "XString" and "XStringSet" (<code>Biostrings</code>)	55
3.2.7	The Package <code>SNPRelate</code>	56
3.3	Data Input and Output	58
3.3.1	Reading Text Files	58
3.3.2	Reading Spreadsheet Files	59
3.3.3	Reading VCF Files	60
3.3.4	Reading PED and BED Files	65
3.3.5	Reading Sequence Files	66
3.3.6	Reading Annotation Files	67
3.3.7	Writing Files	67
3.4	Internet Databases	68
3.5	Managing Files and Projects	69
3.6	Exercises	71
4	Data Manipulation	73
4.1	Basic Data Manipulation in R	73
4.1.1	Subsetting, Replacement, and Deletion	73
4.1.2	Commonly Used Functions	74
4.1.3	Recycling and Coercion	76
4.1.4	Logical Vectors	77
4.2	Memory Management	78
4.3	Conversions	80
4.4	Case Studies	81
4.4.1	Mitochondrial Genomes of the Asiatic Golden Cat	82
4.4.2	Complete Genomes of the Fruit Fly	83
4.4.3	Human Genomes	84
4.4.4	Influenza H1N1 Virus Sequences	85
4.4.5	Jaguar Microsatellites	87
4.4.6	Bacterial Whole Genome Sequences	87

4.4.7	Metabarcoding of Fish Communities	88
4.5	Exercises	91
5	Data Exploration and Summaries	93
5.1	Genotype and Allele Frequencies	93
5.1.1	Allelic Richness	95
5.1.2	Missing Data	96
5.2	Haplotype and Nucleotide Diversity	98
5.2.1	The Class "haplotype"	98
5.2.2	Haplotype and Nucleotide Diversity From DNA Sequences	101
5.3	Genetic and Genomic Distances	103
5.3.1	Theoretical Background	103
5.3.2	Hamming Distance	103
5.3.3	Distances From DNA Sequences	105
5.3.4	Distances From Allele Sharing	105
5.3.5	Distances From Microsatellites	106
5.4	Summary by Groups	107
5.5	Sliding Windows	110
5.5.1	DNA Sequences	110
5.5.2	Summaries With Genomic Positions	112
5.5.3	Package SNPRelate	113
5.6	Multivariate Methods	114
5.6.1	Matrix Decomposition	115
5.6.1.1	Eigendecomposition	115
5.6.1.2	Singular Value Decomposition	117
5.6.1.3	Power Method and Random Matrices	118
5.6.2	Principal Component Analysis	118
5.6.2.1	adeget	119
5.6.2.2	SNPRelate	121
5.6.2.3	flashpcaR	123
5.6.3	Multidimensional Scaling	124
5.7	Case Studies	125
5.7.1	Mitochondrial Genomes of the Asiatic Golden Cat	125
5.7.2	Complete Genomes of the Fruit Fly	127
5.7.3	Human Genomes	134
5.7.4	Influenza H1N1 Virus Sequences	138
5.7.5	Jaguar Microsatellites	142
5.7.6	Bacterial Whole Genome Sequences	149
5.7.7	Metabarcoding of Fish Communities	152
5.8	Exercises	154

6	Linkage Disequilibrium and Haplotype Structure	157
6.1	Why Linkage Disequilibrium is Important?	157
6.2	Linkage Disequilibrium: Two Loci	159
6.2.1	Phased Genotypes	159
6.2.1.1	Theoretical Background	159
6.2.1.2	Implementation in <i>pegas</i>	160
6.2.2	Unphased Genotypes	162
6.3	More Than Two Loci	163
6.3.1	Haplotypes From Unphased Genotypes	163
6.3.1.1	The Expectation-Maximization Algorithm	164
6.3.1.2	Implementation in <i>haplo.stats</i>	164
6.3.2	Locus-Specific Imputation	167
6.3.3	Maps of Linkage Disequilibrium	168
6.3.3.1	Phased Genotypes With <i>pegas</i>	168
6.3.3.2	SNPRelate	170
6.3.3.3	snpStats	171
6.4	Case Studies	172
6.4.1	Complete Genomes of the Fruit Fly	172
6.4.2	Human Genomes	176
6.4.3	Jaguar Microsatellites	177
6.5	Exercises	180
7	Population Genetic Structure	185
7.1	Hardy-Weinberg Equilibrium	185
7.2	<i>F</i> -Statistics	187
7.2.1	Theoretical Background	187
7.2.2	Implementations in <i>pegas</i> and in <i>mmod</i>	189
7.2.3	Implementations in <i>snpStats</i> and in <i>SNPRelate</i>	193
7.3	Trees and Networks	196
7.3.1	Minimum Spanning Trees and Networks	197
7.3.2	Statistical Parsimony	199
7.3.3	Median Networks	200
7.3.4	Phylogenetic Trees	201
7.4	Multivariate Methods	202
7.4.1	Principles of Discriminant Analysis	202
7.4.2	Discriminant Analysis of Principal Components	203
7.4.3	Clustering	207
7.4.4	Maximum Likelihood Methods	207
7.4.5	Bayesian Clustering	210
7.5	Admixture	214
7.5.1	Likelihood Method	214
7.5.2	Principal Component Analysis of Coancestry	217
7.5.3	A Second Look at <i>F</i> -Statistics	218
7.6	Case Studies	222
7.6.1	Mitochondrial Genomes of the Asiatic Golden Cat	222

7.6.2	Complete Genomes of the Fruit Fly	225
7.6.3	Influenza H1N1 Virus Sequences	234
7.6.4	Jaguar Microsatellites	237
7.7	Exercises	239
8	Geographical Structure	241
8.1	Geographical Data in R	241
8.1.1	Packages and Classes	242
8.1.2	Calculating Geographical Distances	242
8.2	A Third Look at F -Statistics	243
8.2.1	Hierarchical Components of Genetic Diversity	243
8.2.2	Analysis of Molecular Variance	246
8.3	Moran I and Spatial Autocorrelation	250
8.4	Spatial Principal Component Analysis	251
8.5	Finding Boundaries Between Populations	255
8.5.1	Spatial Ancestry (tess3r)	255
8.5.2	Bayesian Methods (Geneland)	257
8.6	Case Studies	259
8.6.1	Complete Genomes of the Fruit Fly	259
8.6.2	Human Genomes	260
8.7	Exercises	263
9	Past Demographic Events	265
9.1	The Coalescent	265
9.1.1	The Standard Coalescent	265
9.1.2	The Sequential Markovian Coalescent	268
9.1.3	Simulation of Coalescent Data	269
9.2	Estimation of Θ	275
9.2.1	Heterozygosity	275
9.2.2	Number of Alleles	275
9.2.3	Segregating Sites	276
9.2.4	Microsatellites	276
9.2.5	Trees	277
9.3	Coalescent-Based Inference	278
9.3.1	Maximum Likelihood Methods	278
9.3.2	Analysis of Markov Chain Monte Carlo Outputs	280
9.3.3	Skyline Plots	282
9.3.4	Bayesian Methods	282
9.4	Heterochronous Samples	284
9.5	Site Frequency Spectrum Methods	286
9.5.1	The Stairway Method	288
9.5.2	CubSFS	289
9.5.3	Popsicle	289
9.6	Whole-Genome Methods (psmc)	292
9.7	Case Studies	293

9.7.1	Mitochondrial Genomes of the Asiatic Golden Cat . . .	293
9.7.2	Complete Genomes of the Fruit Fly	298
9.7.3	Influenza H1N1 Virus Sequences	302
9.7.4	Bacterial Whole Genome Sequences	304
9.8	Exercises	306
10	Natural Selection	309
10.1	Testing Neutrality	309
10.1.1	Simple Tests	309
10.1.2	Selection in Protein-Coding Sequences	310
10.2	Selection Scans	313
10.2.1	A Fourth Look at F -Statistics	313
10.2.2	Association Studies (LEA)	314
10.2.3	Principal Component Analysis (pcadapt)	314
10.2.4	Scans for Selection With Extended Haplotypes	315
10.2.5	F_{ST} Outliers	320
10.3	Time-Series of Allele Frequencies	324
10.4	Case Studies	326
10.4.1	Mitochondrial Genomes of the Asiatic Golden Cat . . .	326
10.4.2	Complete Genomes of the Fruit Fly	327
10.4.3	Influenza H1N1 Virus Sequences	335
10.5	Exercises	338
A	Installing R Packages	341
B	Compressing Large Sequence Files	345
C	Sampling of Alleles in a Population	349
D	Glossary	351
	Bibliography	353
	Index	373