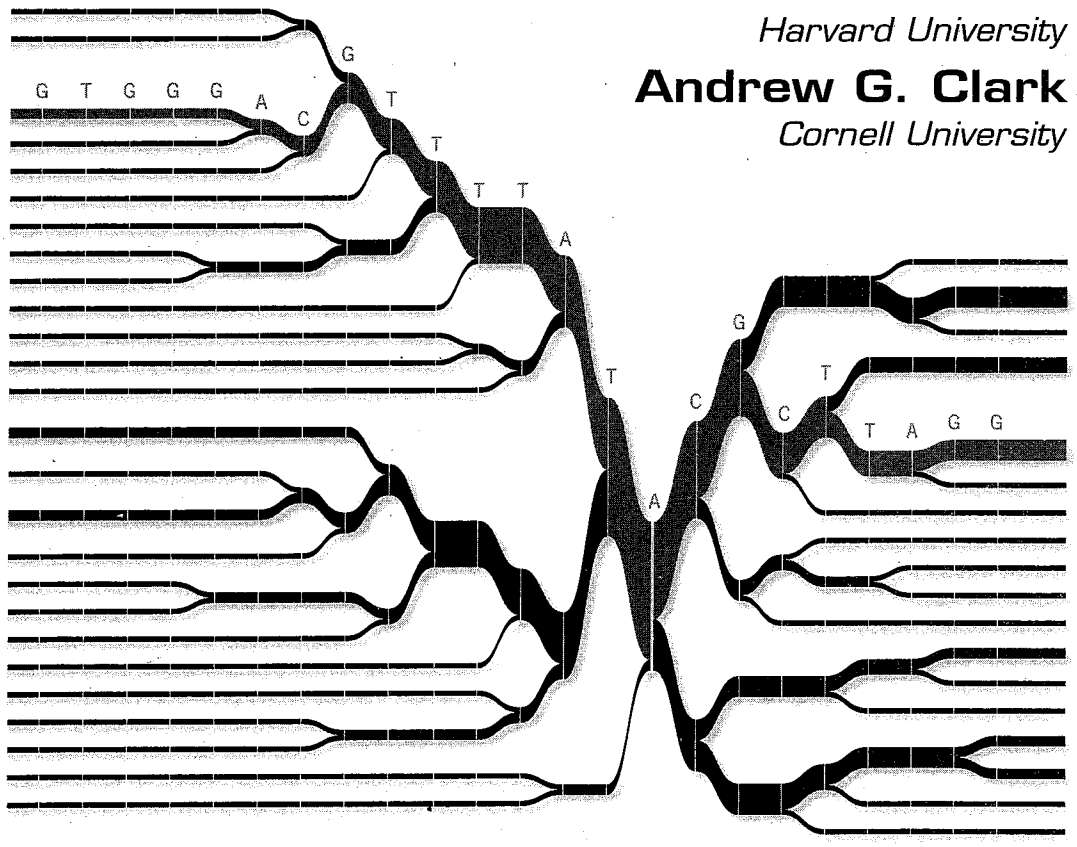
PRINCIPLES OF POPULATION GENETICS

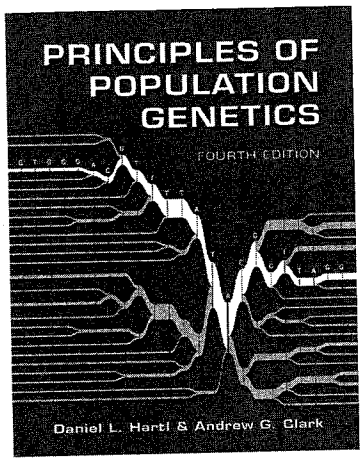
FOURTH EDITION

Daniel L. Hartl
Harvard University

Andrew G. Clark
Cornell University



Sinauer Associates, Inc. Publishers
Sunderland, Massachusetts



The Cover

Bifurcation diagram showing the extent of genetic diversity among a set of haplotypes in a small region of the genome of the malaria parasite *Plasmodium falciparum*. This imaginative visual representation of haplotype diversity was created by Ben Fry, and we are grateful for his permission to use it for the cover. (See Sabeti et al. 2002 for an early application of this style of graphic.)

Principles of Population Genetics, *Fourth Edition*

Copyright © 2007 by Sinauer Associates, Inc.
All rights reserved. This book may not be reproduced in whole or in part without permission from the publisher.

For information, address:

Sinauer Associates, 23 Plumtree Road, Sunderland, MA 01375 U.S.A.
FAX: 413-549-1118
E-mail: publish@sinauer.com
Internet: www.sinauer.com

Library of Congress Cataloging-in-Publication Data

Hartl, Daniel L.
Principles of population genetics / Daniel L. Hartl, Andrew G. Clark. — 4th ed.
p. ; cm.
Includes bibliographical references and index.
ISBN-10: 0-87893-308-5
ISBN-13: 978-0-87893-308-2 (alk. paper)
1. Population genetics. 2. Quantitative genetics. 3. Population genetics—Problems, exercises, etc. I. Clark, Andrew G., 1954- II. Title.
[DNLM: 1. Genetics, Population. 2. Variation (Genetics) QU 450 H331p 2007]

QH455.H36 2007
576.5'8—dc22

2006036153

7 6 5 4 3 2 1

To Christine and Barbara

TABLE OF CONTENTS

PREFACE xii

1 GENETIC AND PHENOTYPIC VARIATION 3

1.1 RELEVANCE OF POPULATION GENETICS 4

1.2 PHENOTYPIC VARIATION IN NATURAL POPULATIONS 5

Continuous Variation: The Normal Distribution 5

Mean and Variance 6

The Central Limit Theorem 8

Discrete Mendelian Variation 10

1.3 MULTIPLE-FACTOR INHERITANCE 12

1.4 MAINTENANCE OF GENETIC VARIATION 14

1.5 MOLECULAR POPULATION GENETICS 15

Electrophoresis 16

Allele Frequencies and Genotype Frequencies 18

Polymorphism and Heterozygosity 20

Allozyme Polymorphisms 21

Inferences from Allozyme Polymorphisms 23

1.6 POLYMORPHISMS IN DNA SEQUENCES 25

Restriction Enzymes 25

The Polymerase Chain Reaction 28

Single Nucleotide Polymorphisms 32

Synonymous and Nonsynonymous Polymorphisms 32

Segregating Sites and Nucleotide Mismatches 34

1.7 UTILITY OF GENETIC POLYMORPHISMS 38

2 Organization of Genetic Variation 45

2.1 Random mating 46
Nonoverlapping Generations 47

2.2 The Hardy-Weinberg Principle 48
Random Mating of Genotypes versus Random Union of Gametes 50
Implications of the Hardy-Weinberg Principle 52

2.3 Testing for Hardy-Weinberg Equilibrium 54
Difficulties in Testing for Hardy-Weinberg Equilibrium 54
Complications of Dominance 62
Frequency of Heterozygotes 66

2.4 Extensions of the Hardy-Weinberg Principle 67
Three or More Alleles 67
X-Linked Genes 70

2.5 Linkage and Linkage Disequilibrium 73
Difficulties in Testing for Linkage Equilibrium 82
Relative Measures of Linkage Disequilibrium: D′ and r^2 83

2.6 Causes of Linkage Disequilibrium 84
Linkage Disequilibrium Due to Population Admixture 85
Linkage Disequilibrium Due to Reduced Recombination 86

3 Random Genetic Drift 95

3.1 Random Genetic Drift and Binomial Sampling 95

3.2 The Wright–Fisher Model of Random Genetic Drift 102

3.3 The Diffusion Approximation 105
An Approach Looking Forward 106
An Approach Looking Backward 110
Absorption Time and Time to Fixation 112

3.4 Random Drift in a Subdivided Population 113

3.5 Effective Population Size 121
Fluctuation in Population Size 121
Unequal Sex Ratio, Sex Chromosomes, Organelle Genes 123
Variance in Offspring Number 126
Effective Size of a Subdivided Population 127

3.6 Gene Trees and Coalescence 128
Coalescent Effective Size 133
Coalescence with Population Growth 134
Coalescent Models with Mutation 136
Applications of Coalescent Methods 137

3.7 Theoretical Implications of Coalescence 138
Coalescent Models with Recombination 140
Linkage Disequilibrium Mapping 143

4 Mutation and the Neutral Theory 151

4.1 Mutation 151
Irreversible Mutation 152
Reversible Mutation 156

4.2 Mutation and Random Genetic Drift 158
Probability of Fixation of a New Neutral Mutation 160

4.3 The Neutral Theory of Molecular Evolution 161

4.4 The Infinite-Alleles Model 162
The Ewens Sampling Formula 166
The Ewens-Watterson Test 168

4.5 Infinite-Sites Model 172
Nucleotide Polymorphism and Nucleotide Diversity 175
Tajima's D Statistic 176
The Fu and Li Test of Fit to Neutral Coalescence 178

4.6 Mutation and Recombination 182
A Model for the Evolutionary Benefit of Recombination 183
Muller's Ratchet 185
Piecewise Recombination in Bacteria 187
Animal Mitochondrial DNA 189

5 Darwinian Selection 199

5.1 Selection in Haploid Organisms 200
Discrete Generations 200
Continuous Time 204
Change in Allele Frequency in Haploids 205
Darwinian Fitness and Malthusian Fitness 206

5.2 Selection in Diploid Organisms 206
Change in Allele Frequency in Diploids 208
Marginal Fitness and Selection with Multiple Alleles 212
Application to the Evolution of Insecticide Resistance 215

5.3 Equilibria with Selection 215
Overdominance 216
Local Stability 220
Heterozygote Inferiority 222
Stable Equilibria with Multiple Alleles 223
Adaptive Topography and the Role of Random Genetic Drift 225

5.4 Mutation-Selection Balance 226
Equilibrium Allele Frequencies 226
The Haldane-Muller Principle 228

5.5 More Complex Types of Selection 229
Differential Selection in the Sexes 229
X-Linked Genes 230
Frequency-Dependent Selection 230
Density-Dependent Selection 231
Fecundity Selection 231
Age-Structured Populations 232
Heterogenous Environments and Clines 232
Diversifying Selection 234
Gametic Selection 236
Meiotic Drive 236
Multiple Loci and Gene Interaction: Epistasis 239
Evolution of Recombination Rate 241
Sexual Selection 241
Kin Selection 243

5.6 Interdeme Selection in Geographically Subdivided Populations 245

5.7 Selection in a Finite Population 248
Weak Selection and the Nearly Neutral Theory 248
Genetic "Draft" 251

6 Inbreeding, Population Subdivision, and Migration 257

6.1 Inbreeding 257
The Inbreeding Coefficient 258
Genotype Frequencies with Inbreeding 259
Genetic Effects of Inbreeding 265
Calculation of the Inbreeding Coefficient from Pedigrees 268
Regular Systems of Mating 272

6.2 Population Subdivision 275
Reduction in Heterozygosity Due to Population Subdivision 276
Average Heterozygosity 278
Wright's F Statistics 281
Linanthus Revisited: Evidence for Selection Associated with Flower Color 285

Inference of Population Structure from Multilocus Genotype Data 286

6.3 The Wahlund Principle 288

Wahlund's Principle and the Fixation Index 291

Genotype Frequencies in Subdivided Populations 291

Relation between the Inbreeding Coefficient and the F Statistics 293

6.4 Assortative Mating 294

6.5 Migration 295

One-Way Migration 295

The Island Model of Migration 297

How Migration Limits Genetic Divergence 299

Estimates of Migration Rates 303

Coalescence-Based Estimates of Migration 303

Migration-Selection Balance 307

7 Molecular Population Genetics 317

7.1 The Neutral Theory and Molecular Evolution 318

Theoretical Principles of the Neutral Theory 318

7.2 Estimating Rates of Molecular Sequence Divergence 321

Rates of Amino Acid Replacement 322

Rates of Nucleotide Substitution 326

Statistical Fitting of Nucleotide Substitution Models 329

7.3 The Molecular Clock 329

Variation among Genes in the Rate of the Molecular Clock 332

Variation across Lineages in Clock Rate 334

The Generation-Time Effect 337

The Overdispersed Molecular Clock and the Neutral Theory 338

The Nearly Neutral Theory 339

7.4 Patterns of Nucleotide and Amino Acid Substitution 340

Calculating Synonymous and Nonsynonymous Substitution Rates 340

Codon Substitution Models 343

Observations of Synonymous and Nonsynonymous Substitution Rates 345

Polymorphism within Species 347

Implications of Codon Usage Bias 349

7.5 Polymorphism and Divergence in Nucleotide Sequence—The McDonald-Kreitman and HKA Tests 351

Polymorphism and Divergence in Noncoding Sequences 354

Impact of Local Recombination Rates 354

Substitution Models for Structural RNA Genes 356

7.6 Gene Genealogies 358

Hypothesis Testing Using Trees 358

7.7 Mitochondrial and Chloroplast DNA Evolution 362

Chloroplast DNA and Organelle Transmission in Plants 363

Maintenance of Variation in Organelle Genomes 364

Evidence for Selection in mtDNA 365

7.8 Molecular Phylogenetics 366

Algorithms for Phylogenetic Tree Reconstruction 366

Distance Methods versus Parsimony 369

Bootstrapping and Statistical Confidence in a Tree 370

Bayesian Methods 370

Trans-Species Polymorphism 371

7.9 Multigene Families 372

Concerted Evolution 374

Subfunctionalization 375

Birth-and-Death Process 376

8 Evolutionary Quantitative Genetics 385

8.1 Types of Quantitative Traits 387

8.2 Resemblance between Relatives and the Concept of Heritability 388

8.3 Artificial Selection and Realized Heritability 394

Contribution of New Mutations to the Response to Selection 396

Prediction Equation for Individual Selection 397

Limits to Selection 400

8.4 Genetic Models for Quantitative Traits 403

Change in Allele Frequency 412

Change in Mean Phenotype 413

Linearity of Response 415

8.5 Components of Phenotypic Variance 416

Genetic and Environmental Sources of Variation 416

Components of Genotypic Variation 422

8.6 Covariance among Relatives 426

Twin Studies and Inferences of Heritability in Humans 430

Estimation of Genetic Variance Components in Natural Populations 432

8.7 Norm of Reaction, Threshold Traits, and Genetic Correlation 433

Norm of Reaction and Phenotypic Plasticity 433

Threshold Traits: Genes as Risk Factors in Disease 437

Genetic Correlation and Correlated Response 439

8.8 Evolutionary Quantitative Genetics 443

Inference of Selection from Phenotypic Data 443

Evolution of Multiple Intercorrelated Traits 444

Random Genetic Drift and Phenotypic Evolution 446

Mutational Variance and Mutation-Accumulation Experiments 448

Mutation-Selection Balance for Quantitative Traits 449

8.9 Genes that Affect Quantitative Traits 452

The Number of Genes Affecting Quantitative Traits 453

Methods for Mapping QTLs 455

9 Population Genomics 469

9.1 Evolution of Genome Size and Composition 470

Organismic Complexity and the C-Value Paradox 470

Base Composition of Genomic DNA 471

9.2 Genome-Wide Patterns of Polymorphism 473

Excess Polymorphism in Subtelomeric Regions 473

Polymorphism and Rates of Recombination 475

Hitchhiking versus Background Selection 477

Linkage Disequilibrium and Haplotype Structures 481

Decline of Linkage Disequilibrium with Genetic Distance 483

9.3 Differences between Species 484

Comparison of Nonsynonymous and Synonymous Divergence 484

Positive Selection 486

Exploiting a Phylogenetic Signal 487

Polymorphism and Divergence 487

Compensated Pathogenic Deviations 489

Structure-Function Analysis 491

9.4 Sexual Selection and the Sex Chromosomes 492

Faster-Male Molecular Evolution 493

Molecular Evolution of Genes in the X Chromosome 495
Haldane's Rule 495
Demasculinization of the X Chromosome 497

9.5 Transposable Elements 498
Diverse Types of Transposable Elements 498
Factors Controlling the Population Dynamics of Transposable Elements 501
Insertion Sequences and Composite Transposons in Bacteria 502
Transposable Elements in Eukaryotes 505
Population Dynamics of Transposable Elements 505
Nonuniformity of Transposition Rates 506
Horizontal Transmission of Transposable Elements 511

10 Human Population Genetics 519

10.1 Human Polymorphism 520
Public SNP Resources and the HapMap Project 521

10.2 Population Genetic Inferences from Human SNPs 524
Ascertainment Bias of SNP Genotypes 524
Departures from Hardy-Weinberg Frequencies 527
Site Frequency Spectrum and Human Population Growth 527
Rooting Human Polymorphism 529
Inference of Inhomogeneities in the Mutation Process 530
Inferences about Male and Female Mutation Rates 531

10.3 Linkage Disequilibrium across the Human Genome 532
The Landscape of Human Linkage Disequilibrium 532
Inferences about Local Rates of Recombination 537

10.4 Population Structure Inferred from Human Polymorphism 539
Multilocus Methods of Inference of Stratification 540
Heterogeneity in Linkage Disequilibrium across Human Populations 542
Linkage Disequilibrium in Admixed Populations: Admixture Mapping 544
Inbred Populations and Homozygosity Mapping 544

10.5 Mendelian Disease and Population Genetics 547
Mutation-Selection Balance 548
Dating the Origin of Mutant Alleles 548

10.6 Genetic Basis for Variation in Risk of Complex Disease 550
Mapping Methods Based on Linkage 550
Linkage Disequilibrium Mapping 552
Genome-Wide Association Studies 553

10.7 Seeking Signatures of Human-Specific Genetic Adaptations 554
Interspecific Divergence 554
McDonald-Kreitman and Poisson Random Field Tests 555
Local Distortions in Linkage Disequilibrium 556
F_{ST} Tests 557
Genome Scans for Selection-Skewed Site Frequency Spectrum 557

10.8 Human Origins 558
Neanderthal Genome Sequence 562

Answers to Chapter-End Problems 567
Bibliography 597
Author Index 629
Subject Index 635

PREFACE

WHY POPULATION GENETICS, WHY NOW?

This Fourth Edition of *Principles of Population Genetics* provides an introduction to population genetics as it exists today. It is intended for students at the undergraduate and graduate levels as well as newcomers from fields like human genetics. The presentation is balanced between theory and observation, and has been written to optimize its use as a teaching aid. Our emphasis is on critical and analytical thinking about real data. We have therefore included numerous problems interspersed within the chapters to encourage students to stop for a moment to think about what they have just learned and how it applies to real situations.

Modern population genetics has been revamped and revitalized by three different but related revolutions in ideas and technology. One revolution was *conceptual* and turned theoretical population genetics on its head. Another was *empirical* and gave population geneticists vast amounts of data to analyze. And the third, *computational*, allowed the conceptual advances and the new data to come together.

The conceptual revolution emerged from the theory of coalescence, which established a framework for thinking about populations and genes in terms of their evolutionary history. This seemingly simple change in perspective was of major importance. Because each of the alleles present in a population at any time descends from an allele that existed in the past, the composition of a sample of alleles taken from a population at any time contains important

information about the evolutionary history of the alleles, and therefore information about the history of the population.

The empirical revolution resulted from genomics, which changed population genetics in three distinct but critically important ways. The first was the determination of the complete DNA sequences of many genomes, including the human genome and that of key model organisms along with those of closely related species. The second was the use of DNA sequencing technology to discover single-nucleotide polymorphisms (SNPs, pronounced *snips*) throughout the genome. Such polymorphisms are important because one of the goals of population genetics is to understand the evolutionary and biological significance of genetic variation. The third contribution of genomics was its high throughput and cost-effective methods for determining the genotypes of hundreds of thousands of SNPs in hundreds or thousands of individuals. This technology provides population geneticists with the substrate for identifying genes affecting complex traits in humans and other organisms. These traits include genetic risk factors for disease, genes affecting polygenic traits in human parasites or agriculturally important plants and animals, and genes that have been targets of natural selection, including those genes whose activities distinguish the physical and mental development and behavior of humans from our closest primate relatives.

The third revolution was computational, including not only vastly increased power for sheer number-crunching but also the explosion of the Internet as a medium for communication among researchers as well as a repository of data and other informational resources. The increased computational power has resulted in widespread application of such methods as Markov-chain Monte Carlo simulation, which allows maximum likelihood and Bayesian methods to be applied to huge data sets for statistical inference on a scale that not long ago would have been considered unimaginable. As one insightful high-school senior has described statistics, "All unrigged data is messy in its own right, and statistics is a collection of methods by which to gather and make sense of that mess artfully in at least one helpful and meaningful way." To make optimal use of statistical resources on the Internet, the text includes references to many vital web resources that connect the material to up-to-the-minute progress in modern population genetics.

- Chapter 1 examines the fundamental observations that motivated the origins of population genetics. It includes the apparent disconnect between dichotomous phenotypic variation of the sort studied by Mendel and continuous phenotypic variation of the sort that so fascinated Galton. The chapter discusses various hypotheses for the origin and maintenance of genetic variation as well as the ways that modern molecular methods are used to discover and quantify the amounts and structure of genetic variation in natural populations.
- Chapter 2 focuses on the relationship between allele frequencies and genotype frequencies, with special emphasis on random mating. We consider statistical pitfalls of testing for random mating both when small

sample sizes and when large-scale simultaneous tests are carried out on hundreds of thousands of SNPs throughout the genome. The chapter also discusses linkage disequilibrium, its causes and measurement, why it matters, and how it is used.

- Chapter 3 considers random genetic drift in the context of the classical Wright-Fisher model. We discuss the mechanisms that cause random genetic drift and examine how diffusion theory was originally deployed to obtain key theoretical results. We then introduce the concept of coalescence and show how this approach is particularly apt for the analysis of molecular data.
- Chapter 4 adds mutation and lays down the foundations for the neutral theory of molecular evolution. We show how mutation and random genetic drift act together to determine the fate of neutral or nearly neutral alleles, and we also consider the characteristics of samples drawn from natural populations at steady state.
- Chapter 5 addresses the mechanisms of Darwinian selection in detail. The chapter brings together the theoretical foundations and empirical observations of the dynamics of gene-frequency change under the action of natural selection. We develop haploid and diploid cases and the concepts of equilibrium, stability, and context dependence. Not only do we develop classical models of mutation-selection balance, but we also consider models of natural selection in many other guises.
- Chapter 6 examines population subdivision and its consequences for the distribution of genetic variation among subpopulations, including the hierarchical F statistics used in estimating these effects. We examine migration in the context of population subdivision, along with the most up-to-date methods for estimating migration rates from natural populations.
- Chapter 7 focuses on molecular evolutionary genetics, currently one of the most active areas for theory and experiment. The unifying theme is the neutral theory; we closely examine the correspondence between the data and theory. Our goal is to give the student a firm grasp of the fundamentals and a deep enough understanding of the principles to identify important gaps in our knowledge.
- Chapter 8 covers the implications of quantitative genetics from an evolutionary perspective. A compelling argument for using quantitative genetics for the study of evolution is that adaptive evolution takes place at the level of the phenotype, and quantitative genetics provides the tools for understanding transmission of phenotypic traits. We give theoretical quantitative genetics special importance because of the paradoxes it raises in contrasting evolution at the levels of the phenotype and of the DNA sequence.
- Chapter 9 is completely new. It covers the exciting field of population genomics, which refers to the analysis of population genetic principles on a genome-wide scale. Included are discussions of genome size and com-

position, genome-wide patterns of polymorphism, linkage disequilibrium, evolutionary divergence, various aspects of the population genetics of speciation, and an extended discussion of the population genetics of transposable elements.

- Chapter 10 is also completely new and shows how the principles of population genetics are actively being applied to human populations to identify genetic risk factors for complex diseases and to understand human evolutionary history and migration across the planet.

Throughout the book we have taken special care to explain the biological motivation behind the theoretical models so that the models do not simply materialize out of thin air, and to explain in plain English the implications of the results. Many concepts are illustrated by numerical examples, using actual data wherever possible. Special topics and examples are often set off from the text as boxed problems whose solutions are explained step by step. Every chapter ends with about 20 problems, graded in difficulty, and solutions worked in full appear at the end of the text.

ACKNOWLEDGMENTS

DLH is deeply indebted to four long-term collaborators—Elena R. Lozovsky, Stanley A. Sawyer, Dyann F. Wirth, and Sarah K. Volkman—as well as to numerous graduate students and postdoctoral fellows, past and present, for their continuing support, as well as for sharing their infectious energy, curiosity, creativity, imagination, and the fruits of their hard work. They are too numerous to name individually, but please consult www.oeb.harvard.edu/hartl/lab/ for a roster as well as brief descriptions of our current research projects. Special thanks to current postdoctoral fellows and especially to current graduate students Trevor Bedford, Kyle Brown, Mauricio Cameiro, Marna Costanzo, Kerry Geiler, Matthew Hegreness, Sarah Kingan, Christian Landry, Bernardo Lemos, Rebekah Rogers, and Martine Zilversmit for their patience and understanding.

AGC would like to thank his current lab members, especially graduate students Erin Hill, Amanda Larracuente, Kirk Lohmueller, Tim Sackton, Sarah Stockwell, and Xu Wang for their inspiring drive, independence, hard work, and their understanding during a particularly challenging rewrite. Other members of my lab group, past and present, have been a great source of motivation to make this book reflect the joy and depth of pursuing a life in population genetics research. Please see www.mbg.cornell.edu/faculty-staff/faculty/clark.cfm for details.

The authors also gratefully acknowledge the professionalism and patience of the Sinauer team, especially Kathaleen Emerson and Sydney Carroll for overtime work and attention to detail, Janice Holabird for clean and uncluttered pages, Chris Small for production, Bobbie Lewis for copyediting, and Andy Sinauer for encouragement and support.

PRINCIPLES OF POPULATION GENETICS

FOURTH EDITION

1.1 Relevance of Population Genetics 4

1.2 Phenotypic Variation in Natural Populations 5

Continuous Variation: The Normal Distribution 5

Mean and Variance 6

The Central Limit Theorem 8

Discrete Mendelian Variation 10

1.3 Multiple-Factor Inheritance 12

1.4 Maintenance of Genetic Variation 14

1.5 Molecular Population Genetics 15

Electrophoresis 16

Allele Frequencies and Genotype Frequencies 18

Polymorphism and Heterozygosity 20

Allozyme Polymorphisms 21

Inferences from Allozyme Polymorphisms 23

1.6 Polymorphisms in DNA Sequences 25

Restriction Enzymes 25

The Polymerase Chain Reaction 28

Single Nucleotide Polymorphisms 32

Synonymous and Nonsynonymous Polymorphisms 32

Segregating Sites and Nucleotide Mismatches 34

1.7 Utility of Genetic Polymorphisms 38

CHAPTER 1

GENETIC AND PHENOTYPIC VARIATION

The science of **population genetics** deals with Mendel's laws and other genetic principles as they apply to entire populations of organisms. The organisms may be human beings, animals, plants, or microbes. The populations may be natural, agricultural, or experimental. The environment may be city, farm, field, or forest. The habitat may be soil, water, or air. Because of its wide-ranging purview, population genetics cuts across many fields of modern biology. A working knowledge of population genetics has become essential in genetics, genomics, evolutionary biology, computational biology, systematics, plant breeding, animal breeding, ecology, natural history, forestry, horticulture, conservation, and wildlife management. A basic understanding of population genetics is also useful in medicine, law, biotechnology, molecular biology, cell biology, sociology, and anthropology.

Population genetics also includes the study of the various forces that result in evolutionary changes in species through time. Individual organisms are characterized by their **genotype**, or their genetic constitution, and by their **phenotype**, or the traits that they manifest. There is often a complex relationship between genotype and phenotype, because phenotype may depend on the interactions of different genes as well as effects of the environment. By defining the genetic framework within which evolution takes place, the principles of population genetics are basic to a broad evolutionary perspective on biology. From an experimental point of view, evolution provides a wealth of testable hypotheses for all other branches of biology. Many oddities in biolo-

gy become comprehensible in the light of evolution: They result from shared ancestry among organisms, and they attest to the unity of life on earth.

1.1 RELEVANCE OF POPULATION GENETICS

Practical applications of population genetics are extensive. Many applications, particularly those relevant to human beings, also have important implications in ethics and social policy. Among the applications of population genetics in medicine, agriculture, conservation, and research are:

- Genetic counseling of parents and other relatives of patients with hereditary diseases.
- Genetic mapping and identification of genes for disease susceptibility in human beings, including breast cancer, colon cancer, diabetes, and schizophrenia.
- Implications of population screening for carriers of disease genes, confidentiality of results, and maintenance of health insurability.
- Statistical interpretation of the significance of matching DNA types found between a suspect and a blood or semen sample from the scene of a crime.
- Design of studies to sample and preserve a record of genetic variation among human populations throughout the world.
- Improvement in the performance of domesticated animals and crop plants.
- Organization of mating programs for the preservation of endangered species in zoos and wildlife refuges.
- Sampling and preservation of germ plasms of potentially beneficial plants and animals that may soon vanish from the wild.
- Interpretation of differences in the nucleotide sequences of genes or amino acid sequences of proteins among members of the same or closely related species.
- Analysis of genes and genomes among diverse species to determine their evolutionary relationships and to test hypotheses about the evolutionary process.

Genetic variation in populations became a subject of scientific inquiry in the late nineteenth century prior even to the rediscovery of the Mendel (1866) paper on heredity. The leading exponent of the study of hereditary differences in human populations was Francis Galton (1822–1911). Galton was a pioneer in the application of statistics to biology. He used statistical methods to study physical traits such as eye color and fingerprint ridges as well as behavioral traits such as temperament and musical ability. Galton was among the first to examine the statistical relations between the distributions of phenotypic traits in successive generations. He is regarded as the founder of biometry, the application of statistics to biological problems. Prior to 1900,

Galton's work was carried out with no knowledge of the theory of inheritance proposed by Gregor Mendel (1822–1884).

1.2 PHENOTYPIC VARIATION IN NATURAL POPULATIONS

Galton and Mendel exemplify opposite approaches to the study of inherited traits. Mendel's point of departure in the study of genetics was **discrete variation**, in which phenotypic differences among organisms can be assigned to a small number of clearly distinct classes, such as round versus wrinkled peas. Galton's point of departure was **continuous variation**, in which the phenotypes of organisms are measured on a quantitative scale, like height or weight, and in which the phenotypes grade imperceptibly from one category into the next. As material for the study of phenotypic variation, Galton's choice was good: Most of the differences among normal people that are visible to the unaided eye are differences in continuous traits—height, weight, skin color, hair color, facial features, running speed, shoe size, and so forth. The same is true of phenotypic variation in other organisms. On the other hand, as material for the study of genetic variation, Mendel's choice was good (Hartl and Orel 1992; Orel 1996): The result of segregation is revealed most clearly in pedigrees of discrete, simple Mendelian traits. By **segregation** we mean that the two forms of a gene present in an individual, say *A* and *a*, separate in the formation of reproductive cells, so that each reproductive cell receives exactly one copy of either *A* or *a*.

Continuous Variation: The Normal Distribution

With continuous traits, not only do the phenotypes grade into one another, but the traits also usually present difficulties for genetic analysis. The problems are of two principal types:

- Most continuous traits are influenced by slightly different DNA sequences in two or more genes, hence the segregation of the differences in one gene in pedigrees is obscured by the segregation of differences in other genes that affect the trait.
- Most continuous traits are influenced by environmental factors as well as by genes, and so genetic segregation is obscured by environmental effects.

These problems are not insurmountable in organisms with a sufficiently high density of genetic markers scattered throughout the genome (the complement of chromosomes) because the genetic markers can be tracked in pedigrees along with the continuous trait of interest. Organisms with sufficiently dense genetic maps include human beings, laboratory animals, and many domesticated animals and crop plants.

In Galton's time, however, studies of continuous traits based on linkage of genetic markers along the chromosome were unknown. Why, then, did

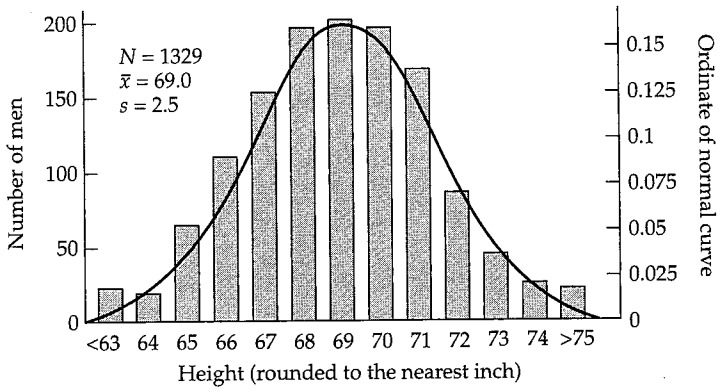


FIGURE 1.1 Distribution of heights among 1329 British men. (Data from Galton 1889.)

Galton focus on continuous traits? The reasons is that they have a sort of regularity—a statistical predictability—of their own. For many continuous traits, when the phenotypes are grouped into suitable intervals and plotted as a bar graph, the distribution of phenotypes conforms closely to the familiar bell-shaped, symmetrical curve known as the **normal distribution**. For example, a bar graph of Galton's data on the heights of 1329 men, rounded to the nearest inch, is plotted in Figure 1.1. The smooth curve is the normal distribution that best fits the data. The equation of the normal curve, more properly called the *normal probability density function*, is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad [1.1]$$

where x ranges from $-\infty$ to $+\infty$, and where $\pi = 3.14159$ and $e = 2.71828$ are constants. The location of the peak of the distribution along the x axis is determined by the parameter μ, which is the **mean**, or average, of the phenotypic values. The degree to which the phenotypes are clustered around the mean is determined by the parameter σ^2, which is the **variance** of the distribution. Mathematically, the variance is the average of the squared difference of each phenotypic value from the mean; that is, it is the average of the values of $(x - \mu)^2$. How μ and σ^2 are estimated from data is considered next.

Mean and Variance

The values of μ and σ^2 are called **parameters**, meaning that they are fixed numerical constants representing some feature or property of a population, in this case the mean and variance, respectively. Although they are constants, their values are unknown, and so they must be estimated from a **sample** chosen to represent the entire population. For the height data, the sample is tabulated in Table 1.1, in which f_i is the number of men whose height is x_i, rounded to the nearest inch. (The fact that the shortest and tallest men are

TABLE 1.1 Heights of 1329 Men

Height interval (i)	Height range (in.)	Nearest inch (x_i)	Number of men (f_i)	$f_i \times x_i$	$f_i \times x_i^2$
1	<63.5	63	23	1,449	91,287
2	63.5–64.5	64	20	1,280	81,920
3	64.5–65.5	65	64	4,160	270,400
4	65.5–66.5	66	110	7,260	479,160
5	66.5–67.5	67	155	10,385	695,795
6	67.5–68.5	68	199	13,532	920,176
7	68.5–69.5	69	203	14,007	966,483
8	69.5–70.5	70	198	13,860	970,200
9	70.5–71.5	71	171	12,141	862,011
10	71.5–72.5	72	88	6,336	456,192
11	72.5–73.5	73	47	3,431	250,463
12	73.5–74.5	74	27	1,998	147,852
13	>74.5	75	24	1,800	135,000
Totals			1,329 (Σf_i)	91,639 ($\Sigma f_i x_i$)	6,326,939 ($\Sigma f_i x_i^2$)

Source: Data from Galton 1889.

grouped in opposite tails of the distribution makes no difference, because these men account for only a small proportion of the total sample.) Also tabulated are the products $f_i \times x_i$ and $f_i \times x_i^2$ as well as their sums.

The mean μ of the distribution is estimated as the mean of the sample, which is conventionally denoted $\bar{x}$ (also sometimes as $\hat{\mu}$):

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \tag{1.2}$$

In this example, $\bar{x} = 91{,}639/1329 = 68.95$ inches.

Likewise, the variance σ^2 of the distribution is estimated as the variance of the sample, which is conventionally denoted s^2 (also sometimes as $\hat{\sigma}^2$):

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - (\bar{x})^2 \tag{1.3}$$

The expression in the middle follows directly from the definition of the variance: It is the average of the squared deviations from the mean, and for any value of x_i, $(x_i - \bar{x})$ is its deviation from the mean. The expression on the right

is identical arithmetically but easier to apply in practice. In the example in Table 1.1, $s^2 = 6{,}326{,}939/1329 - (68.96)^2 = 6.11$. (This value may differ slightly from your own calculation according to the number of significant digits you carried along before rounding off.) If the sample size is small (say, less than 50), then a slightly better estimate of the variance is obtained by multiplying the expression in Equation 1.3 by $n/(n-1)$, where n is the total size of the sample (in this case, 1329).

Closely related to the variance is the **standard deviation** of the distribution, which is the square root of the variance. The standard deviation is a natural quantity to consider in view of the units of measurement. In Table 1.1, for example, each measurement is in inches. The mean is also in inches. However, because the variance is the average of the squared deviations, the variance has the unit of inches-squared, which seems more appropriate for an area than for a height. Taking the square root of the variance restores the correct unit of measure: in this example, inches. The estimate of the standard deviation is conventionally denoted s (also sometimes as $\hat{\sigma}$) and it is calculated as the square root of the quantity in Equation 1.3. In the height example, $s = 2.47$ (which may again differ slightly from your own calculation because of round-off error). The estimate s of the standard deviation is often called the *standard error*.

For a normal distribution, the proportions 68%, 95%, and 99.7% are the proportions of observations expected to fall within 1, 2, or 3 standard errors of the mean, respectively. These emerge directly from Equation 1.1 because the proportion of observations falling with any specified range of x equals the integral of Equation 1.1 across the specified range. For the normal distribution, the integral between the limits $\mu \pm \sigma$ equals 0.6827, that between $\mu \pm 2\sigma$ equals 0.9545, and that between $\mu \pm 3\sigma$ equals 0.9973. In data analysis, $\bar{x}$ and s are used in place of μ and σ. Why do we use two symbols for the mean and two for the standard deviation? Because there is an important difference between $\bar{x}$ and μ and between s and σ. The symbols μ and σ represent the values of the mean and standard deviation in the entire population. The true values of these parameters are unknown and can only be estimated from samples taken from the population. The symbols $\bar{x}$ and s represent the estimates of μ and σ based on a sample, and the different symbols are used to emphasize that the estimates will differ from one sample to the next, and so $\bar{x}$ and s only approximate μ and σ.

Incidentally, the integral of the normal distribution between the limits $\mu \pm 4\sigma$ equals 0.9999; this result says that fewer than one in 10,000 observations falls more than four standard deviations from the mean.

The Central Limit Theorem

Galton was immensely impressed with the observation that many natural phenomena follow the normal distribution. He writes:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "law of frequency of error" [the normal distribution]. Whenever a large sample of chaotic elements is taken in hand and marshaled in the order of their magnitude, this unexpected and most beautiful form of regularity proves to have been latent all along. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

It is, indeed, remarkable to consider that pure, blind chance is the reason for this "unexpected and most beautiful form of regularity." This principle is useful in practice, too. Modern computers can generate uniformly distributed random numbers in a variety of ways. (In a set of uniformly distributed random numbers, one number is as likely to be sampled as any other.) A common way to generate a single sample from a normal distribution is to generate 12 uniformly distributed random numbers on a computer and simply add them up!

The theoretical basis of the normal distribution is known in probability theory as the **central limit theorem**. Roughly speaking, the central limit theorem states that the sum of a large number of independent random quantities always converges to the normal distribution. For our purposes, "independent" in this context means that information about any one of the observations gives no improvement in the ability to predict any other of the observations. A large number of independent random quantities is apparently what Galton meant by "a large sample of chaotic elements." The central limit theorem explains in part why so many continuously distributed traits conform to the normal distribution. Most continuous traits are *multifactorial,* meaning that they are influenced by "many factors," typically several or many genes acting together with environmental factors. Among human beings, for example, the obvious differences between normal people in hair color, eye color, skin color, stature, weight, and other such traits are not usually traceable to single genes. They result from the combined effects of several or many genes as well as numerous environmental effects acting together as "a large sample of chaotic elements," which often produce, in the aggregate, a normal distribution of phenotypes.

It should be emphasized that the "large number" of random elements specified in the central limit theorem need not be excessive. As an example, Figure 1.2 is a bar graph of 100 observations in which each "observation" consists of the sum of nine consecutive random numbers chosen with equal probability from anywhere in the range (–1, +1). For the sum of nine random numbers in this range, the theoretical mean equals 0 and the theoretical standard deviation equals 1.73; the sample values were $\bar{x} = -0.12$ and $s = 1.70$. Expressed as a difference from the mean in multiples of the standard devia-

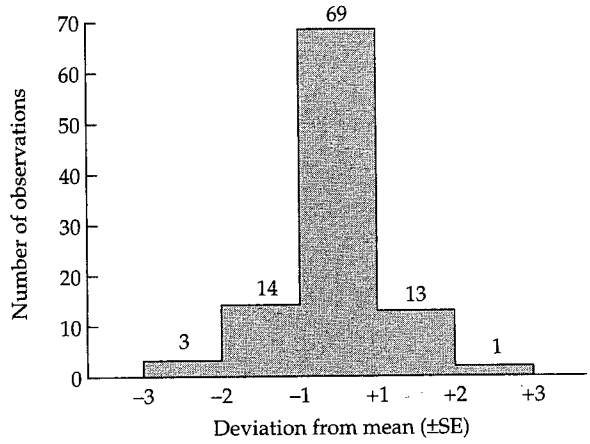


FIGURE 1.2 Distribution of 100 values of the sum of nine random numbers from the interval (–1, +1).

tion, the number of observations in each category is shown at the top of the bar in Figure 1.2. Because the expected numbers are 2.5, 13.5, 68, 13.5, and 2.5, the fit to a normal distribution is obviously very good. In this example, therefore, fewer than 10 "chaotic elements," when added together, yields "this unexpected and most beautiful form of regularity."

PROBLEM 1.1 At an International Health Exhibition in London in 1884, Galton set up an "anthropometric laboratory" that carried out tens of thousands of measurements covering a wide range of human traits. Among the traits was "strength of pull," expressed as the number of pounds that a person could pull with one arm against a resisting force in a sort of arm-wrestling contraption (Galton 1889). The data for 519 males aged 23–26 years fell into the following categories (the number in parentheses is the number of males in each category): 40–50 lbs (10), 50–60 (42), 60–70 (140), 70–80 (168), 80–90 (113), 90–100 (22), 100–110 (24). Using the midpoint of each category as the strength of pull for all males in that category, estimate the mean and standard deviation of strength of pull. Assuming that strength of pull has a normal distribution with parameters equal to these estimates, what is the expected proportion of males whose strength of pull exceeds 112 pounds?

ANSWER The values of x_i are 45, 55, 65, and so forth. Then $\Sigma f_i = 519$, $\Sigma f_i x_i = 38{,}675$, and $\Sigma f_i x_i^2 = 2{,}963{,}375$. Hence, $\bar{x} = 74.5$ lbs, $s^2 = 156.8$ lbs^2, and so $s = 12.5$ lbs. (Answers may differ slightly because of round-off error.) A strength of pull of 112 lbs is three standard errors above the mean; hence a proportion of only $(1 - 0.997)/2 = 0.0015$ (about one in 667) males is expected to have a phenotype exceeding this value.

Discrete Mendelian Variation

Discrete Mendelian variation (also called simple Mendelian variation) refers to phenotypic differences resulting from segregation of the alleles of a single

gene. Environmental effects on the trait are small enough, relative to hereditary differences, that the transmission of alleles determining the trait can be traced through pedigrees. An example of discrete Mendelian variation is the inheritance of red, pink, or white flower color in snapdragon, *Antirrhinum majus* (Figure 1.3). Snapdragons, like human beings, are *diploid* organisms that have two copies of each chromosome, one inherited from each parent. Any gene therefore has a partner on its matching chromosome. Each of the

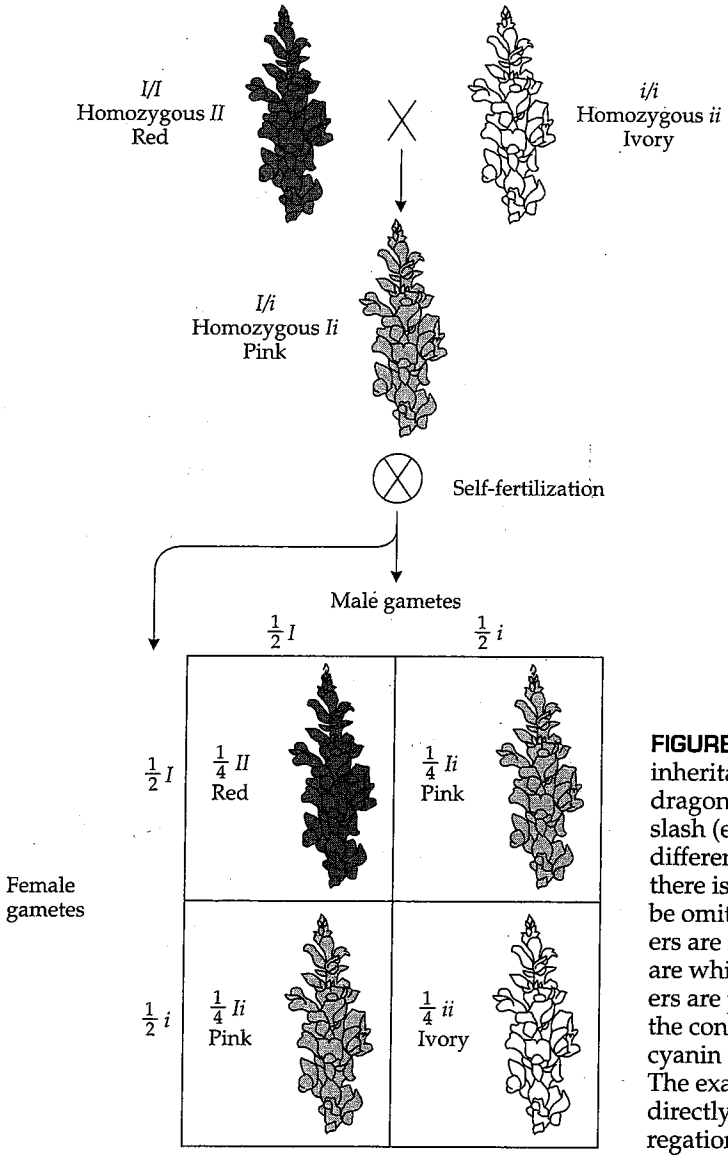


FIGURE 1.3 Simple Mendelian inheritance of flower color in snapdragon (*Antirrhinum majus*). The slash (e.g., *I/I*) separates alleles in different chromosomes, and when there is no ambiguity the slash may be omitted. Homozygygous *II* flowers are red, homozygous *ii* flowers are white, and heterozygous *Ii* flowers are pink. The color results from the concentration of a red anthocyanin pigment in cells of the petals. The example is a classic in showing directly the result of Mendelian segregation in the cross $Ii \times Ii$.

possible forms of a gene encoded in its DNA sequence is called an **allele** of the gene. When the two alleles in an individual are indistinguishable, the genotype of the individual is said to be **homozygous** (*II* or *ii* in Figure 1.3), and when they are different the individual is said to be **heterozygous** (in this example, *Ii*). This example is exceptionally convenient for genetic studies because of the intermediate phenotype of the heterozygote. The result of segregation of the *I* and *i* alleles is outwardly manifested in the 1 : 2 : 1 ratio of plants with red, pink, or ivory-colored flowers.

Natural populations rarely show discrete complex phenotypes that segregate in simple Mendelian fashion of the sort exemplified by flower color in snapdragons. In human populations, for example, although simple Mendelian inheritance does account for many inherited disorders, each of these disorders is individually quite rare. Examples include cystic fibrosis, phenylketonuria, sickle-cell anemia, and hemophilia.

Because most of the variation in phenotype among normal individuals in natural populations is multifactorial, the pattern of inheritance of these traits shows no clear-cut evidence for Mendelian segregation and nothing resembling any of the simple numerical ratios that Mendel originally discovered in his pea-breeding experiments. The absence of such ratios caused a great controversy in the early 1900s immediately after the rediscovery of Mendel's paper. On the one side were the disciples of Galton, called "biometricians," who dismissed the significance of Mendel's discovery, claiming that Mendel's postulated segregating factors were not only irrelevant for continuous traits but also inadequate to explain the observed correlations in traits between relatives. On the other side were the so-called "Mendelians," who argued that segregation of multiple interacting genes could explain continuous traits just as well as discrete traits. The acrimonious dispute between the biometricians and the Mendelians continued for nearly 20 years. The implications of multifactorial inheritance of discrete traits was the focus of a 1918 paper by the statistician Ronald Aylmer Fisher (1890–1962) entitled "The correlation between relatives on the supposition of Mendelian inheritance." The type of model underlying Fisher's analysis is discussed next.

1.3 MULTIPLE-FACTOR INHERITANCE

Fisher examined a mathematical model of multifactorial inheritance and deduced the expected correlations between relatives. He showed that the kinds of data available for continuous traits were not only compatible with Mendelian inheritance but were also predicted by it.

The spirit of Fisher's model is shown in Figure 1.4, which illustrates the genetic variation expected among the progeny of a cross between genotypes that are heterozygous for each of three unlinked genes. (Genes are said to be *unlinked* when they undergo segregation independently of one another, as if they were in different chromosomes.) The alleles of the genes are represented

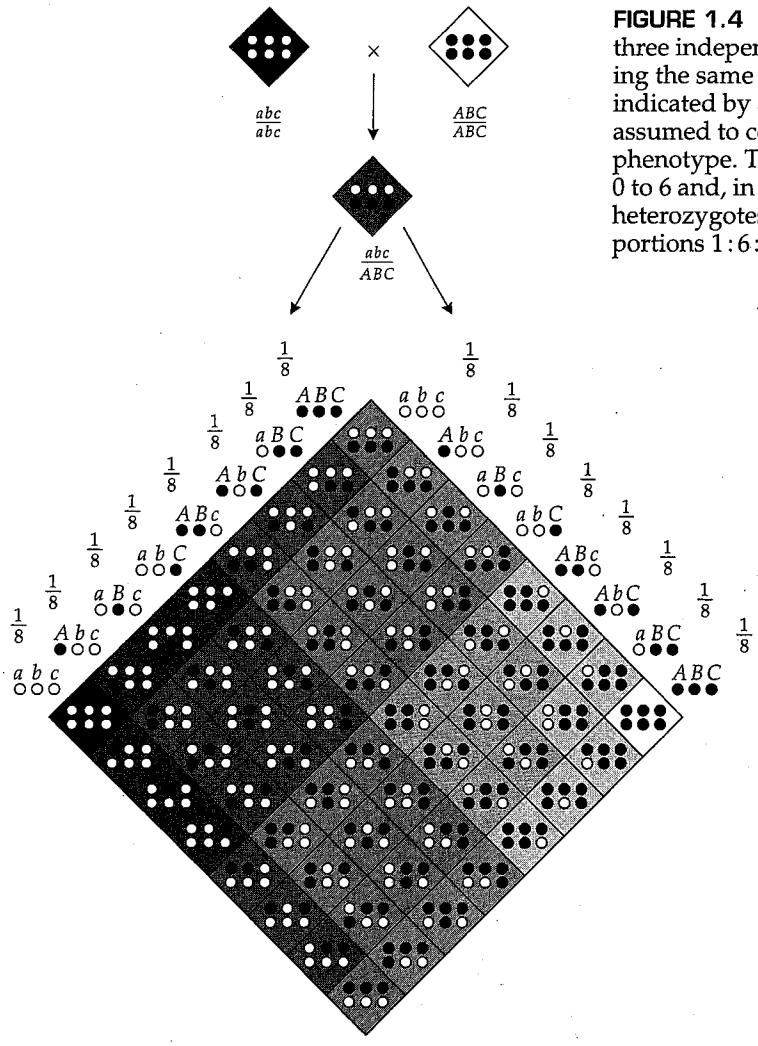


FIGURE 1.4 Result of segregation of three independent pairs of alleles affecting the same trait. Each allele that is indicated by an uppercase letter is assumed to contribute one unit to the phenotype. The phenotypes range from 0 to 6 and, in the cross between triple heterozygotes, are formed in the proportions 1 : 6 : 15 : 20 : 15 : 6 : 1.

A/a, B/b, and *C/c,* and the genetic variation resulting from segregation and independent assortment is evident in the various degrees of shading. If we assume a trait in which each uppercase allele adds one unit to the phenotype and in which each lowercase allele is without effect, then the *aa bb cc* genotype has a phenotype of 0 and the *AA BB CC* genotype has a phenotype of 6. Thus there are seven possible phenotypes (0–6) among the progeny. The distribution of phenotypes is shown in the bar graph in Figure 1.5. The smooth curve is the normal distribution approximating the data, which has a mean of 3 and a variance of 1.5. In Figure 1.4, we have assumed that all of the variation in phenotype results from differences in genotype. If there were

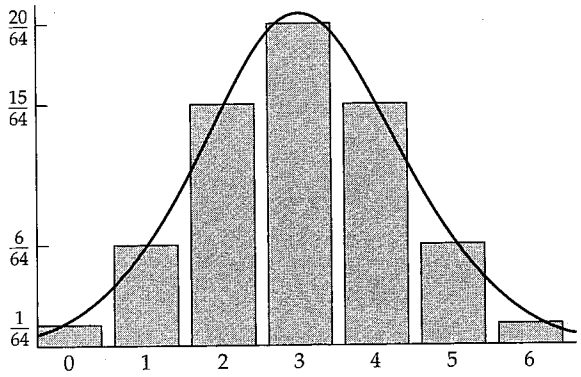


FIGURE 1.5 Distribution of phenotypes from the cross in Figure 1.4 and the approximating normal distribution. The normal curve has mean 3 and variance 1.5.

also random environmental factors affecting the trait, as well as a greater number of genes, then the bars in Figure 1.5 would become less distinct and a normal distribution approximated even better. The result is the central limit theorem at work producing Galton's "supreme law of unreason."

Fisher's model was a good deal more complex than that in Figure 1.4, allowing for differences in the effects of alleles, differences in allele frequency, various types of dominance relations, and the effects of random environmental factors. The work was pathbreaking in demonstrating that continuous variation could be explained by multiple interacting Mendelian factors. Fisher's model was complex for its time and the paper a difficult one. It is not clear even now what practical role Fisher's paper may have played in ending the controversy between the biometricians and the Mendelians. Not many people seem to have read it. (One wag called it a paper you should not read unless you have read it before.) On the other hand, it is the seminal paper that marked the reconciliation of the theories of Galton and Mendel.

1.4 MAINTENANCE OF GENETIC VARIATION

Because Darwin's theory of evolution by means of natural selection requires the presence of genetic variation among individuals, population geneticists have been interested in this issue since the field came into existence in the early 1900s. The main issues were the extent of the differences in genotype from one individual to another, and the processes by which the genetic variation was maintained from one generation to the next. Because the genes underlying multifactorial traits are not revealed by segregation in pedigrees, early population studies were constrained to examine special cases of discrete variation. Classical examples include color or pattern variation within

populations of flowers, insects, or snails; variation in blood groups in humans due to differences in carbohydrate antigens present on the surface of red blood cells recognized by protein antibodies of the immune system; and variation in chromosomes in *Drosophila* due to inversions that could be detected by studying the giant chromosomes present in the larval salivary glands. Each of these examples yielded important insights into evolutionary processes, but all were so different that none could be generalized. Each system also held a possible bias due to the effects of differences in genotype on the relative fitness of the organisms.

Limited as they were, the results were variously interpreted to give support to either of two models for the abundance and maintenance of genetic variation. One view, called the *classical hypothesis*, asserted that genetic variation was uncommon and was composed largely of harmful mutant alleles maintained in the population by a balance between recurrent harmful mutations and negative selection. The other model, called the *balance hypothesis*, supposed that genetic variation was abundant and maintained by selection either favoring heterozygous genotypes or genotypes that were rare. In the classical hypothesis, genetic variation was mostly bad; in the balance hypothesis, mostly good. Each side gave some ground to the other, the classical view conceding some cases of balancing selection, and the balance view conceding the existence of harmful mutations. In the meantime, both hypotheses overlooked another important alternative—that much of the genetic variation in natural populations might have little or no significant effect on the organism's fitness, a model that later became known as the **neutral theory**.

1.5 MOLECULAR POPULATION GENETICS

The classical hypothesis and the balance hypothesis sat across the table glowering at each other through most of the 1950s and 1960s. The issues could not be resolved without an unbiased method for studying genetic variation that could be widely applied to a large number of genes in a variety of organisms. This method finally became possible with the direct study of genes and their products using techniques described in this section, but it came at the price of disconnecting genotype from phenotype. Because the mechanisms of transcription, RNA processing, and translation are relatively free of the gene interactions and environmental effects, the correspondence between DNA sequences and alleles is one-to-one: different alleles have different DNA sequences irrespective of whether the alleles affect phenotype. Likewise, alleles that differ in a protein-coding region may result in different amino acid sequences, irrespective of what the protein does in metabolism or how the difference in sequence affects the organism.

The study of molecules is therefore an efficient way to detect simple Mendelian variation—and therein lies a paradox. As evolutionary biologists, population geneticists are interested in observable phenotypes that are likely

to be subject to natural selection: morphology, rate of development, mating behavior, age of reproduction, longevity, and so forth (in short, the types of traits that attracted Galton). On the other hand, genetic studies are most readily carried out by detecting differences between molecules resulting from simple Mendelian inheritance. The paradox is that differences in molecules among healthy organisms are not usually related in any obvious way to differences in phenotype. Thus, there is a gap in being unable to specify exactly which types of molecular differences underlie the evolutionary process. The irony of the situation is similar to that described by the physiologist Albert Szent-Gyorgyi:

> My own scientific life was a descent from higher to lower dimensions, led by the desire to understand life. I went from animals to cells, from cells to bacteria, from bacteria to molecules, from molecules to electrons. The story had its irony, for molecules and electrons have no life at all. On my way, life ran out between my fingers.

The gap between genotype and phenotype results from the complex interactions between genes and environment in the determination of physiology, development, and behavior. In evolutionary biology, the complexity is even greater because the key issue is the relative ability of organisms to survive and reproduce in their environments. Nevertheless, the disconnect between differences in molecules and evolutionary adaptations is by no means inevitable, permanent, or insurmountable. It is already clear that the study of the relation between genetic variation and evolutionary adaptation must be high on the agenda of evolutionary biology, and already there are many examples in which the relation is quite well established.

Electrophoresis

New and improved methods for studying macromolecules are continually being created, especially for DNA and proteins. Almost as quickly as they appear, alert population geneticists have applied them to studies of genetic variation in natural populations. Although there are many such experimental procedures that differ in a myriad of details, most of the methods are based on novel combinations of a few simple principles.

One of the most widely applied principles for the study of macromolecules is **electrophoresis,** in which macromolecules in solution move in response to an electric field (Smithies 1954, 1995; Shaw 1965; Lewontin and Hubby 1966). Electrophoresis can be used to separate either protein molecules or nucleic acids. The supporting material that holds the macromolecules is usually some sort of gel, which can be in the form of a horizontal slab, a vertical sandwich between two glass plates, or a cylinder contained within the wall of a glass or plastic tube. Opposite sides of the gel are arranged to make contact with a buffered solution and electrodes. Each sam-

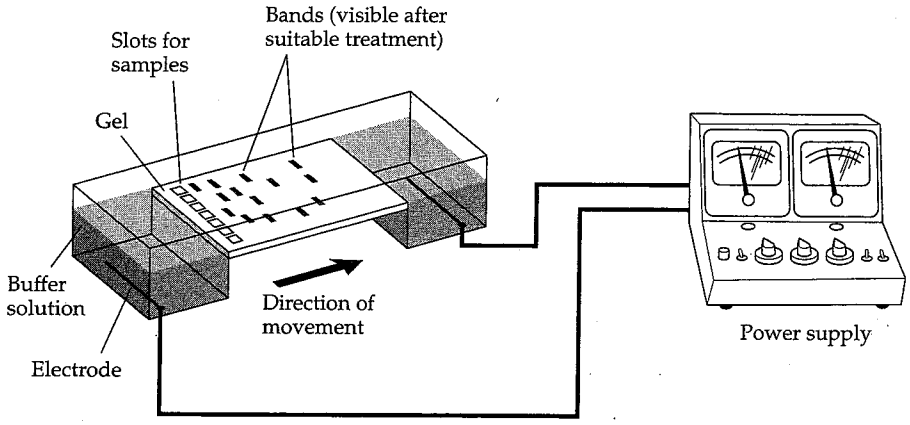


FIGURE 1.6 One type of laboratory apparatus for electrophoresis. The procedure is widely used to separate protein or DNA molecules. In conventional gels, DNA fragments smaller than about 20 kb (1 kb = 1000 nucleotide pairs) migrate approximately in proportion to the logarithm of their molecular weights.

ple of material containing the macromolecules to be separated is placed at or near one end of the slab or tube, and an electric current is applied across the gel for several hours. Molecules in the samples—usually proteins or nucleic acids are of greatest interest—move through the gel in response to the electric field. Molecules of different size and charge move at different rates. Double-stranded DNA molecules move in primarily in relation to their size, whereas protein molecules move primarily in relation to their ionic charge as well as their size. After the electrophoresis is finished, the positions of the molecule or molecules of interest are revealed by any of several procedures.

A typical laboratory setup for protein electrophoresis is diagrammed in Figure 1.6. Protein electrophoresis is used primarily to study enzyme molecules, and the position to which a particular enzyme migrates is revealed by soaking the gel in a solution containing a substrate for the enzyme along with a dye that precipitates where the enzyme-catalyzed reaction takes place. A dark band thus appears in the gel at the position of the enzyme. If the enzyme present in a sample has an amino acid replacement that results in a difference in the overall ionic charge of the molecule, then the enzyme will have a somewhat altered electrophoretic mobility and move at a different rate. The electrophoretic mobility changes because enzymes of the same size and shape move at a rate determined largely by the ratio of the number of positively charged amino acids (primarily lysine, arginine, and histidine) to the number of negatively charged ones (principally aspartic acid and glutam-

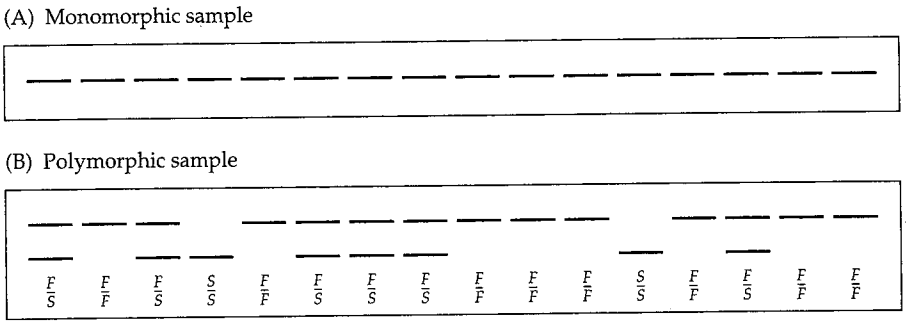


FIGURE 1.7 Monomorphism and polymorphism. (A) Hypothetical gel showing protein monomorphism. All samples have an enzyme with the same electrophoretic mobility. (B) Hypothetical gel showing allozyme polymorphism. Eight samples are homozygous for an allele (F) that codes for a rapidly migrating enzyme; two samples are homozygous for a different allele (S) that codes for a slowly migrating enzyme; and six samples are heterozygous (F/S) and therefore exhibit enzyme bands corresponding to both alleles.

ic acid). Electrophoresis can therefore be used to detect a mutation that results in a difference in electrophoretic mobility of the enzyme it encodes.

One possible result of an electrophoresis experiment is shown in the hypothetical gel in Figure 1.7A, in which all samples manifest an enzyme with the same electrophoretic mobility. The result indicates a *monomorphic sample* because there is only one electrophoretic pattern observed. Another kind of result is shown in Figure 1.7B, in which *polymorphism* is observed in the types of electrophoretic patterns. When polymorphic enzyme bands are observed, genetic tests typically indicate that organisms with only a fast-migrating enzyme are homozygous for a *fast* allele (F/F) and those with only a slow-migrating enzyme are homozygous for a *slow* allele (S/S). Organisms with both enzyme bands are heterozygous for the alleles (F/S). Simple Mendelian inheritance of the polymorphism is indicated by, for example, the finding that matings of two heterozygotes produce, on the average, $\frac{1}{4}$ F/F, $\frac{1}{2}F/S$, and $\frac{1}{4}S/S$ progeny. Two enzyme bands appear in heterozygotes whenever the active enzyme consists of a single polypeptide chain (rather than two or more polypeptide chains aggregated together) because heterozygotes produce a different polypeptide chain from each allele.

Allele Frequencies and Genotype Frequencies

Enzymes that differ in electrophoretic mobility as a result of allelic differences in a single gene are called **allozymes**. Hence, allozyme variation in a population is usually an indication of simple Mendelian genetic variation. As

we shall see later in this chapter, allozyme variation is widespread in almost all natural populations studied by electrophoresis, including organisms such as bacteria, plants, *Drosophila*, mice, and human beings.

To convey some sense of how population genetic data are analyzed, consider a population containing an allozyme polymorphism with F and S alleles at different frequencies. By the **allele frequency** of a specified allele, we mean the proportion of all alleles of the gene that are of the specified type. Suppose we carried out electrophoresis of the enzyme on a sample of 400 members of a population and found 165 F/S, 190 F/S, and 45 S/S. (Here we use the slash to separate the symbol for each allele; if there is no ambiguity, the slash is optional.) In this sample, the observed numbers of F and S alleles are therefore:

纠错：此处基因型个体数量为165F/F

$$F: 2 \times 165 + 190 = 520$$
$$S: 190 + 2 \times 45 = 280$$

The factors of 2 are included for the homozygous genotypes because each FF genotype contains two F alleles and each SS genotype contains two S alleles. The total number of alleles in the sample equals $2 \times 400 = 800$. Therefore, if we let p represent the frequency of the F allele and q represent the frequency of the S allele (with $p + q = 1$ because these are the only alleles of the gene in question), then we can estimate p and q from the observations as:

$$\hat{p} = 520/800 = 0.650$$
$$\hat{q} = 280/800 = 0.350$$

For proportions such as these, the estimated standard errors of the estimated allele frequencies are given by $\sqrt{(\hat{p}\,\hat{q}/n)}$ where n is the number of alleles in the sample. In this case the estimated standard error of $\hat{p}$ (and also of $\hat{q}$) equals $\sqrt{(0.650 \times 0.350\ /\ 800)} = 0.0169$

Note that, if the F and S alleles were combined into genotypes at random (with independence), the expected frequencies of three genotypes can be calculated by multiplication as p^2 FF, $2pq$ FS, and q^2 SS. Therefore, assuming random combination into genotypes, the expected numbers of the three genotypes are:

$$FF: (0.65)^2 \times 400 = 169$$
$$FS: 2 \times 0.65 \times 0.35 \times 400 = 182$$
$$SS: (0.35)^2 \times 400 = 49$$

Hence, the observed numbers in this hypothetical population are very close to those expected with random combinations of alleles. The proportions p^2, $2pq$, and q^2 for the three genotypes when two alleles are combined at random constitute the **Hardy-Weinberg principle**, which is one of the basic principles in population genetics. The Hardy-Weinberg principle is discussed in detail in Chapter 2.

PROBLEM 1.2 Suppose that a random sample of 400 individuals from a different population includes 185 *F/F*, 150 *F/S*, and 65 *S/S* genotypes. Estimate the allele frequency p of *F* and q of *S*. Assuming random combinations of alleles in the genotypes, what are the expected numbers of the three genotypes? Do the observed data seem to fit the expectations?

ANSWER Among the total of 800 alleles, the observed number of *F* alleles is $2 \times 185 + 150 = 520$ and that of *S* alleles is $150 + 2 \times 65 = 280$. Therefore, $\hat{p} = 520/800 = 0.65$ and $\hat{q} = 280/800 = 0.35$. Note that the estimated allele frequencies are the same as in the earlier example, even though the observed numbers of the genotypes are different. With random combinations of alleles in the genotypes, the expected numbers are again 169 *F/F*, 182 *F/S*, and 49 *S/S*. Compared with the observed numbers, there appear to be too many homozygous genotypes and too few heterozygous genotypes. A statistical method for deciding whether or not the fit is satisfactory is discussed in Chapter 2.

Polymorphism and Heterozygosity

Polymorphism of a gene in a sample is usually of interest only insofar as it indicates polymorphism of the gene in the population as a whole. In a population, a **polymorphic gene** is one for which the most common allele has a frequency of less than 0.95. In practical terms, for a gene with two alleles, this definition implies that a random sample as small as 11 individuals from a population with Hardy-Weinberg genotype frequencies (p^2, $2pq$, and q^2) would be expected to include at least one heterozygous genotype, because $2 \times 0.05 \times 0.95 \times 11 = 1$. Observed frequencies of heterozygous genotypes for genes encoding enzymes vary widely but are typically around 15% in vertebrates and 7% in vertebrates (Figure 1.8). The cutoff at 0.95 is arbitrary, but it serves to focus attention on those genes in which allelic variation is common. In any large population, rare alleles are observed for virtually every gene. An allele is considered a *rare allele* if its frequency is less than 0.005; in human beings, between one and two people per thousand are heterozygous for rare alleles of any gene. Many rare alleles are deleterious and are presumably maintained in the population by recurrent mutation. The definition of polymorphism is an attempt to focus on genes that have alleles with frequencies too high to be explained solely by recurrent mutation to harmful alleles. With the 0.95 definition of polymorphism given above, and if alleles are combined at random into genotypes, then roughly 10% of a population with Hardy-Weinberg frequencies is heterozygous for the most common allele, because $2 \times 0.95 \times 0.05 = 0.095 \approx 10\%$.

更正：此处应该是无脊椎动物15%，脊椎动物7%

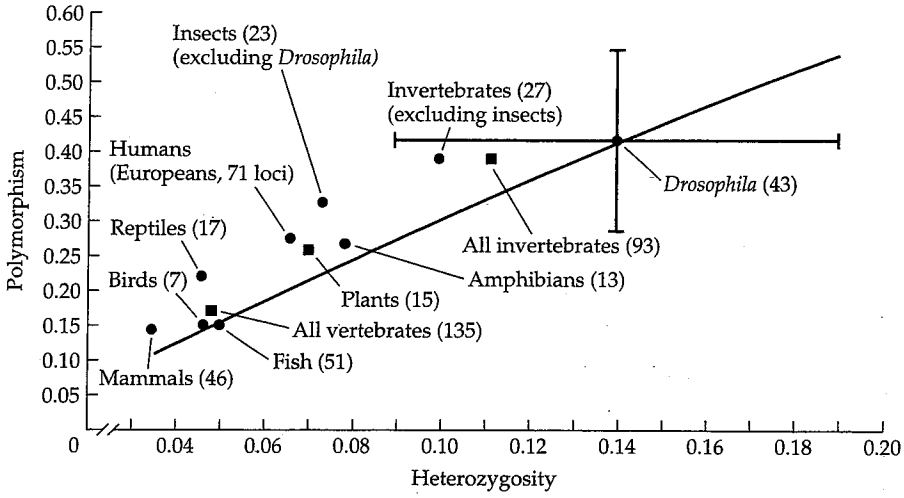


FIGURE 1.8 Estimated levels of heterozygosity and proportion of polymorphic genes derived from allozyme studies of various groups of plants and animals. The curve denotes the theoretically expected relation under the dubious assumption that all allozyme variation is selectively neutral. The number of species studied is shown in parenthesis beside each point. Squares denote averages for plants, invertebrates, and vertebrates. The bars across the *Drosophila* point indicate the standard error within which about 68% of the species are expected to fall. Other groups have similarly large standard errors. (Data from Nevo 1978.)

Allozyme Polymorphisms

Figure 1.8 summarizes the results of electrophoretic surveys of 14 to 71 (mostly around 20) genes in populations of 243 species. Each point in the figure gives the type of organism studied and the number of species examined. The axis labeled Polymorphism refers to the estimated proportion of genes that are polymorphic by the 0.95 criterion. The axis labeled Heterozygosity refers to the average heterozygosity in each group. The average heterozygosity is the estimated proportion of genes expected to be heterozygous in an average organism; it is estimated as the proportion of heterozygous genotypes for each gene averaged over all genes. For example, the data for Europeans include an English population in which 10 enzyme genes were examined (Harris 1966). Of the 10 genes, three were found to be polymorphic, from which the estimated proportion of polymorphic genes in the genome is 3/10 = 0.30. The observed proportion of heterozygous genotypes for each of the three polymorphic genes was 0.509 (for red-cell acid phos-

phatase), 0.385 (for phosphoglucomutase), and 0.095 (for adenylate kinase); the average heterozygosity in this sample—taking into account the additional seven genes for which the observed heterozygosity was 0—is therefore $(0.509 + 0.385 + 0.095 + 7 \times 0)/10 = 0.099$. A more extensive electrophoretic survey of 104 genes in a sample including all major human races gave estimates of polymorphism of 0.32 and heterozygosity of 0.06 (Harris et al. 1977).

The vertical and horizontal bars on the point corresponding to *Drosophila* indicate the size of the standard error of the estimate. The bars indicate the limits of polymorphism and heterozygosity within which about 68% of the species are expected to fall. Among *Drosophila* species, approximately 68% have a proportion of polymorphic genes in the range 0.30–0.56 and an average heterozygosity in the range 0.09–0.19. If such bars were attached to each point, their lengths would be comparable to those for *Drosophila*, indicating substantial variability in polymorphism and heterozygosity among species within groups.

Figure 1.8 indicates a positive relationship between amount of polymorphism and degree of heterozygosity. This relationship is as expected because the greater the fraction of polymorphic genes in a population, the more genes that are expected to be heterozygous on the average. Consider an idealized population in which each new mutation encodes a protein whose electrophoretic mobility is distinct from all others present in the population and in which each new mutant allele is selectively neutral (that is, it has a negligible effects on survival and reproduction). Because of recurrent mutation, the alleles in a population gradually turn over in time as some are lost while others become polymorphic. Under these conditions, and restricting our attention to autosomal genes in a diploid species, the expected proportion of polymorphic loci P is given by

$$\ln[1-P] = \theta \ln(0.05) \approx -3\theta \qquad (1.4)$$

(Kimura and Ohta 1971) where $\theta = 4N\mu$ is the product of the population size (N) and the mutation rate (μ) per gene per generation. (The symbol ln refers to the logarithm in base e.) The value 0.05 emerges from the definition of polymorphism in which the frequency of most common allele is less than 0.95, because $0.05 = 1 - 0.95$.

Under the same assumptions, the expected magnitude of the heterozygosity (H) can be shown to equal

$$H = \frac{\theta}{1+\theta} \qquad (1.5)$$

(Kimura and Crow 1964). Consequently, for genes in an ideal population undergoing successive neutral mutations, the expected relationship between

heterozygosity and polymorphisms can be obtained by eliminating θ between Equations 1.4 and 1.5, with the result that

$$\ln[1-P] = \frac{-3H}{1-H} \tag{1.6}$$

This is the relationship shown by the curve in Figure 1.8.

The overall mean polymorphism in Figure 1.8 is 0.26 ± 0.15, and the mean heterozygosity is 0.07 ± 0.05. Vertebrates have the lowest average amount of genetic variation among the groups in Figure 1.8, plants come next, and invertebrates have the highest. *Drosophila* is the most genetically variable group of higher organisms so far studied, and mammals the least variable. Human beings are fairly typical of large mammals. The one obvious conclusion that can be reached from Figure 1.8 is that allozyme polymorphisms are widespread among higher organisms. Genetic variation is even more prevalent among some prokaryotes. For example, natural isolates of the mammalian intestinal bacteria *Escherichia coli* exhibit levels of genetic polymorphism two or three times greater than vertebrates (Selander et al. 1987).

Although genetic polymorphisms are widespread, they are not universal. For example, both major subspecies of the cheetah *Acinonynx jubatus* are virtually monomorphic (O'Brien et al. 1987). A survey of 49 enzymes among 30 animals from the East African subspecies (*A. j. raineyi*) yielded only two polymorphic genes and estimates of polymorphism of 0.04 and heterozygosity of 0.01; among 98 animals from the South African species (*A. j. jubatus*), the estimate of polymorphism was 0.02 and that of heterozygosity 0.0004. Most unusual was the finding of skin-graft acceptance between unrelated cheetahs from the South African subspecies. Graft acceptance means that the cheetah population is monomorphic for the major histocompatibility locus that triggers graft rejection, which is abundantly polymorphic in other mammals. Apparently, the cheetah, which was worldwide in its range at one time but presently numbers less than 20,000 animals, underwent at least two severe constrictions in population number resulting in the loss of most of its genetic variability.

Inferences from Allozyme Polymorphisms

The generality of estimates of polymorphism based on electrophoresis is somewhat uncertain (Lewontin 1974b, 1991). The amount of polymorphism may be underestimated because conventional electrophoresis fails to detect many amino acid replacements. For example, in a study of 14 myoglobin proteins from various species including cetaceans (whales, dolphins and porpoises), no more than eight could be distinguished by conventional electrophoresis; however, 13 could be distinguished by varying the pH value of the electrophoresis buffer (McLellan and Inouye 1986). Some amino acid

replacements can be detected because they render the enzyme sensitive to high temperatures; a test for temperature sensitivity increased the number of identified alleles of the gene coding for xanthine dehydrogenase in *Drosophila pseudoobscura* from 6 to 37 and increased the estimate of average heterozygosity from 0.44 to 0.73 (Singh et al. 1976). On the other hand, although more elaborate techniques reveal additional alleles of genes known to be polymorphic, thus increasing estimates of heterozygosity, genes classified as monomorphic by means of routine electrophoresis tend to remain monomorphic, and so estimates of polymorphism remain much the same as before.

Electrophoretic surveys might also overestimate the amount of polymorphism because the enzymes typically surveyed are those found in relatively high concentration in tissues or body fluids ("Group I enzymes") and often lack the high substrate specificity of enzymes implicated in central metabolic processes ("Group II enzymes"). For example, among 10 Group I and 11 Group II enzymes in *Drosophila*, estimates of polymorphism and heterozygosity were 0.70 and 0.24 in the former and 0.27 and 0.04 in the latter (Gillespie and Langley 1974). In summary, protein electrophoresis is a convenient method for detecting polymorphisms, but it is difficult to extrapolate from electrophoretic surveys of enzymes to the entire genome because the enzymes may not be representative.

The high levels of polymorphism observed for allozymes immediately cast the classical hypothesis into doubt. This hypothesis asserted that genetic variation consists largely of highly deleterious alleles maintained by recurrent mutation. The classical hypothesis predicts that allozyme polymorphisms should be rare, whereas Figure 1.8 indicates that they are common. But beyond mere estimates of the magnitude of polymorphism and heterozygosity, other data also throw the alternative balance hypothesis into doubt. The balance hypothesis predicts that genetic variation should be common because it is maintained either by selection favoring heterozygous genotypes or selection favoring rare genotypes. This type of selection predicts very strong harmful effects of inbreeding (mating between close relatives), but the inbreeding effects actually observed are relatively mild.

The seemingly good fit between the data in Figure 1.8 and the theoretical curve for neutrality from Equation 1.6 might be taken as support for the neutral theory, but the data conceal a multitude of complications. Some individual genes show too much heterozygosity for their level of polymorphism, and other genes show too little heterozygosity. More refined analyses of the DNA sequences of alleles of individual genes using statistical methods discussed in Chapters 4 and 7 show that Figure 1.8 presents a picture painted with too broad a brush. Among the many loci represented are some for which most polymorphic alleles appear to be slightly deleterious, others for which the polymorphisms appear to be maintained by some form of selection, and still others showing no marked departure from the patterns expected were the polymorphic alleles selectively neutral or nearly neutral.

1.6 POLYMORPHISMS IN DNA SEQUENCES

Electrophoresis is also one of the mainstays for the study of genetic variation in DNA sequences, because DNA molecules are negatively charged and therefore will move in an electric field. Although standard DNA purification procedures typically shear double-stranded DNA molecules at random positions into fragments of about 50 kb (1 kb = 1000 nucleotide pairs), a number of methods can be used to produce fragments of a specific length.

Restriction Enzymes

DNA fragments of a specific size can be produced by means of any of a class of enzymes called **restriction enzymes**, which cleave double-stranded DNA wherever there is a particular, short nucleotide sequence called the enzyme's *restriction site*. Because the cleavage sites are highly specific, the size of any DNA fragment produced is determined by the distance between adjacent restriction sites. Examples of restriction enzymes and their restriction sites are shown in Figure 1.9, where the cuts are made at the positions of the

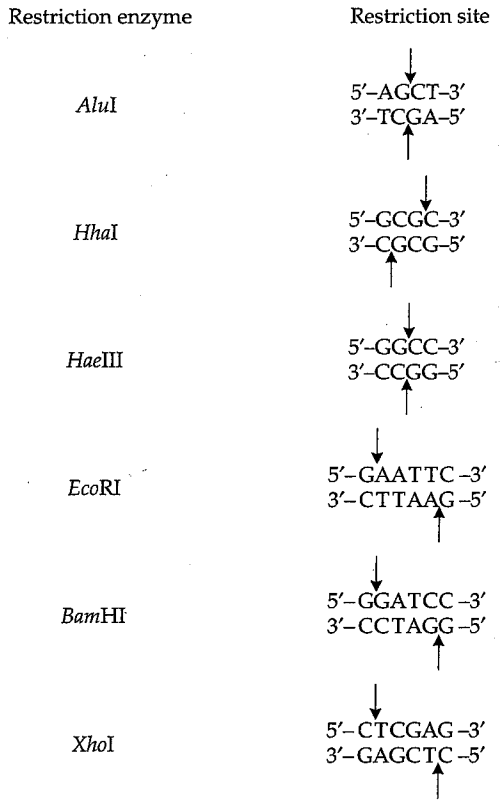


FIGURE 1.9 Restriction enzymes cleave DNA molecules at sites of specific, short nucleotide sequences. More than 500 different restriction enzymes are commercially available. They are essential tools in DNA analysis and gene cloning. The cleavage site in each DNA strand is indicated by the arrow.

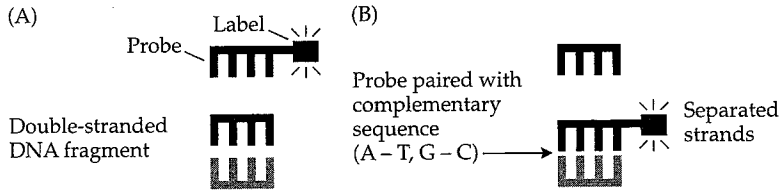


FIGURE 1.10 Nucleic acid probes are based on the principle that individual strands with complementary nucleotide sequences of sufficient length can form stable double-stranded molecules. (A) A probe that has exactly the same nucleotide sequence (region in black) as one strand of a double-stranded DNA molecule. (B) If the DNA strands are separated and then allowed to come together again, then in the presence of an excess of probe molecules, the complementary strand will undergo hybridization with the probe rather than with its original partner.

arrows. For example, the enzyme *Alu*I cuts at sites of the four-nucleotide sequence 5′–AGCT–3′, and *Eco*RI cuts at the six-nucleotide sequence 5′–GAATTC–3′. The nucleotide sequence of only one DNA strand need be specified, because in double-stranded DNA, the nucleotide A pairs with T and the nucleotide G pairs with C. The symbols 5′ and 3′ are used to denote the polarity (left-to-right directionality) of the strands. In double-stranded DNA, each strand has a polarity opposite the other, hence the sequence 5′–GAATTC–3′ is paired with the sequence 3′–CTTAAG–5′. As illustrated in Figure 1.9, most restriction enzymes used in population studies have restriction sites consisting of either four nucleotides or six nucleotides.

Because of the specific cleavage sites, digestion of genomic DNA with a restriction enzyme yields a set of fragments of different sizes according to the distances between adjacent restriction sites. These fragments are separated by size by means of electrophoresis, and then any fragment of interest is identified as illustrated in Figure 1.10. Because complementary nucleotide strands can pair with one another, a stretch of single-stranded DNA is able to pair with a complementary region of a strand in a double-stranded molecule, provided that the strands of the double-stranded molecule are first separated either chemically or by heat. The small stretch of single stranded DNA is usually called a **probe**. A probe can range in size from 24 nucleotides to many thousands of nucleotides, and it is usually labeled in some fashion to be able to emit fluorescent or visible light or to undergo radioactive decay. The label can be attached to one end of the molecule as shown in Figure 1.10A, or it can be incorporated into individual nucleotides along the probe. A probe as short as that shown in Figure 1.10 would not work in practice, because pairing of such short regions is easily disrupted by thermal motion. The diagram will nevertheless suffice to make the point that a DNA (or RNA) probe of suffi-

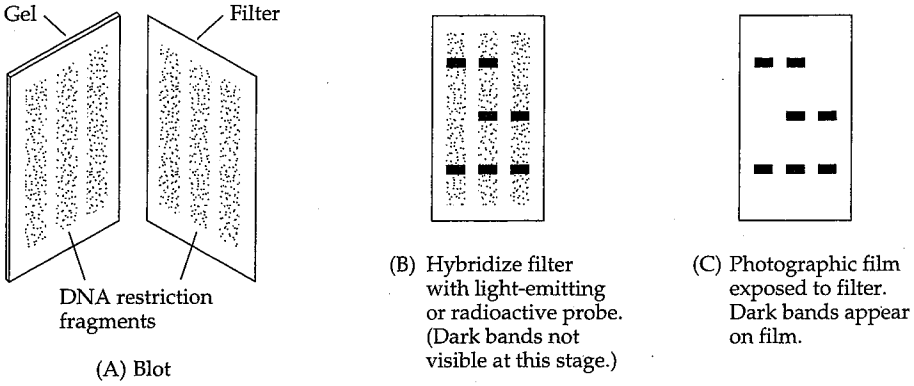


FIGURE 1.11 Southern blot procedure. (A) DNA fragments separated by electrophoresis are transferred and chemically attached to a filter. (B) The filter is mixed with labeled probe DNA, which hybridizes with and sticks to homologous DNA molecules in the filter. (C) After washing, the filter is exposed to photographic film, which develops dark bands caused by radioactive or light emission from the probe.

cient length will hybridize with a complementary (or nearly complementary) sequence in any strand of DNA (or RNA). As noted by the shading in Figure 1.10B, a probe will generally pair only with one strand along a stretch of double-stranded DNA, because the base sequence in the same region of the other strand is identical and therefore not able to pair with the sequence in the probe.

Hybridization of a restriction fragment with a probe is the principle underlying the **Southern blot** procedure illustrated in Figure 1.11. DNA restriction fragments that have been separated by electrophoresis are rendered single-stranded by soaking in a solution of sodium hydroxide, then blotted onto a nitrocellulose or nylon filter where subsequent chemical treatment attaches them (Figure 1.11A). The filter is then bathed in a solution containing labeled probe DNA (part B). As the solution cools, the probe DNA strands form double-stranded molecules with their complementary counterparts on the filter, and careful washing removes all of the probe DNA that has remained unpaired. The filter is sandwiched with photographic film, where light emission or radioactive disintegrations from the bound probe result in visible bands (part C).

Genetic differences resulting in the presence or absence of restriction sites can be identified because they change the length of characteristic restriction fragments. An example is illustrated in Figure 1.12. The upper part of each panel shows the location of restriction sites in the DNA molecules in a diploid genotype. The *a*-type molecule contains one additional restriction site

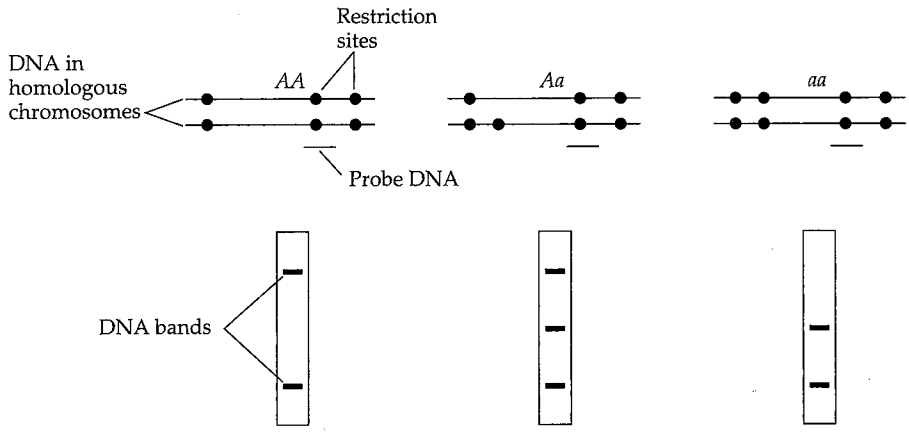


FIGURE 1.12 Restriction fragment length polymorphisms (RFLPs) result from the presence or absence of particular restriction sites in DNA. In this example, the DNA molecule designated A contains three restriction sites, and the one designated a contains four. Genotypes *AA*, *Aa*, and *aa* each yield a different pattern of bands in a Southern blot using the indicated probe DNA.

not present in the *A*-type molecule. The lower part of the figure demonstrates that, with suitable probe DNA, all three genotypes can be distinguished by their pattern of restriction fragments. A difference in the length of a restriction fragment found segregating in natural populations is called a **restriction fragment length polymorphism** or **RFLP**. Because RFLPs are abundant in the genomes of most organisms, they have been widely used to study genetic variation in DNA sequences in natural populations.

The Polymerase Chain Reaction

The **polymerase chain reaction (PCR)** results in the amplification of specific DNA fragments. PCR is of great utility in population genetics, either for the production of probe DNA or for the direct determination of the amount of nucleotide sequence variation present in natural populations. The method is outlined in Figure 1.13. The original DNA sequence to be amplified is shown in black and the newly synthesized DNA strands in gray. The small circles represent short, chemically synthesized sequences of single-stranded DNA (*oligonucleotides*) that are complementary in sequence to the ends of the region to be amplified. The oligonucleotides are called *primer sequences* because they pair with complementary strands at the ends of the sequence to be amplified and are used as primers for chain elongation by DNA polymerase. Primer oligonucleotides are typically 20–30 nucleotides in length. DNA to be used as the template in a PCR reaction is first mixed with both

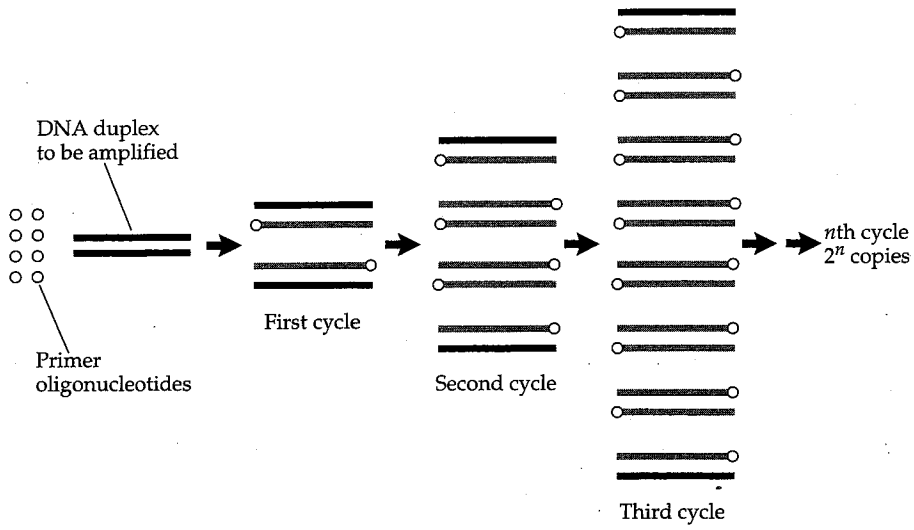


FIGURE 1.13 The polymerase chain reaction (PCR). Short primer oligonucleotides are used as primers to initiate DNA replication from opposite ends of a DNA duplex to be amplified. After each round of replication, the DNA is heated to separate the strands and then cooled to allow new primers to anneal. Repeated rounds of replication result in an exponential increase in the number of target molecules.

primers along with a thermostable DNA polymerase in a buffer solution. The PCR amplification takes place in cycles. In the first cycle, the DNA is heated to separate the strands and then cooled in the presence of a vast excess of the primer oligonucleotides. Then elongation of the primers produces double-stranded molecules. The second cycle of PCR is similar to the first but, after the second cycle, there are four copies of each original molecule. The cycle is repeated from 20 to 30 times, each resulting in a doubling of the number of molecules. The theoretical result of n rounds of amplification is 2^n copies of each template molecule originally present. In practice the reaction does not proceed with perfect efficiency, and the efficiency varies because of the hybridization kinetics of the primers and any tendency for the template to form complex folded structures.

PCR amplification is very useful in generating large quantities of a specific DNA sequence. The main limitation of the technique is that the DNA sequences at the ends of the region to be amplified must be known so that primer oligonucleotides can be synthesized. There are many applications in which this requirement is met. In population genetics, for example, PCR can be used to amplify different alleles present in natural populations.

PROBLEM 1.3 PCR was used to amplify five alleles (designated *a–e*) of the gene *Rh3* coding for a light-sensitive protein in the eye of *Drosophila simulans*, a species of fruit fly closely related to *D. melanogaster*. The resulting DNA fragments were sequenced (Ayala et al. 1993). The data show the nucleotide present at each of 16 polymorphic nucleotide sites found in the first 500 nucleotide sites in the amino acid coding region of the gene; the remaining 484 nucleotide sites were monomorphic in this sample. Any nucleotide site that is an exact multiple of three is at the third position of a codon. In this region of the gene:

(*a*) what proportion of polymorphic nucleotide sites are in third positions of codons? What can you infer from this observation?

(*b*) what proportion of nucleotide sites are polymorphic?

(*c*) why is the binomial standard error $\sqrt{(\hat{p}\,\hat{q}/n)}$ not appropriate for the estimate in part (*b*)?

Nucleotide site in gene

Allele	132	142	162	192	198	201	207	240	246	351	354	372	375	405	417	483
a	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
b	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
c	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
d	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
e	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A

ANSWER (*a*) Among the 16 polymorphic sites, only site 142 is not an exact multiple of three, hence $\frac{15}{16}$ = 94% of the polymorphic sites are in the third codon position. The inference is that many of the nucleotide polymorphisms are silent (synonymous) in that they do not alter the amino acid sequence of the polypeptide. (In fact, all 16 are silent polymorphisms, including the C → T change in 142, which alters the codon from CUA → UUA, both of which code for leucine.) (*b*) A total of $\frac{16}{500}$ = 3.2% of the nucleotide sites are polymorphic in this region of the gene. (*c*) The binomial standard error is not appropriate in this case because the nucleotides within a gene are not independent samples; they are genetically closely linked.

With PCR, the need to know the sequences at the ends of the fragment to be amplified (in order to be able to synthesize the complementary oligonucleotide primers) may seem like a serious limitation, but even this difficulty can be circumvented in a clever method for studying polymorphism known as **amplified fragment length polymorphism (AFLP)**. The method is outlined in Figure 1.14. The first step (part A) is to digest genomic DNA with a restriction enzyme; this example uses the enzyme *Eco*RI, whose restriction site is 5′-GAATTC-3′, and the stars represent the positions of two adjacent restriction sites. (The size of the DNA fragment between the sites, relative to the size of

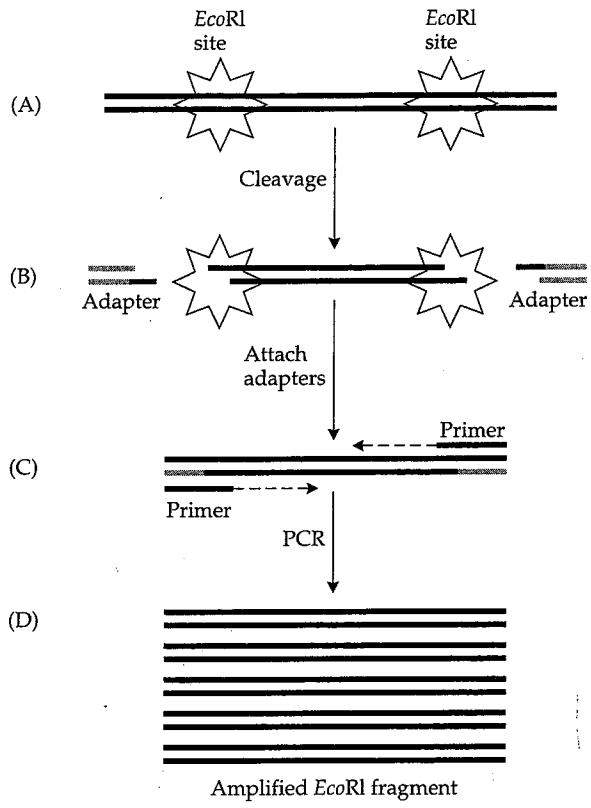


FIGURE 1.14 Amplified fragment length polymorphisms (AFLPs) make use of primer adapters to amplify restriction fragments produced with a particular restriction enzyme, in this case *Eco*RI. (A) Part of the DNA molecule in a chromosome showing the positions of two *Eco*RI restriction sites. (B) After digestion with the restriction enzyme, the fragment is mixed with double-stranded adapters that have single-stranded overhangs complementary to the single-stranded overhangs produced by the restriction enzyme, and then the hybridized adapters are ligated onto the restriction fragment using an enzyme. (C) Primers that are complementary to the adapter sequences are then used to amplify the restriction fragment by means of the polymerase chain reaction. (D) Many copies of the restriction fragment are produced. Normally, the DNA from a single individual produces many different amplified fragments, and any fragment that can be amplified from some individual but not others is an AFLP.

the sites themselves, is not shown to scale.) Digestion yields a large number of restriction fragments flanked by what remains of an *Eco*RI site on each side. *Eco*RI digestion yields a 5′ overhang at each end, with the sequence 5′-AATT (Figure 1.9). These overhangs are long enough that, at low temperature, they can pair with complementary 3′ overhangs of special primer *adapters* (Figure 1.14B), which are attached to the restriction fragments using the enzyme DNA ligase. The resulting fragments (C) are ready for amplification by means of PCR using oligonucleotides that are complementary to the adapters. Note that the same adapter is ligated onto each end, and so a single primer sequence will anneal to both ends and support amplification. There are nevertheless a number of choices concerning the primer sequence. A primer that matches the adapters perfectly will amplify all fragments, but this often results in so many amplified fragments that they are not well separated in the gel. Since a PCR primer must match perfectly at its 3′ end to be elongated, additional nucleotides added to the 3′ end reduce the number of amplified fragments. These primers will amplify only those fragments that, by chance, have a complementary nucleotide immediately adjacent to the *Eco*RI site.

Single Nucleotide Polymorphisms

The ultimate level for the study of genetic polymorphisms is that of the DNA sequence itself, and the smallest unit of polymorphism is the **single nucleotide polymorphism** or **SNP**. A SNP is said to be present at a particular nucleotide site if the DNA molecules in the population frequently differ in the identity of the nucleotide pair that occupies the site. Consider, for example, a single nucleotide site in the protein-coding strand of DNA. Some DNA molecules in a population may have a T (thymidine) nucleotide at this site, whereas other DNA molecules in the same population may have a C (cytosine) nucleotide at the same site. This difference constitutes a SNP. The SNP defines two "alleles" for which there could be three genotypes among individuals in the population, namely, homozygous with T at the corresponding site in both homologous chromosomes, homozygous with C at the corresponding site in both homologous chromosomes, or heterozygous with T in one chromosome and C in the homologous chromosome. The word "allele" is in quotes because any paricular SNP need not be in a coding sequence, or even in a gene. The human genome is thought to contain at least 10 million SNPs, or about one in every 300 base pairs. More than 4 million SNPs have been identified, at which the alternative alleles are both relatively common, and the density among these is about one SNP site every 1000–3000 bp in protein-coding DNA, and about one SNP site every 500–1000 bp in noncoding DNA. How SNPs are used in studies of human population genetics is discussed in Chapter 10.

Synonymous and Nonsynonymous Polymorphisms

One inevitable limitation of protein electrophoresis is that it can detect only those nucleotide polymorphisms in a coding sequence that result in one amino acid being replaced with another in the protein. (In fact, it can detect only the subset of these that alter the protein's charge under the conditions of electrophoresis.) Nucleotide polymorphisms that result in amino acid replacements are known as **nonsynonymous polymorphisms**. There is also a large class of **synonymous polymorphisms** that are present in coding regions but that do not result in an amino acid replacement. These are quite common because the genetic code allows many synonymous nucleotide substitutions. For example, the nucleotide triplet codons TTA, TTG, CTT, CTC, CTA, and CTG all code for the amino acid leucine, and the codons CGU, CGC, CGA, CGG, AGA, and AGG all code for the amino acid arginine. (Only two of the 20 amino acids in the genetic code have no synonymous codons.) Many polymorphisms are also present in noncoding regions of the genome, such as in the region upstream of a coding sequence, in the region downstream of a coding sequence, or in introns that interrupt a coding sequence. Synonymous and noncoding polymorphisms may have subtle effects on the organism, and the polymorphisms may therefore be affected by natural

selection; the polymorphic alleles are synonymous or noncoding only in the sense that they all code for the same amino acid sequence.

An example of extensive synonymous polymorphism in *Drosophila* is illustrated in Figure 1.15 for alleles of the gene coding for alcohol dehydrogenase. This gene has an electrophoretic polymorphism that is widespread in natural populations with two predominant alleles, slow (*Adh-S*) and fast (*Adh-F*). The molecular difference is that the codon for amino acid number 193 in *Adh-S* is AAG (lysine) whereas in *Adh-F* it is ACG (threonine). The enzymes differ not only in electrophoretic mobility. The product of the fast

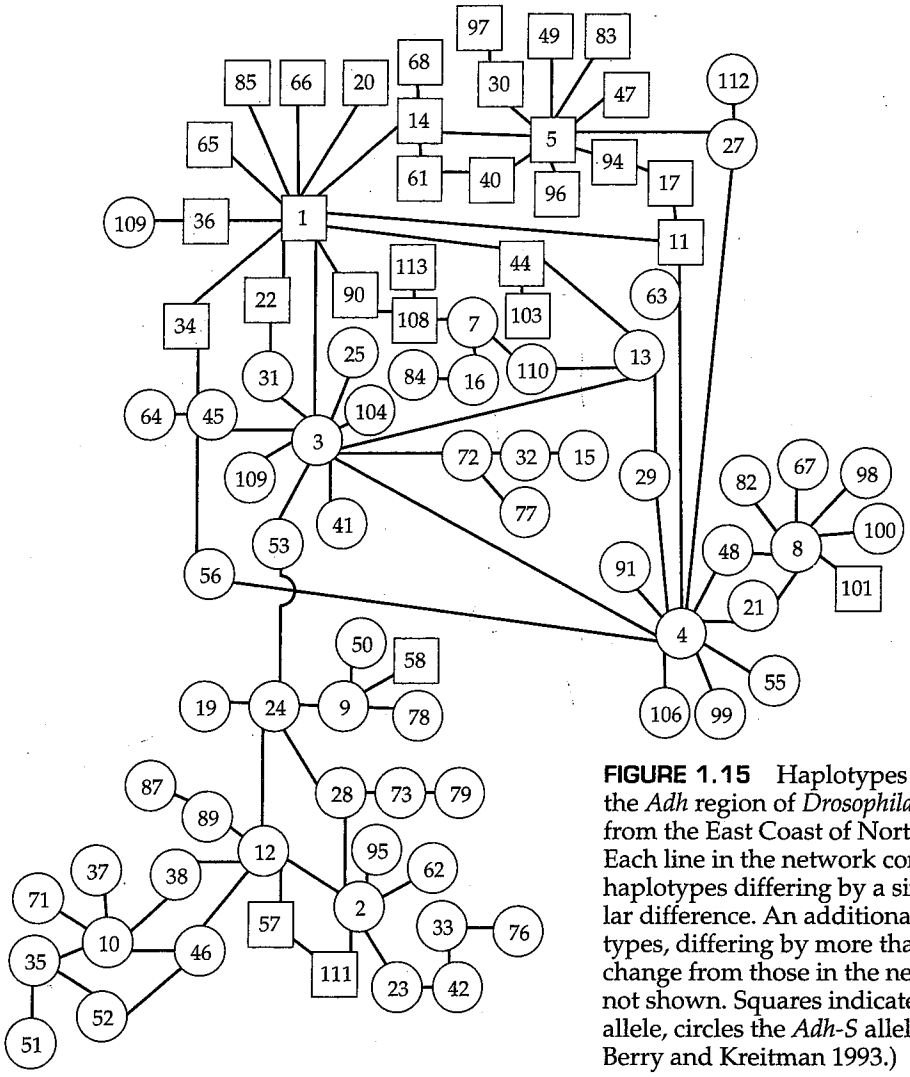


FIGURE 1.15 Haplotypes of alleles in the *Adh* region of *Drosophila melanogaster* from the East Coast of North America. Each line in the network connects two haplotypes differing by a single molecular difference. An additional 20 haplotypes, differing by more than one change from those in the network, are not shown. Squares indicate the *Adh-F* allele, circles the *Adh-S* allele. (From Berry and Kreitman 1993.)

allele has a greater enzymatic activity and is also synthesized in greater amount than that of the slow allele.

The data in Figure 1.15 are derived from studies of the *Adh* region of 1533 flies isolated from 25 populations throughout eastern North America (Berry and Kreitman 1993). A total of 113 haplotypes were identified. A **haplotype** is a unique combination of allelic states of genetic markers present along a single chromosome. The less genetic recombination there is across a region of the genome, or the stronger the selection, the more strongly differentiated the haplotypes are likely to be. The extreme examples are mitochondrial DNA, chloroplast DNA, and the Y chromosome, in which no recombination normally takes place. (See Chapter 4 for a discussion of mitochondrial DNA and chloroplast DNA evolution.)

In Figure 1.15, the haplotypes indicated with squares are *Adh-F* and those with circles are *Adh-S*. The number inside each symbol is the relative abundance of the haplotype (1 being the most frequent, 2 the next most frequent, and so forth). A straight line connecting two haplotypes indicates that they differ by a single nucleotide change. Figure 1.15 includes 93 haplotypes related to at least one other by a singe change; the other 20 haplotypes observed in the study include additional changes. The main point of the *Adh* example is that natural populations contain a great abundance of different types of nucleotide-sequence variation that does not affect amino acid sequence.

Segregating Sites and Nucleotide Mismatches

DNA sequence data contain more information about genetic variation than does protein electrophoresis, because nucleotide polymorphisms are detected even if they are synonymous polymorphisms or are in noncoding DNA, and also because each nucleotide in a set of aligned sequences can be considered individually. (Sequences are said to be *aligned* if corresponding subunits in each sequence, in this case nucleotides, derive from the corresponding subunit in an ancestral sequence.)

By analogy with estimates from protein polymorphism, genetic variation in DNA sequences is conveniently assessed according to the number of nucleotide sites that are polymorphic in a sample of sequences, as well as according to the number of nucleotide sites that are heterozygous. In a set of aligned sequences, the number of *segregating sites* (nucleotide sites that are polymorphic in the sample) is symbolized S and defined as

$$S = \text{Number of nucleotide sites that differ among the aligned sequences} \quad (1.7)$$

The analog of heterozygosity for DNA sequences is the number of nucleotides that differ along any pair of aligned sequences. These are **nucleotide mismatches**, and we will use Π to denote the average number of nucleotide mismatches for every possible pairwise comparison among the aligned sequences. For n sequences, there are $n(n-1)/2$ possible pairwise comparisons. The value of Π for any sample is therefore defined as

$$\Pi = \frac{\text{Total number of nucleotide mismatches}}{\text{Total number of pairwise comparisons}} \tag{1.8}$$

The estimates of both S and Π have a variance due to both random sampling and population history, which are considered in Chapter 4.

To relate S and Π to underlying parameters that affect genetic variation, such as $\theta = 4N\mu$ where again N is the size of an idealized population and μ the mutation rate, one needs a model of how the DNA sequences evolve. One of the simplest models is the *infinite-sites model*, in which the sequence is assumed to consist of a very large (infinite) number of nucleotide sites with no recombination, in which each nucleotide substitution occurs at a different site and is selectively neutral. After a sufficiently long time, the population eventually reaches steady state at which both S and Π are constant, but the individual sequences are slowly turning over owing to recurrent mutation and random loss. At this steady state, it can be shown that the expected values of S (Watterson 1975) and Π (Kimura 1968) for a sample of n sequences are given by

$$E(S) = \theta\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n-1}\right) \tag{1.9}$$

$$E(\Pi) = \theta \tag{1.10}$$

These expressions are derived and discussed later in this book, but we introduce them now to show how they help to tie together some basic measures of DNA sequence variability. The values of the sum of reciprocals in Equation 1.9 are given for small values of n in Table 1.2. For values of n larger than about 20, the sum of the reciprocals equals approximately $0.577 + \ln(n-1)$ (Nei 1987). It is important to emphasize that, in Equations 1.9 and 1.10, while $\theta = 4N\mu$, the value of μ corresponds to the mutations rate across the entire sequence. In other words, μ equals the average rate of mutation per nucleotide multiplied by the number of nucleotides in each sequence being compared.

At steady state in the infinite-sites model, Π gives one estimate of θ (see Equation 1.10) and S gives another estimate through Equation 1.9 as

$$\theta = S \Big/ \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n-1}\right) \tag{1.11}$$

Testing for equality between these two estimates is one of many ways to detect departures from steady state in the infinite-sites model (Tajima 1989). For example, certain types of selection (or recent population growth) result in an excess of rare alleles, and so the estimate of θ from Equation 1.10 will be smaller than that from Equation 1.11. Likewise, other types of selection (or a recent decrease in population size) result in too few rare alleles, in which case the estimate of θ from Equation 1.10 will be greater than that from Equation

TABLE 1.2 Sums of Reciprocals

n	$\Sigma(1/i)^a$	n	$\Sigma(1/i)$
2	1.000	12	3.020
3	1.500	13	3.103
4	1.833	14	3.180
5	2.083	15	3.252
6	2.283	16	3.318
7	2.450	17	3.381
8	2.593	18	3.440
9	2.718	19	3.495
10	2.829	20	3.548
11	2.929	21	3.598

[a]*Note*: $\Sigma(1/i)$ runs from $i = 1$ to $i = n - 1$.

1.11. This and other tests for departures from selective neutrality are discussed in greater detail in Chapter 4.

The data in Problem 1.3 are typical and can be used to exemplify the calculations. There we considered $n = 5$ sequences each of length 500 nucleotides, and found that the observed number of segregating sites was $S = 16$. Hence the estimate of θ from Equation 1.10 is

$$\theta = 16 / \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 7.68 \qquad (1.12)$$

The average number of nucleotide mismatches Π is the average number of nucleotide differences between all possible pairs of sequences in the sample. For the data in Problem 1.3, the sample of 5 sequences allows $(5 \times 4)/2 = 10$ possible pairwise comparisons. The pairwise comparisons may be considered for each nucleotide in turn. For the polymorphic sites in Problem 1.3, the number of pairwise mismatches totals 6 ($= 2 \times 3$) for six sites (sites 132, 142, 246, 351, 405, and 483), it totals 4 ($= 1 \times 4$) for nine sites (sites 162, 198, 201, 207, 240, 354, 372, 375, and 417), and it totals 7 for one site (site 192). Among the 484 monomorphic nucleotides in Problem 1.3, the number of pairwise mismatches is 0. The average number of pairwise mismatches per pairwise comparison is therefore

$$\Pi = \left(\frac{(6 \times 6) + (4 \times 9) + (7 \times 1) + (0 \times 484)}{10}\right) = 7.90 \qquad (1.13)$$

For the data in Problem 1.3, therefore, the estimate of $\theta = 7.68$ (Equation 1.12, based on S in Equation 1.11) is very close to that of $\theta = 7.90$ (Equation 1.13, based on Π in Equation 1.10). The agreement is very good, but on the other hand the sample size is very small. In this kind of comparison, the values of S and Π depend on the length of the sequences being analyzed, and this length will differ from gene to gene and from one study to the next. How the effects of differing sequence lengths can be removed is examined in Chapter 4.

Estimates of segregating sites and nucleotide mismatches can also be carried out with restriction-site data in the form of restriction fragment length polymorphisms (RFLPs). The simplest way to proceed is to analyze the restriction sites in turn. Each monomorphic restriction site is regarded as identifying six adjacent monomorphic nucleotides (or four monomorphic nucleotides, if the enzyme has a four-base restriction site). Each polymorphic restriction site is regarded as identifying five monomorphic nucleotides and one segregating nucleotide (or three monomorphic sites and one segregat-

ing site, if the enzyme has a four-base restriction site). In other words, each restriction site polymorphism is assumed to result from segregation of a single nucleotide difference in the restriction site. Pairwise comparisons to estimate Π are also carried out under this assumption. The reasoning is illustrated in the following problem.

PROBLEM 1.4 Restriction-site variation was studied in the region of the gene for alcohol dehydrogenase (*Adh*) in a population of *D. melanogaster* descended from animals trapped at a Dutch fruit market in Groningen (Cross and Birley 1986). The region contained a total of 23 sites for five restriction enzymes, each having a six-base restriction site. A total of 16 sites were cut in all flies in the sample. The accompanying table documents the presence (+) or absence (–) of each of the seven polymorphic sites in a sample of 10 chromosomes. Assuming that only one nucleotide is altered for each restriction site that is lost, estimate the value of θ based on the number of segregating nucleotide sites S and on the average number of nucleotide mismatches Π. Do these estimates seem to be equal for these data?

*Bam*HI	*Hind*III	*Pst*I	*Xho*I	*Pst*I	*Eco*RI	*Eco*RI
+	–	–	+	+	–	–
+	–	–	–	–	+	+
–	–	+	–	–	+	–
–	+	–	+	–	+	+
–	+	–	+	–	+	+
–	+	–	+	–	+	+
–	+	–	+	–	+	+
–	+	–	+	–	+	+
–	+	–	+	–	+	+
–	–	–	+	–	+	–

ANSWER Consider first the segregating sites. The 16 monomorphic restriction sites identify 16 × 6 = 96 monomorphic nucleotide sites, whereas the 7 polymorphic restriction sites identify 7 × 5 = 35 monomorphic nucleotide sites and 7 × 1 polymorphic (segregating) sites, assuming only one nucleotide is altered for each restriction site that is lost and that all DNA molecules that are missing a particular restriction site have the same haplotype of nucleotides at that site. Altogether, there are 138 assayed nucleotide sites of which $S = 7$ are segregating sites. Because $n = 10$, the denominator in Equation 1.11 equals 2.829 (from Table 1.2). The estimate of θ based on S is therefore $\theta = 7/2.829 = 2.47$. For estimating Π, there are $(10 \times 9)/2 = 45$ pairwise comparisons, and a restriction site with i "plus" and $(10 - i)$ "minus" means that the segregating nucleotide site results in $i \times (10 - i)$ pairwise mismatches. Therefore, the total number of mismatches for each of the restriction sites, from left to right, equals 16, 24, 9, 16, 9, 9, and 21, respectively, totaling 104. Therefore, the estimate of $\Pi = 104/45 = 2.31$, which is also the estimate of θ based on Π. The estimates $\theta = 2.47$ and $\Pi = 2.31$ are in very good agreement. However, the sample size is too small to generalize this conclusion.

1.7 UTILITY OF GENETIC POLYMORPHISMS

Whether studied through allozymes or nucleotide sequences, natural genetic variation has many uses. Genetic variation provides a set of built-in markers for the genetic study of organisms in their native habitats, including organisms for which domestication or laboratory rearing is unfeasible or for which conventional genetic manipulation is impossible.

Genetic polymorphisms are useful in investigating the genetic relationships among subpopulations in a species. The principle is that alleles are shared among subpopulations because of migration, and therefore similarity in allele frequencies among subpopulations can be used to estimate the rate of migration (see Chapter 6). Within subpopulations, alleles are shared because of common ancestry. For example, the Ainu people of Northern Japan have numerous Caucasoid-like features, including their facial features, light skin, and hairy bodies, yet their genetic polymorphisms clearly show them to be more closely related to other Mongoloid groups (Watanabe et al. 1975). Among the most informative alleles, the Ainu people possess the *D*(*Chi*) allele of transferrin protein and the Di^a allele of the Diego blood group, both of which are virtually restricted to Mongoloid populations. Conversely, the Ainu people lack several alleles that are polymorphic in Caucasoids.

From a practical point of view, genetic polymorphisms are useful in human populations as genetic markers that may be genetically linked to harmful genes that cause disease (see Chapter 10). In kinships with a family history of the disease, the genetic markers can be used to determine which members of the kindred are likely to be carriers of the harmful gene. The markers can also be used in early diagnosis of persons likely to be affected. RFLPs and other types of DNA polymorphisms that are linked to disease genes have also demonstrated their utility as probes for identifying recombinant DNA clones containing the defective genes. The nearby genetic markers enable the defective gene and its function to be identified, thus serving as a first step in the search for effective treatments.

Particularly useful in population genetics are DNA markers with a large number of alleles of moderate frequency. In most organisms, many regions of the genome have multiple alleles consisting of a short sequence of bases repeated in tandem. Multiple alleles result because the number of copies of the repeated sequence may differ from one chromosome to the next. The genotypes are even more variable because each genotype carries two alleles. One of the practical applications of the use of such polymorphisms is in **DNA fingerprinting**, in which the alleles in the DNA from a suspect are matched with those from a crime-scene sample. The examination of a sufficient number of such highly variable regions provides a basis for distinguishing one person from another because no two people (with the exception of identical twins) have the same genotype. Genetic variability of this sort is used in determining paternity as well as in criminal investigations.

DNA fingerprinting has also been applied to studies of the natural mating systems of plants and animals because, with the large number and high specificity of DNA types, close relatives can be detected in populations. In behavioral studies, DNA typing can determine whether organisms that perform mutually altruistic acts are genetically related. Polymorphisms of other types can also be informative about mating systems. For example, the observed frequencies of genotypes can be used to estimate the amount of self-fertilization in populations of monoecious plants or hermaphroditic animals.

From the standpoint of evolutionary biology, sequences of genes and patterns of polymorphism can be used to make inferences about evolutionary history and about the evolutionary process. There is in fact an international project called the DNA barcoding project that aims to catalog unique DNA sequences that can be used to identify species of virtually any organism by means of an ever-expanding database. The sequences of macromolecules contain within themselves a record of their evolutionary history. Organisms with a shared ancestry usually have similar gene sequences. Conversely, similarity in sequence can be regarded as a measure of shared ancestry. As an index of shared ancestry, sequence similarity provides a means of inferring the ancestral relationships among a group of organisms (*molecular phylogenetics*, discussed in Chapter 7). The rates and patterns of change in sequence within species and between closely related species also contain a record of evolutionary forces at work. Population genetics has evolved from a relatively data-poor field to a relatively data-rich field, and numerous new methods of data analysis and hypothesis testing have been developed.

SUMMARY

1. Galton studied mostly continuous traits, including height and weight, which are measured on a quantitative scale, whereas Mendel studied discrete variation, including round versus wrinkled peas, resulting from segregation of the alleles of a single gene.
2. In natural populations, most continuous traits are multifactorial, which means that they are determined by the combined effects of multiple genetic and environmental factors.
3. Multifactorial traits require special methods to study their genetic basis, whereas simple Mendelian variation is the rule for genes and their products.
4. For a simple Mendelian trait controlled by two alleles in a population undergoing random mating, there is a simple relation expected between the allele frequencies (p, q) and the genotype frequencies (p^2, $2pq$, q^2).
5. Protein polymorphisms are frequent in most natural populations, and many alleles for protein variants are common.

6. Widespread polymorphism casts doubt upon the classical theory of genetic variation, which posited that most genetic variation was due to rare, harmful alleles maintained by recurrent mutation. Other types of data also cast doubt on the balance hypothesis, which held that most polymorphisms were maintained owing to a greater fitness of the heterozygous genotype.
7. Neither the classical nor the balance model considered that many genetic polymorphisms might have virtually no effect on the ability of the organism to survive and reproduce (the neutral theory).
8. The ultimate level of genetic variation consists of differences in the nucleotide sequence of DNA at corresponding positions in the chromosomes of different individuals. The human genome consisting of approximately 3 billion nucleotide pairs is estimated to contain about 10 million single-nucleotide polymorphisms.
9. Many single-nucleotide polymorphisms occur in parts of the genome that do not code for proteins. In protein-coding regions, nonsynonymous polymorphisms change the amino acid sequence, whereas synonymous polymorphisms do not change the amino acid sequence.
10. Nucleotide variation along a set of aligned DNA sequences can be quantified according to the number of segregating nucleotide sites and the average number of mismatches between a random pair of sequences. Comparison of these measures is one of many ways to make inferences about the historical evolutionary forces that have been acting upon a gene.
11. Genetic polymorphisms are important in almost every aspect of modern biology including human genetics, molecular and cellular biology, plant and animal breeding, and wildlife management and conservation. Applications to human populations range from the identification of genetic risk factors for complex diseases, to the implication of guilt in criminal cases through DNA fingerprinting.

PROBLEMS

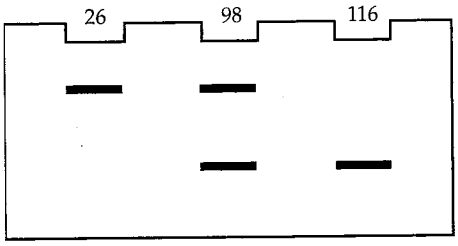
1. Shown here is a table of a phenotypic values measured on each member of a random sample of 100 individuals. Estimate the mean, variance, and standard deviation in the population from which the sample was drawn.

82	80	106	102	82	94	74	123	93	110
102	112	105	97	125	96	89	105	111	97
106	116	127	140	117	94	130	82	79	80
91	114	81	128	73	130	95	94	98	109
99	96	109	71	90	95	107	92	112	110
87	101	113	117	97	80	139	108	107	103
120	86	90	67	88	87	120	124	112	107

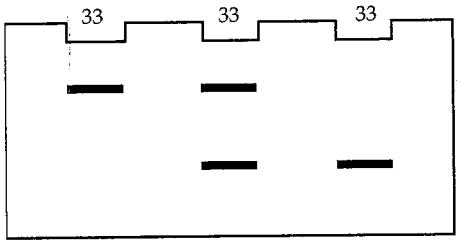
120	101	104	97	72	106	113	88	120	99
106	75	100	82	98	126	103	118	120	104
106	73	88	142	89	96	112	95	99	79

2. Assuming that the sample in Problem 1 was drawn from a normal distribution, what number of individuals in the sample is expected to have a phenotypic value exceeding the mean plus one standard deviation? What number of individuals in the sample is expected to have a phenotypic value smaller than the mean minus one standard deviation? How do these predictions compare with the observations?
3. Assuming that the sample in Problem 1 was drawn from a normal distribution, what number of individuals in the sample is expected to have a phenotypic value exceeding the mean plus two standard deviations? What number of individuals in the sample is expected to have a phenotypic value smaller than the mean minus two standard deviations? How do these predictions compare with the observations?
4. A typical means for using a computer to generate normally distributed random numbers is to choose 12 uniform random numbers and add them up. After scaling the sum by a constant that depends on the mean and the variance of the uniform distribution, the result represents a sample from a normal distribution. Why does this approach work?
5. One statement of the central limit theorem is that the sum of independent, identically distributed random variables has a limiting normal distribution. If the random variables that are added are not independent, but instead exhibit a positive correlation in successive draws, how would you expect the sum deviate from the normal distribution predicted by the central limit theorem?
6. Protein electrophoresis is carried out using blood samples from wild mice for a gene with two alleles whose product runs either fast or slow in the gel. Among 100 females, 64 have only a fast band, 4 have only a slow band, and 32 have both a fast and a slow band. Among 100 males, however, 80 have only a fast band and 20 have only a slow band, with no apparent heterozygous genotypes. What mode of inheritance might explain the absence of heterozygous males?
7. Many proteins exist in an active form only as dimers, with the two subunits held together by hydrogen bonds or covalent cysteine bridges. If an enzyme is active only as a dimer, then in a population with two alleles encoding fast and slow migrating forms of the protein, what kind of banding pattern would you expect from tissue from a heterozygous genotype?

8. The diagram shows the result of electrophoresis of tissue samples from 240 Canadian geese, *Branta canadensis*, stained for the enzyme aldehyde oxidase. The samples were placed in the depressions ("wells") shown at the top of the gel, and as electrophoresis took place the proteins migrated in a downward direction. The numbers are the number of individuals in the sample showing each banding pattern. Estimate the allele frequencies of the fast and slow alleles, and calculate the expected genotype frequencies assuming Hardy-Weinberg proportions. Do the observed results seem to fit these expected values?



9. The accompanying gel diagram summarizes the results of electrophoresis of 99 individuals from a population of pointed phlox (*Phlox cuspidata*). Does this population appear to be in Hardy-Weinberg proportions?



10. A gene with two alleles in a population has genotype frequencies P, Q, and R for the genotypes AA, Aa, and aa, respectively, Show that, if the population is in Hardy-Weinberg proportions, then the expected relation between the genotype frequencies is $PR = Q^2/4$.
11. How many copies of a fragment of DNA should be present after 30 rounds of PCR, assuming perfect efficiency?
12. Four sequences of a 1200 bp DNA fragment from a gene gave the following counts of pairwise differences: 4, 7, 5, 3, 6, 5. What is the estimate of the number of pairwise mismatches for this sample?

13. The nucleotide sequences shown here are the complete set of polymorphic nucleotides found in a region of 5 kb in a sample of six chromosomes from corn, *Zea mays*. The polymorphic nucleotides are not adjacent, as shown here, but scattered throughout the 5 kb. What is the number of segregating sites S and the average number of pairwise mismatches Π among these sequences? Estimate θ from both S and Π. Do these estimates seem to be consistent with one another? If this result were found for a larger sample, what might it suggest about the selective forces acting on the sequence polymorphisms?

```
GCCTT  TATGG  CCTGT  ATGAG
ACTAT  TAAGG  CTTGT  TTGAT
ACCAC  TGTCG  CCCGT  ACGCG
GTCAT  TGTGG  TCCTC  TTGAG
GCTTT  TATGA  CCTTT  ATAAG
ACCAT  CATGA  CCTTT  ATAAT
```

14. In forensic applications of genetics, if the DNA fingerprint taken from a crime-scene sample and that of a suspect do not match, the confidence one has in the conclusion is much greater than if the DNA fingerprints do match. What conclusion would be drawn, and why can one have confidence in this conclusion?

2.1 Random Mating 46

Nonoverlapping Generations 47

2.2 The Hardy-Weinberg Principle 48

Random Mating of Genotypes versus Random Union of Gametes 50

Implications of the Hardy-Weinberg Principle 52

2.3 Testing for Hardy-Weinberg Equilibrium 54

Difficulties in Testing for Hardy-Weinberg Equilibrium 54

Complications of Dominance 62

Frequency of Heterozygotes 66

2.4 Extensions of the Hardy-Weinberg Principle 67

Three or More Alleles 67

X-Linked Genes 70

2.5 Linkage and Linkage Disequilibrium 73

Difficulties in Testing for Linkage Equilibrium 82

Relative Measures of Linkage Disequilibrium: D′ and r^2 83

2.6 Causes of Linkage Disequilibrium 84

Linkage Disequilibrium Due to Population Admixture 85

Linkage Disequilibrium Due to Reduced Recombination 86

CHAPTER 2

ORGANIZATION OF GENETIC VARIATION

The word *population* has so far been used in an informal, intuitive sense to refer to a group of organisms belonging to the same species. Further discussion and clarification of the concept is necessary at this time. In population genetics, the word **population** does not usually refer to an entire species; it refers instead to a group of organisms of the same species living within a sufficiently restricted geographical area so that any member can potentially mate with any other member of the opposite sex. Precise definition of such a unit is difficult and varies from species to species because of the almost universal presence of some sort of *geographical structure* in species—some typically nonrandom pattern in the spatial distribution of organisms. Members of a species are rarely distributed homogeneously in space: there is almost always some sort of clumping or aggregation, some schooling, flocking, herding, or colony formation. Population subdivision is often caused by environmental patchiness, areas of favorable habitat intermixed with unfavorable areas. Such environmental patchiness is obvious in the case of, for example, terrestrial organisms on islands in an archipelago, but patchiness is a common feature of most habitats—freshwater lakes have shallow and deep areas, meadows have marshy and dry areas, forests have sunny and shady areas. Population subdivision can also be caused by social behavior, as when wolves form packs. Even the human population is clumped or aggregated—into towns and cities, away from deserts and mountains. The geographical boundaries matter, because the members of most species, humans included, are likely to choose their mate from the local area in which they live.

The local interbreeding units of possibly large, geographically structured populations are of some interest because it is within such local units that adaptive evolution takes place through systematic changes in allele frequency. Such local interbreeding units—often called **local populations** or **demes**—are the fundamental units of population genetics. Local populations are the actual, evolving units of a species. Unless otherwise specified (or clear from context), the term *population* as used in this book means *local population*. Local populations are sometimes also referred to as *Mendelian populations* or *subpopulations*.

2.1 RANDOM MATING

In sexual organisms, genotypes are not transmitted from one generation to the next. Genotypes are broken up in gamete formation by the processes of segregation and recombination, and they are assembled anew in each generation in fertilization: Genotypes give rise to gametes, which unite and begin a new generation of genotypes. The proportion of a specified genotype in a population is the **genotype frequency**. The formation of a genotype in newly fertilized eggs is determined by the opportunity for the relevant gametes to come together in fertilization, and the opportunity for gametes to come together in fertilization is determined by the matings that take place among organisms of reproductive age in the previous generation. To put the matter in a slightly different way, the genotypes of the mating pairs determine the genotypes of the progeny. Furthermore, one can derive mathematical relationships between the frequencies of mating pairs and the frequencies of progeny genotypes. These mathematical relationships are usually deduced from models in which the types of matings in the population are specified. One of the important models in population genetics is that of **random mating**, in which mating pairs have the same frequencies as if they were formed by random collisions between genotypes. The chance that an organism mates with another having a prescribed genotype is therefore equal to the frequency of the prescribed genotype in the population. For example, suppose that in some population the genotype frequencies of *AA*, *Aa*, and *aa* are 0.16, 0.48, and 0.36, respectively; if mating is random, *AA* males mate with *AA*, *Aa*, and *aa* females in the proportions 0.16, 0.48, and 0.36, respectively; these same proportions apply to the mates of *Aa* and *aa* males.

Superficial appearances to the contrary, random mating is not a simple or trivial process. One complication is that random mating depends on the trait. Mating can be random with respect to some traits but nonrandom with respect to other traits at the same time and in the same population. For example, it is perfectly consistent for a human population to undergo random mating with respect to blood groups, allozyme phenotypes, restriction fragment length polymorphisms, and many other characteristics, but at the same time to engage in nonrandom mating with respect to other traits such as skin

color and height. A second complication is population substructure. Paradoxical as it may seem, random mating may be observed within each of the subpopulations constituting a larger population, but random mating may still fail to hold in the population as a whole. (The reason for this paradox is discussed in Chapter 6.) In spite of these and other complications, random mating plays an important role in models in population genetics because random mating serves as a convenient standard of comparison for more complex mating systems.

Nonoverlapping Generations

One of the most important mathematical models in population genetics is the *nonoverlapping generation model,* in which the cycle of birth, maturation, and death includes the death of all organisms present in each generation before the members of the next generation mature. The nonoverlapping generation model is diagrammed in Figure 2.1. The model applies literally only to organisms with a very simple sort of life history, such as certain short-lived insects or annual plants that have a short growing season. In such plants, all members of any generation germinate at about the same time, mature together, shed their pollen, are fertilized almost simultaneously, and die immediately after producing the new generation. This sort of hypothetical population, with its simple life history, is used in population genetics as a first approximation to populations that have more complex life histories. Although at first glance the model seems hopelessly oversimplified, calculations of expected

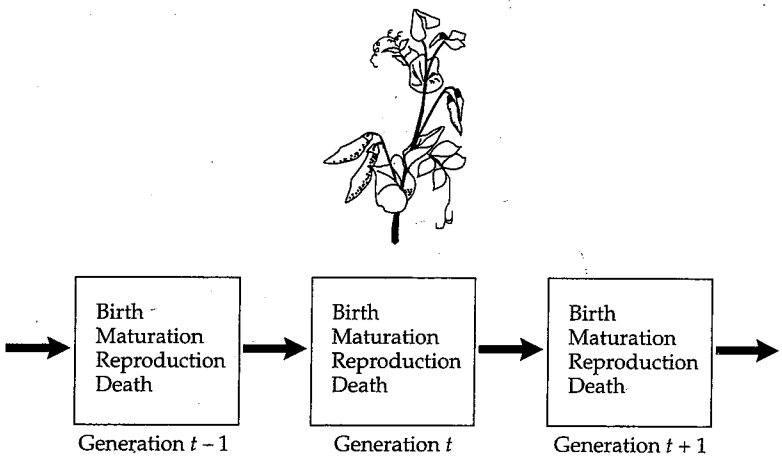


FIGURE 2.1 The nonoverlapping generation model. The life history of the organism is assumed to be like that of an annual plant (or any short-lived organism), and the generations are assumed to be separated in time (discrete generations). Although the model is simple, it provides a convenient first approximation to populations with more complex life histories.

genotype frequencies based on this model are adequate for many purposes. In some applications, the nonoverlapping generation model turns out to be a useful approximation even for populations with a long and complex life history, such as human beings.

2.2 THE HARDY-WEINBERG PRINCIPLE

Genotype frequencies are determined in part by the pattern of mating. In this section, we consider the consequences of random mating in the model with nonoverlapping generations. To deduce the genotype frequencies under random mating, additional assumptions are needed. First, the allele frequencies should not change from one generation to the next because of systematic evolutionary forces, the most important of which are mutation, migration, and natural selection. For the moment, these evolutionary forces are assumed to be absent or negligibly small in magnitude. (Their effects are discussed in Chapters 4 and 5.) Second, the population must be large enough in size that the allele frequencies are not subject to change merely because of sampling error. Variation in allele frequency owing to sampling error in small populations is called *random genetic drift* and is the subject of Chapter 3. Although random genetic drift is present unless the population is infinite in size, the magnitude of the effect on allele frequency over a small number of generations is usually sufficiently small that the process can be ignored if population size is 500 or more. The qualifier "over a small number of generations" is important because the effects of random genetic drift are cumulative. Considered over a sufficiently large number of generations, random genetic drift can be important even in populations of size 10^6 or more.

Before proceeding further, it may be helpful to summarize the assumptions that we are making:

- The organism is diploid.
- Reproduction is sexual.
- Generations are nonoverlapping.
- The gene under consideration has two alleles.
- The allele frequencies are identical in males and females.
- Mating is random.
- Population size is very large (in theory, infinite).
- Migration is negligible.
- Mutation can be ignored.
- Natural selection does not affect the alleles under consideration.

Collectively, these assumptions summarize the *Hardy-Weinberg model*, named after the English mathematician G. H. Hardy (1877–1947) and the German physiologist Wilhelm Weinberg (1862–1937), who, in 1908, independently formulated the model and deduced its theoretical predictions of genotype frequency.

In the Hardy-Weinberg model, the mathematical relation between the allele frequencies and the genotype frequencies is given by

$$AA: p^2 \qquad Aa: 2pq \qquad aa: q^2 \tag{2.1}$$

in which p^2, $2pq$, and q^2 are the frequencies of the genotypes *AA*, *Aa*, and *aa* in zygotes of any generation, p and q are the allele frequencies of *A* and *a* in gametes of the previous generation, and $p + q = 1$. The frequencies displayed in Equation 2.1 constitute the **Hardy-Weinberg principle** or the **Hardy-Weinberg equilibrium (HWE)**.

One rationale for the Hardy-Weinberg principle displayed in Equation 2.1 is based on the outcome of repeated and independent trials. With random mating, the choices of male gamete and female gamete are independent trials, and so pairs of gametes carrying the alleles *AA*, *Aa*, or *aa* are expected in proportions given by $(p\,A + q\,a)^2 = p^2\,AA + 2pq\,Aa + q^2\,aa$. A graphical illustration of the rationale of independent trials is shown in Figure 2.2. The chance of two *A*-bearing gametes coming together is $p \times p = p^2$ and that of two *a*-bearing gametes coming together is $q \times q = q^2$; for the heterozygote, the chance is $p \times q + q \times p = 2pq$ because the female gamete could carry *A* and the male gamete carry *a*, or the other way around.

Male gametes

Allele	*A*	*a*
Frequency	p	q

Female gametes: Allele	Frequency		
A	p	*AA* p^2	*Aa* pq
a	q	*aA* qp	*aa* q^2

Summed frequencies in zygotes:

AA: $P' = p^2$
Aa: $Q' = pq + qp = 2pq$
aa: $R' = q^2$

FIGURE 2.2 Cross-multiplication square showing Hardy-Weinberg frequencies resulting from random mating with two alleles. Such a square is often referred to as a *Punnett square*, named in honor of early geneticist Reginald C. Punnett (1875–1967).

PROBLEM 2.1 The assumption that random mating of individuals is equivalent to random union of gametes (see Figure 2.2) can be used to explore the consequences of random mating when the allele frequencies are different in males and females. Imagine an autosomal gene with alleles A and a in a population in which the allele frequencies in females are p and q, respectively, with $p + q = 1$, and in which the corresponding allele frequencies in males are p' and q' ($p' + q' = 1$). After one generation of random mating, what are the genotype frequencies in females and males? What are the allele frequencies in females and males? What does this result imply about HWE in subsequent generations?

ANSWER For this situation a Punnett square like that in Figure 2.2 would have allele frequencies p and q across the top, since these are the frequencies of A and a in female gametes; and it would have p' and q' along the side, since these are the frequencies of A and a in male gametes. After one generation of random mating, therefore, the genotype frequencies are pp' (AA), $pq' + qp'$ (Aa), and qq' (aa). These frequencies apply equally to female and male offspring, since both parents contribute equally to the inheritance of an autosomal gene. The allele frequencies in both sexes in the offspring generation are therefore $p^* = pp' + (pq' + qp')/2 = (2pp' + pq' + qp')/2 = [p(p' + q') + p'(p + q)]/2 = (p + p')/2$ for A and likewise $q^* = (q + q')/2$ for a. (Note that these values are the averages of the allele frequencies in females and males of the previous generation.) Therefore, one generation of random mating is sufficient to equalize the allele frequencies in the sexes, and in subsequent generations the genotype frequencies will be given by the HWE with $p = p^*$ and $q = q^*$.

Random Mating of Genotypes versus Random Union of Gametes

Figure 2.2 implicitly assumes the important premise that random mating of genotypes is equivalent to random union of gametes. A demonstration of this premise in the case of two alleles is outlined in Table 2.1, in which pairs of genotypes are chosen at random to form matings. The genotype frequencies of AA, Aa, and aa in the parental generation are written as P, Q, and R, respectively, where $P + Q + R = 1$. In terms of the genotype frequencies, the allele frequencies p of A and q of a are as follows:

$$p = (2 \times P + Q)/2 = P + Q/2$$
$$q = (2 \times R + Q)/2 = R + Q/2 \qquad [2.2]$$

Note that $p + q = P + Q + R = 1.0$; this result is a consequence of the fact that the gene has only two alleles.

With two alleles of a gene, there are six possible types of matings. When mating is random, these matings take place in proportion to the genotypic frequencies in the population, and the types of mating pairs are given by successive terms in the expansion of $(P\,AA + Q\,Aa + R\,aa)^2$. For example, the pro-

TABLE 2.1 Demonstration of the Hardy-Weinberg Principle

Mating	Frequency of mating (parents)	Frequency of zygotes (progeny) AA	Aa	aa
$AA \times AA$	P^2	1	0	0
$AA \times Aa$	$2PQ$	$\frac{1}{2}$	$\frac{1}{2}$	0
$AA \times aa$	$2PR$	0	1	0
$Aa \times Aa$	Q^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$Aa \times aa$	$2QR$	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times aa$	R^2	0	0	1
	Totals (next generation)	P'	Q'	R'

therefore

$$P' = P^2 + \frac{2PQ}{2} + \frac{Q^2}{4} = (P + \tfrac{Q}{2})^2 = p^2$$
$$Q' = \frac{2PQ}{2} + 2PR + \frac{Q^2}{2} + \frac{2QR}{2} = 2(P + \tfrac{Q}{2})(R + \tfrac{Q}{2}) = 2pq$$
$$R' = \frac{Q^2}{4} + \frac{2QR}{2} + R^2 = (R + \tfrac{Q}{2})^2 = q^2$$

portion of $AA \times AA$ matings is $P \times P = P^2$. Similarly, the proportion of $AA \times Aa$ matings is $2 \times P \times Q$ because the mating can b between either an AA female and an Aa male (proportion $P \times Q$) or else between an Aa female and an AA male (proportion $Q \times P$). The frequencies of these and the other types of matings are given in the second column of Table 2.1.

The genotypes of the zygotes produced by the matings are given in the last three columns of Table 2.1. The offspring frequencies follow from Mendel's law of segregation, which states that an Aa heterozygote produces an equal number of A-bearing and a-bearing gametes. The AA and aa homozygotes produce only A-bearing and only a-bearing gametes, respectively. Thus, the mating $AA \times aa$ produces all Aa zygotes, the mating $AA \times Aa$ produces $\frac{1}{2}$ AA and $\frac{1}{2}$ Aa zygotes, the mating $Aa \times Aa$ produces $\frac{1}{4}$ AA, $\frac{1}{2}$ Aa, and $\frac{1}{4}$ aa zygotes, and so forth.

The genotype frequencies of AA, Aa, and aa zygotes after one generation of random mating are denoted in Table 2.1 as P', Q', and R', respectively. These values are calculated as the sum of the cross-products shown at the bottom of the table. The genotype frequencies simplify to $P' = p'$, $Q' = 2pq$, and $R' = q^2$, where p and q are the allele frequencies given in Equation 2.2. Note that the parental genotype frequencies—P, Q, and R—were completely arbitrary except for the requirement that $P + Q + R = 1$. Therefore, the Hardy-Weinberg frequencies are attained after one generation of random mating irrespective of the genotype frequencies in the parental generation.

PROBLEM 2.2 In an experimental population of *D. melanogaster*, the genotype frequencies for two alleles, $E6^F$ and $E6^S$, of the gene coding for esterase-6 were found to be consistent with Hardy-Weinberg proportions with allele frequencies of 0.3579 for $E6^F$ and 0.6421 for $E6^S$ (Mukai et al. 1974). Assuming that all of the assumptions of the Hardy-Weinberg model hold, particularly those pertaining to random mating in a large population with no mutation, selection, or migration, make a table of mating frequencies similar to Table 2.1 for the esterase-6 alleles. Then calculate the genotype frequencies expected in the next generation along with the corresponding allele frequencies.

ANSWER The Hardy-Weinberg frequencies among parents are *FF*: 0.1281; *FS*: 0.4596, and *SS*: 0.4123. Therefore, the expected frequencies of the matings are: *FF* × *FF* (0.0164); *FF* × *FS* (0.1177); *FF* × *SS* (0.1056); *FS* × *FS* (0.2112); *FS* × *SS* (0.3790); and *SS* × *SS* (0.1700). The expected genotype frequencies among the zygotes are, for *FF*, 0.0164 + 0.1177/2 + 0.2112/4 = 0.1281; for *FS*, 0.1177/2 + 0.1056 + 0.2112/2 + 0.3790/2 = 0.4596; and for *SS*, 0.2112/4 + 0.3790/2 + 0.1700 = 0.4123; note that these are the same as in the parental generation. The allele frequencies of *F* and *S* are again 0.3579 and 0.6421, respectively.

Implications of the Hardy-Weinberg Principle

The Hardy-Weinberg principle has provided the foundation for many theoretical and experimental investigations in population genetics. However, the theory is far from profound, and the applicability is far from universal. Hardy especially seems to have regarded the Hardy-Weinberg principle as virtually self-evident. He wrote, "I should have expected the very simple point which I wish to make to have been familiar to biologists." In fact, it was familiar to some biologists—the basic principle had been noted as early as 1903 by the Harvard geneticist William E. Castle (1867–1962). Castle's work was little known, however, and Hardy was writing to counter an argument put forth against Mendelism that phenotypic ratios of 3 dominant to 1 recessive should be encountered frequently in natural populations if the mechanism of Mendelian heredity were generally applicable. The immediate implication of the Hardy-Weinberg principle was to refute the 3 : 1 argument by showing that the genotypic ratio of *A*– : *aa* is determined by the allele frequencies and has no special tendency to attain one particular ratio as any other. (The dash in a genotypic symbol is a wild card symbolizing any of the possible alleles; for example, the genotype symbolized *A*– includes both *AA* and *Aa*.)

Beyond the virtue of simplicity, why would anyone want to consider a model based on so many restrictive and seemingly incorrect assumptions? And in what sense can such a simple model be considered fundamental? Among several reasons, two stand out. First, the Hardy-Weinberg model is a

reference model in which there are no evolutionary forces at work other than those imposed by the process of reproduction itself. In this sense, the model is similar to models in mechanical physics where objects fall through the sky without wind resistance or roll down inclined planes without friction. The model affords a baseline for comparison with more realistic models in which evolutionary forces can change allele frequencies. Perhaps more importantly, the Hardy-Weinberg model separates life history into two intervals: gametes combining to produce zygotes, and zygotes maturing to become adults. In constructing more complex and realistic models, one can often introduce the complications into the zygotes-to-adults part of the life cycle. This is the usual approach for considering, for example, the effects of migration into the population or the effects of differential survival among the genotypes. With all sources of change in allele frequency accounted for in the zygotes-to-adults component, the gametes-to-zygotes component follows from the principle that random union of gametes and results in the Hardy-Weinberg proportions among zygotes. In other words, the Hardy-Weinberg model is fundamental in the sense that the approach of tracking allele and genotype frequencies through time can be generalized to more realistic situations.

One of the most important implications of the Hardy-Weinberg principle emerges when we calculate the allele frequencies of A and a in the next generation from the formulas for P', Q', and R' in Table 2.1. Using the result in Equation 2.2, the allele frequency of A among the zygotes equals $P' + Q'/2 = p^2 + 2pq/2 = p^2 + pq = p(p + q) = p$. Likewise, the allele frequency of a among zygotes equals $R' + Q'/2 = q^2 + 2pq/2 = q$. Thus, the allele frequencies in the next generation are exactly the same as they were the generation before. With random mating, the allele frequencies remain the same generation after generation. In any generation, therefore, the genotype frequencies are p^2, $2pq$, and q^2 for AA, Aa, and aa, respectively, as given in Equation 2.1. The constancy of allele frequency—and therefore of the genotypic composition of the population—is the single most important implication of the Hardy-Weinberg principle. The constancy of allele frequencies implies that, in the absence of specific evolutionary forces to change allele frequency, the mechanism of Mendelian inheritance, by itself, keeps the allele frequencies constant and thus preserves genetic variation. A second item of interest is that the Hardy-Weinberg frequencies are attained in just one generation of random mating if the allele frequencies are the same in males and females. This, however, is true only with nonoverlapping generations; in populations with more complex life histories, the Hardy-Weinberg frequencies are attained gradually over a period of several generations.

The principle that Mendelian inheritance preserves genetic variation has important implications for evolutionary theory. In Darwin's time, the prevailing view of heredity was one of *blending*, in which all of the offspring of a mating were assumed to have a hereditary composition equal to the mean of that of the parents. In this scenario, the genetic variation in a population

rapidly disappears. The problem was pointed out by Darwin's critics and understood by Darwin himself to be a potentially fatal flaw in his theory. He handled the problem by assuming a very high rate of appearance of new genetic variation through the effects of the environment on the genome.

It was not until the 1930s that Mendelian inheritance was finally incorporated into evolutionary theory in what was then called the *modern synthesis*. In the modern synthesis, Darwin's idea of a high mutation rate generating genetic variation that was quickly dissipated by the mechanism of heredity was replaced with a model in which a low mutation rate generated genetic variation that tends to be preserved, and therefore accumulated, by the Mendelian mechanism of inheritance.

It is important to note here that conventional statistical tests for Hardy-Weinberg proportions (such as those discussed below) are not very sensitive to deviations from the expected genotype frequencies. Consequently, the mere fact that observed genotype frequencies may happen to fit the Hardy-Weinberg proportions cannot be taken as evidence that all of the assumptions underlying the model are valid. The most that can be concluded is that, whatever departures from the assumptions there may be, they are not sufficiently large to result in deviations from HWE that are detectable with conventional statistical tests.

2.3 TESTING FOR HARDY-WEINBERG EQUILIBRIUM

As a concrete example of the application of the Hardy-Weinberg principle, we will use a common polymorphism in the human gene encoding a multifunctional transcription factor protein known as p53. This protein is one of the genome's key players in preventing cancer, as evidenced by the observation that mutations in p53 are among the most common genetic changes found in cancers. The p53 protein is activated by DNA damage or other problems in DNA synthesis. When activated, p53 turns on transcription of a number of other genes encoding proteins that interrupt the cell cycle until the damage can be repaired; if the damage cannot be repaired, the proteins cause the cell to undergo programmed cell death. The p53 protein also helps regulate the formation of blood vessels and is a mediator of pregnancy by estrogen and progesterone activation.

The human population has a common amino acid polymorphism in p53, in which the amino acid at position 72 can be either arginine (Arg) or proline (Pro). It will be convenient to call these the corresponding alleles *Arg* and *Pro*. In homozygous genotypes, the *Arg* allele is weakly associated with cutaneous melanoma (Shen et al. 2003), whereas the *Pro* allele is weakly associated with recurrent spontaneous abortions (Pietrowski et al. 2005).

In one study of the p53 polymorphism among 318 Caucasians (Pietrowski et al. 2005), the observed number for each genotype was 166 *Arg/Arg*, 120 *Arg/Pro*, and 32 *Pro/Pro*. To determine whether these genotype frequencies

are in accord with HWE, the allele frequencies of *Arg* and *Pro* must first be estimated. The estimated allele frequency $\hat{p}$ of *Arg* is $(2 \times 166 + 120)/(2 \times 318) = 0.7107$ and the frequency $\hat{q}$ of *Pro* is $(120 + 2 \times 32 + 120)/(2 \times 318) = 0.2893$. 这儿在计算Pro的频率时，是不是多加了个120 Were the population in HWE, we would expect the genotype frequencies of *Arg/Arg*, *Arg/Pro*, and *Pro/Pro* to be p^2, $2pq$, and q^2, respectively, where p and q are the allele frequencies in the underlying population from which the sample was drawn. Because p and q are parameters, their true values are unknown. However, in testing for HWE we can substitute the estimated values to obtain the expected proportions *Arg/Arg*: $(0.7107)^2 = 0.5051$, *Arg/Pro*: $2(0.7107)(0.2893) = 0.4112$, and *Pro/Pro*: $(0.2893)^2 = 0.0837$, respectively. Because the sample size is 318, the expected numbers of the *Arg/Arg*, *Arg/Pro*, and *Pro/Pro* genotypes are $0.5051 \times 318 = 160.6$, $0.4112 \times 318 = 130.8$, and $0.0837 \times 318 = 26.6$, respectively.

At this point, it is convenient to tabulate the data into three columns:

Arg/Arg	Observed number 166	Expected number 160.6
Arg/Pro	Observed number 120	Expected number 130.8
Pro/Pro	Observed number 32	Expected number 26.6

With the data so arrayed, it is evident that the fit between the observed numbers and the expected numbers, though not perfect because of chance statistical fluctuations in the number of each genotype that may be included in any given sample, is nevertheless very close. To verify this conclusion, we will apply a conventional statistical test to the data in order to assess quantitatively the closeness of fit. A test commonly employed in population genetics is called the **chi-square test**, which is based on the value of a number, called χ^2, calculated from the data as

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{2.3}$$

where *observed* and *expected* refer to the observed and expected *numbers* in any genotypic class, and the Σ sign denotes that the values are to be summed over all genotypic classes. In the case at hand,

$$\begin{aligned}\chi^2 &= (166 - 160.6)^2/160.6 \\ &\quad + (120 - 130.8)^2/130.8 \\ &\quad + (32 - 26.6)^2/26.6 \\ &= 2.17\end{aligned}$$

To be completely unambiguous, some statisticians prefer use of the symbol χ^2 for the realized value of the test statistic defined Equation 2.3, in order to distinguish between the test statistic and the true χ^2 distribution itself. The distinction should certainly be kept in mind, but we will not recognize it formally with different symbols.

此处怀疑印刷错误。“为了完全避免意义上的含糊不清，许多统计学家倾向于使用另一个符号表示方程（2.3）的数值，以示和卡方分布的区分“，可是另一个符号为什么还是卡方的符号呢？

Associated with any χ^2 value is a second number called the **degrees of freedom** for that χ^2. In general, the number of degrees of freedom (df) associated with a χ^2 value equals

$$df = \text{Number of classes of data} - \text{Number of parameters estimated from the data} - 1$$

In the *p53* example, there are three classes of data and one parameter (p) estimated from the data, and so $df = 3 - 1 - 1 = 1$. Note that a degree of freedom is not subtracted for estimating q because of the relation $q = 1 - p$; that is, once p has been estimated, the estimate of q is automatically determined, and so we deduct just the one degree of freedom corresponding to p.

Calculation of χ^2 and its associated degrees of freedom is carried out in order to obtain a number for assessing goodness of fit, and this number can be determined from Figure 2.3. To use the chart, find the value of χ^2 along the horizontal axis, then move vertically from this value until the line for the number of degree of freedom is intersected, then move horizontally from the point of intersection to the vertical axis and read the corresponding probability value P. In our case, with $\chi^2 = 2.17$ and one degree of freedom, the corresponding probability value is about $P = 0.14$. (Rather than use Figure 2.3 for the P value, you can search for "chi-square calculator" on your favorite web browser, then pick a site and follow the instructions.)

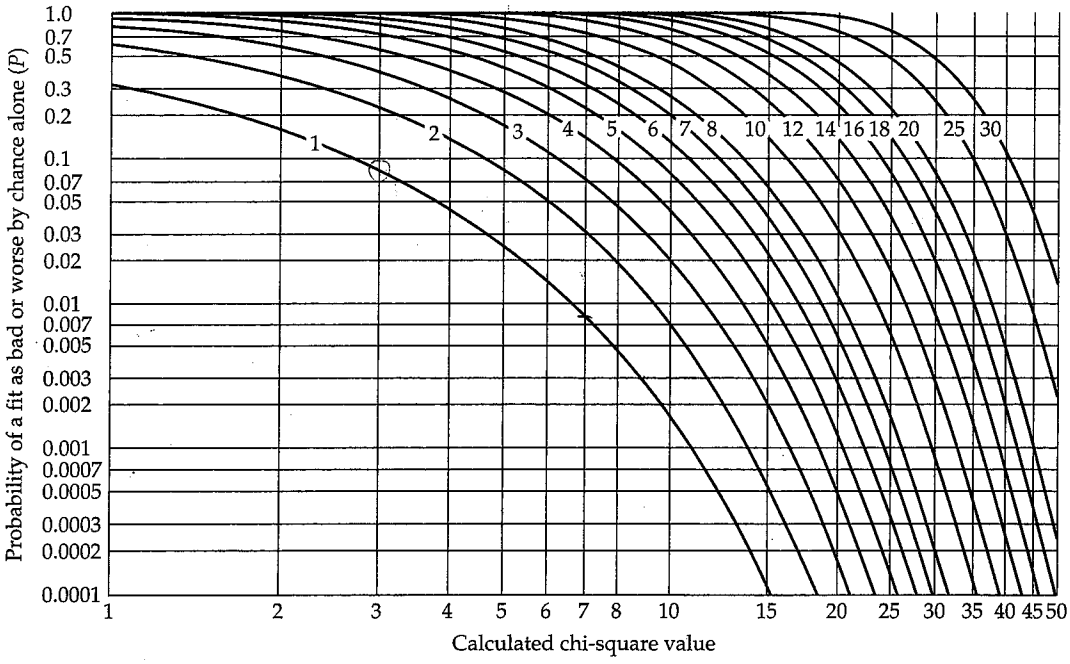
The probability associated with a particular χ^2 test has the following interpretation: The P value is the probability that chance alone could produce a deviation between the observed and expected values at least as great as the deviation actually realized. Thus, if the probability is large, it means that chance alone could account for the deviation, and it strengthens our confidence in the validity of the model used to obtain the expectations—in this case, the Hardy-Weinberg model. On the other hand, if the probability associated with the χ^2 is small, it means that chance alone is not likely to lead to a deviation as large as actually realized, and it undermines our confidence in the validity of the model. Where exactly the cutoff should be between a "large" probability and a "small" one is, of course, not obvious, but there is an established guideline to follow. If the probability is less than 0.05, then the goodness of fit is considered sufficiently poor that the model is judged invalid for the data; alternatively, if the probability is greater than 0.05, the fit is considered sufficiently close that the model is not rejected. Because the

FIGURE 2.3 Graph of χ^2. To use the graph, find the value of χ^2 along the horizontal axis, then read the probability value for the appropriate number of degrees of freedom from the vertical axis. Values of χ^2 smaller than 1 are not shown, because these are never significant. ▶

probability in the *p53* example is 0.14, which is greater than 0.05, we have no reason to reject the hypothesis that the genotype frequencies are in Hardy-Weinberg proportions for this gene.

PROBLEM 2.3 The gene *CCR5* encodes a protein coreceptor used by the AIDS virus for entry into certain white blood cells. Many populations are polymorphic for a deletion of part of the coding sequence that results in an inactive protein. This polymorphism was originally discovered among persons infected with the virus who had remained free of the AIDS disease for at least 10 years. The protective effect of the deletion, denoted *CCR5*Δ, is at least a factor of two. In one study of 338 individuals from Denmark and nearby Germany (Lucotte and Mercier 1998), the observed numbers of the genotypes were as follows: 265 nonmutant *CCR5/CCR5* homozygotes, 66 *CCR5/CCR5*Δ heterozygotes, and 7 *CCR5*Δ/*CCR5*Δ homozygotes. Estimate the allele frequencies of *CCR5* (p) and *CCR5*Δ (q) and carry out a chi-square test of goodness of fit between the observed genotype frequencies and their Hardy-Weinberg expectations. Is there any reason to reject the hypothesis of Hardy-Weinberg proportions for this gene?

ANSWER $\hat{p} = 0.882$ and $\hat{q} = 0.118$. The expected numbers of *CCR5/CCR5*, *CCR5/CCR5*Δ, and *CCR5*Δ/*CCR5*Δ are 262.9, 70.4, and 4.7, respectively. The $\chi^2 = 1.42$ with one degree of freedom. The associated probability from Figure 2.3 is about 0.25, so there is no reason to reject the hypothesis of HWE.



Difficulties in Testing for Hardy-Weinberg Equilibrium

Testing for HWE is important. For example, in studies in human genetics, deviations from HWE of a marker gene among individuals affected with a genetic disease may be helpful in identifying the location of a disease-susceptibility allele near the marker gene along the chromosome (Nielsen et al. 1999). Deviations from HWE may also alert the investigator to possible genotyping errors (Xu et al. 2002). Yet tests for HWE are often unreported or botched (Salanti et al. 2005). There are also complications that can invalidate the simple chi-square calculation in Equation 2.3. Some of the ways of dealing with these complications are examined next.

SAMPLE SIZE TOO SMALL Problem 2.3 illustrates one of the issues that can arise in testing for HWE. The allele frequency of *CCR5Δ* is sufficiently small that, even for a sample size of 338, the observed number of homozygotes was only 7 and the expected number is only 4.7. With such small numbers, chance alone can have a substantial effect on the composition of any actual sample. This is a problem for the formula in Equation 2.3, because this expression has a chi-square distribution only when each of the classes of data has a sufficiently large expected number. What "sufficiently large" means is a matter of judgment, but most statisticians agree that the standard chi-square test should not be trusted when any of the expected numbers is smaller than 5. The *CCR5Δ* example violates this convention. In such cases, many statisticians recommend calculating a chi-square value that is somewhat more conservative than that in Equation 2.3, namely,

$$\chi^2 = \sum \frac{(|observed - expected| - 0.5)^2}{expected} \tag{2.4}$$

In this expression, the vertical bars mean absolute value (the magnitude of the enclosed number disregarding the sign). The fact that 0.5 is subtracted from each difference in the numerator before taking the square serves to reduce the value of the chi-square. In the *CCR5Δ* example, Equation 2.4 yields $(2.1 - 0.5)^2/262.9 + (4.4 - 0.5)^2/70.4 + (2.3 - 0.5)^2/4.7 = 0.915$. This value is not found on the chart in Figure 2.3 because the graph begins at $\chi^2 = 1$ for clarity. In practice, values of χ^2 smaller than 1 are often found (as in the present example), but they are never significant. In this case, a χ^2 of 0.915 with one degree of freedom has a corresponding P value of 0.34. This correction for small sample size has limitations, because if any of the expected numbers is too close to 0, then the correction in Equation 2.4 is unreliable.

EXACT TEST FOR HWE If the sample size is small enough, then it is possible to calculate the exact probability of all possible sample configurations. To be concrete, consider a gene with two alleles A and a, and let the observed

numbers of *AA*, *Aa*, and *aa* in one possible sample be n_{11}, n_{12}, and n_{22}, respectively. The total sample size is therefore $n = n_{11} + n_{12} + n_{22}$, and the observed numbers of *A* and *a* alleles are $n_1 = 2 \times n_{11} + n_{12}$ and $n_2 = n_{12} + 2 \times n_{22}$, respectively. We wish to calculate the probability of any sample configuration (n_{11}, n_{12}, n_{22}) for a fixed sample size n and fixed allele counts n_1 and n_2. Since the allele counts are fixed, any sample is uniquely specified by the number of heterozygotes observed. In fact, the exact probability of the sample configuration (n_{11}, n_{12}, n_{22}), conditional on the allele counts (n_1, n_2), is given by

$$\Pr\{n_{12} | n_1, n_2\} = \frac{n!/(n_{11}!n_{12}!n_{22}!)}{(2n)!/(n_1!n_2!)} 2^{n_{12}} \tag{2.5}$$

(Emigh 1980; Weir 1996; and see Guo and Thompson 1992 for a multiple-allele version). Once these conditional probabilities have been calculated for all possible values of n_{12}, they are arranged in increasing order, and a cutoff is chosen such that the cumulative probability of all outcomes above the cutoff equals 0.05 (or the number nearest to, but smaller than, 0.05). If the observed genotype counts fall below the cutoff, the hypothesis of HWE is rejected.

As an example, consider a sample of size $n = 8$ diploid individuals with fixed allele counts of $n_1 = 8$ and $n_2 = 8$. Then there are only five possible sample configurations (n_{11}, n_{12}, n_{22}), which are given below along with their probabilities calculated from Equation 2.5.

(0, 8, 0)	Pr = 0.01989
(1, 6, 1)	Pr = 0.27848
(2, 4, 2)	Pr = 0.52215
(3, 2, 3)	Pr = 0.17404
(4, 0, 4)	Pr = 0.00544

These should be arranged in increasing order of probability and the cumulative probabilities calculated, as follows:

(4, 0, 4)	Pr = 0.00544	Cumulative Prob = 0.00544
(0, 8, 0)	Pr = 0.01989	Cumulative Prob = 0.00544 + 0.01989 = 0.02533
(3, 2, 3)	Pr = 0.17404	Cumulative Prob = 0.02533 + 0.17404 = 0.19937
(1, 6, 1)	Pr = 0.27848	Cumulative Prob = 0.19937 + 0.27848 = 0.47785
(2, 4, 2)	Pr = 0.52215	Cumulative Prob = 0.47785 + 0.52215 = 1.0000

In each row, the cumulative probability value corresponds to the *P* value of observing a fit as bad (or worse) than the sample configuration given in that row. Hence, an observed sample configuration of (4, 0, 4) would lead to rejec-

tion of the hypothesis of HWE with a significance level of 0.00544, and an observed sample configuration of (0, 8, 0) would lead to rejection of the hypothesis of HWE with a significance level of 0.02533.

As another example, consider again the *CCR5* data in Problem 2.3 in which $(n_{11}, n_{12}, n_{22}) = (265, 66, 7)$, $n_1 = 596$, and $n_2 = 80$. There are exactly 41 sample configurations that are compatible with $n_1 = 596$ and $n_2 = 80$, which have the form $(n_{11}, n_{12}, n_{22}) = (298 - x, 2x, 40 - x)$, where x can assume the values 0, 1, 2, ..., 40. Each of these possible samples has a probability of occurrence given by Equation 2.5 and a deviation from HWE given by Equation 2.3. (Here we are using the chi-square only as a measure of the magnitude of the deviation, without assuming that the values are actually distributed as χ^2.) Among the 41 possibilities, 37 yield chi-square values as great or greater than the observed value, and these samples have a cumulative probability of 0.290. This is the exact P value. As we have seen, the conventional chi-square statistic in Equation 2.3 yields a P value of 0.25, and the chi-square adjusted for small sample size in Equation 2.4 yields a P value of 0.34. None of the values are statistically significant, but the example shows that Equation 2.3 yields a P value that is too small, whereas Equation 2.4 yields one that is somewhat too large. The message is that the P values from Equations 2.3 and 2.4 are best regarded as approximations whose accuracy improves with sample size.

The exact test is the most common test of significance for departures from Hardy-Weinberg in small samples, and in practice one calculates the P values using either a standard statistics package or any of a number of web-based calculators that can be accessed by searching for "exact test for Hardy-Weinberg."

PERMUTATION TEST FOR HWE In some cases it is convenient to test for HWE by comparing the sample with random permutations of the data. For example, in a sample of 8 diploid individuals in which $n_1 = 8$ and $n_2 = 8$, one would consider a large number of random permutations of (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16), where the even numbers represent one allele (A) and the odd numbers the other allele (a). Each successive pair of numbers would then constitute one diploid genotype in the sample. For example, one random permutation is (15, 12, 1, 4, 2, 16, 11, 8, 5, 13, 6, 3, 10, 7, 9, 14), which corresponds to the genotypes *aA, aA, AA, aA, aa, Aa, Aa, aA,* or $(n_{11}, n_{12}, n_{22}) = (1, 6, 1)$. For 16 elements, there are more than 10^{13} possible permutations. Each random permutation yields a possible sample configuration (n_{11}, n_{12}, n_{22}) whose chi-square can be compared with an observed value, and with a large number of random permutations, the proportion of samples whose chi-square is as large or larger than that observed approximates the P value.

To take a more complex example, consider again the *CCR5* case in Problem 2.3, where $n_1 = 596$ and $n_2 = 80$. In this case the vector to be randomly

permuted has 596 + 80 = 676 elements, (1, 2, 3, ..., 676), where the integers less than or equal to 596 represent the nonmutant *CCR5* allele and those greater than 596 represent the *CCR5Δ* allele. Again, each successive pair of integers represents a single diploid genotype in the sample. Among 1000 random permutations, in 294 cases the chi-square was as large or larger than that observed, yielding a P value of $P = 0.294$. This is in good agreement with the exact value of 0.290 calculated earlier, which is not surprising because a large sample of random permutations simulates what the exact test does exactly.

Random permutations are particularly useful when there are multiple alleles, many of them rare, because some heterozygous genotypes will be rare and some homozygous genotypes not represented in the sample. An alternative for such cases is not to compare the individual genotype frequencies with their expectations, but to compare the total number of heterozygous genotypes and homozygous genotypes with the numbers expected under HWE. (Details of HWE for multiple alleles are discussed below.) Finally, there is also an exact test that generalizes Equation 2.5 to multiple alleles (Guo and Thompson 1992).

PROBLEMS OF MULTIPLE TESTS One final issue to be addressed is that of multiple tests. The high-throughput genotyping capabilities brought about by genomics has made it possible to assay millions of polymorphisms simultaneously in large samples. For example, the human genome contains an estimated 10 million single-nucleotide polymorphisms (SNPs)—about 1 SNP every 300 base pairs—of which four million have been identified. Each SNP consists of a site in the DNA at which the particular nucleotide pair differs among chromosomes, and in which the least frequent variant (called the *minor allele*) is relatively common. The most useful SNPs are thought to be those for which the minor allele has a frequency greater than 5%. These SNPs are of great interest because they may allow systematic, genome-wide tests of association between common genetic variants and common diseases including heart disease, diabetes, autoimmunity disease, Alzheimer disease, and many others.

A pioneering study was based on comprehensive genotyping of more than a million SNPs among a sample of 269 individuals from four populations (The International HapMap Consortium 2005). This amount of data presents a challenging statistical problem. For example, if one million independent SNPs were each tested for HWE, then 50,000 SNPs would lead to rejection of the hypothesis at the $P = 0.05$ level purely due to chance. For the same reason, in disease-association tests, 50,000 SNPs would appear to be associated with each disease, for no reason other then chance variation in the samples. Furthermore, because of genetic linkage effects, nearby SNPs are not necessarily independent of one another. In such tests, the hypothesis

being tested—called the *null hypothesis* and symbolized H_0—is that there is no departure from HWE or no association between a SNP and a disease. The significance level of a test is the conditional probability that the null hypothesis is rejected when it is, in fact, true; or, in symbols, Pr{Rejection | H_0 is true}. But to reject the hypothesis that a SNP has no effect is another way of saying that the SNP *does* has an effect. Such a spurious inference is known as a false positive.

One approach to the multiple-testing problem is to make the threshold value for significance more stringent (for example, require a larger chi-square value), so that the probability of any single false positive among all the SNPs tested equals 0.05. For example, if we carried out m independent tests with a cutoff P value of 0.05, then the expected number of false positives would be $0.05 \times m$. This suggests that the appropriate cutoff should be adjusted to $P = 0.05/m$, because then the expected number of false positives is $(0.05/m) \times m = 0.05$. This is the so-called Bonferroni correction, named after the Italian statistician Carlo Emilio Bonferroni (1892–1960). But a cutoff of $P = 0.05/m$ magnifies the opposite problem. In making the null hypothesis harder to reject, only the SNPs with the largest effects are detected, and those with smaller effects escape notice along with those of no effect that are not rejected.

Another important consideration in multiple tests is the extent to which a rejection caused by a statistically significant result implies that the null hypothesis is actually false. In association tests, the probability that the null hypothesis of no association is in fact true, when the data are statistically significant and therefore cause the null hypothesis to be rejected, is called the false discovery rate. This is quite distinct from the false positive rate, as can be seen by expressing the false discovery rate as the conditional probability Pr{H_0 is true | Rejection}.

The most effective cutoffs for mutiple tests are those that strike a balance between the false positive rate [Pr{Rejection | H_0 is true}] and the false discovery rate [Pr{H_0 is true | Rejection}], because they balance finding true associations against discovering false ones. Further details are beyond the scope of this book, but can be found in Storey and Tibshirani (2003) and Verhoeven et al. (2005). Purely statistical approaches can go only so far, however. As stated in a paper by The International HapMap Consortium (2005): "Multiple replications in large samples provide the most straightforward path to identifying robust and broadly relevant associations."

Complications of Dominance

Dominance obscures a one-to-one relation between phenotype and genotype, but the allele frequencies can still be estimated if one is willing to assume HWE. For a polymorphic gene with two alleles in which one of the alleles is dominant, only two phenotypic classes can be distinguished—the dominant phenotype and the recessive phenotype. An example is found in

the Rhesus (Rh) blood groups, which are products of two closely linked genes denoted *RhD* and *RhCE* that originated as a gene duplication during primate evolution (Matassi et al. 1999). The human *RhD* allele encodes a product that elicits formation of antibodies in Rhesus monkeys, whereas a common mutant allele, *Rhd*, is a deletion. When the phenotype of individuals is assayed by means of cross-reaction with the anti-D antibody, blood cells from genotypes *RhD/RhD* and *RhD/Rhd* genotypes cross react and are said to be Rh^+ (Rh positive), whereas blood cells from *Rhd/Rhd* genotypes fail to react and are said to be Rh^- (Rh negative).

Among American Caucasians, the frequency of Rh^+ is about 85.8% and the frequency of Rh^- is about 14.2% (Roychoudhury and Nei 1988). Given only the phenotype frequencies, the data cannot be used to calculate the genotype frequencies because we have no way of knowing what proportion of Rh^+ phenotypes are *RhD/RhD* and what proportion are *RhD/Rhd*. However, if we are willing to assume random mating, then the relative proportions *RhD/RhD* and *RhD/Rhd* genotypes are given by the Hardy-Weinberg principle. Assuming random mating and HWE, the genotype frequencies are given by p^2, $2pq$, and q^2, where p is the allele frequency of *RhD* and q that of the *Rhd* deletion. An estimate of q can therefore be obtained by setting $q^2 = 0.142$ (the frequency of the homozygous recessive phenotype), and so $\hat{q} = \sqrt{0.142} = 0.3768$. More generally, if R is the frequency of homozygous recessive genotypes found in sample of n organisms, then $\hat{q}$ and its standard error are estimated as

$$\hat{q} = \sqrt{R}$$
$$SE(\hat{q}) = \sqrt{\frac{1-R}{4n}} \qquad (2.6)$$

The expression for the standard error comes from the large-sample formula that the variance of a function $f(x)$, $Var[f(x)]$, is given by $[df(x)/dx]^2 \times Var(x)$. In this case, we know the variance of R: It is the binomial variance $R(1-R)/n$, and what we desire is the variance of $f(R) = \sqrt{R}$. Since $d\sqrt{R}/dR = (\frac{1}{2}\sqrt{R})$, it follows that $Var(\sqrt{R}) = (\frac{1}{2}\sqrt{R})^2 \times R(1-R)/n = (1-R)/4n$. Taking the square root of both sides yields the standard error in Equation 2.6.

With $\hat{q}$ estimated from Equation 2.6 as 0.3768, then $\hat{p} = 1 - 0.3768 = 0.6232$, and the frequencies of *RhD/RhD*, *RhD/Rhd*, and *Rhd/Rhd* are expected to be $p^2 = (0.6232)^2 = 0.3884$, $2pq = 2(0.6232)(0.3768) = 0.4696$, and $q^2 = (0.3768)^2 = 0.1420$, respectively. The proportion of Rh^+ individuals that are actually heterozygous is therefore $0.4696/(0.4696 + 0.3884) = 54.7\%$. However, when there is dominance, there is no possibility for a χ^2 test of goodness of fit to HWE because there are 0 degrees of freedom. Because of the lack of degrees of freedom, the calculated frequencies of Rh^+ and Rh^- ($0.3884 + 0.4696 = 0.858$ and 0.142, respectively) fit the observed frequencies exactly.

PROBLEM 2.4 The Basque people, who live in the Pyrenees mountains between France and Spain, have one of the highest frequencies of the *Rhd* deletion so far reported. In one study of 400 Basques, 230 were found to be Rh^+ and 170 Rh^- (Mourant et al. 1976). Assuming HWE, estimate the frequencies of the *RhD* and *Rhd* alleles, the genotype frequencies, and the proportion of Rh^+ individuals who are heterozygous *Dd*. What is the standard error of the estimate $\hat{q}$?

ANSWER $\hat{q} = \sqrt{170/400} = 0.65$, $\hat{p} = 0.35$, and the estimated genotype frequencies of *RhD/RhD*, *RhD/Rhd*, and *Rhd/Rhd* are 0.121, 0.454, and 0.425, respectively. The proportion of *RhD/Rhd* among Rh^+ phenotypes in the Basque population is 0.454/(0.121 + 0.454) = 79%. The standard error of $\hat{q}$ equals $\sqrt{[(1-0.425)/1600]} = 0.02$.

The Hardy-Weinberg principle also finds application in studies of industrial melanism, one of the most famous and best-studied cases of evolution in action (Kettlewell 1973). **Industrial melanism** refers to the evolution of black (melanic) color patterns in several species of moths that accompanied progressive pollution of the environment by coal soot during the industrial revolution. (The various color forms of the moths are known as **morphs**.) The evolution of melanism has been observed in Great Britain, West Germany, Eastern Europe, the United States, and in other heavily industrialized areas. The species that evolve melanism are typically large night-flying moths. Of nearly 800 species of large moths in the British Isles, where industrial melanism has been most intensively studied, about 100 species are industrial melanics (Bishop and Cook 1975). The best known of these are the peppered moth (*Biston betularia*) and the scalloped hazel moth (*Gonodontis bidentata*). The inference that selection has driven the evolution of industrial melanism is strongly supported by the observation that, in both the United Kingdom and the United States, improved air quality due to emissions regulations has been accompanied by a decrease in the frequency of the melanic forms (Grant et al. 1998). In fact, the decrease in frequency is more thoroughly documented than the previous increase (Grant 1999).

The agent of selection was once universally thought to be bird predation, because the morphs are dramatically different in conspicuousness on different background (Figure 2.4). The light forms are concealed on normal bark, whereas the dark forms are concealed on bark that is blackened with soot. The problem is that, while birds are visual predators, the moths fly only at night, and when they rest during the day, they do not rest on the trunks of trees. These concerns tend to undermine classic experiments showing differential bird predation by deliberately placing the different morphs at high density on light and dark tree trunks in the daylight hours (Majerus 1998, Coyne 1998).

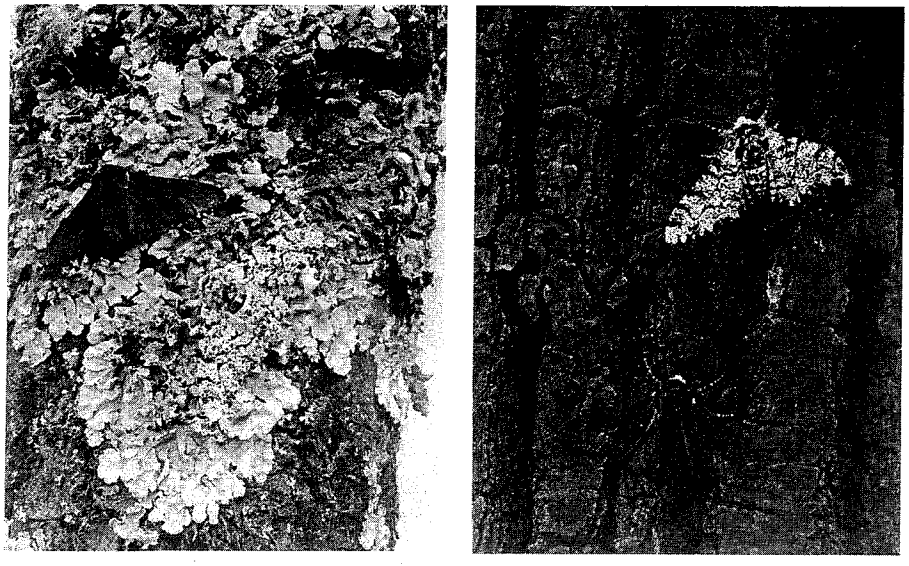


FIGURE 2.4 Melanic and nonmelanic moths, showing camouflage of light moths on light background and dark moths on dark. (Photograph by H. B. D. Kettlewell.)

But these are not the only experiments implicating differential bird predation as a selective agent (Majerus 1998; Grant 1999). On the other hand, bird predation may not be the whole story, because there are also environmental correlates of the decline in frequency of the dark morph, most notably a reduction in the level of atmospheric sulfur dioxide (Grant et al. 1998).

PROBLEM 2.5 In most cases in which the genetic basis of industrial melanism has been analyzed, the melanic color pattern has been found to be due to a single dominant allele. In one study of a heavily polluted area near Birmingham, England, Kettlewell (1956) observed a frequency of 87% melanic *Biston betularia.* Estimate the frequency of the dominant allele leading to melanism in this population and the frequency of melanics that are heterozygous.

ANSWER The observed frequency of homozygous recessives is $R = 0.13$, and so the frequency of recessive allele is estimated as $\hat{q} = \sqrt{(0.13)} = 0.36$. Assuming random mating, the expected frequencies of dominant homozygotes, heterozygotes, and recessive homozygotes are 0.41, 0.46, and 0.13, respectively. The proportion of melanics that are heterozygous is 0.46/0.87 = 52.9%.

Frequency of Heterozygotes

The Hardy-Weinberg principle also has important implications for the frequency of heterozygotes carrying rare recessive alleles. The graphs in Figure 2.5 depict the frequencies of *AA*, *Aa*, and *aa* in a population in HWE. The heterozygotes are most frequent when the allele frequencies are 0.5. Suppose that the allele *a* is a recessive, and consider the curves as the allele frequency of *a* goes toward 0. As *a* becomes rare, the frequencies of recessive homozygotes and heterozygotes both decrease, but the frequency of the recessive homozygote is much lower. As the frequency of *a* goes to 0, the frequency of recessive homozygotes goes to 0 at a rate of q^2, whereas the frequency of heterozygotes goes to 0 at a rate of $2pq$. The result is that the ratio of heterozygotes to recessive homozygotes increases without limit as the recessive allele becomes rare.

To illustrate the principle, suppose $q = 0.10$; then $2pq/q^2 = 18$, meaning that there are 18 times as many heterozygotes as recessive homozygotes. For $q = 0.01$, to take a more extreme example, the ratio is 198; and for $q = 0.001$, the ratio is 1998. These examples demonstrate that when a recessive allele is rare, most genotypes containing the rare allele are heterozygous.

More generally, the ratio of heterozygous to homozygous genotypes equals $2pq/q^2 = 2/q - 2$ which, for small q, is very close to $2/q$. Consequently, the excess of heterozygotes over homozygotes becomes progressively greater as the recessive allele becomes more rare. To take a real example, consider cystic fibrosis, an autosomal-recessive defect in chloride transport characterized by abnormal glandular secretions, impaired digestion, frequent respiratory infections, and other serious symptoms. The frequency of the homozygous recessive genotype in newborn Caucasians is approximately 1 in 1700. For this allele, $\hat{q} = \sqrt{1/1700} = 0.024$. Assuming random mating, the frequency of heterozygotes is estimated as $2(0.024)(1 - 0.024) = 0.047$, or about 1 in 21. In other words, although only 1 person in 1700 is actually affected with cystic fibrosis, 1 person in 21 is a heterozygous carrier of the harmful allele.

PROBLEM 2.6 Phenylketonuria is a defect in phenylalanine metabolism caused by lack of a functional allele encoding the enzyme phenylalanine hydroxylase. Over 200 defective alleles have been identified, and most affected individuals are actually heterozygous for two different defective alleles. The condition affects about 1 in 10,000 newborn Caucasians. Estimate the frequency of heterozygotes for the normal and a defective allele under the assumption of random mating.

ANSWER $\hat{q} = \sqrt{(1/10{,}000)} = 0.01$, and therefore $\hat{p} = 0.99$. The frequency of heterozygous genotypes (carriers) is estimated as $2\hat{p}\hat{q} = 0.0198 \approx 2\%$. Hence, about 1 person in 50 carries a defective allele.

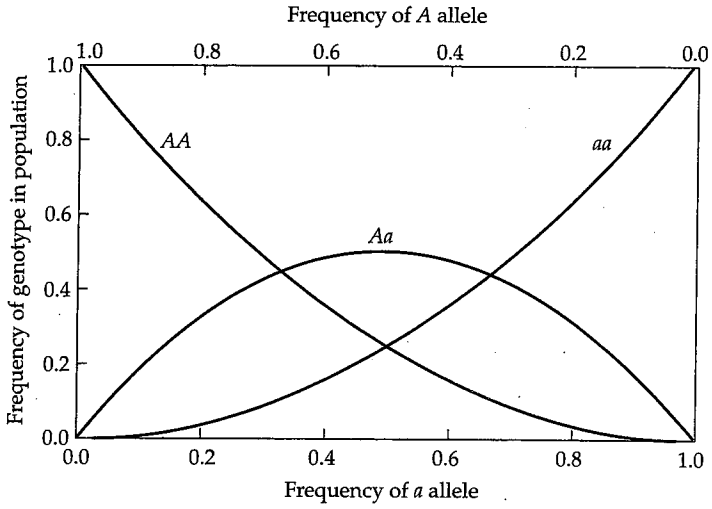


FIGURE 2.5 Frequencies of *AA*, *Aa*, and *aa* genotypes with HWE. Note that, as either allele becomes more rare, the frequency of homozygotous genotypes for that allele is much lower than the frequency of heterozygous genotypes.

2.4 EXTENSIONS OF THE HARDY-WEINBERG PRINCIPLE

In this section we extend the Hardy-Weinberg principle to multiple alleles and to genes located on the X chromosome.

Three or More Alleles

Genotype frequencies under random mating for genes with three alleles are shown in Figure 2.6. Here it is convenient to label the alleles as A_1, A_2, and A_3 and the corresponding allele frequencies as p_1, p_2, and p_3. Because there are only three alleles, $p_1 + p_2 + p_3 = 1$. With three alleles there are six diploid genotypes, and under random mating their expected frequencies are as follows:

A_1A_1	p_1^2	A_1A_2	$2p_1p_2$	A_2A_2	p_2^2
A_1A_3	$2p_1p_3$	A_2A_3	$2p_2p_3$	A_3A_3	p_3^2

These frequencies can be obtained by expanding $(p_1 A_1 + p_2 A_2 + p_3 A_3)^2$, which the cross-multiplication square in Figure 2.6 does automatically.

Application of Figure 2.6 can be illustrated with the familiar ABO blood groups in humans. These red-cell antigens are by far the most important in blood transfusions, and they are controlled by the product of a single gene in chromosome 9. More than 70 molecularly distinct alleles are known, but most of them can be assigned to one of three broad classes known as I^O, I^A, and I^B (Yip 2002). The I^A and I^B alleles encode transferase enzymes that attach

FIGURE 2.6 Punnett square showing Hardy-Weinberg frequencies for three autosomal alleles.

		Male gametes		
	Allele	A_1	A_2	A_3
	Frequency	p_1	p_2	p_3
Female gametes: Allele	Frequency			
A_1	p_1	A_1A_1 p_1^2	A_1A_2 p_1p_2	A_1A_3 p_1p_3
A_2	p_2	A_2A_1 p_2p_1	A_2A_2 p_2^2	A_2A_3 p_2p_3
A_3	p_3	A_3A_1 p_3p_1	A_3A_2 p_3p_2	A_3A_3 p_3^2

Summed frequencies in zygotes:
$A_iA_i : p_i^2$
$A_iA_j : 2p_ip_j$

different substrates to a complex carbohydrate, which is the basis of the antigenic difference between A red cells and B red cells. The diagnostic difference between the I^A and I^B classes of alleles consists of four amino acid replacements, and that distinguishing the I^O class is a single-nucleotide deletion in the initial part of the coding sequence, which shifts the translational reading frame and results in an inactive product.

Since blood type is determined by presence or absence of the A and B antigens, the genotypes I^AI^A and I^AI^O have blood type A, genotypes I^BI^B and I^BI^O have blood type B, genotype I^OI^O has blood type O, and genotype I^AI^B has blood type AB. The situation is essentially one of three alleles, complicated at the phenotypic level by the dominance of I^A and I^B over I^O. In one test for presence of A and B red-cell antigens among 6313 Caucasians in Iowa City, the counts of blood types A, B, O, and AB were 2625, 570, 2892, and 226, respectively (Mourant et al. 1976). The best estimates of allele frequency in this case are $\hat{p}_1 = 0.2593$ (for I^A), $\hat{p}_2 = 0.0625$ (for I^B), and $\hat{p}_3 = 0.6755$ (for I^O). (Estimation of allele frequencies for the ABO blood groups is complicated because of dominance and makes use of a method known as the EM algorithm; see, for example, Cavalli-Sforza and Bodmer 1971 and Vogel and

Motulsky 1986.) The expected (and observed) numbers of the four blood-type phenotypes are therefore:

A:	$[(0.2593)^2 + 2 \times 0.2593 \times 0.6755] \times 6313 = 2636.0$	(observed 2625)
B:	$[(0.0652)^2 + 2 \times 0.0652 \times 0.6755] \times 6313 = 582.9$	(observed 570)
O:	$[(0.6755)^2 \times 6313] = 2880.6$	(observed 2892)
AB:	$(2 \times 0.2593 \times 0.0652) \times 6313 = 213.5$	(observed 226)

The χ^2 for goodness of fit to Hardy-Weinberg proportions is 1.11. There is one degree of freedom for this test: 4 (to start with) – 1 (for fixing the total at 6313) – 1 (for estimating $\hat{p}_1$ from the data) – 1 (for estimating $\hat{p}_2$ from the data); note that a degree of freedom is not deducted for estimating $\hat{p}_3$ because $\hat{p}_3 = 1 - \hat{p}_1 - \hat{p}_2$. (More generally, when there are n alleles and m possible phenotypes ($m > n$), then the number of degrees of freedom in a chi-square test for HWE is $m - 1 - (n - 1) = m - n$.) For a χ^2 of 1.11 with one degree of freedom, the associated probability from Figure 2.3 is about 0.29, and so the Iowa City population gives no evidence against Hardy-Weinberg proportions for this gene.

PROBLEM 2.7 In a sample of 1617 Spanish Basques, the numbers of A, B, O, and AB blood types observed were 724, 110, 763, and 20, respectively. The best estimates of allele frequency are $\hat{p}_1 = 0.2661$ (for I^A), $\hat{p}_2 = 0.0411$ (for I^B), and $\hat{p}_3 = 0.6928$ (for I^O). Calculate the expected numbers of the four phenotypes and carry out a χ^2 test for goodness of fit to the Hardy-Weinberg expectations.

ANSWER The expected numbers of A, B, O, and AB are 710.7, 94.8, 776.1, and 35.4, respectively. The χ^2 equals 9.61 with one degree of freedom, for which the corresponding probability is 0.002. Because a deviation as large or larger than that observed would be expected by chance in only 0.002 samples (that is, about 1 in 500), there is very good reason to reject the hypothesis that the genotypes are in Hardy-Weinberg proportions in this population. The reason for the discrepancy is not known. One likely possibility is migration into the population by people with allele frequencies that are significantly different from those among the Basques themselves.

More generally, in a population that is undergoing random mating for a gene with n alleles $A_1, A_2, \ldots A_n$ having respective frequencies $p_1, p_2, \ldots p_n$ (with $p_1 + p_2 + \cdots + p_n = 1$), then the expected genotype frequencies with HWE are

$$\begin{array}{ll} p_i^2 & \text{for } A_iA_i \text{ homozygotes} \\ 2p_ip_j & \text{for } A_iA_j \text{ heterozygotes} \end{array} \qquad (2.7)$$

The expressions in Equation 2.7 may be applied to data on allozyme polymorphisms in *Drosophila persimilis* in California. One sample of 108 adult flies from the Fish Creek population included four alleles of the gene *Xdh*, which codes for xanthine dehydrogenase. We may call the alleles *Xdh-1*, *Xdh-2*, *Xdh-3*, and *Xdh-4*; their respective frequencies were estimated as $\hat{p}_1 = 0.08$, $\hat{p}_2 = 0.21$, $\hat{p}_3 = 0.62$, and $\hat{p}_4 = 0.09$ (Prakash 1977). With four alleles, there are four possible homozygotes (for example, *Xdh-1/Xdh-1*) and six possible heterozygotes (for example, *Xdh-1/Xdh-2*). In a random-mating population, the frequency of any homozygous genotype is expected to be the square of the corresponding allele frequency, for example, p_1^2 for *Xdh-1/Xdh-1*; and the frequency of any heterozygous genotype is expected to be two times the product of the corresponding allele frequencies, for example, $2p_1p_2$ for *Xdh-1/Xdh-2*. The Hardy-Weinberg frequencies for all 10 possible genotypes can be obtained by expanding the expression (0.08 *Xdh-1* + 0.21 *Xdh-2* + 0.62 *Xdh-3* + 0.09 *Xdh-4*)^2. Note that this is an example in which the expected number of some of the genotypes is small (< 1 in many cases), so a test for HWE would have to be based on the exact probabilities or on random permutations.

PROBLEM 2.8 Four alleles of the gene *Adh* coding for alcohol dehydrogenase were found in a Texas population of *Phlox cuspidata* (Levin 1978). The alleles may be designated *Adh-1*, *Adh-2*, *Adh-3*, and *Adh-4*. Their frequencies were estimated as 0.11, 0.84, 0.01, and 0.04, respectively. What are the expected Hardy-Weinberg proportions of the 10 genotypes?

ANSWER *Adh-1/Adh-1*: $0.11^2 = 0.0121$; *Adh-1/Adh-2*: $2(0.11)(0.84) = 0.1848$; *Adh-2/Adh-2* = $0.84^2 = 0.7056$; *Adh-1/Adh-3* = $2(0.11)(0.01) = 0.0022$; *Adh-2/Adh-3* = $2(0.84)(0.01) = 0.0168$; *Adh-3/Adh-3* = $0.01^2 = 0.0001$; *Adh-1/Adh-4* = $2(0.11)(0.04) = 0.0088$; *Adh-2/Adh-4* = $2(0.84)(0.04) = 0.0672$; *Adh-3/Adh-4* = $2(0.01)(0.04) = 0.0008$; *Adh-4/Adh-4* = $0.04^2 = 0.0016$. It should be pointed out that the observed genotype frequencies were nowhere near the Hardy-Weinberg expectations because *Phlox cuspidata* undergoes a substantial frequency of self-fertilization (about 78%), which violates the assumption of random mating. The question of how to deal with such departures from random mating is discussed in Chapter 6.

X-Linked Genes

An important exception to the rule that diploid organisms contain two alleles of every gene applies to genes on the X and Y chromosomes. In mammals and many insects, females have two copies of the X chromosome whereas males have one X chromosome and one Y chromosome. The X and Y chro-

mosomes segregate, and so half the sperm from a male carry the X chromosome and half carry the Y chromosome. Although the Y chromosome carries very few genes other than those involved in the determination of sex and male fertility, the X chromosome carries as full a complement of genes as any other chromosome. Genes on the X chromosome are called X-linked genes, and the important consequence of X linkage is that a recessive allele on the X chromosome in a male is expressed phenotypically because the Y chromosome lacks any compensating allele. For X-linked genes with two alleles, therefore, there are three female genotypes (*AA*, *Aa*, and *aa*) but only two male genotypes (*A* and *a*).

The consequences of random mating with two X-linked alleles are shown in Figure 2.7, where the alleles are denoted X^A and X^a. Note that in females, which have two X chromosomes, the genotype frequencies are as given by the Hardy-Weinberg principle in Equation 2.1; in males, which have only one X chromosome, the genotype frequencies are equal to the allele frequencies.

		Male gametes		
		X-bearing		Y-bearing
	Allele	X^A	X^a	
	Frequency	p	q	

Female gametes: Allele	Frequency			
X^A	p	X^AX^A p^2	X^AX^a pq	X^AY p
X^a	q	X^aX^A qp	X^aX^a q^2	X^aY q

Summed frequencies in zygotes:

Females		Males	
X^AX^A:	p^2	X^AY:	p
X^AX^a:	$2pq$	X^aY:	q
X^aX^a:	q^2		

FIGURE 2.7 Consequences of random mating with X-linked genes. Genotype frequencies in females equal the Hardy-Weinberg frequencies, and genotype frequencies in males equal the allele frequencies in gametes.

PROBLEM 2.9 Near the tip of the short arm of the X chromosome is a gene, *PBDX* (also called Xg), that encodes a blood cell glycoprotein that can be identified using an appropriate antibody (Ellis et al. 1994). One allele (call it *A*) produces sufficient gene product to be detected, whereas the other allele (call it *a*) produces too little product to be detected. Hence blood cells from females of genotype *AA* or *Aa* and from males of genotype *A* have the antigen detected by the antibody and are said to be Xg-positive, whereas blood cells from females of genotype *aa* and from males of genotype *a* are Xg-negative. In one sample of 2082 British people, 967 Xg-positive females and 667 Xg-positive males were identified, along with 102 Xg-negative females and 346 Xg-negative males (Race and Sanger 1975). The best estimates of allele frequency are $\hat{p} = 0.675$ (for *A*) and $\hat{q} = 0.325$ (*a*). Calculate the expected numbers in the four phenotypic classes, assuming random-mating proportions, and carry out a χ^2 test for goodness of fit. (The number of degrees of freedom in this case is 1: there are four degrees of freedom to start with; one must be deducted for using the observed number of males in calculating the expectations for males; one must be deducted for using the observed number of females in calculating their expectations; and one more must be deducted for estimating *p* from the data.)

ANSWER The expected numbers of Xg-positive and Xg-negative males are $0.675 \times 1013 = 683.8$ and $0.325 \times 1013 = 329.2$, respectively. The expected numbers of Xg-positive and Xg-negative females are $[0.675^2 + 2(0.675)(0.325)] \times 1069 = 956.1$ and $0.325^2 \times 1069 = 112.9$, respectively. The χ^2 equals 2.45, which, as noted above, has one degree of freedom. The associated probability is about 0.12 (see Figure 2.3), and so there is no reason to reject the hypothesis of random-mating proportions.

One of the important features of random mating for X-linked genes is that phenotypes resulting from a recessive allele will be more common in males than in females. In Problem 2.9, for example, the proportion of Xg-negative males is $346/1013 = 34\%$, whereas the proportion of Xg-negative females is only $102/1069 = 10\%$. There is always an excess of affected males because q (which equals the proportion of males with the recessive phenotype) will always be greater than q^2 (which is the proportion of females with the recessive phenotype). Indeed, the discrepancy grows larger as the recessive allele becomes more rare (Figure 2.8). For example, with the X-linked "green" type of color blindness, $q = 0.05$ in Western Europeans, and so the ratio of affected males to affected females is $q/q^2 = 1/q = 1/0.05 = 20$. In contrast, for the X-linked "red" type of color blindness, $q = 0.01$, and so in this case the ratio of affected males to affected females is $1/0.01 = 100$.

The calculations in Figure 2.7 are valid only if the allele frequencies are identical in eggs and sperm. When they differ, approximate equality of allele frequencies in the sexes is attained gradually, not in two generations as for an autosomal gene (see the earlier Problem 2.1). The reason is that males receive

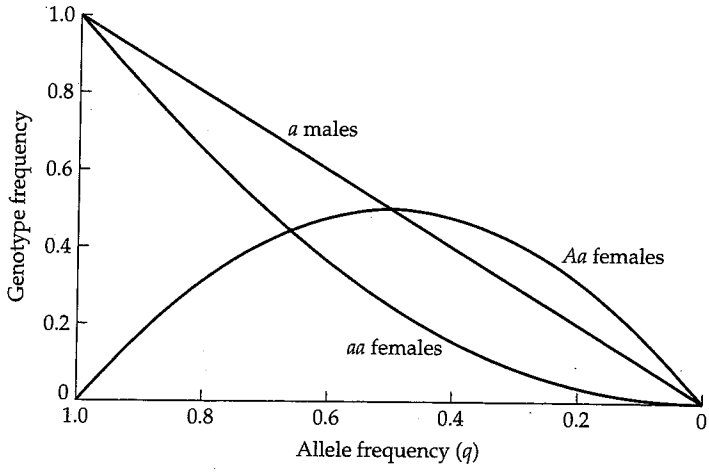


FIGURE 2.8 Random-mating frequencies for an X-linked recessive allele *a*, for males, homozygous females, and heterozygous females. When the *a* allele has a frequency of about $q = 0.1$ or smaller, the ratio of heterozygous to homozygous females is approximately $2/q$, and that of *a* males to *aa* females is approximately $1/q$.

their X chromosome only from their mothers, and so the allele frequency in males in any generation is equal to that in females of the previous generation. Because female progeny receive an X chromosome from each parent, however, the allele frequency in females is equal to the average of the sexes in the previous generation. Together, these considerations imply that the absolute value of the difference in allele frequency between the sexes decreases by 50% in each successive generation. It is necessary to specify the absolute value because the difference in allele frequency alternates in sign owing to the one-generation lag between female allele frequencies and male allele frequencies. If in any generation the allele frequency in females is larger than that in males, then in the next generation the allele frequency in females is smaller than that in males.

Before leaving the subject of X-linkage, it is necessary to point out that certain species—among them, birds, moths, and butterflies—have the sex-chromosome situation reversed. In these species, females are XY and males XX. The consequences of random mating are the same as otherwise, except that the sexes are reversed.

2.5 LINKAGE AND LINKAGE DISEQUILIBRIUM

With random mating, the alleles of any gene are combined at random into genotypes according to frequencies given by the Hardy-Weinberg propor-

FIGURE 2.9 Random association between two alleles of each of two genes, showing expected gametic frequencies when the alleles are in linkage equilibrium.

		Alleles of A gene	
	Allele	A	a
	Frequency	p_A	q_a
Alleles of B gene: Allele	Frequency		
B	p_B	AB $p_A p_B$	aB $q_a p_B$
b	q_b	Ab $p_A q_b$	ab $q_a q_b$

tions. To be specific, imagine a gene with two alleles, call them A and a, at frequencies p_A and q_a, respectively, where $p_A + q_a = 1$. Then the Hardy-Weinberg principle tells us that genotypes AA, Aa, and aa are expected in the proportions p_A^2, $2p_Aq_a$, and q_a^2, respectively, provided that mating is random.

Similarly, we may consider a different gene with alleles B and b at frequencies p_B and q_b, respectively, where $p_B + q_b = 1$. Then the Hardy-Weinberg principle tells us again that the genotype frequencies of BB, Bb, and bb are expected in the proportions p_B^2, $2p_Bq_b$, and q_b^2, respectively, provided that mating is random. Thus, the A allele is in random association with the a allele, and the B allele is in random association with the b allele. Strange as it may seem, the alleles of the A gene may nevertheless fail to be in random association with the alleles of the B gene. The precise meaning of "random association" is illustrated in Figure 2.9. In this figure the squares refer to the alleles present in gametes, not to genotypes as in earlier diagrams. When the alleles of the genes are in random association, the frequency of a gamete carrying any particular combination of alleles equals the product of the frequencies of those alleles. Genes that are in random association are said to be in a state of **linkage equilibrium**, and genes not in random association are said to be in **linkage disequilibrium**. With linkage equilibrium, therefore, the gametic frequencies are:

$$\begin{aligned} AB&: \quad p_A \times p_B \\ Ab&: \quad p_A \times q_b \\ aB&: \quad q_a \times p_B \\ ab&: \quad q_a \times q_b \end{aligned} \tag{2.8}$$

With random mating and the other simplifying assumptions listed earlier (including a large population with no mutation, migration, or selection), linkage equilibrium between genes is eventually attained. However, linkage equilibrium is attained gradually, and the rate of approach can be very slow. The slow approach to linkage equilibrium stands in contrast to the attainment of HWE with alleles of a single gene, which typically requires just one generation (when generations are nonoverlapping) or a relatively small number of generations (when generations are overlapping).

The rate of approach to linkage equilibrium depends on the rate of recombination in genotypes heterozygous for both genes. There are two types of double heterozygotes:

$$AB/ab \text{ and } Ab/aB$$

In the first case, the genotype was formed by the union of an AB gamete with an ab gamete. In the second case, the genotype was formed by the union of an Ab gamete with an aB gamete. As we shall soon see, the frequencies of these two types of doubly heterozygous genotypes are not always equal.

Consider the genotype AB/ab. The gametes produced by this genotype are of four types: (1) AB, (2) Ab, (3) aB, and (4) ab. Gametic types 1 and 2 are known as **nonrecombinant gametes** because the alleles are associated in the same manner as in the previous generation (specifically, A with B and a with b). Gametic types 3 and 4 are known as **recombinant** gametes because the alleles are associated differently than in the previous generation (specifically, A with b and a with B).

Because of Mendelian segregation, the frequency of gametic type 1 equals that of type 2, and the frequency of gametic type 3 equals that of type 4. That is, the two nonrecombinant gametes are formed in equal frequencies, and the two recombinant gametes are formed in equal frequencies. However, the overall frequency of recombinant gametes (type 3 + type 4) does not necessarily equal the overall frequency of nonrecombinant gametes (type 1 + type 2) except in special cases. The term **frequency of recombination**, usually symbolized r, refers to the proportion of recombinant gametes produced by a double heterozygote. Suppose, for example, that the genotype AB/ab produces gametes AB, ab, Ab, and aB in the proportions 0.38, 0.38, 0.12, and 0.12, respectively. Then the frequency of recombination between the genes is $r = 0.12 + 0.12 = 0.24$.

The frequency of recombination between genes depends on whether they are present in the same chromosome and, if so, on the physical distance between them. For genes in different chromosomes, the frequency of recombination is $r = 0.5$ because the four possible gametic types are produced in equal frequency. For genes in the same chromosome, the frequency of recombination depends on their distance apart, because each chromosome aligns side-by-side with its partner chromosome in meiosis and can undergo a sort of breakage and reunion resulting in an exchange of parts between the part-

ner chromosomes. The closer two genes are, the less likely that a breakage and reunion takes place in the region between the genes; the farther apart two genes are, the more likely such an event becomes. The smallest possible frequency of recombination is $r = 0$, which would imply that the two genes are so close together that a break never takes place between them. The largest possible frequency of recombination is $r = 0.5$, which is found when genes are very far apart in the same chromosome or, as noted above, when they are in different chromosomes. Genes for which the frequency of recombination is less than 0.5 must necessarily be on the same chromosome, and such genes are said to be **linked**.

To sum up, if the frequency of recombination between the A and B genes is denoted r, then the genotype AB/ab produces the following types of gametes:

AB with frequency $(1 - r)/2$

ab with frequency $(1 - r)/2$

Ab with frequency $r/2$

aB with frequency $r/2$

The situation in the Ab/aB genotype is much the same, but there is one important difference. In this case, the AB and ab gametes are the *recombinant types*, and the Ab and aB gametes are the *nonrecombinant types*. Thus, the genotype Ab/aB produces the following types of gametes:

AB with frequency $r/2$

ab with frequency $r/2$

Ab with frequency $(1 - r)/2$

aB with frequency $(1 - r)/2$

PROBLEM 2.10 Consider two linked genes that have a frequency of recombination of $r = 0.005$. (In the human genome, this represents a physical distance of about 5 kb.) What types and frequencies of gametes would be produced by an individual of genotype AB/ab? By an individual of genotype Ab/aB?

ANSWER The AB/ab genotype produces gametic types AB, ab, Ab, and aB in proportions $(1 - 0.005)/2 = 0.4975$, $(1 - 0.005)/2 = 0.4975$, $0.005/2 = 0.0025$, and $0.005/2 = 0.0025$, respectively. The Ab/aB genotype produces exactly the same gametic types, but their frequencies are 0.0025, 0.0025, 0.4975, and 0.4975, respectively. (Actually, the frequency of recombination in human females is, on the average, about 1.6 times greater than that in males.)

The frequency of recombination between genes is important in population genetics because it governs the rate of approach to linkage equilibrium. To be precise, consider a population in which the actual frequencies of the chromosome types among gametes are as follows:

$$\begin{aligned} AB&: \quad P_{AB} \\ Ab&: \quad P_{Ab} \\ aB&: \quad P_{aB} \\ ab&: \quad P_{ab} \end{aligned}$$

where $P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$. In terms of the gametic frequencies, linkage equilibrium is defined as the state in which $P_{AB} = p_A p_B$, $P_{Ab} = p_A q_b$, $P_{aB} = q_a p_B$, and $P_{ab} = q_a q_b$ (see Figure 2.9).

Suppose that the genes are not in linkage equilibrium. To determine how rapidly linkage equilibrium is approached, we need to deduce the gametic frequencies in the next generation. Consider first the AB gamete. In any one generation, a chromosome carrying AB either could have undergone recombination between the genes (an event with probability r, where r is the frequency of recombination), or could have escaped recombination between the genes (an event with probability $1 - r$). Among the AB chromosomes that did not undergo recombination, the frequency of AB is the same as it was in the previous generation (P_{AB}); and among the chromosomes that did undergo recombination, the frequency of AB chromosomes is simply the product of the frequencies of the A and B alleles in the previous generation ($p_A p_B$), because the recombination joins alleles from two independent chromosomes. Therefore, the frequency of AB in any generation, call it P_{AB}', is related to the frequency P_{AB} in the previous generation by the equation

$$\begin{aligned} P_{AB}' &= (1 - r) \times P_{AB} && \text{[for the nonrecombinants]} \\ &\quad + r \times p_A p_B && \text{[for the recombinants]} \end{aligned}$$

Subtraction of $p_A p_B$ from both sides leads to

$$P_{AB}' - p_A p_B = (1 - r)(P_{AB} - p_A p_B) \tag{2.9}$$

Equation 2.9 becomes simplified somewhat by defining D as the difference $P_{AB} - p_A p_B$. Then D_n is the value of D in the nth generation, and Equation 2.9 implies that $D_n = (1 - r)D_{n-1}$. The solution of this equation is found by successive substitution as

$$D_n = (1 - r)D_{n-1} = (1 - r)^2 D_{n-2} = \cdots = (1 - r)^n D_0 \tag{2.10}$$

where D_0 is the value of D in the founding population.

Because $1 - r < 1$, $(1 - r)^n$ goes to zero as n becomes large, but how rapidly $(1 - r)^n$ goes to zero depends on r; the closer r is to zero, the slower the rate.

This pattern of decrease is known as *geometric,* and it is closely approximated by exponential decay because $(1-r)^n \approx e^{-rn}$, for small r. The geometric decay in linkage disequilibrium is illustrated in Figure 2.10. Recall here that $r = 0.5$ corresponds either to genes far apart in the same chromosome or to genes in different chromosomes. The key point is that *linkage disequilibrium does not require genes to be physically linked.* Linkage disequilibrium can occur even for genes in different chromosomes. For example, if one population is fixed for the alleles A and B of genes in different chromosomes, and another population is fixed for the alternative alleles a and b, then were the populations to fuse, the gametes would initially consist exclusively of AB and ab. This is an extreme form of linkage disequilibrium, and it would dissipate according to Equation 2.10 with $r = 0.50$. Because linkage disequilibrium does not require physical linkage, some authors choose to call linkage disequilibrium by the term *gametic phase disequilibrium.*

Because $(1-r)^n$ goes to zero, D goes to zero, and therefore P_{AB} goes to $p_A p_B$ unless there are other offsetting processes. Analogous arguments hold for gametes containing Ab, aB, or ab, and so P_{Ab}, P_{aB}, and P_{ab} go to $p_A q_b$, $q_a p_B$, and $q_a q_b$, respectively. Thus, linkage equilibrium is attained at a rate determined by the value of r.

The value of D that holds for $P_{11} - p_1 q_1$ also holds for the other possible gametes, as follows:

$$\begin{aligned} P_{AB} &= p_A p_B + D \text{ (which implies that } D \geq -p_A p_B) \\ P_{Ab} &= p_A q_b - D \text{ (which implies that } D \leq p_A q_b) \\ P_{aB} &= q_a p_B - D \text{ (which implies that } D \leq q_a p_B) \\ P_{ab} &= q_a q_b + D \text{ (which implies that } D \geq -q_a q_b) \end{aligned} \tag{2.11}$$

The quantity D is often called the **linkage disequilibrium parameter.** The implications about the magnitude of D in parentheses in the equations above follow from the fact that each of the gametic frequencies P_{AB}, P_{Ab}, P_{aB}, and P_{ab} must be greater than or equal to 0. Because of these implications, the minimum and maximum values of D must satisfy

$$\begin{aligned} D_{min} &= \text{the larger of } -p_A p_B \text{ and } -q_a q_b \\ D_{max} &= \text{the smaller of } p_A q_b \text{ and } q_a p_B \end{aligned} \tag{2.12}$$

Furthermore, it follows from the Equations 2.11 that D can also be written as

$$D = P_{AB} P_{ab} - P_{Ab} P_{aB} \tag{2.13}$$

With random mating and no countervailing forces, the value of D changes according to Equation 2.10, and $D = 0$ corresponds to linkage equilibrium.

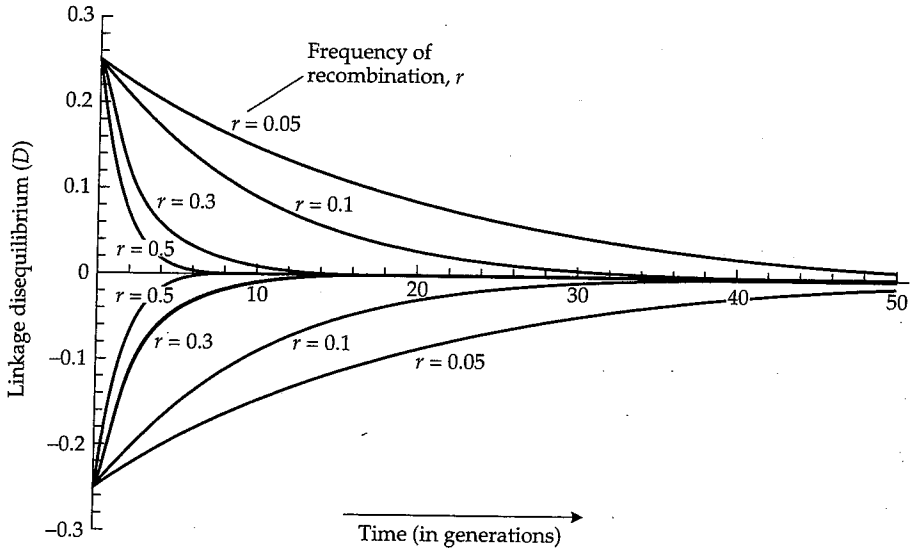


FIGURE 2.10 Linkage disequilibrium between genes gradually disappears when mating is random, provided there is no countervailing force building it up. The rate of approach to linkage equilibrium depends on the recombination frequency between the genes. The disappearance of linkage disequilibrium is gradual even with free recombination ($r = \frac{1}{2}$). In these examples, the frequencies of both alleles at both loci equal $\frac{1}{2}$, and the initial linkage disequilibrium is either at its maximum ($D = 0.25$) or minimum ($D = -0.25$) value, given these allele frequencies.

Linkage disequilibrium is often observed between closely linked single nucleotide polymorphisms or SNPs (The International HapMap Consortium 2005). As an example, consider SNPs in the coding sequence of each of two closely linked genes in human chromosome 4 that encode the proteins for glycophorin A and glycophorin B found on the surface of red blood cells. One SNP is an A to G substitution, which results in an amino acid replacement of serine to leucine (one of two amino acid replacements that distinguish the M and N forms of glycophorin A); the other SNP is a T to C substitution, which results in the amino acid replacement methionine to threonine that distinguishes the S and s forms of glycophorin B. For each SNP considered individually, the genotypes are in HWE. In particular, a sample of 1000 British people yielded genotype counts of 298 *AA*, 489 *AG*, and 213 *GG* for the SNP in glycophorin A, and 99 *TT*, 418 *TC*, and 483 *CC* for the SNP in glycophorin B. The chi-square values for goodness of fit to HWE are 0.22 and 0.38, respectively.

From these SNP data, the allele frequencies can be estimated as $\hat{p}_A = 0.5425$ and $\hat{q}_a = 0.4575$ for the *A* and *G* alleles of the SNP for glycophorin A, and $\hat{p}_B = 0.3080$ and $\hat{q}_b = 0.6920$ for the *T* and *C* alleles of the SNP for glycophorin B. There are four possible haplotypes (a **haplotype** is the combination of alleles present in a chromosome), namely *A T*, *A C*, *G T*, and *G C*, and were the SNPs in linkage equilibrium, the haplotype frequencies would be $p_A p_B$, $p_A q_b$, $q_a p_B$, and $q_a q_b$, respectively. Therefore, among the 1000 haplotypes (a total of 2000 chromosomes), the observed (*obs*) and expected (*exp*) numbers are as shown in the third column below (the second column gives the observed numbers):

AT	*obs*: 474	*exp*: $0.5425 \times 0.3080 \times 2000 = 334.2$
AC	*obs*: 611	*exp*: $0.5425 \times 0.6920 \times 2000 = 750.8$
GT	*obs*: 142	*exp*: $0.4575 \times 0.3080 \times 2000 = 281.8$
GC	*obs*: 773	*exp*: $0.4575 \times 0.6920 \times 2000 = 633.2$

The χ^2 for goodness of fit is 184.7 with one degree of freedom: 4 (to start with) – 1 – 1 (for estimating p_1 from the data) – 1 (for estimating q_1 from the data) = 1. The associated probability is very much less than 0.0001. This result means that chance alone would produce a fit as poor or poorer in substantially less than one time in ten thousand, and so the hypothesis that the loci are in linkage equilibrium can confidently be rejected.

To quantify the amount of linkage disequilibrium, we must estimate the haplotype frequencies corresponding to P_{AB}, P_{Ab}, P_{aB}, and P_{ab}:

AT: $\hat{P}_{AB} = 474/2000 = 0.2370$

AC: $\hat{P}_{Ab} = 611/2000 = 0.3055$

GT: $\hat{P}_{aB} = 142/2000 = 0.0710$

GC: $\hat{P}_{ab} = 773/2000 = 0.3865$

Thus, *D* can be estimated from Equation 2.13 as $\hat{D} = \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} = 0.07$. From Equation 2.12, D_{max} is given by $p_A q_b$ or $q_a p_B$, whichever is smaller; in this case, $p_A q_b = 0.38$ and $q_a p_B = 0.14$, hence $D_{max} = 0.14$. Therefore, $\hat{D}/D_{max} = 0.07/0.14 = 50\%$, and so we conclude that the amount of disequilibrium between the SNPs in the genes for glycophorin A and glycophorin B is about 50% of its theoretical maximum. In most local populations of sexual organisms that regularly avoid extreme inbreeding (mating between relatives) values of *D* are typically zero or close to zero (indicating linkage equilibrium) unless the genes are very closely linked.

Another widely used measure of linkage disequilibrium is related to but distinct from D. This measure is usually symbolized as r^2, which can potentially cause confusion because the symbol r is also widely used for the frequency of recombination between genes. In the context of linkage disequilibrium, the square symbol in r^2 is extremely important, because it signals a measure of linkage disequilibrium rather than a measure of recombination. The value of r^2 is defined as

$$r^2 = D^2/(p_A q_a p_B q_b) \qquad (2.14)$$

There is a nice intuitive biological interpretation of r^2 in that its square root (that is, $\sqrt{r^2}$) is the correlation coefficient in allelic state between alleles in the same gamete. The value of r^2 is also useful for calculating the χ^2 value from the counts of haplotypes, because the value of χ^2 is numerically equal to r^2N, where N is the total number of chromosomes examined. This application of r^2 is illustrated in the problems below.

PROBLEM 2.11 Natural populations of *Drosophila melanogaster* are polymorphic for coding SNPs that result in amino acid replacements in the enzymes esterase 6, esterase C, and octanol dehydrogenase. An order to avoid ambiguity, we will use the symbols *A*, *B*, and *C* to denote the majority (most frequent) nucleotide of each SNP, and the symbols *a*, *b*, and *c* to denote the minority (least frequent) nucleotide of each SNP. The *A*, *a* and *B*, *b* SNPs are rather loosely linked ($r = 0.122$), whereas the *B*, *b* and *C*, *c* SNPs are tightly linked ($r = 0.002$). The recombination fractions are those in females, as recombination does not take place in males of this species. For 489 chromosomes examined from a population in North Carolina, Mukai et al. (1974) used protein electrophoresis to identify the SNPs and found the following haplotypes:

ABC	264	*aBC*	152
ABc	13	*aBc*	7
AbC	29	*abC*	15
Abc	8	*abc*	1

Carry out a chi-square test to determine whether there is significant linkage disequilibrium between the *A*, *a* SNP and the *B*, *b* SNP.

ANSWER The observed numbers of the four haplotypes *AB*, *Ab*, *aB*, and *ab* are 277, 37, 159, and 16, respectively, and so their frequencies are $P_{AB} = 0.5665$, $P_{Ab} = 0.0757$, $P_{aB} = 0.3251$, and $P_{ab} = 0.0327$. The allele frequencies in the sample are $p_A = 0.6421$, $q_a = 0.3579$, $p_B = 0.8916$, and $q_b = 0.1084$, and the estimated value of $D = P_{AB}P_{ab} - P_{Ab}P_{aB} = -0.0061$. The r^2 from Equation 2.14 equals 0.001659, and therefore the $\chi^2 = 0.001659 \times 489 = 0.81$. This χ^2 has one degree of freedom, and the associated probability is about 0.37. Thus, there is no reason to reject the hypothesis that the *A*, *a* and *B*, *b* SNPs are in linkage equilibrium in this population.

PROBLEM 2.12 Determine whether there is significant linkage disequilibrium for the *B*, *b* and *C*, *c* SNPs using Equation 2.14 and the data in Problem 2.11.

ANSWER For the data given in Problem 2.11, the observed numbers of the haplotypes *B C*, *B c*, *b C*, and *b c* are 416, 20, 44, and 9, respectively. The estimated allele frequencies of *B*, *b*, *C*, and *c* are $p_B = 0.8916$, $q_b = 0.1084$, $p_C = 0.9407$, and $q_c = 0.0593$, respectively, and the estimated $\hat{D} = 0.0120$. Thus, $r^2 = (0.0120)^2/(0.8916 \times 0.1084 \times 0.9407 \times 0.0593) = 0.026609$. Consequently, $\chi^2 = 0.026609 \times 489 = 13.0$ with one degree of freedom, for which the associated probability is 0.0003. Thus, there is significant linkage disequilibrium between these SNPs. The value of D_{max} is the smaller of 0.053 and 0.102, and so $D_{max} = 0.053$. The magnitude of the linkage disequilibrium, relative to its theoretical maximum, is 0.012/0.053 = 22.6%

PROBLEM 2.13 Use Equation 2.14 to evaluate the statistical significance of the linkage disequilibrium between an A versus C single nucleotide polymorphism in the gene for alcohol dehydrogenase in *Drosophila melanogaster* and the presence or absence (+ or –) of an *Eco*RI restriction site located 3500 nucleotides downstream. The data are from a population descended from animals trapped at a Dutch fruit market in Groningen (Cross and Birley 1986).

SNP A, *Eco*RI +:	22
SNP A, *Eco*RI –:	3
SNP C, *Eco*RI +:	4
SNP C, *Eco*RI –:	5

ANSWER $\hat{D} = 0.085$ and $\chi^2 = r^2N = (0.453)^2 \times 34 = 7.0$ with one degree of freedom; the associated probability value is approximately 0.008. The linkage disequilibrium is statistically significant and has a value of 49% of its maximum possible value.

Difficulties in Testing for Linkage Equilibrium

All of the difficulties in testing for Hardy-Weinberg equilibrium that were discussed earlier in this chapter are multiplied when it comes to testing for linkage equilibrium. The term "multiplied" can be taken literally. The problem of too small expected numbers is multiplied because the expected frequencies of the gametes are products of the corresponding allele frequencies. This means that the large-sample chi-square test for linkage equilibrium will often be inappropriate, and an exact test (Weir 1996), analogous to the exact test for HWE in Equation 2.5, will be necessary, or else a permutation test analogous to those discussed earlier in the context of HWE.

Difficulties associated with multiple test are also multiplied when it comes to linkage equilibrium, because for n SNPs there are $n(n-1)/2$ possi-

ble pairwise associations, and $n(n-1)(n-2)/6$ possible three-way associations. In other words, the number of pairwise tests grows as n^2 and the number of three-ways tests as n^3. This emphasizes the importance of striking a proper balance between the false positive rate and the false discovery rate, so as not to miss too many true associations, but at the same time to minimize the number of false associations. As with HWE, multiple replications in large samples is probably stronger evidence of a robust association than simple statistical significance in a single sample.

All of the previous examples of estimating linkage disequilibrium are based on actual counts of the four gamete types. Often we know only the counts of genotypes. For two loci, each with two alleles, there are typically nine distinguishable genotypes: *AA BB, AA Bb, AA bb, Aa BB, Aa Bb, Aa bb, aa BB, aa Bb,* and *aa bb*. However, the *Aa Bb* double heterozygous genotype is ambiguous in regard to gametic phase because some individuals will be *A B/a b* and others *A b/a B*; these two classes will have equal frequency only if $D = 0$. To estimate the frequencies of the four gametic types in this situation, one needs to use the information from the eight unambiguous genotypes as well as the double heterozygotes to obtain estimates of the gametic frequencies that best fit the overall data. If we are willing to assume that the population is in Hardy-Weinberg equilibrium, then there is a maximum likelihood estimator for the gamete frequencies (Hill 1974). If we are not willing to assume Hardy-Weinberg equilibrium, then an estimation method is needed that takes departure from Hardy-Weinberg frequencies and linkage disequilibrium into account simultaneously (Weir 1996; Schaid 2004). Software for making these estimates can be found on the web under such names as LDhat, Haploview, and SAS/Genetics.

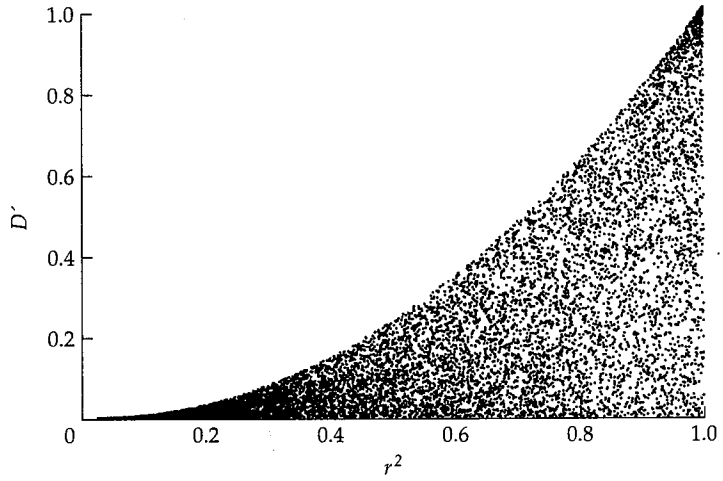
Relative Measures of Linkage Disequilibrium: D′ *and* r²

The value of D as a measure of linkage disequilibrium has the limitation that it depends on the allele frequencies, and its minimum and maximum values are defined in Equation 2.12. For this reason, the magnitude of linkage disequilibrium is often described by a quantity usually called D', which is defined as

$$D' = D/D_{max} \text{ if } D \text{ is positive}$$
$$D' = D/D_{min} \text{ if } D \text{ is negative} \qquad (2.15)$$

The use of D' as a measure of linkage disequilibrium allows comparison of relative values across genomic regions or even among organisms. Another often-used measure of linkage disequilibrium makes use of the quantity $r^2 = D^2/(p_A q_a p_B q_b)$ defined in Equation 2.14. Recall that the square root of r^2 ($\sqrt{r^2}$) is equal to the correlation coefficient in the allelic state between alleles in the same gamete.

Both D' and r^2 are used to describe the amount of linkage disequilibrium because they capture somewhat different aspects of the gametic associations.



2.11 Relationship between D' and r^2 for 10,000 random, uniformly distributed values of the gametic frequencies.

This is apparent in Figure 2.11, which shows a plot of D' against r^2 for 10,000 random, uniformly distributed values of P_{AB}, P_{Ab}, P_{aB}, and P_{ab}. It is clear that, when D' is close to 0, r^2 is also close to 0. However, as D' increases, r^2 can take on any value between 0 and $(D')^2$. The greater range of r^2 results from the fact that r^2 depends not only on D but also on the allele frequencies. The biological reason why D' and r^2 differ will become clear in the next section.

2.6 CAUSES OF LINKAGE DISQUILIBRIUM

Linkage equilibrium is a result of history. It reflects the shared ancestry of haplotypes present in any contemporary population. To understand its causes, it is convenient to consider first a region of chromosome with no recombination. Suppose that the region is initially monomorphic for nucleotides at two different sites; we will designate the nucleotides at these sites as A and B. At some time the A nucleotide may mutate to a, and either because of chance or selection, the a nucleotide increase in frequency. The haplotypes in the population will then be AB and aB. Now suppose that a mutation of B to b takes place in a chromosome carrying aB. Then the resulting population will contain three haplotypes: AB, aB, and ab. Note that the Ab haplotype is missing. Its frequency is 0, and in the absence of recombination or recurrent mutation, its frequency must remain 0. Because of the missing haplotype, the value of $D' = 1$. The value of r^2, on the other hand, depends on when the B to b mutation occurred. If it took place early in the lineage carrying the a mutation, then the correlation between a and b will be high, but if it took place

later in the genealogy, then the correlation between a and b will be low. To be concrete, consider the haplotype frequencies $P_{AB} = 0.5$, $P_{Ab} = 0$, $P_{aB} = 0.01$, and $P_{ab} = 0.49$; this represents a situation in which the b mutation arose early in the a lineage, and so most chromosomes carrying a also carry b: in this case, $r^2 = 0.96$. Contrast this example with one in which $P_{AB} = 0.5$, $P_{Ab} = 0$, $P_{aB} = 0.49$, and $P_{ab} = 0.01$; this represents a situation in which the b mutation arose late in the a lineage, and so few chromosomes carrying a also carry b: in this case, $r^2 = 0.01$. Hence, D' is a measure of linkage disequilibrium that is mainly influenced by the amount of recombination, whereas r^2 also captures information about when and where in the genealogy of the haplotypes the mutations occurred. This difference explains why D' and r^2 are complementary measures of linkage disequilibrium, and also why r^2 can assume a range of values for any given value of D' (see Figure 2.11). Both r^2 and D' have the unfortunate property that they obscure the direction of linkage disequilibrium, because the original positive or negative sign of D is lost.

Linkage Disequilibrium Due to Population Admixture

Linkage disequilibrium between closely linked genes can be caused by linkage disequilibrium in the founding population that has not yet had time to dissipate due to a low frequency of recombination (Equation 2.10). Another source of linkage disequilibrium is natural selection, provided that there is a sufficient selective advantage of some gametic types over others to overcome the natural tendency for D to go to zero.

Linkage disequilibrium can also arise as an artifact of admixture of subpopulations that differ in allele frequencies. Organisms that are subdivided into local populations are said to have **population substructure**. An example of linkage disequilibrium arising from subpopulation admixture is illustrated in Table 2.2. In this example, subpopulation 1 and subpopulation 2 are

TABLE 2.2 Linkage Disequilibrium from Admixture of Subpopulations

Chromosome	Frequency	Subpopulation 1	Subpopulation 2	Equal mixture
AB	P_{AB}	0.0025	0.9025	0.4525
Ab	P_{Ab}	0.0475	0.0475	0.0475
aB	P_{aB}	0.0475	0.0475	0.0475
ab	P_{ab}	0.9025	0.0025	0.4525
$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$	0	0	0.2025	
D_{min}	−0.0025	−0.0025	−0.2500	
D_{max}	0.0475	0.0475	0.2500	

both in linkage equilibrium for the *A*, *a* and *B*, *b* polymorphisms, which could consist of SNPs, RFLPs, or any other kind of genetic marker. Subpopulation 1 has an allele frequency of 0.05 for both *A* and *B*, and subpopulation 2 has an allele frequency of 0.95 for both *A* and *B*, and for each subpopulation $D' = 0$ and $r^2 = 0$. An equal mixture of organisms from both subpopulations has the gametic frequencies shown in the last column of Table 2.2. The allele frequencies of *A* and *B* are still 0.50 in the mixed population, but there is substantial linkage disequilibrium between the alleles, as shown at the bottom of the table. In the mixed population, D' equals $0.2025/0.2500 = 0.81$ and $r^2 = 0.66$. The sole cause of the disequilibrium is the differing allele frequencies in the subpopulations. Furthermore, the considerations in Table 2.2 make no assumption that *A* and *B* are on the same chromosome, hence linkage disequilibrium may result from population admixture even for genes on different chromosomes. If subpopulations become permanently mixed and undergo random mating, then Equation 2.10 implies that the induced linkage disequilibrium is expected to decrease at the rate r per generation, where r is the recombination fraction between the *A* and *B* genes, which for unlinked genes equals $\frac{1}{2}$.

Linkage Disequilibrium Due to Reduced Recombination

Several examples in which linkage disequilibrium is found in natural populations as a result of reduced recombination should be mentioned here. One case involves certain inversions that are polymorphic in populations of certain species of *Drosophila*, most notably *D. pseudoobscura* and *D. subobscura* and their relatives. A chromosome with an **inversion**, as the name implies, has a certain segment of its genes in reverse of the normal order. Because of the inverted segment, the process of chromosome breakage and reunion in meiosis cannot be completed in the normal manner, with the result that the alleles in the inverted segment are usually unaffected by recombination and so they remain linked together. Because inversions prevent recombination, each inversion represents a sort of "supergene," and natural selection accumulates beneficially interacting alleles within each inversion. The beneficially interacting alleles are said to show **genetic coadaptation**.

The disequilibrium caused by inversions can slowly be eroded by a process known as *gene conversion*, in which a double-stranded break in a DNA molecule, having been enlarged into a gap by nuclease activity, is repaired using the homologous DNA molecule as a template. This process allows genetic information to flow from one chromosome to another, even in a region that is inverted. Schaeffer and Anderson (2005) studied the result of gene conversion among inverted chromosomes in *D. pseudoobscura*. They found that the rate of gene conversion is approximately 3.4×10^{-6} conversion events per nucleotide per generation (two orders of magnitude greater than the nucleotide mutation rate) and that the average length of a converted tract is about 200 base pairs. Nevertheless, polymorphisms within the inversions in *D. pseudoobscura* still show linkage disequilibrium, but the magnitude is

somewhat less in the central region than within 100 kb of the inversion breakpoints. Closer to the inversion breakpoints (within about 20 kb), no gene conversion events were detected.

Another situation in which reduced recombination is important in maintaining linkage disequilibrium occurs in plants that ordinarily undergo some frequency of self-fertilization. The process of self-fertilization is an extreme form of *inbreeding* (mating between relatives), because in this case the relative is one's own self. In Chapter 6 we will explain why inbreeding reduces the frequency of heterozygous genotypes, and in the context of linkage it reduces the frequency of the double heterozygotes AB/ab and Ab/aB, which are essential for recombination to take place to reduce the magnitude of linkage disequilibrium. In extreme cases, when complete self-fertilization is practiced, there are no double heterozygotes, and each self-fertilizing line behaves as true-breeding clone that undergoes genetic divergence from other such self-fertilizing lines as mutations accumulate. Across the population as a whole, each pair of SNPs will be in complete linkage disequilibrium ($D' = 1$).

On the other hand, even a small amount of outcrossing (mating between nonrelatives) can result in a significant reduction in linkage disequilibrium, because the progeny of a cross between two homozygous lines are heterozygous for all SNPs at which the parental lines differ. In such a hybrid, every recombination event breaks up some block of SNPs that had been strongly correlated. The diminutive flowering plant *Arabidopsis thaliana*, a member of the mustard family widely used as a model organism in plant biology, serves as an example. In natural populations of *A. thaliana*, about 99% of the plants result from self-fertilization in the previous generation and only about 1% from outcrossing. Although the inbreeding is very intense, the amount of outcrossing is sufficient to reduce the average size of a **haplotype block**, a term referring to a group of linked SNPs that show significant linkage disequilibrium as measured by r^2 (Equation 2.14). Across the genome of *A. thaliana*, the length of the average haplotype block is on the order of 100 kb, and SNPs at a distance of greater than or equal to 250 kb show no greater linkage disequilibrium than SNPs in different chromosomes (Nordborg et al. 2002). Although 100 kb is small relative to the total genome size of 125 Mb, haplotype blocks of 100 kb are very large compared to those in organisms that do not undergo such extreme inbreeding. For example, the average size of haplotype blocks in the maize or *Drosophila* genome (both outcrossers) is on the order of a few kb (Tenaillon et al. 2001; Long et al. 1998), and in the dog genome and the human genome the average length is on the order of a few tens of kb (Lindblad-Toh et al. 2005; The International HapMap Consortium 2005). The length of haplotype blocks differs from one region of the genome to the next, due in part to different levels of recombination. Haplotype blocks are expected to be more nearly equal in length when adjusted for recombination frequency, and in organisms in which the recombination frequencies have been estimated with some precision, this is what has been observed (The International HapMap Consortium 2005).

SUMMARY

1. For a gene with two alleles A and a in a random-mating population, the expected genotype frequencies of AA, Aa, and aa are given by p^2, $2pq$, and q^2, respectively, where p and q are the allele frequencies of A and a, respectively, with $p + q = 1$. These expected genotype frequencies constitute the Hardy-Weinberg equilibrium (HWE).
2. One of the principal implications of the HWE is that the allele frequencies and the genotype frequencies remain constant from generation to generation, hence genetic variation is maintained.
3. The HWE has simple extensions to multiple alleles and X-linked genes.
4. Statistical tests for HWE are often based on the chi-square test for goodness of fit, but this test is compromised when the sample size is too small or multiple tests are performed. For small samples, HWE can be tested by exact tests or permutation tests.
5. When large numbers of statistical tests are performed, it is important to choose the significance criterion to achieve a suitable balance between the rate of false positives (the probability of rejection, given the null hypothesis is true) and the false discovery rate (the probability that the null hypothesis is true, given that it is rejected).
6. Linkage is the tendency for the alleles of genes that are sufficiently close together in the same chromosome to remain together in inheritance. Linkage disequilibrium refers to any nonrandom association between the alleles of different genes.
7. Linkage disequilibrium is measured by the linkage disequilibrium parameter $D = P_{AB}P_{ab} - P_{Ab}P_{aB}$, where P_{AB} represents the frequency of the gametes carrying the alleles A and B, and so forth for the other symbols.
8. Linkage disequilibrium is often reported as D', which represents the value of D as a fraction of its maximum (or minimum) possible value, or as r^2, where $\sqrt{r^2}$ is the correlation coefficient in state of alleles in the same gamete. D' and r^2 capture different aspects of the nonrandom association between alleles in gametes.
9. Linkage disequilibrium is typically found between the alleles of genes that are close together along the chromosome. It is also common between the alleles of genes prevented from undergoing recombination through a chromosomal abnormality, such as an inversion. Organisms that undergo high levels of inbreeding, as through self-fertilization, may also display significant linkage disequilibrium, owing to the reduced opportunities for recombination.
10. The alleles of genes that are not linked (even alleles of genes in different chromosomes) may show linkage disequilibrium when populations with different gametic frequencies are mixed together by population fusion of high levels of migration.

PROBLEMS

1. In certain groups of mosquitoes, including species of *Aedes* and *Culex*, sex is determined by the alleles *M* and *m* of a single autosomal gene. Genotype *Mm* is male and genotype *mm* is female. Since matings are always *Mm* × *mm*, the sex ratio is maintained at 1 : 1. How is it possible for there to be random mating for a gene in the same chromosome as the sex-determining gene, when mating for the sex-determining gene itself is completely nonrandom?

2. The table below shows the number of individuals affected, per million individuals, for each of several autosomal recessive traits. Assuming random mating proportions in the population, what is the expected frequency of individuals who are heterozygous for each of the recessive alleles?

Trait	Number of individuals affected per million individuals
(a)	2786
(b)	658
(c)	287
(d)	160
(e)	102

3. The accompanying table shows the observe numbers of *AA*, *Aa*, and *aa* genotypes in samples of size 100 from each of four populations. Calculate the chi-square value for goodness of fit to Hardy-Weinberg proportions and the associated *P* value for each sample. For which samples can the hypothesis of Hardy-Weinberg proportions be rejected?

Population	*AA*	*Aa*	*aa*
(a)	8	53	39
(b)	9	61	30
(c)	13	58	29
(d)	18	35	47

4. Charles Darwin could have discovered segregation had he known what to look for, because Mendelian segregation occurred in at least one of his own experiments. Darwin (cited in Iltis 1932) studied flower shape in the snapdragon *Antirrhinum majus*. In a cross between a true-breeding strain with regular (peloric) flowers and a true-breeding strain with irregular (normal) flowers, all of the F_1 progeny had normal flowers. Crosses of $F_1 \times F_1$ yielded 88 normal and 37 peloric plants. Perform a chi-square test assuming a 3 : 1 ratio of phenotypes in the F_2 generation. Is the peloric or normal allele dominant?

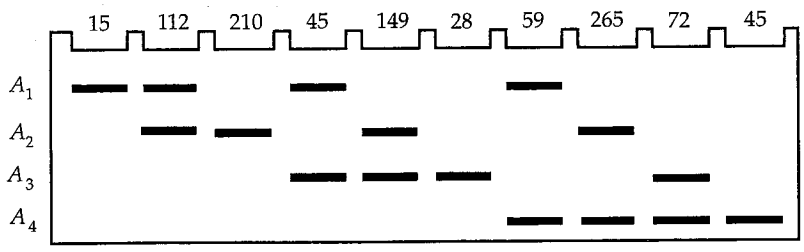
5. For a mating between genotypes that are heterozygous for the alleles of each of three unlinked genes, when one allele of each gene is dominant, there are eight phenotypic classes among the offspring. What are the

expected phenotypic ratios? Mendel carried out such an experiment and obtained the phenotypic ratio 269 : 98 : 86 : 88 : 30 : 34 : 27 : 7 among a total of 639 progeny. (He commented that this particular experiment required the most time and effort of any of his crosses.) Calculate the chi-square and associated probability.

6. When mating is random for a gene with two alleles A and a at frequencies p and q, show that the genotype frequencies of AA, Aa, and aa are approximately $1 - 2q$, $2q$, and 0 when q is so small that q^2 is approximately 0.
7. For a trait due to a rare X-linked recessive allele, show that the frequency of heterozygous carrier females is approximately equal to two times the frequency of affected males. Calculate the exact frequencies for an X-linked recessive allele with an allele frequency of 0.1.
8. What is the analogue of the Hardy-Weinberg principle for a gene with two alleles in a so-called autotetraploid, in which each chromosome is present in four copies that segregate randomly during meiosis so that each gamete receives two of the chromosomes.
9. A number of relationships between the allele frequencies of two alleles follows from the definition of p and q such that $p + q = 1$. Show that

   (a) $p^2 + pq = p$
   (b) $p - q = 1 - 2q$
   (c) $p^2 + q^2 = 1 - 2pq$
   (d) $(p - q)^2 = 1 - 4pq$

10. The accompanying gel diagram shows the banding patterns observed for an esterase enzyme in 1000 individuals of the black-tailed jackrabbit, *Lepus californicus*, in which four alleles A_1–A_4 encode polypeptide chains that differ in their electrophoretic mobility. Estimate the allele frequency for each allele, the expected number of each genotype assuming Hardy-Weinberg proportions, the chi-square value for goodness of fit, the number of degrees of freedom, and the corresponding P value. Interpret the P value in terms of whether it supports or rejects the hypothesis of Hardy-Weinberg proportions.



11. In a population undergoing random mating for a single gene with a dominant and recessive allele, show that the allele frequency of the recessive allele among individuals with the dominant phenotype is $q/(1 + q)$, where q is the allele frequency of the recessive in the whole population. Then show that the frequency of homozygous recessive genotypes from dominant × dominant matings is $[q/(1 + q)]^2$, and from dominant × recessive matings is $q/(1 + q)$. Note that the latter is equal to the square root of the former. These proportions are called *Snyder's ratios* and were once used to test human traits for simple recessive inheritance.

12. Consider a gene A with alleles A_1 and A_2 at frequencies x_1 and x_2, and a different gene B in the same population with alleles B_1, B_2, and B_3 at frequencies y_1, y_2, and y_3. What are the expected frequencies of gametes with linkage equilibrium assuming that $x_1 = 0.3$, $y_1 = 0.2$, and $y_2 = 0.3$?

13. For a gene two alleles A and a, and another gene in the same population with alleles B and b, let p_A and q_a be the allele frequencies of A and a, and let p_B and q_b be those of B and b. Set $p_A = 0.7$ and $P_B = 0.3$.
    (a) What are the expected frequencies of all possible gametes assuming linkage equilibrium?
    (b) What are the expected frequencies of all possible gametes if there is linkage disequilibrium with D equal to 50% of its theoretical maximum?

14. The accompanying table shows the estimated gametic frequencies for the alleles of two genes in a number of populations. For each population, calculate the values of D' and r^2. Which populations show the least amount of linkage disequilibrium? Which show the greatest amounts of linkage disequilibrium? Are there any that show relatively large linkage disequilibrium according to D' but not according to r^2?

Population	P_{AB}	P_{Ab}	P_{aB}	P_{ab}
(a)	0.2598	0.5362	0.0792	0.1248
(b)	0.0008	0.0196	0.0694	0.9102
(c)	0.7332	0.0082	0.0230	0.2356
(d)	0.2363	0.3029	0.2183	0.2425
(e)	0.0237	0.3460	0.5574	0.0729

15. To see how population admixture can cause linkage disequilibrium (even for genes in different chromosomes), consider the three situations in the accompanying table. Each example gives the gametic frequencies in two subpopulations, both of which are in linkage equilibrium for the alleles A, a and B, b of two genes. The gametic frequencies in the two populations are denoted *P1* and *P2* with appropriate subscripts for the alleles in the gametes. For each example, verify the absence of linkage disequilibrium in each original subpopulation. Then assume that the

two subpopulations in each example undergo admixture in equal proportions, and calculate the values of D' and r^2 in the resulting admixed population.

Example	(a)	(b)	(c)
$P1_{AA}$	0.0734	0.7220	0.0277
$P1_{Ab}$	0.6860	0.0082	0.0628
$P1_{Ab}$	0.0232	0.2667	0.2786
$P1_{ab}$	0.2174	0.0031	0.6309
$P2_{AA}$	0.4082	0.0132	0.0281
$P2_{Ab}$	0.3380	0.0621	0.2959
$P2_{Ab}$	0.1388	0.1622	0.0587
$P2_{ab}$	0.1150	0.7625	0.6173

3.1 Random Genetic Drift and Binomial Sampling 95

3.2 The Wright–Fisher Model of Random Genetic Drift 102

3.3 The Diffusion Approximation 105

An Approach Looking Forward 106

An Approach Looking Backward 110

Absorption Time and Time to Fixation 112

3.4 Random Drift in a Subdivided Population 113

3.5 Effective Population Size 121

Fluctuation in Population Size 121

Unequal Sex Ratio, Sex Chromosomes, Organelle Genes 123

Variance in Offspring Number 126

Effective Size of a Subdivided Population 127

3.6 Gene Trees and Coalescence 128

Coalescent Effective Size 133

Coalescence with Population Growth 134

Coalescent Models with Mutation 136

Applications of Coalescent Methods 137

3.7 Theoretical Implications of Coalescence 138

Coalescent Models with Recombination 140

Linkage Disequilibrium Mapping 143

CHAPTER 3

RANDOM GENETIC DRIFT

In every living organism, more gametes are formed than can possibly survive. This is one of the principle tenets of Darwin's theory of natural selection. Which particular gametes survive and which perish are determined partly by chance: the luck of the draw. The element of randomness implies that chance alone can change allele frequency from generation to generation. Because the sampling process does not change the allele frequencies in any predetermined way, this process is known as **random genetic drift**. The subtlety and importance of random genetic drift are the subject of this chapter.

3.1 RANDOM GENETIC DRIFT AND BINOMIAL SAMPLING

To introduce the process of random genetic drift, we first consider a large population at Hardy-Weinberg equilibrium with alleles A and a at equal frequencies $p = q = \frac{1}{2}$. In this population, the genotype frequencies are $\frac{1}{4}$ AA, $\frac{1}{2}Aa$ and $\frac{1}{4}$ aa. Suppose the population were to "crash," and that only four randomly chosen individuals survive to perpetuate the group. It is possible, by chance alone, that the survivors will consist of 4 AA individuals: this possibility has a probability of $(\frac{1}{4})^4 = \frac{1}{256}$. Similarly, it is possible that all four will be aa. Any other possible combination of genotypes could be realized, and it is not difficult to work out the probability for each combination. If the size of the new colony remains at just four individuals in each generation, this type of random sampling occurs each generation. In any reproductive cycle, there

is an opportunity for a possibly large change in gene frequency caused purely by the sampling process. One consequence of random drift soon becomes clear: Eventually the population will have either all A alleles or all a alleles. The reason is that, once the population reaches such a "fixation" state, it is stuck. Only new mutations or migrants into the population can reintroduce variation.

In the example above, we sampled four diploid individuals each generation. If mating takes place at random, sampling four diploid individuals is completely equivalent to sampling eight haploid gametes. When eight gametes are drawn at random from a population with $p = \frac{1}{2}$, there are nine possible outcomes, having 0, 1, 2, 3, ... 8 copies of the A allele and the remaining copies being the a allele. The probability of each of the nine possibilities is given by the **binomial distribution**, corresponding to successive terms in the expansion of $(\frac{1}{2} A + \frac{1}{2} a)^8$. The probability of fixation of the A allele in the next generation corresponds to the probability of drawing eight copies of the A allele. Since each successive draw is considered independent and has a chance of $\frac{1}{2}$ of yielding an A, this implies that the probability of drawing eight consecutive A alleles is $(\frac{1}{2})^8 = \frac{1}{256}$. The result is identical to the probability of drawing four AA genotypes calculated earlier, and it illustrates the principle that, with random mating. random sampling of diploid individuals is equivalent to random sampling of twice as many haploid gametes.

The process of sampling gametes from a finite population is depicted in Figure 3.1. The assumptions are the same as those that yield Hardy-Weinberg frequencies, but in this case the allele frequencies may change from generation to generation because of chance variation due to the finite population size. In the model in Figure 3.1, the reproducing adults in each generation comprise N diploid individuals. These individuals produce an essentially infinite pool of gametes in which the allele frequencies are the same as those in the adults. From this infinite pool of gametes, $2N$ are drawn and united at random to form the zygotes of the next generation. This model of the sam-

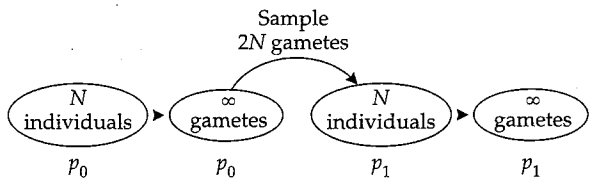


FIGURE 3.1 The gene frequencies and sampling that occur in the Wright-Fisher model. Initially there are N diploid adults with a gene whose frequency is p_0. The adults produce an infinite number of gametes having the same allele frequency. From this pool, $2N$ gametes are drawn at random to constitute the N diploid individuals for the next generation.

pling process yields a binomial distribution of all possible combinations of A and a.

To take a specific example, a population of nine diploid organisms arises from a sample of just 18 gametes, but the gametes can be thought of as being sampled from an essentially infinite pool of gametes. Because small samples are frequently not representative, an allele frequency in the sample may differ from that in the entire pool of gametes. Suppose, for example, that a pool of gametes contains the alleles A and a at frequencies p and q, respectively, with $p + q = 1$. Then if $2N$ gametes are drawn at random to produce the zygotes of the next generation, the probability that the sample contains exactly j alleles of type A is the binomial probability

$$\Pr\{j \text{ alleles of type } A\} = \binom{2N}{j} p^j q^{2N-j} = \frac{(2N)!}{j!(2N-j)!} p^j q^{2N-j} \qquad [3.1]$$

where j can take on any integer value between 0 and $2N$. The binomial coefficient (in parentheses in the middle expression) is often read as "two N choose j," because it is the number of ways that exactly j elements can be chosen from a total of $2N$. After one generation of random sampling as embodied in Equation 3.1, the new allele frequency of A in the population (call it p') is given by $j/(2N)$ because, by definition, the allele frequency of A equals the number of A alleles (in this case j) divided by the total (in this case $2N$). In the subsequent generation, the sampling process occurs anew according to Equation 3.1 with p replaced by p' and q by $1 - p'$. In this way, the allele frequency can change at random from generation to generation.

Computer-generated examples based on random sampling according to Equation 3.1 are shown in Figure 3.2. Each line in Figure 3.2A gives the number of A alleles in 20 successive generations of random genetic drift in a population of size $N = 9$ (so $2N = 18$). As you can see, individual populations behave very erratically. In seven populations, the A allele becomes **fixed** (that is, $p = 1$); in five populations, A becomes **lost** (that is, $p = 0$). In the other eight populations remain *unfixed* or **segregating** for both A and a; however, the final allele frequency among the unfixed populations is as likely to be one value as any other. Figure 3.2B shows the same kind of simulation, except now with $2N = 100$. With a larger population size, the rate at which populations go to fixation is obviously slower. The principal conclusion from Figure 3.2 is that allele frequencies behave so erratically in any one population that prediction is virtually impossible.

Although changes in allele frequency due to random genetic drift in any individual population may defy prediction, the *average* behavior of allele frequencies in a large number of populations can be predicted. Consider a large number of populations all starting at the same time with the same allele frequency and same population size N. Each of these populations is assumed to undergo drift independently of the other populations. Except for their finite

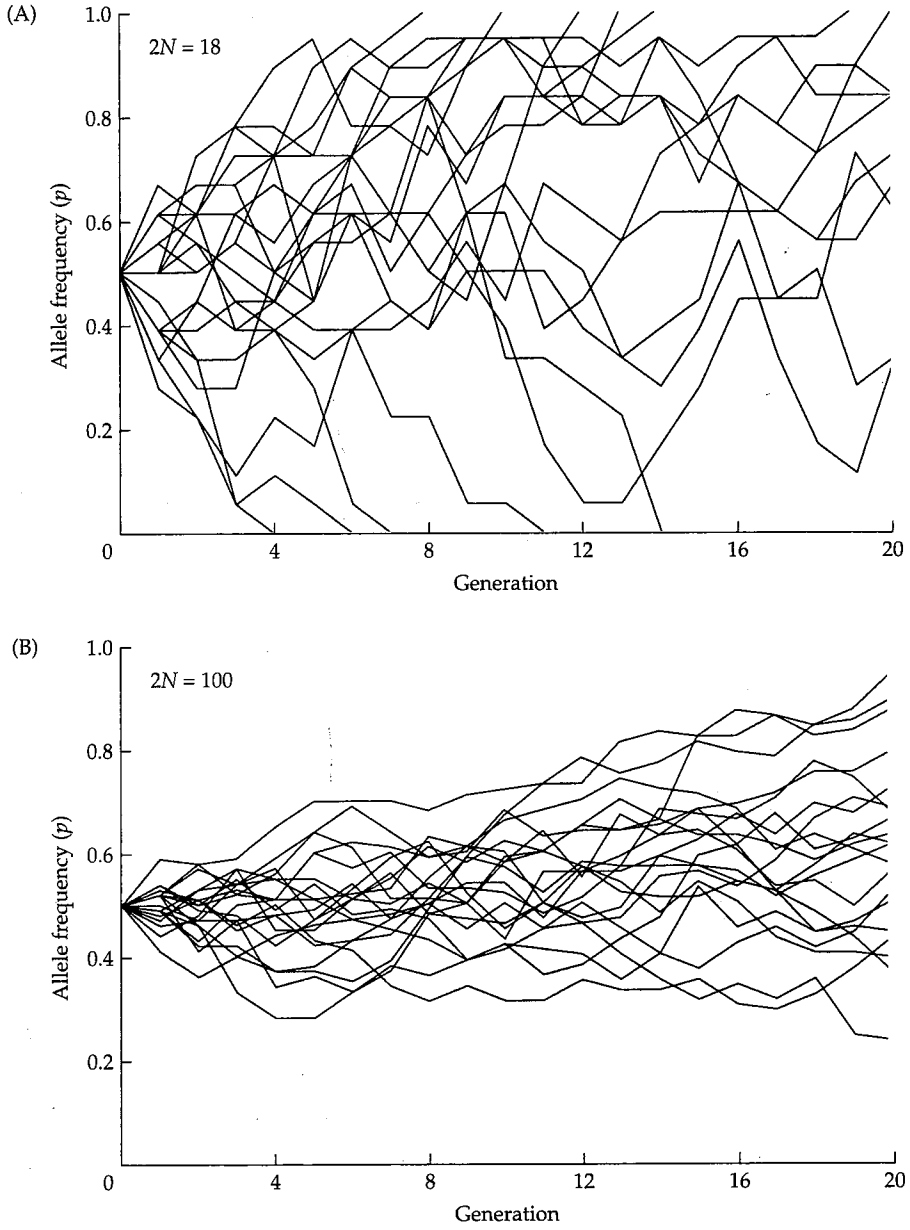


FIGURE 3.2 Computer simulations of the Wright-Fisher model of random genetic drift. Each line represents a population of size (A) $2N = 18$ or (B) $2N = 100$, simulated for 20 generations. In each generation, alleles are sampled from an infinite pool of gametes. An allele frequency of $p = 0.5$ in A implies that there are nine copies of the *A* allele, and nine copies of the *a* allele. In B, an allele frequency of 0.5 implies 50 copies of each allele. Note that the larger population size in B results in smaller fluctuations in allele frequency and a slower rate of fixation.

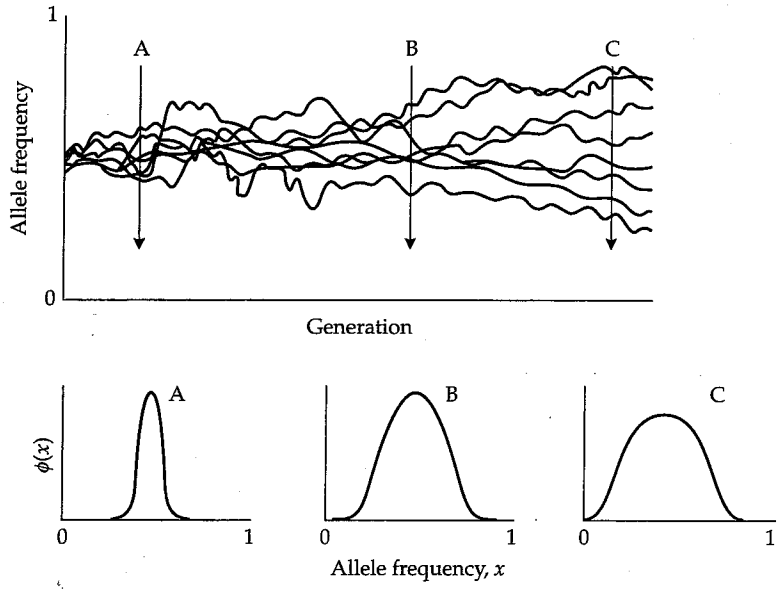


FIGURE 3.3 The implications of random genetic drift can be appreciated by imagining a large collection of subpopulations undergoing the process of repeated sampling. As the top part of the figure indicates, the allele frequency in each subpopulation changes erratically, and the allele frequencies in different subpopulations tend to drift apart. At time intervals, a snapshot of the subpopulations would produce a distribution of allele frequencies, the variance of which increases over time.

size, the subpopulations are assumed to satisfy all the assumptions of the Hardy-Weinberg model, with the additional stipulations that (1) the number of males and females is equal, and (2) each individual has an equal chance of contributing successful gametes to the next generation. The key point—illustrated in Figure 3.3—is that we can describe how these populations change in allele frequency by considering time slices through the graph and tallying a histogram of the counts of populations having each specified allele frequency. Initially, the populations will all be close to the starting allele frequency. As time passes, the populations "drift" apart, and eventually they become spread over all possible allele frequencies. Finally, as we will see, each population must go to fixation for one allele or the other.

To appreciate why ultimate fixation or loss is inevitable, consider an infinitely long bowling alley with minor imperfections that displace an imaginary weightless bowling ball one way or the other. Since the ball has no mass, it has no momentum, and hence at every instant is subject willy-nilly to the bumps and shallows of the alley. This means that, like allele frequencies, the future of the bowling ball depends only on its current position, not

on how it got there. The gutters represent the fixation states of $p = 0$ and $p = 1$. Once the ball goes into the gutter, it cannot get out again. The imperfections keep the ball from rolling in a straight line, and eventually it rolls into one or the other gutter. In this analogy, the size of the population corresponds to the width of the bowling alley; a larger population implies a wider alley. The imperfections still deflect the ball but, in proportion to the width of the alley, the ball's zigs and zags are of a smaller magnitude. Consequently, the ball remains out of the gutter for a longer time, analogous to the longer time to fixation for a larger population, but eventually the ball will end up in the gutter.

For a full understanding of random genetic drift, we must learn how to deduce the distributions of allele frequencies plotted in Figure 3.3. We just described what would happen after one generation—the set of populations would have a range of allele frequencies as described by the binomial distribution in Equation 3.1. The binomial distribution gives us the probability that a population has allele frequency p' after one generation of drift. If we consider 1000 populations all starting at p, the binomial distribution gives us the fraction of those populations with allele frequency p'. What about the following generation? For each population, one can imagine the whole sampling process as starting over again. Because no population remembers where it was the previous generation, the binomial sampling occurs anew in each generation. But because the allele frequency changes, the new allele frequency must be used in Equation 3.1. For 1000 populations, Equation 3.1 would have to be applied to each one individually, and then summed across these distributions to obtain the overall probability of each possible outcome of random drift. Fortunately, there is an easier approach that is described after we examine the following experiment.

An actual experiment designed along the lines of Figure 3.3 yielded the results shown in Figure 3.4. The graph shows the history of 19 generations of random genetic drift in 107 subpopulations of *Drosophila melanogaster*. Each subpopulation was initiated with 16 heterozygous bw^{75}/bw flies (bw = brown eyes) and maintained at a constant size of 16 individuals by randomly choosing eight males and eight females to produce the next generation. Each histogram in Figure 3.4 gives the number of subpopulations containing 0, 1, 2, ... 32 bw^{75} alleles. The pattern of change in allele frequency in Figure 3.4 may at first appear to be complicated, but in reality a simple thing is happening. The initially humped distribution of allele frequency gradually becomes flat as populations fixed for bw^{75} or bw begin to pile up at the boundaries. The piling up occurs because, once an allele has been fixed or lost, it remains fixed or lost since mutation is negligible over such a small number of generations in small populations. After 19 generations, most of the subpopulations are fixed for one allele or the other, and among the unfixed populations, the distribution of allele frequencies is essentially flat.

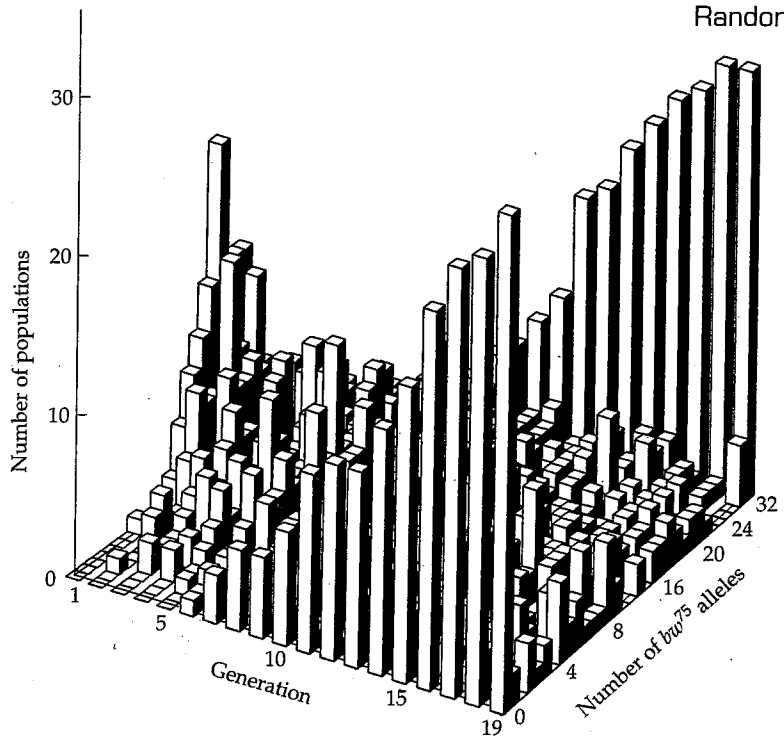


FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents of the next generation. The horizontal axis of each curve gives the number of bw^{75} alleles in the population, and the vertical axis gives the corresponding number of populations. (Data from Buri 1956.)

PROBLEM 3.1 Consider a self-pollinating plant population consisting of a single heterozygous (Aa) individual on a small barren island. Suppose the plant reproduces and dies, so that the generations are discrete and the population can only consist of a single plant. What is the probability that the population is homozygous at this genetic locus by the second generation?

ANSWER The chance that the first generation offspring is AA is $\frac{1}{4}$ and the chance that it is aa is also $\frac{1}{4}$, so the chance of fixation in one generation is $\frac{1}{2}$. If the first generation offspring is Aa, then the probability of fixation in the second generation (given that the population is not fixed in the first generation) is again $\frac{1}{2}$. The probability of not fixing in generation 1 and then fixing in generation 2 is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Add to this the chance of fixing in one generation ($\frac{1}{2}$), and we get $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ as the probability of fixation by two generations. Note that the probability of *not* going to fixation each generation is $\frac{1}{2}$, and so the chance of not fixing for two generations is $\frac{1}{2} \times \frac{1}{2}$, which equals $1 - \frac{3}{4}$.

3.2 THE WRIGHT-FISHER MODEL OF RANDOM GENETIC DRIFT

The model of random genetic drift with binomial sampling described in Equation 3.1 is known as the **Wright-Fisher model** because Fisher (1930) and Wright (1931) derived the expected distribution of allele frequencies among subpopulations. Although neither author formulated the problem in the manner used here, our approach makes the problem much simpler and gives the same results. If a population contains $2N$ alleles among which two alleles A and a may be present, then the state of the population can be described by the number of A alleles in the population. The possible states are then 0, 1, 2,... $2N$. The states 0 and $2N$ are special in that these are fixation states, and once the population get into either of these states, it cannot leave unless there is a new mutation (and for the moment we exclude this possibility). The states 0 and $2N$ are called *absorbing states*. From any nonfixed allele frequency, it is possible for the population to drift to any other allele frequency. However, the population is more likely to remain close to its present state than to take a large jump. To use an example from Figure 3.4, if $2N = 32$, then the chance of drifting from 30 copies of gene A to 29 copies in one generation is 0.186, whereas the chance of drifting to 27 copies is 0.033. The probability of the population drifting from a state having i copies to j copies of allele A is known as the *transition probability*. The transition probability for the Wright-Fisher model is obtained directly from the binomial distribution (see Equation 3.1). In particular, if a population has i copies of allele A and $2N - i$ copies of allele a, then the transition probability, T_{ij}, of going from i copies of A to j copies of A in one generation of random genetic drift is given by

$$T_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(\frac{2N-i}{2N}\right)^{2N-j} = \frac{(2N)!}{j!(2N-j)!}p^{j}q^{2N-j} \qquad (3.2)$$

where $p = i/2N$ is the initial allele frequency of A and $q = (2N - i)/2N$ is the initial allele frequency of a.

The transition probabilities can be put in a square matrix **T**, with elements T_{ij} giving the transition probability from state i to state j for $i, j = 0,1,2, \ldots 2N$. The matrix **T** contains everything that is needed to predict the expected distribution of populations like those in Figure 3.4 over a series of generations. This type of model, expressed in terms of discrete states with fixed probabilities of going from one state to another, is known as a **Markov chain**, and it has some very elegant mathematical properties. Iterations of the Wright-Fisher model give the expected outcome of a pure drift process (Figure 3.5). In a few minutes, we will use the Wright-Fisher model to show an important result regarding fixation probabilities.

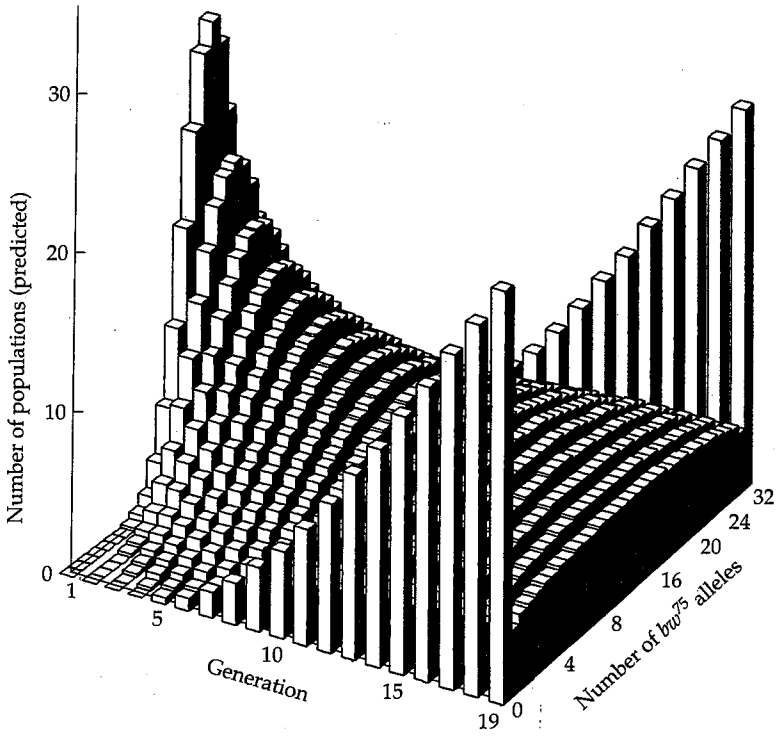


FIGURE 3.5 Prediction of the Wright-Fisher model for the distribution of allele frequencies $\phi(p, x; t)$ in subpopulations of size $N = 16$, where x represents the allele frequency in generation t. Time runs for 19 generations, and all subpopulations start with an initial allele frequency of $p = 0.5$. The values of $\phi(p, x; t)$ were generated by successive multiplication of the Markov transition probability matrix, whose entries are given by the binomial distribution in Equation 3.2. The model with $2N = 32$ predicts that fewer populations have fixed by generation 19 than actually did go to fixation in the experiment in Figure 3.4. This is because the variance in offspring number is about 70% greater than that assumed in the Wright-Fisher model.

PROBLEM 3.2 Consider a population of four diploid individuals. Calculate the probability that a population with four copies of allele A (allele frequency $p = \frac{1}{2}$) drifts in one generation to having three copies. What is the probability that the population will have four copies of A? Five copies? Now consider a population of the same size, but initially with two copies of A. What is its probability of drifting to one, two, or three copies?

ANSWER Applying Equation 3.2, we get $T_{43} = [8!/(5!3!)](\frac{1}{2})^8 = 7/32 = 0.219$. $T_{44} = [8!/(4!4!)]\,(\frac{1}{2})^8 = 70/256 = 0.273$. $T_{45} = 0.219$ ($= T_{4,3}$). (Note that the binomial distribution is symmetric when $p = \frac{1}{2}$ so there is equal probability for samples that are symmetrically divergent from $p = \frac{1}{2}$.) In the case when the initial frequency is $\frac{2}{8}$ we get $T_{21} = [8!/1!7!)]\,(\frac{1}{4})(\frac{3}{4})^7 = 0.267$, $T_{22} = [8!/(2!6!)]\,(\frac{1}{4})^2(\frac{3}{4})^6 = 0.311$, and $T_{23} = [8!/(3!5!)]\,(\frac{1}{4})^3(\frac{3}{4})^5 = 0.208$.

PROBLEM 3.3 An alternative formulation of random genetic drift is due to Moran (1958). This model has considerable intuitive appeal, and it also permits explicit expressions to be derived for various quantities of evolutionary interest (Ewens 2004). The Moran model strictly applies only to haploid populations, but to make it comparable to the Wright-Fisher model we will suppose a population of $2N$ haploid individuals. In each generation, the drift process begins by sampling two individuals at random. The sampling is carried out "with replacement," so the same individual could be chosen twice. If the two sampled individuals are different, pick one of them at random to produce a single offspring, and return the parent and the offspring to the population; discard the other individual. If the two sampled individuals are the same, then return only the offspring to the population. In the Moran model, if a population of $2N$ haploid individuals contains i of type A and $2N - i$ of type a, then the only nonzero transition probabilities are T_{ij} with $j = i - 1$, $j = i$, or $j = i + 1$. These transition probabilities are given by

$$T_{ii} = \frac{i^2 + (2N - i)^2}{(2N)^2} = p^2 + q^2$$
$$T_{ij} = \frac{i(2N - i)}{(2N)^2} = pq \text{ for } j = i + 1 \text{ or } j = i - 1 \qquad [3.3]$$

Calculate the transition probabilities in the Moran model for the examples in Problem 3.2.

ANSWER Applying Equation 3.3 to the case $p = \frac{4}{8}$, we obtain $T_{34} = 0.25$, $T_{44} = 0.50$, $T_{45} = 0.25$, and for the case $p = \frac{2}{8}$ we obtain $T_{34} = 0.1875$, $T_{44} = 0.6250$, $T_{45} = 0.1875$. Unlike the Wright-Fisher model, the transition probabilities, for either keeping the same number of alleles or else for increasing or decreasing by exactly 1, sum to 1.

Both the Wright-Fisher model and the Moran model incorporate an important feature of random genetic drift. It is that the magnitude of random change in allele frequency is greater when the allele frequency is $\frac{1}{2}$ than when the allele frequency is more skewed. The changes are greater because the variance in the sampling distribution is greatest when $p = \frac{1}{2}$. In the Wright-Fisher model, the variance in allele frequency from one generation of random genetic drift is given by $pq/(2N)$, corresponding to the variance of the proportion in a binomial distribution. The variance drops to zero at $p = 0$ and $p = 1$. In the Moran model. the variance resulting from a single birth/death event is $2pq/(2N)^2$. This looks very different from the variance in the Wright-Fisher model, however multiplying by a factor of $2N$ is needed to convert the indi-

vidual births and deaths into replacement of the entire population. There is still a factor of 2 in the numerator, which reflects the subtle fact that the variance in offspring number per individual is exactly twice as large in the Moran model as in the Wright-Fisher model (Ewens 2004). In either formulation of random drift, the variance formula makes it clear that a large population will change allele frequency more slowly than a smaller population, because the sampling variance varies as the reciprocal of population size.

PROBLEM 3.4 Simulating random drift can be a very time-consuming proposition. If one wants to simulate a population of 1000 individuals for 1000 generations, one has to draw 10^6 random numbers and for each decide whether to accept or reject each genotype. Kimura (1980b) came up with a shortcut that relates very closely to how the diffusion approximation works (see the next section). The trick is to use the recursion: $p' = p + (2U - 1)\sqrt{(3pq/2N)}$, where U is a random number uniformly distributed in the range between 0 and 1. In each generation, you pick a random number U, and then calculate the realization of the next generation's allele frequency from the above recursion. Why does this approach work? (*Hint:* The variance in a uniform distribution is the square of the range divided by 12.)

ANSWER The expression $2U - 1$, where U is a number between 0 and 1, yields a value from −1 to +1, or a range of 2. The range of $(2U - 1)\sqrt{(3pq/2N)}$ is therefore $2\sqrt{(3pq/2N)}$. Squaring this expression and dividing by 12, the variance of this uniform random variable equals $pq/2N$, which is exactly that from a binomial sampling distribution. Each generation the allele frequency has an equal chance of increasing or decreasing, and the variance in the allele frequency change is $pq/2N$. Even though the distribution of change in allele frequency is uniform in the pseudosampling simulation instead of binomial (as it is in the Wright-Fisher model), this process can reproduce most of the results of the complete brute-force simulation at a tiny fraction of the computer time. The trade-off is that one must be a little careful when near the fixation states, because the algorithm as described can yield allele frequencies less than 0 or greater than 1.

3.3 THE DIFFUSION APPROXIMATION

The pattern of change in allele frequency shown in Figure 3.4 is very nearly that expected theoretically for an ideal population, as shown in Figure 3.5. This distribution was obtained by successive multiplication of a matrix whose elements are given by the transition probabilities in Equation 3.2. Although the full-blown theory of random genetic drift requires mathematics beyond the scope of this book (see Kimura 1955, 1964, 1976; Wright 1969; Crow and Kimura 1970; Kimura and Ohta 1971; Ewens 2004), in the next section we provide an introductory tidbit to impart the flavor. If you are a student with no background in calculus, the discussion may seem quite

mysterious, but please do not be discouraged because a detailed understanding is not necessary to understand the rest of this chapter or anything later in the book.

An Approach Looking Forward

An elegant alternative to successive matrix multiplication is based on a diffusion approximation (Fisher 1922; Wright 1945; Kimura 1957, 1964). The diffusion approximation assumes that random drift disperses allele frequencies among subpopulations in a manner analogous to heat diffusing through a metal rod or tiny particles diffusing under Brownian motion (Kolmogorov 1931). The idea is to assume that the subpopulations are large enough that the allele frequencies change smoothly through time, not in large jumps. Then the statistical distribution of allele frequencies at any time is a continuous function that we may denote as $\phi(p, x; t)$, where x represents the allele frequency at time t among a large number of segregating populations ($0 < x < 1$), and p is the initial frequency among these populations. The theoretical problem is to formulate an equation that describes how $\phi(p, x; t)$ changes under random genetic drift, and to solve the equation. At any time t, the function $\phi(p, x; t)$ is a smooth, continuous function approximating the histogram of allele frequencies among the subpopulations in Figure 3.5, except that $\phi(p, x; t)$ pertains only to the unfixed subpopulations still segregating for A and a.

There are actually two approaches for obtaining a diffusion equation, each of which has advantages and limitations. One approach is to ask how the distribution $\phi(p, x; t)$ changes as we go forward in time. To explain the meaning of the equation, we will allow x and t to change only in small, discrete increments of Δx and Δt. There are two reasons why the state x could change in the time Δt. One is random genetic drift, the other is a systematic force that might include mutation or selection. We will assume that A is the favored allele, and define $M(x)$ as the probability that x increases by the amount Δx because of the systematic force. The force of random drift is measured by the probability $V(x)$ that x changes because of drift, either decreasing by the amount Δx with probability $V(x)/2$ or increasing by the amount Δx with probability $V(x)/2$. In any time interval Δt, therefore, the probability that x remains at x equals $1 - M(x) - V(x)$.

The reasoning is outlined in Table 3.1. Because changes in state are limited to $+\Delta x$ or $-\Delta x$, a subpopulation can be in state x at time $t + \Delta t$ only if it was in state $x + \Delta x$, x, or $x - \Delta x$ at time t, and these have probabilities proportional to $\phi(p, x + \Delta; t)$, $\phi(p, x; t)$, and $\phi(p, x - \Delta x; t)$, respectively. A subpopulation in state $x - \Delta x$ can change to state x with probability $M(x - \Delta x) + V(x - \Delta x)/2$ according to whether it was pushed by a systematic force (for example, mutation or selection), or else changed randomly because of random drift. A subpopulation in state $x + \Delta x$ can change to state x with probability $V(x - \Delta x)/2$ due to random drift. Finally, a subpopulation in state x can remain in state x with probability $1 - M(x) - V(x)$. The required function for $\phi(p, x; t)$

TABLE 3.1 Random Genetic Drift Looking One Generation Forward in Time

Possibilities for frequency after t generations	Probability of specified frequency after t generations	Possibilities to change to x in next interval Δt	Probability of specified change in next interval Δt
$x - \Delta x$	$\phi(p, x - \Delta x; t)$	$x - \Delta x \rightarrow x$ by systematic force	$M(p)$
	$\phi(p, x - \Delta x; t)$	$x - \Delta x \rightarrow x$ by random drift	$V(p)/2$
$x + \Delta x$	$\phi(p, x + \Delta x; t)$	$x + \Delta x \rightarrow x$ by random drift	$V(p)/2$
x	$\phi(p, x; t)$	x remains at x	$1 - M(p) - V(p)$

obtained by summing the products of columns 2 and 4 in Table 3.1, which after some simplification yields the difference equation

$$\begin{aligned}
&\phi(p,x;t+\Delta t)-\phi(p,x;t)= \\
&\quad -\left[M(x)\phi(p,x;t)-M(x-\Delta x)\phi(p,x-\Delta x;t)\right] \\
&\quad +\frac{1}{2}\left\{\left[V(x+\Delta x)\phi(p,x+\Delta x;t)-V(x)\phi(p,x;t)\right]\right. \\
&\quad \left.-\left[V(x)\phi(p,x;t)-V(x-\Delta x)\phi(p,x-\Delta x;t)\right]\right\}
\end{aligned}$$

On the left-hand side of the equal sign is the change in ϕ ($\Delta\phi$) for a given change in t (Δt). On the right-hand side, the first term is the change in $M\phi$ ($\Delta M\phi$) for a given change in x (Δx), and the second term is the change in the change in $V\phi$ ($\Delta\Delta V\phi$) for a two-step change in x ($\Delta\Delta x$). In symbols, the difference equation can be written as

$$\frac{\Delta\phi(p,x;t)}{\Delta t}=-\frac{\Delta\left[M(x)\phi(p,x;t)\right]}{\Delta x}+\frac{1}{2}\frac{\Delta\left\{\Delta\left[V(x)\phi(p,x;t)\right]\right\}}{\Delta(\Delta x)}$$

At this point we can take the limit as $\Delta t \rightarrow 0$ and $\Delta x \rightarrow 0$ (as we also overlook a number of technical details) to obtain what us called the **Kolmogorov forward equation**:

$$\frac{\partial\phi(p,x;t)}{\partial t}=-\frac{\partial\left[M(x)\phi(p,x;t)\right]}{\partial x}+\frac{1}{2}\frac{\partial^2\left[V(x)\phi(p,x;t)\right]}{\partial x^2} \qquad [3.4]$$

This is a p rtial differential equation, and given some initial function $\phi(p, x; 0)$, it can be lved (though not easily) for $\phi(p, x; t)$. We have not yet specified

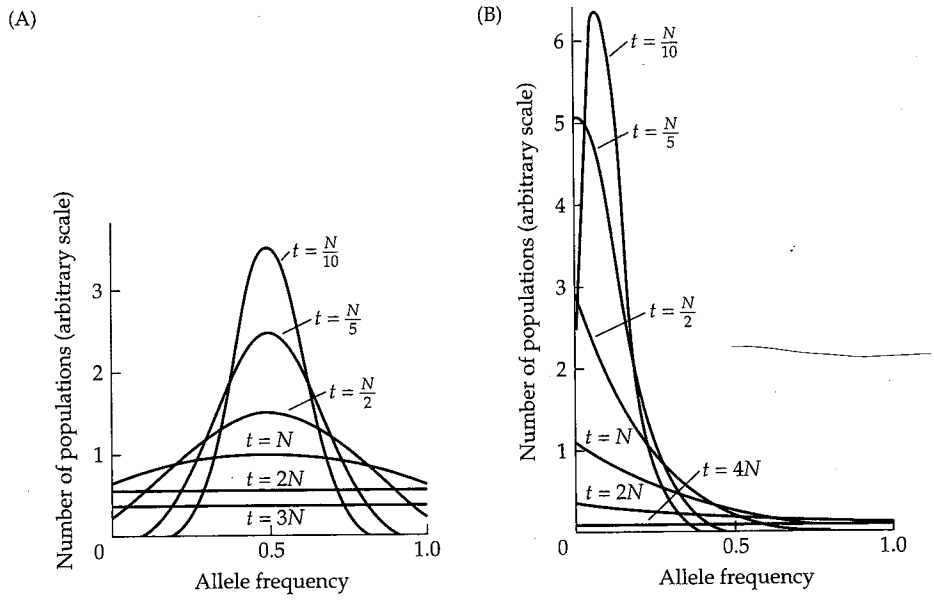


FIGURE 3.6 Theoretical results of random genetic drift. (A) Initial allele frequency = 0.5. (B) Initial allele frequency = 0.1. The curves have been scaled so that the area under each curve is equal to the proportion of populations in which fixation or loss has not yet occurred. The curves are therefore the distributions of allele frequencies among segregating populations. (From Kimura 1955.)

$M(x)$ or $V(x)$ in terms that have any relation to population genetics. The function $M(x)$ is a symbol for the change in allele frequency that occurs in one generation due to any systematic force such as mutation, migration, or selection. The function $V(x)$ also has a straightforward biological interpretation; $V(x)$ is the variance in allele frequency after one generation of binomial sampling of $2N$ alleles according to Equation 3.1; hence $V(x) = x(1-x)/(2N)$.

Many aspects of Equation 3.4 were explored by Wright (1931), and the formal solution to this equation, found by Kimura (1955), required some heavy mathematics. For our purposes, some graphs will illustrate the important properties of the forward diffusion equation. The solutions for $M(x) = 0$ are the curves plotted in Figure 3.6, which show the theoretical distributions of allele frequency among unfixed populations after various times (t) measured in units of N generations. In Figure 3.6A, all populations have an initial allele frequency of $\frac{1}{2}$, as in the actual populations in Figure 3.4; after about $t = 2N$ generations, the distribution of allele frequency is essentially flat, and by this time about half the populations are still unfixed. The distributions in Fig-

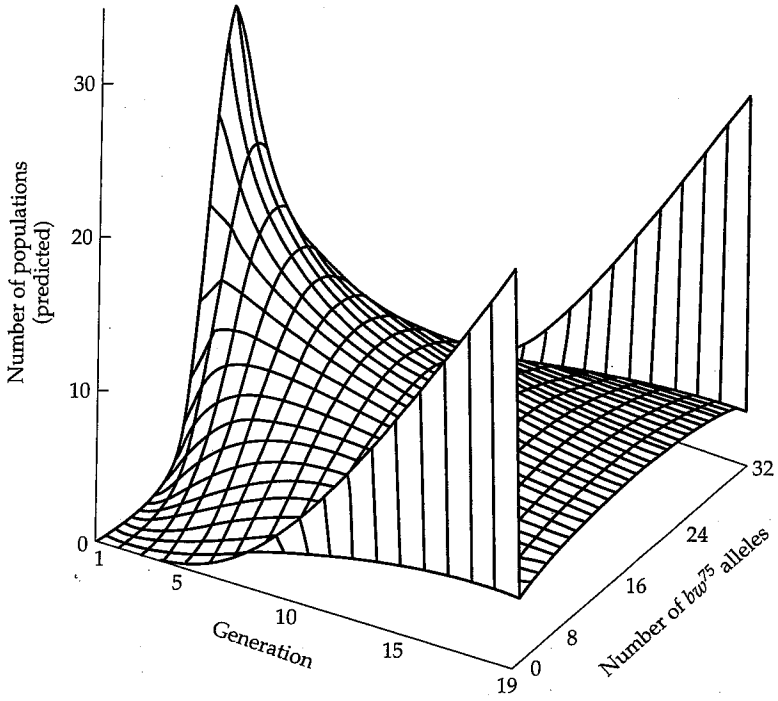


FIGURE 3.7 Kimura's (1955) solution to the diffusion equation for the particular case of $N = 16$. This is the three-dimensional view of Figure 3.6, and represents the diffusion approximation to the exact solution obtained from the Wright-Fisher model in Figure 3.5.

ure 3.6 refer only to those populations that are unfixed; as time goes on, more and more of the populations become fixed, and the distributions progressively pile up at 0 and 1, as in the histograms in Figure 3.4. Indeed, in Figure 3.6, the area under each curve is equal to the proportion of unfixed populations, which becomes progressively smaller. In particular, the rate at which the height of the distribution decreases once it becomes flat is about $1/(2N)$ per generation. To illustrate that the diffusion approximation and the Wright-Fisher model give very similar results, Figure 3.7 shows the diffusion approximation for the data in Figure 3.4, with $2N = 32$, $p_0 = \frac{1}{2}$, and t running from generation 1 through generation 19.

Figure 3.6B shows what happens when the initial allele frequency is 0.1; here the distributions are highly asymmetrical, and the distribution of allele frequency does not become flat until about $t = 4N$ generations, by which time only about 10% of the populations remain unfixed. Once a flat distribution of allele frequency is reached, the distribution remains flat, but random drift continues until fixation or loss has occurred in all populations.

An Approach Looking Backward

To find another equation for $\phi(p, x; t)$, we may also look backward in time to the beginning of the process and consider what may have happened in the very first increment of time Δt. Since the subpopulations initially all begin with an allele frequency of p, in the first time increment Δt a particular subpopulation could change its state to a frequency of $p + \Delta p$, or it could change its state to $p - \Delta p$, or it could remain at p. These possibilities have relative probabilities $M(p) + V(p)/2$, $V(p)/2$, and $1 - M(p) - V(p)$, where again $M(p)$ measures the strength of any systematic force tending to increase the allele frequency and $V(p)$ measures the variance in allele frequency due to random genetic drift.

The bookkeeping is shown in Table 3.2. If p changed state to $p + \Delta p$ in the first time increment, then the probability of the subpopulation achieving state x in the subsequent $t - \Delta t$ time units is proportional to $\phi(p + \Delta p, x; t - \Delta t)$. Similarly, going from state $p - \Delta p$ to state x in $t - \Delta t$ time units has a probability proportional to $\phi(p - \Delta p, x; t - \Delta t)$. Finally, going from state p at time Δt to state x at time t has a probability proportional to $\phi(p, x; t - \Delta t)$. The relevant equation for $\phi(p, x; t)$ is obtained by summing the products of columns 2 and 3 in Table 3.2. After some rearrangement we obtain

$$\begin{aligned}
&\phi(p,x;t) - \phi(p,x;t-\Delta t) = \\
&M(p)\left[\phi(p+\Delta p,x;t-\Delta t) - \phi(p,x;t-\Delta t)\right] \\
&+\frac{V(p)}{2}\left\{\left[\phi(p+\Delta p,x;t-\Delta t) - \phi(p,x;t-\Delta t)\right]\right. \\
&\left.-\left[\phi(p,x;t-\Delta t) - \phi(p-\Delta p,x;t-\Delta t)\right]\right\}
\end{aligned}$$

As before, the left hand side is equal to the change in ϕ ($\Delta\phi$) for a given change in t (Δt). On the right-hand side, the first term is $M(p)$ times the change in ϕ ($M\Delta\phi$) for a given change in p (Δp), and the second term is $V(p)$ times the change in the change in ϕ ($V\Delta\Delta\phi$) for a two-step change in p ($\Delta\Delta p$). In these terms, the difference equation can be written as

$$\frac{\Delta\phi(p,x;t)}{\Delta t} = M(p)\frac{\Delta\phi(p,x;t)}{\Delta p} + \frac{V(p)}{2}\frac{\Delta\left(\Delta\phi(p,x;t)\right)}{\Delta(\Delta p)}$$

Once again we will ignore some technical requirements and simply assert that, in the limit as $\Delta t \to 0$ and $\Delta p \to 0$, the difference equation converges to a partial differential equation called the **Kolmogorov backward equation**:

$$\frac{\partial\phi(p,x;t)}{\partial t} = M(p)\frac{\partial\phi(p,x;t)}{\partial p} + \frac{V(p)}{2}\frac{\partial^2\phi(p,x;t)}{\partial p^2} \tag{3.5}$$

For answering questions of population genetic interest in random drift, the Kolmogorov backward equation (see Equation 3.5) is often more useful

TABLE 3.2 Random Genetic Drift Looking Backward at the First Generation

Possibilities for change in first generation	Probability of specified change in first-generation	Probability of changing to x in remaining $t - \Delta t$ generations
$p \rightarrow p + \Delta p$ by systematic force	$M(p)$	$\phi(p + \Delta p, x; t - \Delta t)$
$p \rightarrow p + \Delta p$ by random drift	$V(p)/2$	$\phi(p + \Delta p, x; t - \Delta t)$
$p \rightarrow p - \Delta p$ by random drift	$V(p)/2$	$\phi(p - \Delta p, x; t - \Delta t)$
$p \rightarrow$ remains at p	$1 - M(p) - V(p)$	$\phi(p, x; t - \Delta t)$

than the forward equation (see Equation 3.4). The quantities of interest include the probability of ultimate fixation of an allele, the average time to fixation of alleles that are eventually fixed, and so forth. To give a sense of how the backward equation is used for these purposes, imagine the form of Equation 3.5 at a time so advanced that the distribution of allele frequencies $\phi(p, x; t)$ is no longer changing. Since random drift will continue to change the allele frequencies as long as any subpopulations are still polymorphic, the statement that $\phi(p, x; t)$ is no longer changing means that all subpopulations have become fixed for one allele or the other, which furthermore implies that the left-hand side of Equation 3.5 equals 0 and that the right-hand side no longer depends on either x (because no populations are still segregating) or t. To emphasize that we are now dealing with a function of a single variable, population geneticists often rewrite this form of Equation 3.5 as

$$0 = M(p)\frac{du(p)}{dp} + \frac{V(p)}{2}\frac{d^2u(p)}{dp^2} \tag{3.6}$$

In this equation, the symbol d is used instead of ∂ to emphasize that $u(p)$ is a function of a single variable. In words, $u(p)$ is the probability of ultimate fixation of the allele A, given an initial frequency of p. Alternatively, $u(p)$ may be interpreted as the proportion of all subpopulations in which A eventually becomes fixed. In the case of pure random drift with no systematic force, $M(p) = 0$. Equation 3.6 then becomes

$$0 = \frac{V(p)}{2}\frac{d^2u(p)}{dp^2} \tag{3.7}$$

This equation defines a family of curves, but the one of interest in population genetics has the property $u(0) = 0$, which says that an allele that does not exist cannot be fixed, and the property $u(1) = 1$, which says that an allele that is already fixed is eventually fixed.

PROBLEM 3.5 For an initial frequency of the A allele of p ($0 < p < 1$), show that $u(p) = p$ is a solution of the differential equation (see Equation 3.7).

ANSWER What needs to be shown is that Equation 3.7 is satisfied when $u(p) = p$. Although $V(p) = p(1 - p)/2N$, this is not relevant to the solution. The solution follows from the fact that, when $u(p) = p$, then $du(p)/dp = 1$ and $d^2u(p)/dp^2 = 0$. Hence $u(p) = p$ is a solution of Equation 3.7 so long as $V(p) \neq 0$. The biological meaning of $u(p) = p$ is that, because of random genetic drift, an allele present in a population at frequency p will ultimately be fixed with probability p and lost with probability $1 - p$, provided it has no effects on the organisms's ability to survive and reproduce (such alleles are often called *selectively neutral alleles*).

Absorption Time and Time to Fixation

For a selectively neutral allele, as indicated in Problem 3.5, the probability of ultimate fixation is equal to its initial allele frequency. Many other important results also follow from an analysis of the Kolmogorov backward equation (see Equation 3.5). These include the expected time for a neutral allele to go to fixation (given that it is eventually fixed) or to loss (given that it is eventually lost). Assuming an initial allele frequency p, Kimura and Ohta (1969) showed that the mean time [$\bar{t}_1(p)$, in generations] until the allele is fixed (given that it is eventually fixed) is

$$\bar{t}_1(p) = -4N\left(\frac{1-p}{p}\right)\ln(1-p) \tag{3.8}$$

Similarly, they showed that the mean time to loss $\bar{t}_0(p)$ (given that the allele is eventually lost) is

$$\bar{t}_0(p) = -4N\left(\frac{p}{1-p}\right)\ln(p) \tag{3.9}$$

Combining Equations 3.8 and 3.9, the mean persistence time of an allele [$\bar{t}(p)$, the average length of time that a population is segregating for A and a] is given by $\bar{t}(p) = p\bar{t}_1(p) + (1-p)\,\bar{t}_0(p)$, which equals

$$\bar{t}(p) = -4N\left[(1-p)\ln(1-p) + p\ln(p)\right] \tag{3.10}$$

Figure 3.8 shows the average times to fixation, loss, and persistence of a neutral allele. An allele is expected to remain in a population for the longest time when its initial frequency is $\frac{1}{2}$. When $p = \frac{1}{2}$, the average time that a population remains unfixed is about $2.77N$ generations.

Equations 3.8 and 3.9 are of particular interest when $p = 1/(2N)$, that is, when a new neutral mutation has just occurred and there is only one copy in

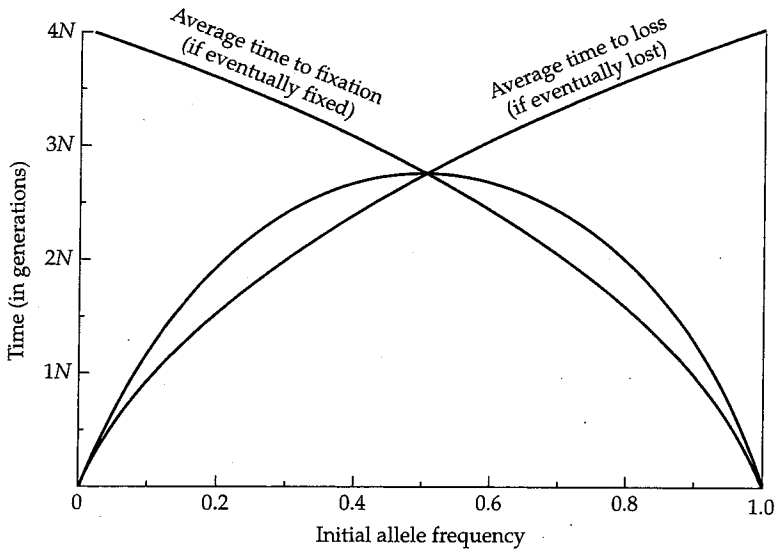


FIGURE 3.8 Average persistence of a neutral allele in an ideal diploid population of size N, plotted against initial allele frequency.

the population. In this case, the probability of eventual fixation is $1/(2N)$, and, given that the allele is eventually fixed, the average time to fixation is approximately $4N$ generations. On the other hand, the probability that a new neutral mutation is eventually lost is $1 - 1/(2N)$, and, given that the allele is eventually lost, the average time to loss is approximately $2\ln(2N)$ generations. In other words, new neutral alleles that are eventually fixed usually take a long time to be fixed, whereas those that are lost are lost very quickly. For the specific example of $N = 500$, the average new neutral mutation that is eventually fixed requires 2000 generations to be fixed, whereas the average new neutral mutation that is destined to be lost requires fewer than 14 generations to be lost.

3.4 RANDOM DRIFT IN A SUBDIVIDED POPULATION

Most real populations are subdivided into smaller units, for example, humans are concentrated in cities, towns, and villages; animals form herds, flocks, or schools; and plants are aggregated into stands. This kind of subdivision is reminiscent of the population structure in Figure 3.5, except that, in nature, the subpopulations are not genetically isolated from one another owing to some *migration*, or movement, of individuals among the subpopulations, which results in **gene flow**, or exchange of genes, between them.

Nevertheless, random genetic drift will tend to cause subpopulations to undergo differentiation in their allele frequencies, even in the face of some gene flow. To see this point, consider the four subpopulations diagrammed in Figure 3.9. Each begins with an allele frequency of $p = \frac{1}{2}$, and each undergoes random drift independently following binomial sampling (see Equation 3.2).

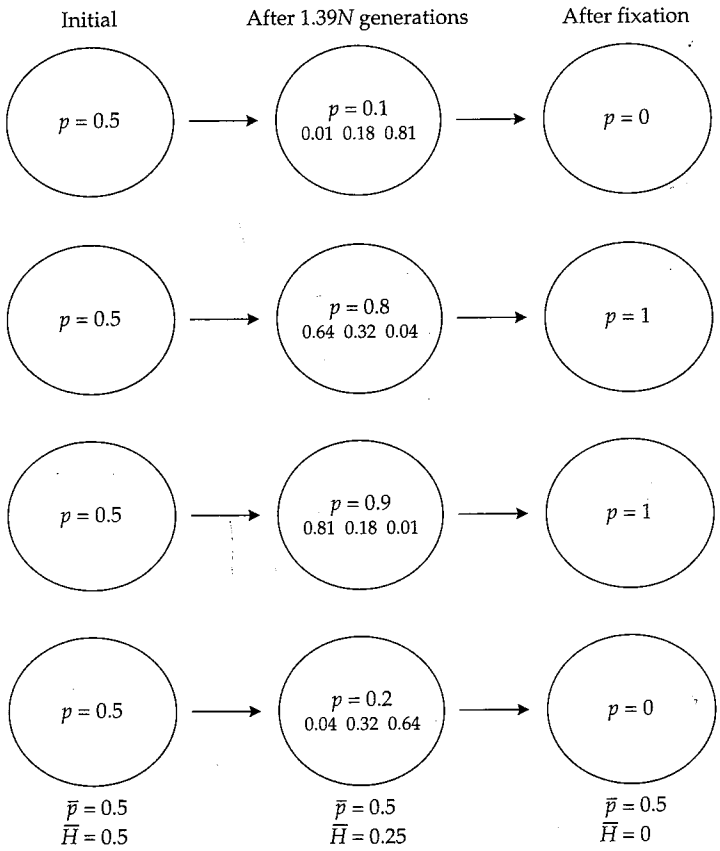


FIGURE 3.9 Schematic showing a set of four subpopulations undergoing the process of random genetic drift. Initially the allele frequency is 0.5 in all four subpopulations, and the average heterozygosity is also 0.5. As the subpopulations drift in allele frequency, the average allele frequency is expected to remain the same (indicated by $\bar{p}$ remaining at the value 0.5), but the average heterozygosity decreases. For the intermediate generation when $t = 1.39N$ generations, the allele and genotype frequencies in each subpopulation are given, as well as the average allele frequency and heterozygosity across subpopulations. By this time the average heterozygosity is reduced to 50% of the value expected without population subdivision. Ultimately, all subpopulations go to fixation, half fix one allele and half fix the other, so the average allele frequency is still 0.5, whereas the heterozygosity is zero.

We assume that random mating takes place within any particular subpopulation (call it subpopulation number i). Therefore, if the allele frequencies of A and a in the ith subpopulation are denoted p_i and q_i, then the genotype frequencies of AA, Aa, and aa are given by the familiar Hardy-Weinberg principle as p_i^2, $2p_iq_i$ and q_i^2. Furthermore, picture the situation in Figure 3.9 at a time so advanced that all subpopulations are fixed for one allele or the other. Within the ith subpopulation, therefore, either p_i equals 0 or else p_i equals 1. The genotype frequencies of AA, Aa, and aa in that subpopulation are either 0, 0, and 1 (if $p_i = 0$), or 1, 0, and 0 (if $p_i = 1$). These genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. Thus, within any one subpopulation in Figure 3.9, the frequency of heterozygotes is that expected with random mating.

The total population in Figure 3.9 is composed of the aggregate of the four subpopulations, and in the total population there is a deficiency of heterozygous genotypes. Suppose that we were unaware of the subpopulation structure and sampled from the total population as if it were a single randomly mating population. If we were to sample randomly from the far right of Figure 3.9, when no populations are still segregating, we would obtain an allele frequency of $p = \frac{1}{2}$. Assuming Hardy-Weinberg equilibrium, we would naively expect a fraction $2pq = \frac{1}{2}$ of the genotypes to be heterozygous. In fact, we would have sampled no heterozygous genotypes at all! This rather paradoxical result—that there is a deficiency of heterozygotes in the total population even though random mating occurs within each subpopulation—is a consequence of the random genetic drift of allele frequencies among subpopulations due to their finite size. The extreme case when each subpopulation is fixed is easy to understand: A population with allele frequency $\frac{1}{2}$ could only be made up of two subpopulations fixed for A and two subpopulations fixed for a; the average allele frequency is $\frac{1}{2}$, but the total population has no heterozygotes.

We are now in a position to quantify the manner in which subpopulations diverge in allele frequency under random genetic drift. To do this efficiently, we need to introduce a concept known as allele identity by descent. Two alleles are **identical by descent** if they are replicas (by DNA replication) of a gene present in some previous generation. This definition does not speak for itself, because if one goes far enough backward in time, every pair of alleles must be identical by descent, and so the concept may seem vacuous. The way out of this trap is to choose some arbitrary time in the past, which may be recent or remote according to the application, and declare that at this time every allele is distinct from every other allele. In this fashion, any earlier identity by descent is erased, and therefore the identity by descent spoken of in the definition is common ancestry through DNA replication since that arbitrary time in the past when every allele was declared distinct.

The concept of identity by descent is useful because it allows us to distinguish two types of homozygous genotypes. In particular, the A alleles in a

homozygous AA genotype could be alleles that are not identical by descent (which means that these alleles both existed in the population at the time when every allele was declared distinct), or they could be identical by descent (which means that they originated by DNA replication of a single A allele since that time). In some cases alleles may be indistinguishable by means of an experimental procedure (for example, protein electrophoresis), but their status in regard to identity by descent is unknown. Such alleles are said to be *identical by kind* or *identical by state*.

The probability that the alleles in an individual are identical by descent is often denoted F, following Wright (1922) who called it the **fixation index**. In the context of population subdivision, F as used in this chapter is the same quantity that in Chapter 6 will be denoted F_{ST}. In this chapter we will drop the subscript because we will want to track changes in F_{ST} through time, and in this case the probability of allele identity by descent in generation t is conveniently represented as F_t.

Now we can be more specific about what we mean by saying that one can choose some arbitrary time in the past and declare that at this time every allele is distinct. In the context of population subdivision as illustrated in Figure 3.9, the time in the past when the alleles are declared as distinct is in the initial populations, when the population subdivision first takes place, and all subpopulations have the same allele frequencies. In symbols, we declare that, at time $t = 0$ when the subpopulations are first established, $F_t = 0$. As time goes on, and each subpopulation undergoes random drift, the genotype frequencies in each subpopulation will satisfy the Hardy-Weinberg principle because mating within subpopulations is random. However, the allele frequencies among the subpopulations will change because of random genetic drift, and moreover the value of F_t will gradually increase as more and more alleles within any subpopulation become identical by descent owing to common ancestry.

The rate of increase in F_t can be calculated with the aid of the diagram in Figure 3.10. This figure shows the $2N$ alleles in a breeding population of generation $t - 1$. In sampling alleles for generation t, the first allele chosen may be any of those present in generation $t - 1$ with equal probability. Having chosen the first allele, the probability that the second allele chosen is of the same type as the first is $1/(2N)$ (in which case $F = 1$), because this is the frequency of each allelic type in the pool of gamates; the probability that the second chosen allele is of a different type from the first is accordingly $1 - 1/(2N)$ (in which case $F = F_{t-1}$). Putting these two possibilities together, the relationship between F_t and F_{t-1} is seen to be

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1} \tag{3.11}$$

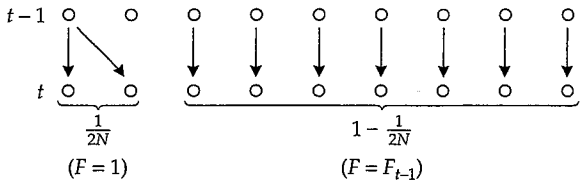


FIGURE 3.10 Diagram illustrating the reasoning behind the recursion for F in a finite population. When the gametes are drawn to make up the population at generation t, there is a chance $1/(2N)$ that any pair of alleles will have been identical in generation $t-1$. If this happens, the probability of identity is 1. For the allele pairs drawn in generation t from two distinct alleles at generation $t-1$ [the probability of this happening is $1-1/(2N)$], the probability of identity is F_{t-1}. Adding the probabilities of these two events, we get $F_t = 1/(2N) + [1-1/(2N)]F_{t-1}$.

Multiplying both sides by –1 and then adding 1 to each side leads to

$$1-F_t = 1-\frac{1}{2N}-\left(1-\frac{1}{2N}\right)F_{t-1} = \left(1-\frac{1}{2N}\right)\left(1-F_{t-1}\right)$$

and so

$$1-F_t = \left(1-\frac{1}{2N}\right)^t\left(1-F_0\right) \tag{3.12}$$

or, when $F_0 = 0$,

$$F_t = 1-\left(1-\frac{1}{2N}\right)^t \tag{3.13}$$

Figure 3.11 shows the rapid increase of F_t in small populations. Even though the genotype frequencies in each individual subpopulation are in Hardy-Weinberg proportions, the frequency of homozygous genotypes in the overall population steadily increases. Conversely, as the frequency of homozygous genotypes increases, the frequency of heterozygous genotypes decreases until, when $F_t = 1$, there are no heterozygous genotypes left and all subpopulations are fixed for either A or a. At any time, the average frequency of heterozygous genotypes among the subpopulations, H_t, relative to what it would be without population subdivision, H_0, decreases linearly in F_t, hence we have $H_t/H_0 = 1-F_t$, or $H_t = (1-F_t)H_0$. Solving Equation 3.13 for $1-F_t$ and substituting, we obtain

$$H_t = \left(1-\frac{1}{2N}\right)^t H_0 \approx H_0 e^{-t/2N} \tag{3.14}$$

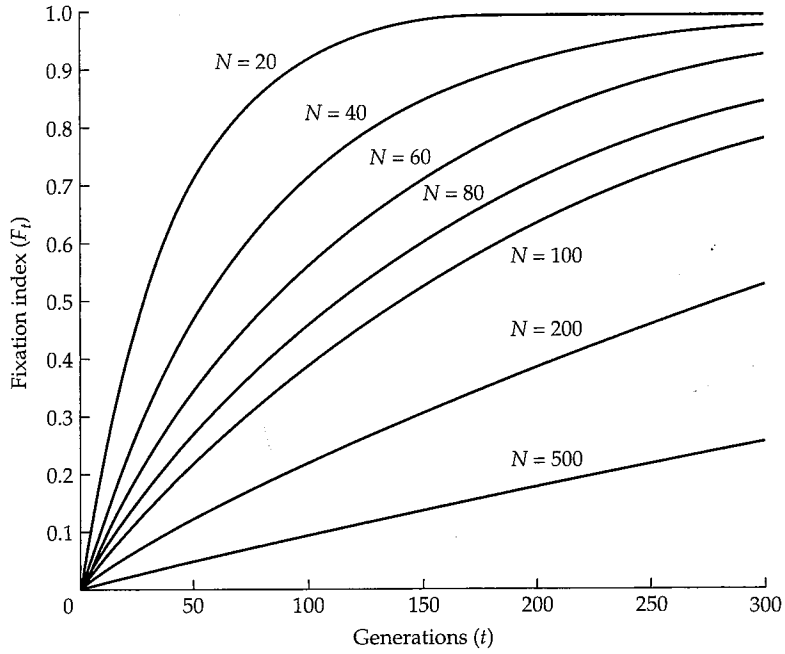
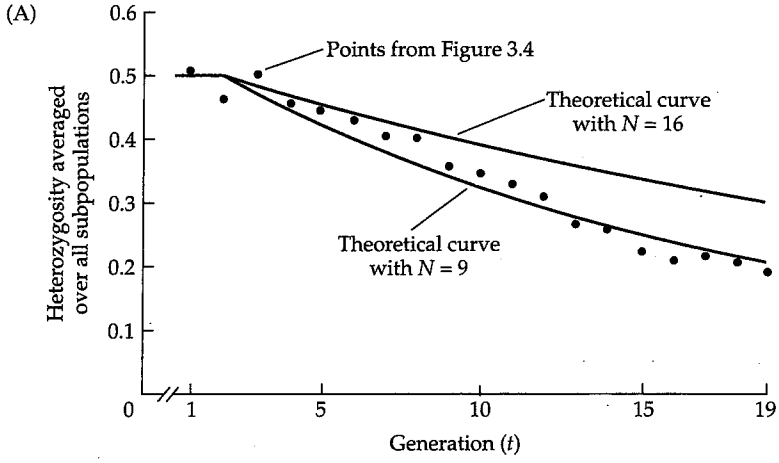


FIGURE 3.11 Increase of F_t in ideal populations as a function of time and effective population size N.

We emphasize again that each individual subpopulation undergoes random drift and remains in approximate Hardy-Weinberg proportions, and that the symbol H_t represents a sort of "virtual heterozygosity" in which the frequency of heterozygous genotypes is averaged across many subpopulations. Equation 3.14 shows that pure random drift should result in the heterozygosity decreasing at a geometric rate, since H_t is multiplied by the constant $[1 - 1/(2N)]$ each generation. Experimental tests of this prediction are shown in Figure 3.12. Figure 3.12A shows how the heterozygosity averaged across the subpopulations in Figure 3.4 declines over generations, but the theoretical curve when $N = 16$ does not fit the data very well. In fact, the rate of decline of heterozygosity is greater than the theoretical expectation, as though the population size were smaller than $N = 16$. In other words, the populations in Figure 3.4 decrease in heterozygosity as if each had a population size of $N = 9$ rather than its actual size of $N = 16$. We call $N = 9$ the *effective size* of the subpopulations, as distinct from the actual size (see Section 3.5). The theory also predicts that the allele frequency, averaged across populations, is not expected to change, and the data agree with this aspect of the theory quite well (Figure 3.12B).



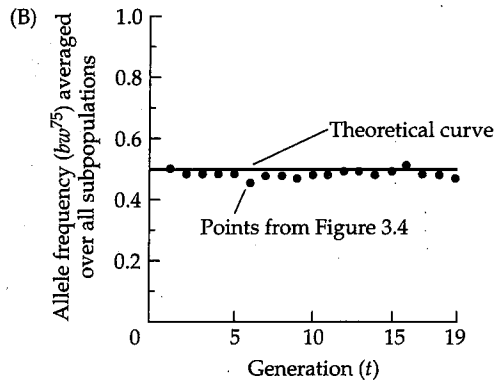


FIGURE 3.12 Theoretical curves for average heterozygosity among subpopulations (A) with $N = 9$ or $N = 16$, along with actual values (plotted as points) from the experiment in Figure 3.4. Part (B) shows the theoretically expected average allele frequency among the 107 subpopulations and the observed average. (Data from Buri 1956.)

PROBLEM 3.6 Use Equation 3.14 to determine the average length of time it would take for a finite population of size N to reduce its initial heterozygosity by a factor of two.

ANSWER Set $H_t = \frac{1}{2} H_0 = H_0 e^{-(t/2N)}$. Now divide both sides by H_0 and take the natural logarithm (base e) to obtain $\ln(\frac{1}{2}) = -t/(2N)$, or $t = -2N \ln(\frac{1}{2}) = 1.39N$ generations. In words, this result says that it requires an average of $1.39N$ generations to halve the heterozygosity, whatever its initial value. Fisher (1918) showed that it also takes $1.39N$ generations to halve what he called the *genic variance* in the population. Since the variance of a binomial sample is $pq/2N$, and the average heterozygosity in a population decreases in proportion to the variance in allele frequency among subpopulations, it follows that the average heterozygosity decreases at the same rate as the variance in allele frequency among subpopulations increases.

Several important consequences of the population structure in Figure 3.9 can now be summarized. First, although each subpopulation is finite in size, we can imagine so many of them that the size of the total population is effectively infinite. For an infinite population, the allele frequencies must remain constant. That is, even though the allele frequency in any individual subpopulation may change willy-nilly due to random genetic drift, the overall average allele frequency of A among subpopulations remains p_0, where p_0 represents the allele frequency of A in the initial populations. Figure 3.12B shows an experimental demonstration of the constancy of average allele frequency. Since F_t is the probability of identity by descent of the two alleles in an individual in generation t, the probability that the two alleles in an individual in generation t are not identical by descent is $1 - F_t$. Because p_0 is the overall allele frequency of A, averaged across all subpopulations, the probability that a randomly chosen individual will be genotypically AA is $p_0^2(1 - F_t)$ [for the case of nonidentity by descent] + p_0F_t [for the case of identity by descent], which equals $p_0^2(1 - F_t) + p_0F_t$. Similarly, the probability that the individual will be Aa equals $2p_0q_0(1 - F_t)$, and likewise the probability that the individual will be aa equals $p_0^2(1 - F_t) + q_0F_t$. To summarize, the average genotype frequencies among subpopulations at any time t have the expected values:

$$AA: \; p_0^2(1 - F_t) + p_0F_t \quad = p_0^2 + p_0q_0F_t \tag{3.15a}$$

$$Aa: \; 2p_0q_0(1 - F_t) \quad = 2p_0q_0 - 2p_0q_0F_t \tag{3.15b}$$

$$aa: \; q_0^2(1 - F_t) + q_0F_t \quad = q_0^2 + p_0q_0F_t \tag{3.15c}$$

where $q_0 = 1 - p_0$ is the average frequency of a, averaged across all subpopulations.

Note that, while each individual subpopulation maintains Hardy-Weinberg frequencies, the average genotypic frequencies in the total population are different because there is an excess of homozygotes and a deficiency of heterozygotes. Equation 3.13 implies that the average heterozygosity among subpopulations at time t equals $2p_0q_0(1 - F_t) = 2p_0q_0[1 - 1/(2N)]^t$, and this is the theoretical curve plotted in Figure 3.12A (with $p_0 = q_0 = \frac{1}{2}$). Additionally, the comment about the variance in the answer to Problem 3.6 can be stated in symbols by saying that, at any time t, the expected variance in allele frequencies among the subpopulations equals $2p_0q_0F_t$.

Since F_t eventually goes to 1, all subpopulations eventually become fixed for one allele or the other (see Equations 3.15). Because the average allele frequency of A remains p_0 even when all subpopulations have become fixed, the proportion of subpopulations that eventually become fixed for A must be p_0, and the proportion that eventually become fixed for a must be q_0. Stated another way, the probability of ultimate fixation of an allele in any ideal subpopulation is equal to the frequency of that allele in the initial population.

This point follows from the diffusion approximation (see Problem 3.5) and is illustrated in the experiment in Figure 3.4, where $p_0 = \frac{1}{2}$; in this case, by generation 19, a total of 58 populations have become fixed, among which 30 are fixed for the *bw* allele and 28 fixed for the bw^{75} allele.

3.5 EFFECTIVE POPULATION SIZE

As we saw in the *Drosophila* experiments in Figure 3.12, populations generally fluctuate in allele frequency by an amount greater than $pq/(2N)$. No real population can be expected to satisfy the assumptions of a theoretically ideal population in all respects. Hence, in any actual case, there must be corrections for such complications as fluctuations in population size, unequal numbers of males and females, skewed distributions in family size, population structure, and so forth (Crow and Kimura 1970; Ewens 2004). The effects of these complicating circumstances on the change in allele frequencies and rates of allele fixation can be approximated by calculating the *effective size* of the population and using this value in the theory for an ideal population. That is, the **effective population size** of an actual population is the number of individuals in a theoretically ideal population having the same magnitude of random genetic drift as the actual population. There are three kinds of effective population size based on how we choose to measure "magnitude," namely: (1) the change in probability of identity by descent (F), (2) the change in variance in allele frequency, or (3) the rate of loss of heterozygosity. These are called the *inbreeding effective size*, the *variance effective size*, and the *eigenvalue effective size*, respectively.

Wright (1931) first worked out the effective population size by considering the increase in identity by descent in various situations. As noted, the effective population size can also be calculated by determining the rate of change in variance in allele frequency among subpopulations, and Kimura and Crow (1963) first applied this approach to the problem of overlapping generations. Usually, the inbreeding effective size and the variance effective size are the same, but exceptions do occur. Similarly, the variance effective size and the eigenvalue effective size can be distinct (Ewens 1982, 2004). Some of the various factors that require calculation of an effective population size will now be illustrated. We will focus on the inbreeding effective size because this concept is the most widely used.

Fluctuation in Population Size

Correction for fluctuating population size is important because natural populations actually do change in size, sometimes by a factor of 10 or more in a single generation. For the sake of simplicity, assume that the population is ideal in all respects except that its size is not constant. We will consider the situation over just two generations. Suppose that the population sizes in two

successive generations are N_0 and N_1. The arguments laid out in Figure 3.10 imply that

$$1-F_2=\left(1-\frac{1}{2N_1}\right)(1-F_1) \tag{3.16}$$

and

$$1-F_1=\left(1-\frac{1}{2N_0}\right)(1-F_0) \tag{3.17}$$

Substituting from the second equation into the first leads to

$$1-F_2=\left(1-\frac{1}{2N_1}\right)\left(1-\frac{1}{2N_0}\right)(1-F_0) \tag{3.18}$$

By analogy with the constant N case, it is appropriate to try to express this equation in the general form

$$1-F_t=\left(1-\frac{1}{2N}\right)^t(1-F_0) \tag{3.19}$$

where N is now the *effective* population size, usually symbolized as N_e. In our example $t = 2$, so

$$1-F_2=\left(1-\frac{1}{2N}\right)^2(1-F_0) \tag{3.20}$$

Setting the two expressions for $1-F_2$ equal to each other, we obtain

$$\left(1-\frac{1}{2N}\right)^2=\left(1-\frac{1}{2N_0}\right)\left(1-\frac{1}{2N_1}\right) \tag{3.21}$$

from which $1/N = \frac{1}{2}(1/N_0 + 1/N_1)$ turns out to be an excellent approximation. In general,

$$\frac{1}{N_e}=\frac{1}{t}\left(\frac{1}{N_0}+\frac{1}{N_1}+\cdots+\frac{1}{N_{t-1}}\right) \tag{3.22}$$

and so the effective size N_e is the **harmonic mean** of the actual numbers—the reciprocal of the average of the reciprocals. As illustrated in the problem below, the harmonic mean tends to be dominated by the smallest terms. In biological reality, this means that a single period of small population size, called a **bottleneck**, can result in a serious loss in heterozygosity. Population bottlenecks are thought to account for the very low levels of polymorphism found in extant populations of the elephant seal (Bonnell and Selander 1974) and the cheetah (O'Brien et al. 1985, 1987). A severe population bottleneck

often occurs in nature when a small group of emigrants from an established subpopulation founds a new subpopulation; the accompanying random genetic drift is then known as a **founder effect** (see Holgate 1966; Nei et al. 1975; Chakraborty and Nei 1977; Neel and Thompson 1978). Founder effects in human populations have implications in medical genetics, because human populations derived from small numbers of founders may have an elevated incidence of an otherwise rare genetic disorder. Examples include Tay-Sachs diseases in Ashkenazi Jews, diastrophic dystrophy in Finns, familial hyperchylomicronemia in Quebecois, and congenital total color blindness in Pingelap Islanders (reviewed in Scriver 2001). In addition to reducing the effective population size, and thereby increasing F, population bottlenecks and founder effects may affect many other aspects of the genetic variation, including causing a reduced number of alleles, a distorted distribution of allele frequencies, and an increase in linkage disequilibrium.

PROBLEM 3.7 Suppose a population went through a bottleneck as follows: $N_0 = 1000$, $N_1 = 10$, and $N_2 = 1000$. Calculate the effective size of this population across all three generations.

ANSWER Using Equation 3.22, we get $1/N_e = (\frac{1}{3})(\frac{1}{1000} + \frac{1}{10} + \frac{1}{1000}) = 0.034$, or $N_e = 1/0.034 = 29.4$. The average effective number over the three-generation period is only 29.4, whereas the arithmetic average number of individuals is $(\frac{1}{3})(1000 + 10 + 1000) = 670$.

Unequal Sex Ratio, Sex Chromosomes, Organelle Genes

A second important case in which the effective size of a nonideal population can readily be calculated concerns sexual populations in which the number of males and females is unequal. This inequality creates a peculiar sort of bottleneck; because half of the alleles in any generation must come from each sex, any departure of the sex ratio from equality will enhance the opportunity for random genetic drift. This situation is important in wildlife management, where, for many game animals (pheasants and deer come immediately to mind), the legal bag limit for males is much larger than for females. Although some management goals are served by such hunting regulations (for example, the species involved are usually polygamous, so one male can fertilize many females and overall actual population size can be maintained), it must be remembered that the resultant inequality in sex ratio reduces the effective population size. Specifically, if a sexual population consists of N_m males and N_f females, the actual size is

$$N_a = N_m + N_f \tag{3.23}$$

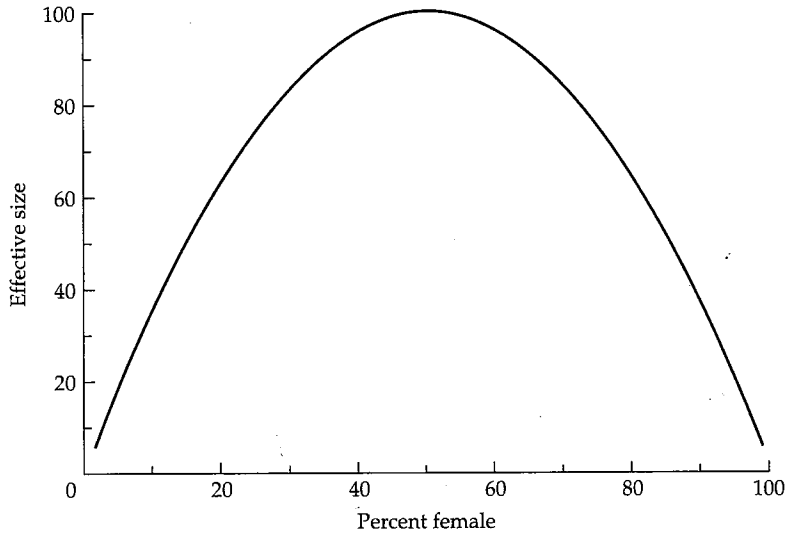


FIGURE 3.13 Effective size falls off rapidly in populations with a skewed sex ratio.

However, the effective population size is

$$N_e = \frac{4N_m N_f}{N_m + N_f} \tag{3.24}$$

Figure 3.13 shows the relationship between sex ratio and the reduction in effective population size. To take a realistic example, if hunting is permitted to a level at which the number of surviving males is one-tenth the number of females, then the effective population size is a mere one-third of the actual number of individuals in the population.

A related problem is the effective population size for an X-linked gene, in which case $\frac{2}{3}$ of the X chromosomes in any generation come from females in the previous generation and $\frac{1}{3}$ come from males. The variance effective population size for an X-linked gene is

$$N_e = \frac{9N_m N_f}{4N_m + 2N_f} \tag{3.25}$$

Equation 3.25 can be justified by noting that the sampling variance for the X chromosomes from males is $p_m q_m / N_m$, whereas the sampling variance for X chromosomes from females is $p_f q_f / 2N_f$, in which p_m and p_f are the frequencies of allele A in males and females, respectively. The frequency of an A-bearing X chromosome in the population is

$$p = \frac{1}{3}p_m + \frac{2}{3}p_f \tag{3.26}$$

Now we use the fact that, if a and b are constants and X and Y are independent random variables, then $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$. In this case $a = \frac{1}{3}$, $b = \frac{2}{3}$, and the variances of p_m and p_f are the binomial variances, and so

$$Var(p) = \frac{1}{9}\left(\frac{p_m q_m}{N_m}\right) + \frac{4}{9}\left(\frac{p_f q_f}{2N_f}\right) \tag{3.27}$$

At steady state, $p_m = p_f = p$ and $q_m = q_f = q$. Making these substitutions and factoring pq results in

$$Var(p) = pq\left(\frac{1}{9}\frac{1}{N_m} + \frac{4}{9}\frac{1}{2N_f}\right) = \frac{pq}{2\left[\dfrac{9N_m N_f}{4N_m + 2N_f}\right]} \tag{3.28}$$

The term in the square brackets corresponds to the N_e in Equation 3.25. It shows why this is a variance effective size: The binomial sampling variance in an ideal population is $pq/(2N_e)$.

PROBLEM 3.8 What is the effective population size for mitochondrial DNA? (Assume transmission is exclusively from mothers to all offspring.) What is the effective population size for a gene on the Y chromosome, given that the population consists of N diploid individuals and is in all respects a theoretically ideal population? (Assume XX individuals are female and XY individuals are male.)

ANSWER Mitochondrial DNA is transmitted essentially exclusively by females, and therefore the chance of drawing two mtDNAs that are identical by descent is $1/N_f$, where N_f is the number of females in the population. However, the probability that two randomly chosen autosomal genes are identical by descent is $1/(2N_e)$. Equating $1/(2N_e) = 1/N_f$ yields $N_e = N_f/2$ as the effective size for the population of mitochondrial DNA molecules. Since $N_f = N/2$ in an ideal population, the effective size for mitochondrial DNA, relative to an autosomal gene in an ideal population, is $N/4$. Similarly, the effective population size for the Y chromosome is $N_m/2$, where N_m is the number of males in the population. As with mitochondrial DNA, the effective size for Y chromosomal DNA, relative to an autosomal gene in an ideal population, is $N/4$. Note that, when $N_f = N_m$, even though mtDNA is present in all individuals, whereas the Y chromosome is present only in males, the effective size of mtDNA is not larger than that of the Y chromosome. The effective size depends on the sampling properties of a gene, which depends not only on how many individuals carry the gene but also on the gene's mode of transmission.

Variance in Offspring Number

An ideal population is one in which each breeding individual has an equal chance of contributing offspring to the next generation. Technically, this means that the statistical distribution of the number of offspring per individual is a binomial distribution with mean 1 and variance $1 - 1/N$. The distribution is binomial because its range is the fixed interval $[0, N]$, owing to the fact that no individual can have more than N progeny. If N is reasonably large, this binomial distribution is virtually identical to a Poisson distribution with mean and variance equal to 1. Nevertheless, the assumption that each individual has the same distribution of number of progeny is usually unrealistic because, in real organisms, breeding individuals can manifest large differences in their number of progeny. A more realistic model is one in which there are N individuals in the population and in which the ith individual ($i = 1, 2, \ldots, N$) produces n_i offspring. In this situation, the effective size of the population is defined as the reciprocal of the probability P that two randomly chosen gametes in the next generation come from the same parent in the previous generation (Crow and Kimura 1970). We will denote the mean and variance of the distribution of offspring number as ξ (Greek xi) and σ^2, respectively. With these definitions,

$$\xi = \frac{\Sigma n_1}{N} \text{ and } \sigma^2 = \frac{\Sigma n_1}{N} - \left(\frac{\Sigma n_1}{N}\right)^2 \tag{3.29}$$

The probability P that two randomly chosen gametes come from the same parent is given by

$$P = \frac{\Sigma \binom{n_i}{2}}{\binom{N\xi}{2}} = \frac{\Sigma n_i(n_i - 1)}{N\xi(N\xi - 1)} = \frac{\Sigma n_i^2 - \Sigma n_i}{N\xi(N\xi - 1)} \tag{3.30}$$

The rationale for Equation 3.30 is that the numerator is the number of ways that two randomly chosen alleles can be present in offspring from the same parent, and the denominator is the number of ways that two randomly chosen alleles can have any parents. Substitution of Equation 3.29 into Equation 3.30 and a little rearrangement yields

$$P = \frac{(\sigma^2 / \xi) + (\xi - 1)}{N - 1}$$

But since $N_e = 1/P$ by definition, we can write

$$N_e = \frac{N-1}{(\sigma^2/\xi)+(\xi-1)} \tag{3.31}$$

and so, when $\xi = 1$, N_e is approximately equal to N/σ^2. Therefore, a large variance in offspring number reduces the effective population size by a factor of $1/\sigma^2$, thereby speeding up the process of random genetic drift. The flip side of this principle suggests a management strategy for endangered species: Loss of genetic variation can be reduced when the variance in offspring number is minimized, because if σ^2 is smaller than 1, the effective population size can be larger than the actual population size.

Variance in offspring number can have a large effect on random genetic drift, as can be seen in particularly important cases in which genes are transmitted by different mechanisms in males and females (for example, in the X and Y chromosomes, or in mitochondrial and chloroplast DNA). Generally, even for nuclear genes, the variance in offspring number of males is far greater than that of females, and one particular consequence is that the effective size for the Y chromosome is much smaller than the theoretical value of $N_m/2$ implied by Problem 3.8.

Effective Size of a Subdivided Population

Finally we will consider a model in which a population is subdivided into D subpopulations (**demes**), each consisting of N diploid individuals, with migration among demes measured by a quantity m equal to the probability that a randomly chosen allele in any deme originates from one of the remaining $D-1$ demes. The population subdivision creates a situation in which two levels of random drift take place simultaneously. There is a drift process within each deme, which takes place relatively rapidly, and another drift process in the population as a whole, which takes place more slowly. Since the mathematics is somewhat rough going (Wakeley 1999, 2000), we shall present only the main result, which is that, when D is reasonably large, the effective population size of the entire population is given by

$$N_e = ND\left(1+\frac{1}{4Nm}\right) \tag{3.32}$$

In this equation, the factor ND comes from the within-deme phase of the random genetic drift, and the factor $1 + 1/(4Nm)$ comes from the among-deme phase. An interesting and important feature of the model is that, unless $4Nm$ is very large, the effective population size (N_e) is larger than the actual population size (ND). This seeming paradox results from the population subdivision. When there are many demes connected by low rates of migration, then even if one knew which allele in some deme is destined ultimately to become the common ancestor of all alleles in the population at some future time, the process by which this lucky allele spreads among the

demes takes a very long time. To put this in another way, when a population is subdivided, it can take a very long time for any two alleles in different demes to trace their lineages back to some common allele in a single deme in some ancestral population.

3.6 GENE TREES AND COALESCENCE

The Wright-Fisher model established a way of thinking that dominated population genetics for about 50 years, considering the genealogies of alleles as they proceed forward in time. It is equally valid to think about the ancestry of alleles as the genealogies proceed *backward* in time, and for some purposes this manner of thinking is much more powerful. A set of alleles sampled from a population furnishes more than an estimate of current allele frequencies. Each allele in the sample has an ancestral history dating back hundreds or thousands of generations. It is possible that a pair of alleles, sampled today, may have come from identical copies of the same allele produced by the same individual just a few generations ago, or the alleles may have had a common ancestor hundreds of generations ago. The term **coalescence** refers to the process in which, looking backward in time, the genealogies of two alleles merge at a common ancestor. In a sample of k alleles, for example, the first coalescence (looking backward in time) merges the k contemporary genealogies into $k-1$ ancestral genealogies, and the second coalescence merges these into $k-2$ genealogies, and so forth, until there remains but a single common ancestor for the whole sample of alleles. The idea of coalescent analysis is to consider the ancestral history of genes in a sample by developing a model for the time intervals between each coalescence (Kingman 1980, 1982a,b, 2000; Hudson, 1983; Rosenberg and Nordborg 2002).

To understand how the coalescent process works, consider in Figure 3.14 what happens as time moves forward in time (down the page). In each generation, there are a number of alleles in the population, and these alleles may be replicated and be present in the following generation; in some cases, however, an allele leaves no descendants and is lost from the population. By chance, some alleles may be sampled twice in constituting the next generation, and the probabilities of these events are the same as those under the Wright-Fisher model of random genetic drift. By a repetition of this process over time, one of the original alleles will eventually become fixed in the population. In the absence of mutation, the population would therefore be fixed for the same allele; however, because mutation may occur during the process, the alleles observed at the present will not all be identical in nucleotide sequence, even though they all descended from a single common ancestral allele.

In reality, we do not usually have the genealogical information enabling us to follow all the alleles through time in a population. Typically what we have is a single "snapshot" represented by a small sample of alleles taken at

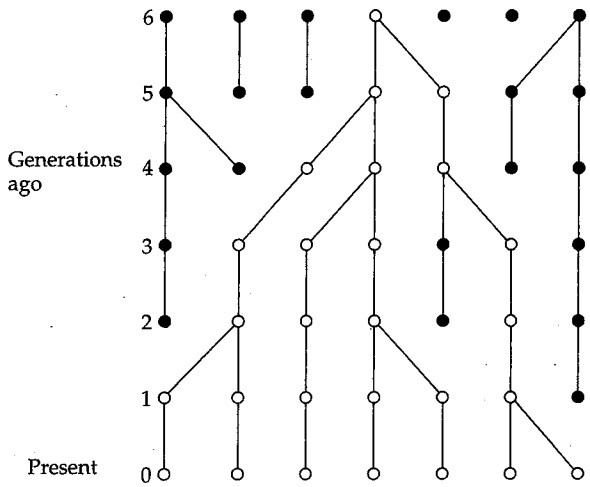


FIGURE 3.14 Diagram showing paths of ancestry of a set of alleles sampled in the present generation. The population is represented as having a constant size. The alleles in the original population are represented at the top. As generations progress forward in time (downward in the diagram), many alleles leave no descendants and therefore go extinct. Eventually one allele goes to fixation. Considering this process in reverse (bottom to top in the diagram), the sample observed in the present generation undergoes a series of coalescent events in which the k alleles present in the present generation had only $k - 1$ ancestral alleles. The coalescences continue backward in time until there is only one ancestral allele. The filled circles represent alleles present in previous generations that have left no descendant alleles in the present generation.

the present time. Now consider Figure 3.14 again, but this time look at what happens when we go backwards in time (up the page). We start with the $k = 7$ alleles in the sample at generation 0. In going from generation 0 to generation 1 (one generation back), we see that the genealogies of the two rightmost alleles "coalesced" into a single ancestral allele. As we go further back in time, the number of ancestral alleles has to either remain the same or decrease, and each reduction in the number of ancestral alleles is called a *coalescent event.*

Figure 3.14 illustrates one reason why coalescent reasoning is so powerful. If we were to study the process in Figure 3.14 going forward in time by means of computer simulation, many of the alleles that are tracked represent wasted computation, since they do not have descendants in the present generation (generation 0). These alleles are denoted by filled circles, and in this case there are 22 of them. On the other hand, if we were to study the same process going backward in time, none of the alleles tracked would be wasted, because each allele present in any generation must trace back to some allele

present in the previous generation. These alleles are represented by open circles, and in this case there are 27 of them. In other words, the forward simulation wastes nearly half its time generating alleles (22 alleles among a total of 49) that are of no interest because they do not contribute to the ancestry of the alleles present in the current population. This is not a great price in the present case, when the sample size is small, but in samples of hundreds of alleles, the vast majority of lineages simulated in the forward direction are unnecessary. In fact, in an original population of size $2N$ that has evolved for long enough that one of the alleles has gone to fixation in the contemporary population, it is unnecessary to deal with any of the original $2N - 1$ lineages that eventually go extinct.

Since we are interested in the time required for a pair of genealogies to coalesce, we need a model from which the coalescence times can be derived. Let us consider the immediate ancestry of two alleles. The probability that the two alleles came from the same allele in the previous generation is $1/(2N)$ (in a diploid population of size N), so the chance that they came from two distinct alleles the previous generation is $1 - 1/(2N)$. Similarly, the probability that three alleles in any generation originate from three distinct ancestral alleles in the previous generation is Pr{alleles 1 and 2 have distinct ancestors}Pr{allele 3 has a different ancestor from those of allele 1 and allele 2} = $[1 - 1/(2N)][1 - 2/(2N)]$. In general, the probability that k alleles had k distinct parental alleles the previous generation is

$$\Pr(k) = \prod_{i=1}^{k-1}\left(1 - \frac{i}{2N}\right) \approx 1 - \frac{\binom{k}{2}}{2N} \tag{3.33}$$

In each generation the sampling process occurs independently of what happened before, and so the probability that k alleles had k distinct parental alleles two generations ago is the square of the right-hand side of Equation 3.33. Consider two alleles again. Suppose we wish to know the probability that the common ancestor of these two alleles occurred exactly $t + 1$ generations ago. In this case there must have been no coalescence (i.e., two distinct ancestral lineages exist) for t generations, and then, in the next preceding generation, a coalescence occurred. The chance of two alleles not coalescing for t generations is $[1 - 1/(2N)]^t$, and the chance that they coalesce in the next generation is $1/(2N)$. The desired probability is the product of these or

$$\begin{aligned}&\Pr\text{(two alleles had common ancestor } t{+}1 \text{ generations ago)}\\ &= \frac{1}{2N}\left[1 - \left(\frac{1}{2N}\right)\right]^t \approx \frac{1}{2N}e^{-t/(2N)}\end{aligned} \tag{3.34}$$

The exponential is an approximation that is quite good when $1/(2N)$ is small. This distribution has a mean of $2N$ generations and a variance of $4N^2$. Note that the confidence interval around the mean time is not very tight, since the standard deviation of the distribution ($2N$) is equal to the mean.

Returning to our sample of k alleles, the probability that the k alleles do not coalesce for t generations, and then one pair coalesces to give $k-1$ alleles at $t+1$ generations ago is as follows:

$$= \Pr(k)^t\left[1-\Pr(k)\right]$$
$$\approx \frac{\binom{k}{2}}{2N}\exp\left[-\frac{\binom{k}{2}}{2N}t\right] \qquad (3.35)$$

This approximation is valid if $k << N$ (that is, if the sample size is much smaller than the population size, which is usually the case). The distribution in Equation 3.35 has a mean and variance given by

$$\text{Mean} = \frac{4N}{k(k-1)}\text{generations} \qquad \text{Variance} = \frac{16N^2}{\left[k(k-1)\right]^2}\text{generations}^2 \qquad (3.36)$$

Felsenstein, in his book *Inferring Phylogenies* (2004), gives a wonderful bugs-in-a-box analogy for the coalescent process that makes it simple and memorable, and we quote it with his permission.

> We can make a physical analogy (if a somewhat fanciful one) by considering a box containing hyperactive, indiscriminate, voracious, and insatiable bugs. We put k bugs into the box. They run about without paying any attention to where they are going. Occasionally two bugs collide. When they do, one instantly eats the other. Being insatiable, it then resumes running as quickly as before. It is obvious that the number of bugs falls from k to $k-1$, to $k-2$, as the bugs coalesce, until finally only one bug is left. ... The analogy is actually fairly precise. The number of pairs of bugs that can collide is $k(k-1)/2$. If there are $2N$ "places" in the box that can be occupied, the probability of a collision will be proportional to $k(k-1)/4N$. The size of the population corresponds to the size of the box. A box with twice as many "places" will slow the coalescence process down by a factor of two. So a simpleminded physical analysis of the bugs-in-a-box process will have the Kingman coalescent [our Equation 3.35] as the probability distribution of its outcomes. (p. 460)

Figure 3.15 shows what the gene genealogy is expected to be in the case of five alleles ($k = 5$). The genealogy is depicted in two forms, both of which are common in the literature. On the right, the tip of each line represents an allele in the original sample, and moving upwards (backward in time), each node (vertex) represents a coalescence to an ancestral allele. In the depiction on the left, the tips again represent the sampled alleles, but now each coalescent event is represented as a horizontal line. Going backwards in time (up the page), the coalescent time for i alleles to coalesce to $i-1$ alleles is denoted T_i ($i = 2$ to 5). The probability distribution of these coalescent times is given by Equation 3.35, and the expected values are shown in Figure 3.15. Starting with five alleles, the first coalescence is expected to occur $2N/10$ generations

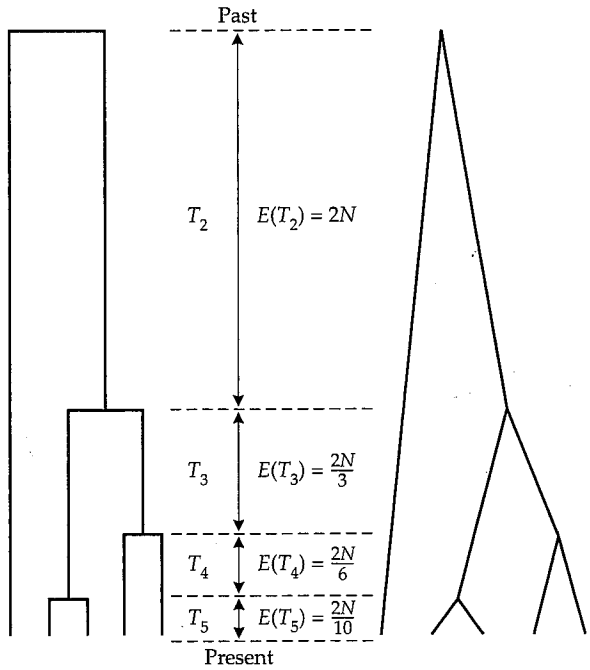


FIGURE 3.15 Two completely equivalent ways of illustrating the coalescences in a gene tree. On the left, the coalescent events are represented as horizontal lines, on the left they are represented as nodes. In any each generation, if there are k alleles present, the expected time back to the next coalescence is given by $4N/[k(1-k)]$. For example, starting with five alleles, the expected time back to the first coalescence is $4N/[(5)(4)] = 2N/10$. Note that the successive times get longer. When there are only two alleles, the time back to the final coalescence is $2N$ generations.

ago, the next at $2N/6$ generations prior to that, and so on. The distribution of each of these time intervals is exponential, with ever-increasing means as one goes back in time.

Note that the coalescent times become longer as one goes back farther in time, and the last coalescent time (from 2 alleles to 1) is the longest. This pattern is typical of coalescence in a population of constant size. Quantitatively, it requires a fraction $(1-1/n)/(1-1/k)$ of the total time for the last n of k alleles in a sample to coalesce (Felsenstein 2004). From this relationship it is easy to see that, if k is reasonably large (say, $k \geq 10$), almost half the time is required to coalesce the last two alleles ($n = 2$).

For a sample of k alleles, the time to the coalescence of all of the alleles (i.e., the most recent time that the sample of k alleles shared a common ancestor) is

$$t = 4N(1-1/k) \tag{3.37}$$

with variance

$$V = 4N^2 \prod_{i=2}^{k} \frac{1}{\binom{i}{2}^2} \tag{3.38}$$

(Kingman 1982a,b; Tajima 1983). As the sample size k increases toward the total population size, t approaches $4N$, which equals the expected time to fixation for a newly arisen neutral mutation that is destined to be fixed.

What is the probability that the most recent common ancestor of the sample (i.e., the allele to which all the lineages coalesce) to be also the most recent common ancestor of all the alleles in the entire population? The answer is given by the ratio $(k-1)/(k+2)$ (Rosenberg and Nordborg 2002). This probability is surprisingly large even for relatively small values of k. For example, for $k = 5$ it is already 67%, and for $k = 9$ it is 80%, and finally for $k = 19$ it is 90%. In other words, the most recent common ancestor in a sample of only 19 alleles has a 90% chance of being the most recent common ancestor of all the alleles in a population, no matter whether the population size is five hundred, a thousand, or a million.

Coalescent Effective Size

Equation 3.37 says that, as k increases toward $2N$, the expected time for all of the alleles in a population to coalesce into a single common ancestral allele is approximately $4N$. This is the *coalescent effective size*, which (when it exists) equals the inbreeding effective size (Sjodin et al. 2005).

In order to explain when a coalescent effective size does or does not exist, we need to consider the matter of time scale. Equation 3.35 gives the probability distribution of coalescent times, and note that we can eliminate an explicit dependence on $2N$ by measuring time t in units of $2N$ generations. With this scaling of time, the mean and variance of the coalescence times in Equation 3.36 become $2/[k(k-1)]$ generations and $4/[k(k-1)]^2$ generations^2, respectively. Likewise, the time to coalescence of all alleles in a population becomes 2. (Since the units of time are now $2N$ generations, the actual magnitude of 2 units on this time scale is $2 \times 2N = 4N$ generations.)

The coalescent effective size exists for any population process whenever time can be rescaled by a constant to reproduce the standard coalescent embodied in Equation 3.35. We have already considered one such example in which the variance in offspring number does not conform to the binomial distribution assumed in the Wright-Fisher model (see Equation 3.31). To simply matters, set the mean number of offspring per individual $\xi = 1$ in Equation 3.31 (which means that the population remains constant in size). In this case the effective size is, to a close approximation, $N_e = N/\sigma^2$ where σ^2 is the variance in offspring number; hence the standard coalescent process is recovered if we scale time in units of $2N/\sigma^2$ generations. This means that a large variance in offspring number reduces the effective population size and speeds up the process of random genetic drift.

There are many processes in which time can be rescaled to recover the standard coalescent process. These include certain models of population growth, age structure, and geographical structure (Emerson et al. 2001; Sagitov and Jagers 2005; Sjodin et al. 2005). The decisive factor is the time scale.

Many processes occur on an ecological time scale, which is typically shorter than the coalescent time scale. Among these processes are changes in population age structure or geographical structure when the rate of migration is sufficiently high. In such cases the "fast" process can simply be averaged. Even though the process does affect the coalescent times, it is only through a scaling factor analogous to the σ^2 in the case of variance in offspring number. Conversely, some processes occur on a geological time scale, which is typically longer than the coalescent time scale, and in this case the "slow" process can be ignored.

Trouble arises when the population processes occur on a time scale that is comparable to that of the coalescent time scale, because then there is no linear rescaling of time that yields the Wright-Fisher model, and so there is no coalescent effective size. What happens is that the effective size changes as the population evolves. As a specific example, Sjodin et al. (2005) consider a case in which the population size fluctuates randomly between $N = 10^3$ and $N = 10^5$ and show that, when the probability of a fluctuation is between 10^{-6} and 10^{-2} per generation, then no coalescent effective population size exists. The reason is that the changes in population size occurring on this time scale affect the coalescent times in a nonlinear and random manner. On the other hand, for probabilities greater than 10^{-2} the fluctuations are fast enough that the coalescent effective size equals the average effective size, and for probabilities smaller than 10^{-6} the fluctuations are slow enough that the coalescent effective size equals the initial size.

Coalescence with Population Growth

Changes in population size affect the probability distribution of the coalescent times. In populations of constant size, the coalescent trees typically have a sparse number of nodes and relatively long branches near the root (the root is the allele at which all lineages in the sample ultimately coalesce, the most recent common ancestor of all the alleles in the sample). This pattern can be seen in the "expected" coalescent tree in Figure 3.15.

In a population that is growing exponentially from an initial size of $N(0)$, the size at any time t is given by $N(t) = N(0)\exp(rt)$, where r is the exponential rate of growth. If both $N(0)$ and r are large, the coalescent trees are distorted such that they will have more and shorter branches near the root. The reason is that, in an expanding population, it takes longer for alleles to "find" each other starting at the tips and going back in time. Felsenstein's (2004) bugs-in-a-box analogy helps to understand this point, because in a box that is expanding in size, the bugs will take longer to bump into each other. In the extreme case of a very large $N(0)$ and r, the coalescences all take place very near the root, resulting in what is often called **star phylogeny**. (Thinking forward in time about a rapidly expanding population, the lineage of each allele has a very low probability of extinction, and so all the lineages persist and a graph of their genealogical relations resembles a star.)

Because coalescent trees in a rapidly expanding population have many short branches near the root, it is reasonable to suppose that, as r decreases, the coalescent events will be less concentrated near the root and will gradually move away from the root toward the tips. The number of branches near the root will decrease, and the branch lengths near the root will become longer until, when $r = 0$, the coalescent trees will assume the form of those with a constant population size.

The pattern of branching in coalescent trees can be used to make inferences about historical patterns of population growth (Pybus et al. 1999; Emerson et al. 2001). We will illustrate this point using a method proposed by Pybus et al. (1999) for estimating the growth rate of viral populations, in their case human immunodeficiency virus type 1 (HIV-1). They observed that, in simulated coalescent trees in samples of size $k = 400$ from populations with a constant population size, the vast majority (= 95%) of trees had three or fewer nodes between the root and the mid-depth point of the tree. (The *mid-depth point* of a tree is halfway between the tree's root and its tips.) A smoothed version of their simulation results is shown in Figure 3.16. This pattern is not expected in population that is undergoing rapid growth with an exponential rate of increase of r. In a growing population more nodes are expected near the root. Furthermore, the proportion of trees with three or

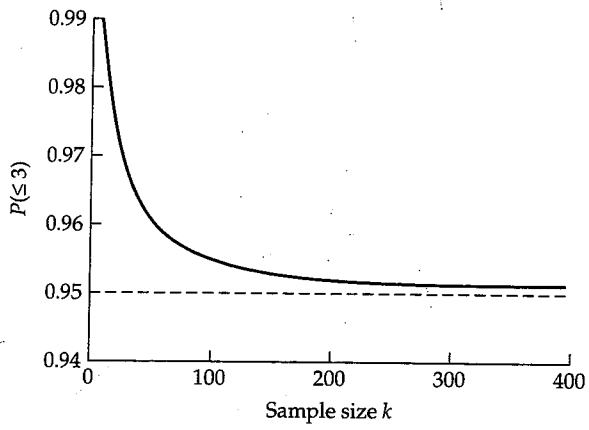


FIGURE 3.16 Random coalescent trees tend to have long branches near the root. In this graph, $P(\leq 3)$ represents the proportion of random coalescent trees that have three or fewer coalescent events between the root and the midpoint of the tree, as a function of the sample size. Even for rather large samples, more than 95% of the random trees have $P(\leq 3)$. This pattern contrasts with coalescent trees in populations that are growing, in which the number of coalescent events near the root tends to be greater. (Based on simulation results of Pybus et al. 1999.)

fewer nodes between the root and the mid-depth point is linearly related to the logarithm of the product of the current population size and the rate of exponential growth, which Pybus et al. (1999) show how to estimate in the case of HIV-1.

The mid-depth method has the drawback that it treats the coalescent tree as known, when it would be more appropriate to average over all possible coalescent trees compatible with the data, each weighed in proportion to its probability. Other methods for estimating population growth rates from coalescent trees are examined in Emerson et al. (2001). By comparing data to simulations, it is also possible to fit more complex demographic histories, such as past population bottlenecks (Thornton and Andolfatto 2006).

Coalescent Models with Mutation

The principles embodied in Equation 3.35 allows us to generate simulated gene genealogies whose branch lengths correspond to the assumptions of the Wright-Fisher model. It is important to emphasize that we do not usually know the *true* ancestral relationships among the alleles. The only cases in which the true ancestries are known come from experimental evolution studies carried out in the laboratory with viruses or microbes in which a sample of genomes is isolated and preserved by freezing at intervals during the process. In other contexts when we want to make inferences about a single sample, we simulate a large number of genealogies consistent with the composition of the sample and then make inferences based on the relative likelihoods of these genealogies.

To put the method more formally, what we are interested in is maximizing the likelihood L of observing the actual data D (typically DNA sequences) across all possible genealogies, given some model of mutation and population processes (Rosenberg and Nordborg 2002; Felsenstein 2004). Formally, we can write

$$L = \sum_{G} \Pr\{D \mid G, \mu\} \Pr\{G, \alpha\} \tag{3.39}$$

where G represents any particular genealogy, μ is the set of parameters that define the mutation model, and α is the set of parameters that characterize the population process (i.e., population size, growth rate, number of demes, migration rate, and so forth).

Equation 3.39 is usually intractable analytically, and therefore simulation of many thousands of random genealogies is carried out instead. In order to simulate the genealogies and sequences of alleles in a sample, we need to specify some type of mutation model. One widely used model is known as the **infinite-sites model**, in which each allele is considered as a sequence of nucleotides with mutation altering any site in the sequence. If the mutation rate is sufficiently low, then most sites will be monomorphic in the sample, and all polymorphic sites will be segregating for just two nucleotides. Much

of the available data on allelic variation in DNA sequence seems consistent with this view: Few nucleotide sites are segregating for more than two nucleotides. If the DNA sequence is sufficiently long and the frequency of polymorphic sites low, then most of the time new mutations will occur at sites that were previously monomorphic. The infinite-sites model is based on these assumptions. It was developed originally by Kimura (1969, 1971), who considered nucleotides as unlinked, and by Watterson (1975), who took account of the nearly complete linkage among sites.

In order to simulate allele sequence data representing samples drawn from a population obeying the infinite-sites model, Hudson (1990, 1993) showed that one can proceed as follows:

1. Estimate the value of $\theta = 4N\mu$ for the gene or region of interest, where N is the effective population size and μ is the per-site mutation rate; this estimate can be based on the number of segregating sites in the sampled sequences or on the average number of site mismatches when the sequences are compared in all possible pairs. (These methods are discussed in Chapter 4.)
2. For the observed sample of k alleles, draw random numbers with appropriate exponential distributions to construct a gene genealogy such that times of coalescence follow Equation 3.35.
3. Randomly scatter mutations with a Poisson distribution on each branch of the tree, such that the mean number of mutations on each branch is given by μt, where t is the branch length in units of generations. (For example, the first interval from the root in Figure 3.15 has an expected length of $2N$ generations; hence each branch from the root has an expected number of mutations of $2N\mu = \theta/2$.
4. Repeat steps 2 and 3 some 10,000 or more times, and estimate the likelihood of observing the actual data across all genealogies according to Equation 3.39.

Applications of Coalescent Methods

A typical (and pioneering) application of this approach is found in Hudson et al. (1994). These authors examined a sample of size $k = 10$ allelic sequences, each of length 1.4 kb, encoding the enzyme superoxide dismutase in a Spanish population of *Drosophila melanogaster*. Among the 10 alleles, five were identical in sequence, whereas the others were all different and contained a total of 55 polymorphic sites. It seems unlikely that this configuration of polymorphisms in a sample would occur by chance, and it is tempting to test this hypothesis by sprinkling 55 polymorphisms at random among the 10 sequences. But this is invalid, because the samples are related through ancestry, and their possible genealogies need to be taken into account.

To test the hypothesis that the samples could arise by chance if all the polymorphisms were selectively neutral, the authors simulated 10,000 sam-

ples with $k = 10$. Rather than assigning mutations to each branch according to a Poisson distribution, they randomly placed 55 mutations on the genealogy in proportion to the branch lengths. The former is technically correct, but the latter is justified unless the number of mutations is very small (Wall and Hudson 2001; Depaulis et al. 2001). For each simulated genealogy, they checked whether it contained a set of five identical alleles. Finding that only about 1% of the simulated genealogies yielded sequences with this characteristic, they felt justified in rejecting the neutral Wright-Fisher model and suggested that the high frequency of one allele resulted either from selection of the gene itself or selection of some linked gene.

As another example of the application of coalescent methods, consider the interpretation of a fragment of mitochondrial DNA that had been amplified from a skeletal Neanderthal bone dated to 30,000–100,000 years ago (Krings et al. 1997). The Neanderthal sequence was compared with the mtDNA of 986 modern humans, with the result that the most recent common ancestor of the Neanderthal sample and that of modern humans was estimated to be several times more ancient than that of the human sequences themselves. This was taken as evidence that Neanderthals and the ancestors of anatomically modern humans were separate species that did not undergo interbreeding.

These data were reconsidered by Nordborg (1998) in the light of a different population model using coalescent simulations as well as analytical methods. This analysis confirmed the conclusion that Neanderthals and the ancestors of anatomically modern humans did not fuse in equal numbers to form a single random mating unit. But other models with intermixing of the gene pools cannot be excluded. For example, in a model in which the populations of Neanderthals and the ancestors of anatomically modern humans came together 68,000 years ago, the probability that all Neanderthal mtDNA lineages were lost by random genetic drift is 52% even if Neanderthals comprised as much as 25% of the mixed population. Because the time scale of random drift is smaller for autosomal genes than for mtDNA by a factor of 4 (see Problem 3.8), the probability of loss of all Neanderthal sequences from autosomal genes is much smaller. Nordborg's (1998) calculations imply that, if Neanderthals comprised 25% of a mixed population 68,000 years ago, 90% of our autosomal genes would still be segregating Neanderthal sequences! This example illustrates the limitations of attempting to draw far-reaching inferences from a sample of a single, nonrecombining sequence such as mitochondrial DNA, and it also shows how important it can be to consider various kinds of population models in accounting for observed data.

3.7 THEORETICAL IMPLICATIONS OF COALESCENCE

The coalescent approach can be used to derive many fundamental principles in population genetics. For example, Equation 3.36 defines the expected

length of each interval T_k in a coalescent tree (see also Figure 3.15), and so the expected sum of the branch lengths $E(T)$ for the entire tree is given by

$$E(T) = E\left(\sum_{i=2}^{k} iT_i\right) = \sum_{i=2}^{k} iE(T_i) = \sum_{i=2}^{k} i\frac{4N}{i(i-1)} = 4N\sum_{i=1}^{k-1}\frac{1}{i} \qquad (3.40)$$

The expected number of segregating sites, $E(S)$, in a set of aligned DNA sequences is equal to the product of the mutation rate and the expected length of all the branches in the coalescent tree, or $\mu E(T)$. The expected number of segregating sites in a sample of k aligned sequences is therefore obtained from Equation 3.40 as

$$E(S) = \mu E(T) = 4N\mu\sum_{i=1}^{k-1}\frac{1}{i} = \theta\sum_{i=1}^{k-1}\frac{1}{i} \qquad (3.41)$$

where $\theta = 4N\mu$. This is the expected number of segregating sites in the infinite-sites model, which we shall discuss again in Chapter 4 in a different context. Note that μ is not the mutation rate per nucleotide site; rather it is the mutation rate across the entire length of the DNA sequence.

As another example of the theoretical utility of the coalescent approach, consider a sample of alleles taken from population presently in equilibrium between mutation and random genetic drift, which means that new mutations in each generation occur at the same rate as old mutations are lost due to random drift. Tracing any pair of alleles back to the previous generation, the pair of alleles could either coalesce, with probability $1/(2N)$, or failing to coalesce, one or the other allele could have undergone mutation with probability 2μ. (The factor 2 is necessary because either of the alleles could mutate.) These are the only two events that affect allele identity by descent, and the sum of their probabilities is $1/(2N) + 2\mu$. The probability of identity by descent (F) is therefore the fraction of the time that the alleles coalesce, or

$$F = \frac{\frac{1}{2N}}{\frac{1}{2N} + 2\mu} = \frac{1}{1+\theta} \qquad (3.42)$$

This expression will also be derived again in Chapter 4 using different methods.

Coalescent methods are not limited to the consideration of the Wright-Fisher model. If one can develop a recursion equation for probabilities of recombination, migration, or other such phenomena in the context of a gene tree, then often powerful insights can be derived from coalescence approaches. For our purposes, suffice it to say that the method can generate classical results, often with much less difficulty, and as we have seen in the previous section, the coalescent approach is especially well suited to making infer-

ences about samples drawn from natural populations by averaging across a large number of simulated trees. The exceptional speed with which computers can simulate samples drawn from a neutral coalescent has yielded unprecedented opportunities for testing the correspondence between observed data and theoretical predictions.

PROBLEM 3.9 In a model of pure random drift, the probability distribution for the number of generations back to the first coalescence in a sample of k genes taken from a haploid population of size N is approximately:

$$\Pr\left\{\text{first coalescence occurred } t \text{ generations ago}\right\}$$
$$= ze^{-zt} \text{ where } z = \binom{k}{2} / N$$

From this one can show that the mean number of generations back to the first coalescence is $1/z$. The more genes in the sample, the more likely it will be that a coalescence occurred recently. Calculate the expected time to first coalescence in a population of $N = 450$ for a sample of 10 genes. How many genes would you have to sample to reduce this coalescence time by half?

ANSWER The expected time to first coalescence in a population of $N = 450$ for a sample of 10 genes is

$$N / \binom{k}{2} = 450 / \binom{10}{2} = 450 / (10 \times 9 / 2)$$
$$= 10 \text{ generations}$$

To determine how many genes one would have to sample to halve this coalescence time, solve for

$$5 = 450 / \binom{k}{2}$$

This is equivalent to $90 = k!/[2!(k-2!)]$ or $180 = k(k-1)$. This is a quadratic equation $k^2 - k = 180$ that fits the general form $ax^2 + bx + c = 0$, where $a = 1$, $b = -1$, $c = -180$. The solutions are given by $[-b \pm \sqrt{(b^2 - 4ac)}]/(2a)$, and in this case the solution we want is $k = 13.9$ (the other solution is negative). Hence a sample of size 14 will reduce the initial coalescence time to about 5 generations (4.94 to be exact). If you do not know the quadratic formula, you can also get this answer by trial and error. In any case, increasing the sample from 10 to only 14, we expect to find any pair of alleles only half as divergent from each other.

Coalescent Models with Recombination

Coalescence with recombination is among the most difficult problems in modern population genetics (Rosenberg and Nordborg 2002; Stumpf and McVean 2003). To understand why it is so difficult, consider the trees in Figure 3.17. These are conventional coalescent trees, but the nodes and tips are labeled with the identities of two distinct nucleotide sites. The tips are also labeled 1–4 to identify the individual alleles in the sample. The symbols A and a represent a single-nucleotide polymorphism at the site. The A could be a G nucleotide at the site, for example, and the a a T nucleotide at the same site. Likewise, B and b represent a single-nucleotide polymorphism at the second site.

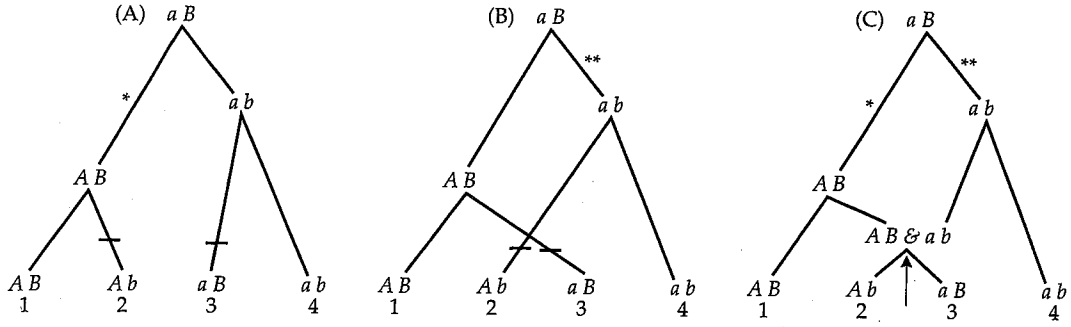


FIGURE 3.17 Coalescence and recombination of single nucleotide polymorphisms (A, a) and (B, b) in samples of size 4. (A) Coalescent tree with regard to A and a, where the asterisk marks the branch with a mutation of a to A. The horizontal tick marks represent an inferred recombination event. (B) Coalescent tree with regard to B and b, where the double asterisk marks the branch with a mutation of B to b. The horizontal tick marks again represent an inferred recombination event. Tree (A) is incompatible with tree (B), but a consistent tree (C) results from an ancestral recombination graph in which a coalescence can depict a recombination event. In this case, the arrow points to the coalescence where the recombination took place, and the recombinant chromosomes recreate their ancestral parental types.

Tree (A) in Figure 3.17 depicts the coalescences of the samples with respect to the A and a pair of alleles. The time of the a to A nucleotide substitution is indicated by the asterisk. Tree (B) depicts the coalescences with respect to the B and b pair of alleles. In this case the time of mutation of B to b is indicated by the double asterisk. Both tree (A) and tree (B) accurately portray the ancestry of the A, a and B, b pairs of alleles, respectively. The problem is that the trees are different. In tree (A), the first coalescence joins samples 1 and 2, whereas in tree (B) it joins samples 1 and 3.

The reason for the discrepancy is that, at the time depicted by the horizontal ticks in (A) and (B), an $A\,B$-bearing chromosome underwent recombination with an $a\,b$-bearing chromosome to yield recombinant $A\,b$ and $a\,B$ products. The tipoff that the nucleotide sites have undergone recombination is that all four possible types of chromosomes ($A\,B$, $A\,b$, $a\,B$, and $a\,b$) are present in the sample. The probability that any sample will contain all four types of chromosomes increases with the size of the sample, so large samples are desirable for detecting recombination. Even with large samples, however, many recombination events are likely to be missed. This test for recombination is valid only if the mutation rate is sufficiently low that each site in an ancestral history can mutate no more than once. If recurrent mutation can occur, which happens in viruses with a high mutation rate like HIV-1, then the detection of recombination becomes more difficult (McVean et al. 2002).

How can the inconsistency between tree (A) and tree (B) in Figure 3.17 be resolved? The conventional method of resolution is shown in tree (C) (Hudson 1990). In this case, as the ancestry of each chromosome is traced back in time, when two chromosomes come together, either of two events can happen: the chromosomes either (1) undergo recombination, or (2) undergo coalescence. In tree (C), the arrow indicates the recombination event, which is depicted by the recombinant chromosomes recreating their ancestral parental types. A graph such as that in C is called *the ancestral recombination graph* for the haplotypes in the sample.

The device of allowing recombination or coalescence at each node resolves the inconsistency in the tree, but at the same time it reveals the complexity of the process. Analytical results are possible only in the simplest of cases (Hudson 1990, 2001). Simulation is an alternative, but suppose you were assigned to simulate the ancestral history of a region of DNA in which there were 50 segregating sites and in which recombination can take place. Intuitively you can see that the task is horrendous. You must generate an ancestral history, sprinkle the branches with mutations at sites along the sequence, decide which nodes will result in coalescence and which result in recombination, and keep track of where in the sequence the mutations and the recombination events took place. The simulation therefore yields coalescent tree whose elements are one-dimensional graphs.

Complex as it is, simulation of coalescence with recombination is possible. The problem is that a random coalescent tree is not even remotely likely to generate a simulated sample of alleles showing characteristics such as linkage disequilibrium that are similar to those in the actual sample. This problem can be dealt with in various ways. One way is to ignore coalescences and estimate the rate of recombination based on various characteristics of the sample itself. Just as the relevant parameter for mutation is $4N\mu$, the relevant parameter for recombination is $4Nr$, where N is the effective population number and r is the rate of recombination per generation. One estimate of $4Nr$ is based on comparing the samples alleles in all possible pairs and for each pair tabulating the number of nucleotide mismatches. The distribution of pairwise mismatches is a basis for estimating the rate of recombination, because the variance in the number of mismatches is reduced by recombination (Wakeley 1997). This point can be appreciated by comparing the mismatch distribution in a sample consisting of *A B*, *A B*, *a b*, and *a b* with that in a sample consisting of *A B*, *A b*, *a B*, and *a b*, the latter of which is indicative of recombination. In both samples the mean number of pairwise mismatches is 1.33, but the variances are 0.89 and 0.22, respectively. The advantage of this approach is that it is completely straightforward; the main limitation is that it does not use all of the information in the sample and so has a larger sampling variance than necessary.

The opposite approach to using summary statistics is to carry out extensive simulations to perform a full likelihood analysis as encapsulated in

Equation 3.39. The problem with this approach, as we have suggested, is that random simulations are extremely unlikely to approximate the actual data. The parameter space is so large that, except in simple cases, a full-blown likelihood approach requires computational capabilities that exceed those of even the most powerful computer systems. This problem has stimulated the implementation of methods that reduce the dimensionality of the problem by collapsing the full data set into to summary statistics, and by zeroing in on portions of the parameter space that is the most relevant. Approximate Bayesian coalescence methods collapse the observed data to summary statistics such as the number of distinct haplotypes or the average number of mismatches between pairs of sequences. For each coalescent tree that is simulated, the same summary statistics are calculated. If the difference between the observed data and simulated sample is small enough (based on a somewhat arbitrary threshold), then the parameters from the simulated sample are accepted. Repeating the simulated sampling many times yields what is called the *posterior distribution* of the parameter estimates.

Two widely used methods that avoid random guesses at the parameter values but instead enable the computer to spend more time in the "most promising" regions of the parameter space are *Markov Chain Monte Carlo* (MCMC) and *sequential importance sampling*. Even these methods are challenged with the large size of currently available data sets. Both of these methods require a criterion for the goodness of fit between the data and the model, and this is itself sometimes difficult to compute. Instead of calculating the full likelihood, a *composite likelihood* is sometimes used, in which the simpler problem of calculating the likelihood for each individual nucleotide site is solved, and then the overall likelihood is taken as the product of the likelihood over the set of nucleotides (Kim and Stephan 2000; Hudson 2001; McVean et al. 2002; Zhu and Bustamante 2005; Carvajal-Rodriguez et al. 2006). This approach assumes statistical independence across the nucleotide sites—an assumption that can hardly be justified—but in practice the method provides a reasonable criterion for acceptance, and it seems to perform much better than might be expected.

Linkage Disequilibrium Mapping

Analysis of coalescence with recombination is important because understanding the consequences of mutation, recombination, and random drift acting simultaneously is at the heart of using human population samples to make inferences about genetic risk factors for multifactorial genetic diseases such as hypertension, diabetes, and schizophrenia (see Chapter 10). The underlying principle is that, in a finite population, the processes of mutation, recombination, and random drift result in *linkage disequilibrium*, the nonrandom association of the alleles along a chromosome, which we have already examined in the context of very large (theoretically infinite) populations in Chapter 2.

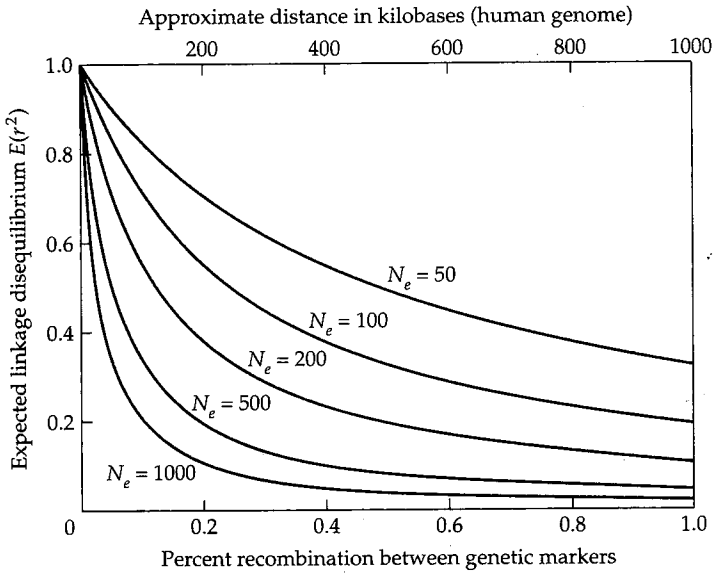


FIGURE 3.18 Linkage disequilibrium (r^2) expected at steady state in the infinite-sites model as a function of the effective population size and the frequency of recombination. The scale in nucleotides across the top is approximate for the average across the human genome, but local recombination rates in the human genome are highly variable.

Quantitative aspects of linkage disequilibrium in the presence of mutation, recombination, and drift are examined in detail in Chapter 9, and we defer the details until then. The main result is summarized in the graphs in Figure 3.18, which show the expected linkage disequilibrium between two genetic markers (for example, SNPs) as a function of the percent recombination between the markers and the effective size of the population. Roughly speaking, in the human genome, a value of 1% recombination corresponds to about 1 Mb of DNA, and using this rule of thumb, a scale in kilobase pairs is shown across the top. You may recall from Chapter 2 that the measure of linkage disequilibrium r^2 has the intuitive meaning that its square root (that is, $\sqrt{r^2}$) is the correlation coefficient between the alleles present in a single chromosome. Hence an $r^2 = 0.2$ implies a correlation coefficient of $\sqrt{0.2} = 0.45$, which is reasonably large. Many of the values plotted in Figure 3.18 have $r^2 = 0.2$. Even for N_e as large as 1000, the equilibrium r^2 is expected to be greater than 0.2 for SNPs separated by 100 kb. This reasoning suggests that the human genome might exhibit substantial linkage disequilibrium across regions at least on the order of several tens of kb, and this expectation has been substantiated by the The International HapMap Consortium (2005).

The analysis of associations between SNPs and complex diseases that have a genetic component is known as **linkage disequilibrium mapping**. This approach will be examined in some detail in Chapter 10. Its goal is to identify individual SNPs that are in or near genes with mutant alleles that predispose to the disease. These mutant alleles are known as *genetic risk factors* for the disease, and the measure of association is the magnitude of the linkage disequilibrium. In essence, such studies examine a large sample of affected individuals (*cases*) and an equally large number of matched, nonaffected individuals (*controls*). The individuals are genotyped for hundreds of thousands of SNPs across the genome, and SNPs are identified that are more frequently found in affected individuals than in unaffected controls. (For such a large number of statistical tests, the issues of false positives and false discovery discussed in Chapter 2 become critical.) SNPs that are significantly associated and reproducible in independent studies are assumed to mark the genomic locations of the genetic risk factors. Any genetic risk factor is likely to show linkage disequilibrium with multiple nearby SNPs, because the genealogical trees of SNPs linked to the genetic risk factor are correlated. Searching among these for the SNP whose inferred genealogy yields the best separation of cases and controls provides a finer level of genetic resolution of the risk factor. These applications are part of the reason that coalescence with recombination is one of the most active fields of research in modern population genetics.

SUMMARY

1. Because of random sampling of gametes in each generation, the allele frequencies in a finite population fluctuate with a theoretical variance equal to $pq/(2N)$. These fluctuations in allele frequency are the basis of random genetic drift.
2. The Wright-Fisher model extends the idea of binomial sampling to multiple generations and implies that, in a population in which the only force acting on allele frequencies is random genetic drift, the probability that an allele will drift to fixation is equal to its initial frequency in the population.
3. Diffusion approximations of the Wright-Fisher model make use of second-order partial differential equations for the distribution of allele frequencies among subpopulations at any given time, when the initial allele frequency among the subpopulations is stipulated. The diffusion approach has yielded important insights into the consequences of drift, such as that, for a new neutral mutation destined to be fixed, the expected time to fixation is $4N$ generations.
4. A useful way to think about random drift is to consider a set of subpopulations of the same size undergoing repeated generations of sampling and drift. Within each of these subpopulations, genotypes are composed

by drawing alleles at random, so that each subpopulation is always in Hardy-Weinberg equilibrium. Averaged across the subpopulations, the frequency of heterozygous genotypes is smaller than that expected from HWE, and the heterozygosity decreases at an average rate of $1/(2N)$ each generation.

5. Real biological populations do not usually fit the Wright-Fisher model because the allele frequencies change more rapidly than would be expected based on the actual population size. The drift model gives better correspondence to reality by calculating the effective population size N_e, which takes into account the sex ratio, variance in offspring number, fluctuation in population size over generations, or population subdivision. In some cases the effective population size can be greater than the actual population size.
6. Many aspects of random drift are greatly simplified by considering the genealogical history of alleles. Looking backward in time, the lineages of alleles come together (coalesce) at points in time when they both originated by replication of a single ancestral allele.
7. The distribution of coalescent times is exponential, which allows computer simulation of coalescent trees to be implemented very easily, yielding the characteristics expected in samples from a population evolving under any particular model. These simulated samples can be compared with actual data to test hypotheses or estimate population parameters.
8. The joint effects of mutation, recombination, and random drift result in a steady state in which the magnitude of linkage disequilibrium is a function of the frequency of recombination and the effective population size. In human populations, the linkage disequilibrium expected opens the possibility of linkage disequilibrium mapping of genetic risk factors for disease.

PROBLEMS

1. Describe the Wright-Fisher model of random genetic drift. The elements T_{ij} of the transition matrix are probabilities. How should T_{ij} be interpreted?
2. Explain how the basic concept behind the Kolmogorov forward equation differs from that underlying the Kolmogorov backward equation.
3. Explain why, in any population model in which random genetic drift plays a role, "backward" computer simulation, which starts with the alleles present in the current population and traces their coalescence backward in time, is computationally much simpler than "forward" simulation, which starts with the alleles present in the original population and simulates random genetic drift moving forward in time.

4. In an ideal diploid population of size 50, what is the probability that a neutral allele present in exactly one copy will be lost in the next generation? What is the answer if the allele is present in two copies?
5. Suppose that a diploid population of size 50 undergoes a change in average heterozygosity across loci from 0.50 to 0.42 in a single generation. Is it plausible to attribute this magnitude of change to random drift alone?
6. How many generations of random genetic drift are required to reduce the expected heterozygosity to 5% of its initial value in a diploid randomly mating population of size 10? Of size 50?
7. An autosomal gene in a colony of 28 Asian wild mice, *Mus castaneus*, undergoes mutation to a new neutral allele. Assuming that the population conforms to the Wright-Fisher model, what is the probability that the allele eventually becomes fixed? What is the probability that it eventually becomes lost? What is the average time to loss, given that the allele is eventually lost. What is the average time to fixation, given that it is eventually fixed?
8. What are the answers to Problem 3.7 if the mutant gene is X-linked and the population consists of equal numbers of males and females? What if the gene is Y-linked?
9. An isolated population of apine edelweiss (*Leontopodium alpinum*) loses half its heterozygosity in 30 generations. What is its effective population size?
10. Remote Pitcairn Island in the South Pacific was settled in 1789 by Fletcher Christian and eight fellow mutineers from *HMS Bounty*, along with a small number of Polynesian women. Although many descendants have left the island in the intervening years, there has been essentially no immigration. Assuming an effective size of 20 in each of the eight generations since the island's settlement, what value of F_t would be expected in today's population from random genetic drift?
11. Show that random genetic drift requires an average of $t = 2N \ln(x)$ generations to reduce the expected heterozygosity from H_0 to H_0/x.
12. A large randomly mating diploid population with two neutral alleles A and a at allele frequencies $p_0 = \frac{1}{3}$ and $q_0 = \frac{2}{3}$, respectively, splits into a large number of isolated subpopulations each of effective size 50. Within each subpopulation mating is random, but the allele frequencies diverge due to random genetic drift. After 69 generations, what are the expected genotype frequencies of AA, Aa, and aa, averaged across the subpopulations? In one of the subpopulations, the allele frequency of A equals 0.3. What are the expected genotype frequencies in this individual subpopulation?

13. Two inbred strains of the azuki bean beetle *Callosobruchus chinensis* are crossed and their progeny allowed to mate at random each generation thereafter. Among 100 single-nucleotide polymorphisms differing in the original inbred strains, what number would be expected to remain segregating after 10 generations, assuming an effective population size of 80 individuals? How many would be expected to remain unfixed after 50 generations?
14. Use Equation 3.14 to show that approximately $2N$ generations of random genetic drift are required to reduce the number of segregating genes by a factor of e ($e = 2.71828\ldots$) given initial allele frequencies close to 0.5.
15. What is the effective population number in a population of African lions, *Panthera leo*, in which each breeding male controls a harem of five females and the total population consists of 200 males and 200 females?
16. What is the effective population size of a herd of 10 dairy cows and 1 bull? What is it for 40 cows and 1 bull? For 10 cows and 2 bulls?
17. What is the variance effective population size for an X-linked gene in a population consisting of 100 females and 10 males? In a population of 10 females and 100 males?
18. In a haploid population of constant effective size 50, what is the probability that two randomly drawn alleles shared a common ancestor exactly 100 generations ago?
19. In a population of effective size 30, how many generations are required on the average to coalesce from 4 alleles to 3? From 3 alleles to 2? From 2 alleles to 1?
20. In a haploid population of effective size 50, how many sequences k should be present in a sample for the first coalescence to have an average time of 10 generations.
21. In the infinite-sites model, if $\theta = 10$, how many segregating sites are expected in a sample of size 10? 20? 50?

4.1 Mutation 151

Irreversible Mutation 152

Reversible Mutation 156

4.2 Mutation and Random Genetic Drift 158

Probability of Fixation of a New Neutral Mutation 160

4.3 The Neutral Theory of Molecular Evolution 161

4.4 The Infinite-Alleles Model 162

The Ewens Sampling Formula 166

The Ewens-Watterson Test 168

4.5 Infinite-Sites Model 172

Nucleotide Polymorphism and Nucleotide Diversity 175

Tajima's D Statistic 176

The Fu and Li Test of Fit to Neutral Coalescence 178

4.6 Mutation and Recombination 182

A Model for the Evolutionary Benefit of Recombination 183

Muller's Ratchet 185

Piecewise Recombination in Bacteria 187

Animal Mitochondrial DNA 189

CHAPTER 4

MUTATION AND THE NEUTRAL THEORY

Several processes can create new types of genetic variation in populations or promote the reorganization of previously existing variation either within genomes or among subpopulations. The ultimate source of genetic variation is **mutation**, by which we mean any heritable change in the genetic material. Mutation therefore includes a change in the nucleotide sequence of a single gene as well the formation of a chromosome rearrangement, such as an inversion or a translocation. Recombination brings mutations of different genes together into the same chromosome, and migration enables mutations to spread among subpopulations. A **transposable element** is a DNA sequence able to replicate and insert into any of a large number of sites in the genome. By insertion in or near a gene, a transposable element can alter the level or pattern of gene expression, and recombination between transposable elements can result in a chromosome rearrangement, for example, an inversion. In this chapter, we consider the processes by which genetic variation is created and examine the expected fate of mutations in natural populations.

4.1 MUTATION

Mutation is the ultimate source of genetic novelty underlying evolutionary change. However, most wildtype genes mutate at a very low rate, typically in the range from 10^{-4} to 10^{-6} new mutations per gene per generation. Even a low mutation rate can create many new mutant alleles because, in a large

population, each of a large number of genes is at risk of mutating. In a population of size N diploid organisms, there are $2N$ copies of each gene, each of which can mutate in any generation. For example, if the mutation rate (probability of mutation) is 10^{-9} per nucleotide pair per generation, then in each human gamete, the DNA of which contains approximately 3×10^9 nucleotide pairs, there would be an average of three new mutations in each generation; each newly fertilized egg would therefore carry, on the average, six new mutations. The present-day human population of approximately 6.5 billion people would therefore be expected to carry about 40 billion new mutations that were not present even one generation earlier.

Irreversible Mutation

Although mutation may create a new allele, the initial frequency of the mutant allele must be very low when the population size is large. To be specific, a single new mutant allele in a diploid population of size N has an initial frequency of $1/(2N)$. New mutations in subsequent generations may augment the number of mutant alleles, but recurrent mutation alone increases the allele frequency of the mutant very slowly. Consider a specific example in which A is the wildtype allele and a the mutant form. If there is exactly one new mutation per generation, and random genetic drift is ignored, then the allele frequency of a increases according to the series $1/(2N), 2/(2N), 3/(2N), \ldots$ and, if N is large (for example, $N = 10^6$), then the increase is very slow indeed. Hence, the tendency for allele frequency to change as a result of recurrent mutation (the **mutation pressure**) is very small. On the other hand, the cumulative effects of mutation over long periods of time can become appreciable.

A useful model for thinking about mutation is the Hardy-Weinberg model of Chapter 2, but with mutation permitted. For the moment, we focus on mutations that have so little effect on the ability of the organism to survive and reproduce that natural selection does not appreciably influence their frequency. We will also assume that mutation is *irreversible*, which means that a cannot reverse-mutate to A. To avoid complications resulting from change in allele frequency due to chance, we will also assume a population that is infinite in size.

Consider a gene with two alleles, A and a, and suppose that A mutates to a at a rate of μ mutations per A allele per generation. In other words, each A allele has a probability of μ of mutating to a in any generation. We will symbolize the allele frequency of A as p and that of a as q and keep track of generations with subscripts. Hence, p_t and q_t are the allele frequencies of A and a, respectively, in the tth generation, where $t = 0, 1, 2, \ldots$ In any generation, $p_t + q_t = 1$ because A and a are the only alleles considered.

Next we will deduce a formula for the allele frequency p_t in terms of the allele frequency p_{t-1} in the previous generation. In generation t, p_t includes all the A alleles in generation t that did not mutate in that generation, and so

$$p_t = p_{t-1}(1 - \mu)$$

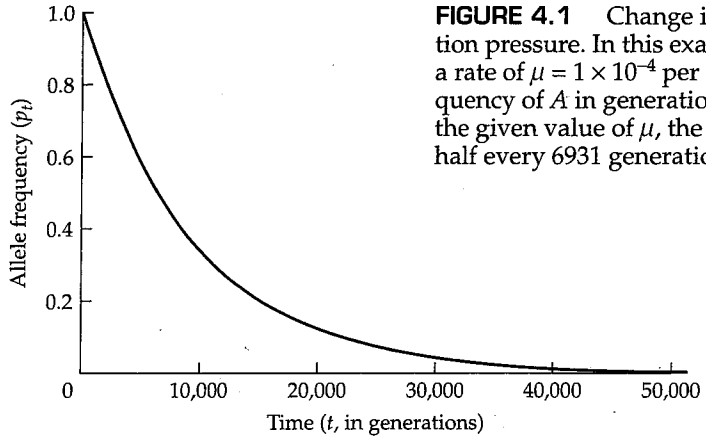


FIGURE 4.1 Change in allele frequency under mutation pressure. In this example, an allele A mutates to a at a rate of $\mu = 1 \times 10^{-4}$ per generation; p_t is the allele frequency of A in generation t. We assume that $p_0 = 1$. With the given value of μ, the allele frequency decreases by half every 6931 generations.

However, by the same reasoning, p_{t-1} includes all A alleles in generation $t-1$ that did not mutate in that generation, and so $p_{t-1} = p_{t-2} \times (1-\mu)$. Substituting this equation into the one above yields

$$p_t = p_{t-2}\,(1-\mu)^2$$

Continuing in the same manner leads eventually to

$$p_t = p_0\,(1-\mu)^t \qquad [4.1]$$

The effect of mutation pressure on allele frequency is illustrated in Figure 4.1 for the case $\mu = 10^{-4}$. The allele frequency of A decreases very slowly, almost linearly at first because the governing term in Equation 4.1, $(1-\mu)^t$, is approximated by $1 - \mu t$ when t is sufficiently small. After 1000 generations, the allele frequency of A is still 0.90; however, at $t = 10{,}000$ generations, $p_t = 0.37$; and at $t = 20{,}000$ generations, $p_t = 0.14$.

One instructive way to analyze Equation 4.1 is to consider the time required to reduce the allele frequency of A by half. To find the "half-life" of the process, set $p_t = 0.5 \times p_0$; this relationship implies that $0.5 = (1-\mu)^t$. Taking logarithms of both sides, we obtain

$$t_{\frac{1}{2}} = \ln(0.5)/\ln(1-\mu) \approx 0.6931/\mu$$

In the example in Figure 4.1, $t_{\frac{1}{2}} = 6931$ generations. A decrease in μ by a factor of 10 increases $t_{\frac{1}{2}}$ accordingly, to approximately 69,310 generations for $\mu = 10^{-5}$ and to approximately 693,100 generations for $\mu = 10^{-6}$. The feeble effect of mutation pressure alone in changing allele frequency is illustrated by the long half-lives calculated for realistic values of the mutation rate.

As noted with reference to Equation 4.1, the approximation $p_t = p_0(1-\mu t)$ is quite accurate for small values of t. With respect to the allele frequency of the mutant allele a, the approximation can also be written as $q_t = q_0 + \mu t$, pro-

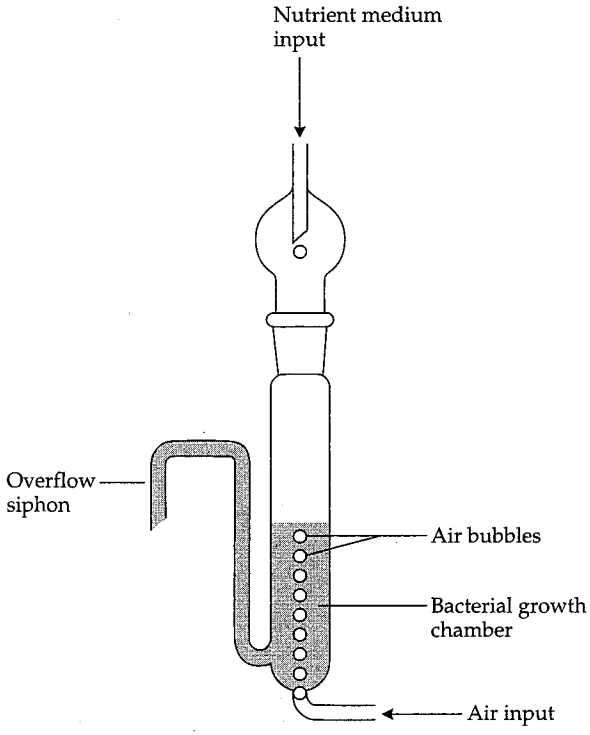


FIGURE 4.2 Diagram of a bacterial chemostat. Nutrient medium drips in at the top, but a constant volume is maintained by means of an overflow siphon. The air coming in at the bottom provides oxygen. At the steady state, the rate of inflow of nutrient equals the rate of outflow. Cells within the chemostat are in a continuous state of division, but the population does not increase in size because, in any interval of time, the number of new cells produced by division is balanced by the number washed out through the siphon.

vided that q_0 is small. This approximation implies that the allele frequency of the a allele increases linearly with time with a slope equal to μ. Because μ is small, however, the linear increase in q_t is difficult to detect experimentally except in very large populations. A large population size can be attained in a bacterial **chemostat**, which is a device for maintaining a population of bacteria in a continuous state of growth and cell division (Figure 4.2). The linear increase in q_t from mutation pressure observed in a chemostat is shown in Figure 4.3. Note the abrupt increase in mutation rate (indicated by the increase in slope) shortly after the addition of caffeine, a bacterial mutagen.

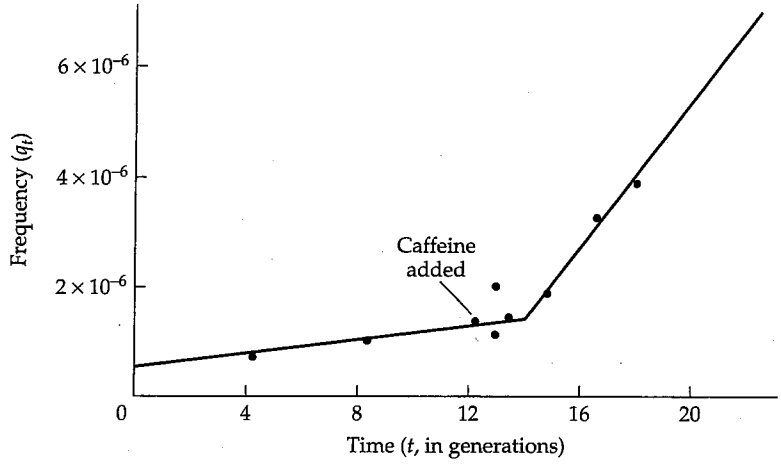


FIGURE 4.3 Estimation of mutation rate in a bacterial chemostat. This example concerns the rate of mutation of a gene in *Escherichia coli* that confers resistance to infection by the bacteriophage T5. The frequency q_t is the frequency of T5-resistant cells after t generations of growth. The mutation rate is estimated as the slope of the straight-line segments. Prior to the addition of caffeine, the slope was $\mu = 7.2 \times 10^{-8}$ per generation. After addition of caffeine at a concentration of 150 mg/l, the slope increased about tenfold to $\mu = 66 \times 10^{-8}$ per generation. In this experiment, the generation time was 5.5 hours. (From Novick 1955.)

PROBLEM 4.1 A genetic factor has been described in *Drosophila mauritiana* that results in the spontaneous deletion of the transposable genetic element *mariner* at a frequency of approximately 1% per generation for each copy (Hartl 2001). In a population containing an autosomal site at which a *mariner* insertion is fixed (homozygous), how many generations would be required for the frequency of flies that are homozygous for a deletion of the element to exceed 5%? Assume that the population is large, that mating is random, that the excision factor is fixed, and that deletion of the element does not affect survival or reproduction.

ANSWER Let p_t be the frequency of chromosomes in which the *mariner* element remains undeleted in generation t, and let $\mu = 0.01$ be the probability of deletion of the element per generation. For this situation, Equation 4.1 applies with $\mu = 0.01$ and $p_0 = 1$. The frequency of deletion homozygotes is greater than 5% when $(1 - p_t)^2 > 0.05$, or $p_t < 1 - \sqrt{(0.05)} = 0.776$. Thus, t should be greater than $\ln(0.776)/\ln(0.99) = 25.2$ generations.

Reversible Mutation

In addition to forward mutation of A to a, the model can also allow reverse mutation from a to A. Mutation pressure on the allele frequency p therefore pushes in both directions: Forward mutation tends to decrease p, reverse mutation tends to increase p. Eventually, an equilibrium is reached in which the frequency p remains constant from generation to generation. At this point, the loss of A alleles from forward mutation is exactly offset by the gain of A alleles from reverse mutation.

To deduce the point of equilibrium, suppose that the rate of forward mutation from A to a is μ per generation and that the rate of reverse mutation from a to A is v per generation. Let p_t and q_t denote the allele frequencies of A and a in generation t, so that $p_t + q_t = 1$. An A allele in generation t can originate in either of two ways. It could have been an A allele in generation $t - 1$ that escaped mutation to a (which happens with probability $1 - \mu$), or it could have been an a allele in generation $t - 1$ that mutated to A (which happens with probability v). In symbols,

$$p_t = p_{t-1}(1-\mu)+(1-p_{t-1})v \tag{4.2}$$

To solve this equation for p_t, note that Equation 4.2 can be written in the form

$$p_t - \frac{v}{\mu+v} = \left(p_{t-1} - \frac{v}{\mu+v}\right)(1-\mu-v) \tag{4.3}$$

Because the relation between p_{t-1} and p_{t-2} is the same as that between p_t and p_{t-1}, the solution to Equation 4.3 is obtained by successive substitutions as

$$p_t - \frac{v}{\mu+v} = \left(p_0 - \frac{v}{\mu+v}\right)(1-\mu-v)^t \tag{4.4}$$

To understand what happens to the allele frequency in the long run, consider Equation 4.4 in the case when t is very large, for example 10^5 or 10^6 generations. Even though $1 - \mu - v$ is ordinarily close to 1, the value of t eventually becomes so large that $(1 - \mu - v)^t$ becomes approximately 0. Thus, the whole right-hand term in Equation 4.4 goes to 0, and so p_t eventually attains a value that remains the same generation after generation. Such a value of p is called an **equilibrium** value, which we will denote by $\hat{p}$. In case of reversible mutation, the equilibrium is found by equating the left-hand side of Equation 4.4 to 0, and therefore

$$\hat{p} = \frac{v}{\mu+v} \tag{4.5}$$

There is a simple intuitive explanation as to why Equation 4.5 gives the equilibrium with reversible mutation. Since A alleles become a alleles at a rate μ per generation, and a alleles revert to A alleles at a rate v per generation, at equilibrium one might expect the ratio of allele frequencies to be the recipro-

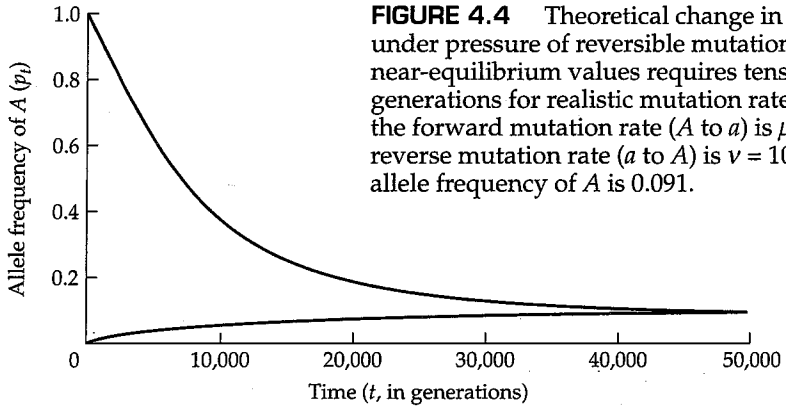


FIGURE 4.4 Theoretical change in allele frequency under pressure of reversible mutation. The attainment of near-equilibrium values requires tens of thousands of generations for realistic mutation rates. In this example, the forward mutation rate (A to a) is $\mu = 10^{-4}$ and the reverse mutation rate (a to A) is $\nu = 10^{-5}$. The equilibrium allele frequency of A is 0.091.

cal of the ratio of the mutation rates, or $\hat{p}/\hat{q} = \nu/\mu$. Because $\hat{q} = 1 - \hat{p}$, this reasoning implies that $\hat{p} = \nu/(\mu + \nu)$, as verified in Equation 4.5.

The manner in which p_t converges to its equilibrium value is shown in Figure 4.4 for the case $\mu = 10^{-4}$ and $\nu = 10^{-5}$. Note that, whatever the initial frequency of A, the allele frequency of A eventually goes to $\hat{p}$, which in this example equals $0.00001/(0.0001 + 0.00001) = 0.091$. Figure 4.4 also indicates that mutation pressure is usually very weak in changing allele frequency, inasmuch as the population requires thousands or tens of thousands of generations to reach equilibrium.

PROBLEM 4.2 The bacterium *Salmonella enterica* has a genetic switching mechanism that regulates the production of alternative forms of a protein component of the cellular flagella. There are two alleles, which we will call A (for the "specific-phase" flagellar property) and a (for the "group-phase" flagellar property). Switching back and forth between A and a takes place rapidly enough that Equation 4.4 can be applied. The transition from A to a has a rate of $\mu = 8.6 \times 10^{-4}$ per generation, and that of a to A has a rate of $\nu = 4.7 \times 10^{-3}$ per generation. These rates are orders of magnitude larger than mutation rates typically observed for other genes. The reason is that the change from A to a and back again does not result from mutation in the conventional sense but from intrachromosomal recombination (Simon et al. 1980). Formally, however, we can treat the system as one with reversible mutation. In cultures initially established with the frequency of A at $p_0 = 0$, Stocker (1949) found that the frequency increased to $p = 0.16$ after 30 generations and to $p = 0.85$ after 700 generations. In cultures initiated with $p_0 = 1$, the frequency decreased to 0.88 after 388 generations and to 0.86 after 700 generations. How do these values agree with those calculated from Equation 4.4 using the estimated mutation rates? What is the predicted equilibrium frequency of A?

ANSWER Note that $v/(\mu + v) = 0.845$. This is the predicted equilibrium frequency (Equation 4.5). Also, $1 - \mu - v = 0.99444$, and this quantity determines the rate of approach to equilibrium. For the cultures with $p_0 = 0$, the predicted values are $p_{30} = 0.845 - (0.845)(0.99444)^{30} = 0.13$ and $p_{700} = 0.845 - (0.845)(0.99444)^{700} = 0.83$. For the cultures with $p_0 = 1$, the predicted values are $p_{388} = 0.845 + (0.155)(0.99444)^{388} = 0.86$ and $p_{700} = 0.845 + (0.155)(0.99444)^{700} = 0.85$. The predicted values are in very good agreement with the observations.

4.2 MUTATION AND RANDOM GENETIC DRIFT

The assumption of a virtually infinite population size is often unrealistic. An improved model makes the population size finite, in which case the change in the frequency of a mutant allele depends not only on the mutation pressure but also on random sampling from generation to generation. The random sampling results in chance changes in allele frequency, a process called *random genetic drift* discussed in some detail in Chapter 3. To understand the effects of random genetic drift when combined with mutation, consider the diagram in Figure 4.5. The squares represent the $2N$ alleles in the adult population in generation t. Each allele is assigned a unique label—$\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{2N}$—to temporarily mask its identity as either A or a. The circles represent the essentially infinite pool of gametes in generation t. In the gamete pool, each labeled allele has a frequency of $1/(2N)$. The squares at the bottom represent two diploid genotypes in generation $t + 1$ formed by random sampling from the pool of gametes. By chance, the two alleles forming a genotype may be replicas of the same allele in the previous generation, for example, $\alpha_i\alpha_i$. Alternatively, the two alleles forming a genotype may come from different alleles in the previous generation, for example, $\alpha_i\alpha_j$.

The random sampling from the gamete pool means that some alleles may be overrepresented in generation $t + 1$, relative to their frequency in generation t, and some alleles may be underrepresented. Indeed, any particular allele has a good chance of being unrepresented in generation $t + 1$, and hence the lineage of that allele is terminated. To be precise, in a population of constant size, each allele in generation t has a chance of approximately $e^{-1} = 0.368$ of not being represented in generation $t + 1$. To understand why, consider the allele designated α_1. The frequency of α_1 in the gamete pool is $1/(2N)$, and the frequency of all other alleles together is therefore $1 - 1/(2N)$. Because the genotypes in generation $t + 1$ are formed by the random selection of $2N$ alleles from the pool of gametes, the distribution of the number of α_1 and non-α_1 alleles present in generation $t + 1$ is given by successive terms in the binomial distribution:

$$\left[\frac{1}{2N}\alpha_1 + \left(1 - \frac{1}{2N}\right)\alpha\right]^{2N} \tag{4.6}$$

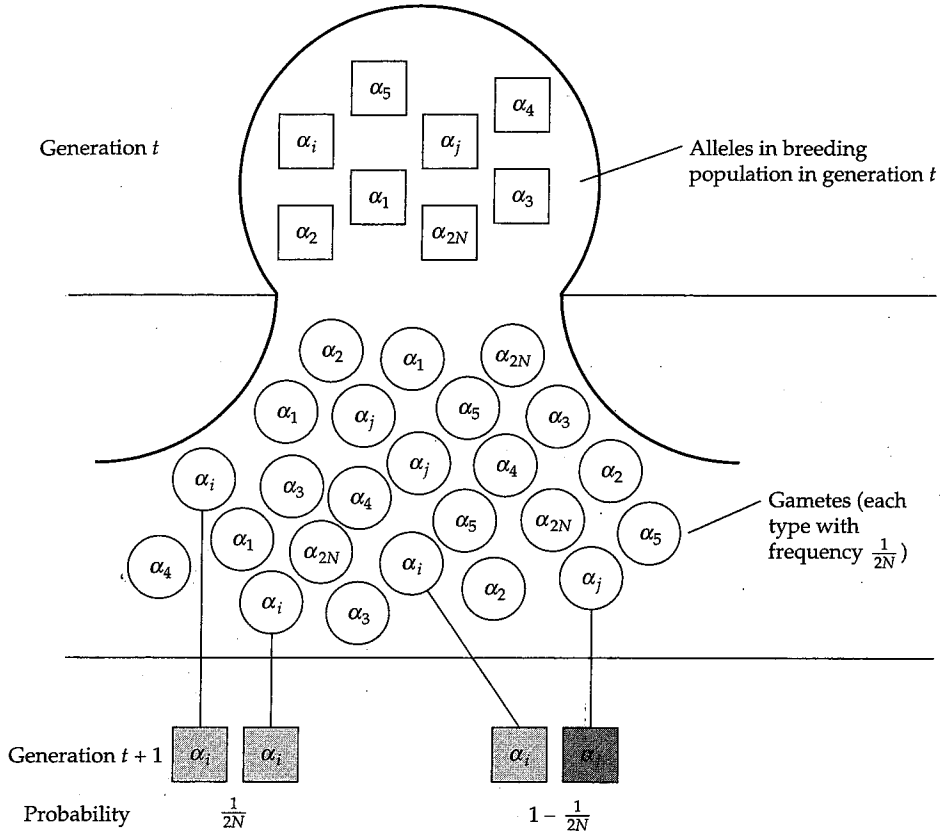


FIGURE 4.5 Random sampling of alleles in a finite population increases the probability of identity by descent. Two randomly chosen alleles, illustrated in the squares at the bottom, may be identical by descent either because they are replicas of the same allele in the immediately preceding generation ($\alpha_i\alpha_i$) or because they are replicas of the same allele in a more remote generation ($\alpha_i\alpha_j$).

in which α represents the collection of all alleles other than α_1. Hence, the probability that α_1 is not represented in generation $t + 1$ is given by

$$\left(1-\frac{1}{2N}\right)^{2N} \approx e^{-1} = 0.368 \qquad [4.7]$$

The approximation is very good even when N is quite small. For example, when $N = 10$, the left-hand side of Equation 4.7 equals 0.358, and, when $N = 20$, the left-hand side equals 0.363.

The important implication of Equation 4.7 is that, owing to random genetic drift, the ancestral lineage of each allele faces a substantial risk of

extinction in each generation. As time goes on, the lineages progressively disappear, one or a few at a time. Eventually, a time is reached at which all lineages except one have become extinct. At that time, every allele in the population is identical by descent with a particular allele present in an ancestral population.

Probability of Fixation of a New Neutral Mutation

The ultimate extinction of all but one lineage implies the answer to the question: What is the probability that a single new mutation eventually becomes fixed in a population of size $2N$? One approach to this problem is illustrated in Figure 4.6. Parts A and B show all the alleles present in the current generation, immediately after a new mutation (shaded circle) has been created. After a sufficient number of generations have passed, each of the alleles in

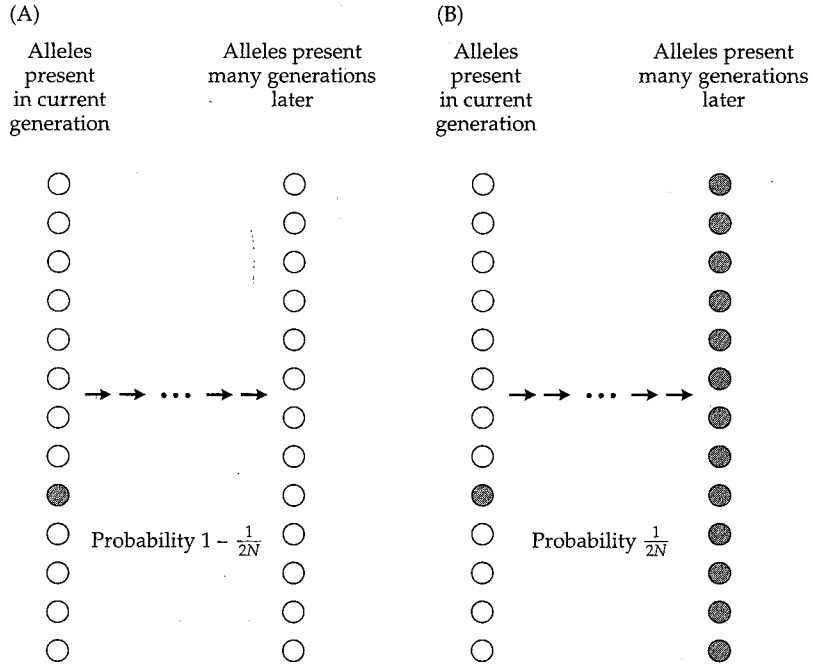


FIGURE 4.6 In a finite population, the lineages of all alleles must trace back to a single allele in some ancestral population. Here, a particular allele of interest in a diploid population of size N is indicated by the shaded circle. (A) The probability the designated allele is not destined to be the common ancestor of all alleles many generations in the future is $1 - 1/(2N)$. (B) The probability the designated allele is destined to be the common ancestor of all alleles many generations in the future is $1/(2N)$. Hence, the probability of ultimate fixation of a newly arising neutral allele is $1/(2N)$.

the descendant population will descend from a single allele, chosen at random, in the current population. In part A, the descendant alleles all derive from one of the nonmutant alleles in the current population; the nonmutant alleles have frequency $1 - 1/(2N)$, and so this is the probability of ultimate fixation of a nonmutant allele. In part B, the descendant alleles all derive from the mutant allele, and so $1/(2N)$ is the probability of ultimate fixation of a new mutant allele. More generally, for neutral alleles, which do not affect the survival or reproduction of the organism, the probability of ultimate fixation of a selectively neutral allele in a finite population is equal to the frequency of the neutral allele in the initial population.

The result that a new neutral mutation has a probability of ultimate fixation of $1/(2N)$ was already deduced in Chapter 3 in two different ways. One was through reasoning based on coalescence; Figure 4.6B, when examined from right to left, is a sort of coalescent diagram without the branches, in which all the shaded alleles at the right coalesce to the shaded ancestral allele at the left. The other approach to this problem in Chapter 3 was through the Kolmogorov backward equation (see Problem 3.5). In Chapter 3 we also used this approach to show that, among the lucky few neutral alleles that are eventually fixed, the process takes a long time—on the average, $4N$ generations.

4.3 THE NEUTRAL THEORY OF MOLECULAR EVOLUTION

The hypothesis that many genetic polymorphisms result from selectively neutral alleles maintained by a balance between the effects of mutation and random genetic drift is known as the **neutral theory** (Kimura 1968a; King and Jukes 1969). Mutation introduces new alleles into a population, and random genetic drift determines whether a neutral allele will ultimately be fixed or lost, though loss is the usual outcome. At equilibrium, there is a balance between mutation and random genetic drift, so that, on the average, each new allele gained by mutation is balanced against an existing allele that is lost.

In essence, the neutrality hypothesis states that many mutations have so little effect on the organism that their influence on survival and reproduction is negligible. The frequencies of neutral alleles are not, therefore, determined by natural selection. Consequently, if the neutrality hypothesis is true, then many polymorphisms may have no particular significance in the adaptation of a species to its environment. From the perspective of adaptation, selectively neutral polymorphisms are mere evolutionary "noise" and, regardless of how much their study may reveal about population structure and random genetic drift, they tell us little or nothing about adaptive genetic changes in evolution. Kimura (1968a) gave the irony a positive spin by noting that "if my chief conclusion [about the prevalence of neutral alleles] is correct, then we must recognize the great importance of random genetic drift . . . in forming the genetic structure of biological populations." Quite so. Although neu-

tral alleles may be unsuitable for the study of genetic adaptation, the very fact that they are invisible to natural selection makes them ideal for mapping the geographical structure of populations and for tracing the ancestral lineages of DNA sequences to make inferences about the genealogical relationships within and among species.

Because the neutrality hypothesis is of fundamental importance in population genetics and evolution, it has been a subject of considerable discussion (for example, Li 1997; Graur and Li 2000; Hartl 2000a; Nei and Kumar 2000; Gillespie 2004). The neutrality hypothesis was put forward in the late 1960s at a time when most of the genome was supposed to have a protein-coding function. Introns and other noncoding sequences were unknown. Today it is clear that only about 1.5% of the mammalian genome codes for proteins. The low coding density affords ample scope for mutations that have little or no effect on fitness, including some (but by no means all) mutations in introns, pseudogenes, spacers between genes, noncoding DNA in the centromeric region of chromosomes, and so forth.

4.4 THE INFINITE-ALLELES MODEL

Many genes have more than two alleles represented among the organisms in a natural population. It is therefore of some importance to determine the expected level of genetic variation under mutation pressure. A convenient measure of genetic variation is the proportion of heterozygous genotypes (the *heterozygosity*). If a gene has a greater heterozygosity than expected from mutation pressure alone, then other forces that operate in nature must tend to preserve genetic variation. On the other hand, if a gene has a smaller heterozygosity than expected, then other forces must tend to eliminate genetic variation.

The heterozygosity of a gene is a function of the number of alleles and their relative frequencies. In principle, the number of alleles of any gene could be very large. For example, a gene coding for a protein of 300 amino acids has a coding sequence 900 nucleotides in length. Because each nucleotide site could be occupied by either an A, T, G, or C, the total number of possible alleles is 4^{900}, which equals about 10^{542}. Hence, we can suppose that every new mutation creates an allele that does not already exist in the population. This is called the **infinite-alleles model** of mutation. The infinite-alleles model is but one way to specify the characteristics of new mutations. Although it represents a somewhat simplified view of mutation, it nevertheless provides a useful standard of comparison for other models or for observed allele frequencies.

In the infinite-alleles model, two alleles that are identical in sequence must also be identical by descent because of the assumption that each mutation creates a unique allele. The concept of **identity by descent** may be made clear with reference to Figure 4.5, where each allele is assigned a unique iden-

tifier, α_1, α_2, α_3, and so forth. At the bottom, the alleles in the genotype $\alpha_i\alpha_i$ on the left are considered as identical by descent because they descend from a single ancestral allele by DNA replication in a previous generation. In this case the DNA replication that produced $\alpha_i\alpha_i$ occurred in the immediately preceding generation. Also in Figure 4.5, the alleles in the genotype $\alpha_i\alpha_j$ might be identical by decent. The different subscripts imply only that they did not derive by DNA replication in the immediately preceding generation, but if they derived by DNA replication in some earlier generation, they are nevertheless identical by descent.

In the literature of population genetics, a genotype in which the alleles are identical by decent is sometimes said to be **autozygous**, and one in which the alleles are not identical by descent is said to be **allozygous**. There is some ambiguity in the concept of identity by descent, because the coalescent process implies that every allele of a gene must ultimately derive from DNA replication of a single ancestral allele at some time in the possibly ancient past. In practice the ambiguity is resolved by picking some reference point in time in the past and arbitrarily declaring that at this point in time no allele is identical by descent with any other.

In the infinite-alleles model, in which each mutation produces a new allele not present in the population, homozygous genotypes must have alleles that are autozygous (identical by descent). To measure the homozygosity, therefore, we need only to calculate the autozygosity. This can agan be done with reference to the finite-population model in Figure 4.5. Define F_t as the probability that, in generation t, two alleles randomly chosen from a population are identical by descent. In the context of Figure 4.5, the randomly chosen alleles are combined in pairs to make genotypes, and so F_t is also the probability of autozygosity in generation t. We will use the $\alpha_i\alpha_i$ and $\alpha_i\alpha_j$ genotypes in generation t in Figure 4.5 to derive an expression for F_t in terms of F_{t-1}, N, and the mutation rate μ. First, consider the genotype $\alpha_i\alpha_i$. What is the probability that this genotype has alleles that are identical by descent? The alleles must be identical by descent provided that neither allele has mutated in the course of one generation, and so the probability of identity by descent in this case is $(1-\mu)^2$. Now consider the genotype $\alpha_i\alpha_j$. These alleles are identical by descent only if two randomly chosen alleles in generation $t-1$ are identical by descent and if neither allele mutated in the course of one generation. Therefore, the probability of identity by descent in this case is $F_{t-1}(1-\mu)^2$. Because each of the labeled α's in Figure 4.5 has the same frequency in the gamete pool, namely $1/(2N)$, the probability of a combination like $\alpha_i\alpha_i$ is $1/(2N)$ and the probability of a combination like $\alpha_i\alpha_j$ is $1-1/(2N)$. Putting all this together, the recurrence equation for F_t is

$$F_t = \left(\frac{1}{2N}\right)(1-\mu)^2 + \left(1-\frac{1}{2N}\right)(1-\mu)^2 F_{t-1} \tag{4.8}$$

Eventually an equilibrium value of F, call it $\hat{F}$, is attained in which the increase in autozygosity from random genetic drift in any generation is exactly offset by the decrease in autozygosity from new mutations. The equilibrium can be found by equating $F_t = F_{t-1} = \hat{F}$ in Equation 4.8 and solving. Ignoring terms in μ^2 and those in μ/N because they are expected to be negligibly small, the solution is

$$\hat{F} = \frac{1}{1+4N\mu} \tag{4.9}$$

to an excellent approximation. Therefore, the number of selectively neutral alleles increases under mutation pressure until $\hat{F}$ satisfies Equation 4.9. Since it is the equilibrium value of the probability of identity by descent, $\hat{F}$ is also the equilibrium value of the autozygosity. Because of the assumption in the infinite-alleles model that each allele in the population arises only once, all genotypes that are homozygotes must also be autozygous. Therefore, $\hat{F}$ can also be interpreted as the equilibrium value of the proportion of homozygous genotypes.

In Equation 4.9, N should be interpreted as the effective population size, N_e, defined in Chapter 3 as the size of an ideal population that has the same rate of increase in homozygosity as the population in question. In population genetics, the usual symbol for $4N_e\mu$ is θ, so that $\theta = 4N_e\mu$, and Equation 4.9 can be rewritten as

$$\hat{F} = \frac{1}{1+\theta} = \frac{1}{1+4N_e\mu} \tag{4.10}$$

Because any genotype that is not homozygous is heterozygous, it follows that the proportion of heterozygous genotypes in a population is given by $1-\hat{F}$. In the infinite-alleles model, therefore, the heterozygosity is given by Equation 4.10 as

$$1-\hat{F} = \frac{\theta}{1+\theta} = \frac{4N_e\mu}{1+4N_e\mu} \tag{4.11}$$

where again $\theta = 4N_e\mu$. Figure 4.7 shows the equilibrium homozygosity $[1/(1+\theta)]$ and heterozygosity $[\theta/(1+\theta)]$ for a range of values of $\theta = 4N_e\mu$. The illustration shows that there is a rather narrow range of $4N_e\mu$ over which an intermediate level of genetic variation (heterozygosity) is maintained. For example, the equilibrium heterozygosity is in the range 0.2 to 0.8 only when $4N_e\mu$ is in the range 0.25 to 4. In reality, however, as is clear from Figure 1.8, the heterozygosity for electrophoretic variants in protein molecules is smaller than 0.2, and often much smaller than 0.2 (in mammals it is about 0.03).

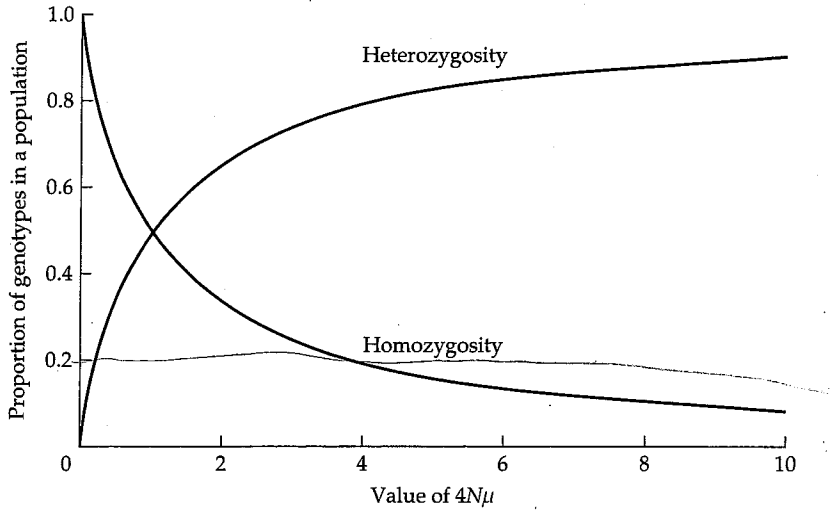


FIGURE 4.7 Plot of average homozygosity and average heterozygosity for the infinite-alleles model. Intermediate values of heterozygosity are maintained over only a small range of $\theta = 4N_e\mu$.

This means that the range of realistic values in Figure 4.7 is confined to the extreme left of the graph, where θ is substantially smaller than 1. In fact, for the protein electrophoresis data in Figure 1.8, the estimated range of θ is 0.03–0.16. In other words, the maximum estimated value of $4N_e\mu$ differs from the minimum by a factor of only about five. This is quite unexpected, inasmuch as the population number in different species can differ by factor of 10^4 or more. The apparently too narrow range of $\theta = 4N_e\mu$ among diverse organisms has been interpreted as implying that the neutrality hypothesis is simply wrong for amino acid polymorphisms (Gillespie 1991). On the other hand, estimates of the effective population number in natural populations are generally imprecise because the studies are very difficult, and estimates of μ, which in this case is the mutation rate to *neutral* amino acid polymorphisms, are even more uncertain. However, the actual distribution of allele frequencies in populations suggests that many segregating amino acid polymorphisms present at low frequency are mildly deleterious and maintained by mutation pressure.

PROBLEM 4.3 It is an oddity that Equations 4.10 and 4.11 predict the equilibrium homozygosity and heterozygosity without explicit reference to the allele frequencies. If the allele frequencies are estimated in a random-mating population, however, then the homozygosity and heterozygosity can be estimated using the Hardy-Weinberg principle from Chapter 2. In particular, the homozygosity is estimated as $\hat{F} = \Sigma p_i^2$, where the sum is over all allele frequencies p_i, and the heterozygosity is estimated as $1 - \hat{F}$. A study electrophoretic protein variants in a Caribbean population of *Drosophila willistoni* (Ayala and Tracy 1974) yielded the following estimated allele frequencies for the loci *Adk-1* (adenylate kinase-1), *Lap-5* (leucine amino peptidase-5), and *Xdh* (xanthine dehydrogenase).

	Adk-1	*Lap-5*	*Xdh*
Allele 1	0.574	0.801	0.446
Allele 2	0.309	0.177	0.406
Allele 3	0.114	0.014	0.092
Allele 4	0.003	0.004	0.034
Allele 5	—	0.004	0.014
Allele 6	—	—	0.004
Allele 7	—	—	0.002
Allele 8	—	—	0.002

Estimate the homozygosity and heterozygosity for each gene, and give the corresponding estimate of θ.

ANSWER The homozygosity estimates are 0.438 for *Adk-1*, 0.673 for *Lap-5*, and 0.373 for *Xdh*, and the corresponding heterozygosities are 0.562, 0.327, and 0.626. Because the equilibrium homozygosity $\hat{F}$ equals $1/(1 + \theta)$ [see Equation 4.10], then θ can be estimated as $(1 - \hat{F})/\hat{F}$, which equals the ratio of the heterozygosity to the homozygosity. For these three genes the estimates of θ are 1.28, 0.49, and 1.68, respectively. These value are substantially greater than the average for electrophoretic polymorphisms in *Drosophila* (see Figure 1.8), which is about $\theta = 0.16$.

The Ewens Sampling Formula

Equation 4.11 shows that the infinite-alleles model has an equilibrium when the heterozygosity equals $\theta/(1 + \theta)$. This is not an "equilibrium" in the usual sense, meaning the absence of change. In reality, it is a dynamic state in which allele frequencies are always changing, new mutations continue to come into the population, alleles previously present are lost, and even alleles that might at one time have been fixed are subject to eventual loss. The term *steady state* is more appropriate for this kind of situation, since the alleles are not maintained at a constant frequency, but rather new ones enter and old ones are lost from the population. The population remains at a steady state in the sense that the number of alleles and the homozygosity (autozygosity in the infinite-alleles model) remain stationary. But if the number of alleles and the level of autozygosity are in a steady state, then it is reasonable to assume that there is also a steady-state distribution of allele frequencies. When there are multiple alleles, the joint distribution of the allele frequencies present in a population is often called the **allele-frequency spectrum** in the

population. When the allele-frequency spectrum is at a steady state, this implies that the most common allele always has a frequency of p_1, and the next most common has a frequency of p_2, and so on. The steady-state allele-frequency spectrum has the curious property that, even though the most common allele is expected to have a frequency of p_1, the *identity* of the most common allele will change with time. In the steady-state population, not all alleles are equally frequent, and F is greater than it would be were all alleles equally frequent.

Consider now the steady-state allele-frequency spectrum from the point of view of an experimenter taking a sample from a population. Let the sample size be n genes, and suppose there are k different alleles in this sample. For example, a sample of size $n = 20$ might consist of $k = 10$ unique alleles, with one allele present six times in the sample, one allele represented four times, two alleles each represented twice, and six alleles each represented only once. Such a description of the sample is called the **allelic configuration** of the sample. A remarkable finding of Ewens (1972) was that the expected allelic configuration of a sample drawn from a steady-state population obeying the infinite-alleles model under neutral mutation and random genetic drift (measured as $\theta = 4N\mu$) is completely determined by the sample size n and the number of observed alleles k. In particular, Ewens (1972) showed that the expected number k of alleles in a sample of size n is a simple function of θ:

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1} \tag{4.12}$$

If θ is very small, $E(k) \approx 1$, whereas for very large θ, $E(k)$ approaches n, implying that for a large enough population with a high enough mutation rate, every allele that is sampled will be different. The form of Equation 4.12 suggests that, as the sample size increases, more alleles will be found, but that there is a diminishing return in finding new alleles when the sample size increases. When $E(k)$ is plotted against θ (Figure 4.8), the increase in the expected number of alleles is greatest for larger sample sizes when the population is highly diverse (large θ).

The infinite-alleles model gives a steady-state prediction of F given θ (because $F = 1/(1 + \theta)$ from Equation 4.10), and a prediction of k from Equation 4.12. Combining these predictions, the expected relation between F and k is plotted in Figure 4.9. The hyperbolic relation is not surprising, because a population with many alleles will generally have a lower probability of identity of a randomly chosen pair of alleles. For $\theta = 1$, the expected F is $\frac{1}{2}$ for all sample sizes, but a larger sample size should yield a greater number of distinct alleles. The curves are not dramatically different for different samples sizes (n), mainly because an increase in sample size reveals a greater number of low-frequency alleles, and low-frequency alleles do not contribute very much to the homozygosity F.

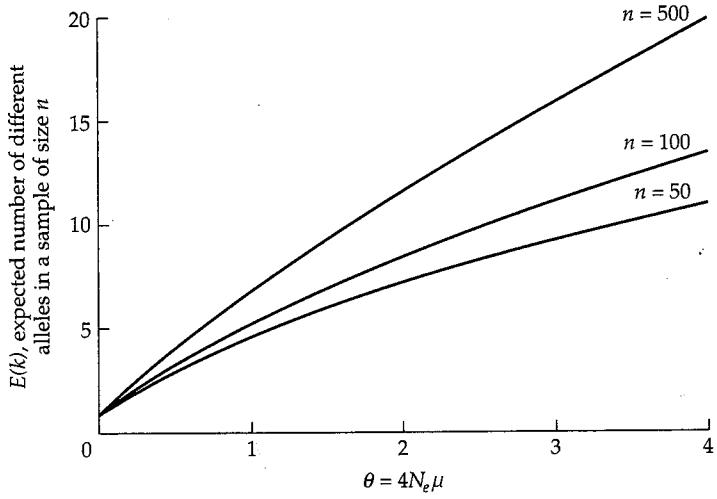


FIGURE 4.8 Relation between θ, the expected number of alleles, and the sample size according to the Ewens sampling theory of a population in steady state under the infinite-alleles model of neutral mutation.

Making use of Ewen's result, Karlin and McGregor (1972) found an explicit formula for the expected allele frequency configuration in samples. In particular, they showed that the probability that a sample of size n containing k distinct alleles will contain exactly n_1 alleles of type 1, n_2 alleles of type 2, $\cdots$, n_k alleles is of type k, is given by

$$\Pr\{n_1, n_2, \cdots, n_k, k\} = \frac{n!\theta^k}{k!n_1n_2\cdots n_kS_n(\theta)} \tag{4.13}$$

where $S_n(\theta) = \theta(\theta + 1)(\theta + 2)\cdots(\theta + n - 1)$. This equation provides the basis for comparing the allele configuration observed in samples with those expected under the infinite-alleles model with neutrality. [See Ewens (2004) for additional and more advanced discussion.]

The Ewens-Watterson Test

The Ewens (1972) paper is one of the landmarks in the history of population genetics. Because it afforded explicit predictions of expected allele configurations in samples assuming only neutral alleles, these predictions could be compared with actual observations to test the neutral theory. Based on the observed and expected configurations, a number of test statistics can be devised to determine whether an observed sample fits the expected values of the neutral model. Tests based on the infinite-alleles model are most appro-

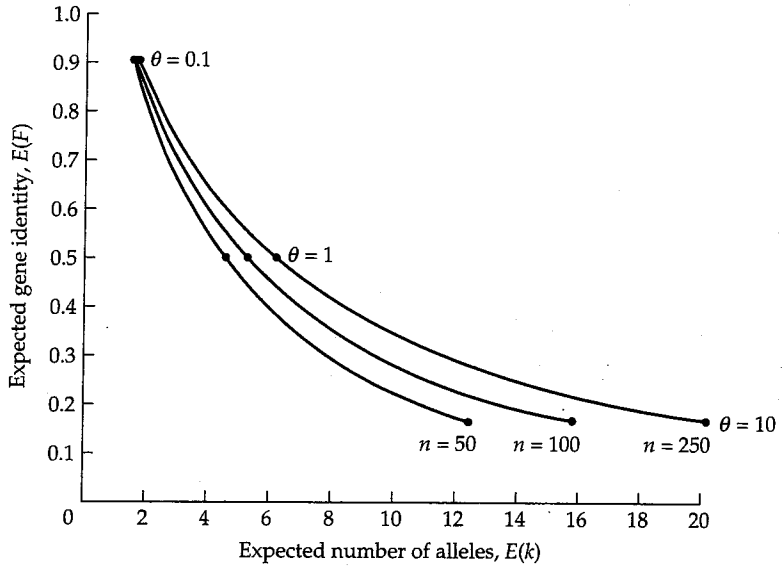


FIGURE 4.9 The infinite-alleles model prediction of the relation between the expected number of alleles $E(k)$ and the expected gene identity (homozygosity) F. The three curves represent a range of values of $\theta = 4N_e\mu$, starting at $\theta = 0.1$ in the upper left, and ending with $\theta = 10$ in the lower right. For the value of $\theta = 1$, the expected F, given by the relation $F = 1/(1 + \theta)$, is $\frac{1}{2}$, regardless of the sample size. Larger sample sizes always lead to larger expected numbers of alleles, but the difference is greater in more diverse populations (those with smaller F).

priate when a large number of alleles can be distinguished, for example, by protein electrophoresis when the actual differences in DNA sequence are unknown. In such situations, tests based on the Ewens sampling formula (see Equation 4.11) must assume that any alleles that cannot be distinguished must be identical by descent, and this can be a serious limitation.

To give the flavor of such tests, we consider two examples. One type of test compares the observed and expected allele frequency configurations in samples, using the Equation 4.13. Figure 4.10 shows histograms of the observed and expected configurations for polymorphic alleles in a human population, where each allele has a different number of copies of a short tandem repeat in a region of DNA. In this particular example there appears to be a slight excess of the common allele; this excess is consistent with any number of causes of departure from the infinite-alleles model.

A second test is based on an approach pioneered by Watterson (1978), which compares the observed homozygosity in a sample with that expected from Equation 4.13. In one study, a sample of 89 homozygous lines of *Drosophila pseudoobscura* were collected at the Gundlach-Bundschu Winery in

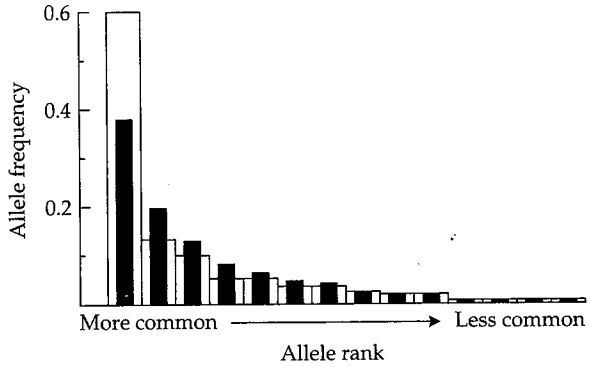


FIGURE 4.10 Observed (open columns) and expected (black bars) allele frequency spectrum of the *HRAS-1* gene in humans, identified by Southern blotting with the *pLM0.8* probe and *Taq*I digests. Observed data are from Baird et al. (1986). The expected distribution was generated using the Ewens sampling formula. In this sample of 490 genes there were 14 distinct alleles, four of which were present in just one individual. (From Clark 1988.)

Sonoma Valley, California (Keith et al. 1985). Homogenized tissue from each of these 89 lines was subjected to sequential electrophoresis (a sensitive means of detecting charge and conformation differences among the protein products), and the gels were treated to reveal differences in xanthine dehydrogenase (*Xdh*) mobility. The authors obtained a common allele that was present in 52 of the lines, one allele that was present in nine lines, one allele that was present in eight lines, two alleles present in four lines each, two alleles that were present in two lines each, and eight singleton or unique alleles.

To test whether the observed configuration fits the expectation, a computer simulation was run to generate realizations of samples from populations that obey the infinite-alleles model, focusing on simulated samples with the same number of alleles as the observed data. An algorithm to do this simulation is described by F. Stewart in Fuerst et al. (1977), but see also Manly (1985). From each computer-generated sample, F was calculated as the sum of the squared allele frequencies. Figure 4.11 shows a histogram of the computer-generated distribution of F, along with an arrow showing where the *Drosophila* sample fell. The sample had an observed F that fell in the upper tail of the distribution, and since so few values of F from the neutral hypothesis were larger than the observed F, the researchers rejected the neutrality hypothesis and argued that the data did not fit the infinite-alleles model satisfactorily. The departure was in the direction of excess homozygosity, but since the populations were probably in Hardy-Weinberg proportions, a clearer way to state the result would be to say that there was a deficiency of heterozygotes for the given number of observed alleles. The deficit means that the common allele is more common than expected, and there are also more rare alleles than expected. This pattern of frequencies is consistent with purifying selection acting to reduce the frequency of deleterious alleles that continually enter the population by mutation. It is also consistent with other scenarios, such as population growth. A growing population has more new mutant alleles than a population that is not growing (because a growing pop-

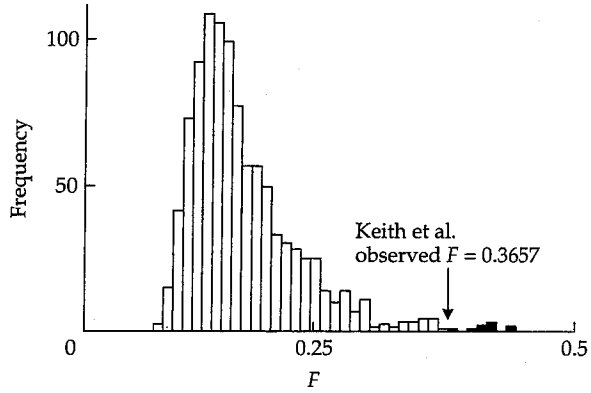


FIGURE 4.11 Computer-generated distribution of F obtained from 1000 samples from a population obeying the assumptions of the infinite-alleles model with $k = 15$ alleles and a sample of size $n = 89$. The mean of F from the simulation was 0.168, which is well below the observed F of 0.366 in the Gundlach-Bundschu sample (Keith et al. 1985). A significant departure of the observed F from the predictions of the model is noted by the small area under the tail of the distribution to the right of the arrow.

ulation has more allele copies at risk of mutating). Therefore, a growing population is expected to have an excess of low-frequency polymorphisms relative to a stable population.

The results of the Ewens-Watterson test can also be reported graphically as in Figure 4.12. Each gene yields a point specified by the number of distinct alleles and the observed F. The two curves represent the 95% confidence

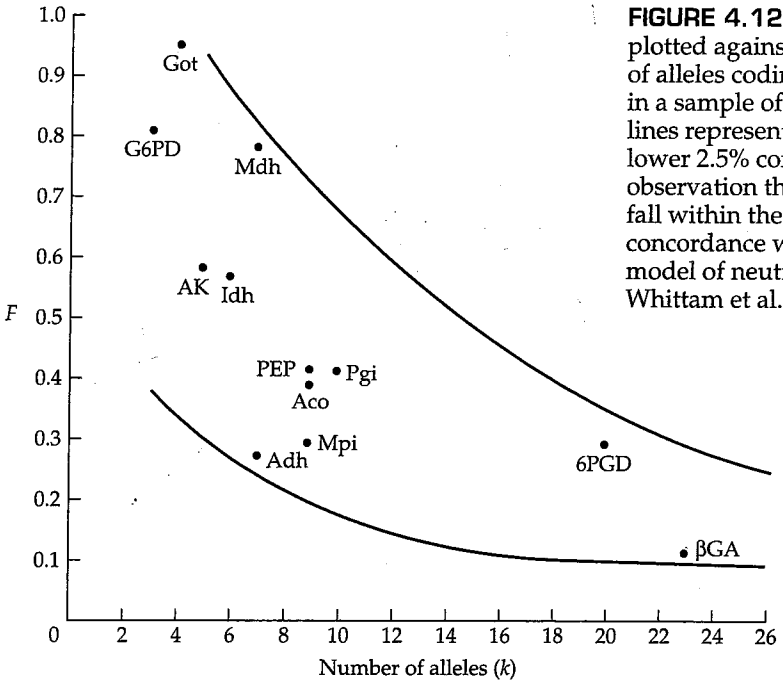


FIGURE 4.12 Gene identity (F) plotted against the observed number of alleles coding for various proteins in a sample of 279 *E. coli*. The solid lines represent the upper 97.5% and lower 2.5% confidence limits, and the observation that all of the tested loci fall within these limits suggests good concordance with the infinite-alleles model of neutral mutation. (From Whittam et al. 1983.)

interval generated by the Ewens sampling theory. A quick check of the concordance of the data with the model can be made by seeing whether points remain in this confidence region. Although *Xdh* in *Drosophila pseudoobscura* provides a dramatic departure from the infinite-alleles model, results like those plotted in Figure 4.12, which show an acceptable fit to neutrality, are more commonly obtained.

4.5 INFINITE-SITES MODEL

When DNA sequence data are available, then the infinite-sites model is intuitively more appealing than the infinite-alleles model. The **infinite-sites model**, we consider a very long sequence of nucleotide sites along a DNA molecule, and we assume that each mutation alters a single nucleotide site. This model directly addresses the type of data that molecular population geneticists can access. Moreover, allelic DNA sequences contain considerable information about the evolutionary history of the alleles, which is hidden in the patterns of similarity and differences. The infinite-sites model was first developed by Kimura (1969, 1971), who considered the nucleotide sites as unlinked, and by Watterson (1975), who took account of the nearly complete linkage among sites. If each mutation changes a different nucleotide site in a model with complete linkage between sites, then the infinite-sites model shares many properties with the infinite-alleles model. We discussed the infinite-sites model briefly in Chapter 1 to exemplify the types of inferences that can be made from the DNA sequences of alleles in a population, and again in Chapter 3 in the context of coalescence. In this section we consider the model in somewhat more detail and examine some of the statistical tests of neutrality based on its implications.

In a long sequence of nucleotides, if the mutation rate is sufficiently low, then most sites will be monomorphic, and all polymorphic sites will be segregating for just two nucleotides. Much of the available data on allelic variation in DNA sequence seems consistent with this view, because few nucleotide sites are segregating for more than two nucleotides. If the DNA sequence is sufficiently long and the frequency of polymorphic sites low, then most of the time new mutations will occur at sites that were previously monomorphic.

To reinforce these ideas, let us consider the very small sample of four aligned allelic DNA sequences shown in Table 4.1. These are made-up sequences, far shorter than would ever be used in practice, and much more diverse than usually found; they are intended to show as clearly as possible some of the types of information that can be extracted from such sequences. For ease of reading, in each column of aligned nucleotide sites, any nucleotide not matching the majority-rule consensus for that site is underlined. In regard to the infinite-alleles model, two types of information are usually extracted:

- The nucleotide sites in the sample that are occupied by two or more nucleotides. These are known as **segregating sites**, and in Chapters 1 and 3 we denoted the number of segregating sites as S. Among the four sample sequences *a–d* each of length 16 nucleotides, there are exactly 8 segregating sites (sites 1, 2, 4, 5, 9, 10, and 11, 12), and so $S = 8$.
- The nucleotide sites in the sample that differ between individual pairs of sequences. These are known as **nucleotide mismatches**, and in Chapter 1 we denoted the average number of nucleotide mismatches among all pairwise comparisons of aligned sequences as Π. Among the four sequences *a–d* there are 6 (i.e., 4 choose 2) pairwise comparisons, namely *a–b*, *a–c*, *a–d*, *b–c*, *b–d*, and *c–d*. Each of these combinations compares 16 nucleotide sites, and among the 6 pairwise comparisons, the number of mismatches is 0 (*a–b*), 4 (*a–c*), 4 (*a–d*), 4 (*b–c*), 4 (*b–d*), and 8 (*c–d*). The total number of pairwise mismatches is therefore $0 + 4 + 4 + 4 + 4 + 8 = 24$ among a total of 6 pairwise comparisons, and so in this example $\Pi = 24/6 = 4$.

With the concepts of segregating sites and nucleotide mismatches in mind, we can proceed to examine some of the properties of the infinite-sites model of neutral evolution. First, consider a sample consisting of only two sequences. In this case, the number of segregating sites S and the average number of number of nucleotide mismatches Π are identical, because there is only one pairwise sequence comparison. For a sample of size 2, Watterson (1975) showed that the probability that the number of segregating sites S equals any number i is given by

$$\Pr\{S = i\} = \frac{1}{(1+\theta)}\left(\frac{\theta}{1+\theta}\right)^i \tag{4.14}$$

where $\theta = 4N_e\mu$. We emphasize that, in this formulation, μ is the mutation rate *across the entire nucleotide sequence.* (Formally, μ can be considered as the sum of the per-nucleotide mutation rate across all the nucleotide sites present in the sequence.)

TABLE 4.1 A Sample of Four Allelic DNA Sequences

	Nucleotide site in the DNA sequence															
Allele	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
a	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
b	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
c	G	A	A	A	C	T	T	T	A	G	G	G	T	C	C	C
d	A	G	A	A	T	C	T	T	G	A	G	G	C	T	C	C

A particular case of Equation 4.14 gives the probability that two sequences have no mismatches ($i = 0$), and hence are identical. Substituting $i = 0$ into Equation 4.14, we obtain

$$\Pr\{S = 0\} = \frac{1}{1+\theta} \tag{4.15}$$

Note that right-hand side of Equation 4.15 for the infinite-sites model is the same as the right-hand side of Equation 4.10 for the steady-state autozygosity in the infinite-alleles model. The reason is that, in a sample of size 2 in both models, the probability that the sequences are identical is also the probability of autozygosity.

From Equation 4.14 for a sample of size 2, it can be shown that the mean and variance in the number of segregating sites S are given by $E(S) = \theta$ and $V(S) = \theta + \theta^2$. As noted, for a sample of size 2, the average number of pairwise mismatches Π is equal to the number of segregating sites, and so $E(\Pi) = \theta$ and $V(\Pi) = \theta^2$ also. The variance $\theta + \theta^2$ requires complete linkage. If the nucleotide sites can undergo recombination, then the variance is reduced. An example obtained from computer simulation is shown in Figure 4.13, which compares the average number of pairwise mismatches per nucleotide site for a simulated set of data with no recombination (larger variance, black bars) and a simulated set of data with free recombination (smaller variance, gray bars). Because of this difference, the relationship between the mean and the

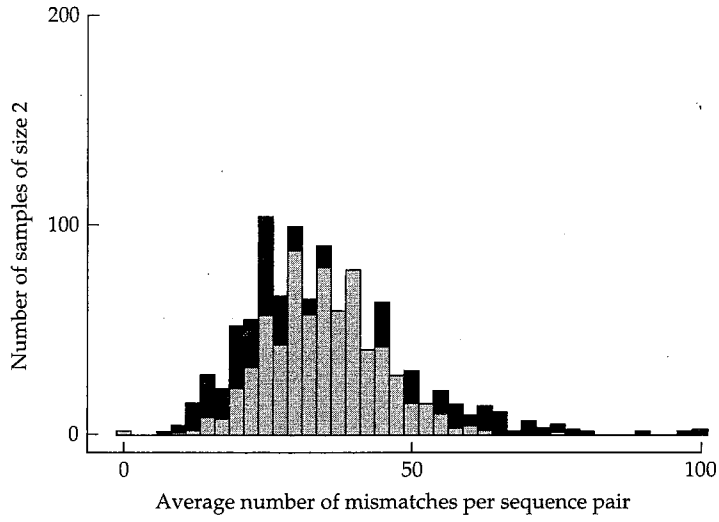


FIGURE 4.13 Equilibrium distribution of the number of mismatches between a pair of alleles. Note that free recombination results in a smaller variance than the case of no recombination.

variance in the mismatch distribution has been used to make inferences about the extent of intragenic recombination (Hudson 1987, Wakeley 1997).

Important sampling properties of the infinite-sites model with neutral evolution and no recombination were first discovered by Watterson (1975), who examined both the number of segregating sites and the average number of pairwise mismatches. The expected number of segregating sites in a sample of size n sequences is given by

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{4.16}$$

This equation was already derived in Chapter 3 (see Equation 3.41) based on the expected total length of the branches in a coalescent tree. Here $\theta = 4N_e\mu$, where μ is the mutation rate across the entire nucleotide sequence. The variance in the number of segregating sites in a sample of size n equals

$$V(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{4.17}$$

This expression for the variance is for the case of no recombination. It turns out that recombination does not affect $E(S)$, but it reduces $V(S)$. In the extreme case of free recombination between adjacent nucleotide sites, the number of segregating sites along the sequence has a Poisson distribution, and in this case the variance equals the mean.

Now consider the average number of pairwise mismatches Π among a set of sequences. An important result is that in a sample of size n at steady state,

$$E(\Pi) = \theta \tag{4.18}$$

Here again in $\theta = 4N_e\mu$, the symbol μ refers to the mutation rate across the entire nucleotide sequence.

When there is no recombination between nucleotide sites, the variance in Π was found by Tajima (1983) to be

$$V(\Pi) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \tag{4.19}$$

where θ is defined as in Equation 4.18. Again, recombination reduces the variance in the average number of pairwise mismatches (see Figure 4.13 for the case $n = 2$). It is not difficult to see intuitively why the variance is reduced—recombination shuffles the variation among alleles, reducing the average number of sites by which random pairs of alleles differ. Wakeley (1997) gives a more advanced treatment of this subject.

Nucleotide Polymorphism and Nucleotide Diversity

One limitation of S and Π as measures of nucleotide variation in a population is that each quantity depends on the length of the sequences that are compared; these differ from gene to gene and from one study to another. The

dependence on sequence length can be eliminated by expressing both the number of segregating sites S and the average number of pairwise mismatches Π as a proportion of the total number of sites. To be specific, suppose that a sample consists of n aligned sequences each of length L. Then the proportion of segregating sites among all sites compared equals S/L. There is no established symbol for this quantity, but it is sometimes called the **nucleotide polymorphism**. For the sake of consistency, we shall denote the nucleotide polymorphism by the symbol S^*, so that $S^* = S/L$. Since L is a constant, it follows that the mean and variance of S^* are given by

$$E(S^*) = E(S)/L \quad V(S^*) = V(S)/L^2 \tag{4.20}$$

Expressions for $E(S)$ and $V(S)$ in the case of no recombination are found in Equations 4.16 and 4.17. Although S^* does not depend on the sequence length, it does depend on the sample size, as is evident from Equation 4.16. On the other hand, the dependence on sample size is fairly weak unless the samples are very small (see Table 1.2).

Similarly, the average proportion of pairwise mismatches in sequences of length L is given by Π/L. This quantity does have an established symbol, namely $\pi = \Pi/L$, and it is called the **nucleotide diversity** (Nei and Li 1979). Because L is a constant, the mean and variance of π are given by

$$E(\pi) = E(\Pi)/L \quad V(\pi) = V(\Pi)/L^2 \tag{4.21}$$

Expressions for $E(\Pi)$ and $V(\Pi)$ in the case of no recombination are given in Equations 4.18 and 4.19.

Tajima's D *Statistic*

Equation 4.16 provides a method for estimating the parameter $\theta = 4N\mu$ based on the number of segregating sites in a sample S. If we define

$$\alpha = \sum_{i=1}^{n-1} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} \tag{4.22}$$

then Equation 4.16 yields the estimate

$$\hat{\theta} = S/a \tag{4.23}$$

Likewise, Equation 4.18 provides a method for estimating θ based on the average number of pairwise mismatches Π, and in this case the estimate is very direct:

$$\hat{\theta} = \Pi \tag{4.24}$$

Tajima (1989) proposed that the difference between the estimates of θ in Equations 4.23 and 4.24 could be used as a test of goodness of fit to the infinite-alleles model, and this test has come into widespread use. The rationale is that the number of segregating sites and the average number of pairwise mismatches differ primarily because the former is indifferent to the relative

frequencies of the polymorphic nucleotides at a site. The two values lead to consistent estimates for θ anyway, unless some evolutionary process causes a discrepancy from the assumptions of the infinite-alleles model. Tajima's test is based on the difference $\Pi - S/a$. If the infinite-alleles model holds (or any discrepancies are too small to invalidate Equations 4.23 and 4.24), then the difference $\Pi - S/a$ will equal 0. The major discrepancies occur in two situations:

- The frequencies of polymorphic nucleotides are too nearly equal. This pattern increases the average number of pairwise differences over its neutral expectation; hence $\Pi - S/a$ is positive. The finding typically suggests either some type of balancing selection, in which heterozygous genotypes are favored, or some type of diversifying selection, in which genotypes carrying the less common alleles are favored. This situation may also happen if the sampled population was formed from a recent admixture of two different populations.
- The frequencies of the polymorphic variants are too unequal, with an excess of the most common type and too many of the less common types. This pattern results in a decrease in the proportion of pairwise differences, so $\Pi - S/a$ is negative. One possible reason for an excess of rare alleles is selection against genotypes carrying deleterious mutant alleles. However, departures from the infinite-sites model do not necessarily imply that natural selection is operating. For example, a population that is growing will also feature an excess rare alleles and a negative value of $\Pi - S/a$.

PROBLEM 4.4 To make use of these ideas concrete, consider the example in Table 4.1. From these data, use Equations 4.23 and 4.24 to obtain estimates of θ based on the number of segregating sites and on the average number of pairwise mismatches. Then calculate $\Pi - S/a$ and interpret the results in terms of how the data depart from the expectations of the infinite-alleles model.

ANSWER For the data in Table 4.1 we have already calculated that $S = 8$ and $\Pi = 4$. In this case, $n = 4$ so that $a = 1 + \frac{1}{2} + \frac{1}{3} = 1.833$. The estimate of θ from Equation 4.23 is therefore $\frac{16}{1.833} = 4.36$, and that from Equation 4.24 is 4.00. Hence, in this example, $\Pi - S/a = 4.00 - 4.36 = -0.36$. Because of the small sample size, there is little justification to carry out a formal statistical test of whether this value is significantly different from 0, but the small discrepancy from 0 suggests no significant excess of rare alleles.. In practice, random coalescence in the infinite-sites model with neutral mutations can be generated using, for example, a program called ms (Hudson 2002). For each simulated sample, one calculates a realization of $\Pi - S/a$, and many such samples produce the null distribution of the test statistic assuming neutrality. If the observed value falls in the upper or lower 5% of the null distribution, then the P-value for the test is regarded as significant ($P < 0.05$).

Data from natural populations often have an excess of segregating sites in which the minority nucleotide is present only once in the sample, constituting what is called a **singleton**. Although excess singletons can result from recent rapid population growth, more often it suggests that the singletons represent slightly deleterious alleles held at low frequency by selection. When this is observed for nonsynonymous nucleotide polymorphisms in protein-coding regions, it is usually interpreted to imply that many amino acid polymorphisms are slightly deleterious and held at low frequency in the population by a balance between selection tending to drive them out and recurrent mutation producing new deleterious alleles.

Tajima's (1989) test is actually based on a normalized version of $\Pi - S/a$, where the magnitude of the difference is expressed as a multiple of the standard deviation of the difference. The resulting statistic is known as **Tajima's *D***:

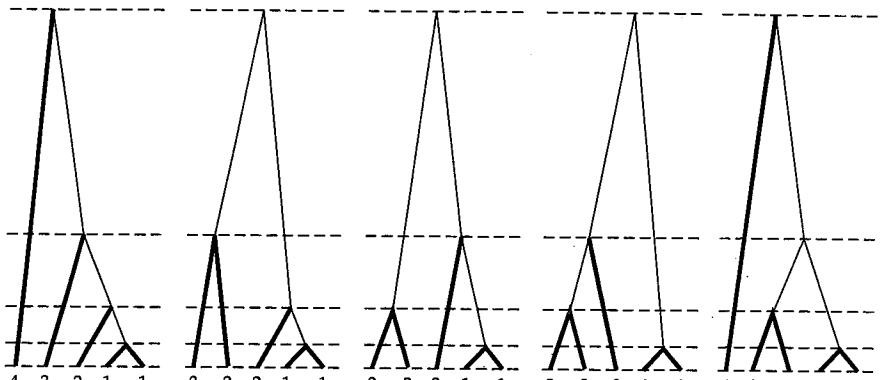
$$D = \frac{\Pi - S/a}{\sqrt{V(\Pi - S/a)}} \tag{4.25}$$

An explicit formula for the denominator in Equation 4.25 can be found in Tajima's paper. The original intent was that the statistical significance of any observed value of Tajima's D could be based on the mean and variance of the distribution of $\Pi - S/a$. But the distribution of $\Pi - S/a$ is very complex, and at the present time most tests of significance are based on comparing an observed value of Tajima's D against simulated values obtained from coalescent simulations as described in the answer to Problem 4.4.

The Fu and Li Test of Fit to Neutral Coalescence

A remarkable property of coalescent trees is the basis of another widely used test of whether a sample configuration of nucleotide polymorphisms is consistent with the neutral infinite-sites model at steady state (Fu and Li 1993). This property is illustrated in Figure 4.14 for a sample of size $n = 5$. For this sample size there are only five basic tree structures based on the pattern of coalescences, which are laid out as in Figure 3.15. There are actually many more trees (120, to be exact) when one takes into account the number of different ways that the tips of the tree can be labeled with the names of alleles in the sample. When this is taken into account, it happens that there are exactly twice as many trees with the structure on the left as any of the others. The fraction of all trees with each of the structures is summarized in the second row of numbers in Figure 4.14.

The trees in Figure 4.14 are "average" trees in the sense that the coalescent times have been shown in proportion to their expected values. Note that some of the branches are thick, others thin. Each of the thick branches is called an **external branch** because it emerges from an internal node and terminates at one of the tips. Each of the thin branches is called an **internal branch** because it connects two internal nodes.



Tree structure					
Expected length of external branch (units of $4N_e$)	$\frac{4}{5}$ $\frac{3}{10}$ $\frac{2}{15}$ $\frac{1}{20}$ $\frac{1}{20}$	$\frac{3}{10}$ $\frac{3}{10}$ $\frac{2}{15}$ $\frac{1}{20}$ $\frac{1}{20}$	$\frac{2}{15}$ $\frac{2}{15}$ $\frac{3}{10}$ $\frac{1}{20}$ $\frac{1}{20}$	$\frac{2}{15}$ $\frac{2}{15}$ $\frac{3}{10}$ $\frac{1}{20}$ $\frac{1}{20}$	$\frac{4}{5}$ $\frac{2}{15}$ $\frac{2}{15}$ $\frac{1}{20}$ $\frac{1}{20}$
Fraction of random coalescent trees with each structure	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Expected sum of external branch lengths	$(\frac{16}{12})\,4N_e$	$(\frac{10}{12})\,4N_e$	$(\frac{8}{12})\,4N_e$	$(\frac{8}{12})\,4N_e$	$(\frac{14}{12})\,4N_e$
Expected sum of internal branch lengths	$(\frac{9}{12})\,4N_e$	$(\frac{15}{12})\,4N_e$	$(\frac{17}{12})\,4N_e$	$(\frac{17}{12})\,4N_e$	$(\frac{11}{12})\,4N_e$

FIGURE 4.14 Coalescent trees for samples of size 5 with the times of coalescence shown in proportion to their expected values. Thick lines denote external branches, thin lines internal branches. The expected length of the external branches, averaged across all trees, equals $4N_e$ generations; this number holds true for coalescent trees with any number of samples.

The property used in the Fu and Li (1993) test is based on the expected total length of the external and internal branches. The first row of numbers, positioned beneath each tip, consists of fractions that indicate the expected length of the external branch to that tip, expressed in units of $4N_e$ generations. For example, the long external branch at the far left has an expected length of $(\frac{4}{5}) \times 4N_e$ generations. The expected branch lengths are based on the fact that the expected time to coalesce from k to $k-1$ alleles is given by $4N_e/k(k-1)$, which is implied by Equation 3.35 in Chapter 3.

For each tree structure, the expected length of all the external branches taken together is given in the third row of numbers. In this case the dependence on $4N_e$ is made explicit. For each tree structure, the multiplier in paren-

theses is the sum of the expected length of the individual external branches for that tree. Taking the tree at the far left as an example,

$$\left[\left(\frac{4}{5}\right)+\left(\frac{3}{10}\right)+\left(\frac{2}{15}\right)+\left(\frac{1}{20}\right)+\left(\frac{1}{20}\right)\right]\times 4N_e=\left(\frac{16}{12}\right)\times 4N_e$$

which is the entry in the third row. The reason that all of the values are expressed with 12 as the common denominator is that it makes it easier to calculate the expected length of the external branches over all possible coalescent trees. In this case, the expected total external branch length is given by

$$\left[\left(\frac{2}{6}\right)\left(\frac{16}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{10}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{8}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{8}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{14}{12}\right)\right]\times 4N_e=4N_e$$

Remarkably, this result is completely general and independent of the sample size.

What about the total length of the internal branches? Equation 3.40 in Chapter 3 says that the expected total length of all the branches is equal to $4N_e \times a$, where a is the sum of reciprocals defined in Equation 4.22. Because the length of the internal branches must equal the difference between the length of all branches and the length of the external branches, it follows that the total length of the internal branches must be $4N_e \times (a-1) = (a-1) \times 4N_e$. In the case of $n = 5$ (see Figure 4.14), $a = \frac{25}{12}$, and therefore the expected length of the internal branches in this case is $[\frac{25}{12} - 1] \times 4N_e = (\frac{13}{12}) \times 4N_e$. This can be verified directly from Figure 4.14 by calculating

$$\left[\left(\frac{2}{6}\right)\left(\frac{9}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{15}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{17}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{17}{12}\right)+\left(\frac{1}{6}\right)\left(\frac{11}{12}\right)\right]\times 4N_e=\left(\frac{13}{12}\right)4N_e$$

The reason for distinguishing external branches from internal braches is that *any mutation along an external branch results in a singleton polymorphism in the sample.* Similarly, every mutation along an internal branch results in a nonsingleton polymorphism in the sample. Therefore, the numbers of singleton and nonsingleton polymorphisms in the sample allow a comparison of the total lengths of the external and internal branches of the coalescent tree.

To do this comparison, we need to consider where mutations may occur along the branches, and for this purpose let μ represent the rate of mutation along the total length of the DNA sequence in each sampled allele, and as usual let $\theta = 4N_e\mu$. The number of mutations along the external branches is usually denoted η_e, and its expected value is given by the product of the expected external branch lengths and the mutation rate, or $4N_e \times \mu$. In other words the expected number of singleton polymorphisms in the sample is given by

$$E(\eta_e) = 4N_e\mu = \theta \tag{4.26}$$

Letting η_i denote the number of mutations along the internal branches, simi-

lar reasoning indicates that the expected number of nonsingleton polymorphisms in the sample is given by

$$E(\eta_i) = (a-1)4N_e\mu = (a-1)\theta \tag{4.27}$$

Equations 4.26 and 4.27 yield estimated θ based either on the number of singleton polymorphisms or the number of nonsingleton polymorphism. Furthermore, for samples sizes of about 10 or greater, η_e and η_i are nearly independent of one another (Li 1997).

The Fu and Li (1993) test is based on the difference between the two estimates of θ implied by Equations 4.26 and 4.27, namely,

$$G = \frac{\eta_e - \eta_i / (a-1)}{\sqrt{V[\eta_e - \eta_i / (a-1)]}} \tag{4.28}$$

where G is the test statistic and V indicates the variance. As with other such tests, perhaps the best way to obtain a P-value is to estimate the null distribution of G from neutral coalescent simulations and then compare the observed value of G with the simulated values.

Fu and Li (1993) suggest that G might be a useful test statistic in cases in which most new mutations are deleterious, because in such cases the number of singletons will be increased relative to the number of nonsingletons. The reasoning is that if most new mutations are harmful, they may appear in samples as singletons, but most of them will be eliminated very quickly. Only the minority of new mutations that are neutral or nearly neutral have much chance of increasing in frequency to a level sufficient to appear in samples as nonsingletons. Hence, in this model, η_e is increased relative to η_i. It should however be noted that the neutral infinite-sites model predicts a large fraction of singletons anyway. Table 4.2 shows the expected proportion of single-

TABLE 4.2 Expected Proportion of Singletons in Samples

Sample size n	E (proportion singletons)	Sample size n	E (proportion singletons)
2	1.000	12	0.331
3	0.667	13	0.322
4	0.545	14	0.314
5	0.480	15	0.308
6	0.438	16	0.301
7	0.408	17	0.296
8	0.386	18	0.291
9	0.368	19	0.286
10	0.353	20	0.282
11	0.341	21	0.278

tons (equal to $1/a$) in samples of various size. The expected proportion of singletons does not drop below 20% until $n = 85$.

4.6 MUTATION AND RECOMBINATION

Recombination reshuffles the alleles created by mutation. On the surface, this seems like a very good thing. Any mechanism that allows parental organisms to generate a myriad of genetic combinations among the progeny would allow a more thorough exploration of combinations of alleles that might be favored by natural selection. This mechanism would help make natural selection more efficient and promote the persistence of genes present in the parents. This argument is so seductive that it seems self-evident, but in fact the argument contradicts the fundamental principles of evolution. It is like saying that everyone should buy lottery tickets, because someone might hit the jackpot. But most lottery tickets are losers, and there often is no jackpot. The expected return on any lottery ticket is negative, and so the strategy eventually leads to financial ruin.

Natural selection operates through the phenotype of the individual, and segregation and recombination break down all combinations of alleles, including combinations of alleles resulting in superior phenotypes. Under artificial selection as practiced by breeders, as we shall see in Chapter 8, segregation and recombination are often handicaps, because the average phenotype of the offspring of superior individuals regresses toward the population average. Furthermore, because of sexual reproduction, any individual contributes only half of its genes to any offspring, whereas if the individual were to reproduce asexually by parthenogenesis, then it would contribute twice as many genes to the offspring. Sexual reproduction therefore has an intrinsic twofold cost as compared with asexual reproduction, and in principle (though not necessarily in practice) an ideal strategy for breeders to keep the best genotypes intact would be to clone them. In population genetic models in which differences in survival among genotypes result in stable polymorphisms among interacting genes, it can be shown that, in the absence of mutation, natural selection favors genetic modifiers that reduce the frequency of recombination (Altenberg and Feldman 1987).

Inasmuch as there is an intrinsic twofold cost of sex, and recombination breaks down favorable combinations of alleles, why is sex and recombination so widespread among eukaryote organisms? There is at present no consensus on this issue, and certainly no definitive data beyond the observation that, with few exceptions (Mark Welch et al. 2004), groups of sexual organisms that have given up sex and become asexual tend to have a short evolutionary lifetime (Judson and Normark 1996).

One hypothesis is that recombination arose as a byproduct of DNA repair, since at the molecular level recombination is initiated by a double-stranded break in DNA, and many of the proteins involved in recombination are also

implicated in DNA repair (Redfield 2001). Another possibility is that recombination is a side effect of chromosome separation in meiosis, since breakage and reunion between DNA molecules is necessary to lock homologous chromosomes together to ensure proper segregation. As long as most mutations are recessive, as supported by abundant evidence, selection for segregation is stronger than selection for recombination (Otto 2003). There are also models for the evolution of recombination based in population genetics, and these are considered next.

A Model for the Evolutionary Benefit of Recombination

Evolutionary biologists have emphasized for a long time that recombination can accelerate the rate of formation of beneficial gene combinations, and it has been suggested that this acceleration is the reason why recombination evolved (Fisher 1930; Muller 1932). A graphical representation of the Fisher-Muller model is illustrated in Figure 4.15. In part A are two large populations, one with no recombination (an asexual species) and one with recombination (a sexual species). Each has three favorable mutations, *a*, *b*, and *c*, which ultimately become incorporated into the genome. In the asexual species, the mutations are incorporated sequentially because each favorable mutation must take place in the genetic background of the one before. The process is shown as being slow because each favorable mutation must be nearly fixed before there is a high chance that the next favorable mutation takes place in the proper genetic background. In experimental populations of bacteria with no recombination, the process in which a lineage containing a favorable mutation displaces other lineages, including those containing less favorable mutations, is known as **clonal interference** (Gerrish and Lenski 1998; Hegreness et al. 2006).

A similar type of interference between favorable alleles occurs to a smaller extent even in the presence of recombination, which is called the **Hill-Robertson effect** (Hill and Robertson 1966). The Hill-Robertson effect occurs because two different favorable mutations (call them *A* and *B*) are likely to arise in different genetic backgrounds, and as the favored alleles increase in frequency, they bring about negative linkage disequilibrium (D in Equation 2.13 in Chapter 2), in which the product of the frequencies of gametes carrying one favorable and one unfavorable allele (Ab and aB) is greater than that of gametes carrying both favorable alleles (AB, virtually nonexistent) or both unfavorable alleles (ab). Because of the genetic backgrounds they are in, the increase in frequency of the favorable alleles also increases the magnitude of the negative linkage disequilibrium, and the effect is greatest with a low frequency of recombination or a small effective population size. Under these conditions there is selection for increased recombination (Otto and Barton 1997, Barton and Otto 2005; Roze and Barton 2006), and when the frequency of recombination is high, different favorable alleles can be brought together in rapid succession (Figure 4.15). On the other hand, if the rate of favorable

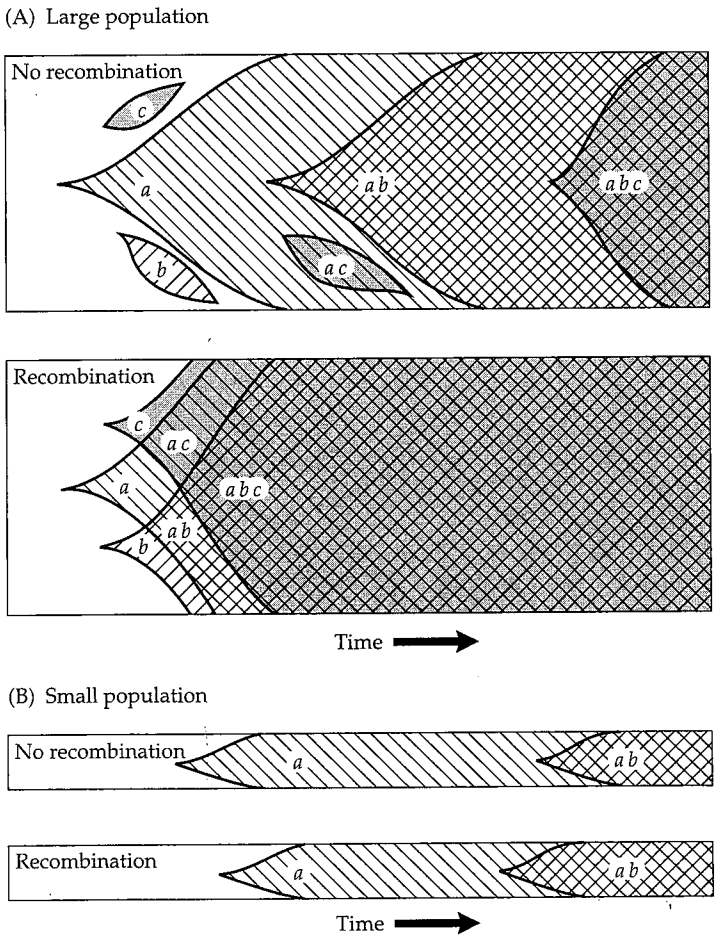


FIGURE 4.15 Model for the evolutionary benefit of recombination. (A) In a large population of an asexual species with no recombination (top panel), the favorable mutations *a*, *b*, and *c* must be incorporated into the genome sequentially because there is no mechanism to bring the favorable mutations together; each favored mutation must reach a high frequency to have a reasonable chance that the next favorable mutation will take place in the proper genetic background. With recombination (bottom panel), recombination between the favorable genes enables the triple mutant *abc* to be formed very rapidly. (B) The beneficial effect of recombination is diminished in a very small population because, in a small population, multiple favorable mutations are unlikely to be present simultaneously. (From Crow and Kimura 1970.)

mutation is so low that an asexual population can fix any favorable allele before the next one occurs, then the advantage of recombination illustrated in Figure 4.15 is very much reduced (Christiansen et al. 1998).

The advantage of recombination in the Fisher-Muller model is also affected by the size of the population. In very small populations, multiple favorable mutations are unlikely to be present simultaneously, and so the fixation of the favorable alleles proceeds sequentially in a sexual as well as in an asexual species (Figure 4.15B). But there is an offsetting process, which is the negative linkage disequilibrium discussed in connection with the Hill-Robertson effect. Although the magnitude of the disequilibrium is small when the favorable alleles are rare, the increase in frequency of these alleles due to selection amplifies the linkage disequilibrium, and genetic factors that increase the frequency of recombination are favored (Barton and Otto 2005). In a single population, the extent of linkage disequilibrium built up diminishes as the population size increases, and so the advantage of recombination decreases. The subdivision of a large population into many smaller subpopulations counteracts this effect, and in this case substantial amounts of linkage disequilibrium can accumulate even in a large population (Martin et al. 2006).

Muller's Ratchet

Clonal interference (in asexual organisms) or the Hill-Robertson effect (in sexual organisms) can decrease the efficiency with which favorable mutations are incorporated into populations. The flip side is that these same processes make it more difficult for a population to get rid of deleterious mutations. The Hill-Robertson effect appears to account for a correlation between intron size and frequency of recombination in *Drosophila* (Carvalho and Clark 1999). In particular, in *D. melanogaster*, the largest introns (>> 80 base pairs) and the smallest introns (< 60 base pairs) tend to be found in genes that are located in regions of the genome with low rates of recombination. This finding suggests that both very large introns and very small introns are deleterious, and that deleterious insertion events that make large introns larger and deleterious deletion events that make small introns smaller are less efficiently eliminated in regions of low recombination, as predicted by the Hill-Robertson effect (Carvalho and Clark 1999). An alternative explanation is that larger introns are favored in regions of low recombination because they allow more recombination to take place (Comeron and Kreitman 2002). This model does not explain why genes in the Y chromosome have large introns, because the Y chromosome never undergoes recombination.

In asexual organisms, the accumulation of deleterious mutations in small populations is known as **Muller's ratchet** (Muller 1964). To understand this process, consider the experiment diagrammed in Figure 4.16. A clone formed from a single bacterial cell is used to inoculate a liquid culture, and after the population has grown to a large size, an aliquot is diluted to such an extent

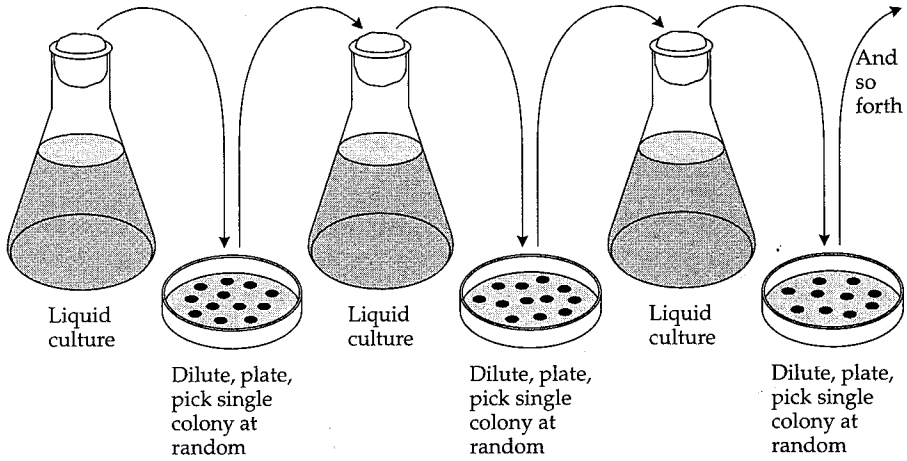


FIGURE 4.16 An experimental procedure demonstrating Muller's ratchet. In each cycle, the population passes through an extreme bottleneck of size $N = 1$, and so any mutations present in the chosen individual immediately become fixed in the population. Over a long period, the genome accumulates many deleterious mutations, including deletions.

that individual cells will be well separated when a small amount of liquid is spread onto the surface of semisolid growth medium in a petri plate. During incubation of the plate, each cell divides multiple times and soon yields a clone of cells, forming a visible colony. At this point a single colony is chosen at random, some of its cells are used to inoculate another flask of liquid medium, and the cycle is repeated.

In each generation, new mutations occur, and we may safely assume that, except for any that are neutral or nearly neutral, the great majority of the other mutations are deleterious. Let us assume for simplicity that, in each cycle of liquid growth, the cells undergo enough rounds of DNA replication that the overall genomic mutation rate to deleterious but nonlethal mutations is 1 per cell per experimental cycle. A deleterious mutation rate of 1 per genome per generation is not unrealistic for higher eukaryotes (Kondrashov 2001). At a deleterious mutation rate of 1 per experimental cycle, the probability that a colony chosen at random has no new deleterious mutations after one cycle is $e^{-1} = 0.37$, after two cycles it is $e^{-2} = 0.14$, and after three cycles it is $e^{-3} = 0.05$. In other words, in the experiment in Figure 4.16, the probability that a colony chosen at random in the third cycle has one or more deleterious mutations is about 95%. What's more, once a colony with a deleterious mutation is chosen, that deleterious mutation becomes fixed in the population, except in the very unlikely event of reverse mutation. This process of fixing

deleterious mutations is the basis of Muller's ratchet. Each deleterious mutation that is fixed sets a new base line, and subsequent deleterious fixations can only make matters worse. Eventually, in the absence of recombination to bring nonmutant combinations of alleles back together again, the genome undergoes mutational degeneration.

In the experiment in Figure 4.16, the repeated population bottleneck of size $N = 1$ is quite extreme, but actual experiments with a range of bottleneck sizes have been carried out using a bacterial virus (bacteriophage $\phi 6$) that has an RNA genome (Poon and Chao 2004). This RNA virus is convenient for such experiments because it has a small genome and a high mutation rate. The experiments verified the accumulation of deleterious mutations owing to Muller's ratchet, and showed that bacteriophage populations allowed to undergo recombination usually outperformed asexual populations, with the greatest advantage of recombination in the smallest populations (Poon and Chao 2004). Muller's ratchet also results in the fixation of spontaneous deletions. In an experiment like that in Figure 4.16 carried out in *Salmonella enterica*, observed deletions ranged in size from 1–200 kb, and the average rate of DNA loss per genome was 0.05 base pairs per generation (Nilsson et al. 2005).

Over an evolutionary time scale, the accumulation of deletions owing to Muller's ratchet can result in extreme reductions in genome size. Remarkable examples are found in bacteria that are obligate intracellular pathogens or symbionts that are supplied with nutrients by their host (Ochman 2005). An example is the bacterial endosymbiont of aphids, *Buchnera aphidicola*, which is transmitted in very small numbers through the aphid oocyte and has no opportunity for recombination. Since its association with aphids 100–250 million years ago, the bacterial genome has undergone many deleterious amino acid substitutions and regulatory changes (Moran 1996; Moran and Degnan 2006). Its genome size has been reduced by deletions to 600 kb, whereas the ancestral genome was approximately ten-fold larger (Ochman 2005).

Piecewise Recombination in Bacteria

Many prokaryotic organisms make use of mechanisms of recombination in which a piece of DNA that is small, relative to the size of the entire genome, is transferred from a donor cell into a recipient cell (Redfield 2001). These mechanisms include *transformation*, in which free DNA is taken up by the recipient from the surrounding medium; *transduction*, in which a DNA fragment is carried from the donor to the recipient by means of a virus particle; and *conjugation*, in which a replica of the chromosome from a donor cell is transferred into a recipient cell by a gradual process requiring cell-to-cell contact, but the chromosome usually breaks before the transfer is complete. Because relatively short patches of the genome participate in recombination, these processes differ in their evolutionary implications from meiotic recombination in eukaryotes. Through mechanisms involving transmissible plas-

Nucleotide site in *phoA* gene

Allele	6	8	1 0	1 0	1 0	1 0	1 0	1 0	1 4	1 4	1 4	1 4	1 4	1 5	1 5	1 5	1 5	1 6	1 7	1 7	1 7	1 8	1 8
	2	3	5	6	6	7	7	8	2	2	7	7	9	0	2	5	6	8	1	6	8	2	5
	7	1	9	1	8	4	7	1	5	8	4	9	7	9	4	1	0	3	2	9	2	6	0
				*				*	*		*												
RM217T	C	A	G	A	C	G	A	G	G	T	G	T	G	T	T	T	T	T	T	G	A	A	T
RM45E	T	A	C	G	C	G	A	G	T	C	A	C	T	C	C	C	C	T	T	G	A	A	T
RM224H	C	G	C	G	T	A	G	T	T	C	A	C	T	C	C	C	C	C	C	A	T	T	C

FIGURE 4.17 Evidence for recombination in the *phoA* gene in natural isolates of *E. coli*. The pair of strains at the top are more similar at the beginning and end of the gene, the pair of strains at the bottom are more similar in the central region. There is significant clustering of the nucleotide sites inscribed in boxes, as expected from recombination. (Data from DuBose et al. 1988.)

mids (extrachromosomal DNA elements) and transposable elements, bacteria can also acquire genes from other species. Although these processes of **lateral gene transfer** are extremely important in the origin and spread of bacteria resistant to multiple antibiotics, they are individually rare events. Normally, genetic exchange between bacteria takes place between individuals of the same species (Ochman et al. 2005).

The main effect of short-patch recombination is that long-range linkage disequilibrium tends to be maintained. For example, in enteric bacteria, such as *Escherichia coli*, which are part of the normal intestinal flora, linkage disequilibrium between allozyme loci is very strong (Whittam et al. 1983). At the level of DNA sequence, however, many genes have an obviously mosaic structure in which different segments have different phylogenetic histories (DuBose et al. 1988). An example from the *phoA* gene, coding for alkaline phosphatase in *E. coli*, is illustrated in Figure 4.17. Among the polymorphic nucleotide sites indicated, the unique nucleotide at each site is inscribed in a box. At the extreme ends of the gene, the alleles from strains *RM217T* and *RM45E* are the most closely related; in the middle of the gene, from nucleotide sites 1425 to 1560, there is a run of polymorphic nucleotides in which the similarity between *RM217T* and *RM45E* is lost, as if this part of the gene had been introduced by recombination with a more distantly related allele. Although short runs of similar or dissimilar nucleotides can also be the result of chance, chance effects can be ruled out by appropriate statistical tests for recombination (Stephens 1985; Sawyer 1989).

The finding that many genes have a mosaic ancestry through recombination seems at first to contradict the finding of significant linkage disequilibrium between more widely separated genes. The paradox is resolved by the fact that each recombination event is local; it replaces a relatively short

stretch of the recipient chromosome, and the linkage phase between more distant alleles is maintained. The *E. coli* chromosome, therefore, consists of clonal segments from a common ancestor, which is called the **clonal frame** (Milkman and Bridges 1990, 1993), interrupted by short segments derived from recombination with diverse other clones. Even though the clonal frames are interrupted by relatively short recombinant segments, their integrity would ultimately be lost unless there were occasional selective events favoring particular genotypes. A clonal frame implies that most genes in the genome will share a common gene tree, their coalescence. The existence of a clonal frame depends on the level of recombination, because high levels of recombination result in different genes having different coalescent histories. Although gene trees in species such as *E. coli* and *Hemophilus influenzae* show good congruence, the gene trees for a sample of genes from *Neisseria meningitidis, Streptococcus pneumoniae, Streptococcus pyogenes,* and *Staphylococcus aureus* show no better congruence with each other than with trees of random topology (Feil et al. 2001).

Animal Mitochondrial DNA

Population genetic studies in animals often focus on the DNA of mitochondria. The mitochondrial genome is informative about parentage because, in most species of animals, it is almost always maternally inherited and rarely if ever undergoes recombination. It is also a small molecule present in abundant quantities in most cells. In animals, mitochondrial DNA (mtDNA) is a circular molecule typically in the range from 15–20 kb in length. It codes for fewer than 40 genes; approximately half code for ribosomal RNA or for transfer RNA used in mitochondrial protein synthesis, and the remaining genes code for proteins used in electron transport or oxidative phosphorylation. In many species, including mammals, parts of the mtDNA sequence evolve very rapidly in comparison with nuclear genes, and hence mtDNA can often be used to make inferences about population structure and recent population history.

An example of the utility of mtDNA in population studies is illustrated in Figure 4.18, which summarizes the result of examining the mtDNA of 87 pocket gophers, *Geomys pinetis,* collected across the geographic range of the species in Alabama, Georgia, and Florida (Avise et al. 1979). The mtDNA from each gopher was digested in turn with each of six restriction enzymes, each cleaving the DNA at a different six-base recognition site. The resulting restriction fragments were separated by electrophoresis and compared among the animals to estimate the number of nucleotide differences affecting the restriction sites.

Among the 87 gophers, there were 23 distinct types of mtDNA, represented by the lowercase letters in Figure 4.18. Each of these types represents a maternal mtDNA lineage, distinct from other lineages. Animals that share an mtDNA type must have a female ancestor in common. The branching network

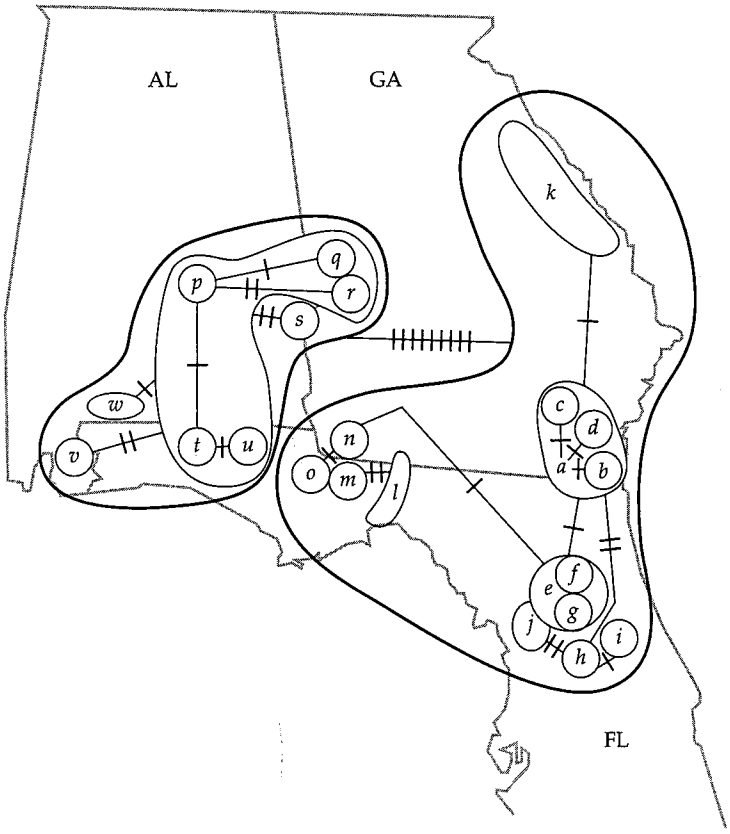


FIGURE 4.18 Lineage relationships between mtDNA types in pocket gophers. The lowercase letters are different mtDNA types grouped according to similarity and superimposed on a geographical map of the collection sites. The tick marks across the connecting lines are the numbers of inferred mutational steps. (From Avise 1994.)

in Figure 4.18 estimates the matriarchal phylogeny of the mtDNA. The straight lines connect related types of mtDNA, and the number of slashes across each line indicates the estimated number of nucleotide differences in the restriction sites between the mtDNA types. Groups of related mtDNA types are enclosed in thin black lines; the thickest lines delineate a western and an eastern subpopulation of gophers whose overall mtDNA sequence differs by an estimated 3%. Between the eastern and western subpopulations, there are 9 nucleotide differences among the sites cleaved by the restriction enzymes.

The mtDNA network in Figure 4.18 also resolves population subdivision within the western and eastern subpopulations. This subdivision is indicated

by the mtDNA types circumscribed by the thin black lines. Some of the mtDNA types such as "*k*" and "*p*" are widespread, whereas others such as "*b*" and "*q*" are more local in their distribution. The local clones usually differ from the most widespread mtDNA type in the region by only one or two nucleotides among the sites cleaved by the restriction enzymes. The example in Figure 4.18 shows that, because of matrilineal inheritance and the absence of recombination in mtDNA, the network of mtDNA types can reveal a great deal about population substructure in natural populations.

At the beginning of this section, we said that animal mtDNA "is almost always maternally inherited and rarely if ever undergoes recombination." About the first point there is little dispute, although there does seem to be one authenticated case of a human male who inherited his mtDNA from his father (Bromham et al. 2003). The issue of recombination is much more unsettled (Piganeau et al. 2004, Tsaousis et al. 2005). Detecting possible rare recombinants in mitochondrial DNA sequences is complicated by the possible effects of parallel mutation, hot spots of mutation, selection, population substructure, sequencing errors, and other technical issues. There are also a large number of types of statistical tests for recombination, each of which perform well under some sets of assumptions but not others (Bruen et al. 2006). Hence the issue of recombination in animal mitochondrial DNA is still up in the air, and where it comes down has implications for long-term mtDNA evolution. However, it seems clear from the available data that the magnitude of mtDNA recombination, if it occurs, is not sufficient to nullify the use of mtDNA as a marker for studies of population substructure or recent population history.

SUMMARY

1. Mutation is the ultimate source of evolutionary novelty, but for most genes the rate of mutation are usually so low that mutation pressure is a weak force for changing allele frequencies.
2. Despite the weak force of mutation pressure, over very long periods of evolutionary time, populations can come into equilibrium between forward and reverse mutation, when the ratio of allele frequencies equals the reciprocal of the ratio of the mutation rates.
3. The neutral theory asserts that many mutations have so little effect on the survival and reproduction of the organism that their fate is determined primarily if not exclusively by random genetic drift. Put forward at a time when most DNA was thought to code for proteins, the neutral theory was at one time highly controversial, but the large amount of noncoding DNA now known to be in introns, pseudogenes, spacers between genes, and so forth may give considerable scope for mutations that are neutral or nearly neutral.
4. The infinite-alleles model assumes that each new mutation yields an allele that is unique to the population, and it is appropriate for situations

in which a large number of alleles can be distinguished without the actual DNA sequences being known.

5. In the infinite-alleles model, alleles that are physically indistinguishable are assumed to be identical by descent. For selectively neutral mutations at steady state, the proportion of heterozygous genotypes is expected to be $\theta/(1 + \theta)$, where $\theta = 4N_e\mu$.
6. The Ewens sampling formula pertains to the expected configuration of alleles in samples from populations that are at steady state under the infinite-alleles model. The sampling formula allows tests of the neutral theory based on comparisons between theoretical predictions and the composition of observed samples.
7. The infinite-sites model is appropriate for DNA sequences and assumes that each new mutation alters a single nucleotide site. For selectively neutral mutations at steady state, the model makes specific predictions about the number of segregating nucleotide sites and the average number of nucleotide differences between pairs of sequences.
8. Predictions based on the infinite-sites model allow statistical tests of neutrality based on various characteristics in observed samples, such as the evenness of the allele-frequency spectrum (Tajima's D) or the number of singleton versus nonsingleton polymorphisms (Fu and Li test).
9. Recombination allows the formation of beneficial combinations of genes, but there is no consensus on the evolutionary origin of recombination. In asexual organisms, clonal interference reduces the efficiency of selection. In sexual organisms, population genetic models demonstrate that selection can amplify negative linkage disequilibrium between favorable mutations, particularly in small populations, and this process favors increased recombination.
10. The genomes of some species of bacteria, such as *E. coli*, show extensive linkage disequilibrium over long genetic distances in spite of the fact that each gene may have a mosaic ancestry owing to intragenic recombination. The apparent paradox results because recombination in bacteria usually involves a short stretch of DNA and the process is infrequent.
11. In animal mitochondrial DNA (mtDNA), maternal transmission and the absence of recombination enable the tracking of mitochondrial lineages to make inferences about population history and substructure. Isolated cases of nonmaternal transmission of mtDNA do occur, and the issue of whether recombination in animal mtDNA is completely absent or merely rare is highly controversial.

PROBLEMS

1. Most protein-coding genes have a forward mutation rate (wildtype to mutant) that is at least an order of magnitude greater than the reverse mutation rate (mutant back to wildtype). Why should this be the case?

2. What is the Hill-Robertson effect and what causes it?
3. What is Muller's ratchet and why is it important in populations that undergo frequent bottlenecks of population size?
4. A classical bacterial experiment demonstrated that mutations occur at random and not in response to specific selection pressures for them. The experiment used sterilized velvet to imprint the geometrical pattern of bacterial colonies on an agar surface in a petri dish (a "plate"), which was used to replicate the pattern by impressing the velvet on sterile nutrient agar in a selective plate containing an antibiotic. Colonies on the original plate giving resistant cells on the selective plate were dispersed into single cells, spread onto a nutrient agar plate without antibiotic, and allowed to multiply into colonies. This procedure was repeated until one or more colonies on the nonselective medium consisted exclusively of antibiotic-resistant cells. How does this experiment prove the point?
5. Estimation of mutation rates from bacterial cultures can be tricky because, if a mutation occurs early in the life of a culture, the final frequency will be very high, but if it occurs late, the final frequency will be low. The *fluctuation test* is a method for getting around this problem by growing many smaller cultures and estimating the mutation rate from the proportion of cultures that contain no mutations using the zero term of the Poisson distribution $P_0 = \exp(-\mu N)$, where P_0 is the proportion of cultures with no mutations, μ is the mutation rate, and N is the average number of cells per culture. In one experiment for bacteriophage T1 resistance, 11/20 cultures contained no mutations and the average number of cells per culture was 5.6×10^8. Estimate μ.
6. If recessive lethals occur independently in *Drosophila* autosomes, and the probability that an autosome contains one or more recessive lethals is 0.35 (a typical figure for chromosomes isolated from natural populations), what is the average number of recessive lethals per chromosome? Assume that the distribution of lethals is Poisson so that the probability of a chromosome containing exactly i lethals is

$$\Pr\{\text{exactly } i \text{ lethals}\} = \frac{m^i}{i!} e^{-m}$$

where m is the mean number of lethals per autosome.
7. The doubling dose of radiation is the quantity of radiation that induces as many mutations as occur spontaneously, so the total mutation rate of organisms exposed to the doubling dose equals two times the spontaneous mutation rate. Below are the induction rates per rad of x-rays (a standard measure of dose) for various genetic end points in irradiated male mice, along with the spontaneous rates. What are the corresponding doubling doses?

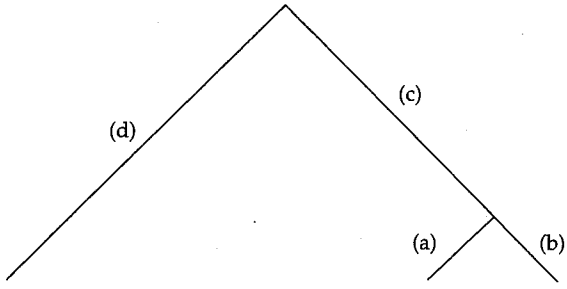
	Induction rate/rad	*Spontaneous rate*
Dominant lethals	5×10^{-4}/gamete	2 to 10×10^{-2}/gamete
Recessive visibles	7×10^{-8}/locus	8×10^{-6}/locus
Reciprocal translocations	1 to 2×10^{-5}/cell	2 to 5×10^{-4}/cell

8. For irreversible mutation with a forward mutation rate $\mu = 5 \times 10^{-6}$, calculate the expected allele frequency p after 10, 100, 1000, and 10,000 generations, assuming $p_0 = 1.0$.
9. If a transposable genetic element becomes fixed at a particular site but undergoes deletion at the rate of 1% per generation, how many generations are required to decrease the frequency of the element at the site to 90%?
10. The following data give the frequency q of bacteria resistant to a bacteriophage after t generations of chemostat growth. At $t = 12$ hours a novel metabolite was added to the medium.
    (a) What is the basal rate of mutation to resistance?
    (b) What is the effect of the novel metabolite on the mutation rate?

t	q	t	q
0	1×10^{-6}	16	7.04×10^{-6}
4	3×10^{-6}	20	7.08×10^{-6}
8	5×10^{-6}	24	7.12×10^{-6}
12	7×10^{-6}		

11. In the model of forward and reverse mutation, what is the equilibrium frequency p of A if
    (a) $\mu = 10^{-5}$ and $v = 10^{-6}$?
    (b) μ is increased tenfold?
    (c) v is increased tenfold?
    (d) both are increased tenfold?
12. In the model of forward and reverse mutation, show that the time required for the allele frequency to go halfway to equilibrium is approximately $t = 0.69/(\mu + v)$ generations. Use the approximation that $\ln(1 - x) \approx -x$ when x is small. What time is required to go halfway to equilibrium when $\mu = 10^{-5}$ and $v = 10^{-6}$?
13. In the model of irreversible mutation, what is the frequency q_t of allele a in generation t if the mutation rate changes from generation to generation? If the equation $q_t = q_0 + \mu t$ is applied to this situation, what value corresponds to μ?

14. A population at steady state in the infinite-alleles neutral model has a homozygosity F equal to 12.5%. What value of $\theta = 4N\mu$ is indicated? With random mating, how many equally frequent alleles would be required to produce the same level of homozygosity?
15. What sample values are compared in Tajima's D, and what is the rationale for the comparison?
16. A sample of size $n = 12$ includes $S = 50$ segregating nucleotide sites. Assuming that the sample conforms to the expectations of the neutral infinite-sites model at steady state, what is the expected average number of pairwise mismatches Π? What is the average number of pairwise mismatches per segregating site? How many pairwise mismatches per segregating site would result from a singleton nucleotide?
17. What sample values are compared in the Fu and Li test, and what is the rationale for the comparison?
18. A sample of size $n = 15$ includes $S = 75$ segregating nucleotide sites. Assuming that the sample conforms to the expectations of the neutral infinite-sites model at steady state, what is the expected number of singleton nucleotide polymorphisms? What is the expected number of nonsingleton nucleotide polymorphisms?
19. For neutral coalescence, show that the expected fraction of polymorphic nucleotides that are singletons equals $1/a$, where $a = 1 + (\frac{1}{2}) + (\frac{1}{3}) + \ldots + [1/(n-1)]$ and n is the number of alleles in the sample. Calculate this fraction for $n = 2, 5, 10, 20, 50$, and 100.
20. The accompanying illustration shows a coalescent tree for a sample of size $n = 3$. What are the expected lengths, in units of generations, along each of the labeled branches? Show that the expected total length of all the branches taken together equals $4Na$, where $a = 1 + \frac{1}{2} = \frac{3}{2}$. Show also that the expected length of all the external branches equals $4N$ and that the expected length of all the internal branches equals $4N(a-1)$. Use the principle that the expected time in generations for k neutral alleles to coalesce into $k-1$ alleles equals $4N/[k(k-1)]$.



21. One might naively think that samples from an infinite-alleles neutral model should contain roughly equal numbers of the alleles represented. But this is far from the truth. The expected sample configurations are very unequal, because the representation of each allele depends on the time in evolutionary history when it was created by mutation and the manner in which its frequency was affected by random genetic drift. To take a specific example, consider a sample of size $n = 6$ from a population evolving according to the infinite-alleles neutral model, and suppose that the sample contains only $k = 2$ different alleles. Let the configuration of alleles in the sample be represented as $(a_1, a_2, a_3, a_4, a_5)$, where a_i is the number of alleles represented exactly i times, with $\Sigma i a_i = 6$. It can be shown from Ewen's sampling formula that the probability of the configuation $(a_1, a_2, a_3, a_4, a_5)$ equals

$$\Pr\{a_1, a_2, a_3, a_4, a_5 \mid k = 2\} = \frac{6!}{274 \times 1^{a_1} 2^{a_2} 3^{a_3} 4^{a_4} 5^{a_5} \times a_1! a_2! a_3! a_4! a_5!}$$

(Equation 9.30 in Ewens 2004). In this case only three sample configurations are possible, namely $x = (1, 0, 0, 0, 1)$, $y = (0, 1, 0, 1, 0)$, and $z = (0, 0, 2, 0, 0)$. Calculate the probabilities of x, y, and z, and the expected proportion of samples in which the numbers of the two alleles are not equal.

22. For the infinite-alleles neutral model, the probability that a sample of size $n = 6$ contains exactly $k = 3$ alleles in the configuration (a_1, a_2, a_3, a_4) is given by

$$\Pr\{a_1, a_2, a_3, a_4 \mid k = 3\} = \frac{6!}{225 \times 1^{a_1} 2^{a_2} 3^{a_3} 4^{a_4} \times a_1! a_2! a_3! a_4!}$$

where a_i is the number of alleles represented i times in the sample, and $\Sigma i a_i = 6$ (Equation 9.30 in Ewens 2004). What sample configurations (a_1, a_2, a_3, a_4) are possible, and what are their probabilities?

5.1 Selection in Haploid Organisms 200

Discrete Generations 200
Continuous Time 204
Change in Allele Frequency in Haploids 205
Darwinian Fitness and Malthusian Fitness 206

5.2 Selection in Diploid Organisms 206

Change in Allele Frequency in Diploids 208
Marginal Fitness and Selection with Multiple Alleles 212
Application to the Evolution of Insecticide Resistance 215

5.3 Equilibria with Selection 215

Overdominance 216
Local Stability 220
Heterozygote Inferiority 222
Stable Equilibria with Multiple Alleles 223
Adaptive Topography and the Role of Random Genetic Drift 225

5.4 Mutation-Selection Balance 226

Equilibrium Allele Frequencies 226
The Haldane-Muller Principle 228

5.5 More Complex Types of Selection 229

Differential Selection in the Sexes 229
X-Linked Genes 230
Frequency-Dependent Selection 230
Density-Dependent Selection 231
Fecundity Selection 231
Age-Structured Populations 232
Heterogenous Environments and Clines 232
Diversifying Selection 234
Gametic Selection 236
Meiotic Drive 236
Multiple Loci and Gene Interaction: Epistasis 239
Evolution of Recombination Rate 241
Sexual Selection 241
Kin Selection 243

5.6 Interdeme Selection in Geographically Subdivided Populations 245

5.7 Selection in a Finite Population 248

Weak Selection and the Nearly Neutral Theory 248
Genetic "Draft" 251

CHAPTER 5

DARWINIAN SELECTION

Thus far in this book, the term *natural selection* has been used in the informal, intuitive sense used by Darwin in *The Origin of Species* (1859):

> Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection.

Modern formulations of natural selection are less literary and usually compacted into a form resembling a logical syllogism:

- In all species, more offspring are produced than can possibly survive and reproduce.
- Organisms differ in their ability to survive and reproduce—in part because of differences in genotype.
- In every generation, genotypes that promote survival in the current environment are present in excess at the reproductive age and thus contribute disproportionately to the offspring of the next generation.

Through natural selection, therefore, alleles that enhance survival and reproduction increase gradually in frequency from generation to generation,

and the population becomes progressively better able to survive and reproduce in the environment. The progressive genetic improvement in populations resulting from natural selection constitutes the process of **evolutionary adaptation**.

In the brief description of natural selection quoted above, Darwin uses the term *individual* three times. The unit of selection is the individual organism—not the species, not the subpopulation, not the sibship. It is the performance of the individual organism that matters. Each individual organism competes in the struggle for existence and survives or perishes on its own. Darwin also used the terms "struggle for existence" and "survival of the fittest" as synonyms for natural selection, but he emphasized that he employed the terms in their widest metaphorical sense to include not only the life of the organism but also the success of the organism in leaving progeny: fecundity is as important as survival. In this chapter, we shall see how Darwin's concept of "survival of the fittest" of individual organisms has been made more formal and quantitative and incorporated into models describing the change in allele frequency under natural selection. These models show that natural selection acts simultaneously on different components of fitness and can operate at different levels of population structure. The modern view of natural selection departs a bit from Darwin's view in acknowledging that natural selection may act at haploid stages, at diploid stages, on pairs of mating genotypes, and probably much more weakly, on groups of individuals.

5.1 SELECTION IN HAPLOID ORGANISMS

Selection acts on the phenotype, not on the genotype, and the total phenotype is determined by many genes that interact with each other as well as with numerous environmental factors. However, in exploring the consequences of selection, it is convenient to focus on changes in the frequency of the alleles of a single gene. We shall begin by examining selection in its simplest form operating in a haploid, asexual organism, such as a species of bacteria. In haploids, selection is realized as differential population growth. The overall process of selection is identical whether population growth is in discrete or continuous generations, but the models have a somewhat different parameterization and it is necessary to relate the models to avoid confusion later.

Discrete Generations

Consider two bacterial genotypes, A and B, that reproduce asexually. For simplicity, we will assume a discrete model of geometric population growth such that $A_t = (1 + a)^t A_0$ and $B_t = (1 + b)^t B_0$, where A_t and B_t are the number of cells of genotype A and genotype B, respectively, at time t. Selection takes place when $a \neq b$. Figure 5.1A is an example in which the growth rates of A and B are $a = 0.41$ and $b = 0.26$, respectively. Both populations increase in size exponentially, but that of A increases faster than that of B. In most cases,

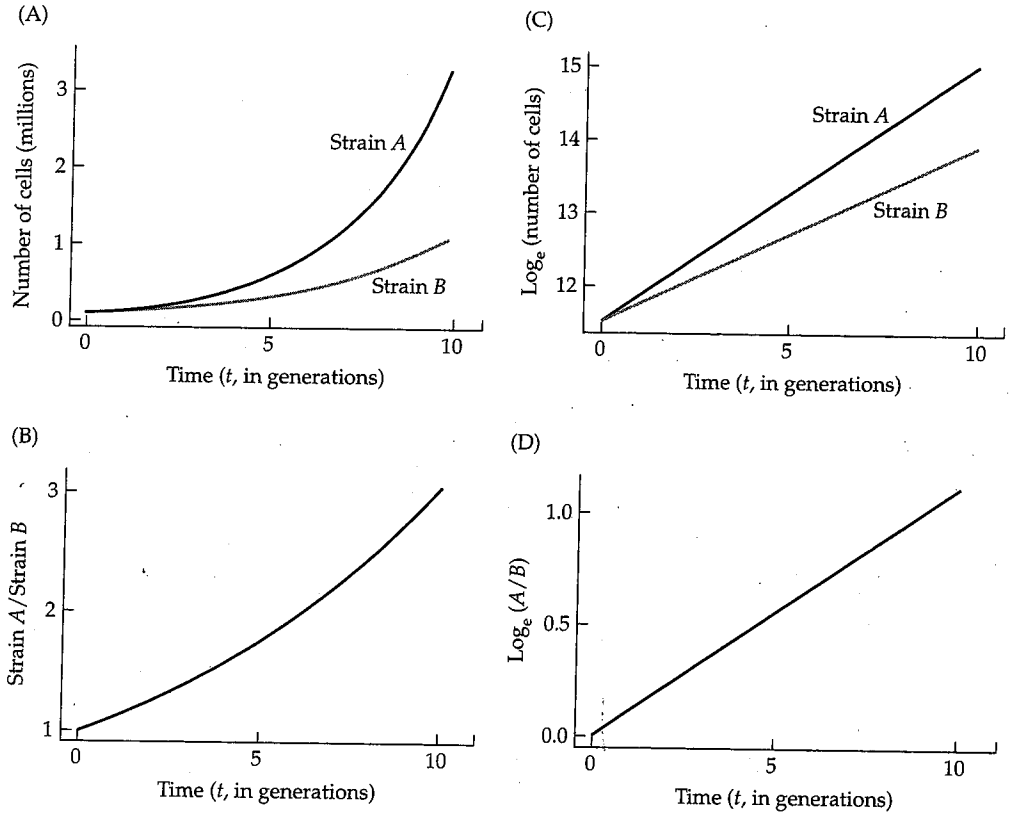


FIGURE 5.1 (A) Population growth of two hypothetical bacterial strains, A and B, in which the growth rates are 41% per generation for A and 26% per generation for B. The initial cell numbers are 10^5 for A and 10^5 for B. (B) Ratio of cell numbers of $A : B$. Because the A population grows faster than the B population, the proportion of A in the total population increases. (C) and (D) are the growth trajectories and their ratio on a log scale.

we are not interested in the actual number of A cells or B cells but in the proportion of all cells that are of type A. Equivalently, we can examine the ratio of the number of A cells to that of B cells at time t, which is given by

$$\frac{A_t}{B_t} = \left(\frac{1+a}{1+b}\right)^t \frac{A_0}{B_0} = w^t \left(\frac{A_0}{B_0}\right) \qquad [5.1]$$

The outcome of selection is determined by the ratio of a to b because, if $a < b$, then the ratio of A cells to B cells decreases until, ultimately, A is lost; conversely, if $a > b$, then the ratio of A cells to B cells increases without limit. Figure 5.1B shows the change in A/B for the example in part A. From a value of 1 at the beginning, the ratio increases to a value of 3 in 10 generations;

these ratios correspond to frequencies of A of 0.50 and 0.75, respectively. Whenever there is geometric growth, plotting on a log scale (Figure 5.1C and D) produces straight lines, and this is often useful for statistical analysis (e.g., estimation of a and b from data).

In Figure 5.1 it is not necessary to specify whether a and b differ because of survivorship or fecundity. All that matters is that they do differ. It is also important that the outcome depends only on the ratio $(1 + a)/(1 + b)$, which means that, in practice, we do not need to know the absolute growth rates of A and B but only their relative values (their ratio). In Equation 5.1, w represents the ratio $(1 + a)/(1 + b)$. The symbol w is conventionally used in discrete models of selection and, in this example, it is the *relative fitness* of genotype A to that of genotype B. In other words, in a haploid organism, the relative fitness equals the ratio of the growth rates.

Although it is sometimes instructive to do so, it is not necessary to keep track of population size in models of selection. The variable of interest is usually the allele frequency and not the population size. Therefore, let p_t and q_t represent the frequencies of genotypes A and B, respectively, in generation t, with $p_t + q_t = 1$. A method to relate the frequencies of A and B in any two successive generations is illustrated in Table 5.1. For ease of discussion, we divide each generation into three phases: birth, selection, and reproduction. In generation $t - 1$, the frequencies of A and B at birth are p_{t-1} and q_{t-1}, respectively. The genotypes A and B are assumed to survive in the ratio $w : 1$, which means that w is the probability of survival of an A genotype relative to that of an B genotype. As before, the absolute probabilities of survival of the genotypes are not relevant. All that matters is the ratio. After selection, the ratio of frequencies of $A : B$ equals $p_{t-1} \times w : q_{t-1} \times 1$. If the surviving genotypes repro-

TABLE 5.1 A Model of Selection in a Haploid Organism, in which w is the Probability of Survival of an A Cell Relative to that of a B Cell

	Genotype	
Generation $t - 1$	A	B
Frequency before selection	p_{t-1}	q_{t-1}
Relative fitness	w	1
After selection	$p_{t-1}w$	q_{t-1}
Generation t	$\dfrac{p_{t-1}w}{p_{t-1}w + q_{t-1}}$	$\dfrac{q_{t-1}}{p_{t-1}w + q_{t-1}}$

Note: The fractions in the bottom line are expressions for the allele frequencies in generation t in terms of those in generation $t - 1$. Although this model assumes differential survival, $w : 1$ could also be the relative probability of reproduction of A and B. More generally, the relative fitness $w : 1$ represents the net output of $A : B$ for the combined effects of differential survival and reproduction.

duce with equal efficiency, then the frequencies at birth in the following generation are given by the expressions across the bottom in Table 5.1; the denominators in these expressions are necessary to make the allele frequencies in generation t sum to 1.

For comparison with Equation 5.1, consider that p_t is the number of A cells in generation t divided by the total; likewise, q_t is the number of B cells divided by the total. Therefore, the ratio p_t/q_t equals the ratio of A cells to B cells in generation t because the denominators cancel. The expressions in Table 5.1 imply that the ratio of p/q in any generation equals w multiplied by the ratio of p/q in the previous generation, and so

$$\frac{p_t}{q_t} = w\frac{p_{t-1}}{q_{t-1}} = w^2\frac{p_{t-2}}{q_{t-2}} = \cdots = w^t\frac{p_0}{q_0} \tag{5.2}$$

The right-hand side of Equation 5.2 is identical to that in Equation 5.1 except that the relative frequencies p and q replace the absolute number of cells of type A and type B. Hence, to deduce the outcome of selection, we do not need to keep track of population size. All we need to know is the relative fitness w and the initial frequencies p_0 and q_0.

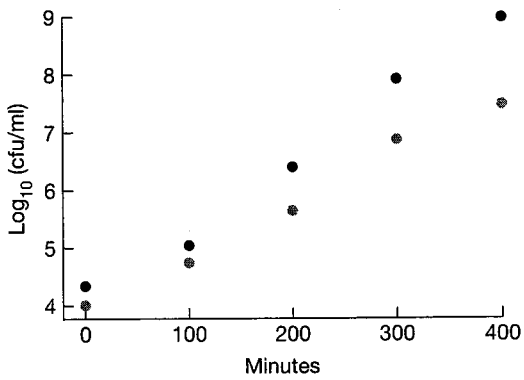
For application to experimental data, Equation 5.2 is often transformed by taking the natural (base e) logarithm:

$$\log\left(\frac{p_t}{q_t}\right) = \log\left(\frac{p_0}{q_0}\right) + t\log(w) \tag{5.3}$$

Equation 5.3 means, for example, that if the values of p_t/q_t are monitored in an experimental population of bacteria over the course of time, then a plot of log (p_t/q_t) against time (in generations) should yield a straight line with slope equal to log w (see Figure 5.1D). This kind of experiment is examined in the following problem.

PROBLEM 5.1 Methicillin-resistant *Staphyloccocus aureus* (MRSA) is a serious pathogen that has seen a particularly rapid spread, and with the spread has come an increase in the diversity of MRSA strains. One particular subclass of MRSA strains that spread especially fast is gentamicin sensitive (GS-MRSA). Laurent et al. (2001) reported experiments designed to test the relative growth rates of GS-MRSA and the older gentamicin-resisitant strains (GR-MRSA). Rather than use a chemostat, they simply grew up the strains in 200-ml flasks of medium, taking samples at intervals to measure density in colony forming units per ml (cfu/ml). From the table of $\log_{10}$(cfu/ml) at different times, calculate the relative growth rates of the two strains in the intervals 0–100 minutes and 300–400 minutes (assume a generation time of 100 minutes):

Min	GR-MRSA	GS-MRSA
0	4.000	4.322
100	4.708	5.041
200	5.633	6.398
300	6.669	7.908
400	7.462	8.968



ANSWER First calculate the proportion of the mixed culture that consists of each strain at each time interval, and get 0.6774, 0.6832, 0.8532, 0.9427, 0.9698 as the proportions that are GS-MRSA at the respective times. Then notice that the natural $\log(p_{100}/q_{100})$ = 0.76865 and $\log(p_0/q_0)$ = 0.74194, giving a difference of 0.0267 = $\log(w)$. In the first 100 minutes, the fitness of GS-MRSA relative to GR-MRSA is $e^{0.0267}$ = 1.027, or a 2.7 percent advantage. In the last 100 minutes, we get $\log(p_{300}/q_{300})$ = 2.80106 and $\log(p_{400}/q_{400})$ = 3.46789. The difference is now 0.6668, so the relative fitness is now $e^{0.6668}$ = 1.95. Early in the experiment the GS-MRSA strain appears not to have been growing in its maximal or log phase, but later there is nearly a twofold growth advantage to the GS-MRSA strain. If we use all the data we get $\log(w)$ = 0.748 per generation. This gives $w = e^{0.748}$ = 2.11. You can see from the data that the ratio of GS-MRSA/GR-MRSA increased 16 fold in four generations, consistent with roughly a twofold growth advantage of GS-MRSA.

Continuous Time

Bacterial populations such as those in Problem 5.1 do not reproduce in discrete generations but instead they reproduce continuously. In a continuous model, the exponential population growth of A and B are governed by the equations $dA_t/dt = a'A_t$ and $dB_t/dt = b'B_t$, where a' and b' are the growth rates. Therefore, $A_t = A_0 \exp^{a't}$ and $B_t = B_0 \exp^{b't}$, and so

$$\frac{A_t}{B_t} = \frac{A_0}{B_0} e^{(a'-b')t} = \frac{A_0}{B_0} e^{mt} \qquad [5.4]$$

Equation 5.4 means that, in a continuous population, the outcome of selection depends on the difference between the exponential growth rates $a'-b'$, which is represented by the symbol m on the right-hand side. The value of

m also measures the relative fitness of strain A to strain B, but in a continuously reproducing population. Comparing Equation 5.4 with Equation 5.1 yields the relation between m and w:

$$m = \ln w \tag{5.5}$$

In other words, the relative fitness with continuous growth m equals the natural logarithm of the relative fitness with discrete reproduction w. Selective neutrality means that $w = 1$ or that $m = 0$. For the values of w estimated in Problem 5.1, the corresponding values of m are 0.0267 and 0.6668, respectively. If w is not too different from 1, then $m = w - 1$ is a reasonable approximation.

Change in Allele Frequency in Haploids

Although the discrete and continuous models are completely equivalent under the transformation in Equation 5.5, the equations for change in allele frequency look rather different. In the discrete model, the change in the frequency of strain A in generation t is given by the difference $p_t - p_{t-1}$, which can be calculated in terms of p_{t-1} from the formulas in Table 5.1. The difference $p_t - p_{t-1}$ is usually symbolized Δp and, for simplicity, the subscripts are usually suppressed. Using the expressions in Table 5.1 and the fact that $q = 1 - p$, we obtain

$$\Delta p = \frac{pw}{pw+q} - p = \frac{pq(w-1)}{pw+q} \tag{5.6}$$

Not surprisingly, p increases if the relative fitness of A is greater than 1 and decreases if the relative fitness of A is smaller than 1. If the relative fitnesses of A and B are equal, then p does not change—provided that the population size is very large (theoretically, it has to be infinite).

The analog of Equation 5.6 in a continuous model contains the derivative dp/dt in place of Δp. This we can obtain from Equation 5.4 with a little trickery. Because A_t/B_t equals p_t/q_t, the derivative of Equation 5.4 with respect to t must equal the derivative of p_t/q_t with respect to t. For simplicity, we will write p and q instead of p_t and q_t. The derivative of Equation 5.4 with respect to t equals mp/q and the derivative of p/q with respect to t equals $(1/q^2)(dp/dt)$. Setting these expressions equal to each other and solving for dp/dt, we obtain

$$\frac{dp}{dt} = pqm \tag{5.7}$$

Where did the denominator go? In a technical sense, it disappeared into the difference between the discrete model and the continuous model. In a practical sense, the absence of a denominator in Equation 5.7 greatly simpli-

fies some of the formulas used to describe the effects of selection. Although they look very different, Equations 5.6 and 5.7 are merely different ways of saying the same thing. In this chapter, we will deal mainly with expressions analogous to Equation 5.6 because they are more easily derived for various types of selection. However, when it is necessary to dispose of a troublesome denominator, we will invoke the continuous model in Equation 5.7 and be rid of it.

Darwinian Fitness and Malthusian Fitness

The distinction between the fitness parameters in the discrete and continuous models has been incorporated into the terminology of population genetics in the terms **darwinian fitness**, which refers to the discrete model, and **malthusian fitness**, which refers to the continuous model. The latter is named after Thomas Malthus (1766–1834), whose views on the implications of continued population growth strongly influenced Darwin's thinking on the subject. A darwinian fitness is conventionally represented by the symbol w, often embellished with a subscript, and malthusian fitness is conventionally represented by the symbol m. In this book, the term *fitness*, when used without qualification, will mean darwinian fitness unless it is clear from the context that some other meaning is intended.

5.2 SELECTION IN DIPLOID ORGANISMS

In haploid organisms, some cells are better at avoiding death, or are able to acquire nutrients more efficiently, or divide faster. All these attributes can be collapsed into a single fitness parameter that expresses the differential growth rate of that haploid genotype. In diploid organisms, the consequences of selection can be manifested in more complicated ways. Starting with the fusion of gametes, the zygote has to divide and grow and develop. From one genotype to another there might be different rates of development and different probabilities of survival to the adult stage. We call this component of selection **viability**. The adults must then be successful at attracting mates, and the differences among genotypes in this ability is manifested as **sexual selection**. When heterozygous individuals produce gametes, there can be departures from Mendelian segregation, and this results in a form of selection called **meiotic drive** or **segregation distortion**. Many marine organisms release gametes in the sea, and there might be differential survival of these haploid cells, in a form of selection called **gametic selection**. Once matings occur, genotypes may produce different numbers of offspring, and we call this **fecundity selection**. To make matters more complicated, the number of progeny that are produced by a pair of genotypes might not be expressed as a simple sum of fecundity effects of each genotype, but might instead have to be expressed as a property of the mating pair of genotypes.

Later on in the chapter we will consider some of the consequences of considering these various components of selection. For many organisms, it appears that differential fecundity may actually be the most important fitness component, but for now, the easiest model to consider first is that of viability selection.

To develop a model of natural selection in diploids, we will start with the model of random mating in Chapter 2, but we incorporate selection by permitting the fitnesses of the genotypes to differ. Recall that in this model we made many assumptions – that the population was infinite so that no genetic drift occurs, that there are no new mutations or migration, and that genotypes mate at random. To these assumptions we now add that fitnesses are constant and selection is assumed to take place only through differential zygote-to-adult survival of the diploid genotypes. We shall use the conventional symbols w_{11}, w_{12}, and w_{22} to represent the darwinian fitnesses of the genotypes *AA*, *Aa*, and *aa*, respectively. If the fitness of each genotype is set equal to its probability of survival, then each fitness is an **absolute fitness** because its value is independent of the fitnesses of the other genotypes. In practice, we usually know only the value of the viability of each genotype relative to that of another genotype chosen as the standard of comparison. When a fitness value is expressed relative to that of another genotype, the fitness is a **relative fitness**. The relative fitness of the genotype chosen as the standard of comparison is arbitrarily assigned the value 1.

To consider a specific example, suppose that the genotypes *AA*, *Aa*, and *aa* have probabilities of survival from conception to reproductive age of 0.75, 0.75, and 0.50, respectively. These are the absolute viabilities of the genotypes. They can be judged realistic or not only if we specify the organism. They may be plausible values if the organism is a mammal or a bird because each offspring has a reasonable chance of survival, but implausible if the organism is an insect or an oyster because, in these organisms, most newborns are destined not to survive. Because selection depends on the relative magnitudes of the viabilities, it is usually most convenient to express the viabilities in relative terms. Taking genotype *AA* as the standard, the relative viabilities of *AA*, *Aa*, and *aa* are 0.75/0.75, 0.75/0.75, and 0.50/0.75, or 1.0, 1.0, and 0.67, respectively. Equivalently, we could choose genotype *aa* as the standard, in which case the relative viabilities are 0.75/0.50, 0.75/0.50, and 0.50/0.50, or 1.5, 1.5, and 1.0, respectively. Usually, the relative viabilities are calculated so that the largest relative viability equals 1.0. The relative viabilities are equal to the relative fitnesses of the genotypes provided that the genotypes are equally capable of reproduction. Viabilities expressed in relative terms are as valid for osprey as for oysters because the relative fitnesses are the same whether the absolute fitnesses are 0.75, 0.75, and 0.50 or 0.00075, 0.00075, and 0.00050. We shall see that the dynamics of how the allele frequencies change depend only on the relative fitnesses.

PROBLEM 5.2 A number of mutations in laboratory organisms have been discovered that lengthen the lifespan of the organisms. One particularly dramatic example is *daf-2*, a gene in the insulin signaling pathway of *Caenorhabditis elegans* which, when mutated, can double the lifespan of the worms. Jenkins et al. (2004) hypothesized that there must be negative effects to the *daf-2* mutation; otherwise such a mutation would have gone to fixation in the population. They sought to measure the total fitness differences between wildtype and *daf-2* mutant worms. They started six replicate populations with equal frequencies of the two homozygous genotypes ($p_0 = 0.5$) and observed frequencies of the *daf-2* allele of 0.28, 0.09, 0.02, and 0 in generations 1, 2, 3, and 4. What was the net fitness of *daf-2* relative to wildtype? (Hint: *C. elegans* is diploid but is a selfing hermaphrodite almost all the time, so these two genotypes can be treated as haploid clones.)

ANSWER For the first generation, the frequency of *daf-2* fell from 0.5 to 0.28, and we obtain natural log(0.28/0.72) = log(0.5/0.5) + log(w). Solving for w, we get 0.389. Likewise, for the transitions from generations 1 to 2, and 2 to 3, we get estimates of $w = 0.254$ and $w = 0.206$. If we do a linear regression of log(p/q) against generation, the slope is −1.3 for an average fitness estimate throughout the experiment of $e^{-1.3} = 0.27$. The *daf-2* allele really does much worse than the wildtype allele, and would be very rapidly eliminated, as it was in the lab population. This example confirms the old adage that if something sounds too good to be true, it probably is too good to be true.

Change in Allele Frequency in Diploids

If we write the allele frequencies of A and a as p_t and q_t, respectively, in generation t, then it is straightforward to derive expressions for the allele frequencies in generation t in terms of the allele frequencies p_{t-1} and q_{t-1} in the previous generation. The subscripts t and $t-1$ are rather cumbersome to carry along in equations, so we will use the symbols p and q for p_{t-1} and q_{t-1}, and the symbols p' and q' for p_t and q_t.

The relation between the allele frequencies in two consecutive generations is deduced in Table 5.2, where the fitnesses w_{11}, w_{12}, and w_{22} are the relative viabilities. In generation $t-1$, the genotype frequencies of AA, Aa, and aa among newly fertilized eggs are given by p^2, $2pq$, and q^2, respectively, assuming random mating. By definition, newly fertilized eggs survive in the ratio $w_{11} : w_{12} : w_{22}$, and so the ratio of $AA : Aa : aa$ among surviving adults is

$$p^2 w_{11} : 2pq w_{12} : q^2 w_{22}$$

To proceed, we need to convert the terms in the above expression into relative frequencies by dividing each term by the sum. The value of the sum is indicated in Table 5.2 as

$$\bar{w} = p^2 w_{11} + 2pq w_{12} + q^2 w_{22} \tag{5.8}$$

TABLE 5.2 Diploid Selection for Survivorship (Viability)

	Genotype			Total
Generation $t-1$	AA	Aa	aa	
Frequency before selection	p^2	$2pq$	q^2	$1 = p^2 + 2pq + q^2$
Relative fitness (viability)	w_{11}	w_{12}	w_{22}	
After selection	p^2w_{11}	$2pqw_{12}$	q^2w_{22}	$\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$
Normalized	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$	
Generation t	$p' = \frac{p^2w_{11} + pqw_{12}}{\bar{w}}$			
	$q' = \frac{pqw_{12} + q^2w_{22}}{\bar{w}}$			

Note: The allele frequencies p and q are those in gametes immediately prior to fertilization. The AA, Aa, and aa zygotes survive to reproductive maturity in the ratio $w_{11} : w_{12} : w_{22}$. All genotypes, as adults, are assumed to have the same reproductive capacity.

The symbol $\bar{w}$ is the mean fitness in the population in generation $t-1$, and it is simply the average fitness of all individuals in the population. Division of each term in the ratio of survivors by $\bar{w}$ yields the genotype frequencies among adults:

$$AA : \frac{p^2w_{11}}{\bar{w}} \quad Aa : \frac{2pqw_{12}}{\bar{w}} \quad aa : \frac{q^2w_{22}}{\bar{w}} \tag{5.9}$$

Among the surviving adults, the AA genotypes produce all A gametes, the Aa genotypes produce $\frac{1}{2}A$ and $\frac{1}{2}a$ gametes, and the aa genotypes produce all a gametes. Hence, the frequencies of the gametes that unite at random to form the zygotes of the next generation are:

$$p' = \frac{p^2w_{11} + pqw_{12}}{\bar{w}} \quad q' = \frac{pqw_{12} + q^2w_{22}}{\bar{w}} \tag{5.10}$$

These are the relations we were after because they express the allele frequencies in any generation in terms of the allele frequencies in the previous generation. From these equations, the outcome of selection can be deduced. Because $q' = 1 - p'$, the recursion that describes how the allele frequencies change is fully specified by giving only the equation for p'.

As in the haploid model, it is often useful to know Δp, which is the difference in allele frequency $p' - p$ resulting from one generation of selection. Subtraction of p from the expression for p' in Equation 5.10 and a little manipulation leads to:

$$\Delta p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{\bar{w}} \tag{5.11}$$

Equation 5.11 is the diploid analog of that in the haploid model in Equation 5.6.

At this point, an example of the use of these equations is in order. We will use data on the change in the frequency of the *Gl* (*Glued* eyes) allele in a laboratory population of *Drosophila melanogaster*, which are plotted in Figure 5.2. The *Gl* allele is lethal when homozygous, so $w_{11} = 0$. The points in Figure 5.2 pertain to the frequency of *Gl* heterozygotes but, because *Gl/Gl* genotypes do not survive, the allele frequency p of *Gl* equals one-half the frequency of *Gl*/+ adults. The points in the figure are each separated by one generation,

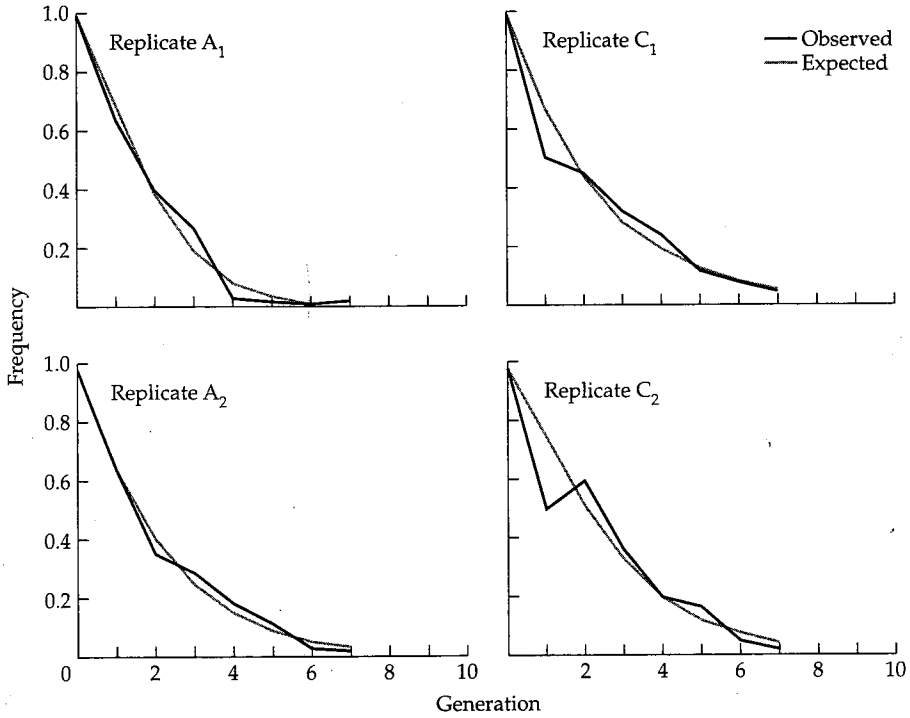


FIGURE 5.2 Change in frequency of adult *Drosophila melanogaster* heterozygous for the dominant mutation *Gl* (*Glued* eyes) in an experimental population. The genotype *Gl/Gl* is lethal. The curves represents the theoretical change in frequency when the ratio of viabilities of *Gl*/+ to +/+ is estimated from the data. The four plots are four independent population replicates. (From Clegg et al. 1976.)

and the initial generation has a frequency of $Gl/+$ adults of 0.67; hence $p_0 = 0.335$ and thus $q_0 = 0.665$. The data from Figure 5.2 were used to estimate the fitnesses of the $Gl/+$ vs. $+/+$ genotypes, and this estimation was done separately for each replicate, giving relative fitness values of 0.383, 0.573, 0.693 and 0.559, for an average of 0.538, relative to a value of $w_{22} = 1.0$ for $+/+$ genotypes. Substituting these values for p, q, w_{11}, w_{12}, and w_{22} into the expression for p' in Equation 5.10 yields

$$p' = \frac{0.335^2 \times 0 + 0.335 \times 0.665 \times 0.538}{0.335^2 \times 0 + 2 \times 0.335 \times 0.665 \times 0.538 + 0.665^2 \times 1} = 0.292$$

Therefore, the predicted frequency of $Gl/+$ adults in generation 1 is $2p' = 0.584$, which is reasonably close to the observed values that range from 0.504 to 0.646 among replicates.

Note that the four panels of Figure 5.2 are rather different, even though they are supposed to be replicates of each other. Statistical tests showed that they are in fact *not* consistent with the same model of selection. Clegg et al. (1976) set up this experiment to follow not only the dynamics of the *Glued* allele itself, but also to follow the way that linked genes change through a process known as *genetic hitchhiking*. Interestingly, the linked genes did not follow a simple trajectory of allele frequencies that one would predict from the *Glued* lethal allele and recombination. Instead, the linked genes changed frequency in a way that could only be explained by additional selected alleles in the genetic background. An important consideration for any experiment dealing with natural selection in the laboratory is that not all variable can be measured, and other unobserved loci are probably under selection too!

PROBLEM 5.3 Balancer chromosomes with multiple inversions have been important in *Drosophila* population genetics because they allow the investigator to isolate single chromosomes from natural populations for study. Dobzhansky and Spassky (1963) estimated the relative viability of 1063 wild-derived *D. pseudoobscura* second-chromosome homozygotes compared to heterozygotes with a balancer, and found no correlation between the homozygous viability and heterozygous viability for random pairs of these wild chromosomes. In order to get at total fitness, the full dynamics of allele frequencies can be followed by observing changes in allele frequency cage populations. Sved and Ayala (1970) developed an estimator of total fitness of whole chromosomes in *Drosophila* that made use of the same kind of balancer chromosome system, but followed the frequencies in cage populations. Because the homozygotes for the balancer (*Ba/Ba*) were lethal, the only two surviving genotypes were *Ba/+* and +/+. In one cage, the frequency of the *Ba/+* heterozygotes was 0.486, and in the next generation this frequency was 0.726. What is the net fitness of *Ba/+* relative to +/+?

ANSWER The allele frequency for the balancer is one-half the heterozygote frequency, so the allele frequency changed from $p = 0.486/2 = 0.243$ to $0.726/2 = 0.363$. If we let the *Ba*/+ genotype have fitness w relative to the +/+ genotype (*Ba*/*Ba* has fitness zero), then the recursion for the wild-type allele is $q' = (wpq + q^2)/(2pqw + q^2)$. Substituting, we get $0.637 = [w(0.243)(0.757) + (0.757)^2]/(w2(0.243)(0.757) + (0.757)^2]$. Solving for w, we get 4.13. Another way to put this is to say that the relative fitness of the wild-type homozygote relative to the balancer heterozygote is $1/(4.13) = 0.242$. Sved and Ayala took pains to try to decompose fitness into viability and fecundity components, and showed that both played an important role in determining fitness differences between second chromosome.

Figure 5.3 shows another important aspect of directional selection. When a new mutation is advantageous, it increases in frequency initially much faster when it is dominant than when it is recessive. The reason is that a rare allele is almost entirely found in heterozygotes, and if the mutation is completely recessive, then its effects on fitness are not manifested nearly to the degree that they are if the allele were dominant. When the allele becomes common, the tables are turned and the dominant allele's frequency changes very slowly, but the common recessive allele changes rapidly in frequency. The same reasoning applies: When a common dominant allele is present in a population, almost all fitnesses are the same, so selection changes allele frequency slowly.

By substituting in the formulae for p' and for q' into $\overline{w}$ one can derive an equation for $\overline{w}'$ that gives a recursion for how the mean population fitness changes over generations. Further algebra shows that $\overline{w}$ in the one-locus two-allele model of selection with constant fitnesses is nondecreasing. One can also show that $\overline{w}$ sits at a local maximum value whenever the population is at equilibrium allele frequencies. This property of the mean fitness is part of Fisher's (1930) *fundamental theorem of natural selection*, which states that the rate of increase in the mean fitness of a population ascribable to gene frequency changes is exactly equal to the additive genetic variance in fitness. It was not until Ewens (1989) that there was a convincing proof of Fisher's principle, in part because of the subtle interpretation of Fisher's statement of the theorem (see Edwards 2002). Somewhat surprisingly, the mean fitness does not always increase in all models, and the equilibrium is often not a maximum of mean fitness.

Marginal Fitness and Selection with Multiple Alleles

We make a slight digression to point out that it is sometimes convenient to think in terms of the marginal fitnesses of the A and a alleles. The **marginal fitness** equals the average fitness of all genotypes containing A or a, respectively, weighted by their relative frequency and the number of A or a alleles they contain. For example, A alleles are found in AA and Aa genotypes in the

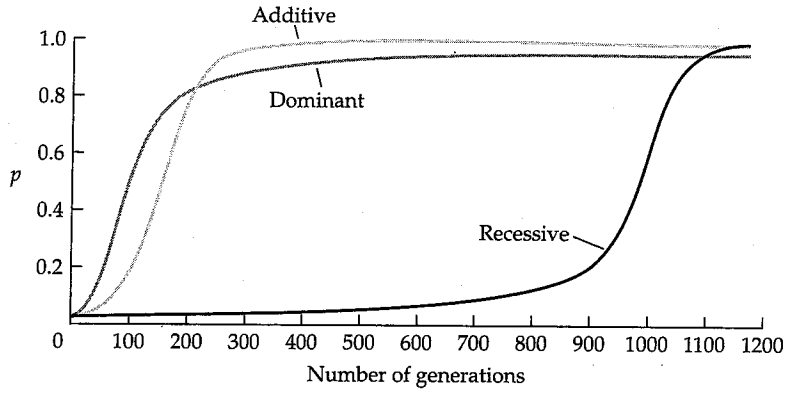


FIGURE 5.3 The change in frequency p of a favorable allele that is either dominant, additive, or recessive in its effect on fitness. The frequency of a favored dominant allele changes most slowly when the allele is common, and the frequency of a favored recessive allele changes most slowly when the allele is rare. In all three examples, the difference in relative fitness between the homozygous *AA* and *aa* genotypes is assumed to be 5%.

proportions p and q and, therefore, the marginal fitness $\overline{w}_1$ of *A*-containing genotypes equals $pw_{11} + qw_{12}$. Similarly, the marginal fitness of *a*-containing genotypes is $\overline{w}_2 = pw_{12} + qw_{22}$. The expression for p' in Equation 5.10 thus becomes $p' = p\,\overline{w}_1 / \overline{w}$, and Equation 5.11 becomes $\Delta p = p(\overline{w}_1 - \overline{w}) / \overline{w}$. This expression makes it clear that any allele increases in frequency if the marginal fitness of genotypes containing the allele ($\overline{w}_1$) is greater than the average fitness in the population ($\overline{w}$). This approach also generalizes readily to multiple alleles: For an allele with frequency p_i and marginal fitness $\overline{w}_i$, the change in frequency in one generation equals

$$\Delta p_i = \frac{p_i(\overline{w}_i - \overline{w})}{\overline{w}} \tag{5.12}$$

Equation 5.12 has some interesting consequences. Note that one condition for equilibrium is that $\overline{w}_i = \overline{w}$. This implies that the marginal fitness of every allele is the same. In the case of three alleles, the marginal fitnesses can be written:

$$\overline{w}_1 = p_1 w_{11} + p_2 w_{12} + p_3 w_{13}$$
$$\overline{w}_2 = p_1 w_{21} + p_2 w_{22} + p_3 w_{23}$$
$$\overline{w}_3 = p_1 w_{31} + p_2 w_{32} + p_3 w_{33}$$

This system of equations consists of three linear equations in three unknowns, and besides the three fixation equilibria [where the allele frequencies (p_1, p_2, and p_3) are either (1,0,0), (0,1,0) or (0,0,1)], this system has either no polymorphic equilibria, one unique equilibrium, or the full range of

allele frequencies are equilibria. (The latter is an aberrant case that happens, for example, when the fitnesses are all equal.) We will postpone statement of the conditions on the fitnesses necessary to maintain all three alleles in a stable equilibrium under this form of selection until we discuss the stability characteristics of the model.

PROBLEM 5.4 An extension of the use of population cages that allows estimation of heterozygote fitness makes use of a chromosome isolated from a natural population and two balancer chromosomes. Gardner et al. (2005) studied 40 different wild chromosomes using the balancers *TM1* and *TM2*. For the wild chromosomes that were lethal when homozygous, the only genotypes that survive are *+/TM1*, *+/TM2* and *TM1/TM2*, where *TM1* and *TM2* are balancer chromosomes (*TM* stands for "third multiply-inverted"). If the relative fitnesses of *+/TM1*, *+/TM2* and *TM1/TM2* are s, t, and u, respectively, find the equilibrium frequencies of + (p_1), *TM1* (p_2), and *TM3* (p_3). What are these frequencies when $s = 0.8$, $t = 0.3$, and $u = 0.7$?

ANSWER: First note that the fitnesses can be written in the form of a symmetric matrix $w_{ij} = w_{ji}$ as follows:

$$\begin{bmatrix} 0 & s & t \\ s & 0 & u \\ t & u & 0 \end{bmatrix}$$

At equilibrium each of the three marginal fitnesses is equal to the mean fitness, hence:

$$\bar{w} = 0p_1 + sp_2 + tp_3$$
$$\bar{w} = sp_1 + 0p_2 + up_3$$
$$\bar{w} = tp_1 + up_2 + 0p_3$$

This is system of three linear equations in three unknowns, and it can be solved by searching the Internet for a method called Cramer's rule. Cramer's rule gives the solutions as:

$$\hat{p}_j = D_j / \sum_{i=1}^{3} D_i$$

where D_1, D_2, and D_3 are the following matrix determinants. (At equilibrium $\bar{w}$ is a constant, and it cancels out in the ratios.)

$$D_1 = \begin{vmatrix} \bar{w} & s & t \\ \bar{w} & 0 & u \\ \bar{w} & u & 0 \end{vmatrix}$$

$$D_2 = \begin{vmatrix} 0 & \bar{w} & t \\ s & \bar{w} & u \\ t & \bar{w} & 0 \end{vmatrix}$$

$$D_3 = \begin{vmatrix} 0 & s & \bar{w} \\ s & 0 & \bar{w} \\ t & u & \bar{w} \end{vmatrix}$$

Solving these determinants, the equations for the $\hat{p}_j$ are

$$\hat{p}_1 = \frac{u(u-s-t)}{u(u-s-t)+t(t-u-s)+s(s-t-u)}$$

$$\hat{p}_2 = \frac{t(t-u-s)}{u(u-s-t)+t(t-u-s)+s(s-t-u)}$$

$$\hat{p}_3 = \frac{s(s-t-u)}{u(u-s-t)+t(t-u-s)+s(s-t-u)}$$

When $s = 0.8$, $t = 0.3$, and $u = 0.7$ these become $\hat{p}_1 = 0.35$, $\hat{p}_2 = 0.45$, and $\hat{p}_3 = 0.20$ These values reproduce almost exactly the equilibrium values observed with one of the wild chromosomes.

Application to the Evolution of Insecticide Resistance

Some of the most dramatic examples of evolution in action result from the natural selection for chemical pesticide resistance in natural populations of insects and other agricultural pests. In the 1940s, when chemical pesticides were first used on a large scale, an estimated 7% of the agricultural crops in the United States were lost to insects. Initial successes in chemical pest management were followed by gradual loss of effectiveness. Today, more than 400 pest species have evolved significant resistance to one or more pesticides, and 13% of the agricultural crops in the United States are lost to insects. Total cost and loss due to insects in 2005 were $1.264 billion (Robinson 2006). In many cases, significant pesticide resistance has evolved in 5 to 50 generations irrespective of the insect species, geographical region, pesticide, frequency and method of use, and other seemingly important variables. Details in actual examples depend on such factors as effective population number and extent of genetic isolation between local populations. The evolution of resistance caused by multiple interacting alleles may be expected to take somewhat longer than single-gene resistance.

PROBLEM 5.5 Resistance to organophosphate and carbamate insecticides in species of the mosquitoes *Culex* and *Anopheles* has been shown to be mediated by four independent mutations in the acetylcholinesterase gene *ace-1* (Weill et al. 2004). The mutations, resulting in G119S (a glycine replaced by a serine at position 119), makes the enzyme insensitive to inhibition by these insecticides. Because the novel form of the protein remains active, this is a gain-of-function mutation, and resistance is therefore dominant. If mosquitoes bearing the resistance allele are 10 times as likely to survive and reproduce as the sensitive allele (relative fitnesses 10:1), how long would it take the allele frequency to rise from 0.01 to 0.50?

ANSWER Substitute $w_{11} = w_{12} = 10$, $w_{22} = 1$, and $p = 0.01$ into Equation 5.10 and calculate the allele frequency the next generation. This calculation is especially easy to do on a spreadsheet, making columns for allele frequencies, genotype frequencies, mean fitness, and allele frequency the next generation. You will find that the allele frequencies are 0.0848, 0.3445, and 0.5617, so in just three generations the frequency exceeds 50%.

5.3 EQUILIBRIA WITH SELECTION

An **equilibrium value** of p in a discrete model is any value for which $\Delta p = 0$. When the allele frequency is at an equilibrium in an infinite population, the allele frequency remains the same generation after generation. Because real populations are finite in size, an allele frequency is subject to chance fluctuations and so cannot usually remain exactly at an equilibrium value. For any

equilibrium, therefore, it is important to consider how the allele frequency behaves when it is close, but not exactly equal, to the equilibrium value. Any equilibrium can be classified as one of several different types according to the behavior of the allele frequency when it is near the equilibrium:

- An equilibrium is said to be **locally stable** if the allele frequency, when it is already close to the equilibrium, moves progressively closer in subsequent generations. A locally stable equilibrium may also be **globally stable**. This term means that the allele frequency always moves toward the equilibrium regardless of where it starts, even if initially far away from the equilibrium. A polymorphism with a stable equilibrium is sometimes called a **balanced polymorphism**.
- An equilibrium is **unstable** if the allele frequency, initially close to the equilibrium, moves progressively farther away in subsequent generations.
- An equilibrium is called **neutrally stable** or *semistable* if the allele frequency has no tendency to change regardless of its initial value. In such a case, every allele frequency represents an equilibrium because $\Delta p = 0$ whatever the value of p. This type of equilibrium is exemplified by the Hardy-Weinberg principle in an infinite population (see Chapter 2).

The concepts of stability can be applied to the case of selection governed by Equation 5.11 in which A is the favored allele. For A to be favored, we need $w_{11} \geq w_{12} \geq w_{22}$, and at least one of the strict inequalities must be true. In such a case, there are only two equilibria, namely $p = 0$ and $p = 1$. Except for $p = 0$ and $p = 1$, when $\Delta p = 0$, it is always true that $\Delta p > 0$. Hence, if p is close to 0, its value increases (moving it farther away from 0), and so the equilibrium at $p = 0$ is unstable. On the other hand, if p is near 1, it moves still closer to 1 (because $\Delta p > 0$), and so the equilibrium at $p = 1$ is locally stable. In this example, p eventually goes to 1 whatever its initial value, and so the equilibrium at $p = 1$ is globally stable also.

Overdominance

In a diploid organism there is the possibility that the heterozygous genotype has a higher fitness than either homozygote. In this case, there is a polymorphic equilibrium in which the equilibrium value of p is between 0 and 1. We call this situation **overdominance** or *heterozygote superiority*. Symbolically, heterozygote superiority means that $w_{12} > w_{11}$ and simultaneously $w_{12} > w_{22}$. With overdominance, $p = 0$ and $p = 1$ are both equilibria because, according to Equation 5.11, $\Delta p = 0$ at these values. There is also a third equilibrium made possible by the fact that $p(w_{11} - w_{12}) + q(w_{12} - w_{22})$ can equal 0. The equilibrium frequency of A is conventionally denoted $\hat{p}$; hence the equilibrium allele frequency of a is $\hat{q} = 1 - \hat{p}$. The equilibrium can be found by solving $\hat{p}\,(w_{11} - w_{12}) + \hat{q}\,(w_{12} - w_{22}) = 0$, from which a little algebra gives

$$\hat{p} = \frac{w_{12} - w_{22}}{2w_{12} - w_{11} - w_{22}} \tag{5.13}$$

Equation 5.13 is often encountered in another form in which the fitnesses are all expressed relative to that of the heterozygote by setting $w_{11} = 1 - s$, $w_{12} = 1$, and $w_{22} = 1 - t$. With these substitutions, Equation 5.13 becomes

$$\hat{p} = \frac{t}{s+t}$$

This relationship makes intuitive sense because it implies that greater selection against *aa* increases the equilibrium frequency $\hat{p}$ of *A*.

The overdominance equilibrium in Equation 5.13 is globally stable, whereas those at $p = 0$ and $p = 1$ are unstable. The time course is indicated in Figure 5.4A, where the arrowheads show the direction of change in allele frequency. Figure 5.4B shows the change in $\bar{w}$ with overdominance. The average fitness in the population is maximized at the stable equilibrium. Maximization of average fitness is a frequent outcome of selection in random-mating populations with constant fitnesses. There are, however, many exceptions when mating is nonrandom, when the fitnesses are not constant, or when

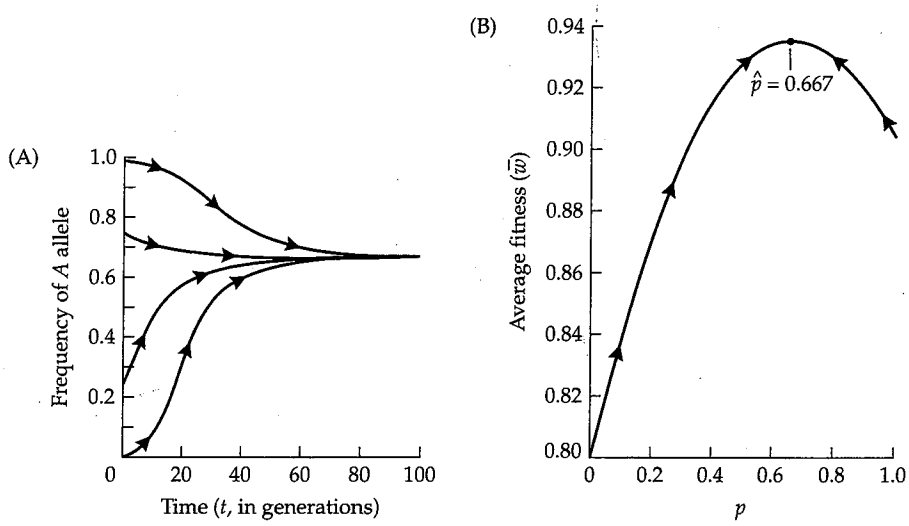


FIGURE 5.4 Selection when there is overdominance. (A) The allele frequencies converge to an equilibrium value irrespective of the initial frequency. In this example, $w_{11} = 0.9$, $w_{12} = 1$, and $w_{22} = 0.8$, and the equilibrium frequency of the *A* allele, $\hat{p}$, is 0.667. (B) Average fitness $\bar{w}$ against p for the same example. Note that $\bar{w}$ is a maximum at equilibrium.

there are interactions between alleles of different genes (Ewens 1979; Curtsinger 1984). Note particularly that $\bar{w}$ is the average fitness *in* the population, not the average fitness *of* the population. The relative survivorships w_{11}, w_{12}, and w_{22} are relevant only to the differential mortality of the genotypes within a population at any given time. The average of the relative survivorships is the average "fitness" $\bar{w}$ *in* the population. However, $\bar{w}$ has no necessary relation to vernacular meanings of "fitness" such as competitive ability, population size, production of biomass, or evolutionary persistence (Haymer and Hartl 1982).

Although overdominance is one mechanism for the maintenance of polymorphisms in natural populations, it has been documented in only a few cases. The classic case is sickle-cell anemia in human beings, which is prevalent in many populations at risk for the type of malaria caused by the mosquito-borne protozoan parasite *Plasmodium falciparum* (Figure 5.5A). The anemia is caused by an allele *S* that codes for a variant form of the β chain of hemoglobin. In persons of genotype *SS*, many red blood cells assume a curved, elongated shape ("sickling") and are removed from circulation. The result is a severe anemia as well as pain and disability because of the accumulation of defective cells in the capillaries, joints, spleen, and other organs. In the absence of intensive medical care, persons of genotype *SS* usually do not survive. The *S* allele is maintained at a relatively high frequency because persons of genotype *AS*, in which *A* is the nonmutant allele, have only a mild form of the anemia but are quite resistant to malaria, perhaps because red blood cells infested with the parasite undergo sickling and are removed from circulation. Homozygous *AA* people are not anemic but, on the other hand, are the most sensitive to severe malaria. The result of the offsetting sickle-cell anemia and malaria resistance is that the heterozygotes have the highest fitness.

In regions of Africa in which malaria is common, the viabilities of *AA*, *AS*, and *SS* genotypes have been estimated as $w_{11} = 0.9$, $w_{12} = 1$, and $w_{22} = 0.2$, respectively (Cavalli-Sforza and Bodmer 1971; Templeton 1982). Substitution into Equation 5.13 leads to a predicted equilibrium allele frequency for *A* of $\hat{p} = 0.89$. Consequently, that of *S* is 0.11. This value is reasonably close to the average allele frequency of 0.09 across West Africa, but there is considerable variation in allele frequency among local populations.

Malaria has been among the most important selective agents acting on the human genome over at least the last 10,000 years. Numerous genes show evidence for selection mediated by malaria. In addition to sickle-cell anemia, another well documented resistance mechanism is associated with deficiency of the red-blood-cell enzyme glucose-6-phosphate dehydrogenase (G6PD), mapped in Figure 5.5B.

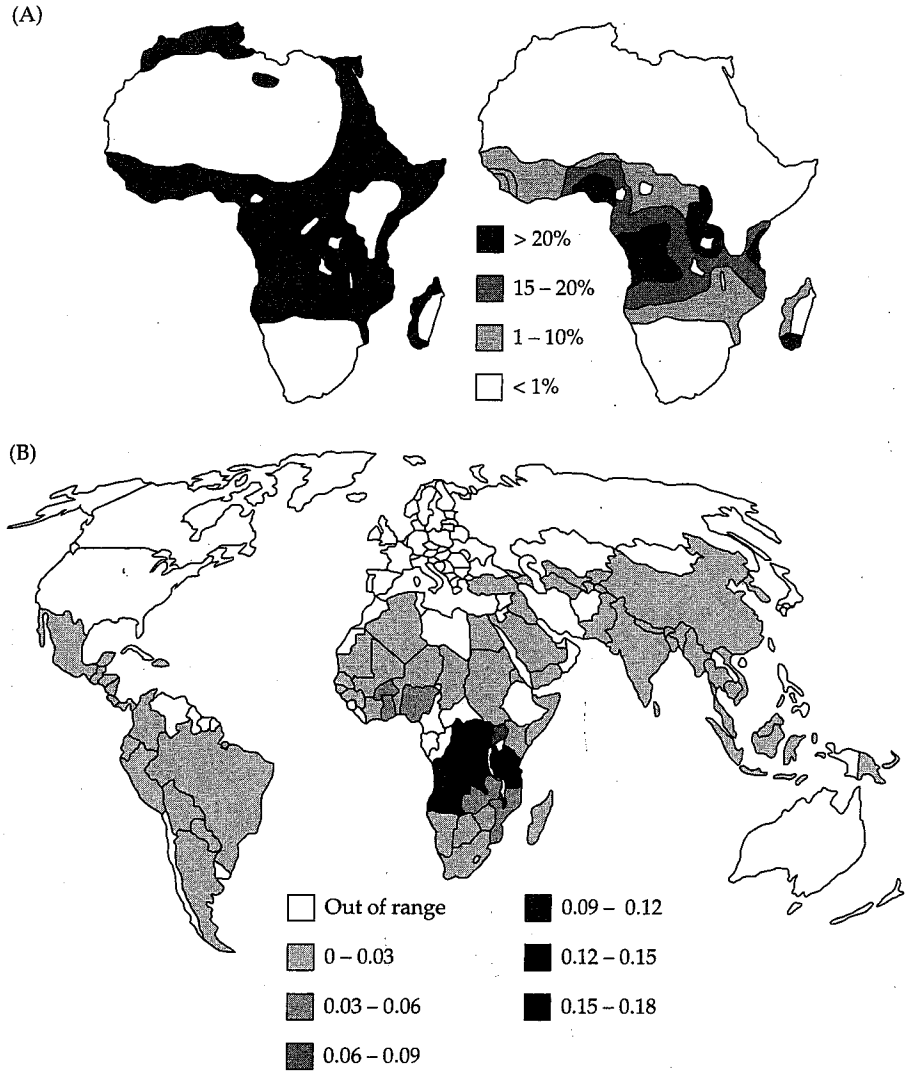


FIGURE 5.5 (A) A map of the distribution of the allele for sickle-cell anemia, and (B) map of the incidence of the *G6PD* deficiency allele. Both mutations are associated with resistance to malaria cause by *Plasmodium falciparum*. (A) The upper left map shows in gray the areas of incidence of falciparum malaria in Africa in the 1920s before mosquito control programs were implemented. The upper right figure shows the distribution of the beta-S globin allele. The global map in (B) shows the frequency of *G6PD* deficiency alleles indicated by the shading. The extensive overlap in the distributions relative to malaria was an early indication that there might be some causal connection.

PROBLEM 5.6 A potentially useful way to estimate the effect on fitness of an environmental insult is to use the cage population approach and estimate fitnesses before and after the insult. In one such study, experimental populations of *Drosophila pseudoobscura* were periodically treated with weak doses of the insecticide DDT. One population was initially polymorphic for five different inversions of the third chromosome. After 13 generations, three of the inversions had essentially disappeared from the population. The two that remained were Standard (*ST*) and Arrowhead (*AR*). Changes in frequency of each inversion were monitored and, from the values for the first nine generations, the relative fitnesses of *ST/ST*, *ST/AR*, and *AR/AR* genotypes were estimated as 0.47, 1.0, and 0.62, respectively (DuMouchel and Anderson 1968). Because the inversions undergo almost no recombination, each type can be considered as an "allele." What equilibrium frequency of *ST* is predicted? What equilibrium value of $\bar{w}$ is predicted?

ANSWER From Equation 5.13, $\hat{p} = (1.0 - 0.62)/(2.0 - 0.47 - 0.62) = 0.42$. (The observed value after 13 generations was 0.43.) The predicted equilibrium value of $\bar{w}$, from Equation 5.8, equals $0.422^2 \times 0.47 + 2 \times 0.42 \times 0.58 \times 1.0 + 0.58^2 \times 0.62 = 0.78$.

PROBLEM 5.7 One of the most widely prescribed drugs for heart patients is the blood anticoagulant warfarin. Its initial use was as a rat poison, and it was initially highly successful, but the effectiveness of the rodenticide gradually diminished because of the evolution of resistance among some target populations. Recently the gene for vitamin K epoxide reductase complex 1 (*VKORC1*) which is the target of warfarin was identified (Rost et al. 2004; Pelz et al. 2005), and there are clear polymorphisms in humans that affect the optimal dose of the drug (Rieder et al. 2005). Among Norway rats in Great Britain, resistance arises due to a mutation in *VKORC1*. If we designate the resistant and sensitive alleles as *R* and *S*, then in the absence of warfarin, the relative fitnesses of *SS*, *SR*, and *RR* genotypes have been estimated as 1.00, 0.77, and 0.46, respectively. In the presence of warfarin, the relative fitnesses have been estimated as 0.68, 1.00, and 0.37, respectively (May 1985). The reduced fitness of the *RR* genotype appears to result from an excessive requirement for vitamin K. Calculate the equilibrium frequency $\hat{q}$ of *R* in the presence of warfarin.

ANSWER From Equation 5.13, the equilibrium frequency $\hat{p}$ of *S* equals $(1.00 - 0.37)/(2 - 0.68 - 0.37) = 0.66$, and so $\hat{q}$ of $R = 0.34$.

Local Stability

Although the curves in Figure 5.4A indicate that the interior equilibrium is locally stable when there is overdominance, an alternative approach can also be applied to the analysis of local stability in models of much greater

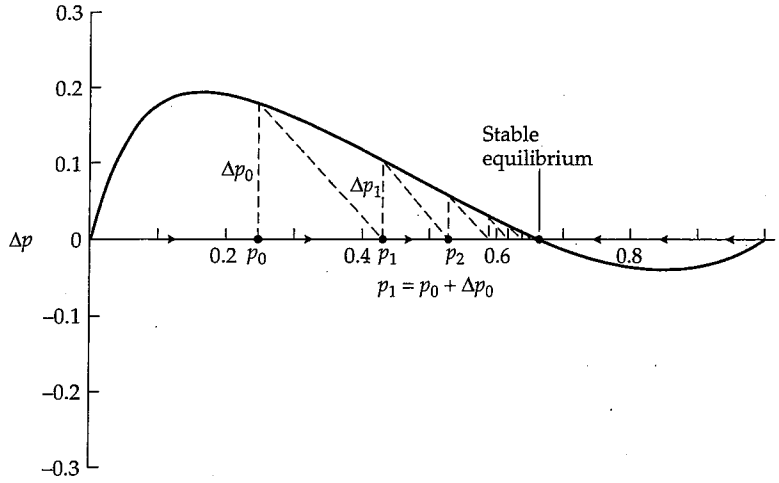


FIGURE 5.6 The change in allele frequency Δp plotted as a function of allele frequency p for a case of overdominance in which $w_{11} = 0.6$, $w_{12} = 1$, and $w_{22} = 0.2$. Starting with an allele frequency p_0 smaller than the equilibrium value, the positive value of Δp_0 indicates that the allele frequency in the next generation, p_1, will be greater than p_0 because $p_1 = p_0 + \Delta p_0$. At an allele frequency of p_1, the value of Δp_1 is also positive, and so p_2 is greater than p_1 because $p_2 = p_1 + \Delta p_1$. The steady increase continues until the population arrives at the equilibrium point $\hat{p}$. The same logic shows that, starting with an initial allele frequency greater than $\hat{p}$, the allele frequency decreases in each succeeding generation and ultimately converges to the equilibrium from the other side.

complexity. It is based on the expression for Δp in Equation 5.11. The local stability of an equilibrium depends on the behavior of Δp for a value of p close to, but not equal to, the equilibrium, as illustrated in Figure 5.6. It is convenient to write $\Delta(p + \varepsilon)$ as the change in allele frequency when the starting point is a small deviation, ε, from any allele frequency p. The function $\Delta(p + \varepsilon)$ can be expanded term by term into an infinite sum:

$$\Delta(p+\varepsilon) = \Delta(p) + \frac{d\Delta(p)}{dp}\varepsilon + \frac{d^2\Delta(p)}{dp^2}\frac{\varepsilon^2}{2!} + \frac{d^3\Delta(p)}{dp^3}\frac{\varepsilon^3}{3!} + \cdots$$

This is known as a *Taylor series expansion*, and it is described in most textbooks on calculus. In evaluating the Taylor series expansion, if ε is sufficiently small, then all terms in ε^2 and higher can be ignored. Therefore, for any value of p, we can approximate $\Delta(p + \varepsilon)$ in terms of Δp itself and its first derivative. Furthermore, if p is one of the equilibrium points, then $\Delta p = 0$ by definition, and so the sign of $\Delta(p + \varepsilon)$ depends on the sign of first derivative of Δp evaluated at the equilibrium in question. By definition, an equilibrium is locally stable if the allele frequency, starting at a point near the equilibrium,

moves ever-closer to the equilibrium. In symbols, this means that $\Delta(p + \varepsilon) < 0$ if $\varepsilon > 0$ and $\Delta(p + \varepsilon) > 0$ if $\varepsilon < 0$. Therefore, any equilibrium point, denoted generically as $\hat{p}$, is locally stable if, and only if,

$$\left.\frac{d\Delta(p)}{dp}\right|_{\hat{p}} < 0$$

where the vertical line and $\hat{p}$ means that the derivative should be evaluated at the equilibrium in question. In practice, calculating the derivative of Δp can be quite tedious without the use of computer software like Maple or Mathematica to do the algebraic manipulations. The result of differentiating Equation 5.11 is that

$$\frac{d\Delta(p)}{dp} = \frac{pqw}{\overline{w}} + \frac{(q-p)(p-\hat{p})w}{\overline{w}} - \frac{2pq(p-\hat{p})^2 w^2}{\overline{w}^2}$$

where $w = w_{11} - 2w_{12} + w_{22}$. With overdominance, $w < 0$. Note that, when $d\Delta p/dp$ is evaluated at $p = 0$ or $p = 1$, both the first and last terms equal 0; when it is evaluated at $p = \hat{p}$, the second and last terms equal 0. The stability analysis proceeds as follows:

$$\text{At } p = 0,\ d\Delta p/dp > 0$$
$$\text{At } p = \hat{p},\ d\Delta p/dp < 0$$
$$\text{At } p = 1,\ d\Delta p/dp > 0$$

Therefore, as is already clear from Figure 5.4A, the equilibrium points at 0, $\hat{p}$, and 1 are unstable, locally stable, and unstable, respectively. This stability analysis is predicated on the assumption of heterozygote superiority, which implies that $w < 0$. Exactly the same equilibrium points are present when there is heterozygote inferiority, but then $w > 0$, which means that the stability property of each equilibrium point is reversed. This situation is discussed next.

Heterozygote Inferiority

Heterozygote inferiority means that the fitness of the heterozygous genotype is smaller than that of both homozygotes: $w_{12} < w_{11}$ and $w_{12} < w_{22}$. An interior equilibrium, given by Equation 5.13, exists in this case also. The analysis in the previous section indicates that this equilibrium is unstable, whereas the equilibria at $p = 0$ and $p = 1$ are both locally (but not globally) stable. An example of heterozygote inferiority is depicted in Figure 5.7A, where the arrows again denote the direction of change in allele frequency. If the initial allele frequency is exactly equal to the equilibrium value (in this example, $\hat{p} = \frac{1}{3}$), then the allele frequency remains at that value. In all other cases, p goes to 1 or 0 depending on whether the initial allele frequency was above or below the equilibrium value.

Figure 5.7B shows the change in average fitness. The unstable equilibrium at $\hat{p} = \frac{1}{3}$ is the minimum average fitness. The shape of the $\overline{w}$ curve has an

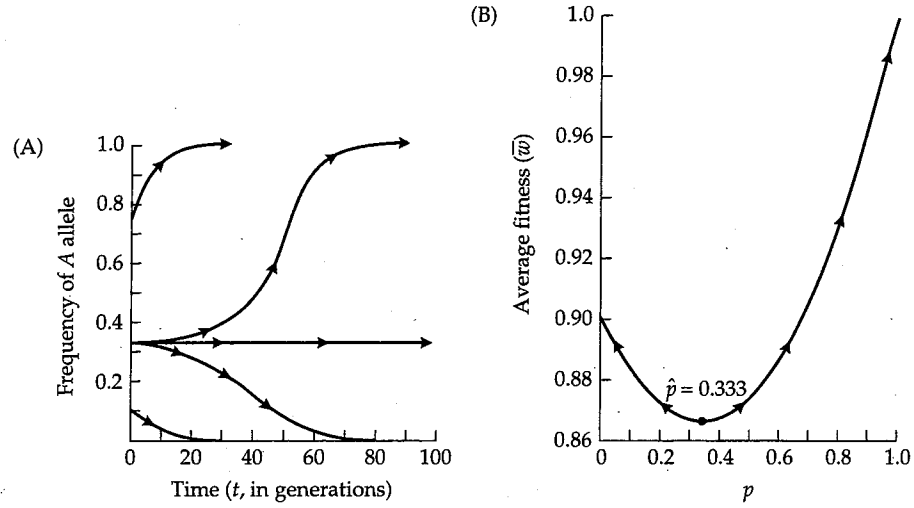


FIGURE 5.7 Selection when there is heterozygote inferiority. (A) The allele frequency goes to 0 or 1 depending on the initial frequency. In this example, $w_{11} = 1$, $w_{12} = 0.8$, and $w_{22} = 0.9$, and there is an unstable equilibrium when the frequency of the A allele is $\hat{p} = \frac{1}{3}$. An infinite population with $p = \frac{1}{3}$ maintains this frequency, but any slight upward change in the frequency of A results in eventual fixation, and any slight downward change in the frequency of A results in ultimate loss. (B) Average fitness $\bar{w}$ against p for the same example. The unstable equilibrium represents the minimum of $\bar{w}$.

important implication that carries over to more complex examples. Imagine a population with an allele frequency near 0, at which $\bar{w} = 0.9$. In terms of average fitness in the population, the population would be better off if the allele frequency were near 1, because then $\bar{w} \approx 1.0$. However, as shown by the direction of the arrows, the population cannot evolve toward $p = 1$. It cannot get through the "valley" because $p = 0$ is a locally stable equilibrium. The population has no way to escape from the equilibrium even though, in doing so, it would eventually end up with a greater average fitness. This consideration would seem to limit the ability of natural selection to increase average fitness in such cases, but one way out of the impass is suggested in the next section.

Stable Equilibria with Multiple Alleles

The presence of multiple alleles complicates the analysis of selection because the number of fitness parameters increases. With n alleles, there are $n(n + 1)/2$ possible genotypes, each with its own fitness. Furthermore, simple generalizations from two-allele theory do not necessarily carry over to multiple alleles. Consider the example of heterozygote superiority. Intuitively, one

might expect that fitnesses yielding stable, multiple-allele polymorphisms would be easy to generate by requiring that each heterozygous genotype has a greater fitness than the homozygous genotypes formed from the constituent alleles. This is not the case, however. If, for n alleles, the fitnesses of the genotypes are assigned at random between 0 and 1, subject to the condition that, for each i and j, $w_{ij} > \max(w_{ii}, w_{jj})$, then only a relatively small proportion of systems with four or more alleles yields a stable polymorphism with all alleles present. For four, five, and six alleles, the percentage of fitness sets yielding a stable equilibrium is 12.6, 1.2, and 0.03, respectively (Lewontin et al. 1978). The reason for the low percentages is that, even if a heterozygote is more fit than its constituent homozygotes, there might be a different homozygote more fit than all three.

How about requiring that each heterozygote be better than *every* homozygote? Surprisingly, this requirement does not help matters much. In this case, for four, five, and six alleles, the percentage of fitness sets yielding a stable equilibrium is 34.3, 10.4, and 1.3, respectively (Lewontin et al. 1978). The point is that polymorphisms with greater than three or four alleles are extremely unlikely to be maintained by selection for simple heterozygous advantage with constant survivorship. If selection is implicated in such a case, models of selection such as diversifying selection or heterogeneous environments are much more plausible. On the other hand, the fitnesses of genotypes in nature are not chosen simultaneously by a random number generator. Each new allele that arises is tested against the resident alleles, and the new allele is able to invade the population if its marginal fitness exceeds the mean fitness of the population. By this process, multiple allele polymorphisms can be accumulated, and the order in which the mutations appear makes a difference (Spencer and Marks 1988).

The possibility of multiple alleles also creates surprising situations in which the outcome of natural selection depends on the order in which the alleles are introduced into the population. Earlier in this chapter we mentioned the sickle-cell hemoglobin polymorphism in Africa and its relation to malaria resistance. People who are homozygous *AA* for the normal allele are susceptible to falciparum malaria, those who are heterozygous *AS* for the sickle-cell allele are resistant to malaria and have a mild anemia, and those who are homozygous *SS* for the sickle-cell allele have a life-threatening anemia. This is a classic case of heterozygote superiority. There is another allele, *C*, found at low frequency in populations in which the *S* allele is prevalent. The *C* allele is also protective against malaria, but the allele is recessive, and so only the *CC* genotypes are resistant. Unlike the *S* allele, the *C* allele does not cause anemia.

The relative survivorship of each of the various hemoglobin genotypes has been estimated based on studies of more than 32,000 people in 72 populations in West Africa (Cavalli-Sforza and Bodmer 1971). The survivorships are given in the following table, which indicates the genotypes that are resist-

ant and those that have severe hemolytic anemia. The survivorships were estimated in a geographical region where malaria was common. Note that the S allele causes a severe anemia in the heterozygous SC genotype, but not so serious as that in the homozygous SS genotype.

Genotype	*AA*	*AS*	*SS*	*AC*	*SC*	*CC*
Survivorship	0.9	1.0	0.2	0.9	0.7	1.3
Health status		Resistant	Anemic		Anemic	Resistant

Inspection of these survivorships reveals a paradox. The CC genotype has the highest fitness, yet the C allele is not fixed. The reason is found in the historical order in which the S and C mutations took place. The A allele is the ancestral type and undoubtedly predated the human settlement of regions subject to malaria. In such a region, the appearance of an S allele creates a heterozygous advantage, and natural selection quickly attains a stable equilibrium at which the ratio of $A : S$ alleles is approximately $8 : 1$. At this equilibrium, the average fitness in the population is $\overline{w} = 0.911$. Now suppose that mutation or migration were to introduce a small number of C alleles. Because C alleles are rare, each is present in either the AC genotype, with probability $\frac{8}{9}$, or in the SC genotype, with probability $\frac{1}{9}$. The average fitness of genotypes heterozygous for C is therefore 0.878, which is smaller than the average fitness in the population. Hence, the frequency of C decreases, and C goes extinct. The C allele has no chance of invading an A/S polymorphism unless the initial frequency of C is sufficiently large. But once C can get established in the population, it eventually becomes fixed.

Adaptive Topography and the Role of Random Genetic Drift

Any graph of $\overline{w}$ against allele frequency is called an **adaptive topography**. The simplest example is Figure 5.7B. In order to generalize the example, try to imagine an adaptive topography in many dimensions with $\overline{w}$ a function of the allele frequencies at many loci. In many dimensions, the adaptive topography is a complex surface upon which there may be "peaks" and "pits" and even "saddle-shaped" regions. The peaks represent locally stable equilibria. Even if natural selection changes the allele frequencies so as to move $\overline{w}$ to the top of some peak, the peak it perches on may not be the highest peak that exists on the whole surface. However, as illustrated in Figure 5.7B, the population may become stuck there because the peak is a locally stable equilibrium.

By what process can a population stranded on a submaximal fitness peak get off the peak? To do so, it has to travel through a nearby valley to a place where natural selection can carry it to the top of an even higher fitness peak. This is something that natural selection acting alone cannot accomplish because it entails a temporary reduction in fitness. There is, however, a

process that can accomplish the task—random genetic drift. In a sufficiently small population, the allele frequencies can change by chance, even producing a reduction in average fitness. Theoretically, random genetic drift can shift a population from a locally stable equilibrium, through a nearby valley, and into a region where it is attracted by another locally stable equilibrium toward a higher fitness peak. Random genetic drift can therefore play a crucial role in evolution by allowing a population to explore the full range of its adaptive topography. This role of random genetic drift has been particularly emphasized by Wright (1977 and earlier) in his proposed shifting balance theory of evolution. Additional discussion of the theory is found in this chapter's section on interdemic selection.

5.4 MUTATION-SELECTION BALANCE

You may recall from Chapter 4 that outcrossing species typically contain a large amount of hidden genetic variability in the form of recessive, or nearly recessive, harmful alleles, each present at a low frequency. Now we can explain why harmful alleles are not completely eliminated. Selection cannot eliminate them because they are continually created anew through recurrent mutation. To be specific, suppose that a is a harmful allele of the wildtype A and that mutation of A to a takes place at the rate μ per generation. Because the allele frequency of a, which we call q, remains small, reverse mutation of a to A can safely be ignored. The calculation of p' carried out to obtain Equation 5.10 is still valid, except that a proportion μ of A alleles mutate to a in each generation. Therefore,

$$p' = \frac{(p^2 w_{11} + pq w_{12})}{\bar{w}}(1-\mu) \tag{5.14}$$

To proceed further, it is convenient to write the relative fitnesses as

$$w_{11} = 1 \qquad w_{12} = 1 - hs \qquad w_{22} = 1 - s$$

The value of s is the selection coefficient against the homozygous aa genotypes and h is the **degree of dominance** of the a allele. If $h = 0$, then a is a complete recessive because AA and Aa have an identical fitness. If $h = 1$, then a is dominant because Aa and aa have an identical fitness. Semidominance means that $h = \frac{1}{2}$, and in this case the allele affects are additive. In mutation-selection balance, we are concerned with harmful alleles that are near the recessive end of the spectrum, and so h will usually be substantially smaller than 0.5.

Equilibrium Allele Frequencies

When selection is balanced by recurrent mutation, there is a globally stable equilibrium at an allele frequency of $\hat{p}$, which is the value of p in Equation 5.14 for which $p' = p$. The equilibrium frequency of the harmful a allele is therefore $\hat{q} = 1 - \hat{p}$. There are two important cases:

- When the harmful allele is a complete recessive ($h = 0$), then

$$\hat{q} \approx \sqrt{\frac{\mu}{s}} \tag{5.15}$$

- When the harmful allele shows partial dominance ($h > 0$), then, to an excellent approximation for realistic values of μ, h, and s,

$$\hat{q} \approx \frac{\mu}{hs} \tag{5.16}$$

Use of these equations is exemplified by Huntington disease in human beings. This severe inherited disorder is characterized by a degeneration of the neuromuscular system that typically appears after age 35. Although the disease itself results from a dominant mutation, the effects on fitness show only partial dominance owing to the late age of onset of the disease. Relative to a value of $w_{11} = 1$ for the homozygous nonmutant genotype, the fitness of the heterozygous genotype has been estimated as $w_{12} = 0.81$ (Reed and Neel 1959). Homozygous mutant genotypes also have the disease, but they are so rare that the equilibrium frequency of the mutant allele is determined by the fitness of the heterozygote. Equation 5.16 with $hs = 0.19$ is appropriate in this example. If we knew either μ or $\hat{q}$, we could estimate the other. In a Michigan population, $q = 5 \times 10^{-5}$ for the Huntington allele (Reed and Neel 1959). Assuming that the population is in equilibrium, we can estimate μ from Equation 5.16 as $\mu = 5 \times 10^{-5} \times 0.19 = 9.5 \times 10^{-6}$. This use of Equation 5.16 illustrates one of the common indirect methods for the estimation of mutation rates in human beings.

The degree of dominance of a harmful allele is a primary factor in determining its equilibrium frequency. Harmful alleles held in mutation-selection balance are rare. Thus the great majority of harmful alleles are present in heterozygous genotypes. Because there are so many heterozygous genotypes, relative to homozygous mutant genotypes, even a small reduction in fitness in the heterozygote has a large effect in decreasing the equilibrium allele frequency. The wavy equals signs in the expressions for the equilibrium under mutation-selection balance actually indicate approximations. These approximations fail near the boundary where $h \approx 0$. In this case we can do a little more algebra to obtain the full solution to the internal equilibrium for one-locus mutation-selection balance:

$$\hat{q} = \frac{\sqrt{(h^2\mu^2 + 2h^2\mu + h^2)s^2 + (4 - 8h)\mu s + (\mu + 1)hs}}{(4h - 2)s}$$

The terms in μ^2 will likely be small enough to ignore, but this expression permits one to consider the region where $h \approx 0$. Figure 5.8 shows the relation between the equilibrium allele frequency and the degree of dominance h for mutations that are homozygous lethal ($s = 1$). Notice how even a small degree of dominance can cause a large reduction in equilibrium frequency. In

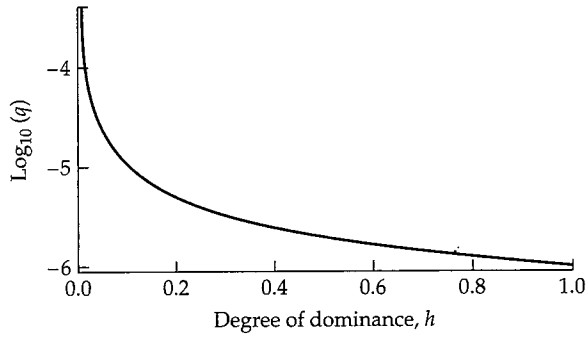


FIGURE 5.8 Allele frequencies maintained at equilibrium by mutation-selection balance as a function of the dominance of the newly arising homozygous-lethal mutations. (From Clark 1998.)

general, for realistic values of μ, s, and h, the value of $\hat{q}$ is typically less than 0.01. Therefore, although mutation-selection balance can account for low-frequency deleterious alleles, it cannot readily account for a harmful allele with a frequency greater than 0.01.

PROBLEM 5.8 To confirm for yourself that a small amount of dominance can have a major effect in reducing the equilibrium frequency of a harmful allele, imagine an allele that is lethal when homozygous ($s = 1$) in a population of *Drosophila*. Suppose that the allele is maintained by mutation-selection balance with $\mu = 5 \times 10^{-6}$. Calculate the equilibrium frequency of the allele for a complete recessive and for partial dominant when $h = 0.025$.

ANSWER For a complete recessive, $\hat{q} = \sqrt{\mu/s} = \sqrt{(5 \times 10^{-6})} = 2.24 \times 10^{-3}$. For partial dominance, $\hat{q} = \mu/hs = (5 \times 10^{-6})/0.025 = 2.00 \times 10^{-4}$. With partial dominance, the equilibrium allele frequency is reduced more than tenfold, and the frequency of homozygous recessive genotypes at equilibrium is reduced more than a hundredfold. It is of interest that $h = 0.025$ is near the average degree of dominance estimated for "recessive" lethals in *Drosophila* (Simmons and Crow 1977).

The Haldane-Muller Principle

The Haldane-Muller principle, named after the geneticists J. B. S. Haldane (1892–1964) and H. J. Muller (1890–1967), deals with the effect of mutation-selection balance on the average fitness of a population. Ignoring recurrent mutation, selection would be able to rid a population completely of a harmful allele, in which case $\hat{q} = 0$ and $\overline{w} = 1$. Because of recurrent mutation, the equilibrium allele frequency is greater than 0. When $h = 0$, the average fitness in the population at equilibrium equals $1 - \hat{q}^2 s = 1 - (\mu/s)s = 1 - \mu$. The reduction in average fitness due to mutation therefore equals $1 - (1 - \mu) = \mu$, which is called the **mutation load**. When a is partially dominant, the muta-

tion load is approximately 2μ because the average fitness at equilibrium is $1 - 2\hat{p}\hat{q}hs - \hat{q}^2 s \approx 1 - 2\mu$. This result is obtained by ignoring terms in $\hat{q}^2$ because they are so small. With or without partial dominance, therefore, the effect of recurrent mutation in reducing the average fitness in the population is independent of how harmful the mutation is. That the effect of recurrent mutation on average population fitness depends only on the mutation rate is the **Haldane-Muller principle**. The implication is that the harmful effect of an increase in the mutation rate is the same irrespective of whether the mutations produced are mildly detrimental or severely harmful. The effects of severe and mild mutations balance out because a more harmful mutation comes to a lower equilibrium frequency.

5.5 MORE COMPLEX TYPES OF SELECTION

Most mutations affect more than one phenotypic attribute of the organism, and so the two-allele model of viability is a simplification that ignores the complications of multiple, or *pleiotropic*, effects of mutations. Such pleiotropic effects may be easy for an experimentalist to miss. For example, a gene affecting embryonic growth rate may also affect age at first reproduction. When the pleiotropic effects act in opposing directions (for example, increasing viability but reducing fertility), the net effect on fitness may be quite small. As a result, mutations with offsetting effects on different components of fitness may remain segregating in a population for many generations.

Additional complications arise because fitness is determined by many genes that interact with each other. Simple models of selection are valid only when the alleles interact in such a way that their effects on fitness are additive or multiplicative across genes. Other complications result when the fitnesses of the genotypes are not constant but variable in time or space. In this section we briefly examine a sample of more complex models. Many of the models are of interest because they can maintain genetic polymorphisms. Although the list is extensive, it is by no means complete. An active area in population genetics research is the determination of means by which natural selection can maintain variation in populations, and this invariably means embracing the complexity of pleiotropic effects on multiple components of fitness imposed by genetic variation.

Differential Selection in the Sexes

Some genes may have different effects in the two sexes. If the fitnesses of genotypes differ between the sexes, then genotypes that are disfavored in one sex may be favored in the other. The offsetting effects increase the opportunity for a balanced polymorphism. The survivorship model of selection can be extended to include this case by supposing that the relative viabilities of the genotypes AA, Aa, and aa are given by w_{11}, w_{12}, and w_{22} in females and by v_{11}, v_{12}, and v_{22} in males. One of the w's and one of the v's

can be set arbitrarily to 1, which leaves four fitness parameters rather than two. A more serious complication is that the allele frequencies in gametes are no longer the same in males and females. If we let p_f and p_m be the allele frequency of A in female and male gametes, respectively, then the genotype frequencies of AA, Aa, and aa in the zygotes are $p_f p_m$, $p_f q_m + q_f p_m$, and $q_f q_m$, respectively, where $q_f = 1 - p_f$ and $q_m = 1 - p_m$. One of the consequences of differential selection in the sexes is that, with an appropriate choice of fitnesses, it is possible to have more than one stable polymorphic equilibrium. A stable equilibrium is also possible with heterozygote inferiority in one sex or with incomplete dominance when selection works in opposite directions in the two sexes.

X-Linked Genes

Genes located in the X chromosome can have the same sort of complications as differential selection in the sexes, but the possibilities for polymorphism are not quite so numerous because there are only three fitness parameters instead of four. If A and a are alleles of an X-linked gene, then there are three genotypes in females (AA, Aa, and aa) and two genotypes in males (either A or a along with the Y chromosome). One fitness parameter in each sex can be set arbitrarily to 1, leaving three free parameters. As with differential selection in the sexes, the allele frequencies differ in eggs and sperm. However, in any generation, the frequency of A in male zygotes equals the frequency of A in female gametes of the preceding generation. If you do not understand why, think about the parental origin of the X chromosome in a male.

Frequency-Dependent Selection

Frequency-dependent selection takes place when fitness is a function of either allele frequencies or genotype frequencies. There is no restriction on the type of frequency dependence except that each Darwinian fitness must be nonnegative. A simple example that illustrates frequency dependence is one in which the fitness of each genotype decreases in proportion to its frequency with a constant of proportionality equal to c:

$$AA : w_{11} = 1 - cp^2 \qquad Aa : w_{12} = 1 - 2cpq \qquad aa : w_{22} = 1 - cq^2$$

In this example, $\Delta p = cpq(q - p)(p^2 - pq + q^2)/\overline{w}$, and so there are equilibria at $p = 0$, $\frac{1}{2}$, and 1. (The factor $p^2 - pq + q^2$ does not have a root for p in the range [0, 1].) A curious feature of this type of frequency-dependent selection is that, at equilibrium, w_{12} is smaller than either w_{11} or w_{22}, so there is heterozygote inferiority; yet $p = \frac{1}{2}$ is a globally stable equilibrium and $\overline{w}$ is a maximum at this equilibrium. The peculiarities of this example are illustrative of frequency-dependent selection in general. Because the fitnesses can be any functions of allele or genotype frequency, nearly anything can happen.

Density-Dependent Selection

Density-dependent selection means that the fitnesses are functions of the population size. Models of density-dependent selection must explicitly include population size and population growth. The classic model of logistic growth says that the population grows initially at an exponential rate, but as the population size approaches the carrying capacity of the environment, K, then the rate slows down as $(K - N)/K$. This produces the differential equation $dN/dt = rN(K - N)/K$. With two haploid genotypes whose numbers at time t are A_t and B_t, the equation for logistic growth becomes:

$$\frac{dA_t}{dt} = r_1 A_t \left(\frac{K_1 - [A_t + B_t]}{K_1} \right) \qquad \frac{dB_t}{dt} = r_2 B_t \left(\frac{K_2 - [A_t + B_t]}{K_2} \right)$$

Each genotype has its own intrinsic rate of increase (r_1 or r_2) and its own carrying capacity (K_1 or K_2), but they affect each other's growth through the total population size ($A_t + B_t$). At any time, the outcome of selection depends on the total population size. When the population size is much smaller then either K_1 or K_2, then the right-hand factor in each growth equation equals approximately 1, and so the selection is determined by the relative values of r_1 and r_2. When the population size becomes approximately equal to the smaller of K_1 or K_2, then the genotype with the smaller carrying capacity stops growing while the other continues, and so the selection is determined by the relative values of K_1 and K_2. Interesting events happen when the selection for r favors one genotype and the selection for K favors the other, especially in situations in which stochastic factors also affect population size or there is a time lag between population size and its affect on growth rate. For further information on these types of models, see Roughgarden (1979), May (1981), Bulmer (1994), and Cohen (1995).

Fecundity Selection

In **fecundity selection**, differences in fitness between the genotypes result from the differing abilities of mating pairs to produce offspring. Because both genotypes in a mating pair contribute to the total number of offspring, the number of fitness parameters potentially equals the number of distinct kinds of mating pairs. For two alleles of one gene, there are nine possible types of mating because reciprocal matings may differ in the expected number of offspring; for example, the expected number of offspring from the mating *Aa* females × *aa* males may differ from that from the mating *Aa* males × *aa* females. The presence of so many fitness parameters complicates the mathematical analysis. An analysis of selection based on individual genotypes, analogous to viability differences, is not possible unless the overall fecundity of any mating pair can be written as either the product or the sum of two parameters, one for each genotype in the mating pair. When this strong sim-

plification does not hold, models of selection with fertility differences become rather complex (Ewens, 1979; Clark and Feldman 1986). Models in which differences in fecundity are combined with differences in survivorship can retain genetic polymorphisms even if there is directional selection in one or the other component of fitness.

Age-Structured Populations

Age-structured populations with overlapping generations present problems even more formidable than those caused by fecundity and survivorship differences in populations with discrete, nonoverlapping generations. In each short interval of time, a new cohort of newborns comes into existence and, as it ages, the fate of each organism in the cohort is governed by the functions l_x, which is the probability of survival from birth to age x, and b_x, which is the probability that an organism of age x (actually in the infinitesimal age interval x to $x + dx$) reproduces. If the functions l_x and b_x maintain the same form over time, then it can be shown that the population eventually reaches a stable age distribution in which the number of organisms in each age group increases or decreases at a constant rate. At the stable age distribution, the overall growth rate of the population is the value of m that satisfies the equation:

$$1 = \int_0^\infty e^{-mx} l_x b_x dx$$

(See Crow and Kimura, 1970, for a derivation.) For this value of m, $dN/dt = mN$, where N is the total population size. In an age-structured population, m corresponds to the intrinsic rate of increase for an exponentially growing population.

So far so good, but genetics complicates this situation enormously. If the l_x and b_x functions differ for different genotypes, then the allele frequencies change through time. As the allele frequencies change, so does the age structure, and the genotype frequencies in each age class may be different. The result is that the age structure may not become stable until selection reaches some equilibrium (possibly fixation). The sorts of complexities that can arise have been examined by Charlesworth (1980).

Heterogeneous Environments and Clines

Heterogeneous environments refer to models in which the relative fitnesses change according to the environment. The environmental heterogeneity may be spatial or temporal or both. Selection of this type can maintain polymorphisms in the absence of overdominance. If each homozygous genotype is favored in a different subset of environments, then there can be **marginal overdominance**, in which the heterozygous genotype has the highest fitness when averaged across all the environments, even though it is not the most fit genotype in any particular environment.

In some cases, the relative fitnesses of the genotypes vary geographically across a more or less smooth environmental gradient, for example, according to latitude, altitude, aridity, or salinity. If sufficiently stable in time, a gradient of selection across a region can result in a gradient of allele frequency across the region. A geographical trend in an allele frequency is called a **cline**. An unusually extreme example of a cline is found in the *hemoglobin–I^1* allele in the eelpout fish *Zoarces viviparus*, the allele frequency of which drops from a value of nearly 1 in the North Sea to a value of nearly 0 in the Baltic Sea (Christiansen and Frydenberg 1974). In human aboriginal populations, there is a cline of increasing frequency of the allele I^B in the ABO blood groups from Southwest to Northeast Europe.

Although clines can result from selection—for example, when one genotype is favored at one extreme of the environmental gradient but disfavored at the other extreme—clines can also result from other processes. Migration is one possibility: Differences in allele frequency in local populations at the extremes of the range may result from chance processes (for example, different founding populations), and migration of organisms from the extremes into the intermediate zone produces the cline.

The strongest evidence that a cline results from selection is when a cline is reproduced in different locations along a similar environmental gradient. A example of parallel clines played out on a grand scale is found in the electrophoretic polymorphism of alcohol dehydrogenase (the *Adh* gene) in *D. melanogaster*. In Eastern North America, the frequency of the Adh^F allele increases as one goes north, whereas DNA polymorphisms flanking *Adh* show no such geographic trend (Berry and Kreitman 1993). The cline is shown in the upper part of Figure 5.9. The frequency of Adh^F is correlated with cooler temperatures and less rainfall in the more northern latitudes. In Australia, as shown in the lower part of Figure 5.9, the frequency of the Adh^F allele increases as one goes south (Oakeshott et al. 1982). This pattern is in apparent contradiction to that in Eastern North America but, because Australia is in the Southern Hemisphere, the clines are actually parallel. Both show an increase in the frequency of Adh^F as one proceeds from the equator toward the polar cap—the North Pole in the Northern Hemisphere and the South Pole in the Southern Hemisphere. On a much smaller geographical scale, in mountainous regions, the frequency of the Adh^F allele shows a clinal increase with altitude, which is again correlated with cooler temperature and less rainfall.

An interesting angle on the physiological basis for this cline was recently suggested (Montooth et al. 2006). It appears that ethanol toxicity is in part mediated by changes in membrane fluidity, and that the sensitivity to the changes in membrane fluidity are highly temperature- and *Adh*-genotype-dependent. Further suggestion that the cline is temperature driven comes from the amazing observation that the entire cline has shifted toward the poles over the last two decades in a manner consistent with the degree of global warming (Umina et al. 2005).

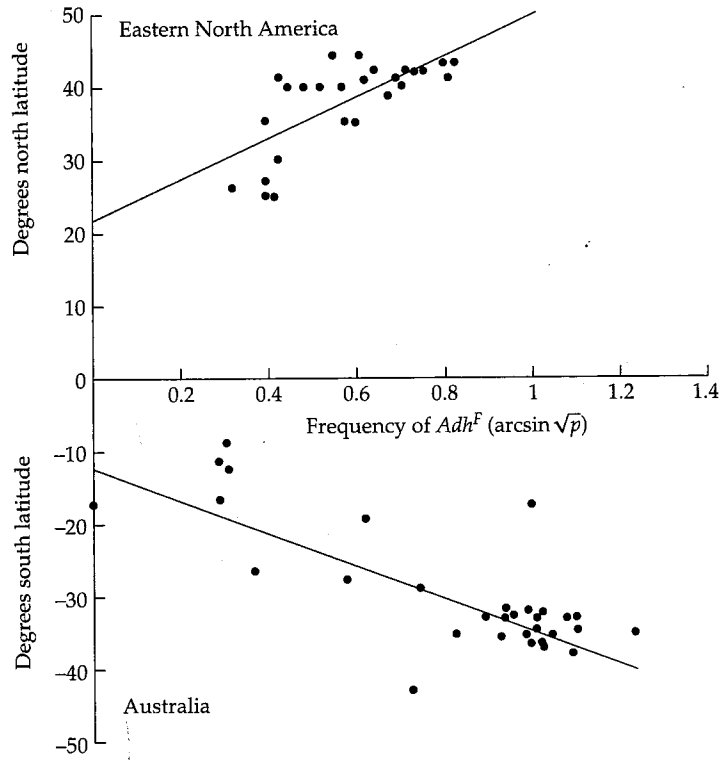


FIGURE 5.9 Parallel clines of the Adh^F (*alcohol dehydrogenase fast*) allele in Eastern North America and in Australia. The allele frequency is given as $\arcsin(\sqrt{p})$, where p is the allele frequency of Adh^F. The angular transformation stretches the scale near the extreme values of p: For values of $p = 0.1$, 0.5, and 0.9, the values of $\arcsin(\sqrt{p})$ are 0.322, 0.785, and 1.249, respectively, where the angles are measured in radians. The angular transformation is often used for proportions because it separates the variance of an estimate from the estimate itself: For a binomial proportion p based on n observations, the variance of p is $p(1-p)/n$, whereas the variance of $\arcsin(\sqrt{p})$, with the angle expressed in radians, is approximately $1/(4n)$. (North American data from Berry and Kreitman 1993; Australian data from Oakeshott et al. 1982.)

Diversifying Selection

The term **diversifying selection** refers narrowly to selection that favors extreme phenotypes. In a normal distribution of phenotypes, for example, diversifying selection means that organisms in the tails of the distribution are favored relative to those in the middle. More generally, diversifying selection refers to any type of selection in which genotypes are favored merely because they are different. Genes under diversifying selection tend to maintain a

relatively large number of alleles. Examples include genes of the major histocompatibility complex in mammals, in which the selective agent is thought to be through resistance to parasitic microorganisms (Satta et al. 1993) and bacterial genes that produce toxins (colicins) that kill other bacteria, in which the selective agent is the destruction of competitors (Riley 1993; Ayala et al., 1994; Wertz and Riley 2004).

Some plants have genes for gametophytic self-incompatibility, in which a pollen grain that carries any self-incompatibility allele is unable to pollinate a plant that carries the same allele. Self-incompatibility of this type implies that no plant can fertilize itself. Because a plant of genotype S_iS_j can produce only S_i and S_j pollen, the pollen cannot fertilize S_iS_j plants. Furthermore, homozygous genotypes are not normally found because their formation would require that S_i pollen fertilize an S_iS_j plant. It is easy to show that there is positive selection for new self-sterility alleles and that, at equilibrium, every allele has the same frequency. For n alleles, if S_i has frequency p_i, then the frequency of S_iS_j genotypes with random mating is $2p_i(1 - p_i)/(1 - \Sigma p_i^2)$. The denominator is necessary because of the absence of homozygous genotypes. The probability that an S_i pollen can be successful in fertilization is therefore the probability of genotypes other than S_iS_j, which equals $1 - 2p_i(1 - p_i)/(1 - \Sigma p_i^2)$. At equilibrium, we must have $p_i(1 - p_i) = p_j(1 - p_j)$. From these expressions follow some important conclusions summarized in Problem 5.9. For more information on gametophytic self-incompatibility systems, see Ioerger et al. (1991) and Uyenoyama (1995). The loss of self-incompatibility is an especially interesting problem, and analysis of the *SI* loci in *Arabidopsis thaliana* (which mostly self-pollinates) and its outcrossing relatives has identified three major haplotypes of *SRK*, suggesting either more than one loss of selfing or ancient recombination (Bechsgaard et al. 2006).

PROBLEM 5.9 Show that $p_i(1 - p_i) = p_j(1 - p_j)$ for all i and j implies that $p_i = p_j = 1/n$, where n is the number of self-incompatible alleles and $n \geq 3$. Use these equilibrium allele frequencies to show that the probability that a pollen grain lands on a compatible style equals $(n - 2)/n$. Finally, show that the probability of successful fertilization by a new mutant S allele, relative to that of any preexisting allele, equals $n/(n - 2)$.

ANSWER $p_i(1 - p_i) = p_j(1 - p_j)$ implies that $p_i - p_j = p_i^2 - p_j^2 = (p_i - p_j)(p_i + p_j)$ so that either $p_i = p_j$ for all i and j or $p_i + p_j = 0$. Because $n \geq 3$, $(p_i + p_j) \neq 1$. Because there are n alleles, we must have $\Sigma p_i = 1$, and so $p_i = 1/n$. The probability of a pollen grain landing on a compatible style is $1 - 2p_i(1 - p_i)/(1 - \Sigma p_i^2) = 1 - 2/n = (n - 2)/n$. A pollen grain containing a newly arising S allele will always land on a compatible style, and so its probability of fertilization, relative to that of a preexisting allele, equals $1/[(n - 2)/n] = n/(n - 2)$. In effect, this is the relative fitness of a new mutation. For $n = 3$, 4, 5, 10, 50, and 100, it equals 3, 2, 1.67, 1.25, 1.04, and 1.02, respectively.

Gametic Selection

Many plants go through a life cycle in which both haploid products of meiosis and the diploid products of fertilization are exposed to selection. In mosses and vascular plants, for example, a diploid organism (the *sporophyte*) produces spores, each of which germinates to form a haploid organism (the *gametophyte*) that reproduces asexually by mitosis. The gametophytes give rise to haploid male and female gametes, which undergo fertilization creating a new diploid generation. In mosses, the prominent stage of the life cycle is the gametophyte, whereas in higher plants, the prominent stage is the sporophyte.

When the haploid phase of the life cycle is exposed to selection, the selection is called **gametic selection**. As a concrete model, suppose that the relative survivorships of A and a gametophytes (the haploid phase) are given by v_1 and v_2, respectively. In the sporophytes (the diploid phase), the survivorships can be written as before as w_{11}, w_{12}, and w_{22}. If p and q are the allele frequencies of A and a at the beginning of the haploid phase, then after the differential haploid mortality has taken place, the frequencies will be $p^* = pv_1/\bar{v}$ and $q^* = qv_2/\bar{v}$, where $\bar{v} = pv_1 + qv_2$. With random fertilization among the gametes, the diploid genotypes AA, Aa, and aa are formed in the proportions p^{*2}, $2p^*q^*$, and q^{*2}, and these survive in the relative proportions w_{11}, w_{12}, and w_{22}. You may verify for yourself that, at the beginning of the haploid phase of the next generation, the allele frequency of A is

$$p' = \frac{p^2 w_{11} v_1^2 + pq w_{12} v_1 v_2}{p^2 w_{11} v_1^2 + 2pq w_{12} v_1 v_2 + q^2 w_{22} v_2^2}$$

This equation has the same form as the equation for p' in Equation 5.10 except that w_{11} is replaced with $w_{11}v_1^2$, w_{12} with $w_{12}v_1v_2$, and w_{22} with $w_{22}v_2^2$. The conditions for fixation or for a stable or unstable equilibrium are therefore determined by the relative magnitude of the composite "fitness" of the heterozygous genotype relative to those of the homozygous genotypes.

Meiotic Drive

A situation analogous to, but distinct from, gametic selection takes place when there is non-Mendelian segregation in the heterozygous genotype. In females, unequal recovery of reciprocal products of meiosis can be caused by nonrandom segregation of homologous chromosomes to the functional egg nucleus, which is why non-Mendelian segregation is known generically as **meiotic drive**. In other cases, the unequal recovery is caused by a gene or genes that act to render gametes carrying the homologous chromosome nonfunctional. Examples include "sperm killers" such as *segregation distortion* in *Drosophila melanogaster* (Charlesworth and Hartl 1978) and the *t* alleles in the house mouse (Lewontin and Dunn 1960; Hammer and Silver 1993) as well as "spore killers" described in filamentous fungi (Raju 1994).

Because meiotic drive acts only in the heterozygous genotype, its effect is to alter the term pqw_{12} in Equation 5.10 for p'. This term comes from the expression $\frac{1}{2} \times 2pqw_{12}$ for the proportion of A-bearing gametes from surviving Aa genotypes, and the $\frac{1}{2}$ is the Mendelian segregation ratio. If the ratio of $A : a$ gametes from Aa heterozygotes is $k : 1 - k$ instead of $\frac{1}{2} : \frac{1}{2}$, then the expression for p' becomes

$$p' = \frac{p^2 w_{11} + 2kpqw_{12}}{\overline{w}} \tag{5.17}$$

where $\overline{w}$ is the average survivorship in the population defined in Equation 5.8. Since A is the driven allele, $k > \frac{1}{2}$. Equation 5.17 is illustrative of meiotic drive even though it requires that the non-Mendelian segregation affect both sexes equally, a case that is not generally found in practice. One implication of the equation is that, unless selection counteracts the meiotic drive, the driven allele goes to fixation (Figure 5.10). In particular, if the relative viabilities are equal, then $p' = p^2 + 2kpq$ and $\Delta p = pq(2k - 1)$, so that $p \to 1$ because $k > \frac{1}{2}$. In some examples of meiotic drive, including *segregation distortion* and the t alleles, the driven allele is lethal when homozygous (Hartl 1970). Assuming that the lethality is completely recessive, the survivorships are $w_{11} = 0$, $w_{12} = 1$, and $w_{22} = 1$. Equation 5.17 implies that $p' = 2kp/(1 + p)$ and so $\Delta p = p[(2k - 1) - p]/(1 + p)$. There is an interior equilibrium at $\hat{p} = 2k - 1$, which intuition suggests (correctly) is locally stable. It is also globally stable (see Figure 5.10). Note that $\hat{p}$ is between 0 and 1 for any value of k between $\frac{1}{2}$ and 1. The calculations for a recessive-lethal driven allele are a special case of the slightly more general model discussed in Problem 5.10.

PROBLEM 5.10 Suppose that the AA genotype has a survivorship given by $1 - s$ relative to a value of 1 for both Aa and aa genotypes. Use Equation 5.17 to show that $\Delta p = pq[(2k - 1) - ps]/(1 - p^2 s)$. Find $\hat{p}$ and define the conditions, in terms of k and s, for which $\hat{p}$ is between 0 and 1. Show also that the equilibrium is locally stable.

ANSWER Equation 5.17 implies that $p' = [p^2(1 - s) + 2kpq]/(1 - p^2 s)$. $\Delta p = p' - p$ simplifies to the formula given. Setting $\Delta p = 0$ yields equilibria at 0, 1, and $\hat{p} = (2k - 1)/s$. For $\hat{p} > 0$, we need $(2k - 1)/s > 0$, or $k > \frac{1}{2}$. For $\hat{p} < 1$, we need $(2k - 1)/s < 1$, or $k < (s + 1)/2$. Note that, as the selection against the A allele becomes smaller (s closer to 0), more values of k result in fixation of the unfavorable A allele and fewer result in an interior equilibrium. The stability of $\hat{p}$ can be deduced by evaluating the derivative of Δp. For this purpose, it is convenient to write Δp as $pqs(\hat{p} - p)/(1 - p^2 s)$. In taking the derivative, remember that any term containing $\hat{p} - p$ becomes 0 when $p = \hat{p}$, so these terms can be neglected. The derivative, evaluated at $\hat{p}$, equals $-\hat{p}\hat{q}s/(1 - \hat{p}^2 s)$, where $\hat{q} = 1 - \hat{p}$. The sign of this number must be negative, and so the equilibrium at $\hat{p}$, when it exists, is locally stable.

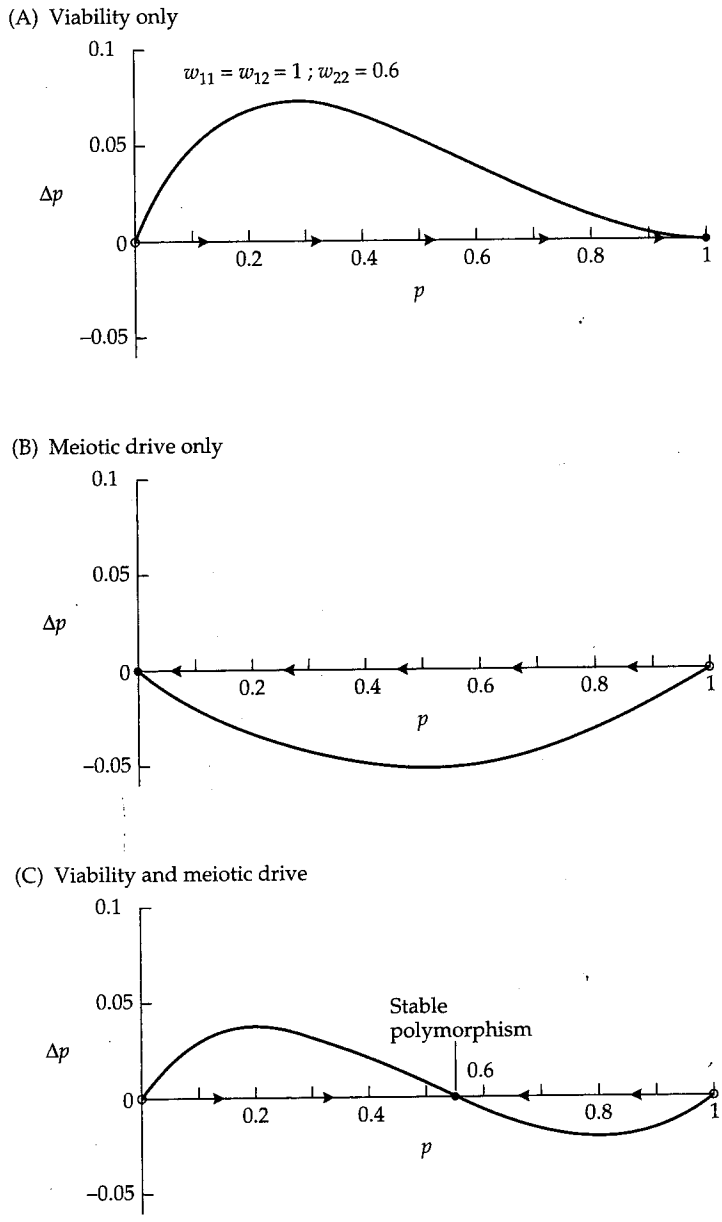


FIGURE 5.10 The balance between meiotic drive and viability selection. (A) Δp versus p for viability alone, when the fitnesses are $w_{11} = w_{12} = 1$ and $w_{22} = 0.6$. With these fitnesses, viability selection would eliminate the a allele. (B) Meiotic drive alone, where the heterozygous genotype Aa produces 40% A-bearing gametes and 60% a-bearing gametes. With meiotic drive alone, the A allele would be lost. (C) Δp versus p when both viability selection and meiotic drive are operating at the same time, using the same fitness and meiotic drive parameters as above. In this example, when both processes operate simultaneously, their offsetting effects create a stable polymorphism.

Multiple Loci and Gene Interaction: Epistasis

With multiple loci, as many types of gametes are possible as there are combinations of alleles. The simplest example is the two-locus, two-allele case, in which the possible gametes are AB, Ab, aB, and ab. In the absence of recombination ($r = 0$), each type of gamete can be regarded as an "allele" of one locus with four alleles. The principles of multiple-allele selection then apply, and some of the "alleles" may be eliminated by selection. The presence of recombination complicates matters because each gametic type is continually recreated by recombination even if it is disfavored by selection. The influence of recombination on the outcome of selection is determined by the recombination fraction and by the degree of interaction between the loci. When selection acts on the phenotype produced by the joint effects of multiple loci, there are two general situations:

- Changes in allele frequency are driven primarily by the selection coefficients and recombination plays a minor role.
- Selection and recombination are about equally important in determining the outcome.

The former is usually the case with weak epistasis and moderate or loose linkage; the latter is more prevalent with strong epistasis and tight linkage. The term **epistasis** is often used in population genetics as a synonym for gene interaction; it applies to any situation in which the genetic effects of different loci that contribute to a phenotypic trait are not additive. How much epistasis is there for fitness in real populations? Opinions vary widely on this point, but it seems that whenever a study is designed to be able to detect epistasis, it is not hard to find abundant examples of gene-gene interactions playing a role in fitness (Figure 5.11).

In the two-locus, two-allele example, the fitnesses (survivorships) of the genotypes can be written as shown in Table 5.3, where it is assumed that the

TABLE 5.3 Two-Locus Fitnesses

Genotype at *A* locus	Genotype at *B* locus: *BB*	*Bb*	*bb*	
AA	w_{11}	w_{12}	w_{13}	w_{AA}
Aa	w_{21}	$w_{22} = 1$	w_{23}	w_{Aa}
aa	w_{31}	w_{32}	w_{33}	w_{aa}
	w_{BB}	w_{Bb}	w_{bb}	

Note: The table assumes that the two types of double heterozygotes, AB/ab and Ab/aB, have the same fitness, w_{22}.

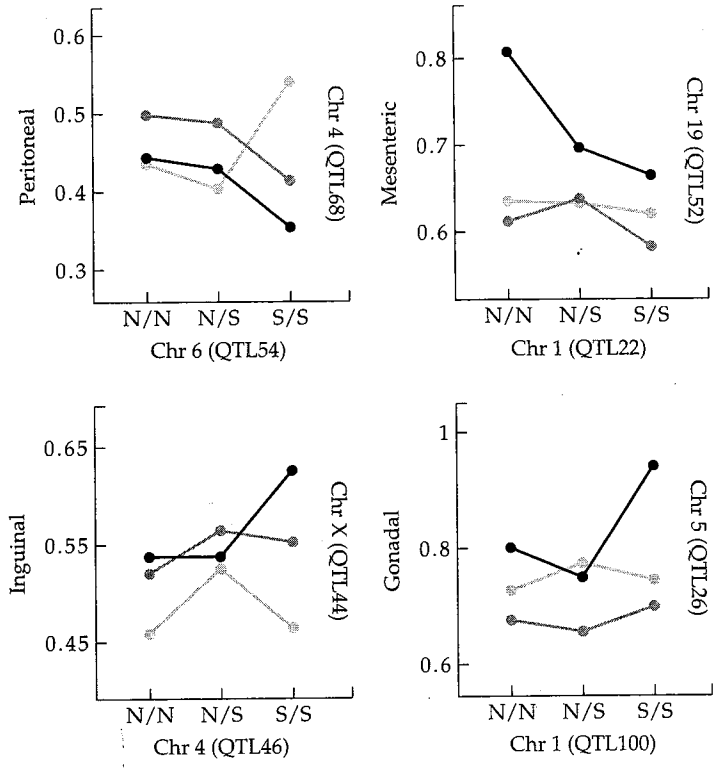


FIGURE 5.11 The mean phenotypes for the 9 genotypes as depicted in Table 5.3 for a series of mouse SNP-pairs. In this study, tests of epistatic interactions were done by measuring fat pad weights in mice from 513 F_2 progeny generated from the SM/J × NZB/BINJ inbred mouse strain cross. The fact that the line segments are not parallel is an indication of the nonadditivity (epistasis) of effects of each given SNP pair. Each QTL is designated according to its relative position along the chromosome. (From Stylianou et al. 2006.)

two types of double heterozygote (AB/ab and Ab/aB) have the same fitness; for convenience, this value is often set at $w_{22} = 1$. For each single-locus genotype, the average survivorship is equal to the weighted average across each genotype at the other locus. In Table 5.3, these averages are denoted w_{AA}, w_{Aa}, and so on. Additivity across loci means that $w_{11} = w_{AA} + w_{BB}$, $w_{12} = w_{AA} + w_{Bb}$, and so forth for all genotypes, including $w_{22} = w_{Aa} + w_{Bb} = 1$. If additivity does not apply across all nine genotypes, then epistasis is said to be present.

When there is strong epistasis and tight linkage, complications abound. With two loci and two alleles at each, there are as many as 15 equilibria. Most of them are unstable, but examples are known in which four interior equilibria are simultaneously stable. The average fitness in the population is not necessarily a maximum at equilibrium, and there are cases in which none of the four stable equilibria is a point of maximum average fitness. In addition, not only is average fitness not necessarily a maximum at equilibrium, natural selection can cause a decrease in average fitness. Despite all this odd behavior, extensive computer simulation and approximate solutions (Ewens 1979) show that if epistasis is not too strong, and linkage is not too tight, then the average fitness in the population usually increases.

Evolution of Recombination Rate

With constant fitnesses, the population can only maintain particularly favorable combinations of alleles when the rate of recombination is low. So long as it is possible to arrive at these favorable combinations, theory has shown that alleles at a modifier locus that controls the rate of recombination will usually be favored only if they reduce the recombination rate. Why then is recombination so prevalent? Many arguments have been raised, but the essential idea is that recombination can increase the efficacy of natural selection. The **Hill-Robertson effect** arises when favorable combinations of alleles are created more rapidly with increasing rates of recombination (so that recombination accelerates fixation of favorable alleles) and conversely, deleterious mutations arising on an otherwise favorable background have their elimination from the population retarded until recombination breaks them away from those favorable background alleles. This recombination can also accelerate the rate of elimination of deleterious alleles. The situation is more complicated when there is epistasis.

Regardless of the form of epistasis, recombination is advantageous when it allows for a more rapid generation of favorable recombinant gametes, as can happen when different individuals in a population carry different advantageous alleles. Recombination accelerates the process of adaptation by allowing these favorable alleles to recombine onto one gamete. But models for recombination modifiers show that genes that foster sex and recombination invade the population only if the rate of adaptive mutation is high. Recently Keightley and Otto (2006) showed that background selection against deleterious mutant alleles provides an advantage to sex and recombination that increases with population size. With low levels of recombination, selection at other loci severely reduces the effective population size and decreases genetic variance in fitness. Natural selection becomes less effective at ridding the population of deleterious alleles when the population size is small and recombination is rare (due to the Hill–Robertson effect). Recombination results in an increase in genetic variance in finite populations and this directly improves the response to selection. A substantial advantage occurs for sex and recombination that is surprisingly insensitive to the form of epistatic interactions between deleterious alleles.

Sexual Selection

It seems that, wherever one looks in nature, animals have physical adornments or behavioral displays to help them in obtaining mates. In some cases, there is direct competition between animals, usually males, as exemplified by the contests of antler clashing in moose or head butting in bighorn sheep. In other cases, there is indirect competition, as seen in the behavioral displays of male peacocks in full plumage strutting their stuff. These are dangerous activities. A bighorn sheep can get his skull fractured or fall off a cliff. The

male peacock is conspicuous, burdened, and preoccupied—vulnerable to any predator.

Darwin (1871) was the first to draw attention to competition for mates as a source of selection not necessarily related to adaptation of the organism to its environment. This type of selection he called **sexual selection**. In the case of direct competition for mates, it is easy to understand that a successful male leaves more progeny than an unsuccessful male, and so alleles promoting the physical adornments, strength, and aggressiveness needed for successful competition for mates are perpetuated even though they may occasionally be detrimental. The example of indirect competition is considerably more subtle because the male is merely advertising. The female does the choosing. One theory for the evolution of male sexual displays is that, in the early stages of their evolution, the displays take advantage of a female preference. The origin of the initial preference is unclear. Darwin suggested that female choosiness and offspring number are both associated with superior nutrition; hence choosy females may, at the beginning, have had more offspring. Whatever the cause, given an initial choosiness among females, males with more effective displays are chosen preferentially as mates, and their offspring receive alleles that create both the displays in the males and the preferences in the females. If these traits are genetically correlated—as, for example, through common hormonal or neurological pathways or through linkage disequilibrium—then selection becomes a self-accelerating process promoting increasingly elaborate displays and increasingly greater choosiness. According to Fisher (1930):

> The two characteristics affected by such a process, namely plumage development in the male, and sexual preference for such developments in the female, must thus advance together, and so long as the process is unchecked by severe counterselection, will advance with ever-increasing speed. In the total absence of such checks, it is easy to see that the speed of development will be proportional to the development already attained. There is thus, in any situation in which sexual selection is capable of conferring a great reproductive advantage, the potentiality of a runaway process which will, however small the beginnings from which it arose, must, unless checked, produce great effects, and in the later stages with great rapidity.

The ever-accelerating process is called **runaway sexual selection**, and the conditions under which it takes place have been studied theoretically (Lande and Arnold 1985; Kirkpatrick and Barton 1995; Iwasa and Pomiankowski 1995).

Because males can produce an excess of sperm and matings typically do not involve nearly as much investment in energy as for a female producing progeny, there arises a tension between the roles of selection optimizing male reproductive fitness and those optimizing female reproductive fitness. **Sexual conflict** is the name given to this topic; it is an area of active research (Chapman 2006). By rearing flies under conditions where they are forced to

be monogamous, there is relaxation of selection on male-male competitive ability, and the males become poorer at courting and at sperm competition (Rice and Holland 2005). Other experimental designs favor increased response to male-male competition by allowing males to evolve in the laboratory without their genes facing selection in females (Rice et al. 2006). The simplest experimental design for this purpose is to allow for selection to occur in a male lineage, such that recombination with female-passed genes does not occur, and females are drawn from outside the selected line in each generation. When this is done, the males become more aggressive, they outcompete nonselected males for matings, and their matings do more damage to females. In addition, when the alleles from such selected males are present in females, they have a negative effect on female fitness.

Kin Selection

One alternative type of selection, called **kin selection**, makes use of an extended concept of "fitness." In kin selection, a positive selection for certain alleles takes place indirectly through enhanced reproduction of the genetic relatives of carriers of the alleles rather than directly through an increased fitness of the carriers themselves. Kin selection has been postulated in attempts to account for the evolution of altruism. A behavior is regarded as altruism if it increases the fitness of other organisms at the expense of one's own fitness. Altruistic behavior is exhibited most dramatically by social insects such as termites, ants, and bees, in which certain worker castes exert their labors for the care, protection, and reproduction of the queen and her offspring but do not reproduce themselves.

A central consideration in kin selection is that relatives have genes in common. Therefore, a gene that causes altruistic behavior can increase in frequency if the increase in the recipient's fitness as a result of altruism is sufficiently large to offset the decrease in the altruist's own fitness. The essentials of the situation can be made clear by considering the case of identical twins. Because identical twins are genetically identical, the reproduction of one's twin is genetically equivalent to reproduction by oneself. Thus, it makes no difference if an altruistic organism decreases its own fitness for the sake of an equal increase in fitness of an identical twin; from an evolutionary point of view, it is an even trade because the combined number of offspring from both twins remains unchanged. By the same token, if an altruistic act decreases the fitness of an organism by an amount less than the increase gained by an identical twin, then the altruism results in a net increase in the combined number of offspring. One would, therefore, expect altruism between identical twins to be favored by natural selection as long as the risk to the altruist is no greater than the benefit to the recipient.

These considerations of identical twins can be extended to other degrees of relationship as well, but the risk to the altruist must be correspondingly smaller than the benefit to the recipient because other types of relatives share

fewer genes than identical twins. The break-even points for altruism toward various degrees of relationship have been wittily summarized by J. B. S. Haldane, who is said to have quipped that he would lay down his life for two brothers, four nephews, or eight cousins. In any case, fitness considerations that take into account not only an organism's own fitness but also the fitness of relatives (other than direct descendants) constitute what is called the **inclusive fitness** of the organism.

To be concrete, suppose that altruism results in a decrease in fitness c of the altruist that is offset by an increase in fitness b in the recipient. The gene for altruism increases in frequency if the ratio of cost to benefit is great enough, relative to the genetic relationship between the altruist and the recipient; that is, the gene for altruism increases in frequency if

$$\frac{c}{b} < r \tag{5.18}$$

as shown first by Hamilton (1964) and discussed in detail by Cavalli-Sforza and Feldman (1978) and Uyenoyama and Feldman (1980). In this context, r is a measure of genetic relationship between the altruist X and the recipient of the altruism Y, defined as

$$r = \frac{2F_{XY}}{(1+F_X)} \tag{5.19}$$

where F_X is the inbreeding coefficient of the altruist X, and F_{XY} is the inbreeding coefficient of a hypothetical offspring of X and Y. As illustrated in Figure 5.12, r equals the probability that two gametes from X and Y contain alleles that are identical by descent, F_{XY}, relative to the probability that two gametes from X contain alleles that are identical by descent, $(1 + F_X)/2$. The cost-

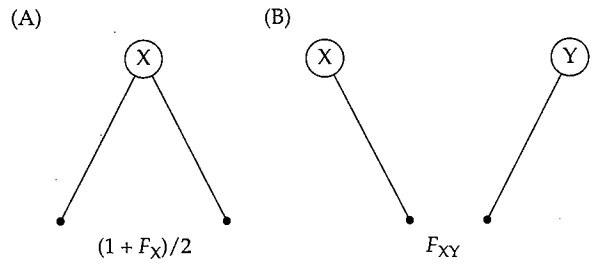


FIGURE 5.12 Definition of the genetic relationship between an altruist X and the recipient of the altruism Y. (A) Two alleles chosen at random from an organism X are identical by descent with probability $(1 + F_X)/2$. (B) Two alleles chosen at random, one from X and the other from Y, are identical by descent with probability F_{XY}, which is the inbreeding coefficient of a hypothetical offspring of X and Y. The ratio of F_{XY} to $(1 + F_X)/2$ is the appropriate measure of genetic relationship in the consideration of kin selection.

benefit tradeoff in Equation 5.19 is generally valid for weak selection when $F_X = 0$ and valid for additive alleles even when $F_X \neq 0$ (Aoki 1981).

5.6 INTERDEME SELECTION IN GEOGRAPHICALLY SUBDIVIDED POPULATIONS

When a population is composed of a set of semi-isolated subpopulations of the same species, then it is possible that the environments may differ across these subpopulations, giving rise to different pressures of natural selection upon allele frequencies. If subpopulations, or demes, composed of certain genotypes are more likely to become extinct and have their vacated habitats recolonized by migrants from other subpopulations composed of other genotypes, then the more successful subpopulations can, in some sense, be considered as having a greater "fitness" than the less successful ones. Since this concept of population fitness is a characteristic of the entire population and not merely the average fitness of the genotypes within it ($\overline{w}$), interdeme selection is outside the realm of most conventional models of selection.

Interdeme selection plays an essential role in the **shifting balance theory** of evolution of Wright (1977 and earlier). In the shifting balance theory, a large population that is subdivided into a set of small, semi-isolated subpopulations (demes) has the best chance for the subpopulations to explore the full range of the adaptive topography and to find the highest fitness peak on a convoluted adaptive surface. If the subpopulations are sufficiently small, and the migration rate between them is sufficiently small, then the subpopulations are susceptible to random genetic drift, which allows them to explore their adaptive topography more or less independently. In any subpopulation, random genetic drift can result in a temporary reduction in fitness that would be prevented by selection in a larger population, and so a subpopulation can pass through a "valley" of reduced fitness and possibly end up "climbing" a peak of fitness higher than the original. Any lucky subpopulation that reaches a higher adaptive peak on the fitness surface increases in size and sends out more migrants to nearby subpopulations, and the favorable gene combinations are gradually spread throughout the entire set of subpopulations by means of interdeme selection.

The shifting balance process includes three distinct phases:

1. An exploratory phase, in which random genetic drift plays an important role in allowing small subpopulations to explore their adaptive topography.
2. A phase of mass selection, in which favorable gene combinations created by chance in the random drift phase become rapidly incorporated into the genome of local subpopulations by the action of natural selection.
3. A phase of interdeme selection, in which the more successful demes increase in size and rate of migration; the excess migration shifts the

> allele frequencies of nearby subpopulations until they also come under the control of the higher fitness peak. The favorable genotypes thereby become spread throughout the entire population in an ever-widening distribution. Where the region of spread from two such centers overlaps, a new and still more favorable genotype may be formed and itself become a center for interdeme selection. In this manner, the whole of the adaptive topography can be explored, and there is a continual shifting of control from one adaptive peak to control by a superior one.

The shifting balance theory has played an important role in evolutionary thinking, in part because of its use of mountain-climbing terms as tropes for stages in the evolutionary progress: "exploration" of the adaptive topography, chance "discovery" of a route to a higher adaptive peak, and ultimately the "conquest" of the highest adaptive peak by the whole species. However, as a comprehensive theory of evolution, many aspects of the theory remain untested. For the theory to work as envisaged, the interactions between alleles must often result in complex adaptive topographies with many peaks and valleys. The population must be split up into smaller subpopulations, which must be small enough for random genetic drift to be important, but large enough for mass selection to fix favorable combinations of alleles. Although migration between demes is essential, neighboring demes must be sufficiently isolated for genetic differentiation to take place, but sufficiently connected for favorable gene combinations to spread. Because of uncertainly about the applicability of these assumptions, the shifting balance process remains a picturesque metaphor that is still largely untested. However, computer simulations have been carried out to investigate the range of magnitudes of the key parameters that are necessary for the shifting balance process to be effective; these parameters include the size of the subpopulations, the rate of migration and range of dispersal of the migrants, the degree of epistasis between genes, and the rate of recombination (Bergman et al. 1995). Some empirical studies have also explored the partitioning of genetic variance within and between groups for traits associated with fitness (Wade and Goodnight 1991).

One important implication of interdeme selection is that alleles that are harmful in themselves may nevertheless be favored because they are beneficial to the group. This principle is illustrated in the model in Table 5.4, where the allele A' is harmful to organisms within demes but favorable to the deme as a whole. Equation 5.11 implies that, within the ith deme, $\Delta q_i = -cq_i(1 - q_i)$ (assuming that $\overline{w} \approx 1$). Averaging across all of the subpopulations, the change in allele frequency resulting from selection within subpopulations, Δq_w, equals $-c\overline{q}(1 - \overline{q})(1 - F)$, where F is the fixation index F_{ST} discussed in Chapter 3. At the same time within-subpopulation selection takes place, interdeme selection favors demes containing A', and the change in allele frequency resulting from between-subpopulation selection, Δq_b, equals $2(b - c)\overline{q}(1 -$

TABLE 5.4 Model of Interdeme Selection

Genotype	AA	AA'	$A'A'$
Frequency in deme i	p_i^2	$2p_iq_i$	q_i^2
Within-population fitness	1	$1-c$	$1-2c$
Between-population fitness of deme i		$1+2(b-c)q_i$	

$\bar{q})F$, as shown by Crow and Aoki (1982). Putting the within-subpopulation and between-subpopulation selection together, the total change in the frequency of A' is

$$\Delta q = \Delta q_w + \Delta q_b = -c\bar{q}(1-\bar{q})(1-F) + 2(b-c)\bar{q}(1-\bar{q})F \qquad (5.20)$$

The terms on the right-hand side can be interpreted by considering the extremes of $F = 0$ and $F = 1$. When $F = 0$, there is no population substructure, which means that all subpopulations have the same allele frequency $\bar{q}$; in this case, the change in allele frequency is just $-c\bar{q}(1-\bar{q})$. At the other extreme, when $F = 1$, each subpopulation is fixed for either A or A', and the proportion fixed for A' equals $\bar{q}$. The between-subpopulation selection is therefore analogous to selection between alleles in a haploid organism in which the fitnesses of A and A' demes are in the ratio $1 : 2(b-c)$. In this case, therefore, the change in allele frequency is $2(b-c)\bar{q}(1-\bar{q})$ (from Equation 5.11, assuming that $\bar{w} \approx 1$).

Equation 5.20 implies that $\Delta q > 0$ if

$$\frac{b-c}{c} > \frac{1-F}{2F} \qquad (5.21)$$

This is the condition necessary for selection between demes to override selection within demes, and the formulation is quite general (Crow and Aoki 1982). A biological interpretation of the inequality in Equation 5.21 can be inferred by comparison with the break-even point for kin selection given in Equations 5.18 and 5.19. Expressing Equation 5.21 in terms of $r = 2F/(1+F)$, which means that $F = r/(2-r)$, yields $c/b < r$; this condition is identical to Equation 5.18. In these models, the equivalence between kin selection and interdeme selection results from the shared remote ancestry of the members of each subpopulation caused by random genetic drift among the subpopulations. The members of each subpopulation are related by kinship, and so interdeme selection is the same phenomenon as kin selection; the break-even point is that at which the benefit b to one's kin through interdeme selection equals the cost to one's self c through direct selection against the A' allele.

5.7 SELECTION IN A FINITE POPULATION

In order to derive the dynamical behavior of alleles under selection, we have ignored lessons learned in previous chapters that populations are finite and that the sampling of alleles over generations can result in substantial allele frequency change. It is important to determine how the dynamics of alleles under selection changes when the population is no longer infinite, so that random drift intervenes as well. The model that we would like to explore couples together the standard model for natural selection with one locus and two alleles:

$$p' = \frac{p^2 w_{11} + pq w_{12}}{\bar{w}}$$

with the usual way that Fisher and Wright used to model random genetic drift, namely as a binomial sampling process:

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}$$

where p_{ij} is the probability of transitioning from i copies to j copies of allele A the next generation, when the current frequency of allele A is $p = i/(2N)$. This is enough to specify the model for forward simulation on a computer, but Wright wanted solutions that were easy to interpret. Recall from Chapter 3 that the diffusion approximation allowed one to consider effects of random drift and other pressures on allele frequencies. Let the fitnesses be additive, also known as genic selection, such that $w_{11} = 2s$, $w_{12} = s$, and $w_{22} = 1$. Wright (1931) also considered the case of bidirectional mutation, with a rate from A to a of u and a mutation rate from a to A of v. With these assumptions, the stationary distribution of allele frequencies under mutation, genic selection, and random genetic drift is given by

$$\phi(x) = Ce^{4N_e sx} x^{4N_e v-1}(1-x)^{4N_e u-1} \tag{5.22}$$

where C is a constant to make the integral from $x = 0$ to $x = 1$ equal 1. Figure 5.13 shows graphically some of the resulting site frequency spectra that arise from this model, showing the conditions when there is much greater (or less) intermediacy of allele frequencies than expected under mutation-drift balance.

Weak Selection and the Nearly Neutral Theory

In a finite population with natural selection operating, the intuition is clear that for a very large population, the classical equations for natural selection will hold, but in a very small population, random drift can get quite strong and the picture is less clear. Another way to approach the problem is to let the population size be constant at some intermediate level and to consider what happens when the selection coefficient gets smaller and smaller. When

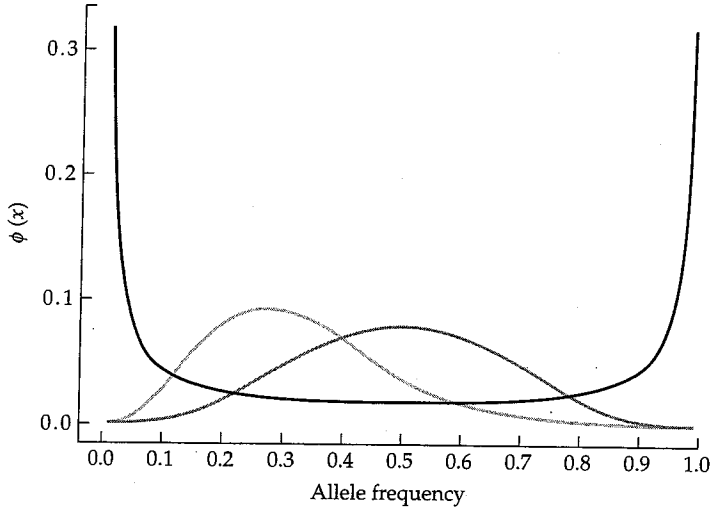


FIGURE 5.13 Site frequency spectra, $\phi(x)$, under a balance of selection, mutation and drift, following Wright's diffusion approximation (see Equation 5.22). The U-shaped solid curve has $4Nv = 4N\mu = 1$ and $4Ns = 0$, so that this represents a simple mutation-drift balance with most alleles being near fixation. The dark gray curve has $4Nv = 4N\mu = 4$ and $4Ns = 0$, so this case also lacks selection, but allele frequencies are intermediate because the population size is large enough that intermediate allele frequency is maintained just by mutation pressure. The light gray curve has $4Nv = 4N\mu = 4$ and $4Ns = -7$, and directional selection pushes the site frequency spectrum toward rarer frequencies.

the fitnesses of all genotypes are nearly equal, there is an interesting tension between selection and drift, and we call this situation **weak selection**. The study of weak selection took off when Tomoko Ohta first developed a theory for the fate of mutations with very small selection coefficients, which she called the *nearly neutral theory* (Ohta 1973). In this first paper, Ohta primarily focused on mutations that are "slightly deleterious" and ignored mutations that were "slightly advantageous." Later, when empirical evidence supported the idea that population also harbored slighted advantageous mutations as well, she modified the nearly neutral theory to include any mutations where $|2Ns| \approx 1$ (Ohta and Tachida 1990; Ohta 1992). Mutations that have this property of being nearly neutral (also called "weak selection") present an interesting challenge because for a large population size, natural selection can dominate their dynamics, but when the population size gets small, the mutations behave as if neutral.

One feature of the weak selection model that is particularly instructive is the probability of ultimate fixation of an allele. Following Kimura (1957), let

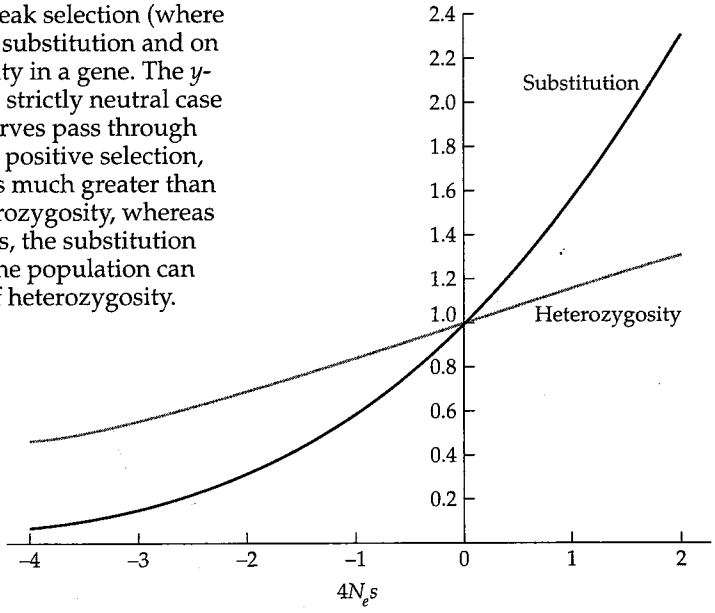


FIGURE 5.14 The effect of weak selection (where $-10 < 2N_es < 10$) on the rates of substitution and on standing levels of heterozygosity in a gene. The y-axis is expressed relative to the strictly neutral case (so that when $2N_es = 0$, both curves pass through 1.0 on the y-axis). Note that for positive selection, the effect on substitution rate is much greater than the impact of selection on heterozygosity, whereas when mutations are deleterious, the substitution rate falls nearly to zero while the population can still retain appreciable levels of heterozygosity.

$u(p)$ be the probability that a mutant allele whose frequency is currently p will ultimately be fixed in the population. If the fitnesses w_{11}, w_{12}, and w_{22} are replaced with $1+s$, $1+s/2$ and 1, then Kimura (1957) showed that:

$$u(p) = \frac{1-e^{-4Nsp}}{1-e^{-4Ns}} \tag{5.23}$$

For a newly arisen mutation when $p = 1/(2N)$, Kimura found that the probability of ultimate fixation of allele A is:

$$\Pr(A \text{ fixed}) = u = \frac{2s}{(1-e^{-4Ns})} \tag{5.24}$$

In the limit as s is reduced to zero (neutrality), the probability of fixation is $u = 1/(2N)$, as expected. Figure 5.14 plots Equation 5.24, and shows the relationship between the selection coefficient s to the probability of fixation with weak selection (scaled by the probability of fixation in the strictly neutral case).

This figure also shows the scaled heterozygosity in a population with weak selection. Without showing the derivation, this heterozygosity (scaled to the expected heterozygosity for a neutral allele integrated over generations until fixation) is:

$$\frac{H_t}{H_0} = \frac{2(4Ns - 1 + e^{-4Ns})}{4Ns(1+e^{-4Ns})} \tag{5.25}$$

Figure 5.14 shows that weakly deleterious alleles ($s < 0$) can have substantial heterozygosities, despite the fact that their probability of fixation is quite low. The reason is that when they are rare, they are mostly in heterozygotes and selection against them is even weaker, so random drift predominates in their dynamics. Selection against deleterious alleles becomes more effective if the deleterious allele attains an intermediate frequency, and so the probability of fixation is very low.

Genetic "Draft"

Context can be of key importance in evolution, and this is certainly true for the fate of selected genes. When many mutations are undergoing selective fixation, they tend to drag flanking variation to fixation with them through a process of **genetic hitchhiking** (Maynard Smith and Haigh 1974). Selection of this sort results in a reduction in the genetic variation for a region surrounding each selected target. A fundamental question raised by John Gillespie (2000) was whether the frequency of such selective events served to dominate the dynamics of neutral variants over the effect of random genetic drift. In an earlier paper, Gillespie (1999) was bothered by the failure of the standard neutral theory to explain why levels of standing variation are so poorly predicted by population size. He reasoned that one possible explanation could be that there are sufficiently many selective sweeps that levels of variability may be dominated by selection at linked sites. Many empirical papers have shown that levels of genetic variation are lower in regions of reduced local recombination rate (Aguadé et al. 1989; Begun and Aquadro 1992), and this could be partly due to hitchhiking and selective sweeps, or it could also be due to the effect of selection on deleterious mutations, which reduces the local effective size (Charlesworth et al. 1993). At any rate, it appears that something decouples population size from levels of variation, and Gillespie's work shows the conditions under which natural selection would have this effect. He furthered showed that the rate of substitution of weakly selected advantageous mutations decreases with increasing population size, whereas that for deleterious mutations increases with population size. The conclusion one draws from this model of weak selective sweeps, which Gillespie has whimsically (and confusingly) called "genetic draft," is that population size and binomial sampling may not be as relevant to a species' evolution as was once thought!

It should be clear by now that even the classical problems of natural selection are far from being totally understood. The complex interplay of mutation, random drift, and selection present many unsolved problems to challenge and engage the imaginative student.

SUMMARY

1. Natural selection occurs when any genetic difference results in differential representation of alleles over generations.
2. In haploid organisms with constant fitnesses, we can write equations that specify the full trajectory of allele frequency changes. In these models, the fitness is equivalent to the malthusian growth parameter for each haploid clone.
3. In a diploid organism, even the simplest model of constant fitnesses does not allow a closed-form equation specifying allele frequencies at arbitrary future generations.
4. Instead, most models of natural selection have specified changes in allele and genotypic frequencies across discrete generations, and the models are represented as one-generation difference equations.
5. The equilibrium behavior is especially of interest and is often easily solved analytically. For one locus and two alleles, when the fitnesses of *AA*, *Aa*, and *aa* (w_{11}, w_{12} and w_{22}) are ordered $w_{11} > w_{12} > w_{22}$, then the *A* allele goes to fixation. With overdominance ($w_{12} > w_{11}$ and $w_{12} > w_{22}$), then there is a stable polymorphic equilibrium. With underdominance ($w_{12} < w_{11}$, $w_{12} < w_{22}$) then there is an unstable polymorphic equilibrium, and the population goes to one fixation state or the other depending on initial genotype frequencies.
6. Mutation-selection balance refers to the maintenance of a harmful allele in a population at a low equilibrium frequency, because in every generation the elimination of preexisting harmful alleles by selection is offset by the introduction of new harmful alleles by mutation.
7. For a completely recessive harmful allele *a*, where the fitness of *aa* is $1 - s$, the equilibrium allele frequency is $\hat{q} = \sqrt{\mu/s}$, where μ is the rate of mutation from *A* to *a*. For a partially dominant allele, where fitnesses of *Aa* and *aa* are $1 - hs$ and $1 - s$, and h is the degree of dominance, the equilibrium allele frequency is approximately $\hat{q} = \mu/hs$.
8. Many other aspects of the life cycle may impact the way that natural selection works in a population, including X-linkage, frequency-dependent selection, differential fecundities, density-dependent selection, and so on.
9. Differential selection across geographically separted demes results in variation in allele frequencies across subpopulations. The interplay of migration and selection in subdivided populations may create conditions in which random genetic drift can produce slightly deleterious combinations of alleles at frequencies high enough for a subpopulation to come under the control of a superior adaptive peak (Wright's shifting balance theory).
10. When selection coefficients are close to satisfying $2Ns = 1$, there is an interplay between selection and drift that produces complex patterns of variation.

PROBLEMS

1. Explain how the average fitness in a population, $\overline{w}$, can be increasing at the same time that the population is going extinct.
2. In many mammals, birds, insects, and other bisexual animal populations, males develop elaborate physical or behavioral features that are used in attracting females. These include large antlers in male moose and elk, and the male peacock's tail. The energy expended in such ornaments is thought to be rewarded by reproductive success in attracting "choosy" females. Explain how male ornamentation and female choosiness can result in the process of *runaway sexual selection*, which leads to extraordinary allocation of resources to the imperatives of sexual attractiveness in males and extraordinary reproductive restraint in females. What can counteract runaway sexual selection?
3. Consider a population of diploid organism with genotypes *AA*, *Aa*, and *aa* for some gene. Suppose that genotype *AA* is an embryonic lethal and that genotype *aa* is fully viable but completely sterile. What genotype frequencies would be found among adults in a population at equilibrium? Do these genotype frequencies require the assumption of random mating?
4. Suppose that in the ith generation of a haploid population the fitnesses of *A* and *a* are given by $s_i : 1$. Show that $p_n/q_n = (p_0/q_0)(s_0 s_1 s_2 \ldots s_{n-1})$. If this is written as $p_n/q_n = (p_0/q_0)s^n$, then how can s be interpreted?
5. If the fitnesses of *AA*, *Aa*, *aa* are 1.0, 0.9, 0.6, and $p_0 = 0.7$, calculate p_1, p_2, and p_3, the allele frequencies after 1, 2, and 3 generations of selection.
6. Calculate the equilibrium allele frequency with overdominance when the fitnesses of *AA*, *Aa*, and *aa* are, respectively:
   (a) 0.300, 1, 0.700.
   (b) 0.930, 1, 0.970.
   (c) 0.993, 1, 0.997.
7. Calculate $\overline{w}$ for $w_{11} = 0.9$, $w_{12} = 1$, $w_{22} = 0.6$, and $p = 0.8$, assuming random mating. Does any other p give a larger $\overline{w}$? Why or why not?
8. If a rare allele that is lethal when homozygous ($s = 1$) decreases in frequency by 1% each generation (i.e., $q' = 0.99q$), then what is the selection coefficient ($h = hs$) against heterozygotes? (Hint: Assume that q and qh are small compared to 1.)
9. If selection is not too intense, an additive gene yielding fitnesses $1 + s$, $1 + s/2$, and 1 in *AA*, *Aa*, and *aa* genotypes will increase in frequency approximately according to $\ln(p_t/q_t) = \ln(p_0/q_0) + (s/2)t$. Calculate the approximate number of generations required to evolve significant insecticide resistance in an insect population when $s = \frac{1}{2}$ and $p_0 = 10^{-5}$. Significant resistance in the population may be taken as $p_t = 10^{-1}$. Show that, when q_0 and q_1 are both close enough to 1 that $\ln(p_t/q_t) \approx \ln(p_t)$and $\ln(p_0/q_0) \approx \ln(p_0)$, then $t \approx (2/s)\ln(p_t/p_0)$.

10. Show that a random mating diploid population with fitnesses $1, 1-s$, and $(1-s)^2$ for AA, Aa, and aa gives the same change in the allele frequency p of A as a haploid population with fitnesses 1 and $1-s$ of A and a.

11. If selection is not too strong, the time required for the allele frequency of a favored dominant allele to change from p_0 to p_t is given approximately by

$$\ln(p_t/q_t) + (1/q_t) = [\ln(p_0/q_0) + (1/q_0)] + st$$

Use this equation to derive the analogous equation for a favored recessive.

12. The following equation has equilibria at $\hat{p} = 0, \frac{1}{2}$, and 1. Classify the equilibria as to stability. If there is a stable equilibrium, is it locally or globally stable?

$$\Delta p = p(\tfrac{1}{2} - p)(1 - p)$$

13. Show that the allele frequency of a recessive lethal in generation n is given by $q_n = q_0/(1 + nq_0)$. (Hint: It is easiest to derive an expression first for $1/q_n$.) How many generations are required to reduce the allele frequency by half?

14. Consider a haploid organism in which genotypes A and a have frequencies p and q ($p + q = 1$) and relative fitnesses 1 and $1 - s$, respectively. Suppose that A mutates to a at a rate of μ per A allele per generation. Deduce the equilibrium frequency of a at the steady-state when there is a balance between mutation and selection.

15. The mutation rate to a dominant gene for neurofibromatosis is approximately 9×10^{-5} and the reproductive fitness of affected individuals is estimated as $\frac{1}{2}$. What is the expected equilibrium frequency of affected individuals at birth?

16. What is the equilibrium frequency of a recessive allele arising with a mutation rate of 4×10^{-6} and a reproductive fitness in homozygotes of 0.8? What would it be if the gene were partially dominant with $h = 0.05$?

17. What is the equilibrium frequency of a complete recessive allele arising with a mutation rate of 10^{-6} when the fitness in homozygous genotypes is 0.4? How much would the equilibrium frequency be reduced if the homozygous genotypes did not reproduce at all?

18. For a lethal allele maintained at an equilibrium frequency of $q = \mu/h$, where h is the selection coefficient against heterozygous genotypes, show that the proportion of heterozygous zygotes resulting from new mutations is approximately equal to h.

19. A polymorphism is said to be *protected* if all of the fixation states are unstable equilibria. Suppose the viabilities of males and females are as follows:

	AA	*Aa*	*aa*
Females	0.9	1	0.8
Males	1.0	v_{12}	0.5

What is the smallest value of v_{12} that ensures a protected polymorphism? (Hint: Some algebra shows that a condition for polymorphism is $w_{12}/w_{11} + v_{12}/v_{11} > 2$ and $w_{12}/w_{22} + v_{12}/v_{22} > 2$.)

20. If an allele a is a recessive lethal in zygotes and the relative fitness of $A : a$ among gametes is $1 - s : 1$, then what is the equilibrium allele frequency of a? Hint: If an equilibrium exists in this situation, it satisfies

$$p' = \frac{p^2 w_{11} v_1^2 + pq w_{12} v_1 v_2}{p^2 w_{11} v_1^2 + 2pq w_{12} v_1 v_2 + q^2 w_{22} v_2^2} = p$$

21. In a *Drosophila* population cage containing a chromosome known as *segregation distorter* (*SD*) that exhibits meiotic drive, the segregation ratio in heterozygous genotypes was about $k = 0.75$. When the chromosomes in the population cage had reached approximate equilibrium, the relative frequencies of the *SD* and non-*SD* chromosomes were approximately 0.11 and 0.89. The SD chromosome is homozygous lethal in both sexes. In this case a balance between viability and meiotic drive is reached when $\hat{p} = (2kw_{12} - 1)/(2w_{12} - 1)$, where $\hat{p}$ is the equilibrium frequency of the *SD* chromosome. Use this equation to estimate the approximate value of w_{12} consistent with these data.

22. Consider a gene with multiple alleles in which each heterozygous genotype is superior to the homozygous genotypes for each of the alleles it contains. Explain why this condition is not sufficient to ensure that all of the alleles will be maintained in a balanced polymorphism.

23. For a random-mating population with genotype frequencies of *AA*, *Aa*, and *aa* equal to p^2, $2pq$, and q^2, respectively, suppose that frequency-dependent selection occurs in such a way that the relative fitnesses of *AA*, *Aa*, and *aa* are given by $w_{11} = c/p^2$, $w_{12} = c/pq$, and $w_{22} = c/q^2$, respectively, where c is a constant. Derive an expression for p' and explain in words what it means. Also explain why this model breaks down when $p = 0$ or $p = 1$.

24. Suppose that the relative viabilities of genotypes $A'A'$, $A'A$, and AA are 0.5, 1, and 0.7, respectively. If the initial frequency of allele A' is 0.05, what will its frequency be when the population reaches an equilibrium? Now suppose that a mutation occurs in the equilibrium population that introduces a novel allele A^*, such that the relative fitnesses of A^*A^*, A^*A', and A^*A are all equal to 0.8. Will the A^* allele increase in frequency in the population? Why or why not?

25. Suppose alleles A_1, A_2, A_3, and A_4 are additive in their effects on fitness and the fitnesses of the homozygous genotypes are 0.8 for A_1A_1, 0.6 for A_2A_2, 0.4 for A_3A_3, and 0.2 for A_4A_4. What are the fitnesses of the heterozygous genotypes? In a random mating population in which all of the alleles are equally frequent, what is the average fitness in the population?

6.1 Inbreeding 257

The Inbreeding Coefficient 258

Genotype Frequencies with Inbreeding 259

Genetic Effects of Inbreeding 265

Calculation of the Inbreeding Coefficient from Pedigrees 268

Regular Systems of Mating 272

6.2 Population Subdivision 275

Reduction in Heterozygosity Due to Population Subdivision 276

Average Heterozygosity 278

Wright's F Statistics 281

Linanthus Revisited: Evidence for Selection Associated with Flower Color 285

Inference of Population Structure from Multilocus Genotype Data 286

6.3 The Wahlund Principle 288

Wahlund's Principle and the Fixation Index 291

Genotype Frequencies in Subdivided Populations 291

Relation between the Inbreeding Coefficient and the F Statistics 293

6.4 Assortative Mating 294

6.5 Migration 295

One-Way Migration 295

The Island Model of Migration 297

How Migration Limits Genetic Divergence 299

Estimates of Migration Rates 303

Coalescence-Based Estimates of Migration 303

Migration-Selection Balance 307

CHAPTER 6

INBREEDING, POPULATION SUBDIVISION, AND MIGRATION

In this chapter we consider some of the deep and important implications of the concept of *identity by descent*, which was introduced briefly in Chapter 3 in the context of random genetic drift. Here we show how this concept illuminates the consequences of mating between relatives, not only for the inbred individuals themselves but for genotype frequencies in the population as a whole. Then we show that population subdivision has similarities to inbreeding, because the members of any finite subpopulation are related to one another, though perhaps remotely in time. Finally we consider migration as an evolutionary process that counteracts the tendency for genetic divergence among subpopulations, and discuss applications of coalescence theory to the analysis of real data to make inferences about the history of mutation, migration, random drift, and natural selection among evolving subpopulations.

6.1 INBREEDING

When a mating take place between individuals that are related, the mating constitutes **inbreeding**, and the progeny that result are said to be **inbred**. In human beings, the closest degree of inbreeding usually encountered in most societies is first-cousin mating. But many plants regularly undergo self-fertilization, and some insects regularly practice brother-sister mating. By definition, relatives share one or more common ancestors in their pedigree, and it

is reasonable to suppose that these common ancestors will contribute disproportionately to the genotype of the offspring of a mating between relatives. But how can this effect be measured? The pioneering insight is due to Wright (1922), who formulated a measure of inbreeding called the *inbreeding coefficient* in terms of the correlation between uniting gametes. A later interpretation of the inbreeding coefficient in terms of probability is more transparent (Cotterman 1940; Malécot 1948), and this is the approach we will adopt.

The Inbreeding Coefficient

To make the discussion specific, consider the pedigree in Figure 6.1. It represents the closest degree of inbreeding possible, namely self-fertilization. The curved lines emanating from individual A in generation 0 represent gametes, which join to produce the individual I in generation 1. The black dots on the lines represent the alleles of a gene present in the gametes, which come together to form the genotype of the locus in individual I. The dots are both black to symbolize that they are **identical by descent**, which means that they arose from replication of the same DNA molecule in a previous generation, in this case in generation 0. The **inbreeding coefficient**, typically denoted by the symbol F, is defined as the probability that the two alleles at a locus in an inbred individual are identical by descent. Later in this chapter we will have a need to denote the inbreeding coefficient as F_{IS}, but for now we do not need the subscript and will suppress it.

The concept of identity by descent conceals a subtlety that requires some explanation. As one traces the ancestries of alleles into the past, their ancestral lineages come together, or *coalesce*, reducing the number of ancestral alleles, until ultimately only one common ancestral allele remains. Details of the coalescent process are examined in Chapter 3. Because of coalescence, every allele shares a common ancestral allele with every other allele, and they all are related through DNA replication. Superficially, the coalescent process seems to undermine the concept of the inbreeding coefficient, or at least to render it ambiguous. The ambiguity can be resolved by choosing some arbitrary time in the past and declaring that, at that time, every allele in the population is to be considered as being not identical by descent with any other. This clears the slate because it sets $F = 0$ for all individuals, and thereafter what the inbreeding coefficient actually refers to is the probability of identity by descent subsequent to the time the slate was cleared.

In Figure 6.1, the arbitrary time when all alleles are defined as distinct (not identical by descent) is generation 0. Hence we can write the genotype of individual A in generation 0 as $\alpha_1\alpha_2$, and so by definition α_1 and α_2 are not identical by descent. The probability that the alleles in the inbred offspring I are identical by descent can then be deduced from first principles. Individual I has any one of four possible genotypes with the following probabilities: $\frac{1}{4}\,\alpha_1\alpha_1$, $\frac{1}{4}\,\alpha_1\alpha_2$, $\frac{1}{4}\,\alpha_2\alpha_1$, and $\frac{1}{4}\,\alpha_2\alpha_2$. In the cases $\alpha_1\alpha_1$ and $\alpha_2\alpha_2$, the alleles are identical by descent, and the individual is said to be **autozygous** (the prefix

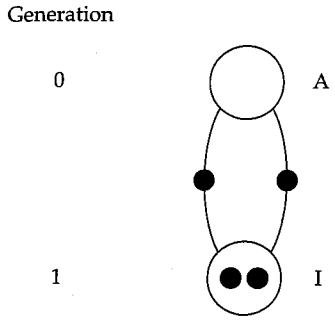


FIGURE 6.1 Pedigree for self-fertilization. Inbred individual I is the result of self-fertilization of the ancestor A. The black dots represent the alleles transmitted by A to I. The inbreeding coefficient of I is defined as the probability that the alleles of a gene in I are identical by descent.

auto means *self*). In the cases $\alpha_1\alpha_2$ and $\alpha_2\alpha_1$, the alleles are not identical by descent, and the individual is said to be **allozygous** (the prefix *allo* means *other*). Note that the concepts of autozygosity and allozygosity have nothing to do with the state of an allele—whether the allele is *A* or *a*, for example. The concepts are concerned only with common ancestry. If the alleles are replicas of a single allele in a common ancestor, they are autozygous; otherwise, they are allozygous.

Because the inbreeding coefficient is equal to the probability of autozygosity, the inbreeding coefficient of individual I is given by $F = \frac{1}{2}$; or to state the result in somewhat different terms, $F = \frac{1}{2}$ is the inbreeding coefficient resulting from one generation of self-fertilization. Two equally valid interpretations of F are:

- F is the probability that any particular gene in an inbred individual has alleles that are identical by descent, or
- F is the overall proportion of genes in an inbred individual that have alleles that are identical by descent.

Because $F = \frac{1}{2}$ in Figure 6.1, this value means that one generation of self-fertilization results in an inbred individual in which 50% of the genes have alleles that are identical by descent (autozygous). Because the span of time in a pedigree is usually short, in this case only one generation, mutation can safely be ignored. Autozygous genotypes must therefore be homozygous, whereas allozygous genotypes can be either homozygous or heterozygous.

Genotype Frequencies with Inbreeding

At the population level, Figure 6.2 illustrates how the concepts of autozygosity and allozygosity are related to those of homozygosity and heterozygosity in a population of inbred organisms. The small circles represent the individuals in a population, with their genotypes indicated for one locus. The population is assumed to be infinite, but to make matters concrete we will focus on these 32 individuals and assume that they are a perfectly rep-

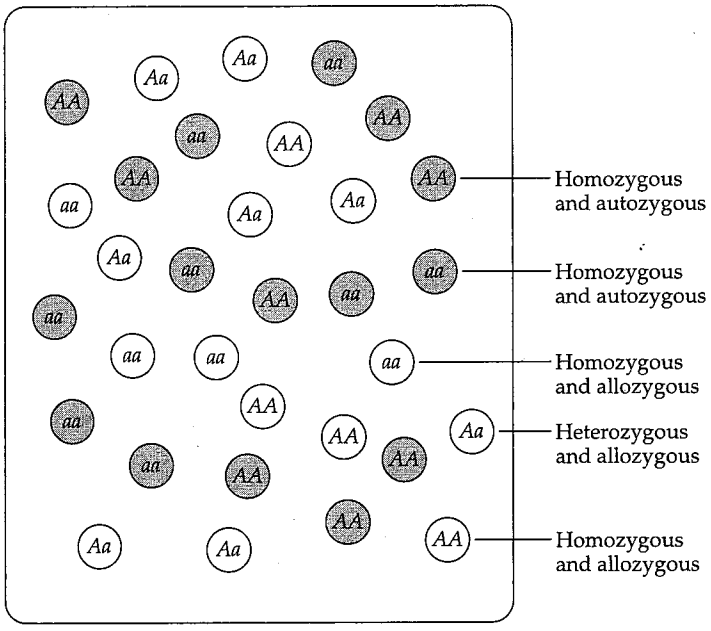


FIGURE 6.2 When there is inbreeding, the alleles in homozygous genotypes may be identical by descent (autozygous, here represented by the shaded circles) or not identical by descent (allozygous). In the absence of mutation, the alleles in heterozygous genotypes must be allozygous.

resentative sample. Some of the circles are shaded, and these represent autozygous genotypes whose alleles are identical by descent. Other circles are unshaded, and these represent allozygous genotypes whose alleles are not identical by descent. Disregarding the possibility of mutation since the time that F was declared to equal 0 (clearing the slate), all autozygous genotypes must be homozygous, but allozygous genotypes may be either homozygous or heterozygous (see Figure 6.2). Since F is the probability of identity by descent, it is also the proportion of individuals whose alleles are autozygous. In this example, $F = \frac{16}{32} = \frac{1}{2}$, which can be determined by counting. Normally one would not be able to distinguish which homozygous genotypes were autozygous and which were allozygous, and here we have chosen them arbitrarily.

The essential point of Figure 6.2 is that two alleles can be *identical by state*, which means that they have the same sequence of nucleotides along the DNA, without being identical by descent. The concept of *identity by descent* pertains to the ancestral origin of an allele and not to its chemical makeup.

Figure 6.2 also illustrates the effect of inbreeding on the genotype frequencies. In this population, the allele frequencies are $p = \frac{16}{32} = \frac{1}{2}$ for A, and $q = \frac{16}{32} = \frac{1}{2}$ for a. These again can be determined by direct counts. With Hardy-Weinberg equilibrium (see Chapter 2), the expected genotype frequencies are $(\frac{1}{2})^2 \times 32 = 8\ AA$, $2(\frac{1}{2})(\frac{1}{2}) \times 32 = 16\ Aa$, and $(\frac{1}{2})^2 \times 32 = 8\ aa$. The genotype counts are actually 12 AA, 8 Aa, and 12 aa. The excess of homozygous genotypes, and deficiency of heterozygous genotypes, are a direct consequence and characteristic of inbreeding.

To understand how inbreeding affects the genotype frequencies, we need only consider the implications of the definition of F for a population of inbred organisms. For this purpose, consider the alleles of a gene present in any one of the inbred organisms. Either of two things must be true: The alleles must either be allozygous (probability $1 - F$) or be autozygous (probability F). If the alleles are allozygous, then the probability that the chosen organism has any particular genotype is simply the probability of that genotype in a random-mating population, because, by chance, the inbreeding has not affected this particular gene. On the other hand, if the alleles are autozygous, then the chosen organism must be homozygous, and the probability of homozygosity for any particular allele is simply the frequency of the allele in the subpopulation as a whole. (Because the alleles in question are autozygous, knowing which allele is present in one chromosome immediately tells you that an identical allele is in the homologous chromosome.) These considerations hold regardless of the number of alleles, but to simplify matters, we consider the case of two alleles A and a at frequencies p and q (with $p + q = 1$). In this case the genotype frequencies are given by

$$AA: \quad p^2(1-F) + pF = p^2 + pqF \tag{6.1a}$$

$$Aa: \quad 2pq(1-F) = 2pq - 2pqF \tag{6.1b}$$

$$aa: \quad q^2(1-F) + pF = q^2 + pqF \tag{6.1c}$$

Equation 6.1a is the probability that an organism has genotype AA; the first term refers to cases in which the alleles are allozygous and the second to cases in which the alleles are autozygous. Similarly, Equation 6.1c is the probability that an organism has genotype aa. Heterozygous Aa genotypes then have the frequency given by Equation 6.1b, since alleles that are heterozygous must be allozygous. The expressions at the far right in Equations 6.1a–c can be obtained by multiplying out those on the left and remembering that $p(1-p) = q(1-q) = pq$.

Applying Equation 6.1 to the example in Figure 6.2, we have already shown that $F = \frac{1}{2}$ and also that $p = q = \frac{1}{2}$. From the expressions on the far right in Equation 6, therefore, the expected numbers of each of the three genotypes are $[(\frac{1}{2})^2 + (\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 12\ AA$, $[2(\frac{1}{2})(\frac{1}{2}) - 2(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 8\ Aa$, and $[(\frac{1}{2})^2 + (\frac{1}{2})(\frac{1}{2})(\frac{1}{2})] \times 32 = 12\ aa$. The genotype frequencies shown in Figure 6.2 are therefore in perfect agreement with those expected in the progeny of

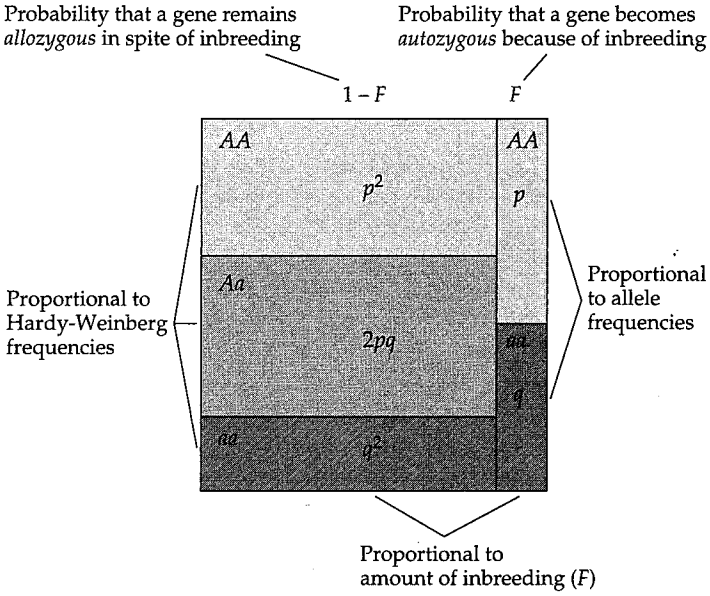


FIGURE 6.3 Graphical representation of the effects of inbreeding on genotype frequencies. Some genes remain allozygous in spite of the inbreeding, and among these the genotype frequencies of *AA*, *Aa*, and *aa* are given by the Hardy-Weinberg principle. Other genes are autozygous because of the inbreeding, and among these the genotype frequencies of *AA* and *aa* are given by the allele frequencies. There are no heterozygotes in the autozygous case because the two alleles present at an autozygous locus are, by definition, identical by descent.

a population of plants in which each individual had undergone one generation of self-fertilization.

The genotype frequencies with inbreeding are summarized graphically in Figure 6.3. The box is divided vertically into two parts, corresponding to genes whose alleles remain allozygous in spite of the inbreeding and those whose alleles are autozygous because of the inbreeding. The division is in the proportion $1 - F : F$. Within the allozygous part of the box, the horizontal panels correspond to the allozygous genotypes *AA*, *Aa*, and *aa*, which are the Hardy-Weinberg frequencies. Within the autozygous part of the box, the horizontal panels correspond to the autozygous genotypes *AA* and *aa*, which are in the proportions $p : q$. Some special cases of Equation 6.1 for the genotype frequencies with inbreeding are given in Table 6.1. When $F = 0$, the genotype frequencies are identical to those of the Hardy-Weinberg principle; and when $F = 1$ (complete inbreeding), all individuals are autozygous, and there is a total absence of heterozygous genotypes.

TABLE 6.1 Genotype Frequencies with Inbreeding

	Frequency in Population		
Genotype	With inbreeding coefficient F	With $F = 0$ (random mating)	With $F = 1$ (complete inbreeding)
AA	$p^2(1-F) + pF$	p^2	p
Aa	$2pq(1-F)$	$2pq$	0
aa	$\underbrace{q^2(1-F)}_{\text{Allozygous genes}} + \underbrace{qF}_{\text{Autozygous genes}}$	q^2	q

Note also from Equation 6.1 that, while inbreeding does change the genotype frequencies in a population, it does not change the allele frequencies. This is true because, for any value of F, the allele frequency of A is given by $[p^2 + pqF] + (\frac{1}{2})[2pq - 2pqF] = p^2 + pq = p(p+q) = p$. This principle requires the assumption that all genotypes have the same fitness, which is to say that no natural selection takes place. If there is selection, then the allele frequencies can change with inbreeding.

Equation 6.1 generalizes to multiple alleles in a strightforward way. If a gene has multiple alleles $A_1, A_2, \ldots A_n$ at respective frequencies $p_1, p_2, \ldots p_n$ (with $p_1 + p_2 + \cdots + p_n = 1$), then in a population with inbreeding coefficient F, the frequencies of A_iA_i homozygotes and A_iA_j heterozygotes are as follows:

$$A_iA_i: \quad p_i^2(1-F) + p_iF = p_i^2 + p_i(1-p_i)F \tag{6.2a}$$

$$A_iA_j: \quad 2p_ip_j(1-F) = 2p_ip_j - 2p_ip_jF \tag{6.2b}$$

CORRELATION BETWEEN UNITING GAMETES Wright's (1922) original conception of the inbreeding coefficient F was as a measure of the correlation between uniting gametes. That this concept is consistent with the probability interpretation is shown for gene with two alleles in Table 6.2. The upper part of the table shows all possible pairs of uniting gametes and their frequencies with inbreeding, with the female gamete on the left and the male gamete on the right. The alleles have been coded with numerical values, A with a value 1 and a with value 0. Any arbitrary numerical values lead to the same conclusion, but the assignments in Table 6.2 simplify the formulas. The bottom part of the table shows how various expected values are calculated, with the goal of deducing $Cov(xy)$, the covariance between x and y, as well as $V(x)$ and $V(y)$, the variances. By definition, the correlation between uniting gametes r_{UG} (UG for uniting gametes) is the ratio of the covariance to the product of the standard deviations, and hence

TABLE 6.2 Correlation between Uniting Gametes

Uniting gametes	Relative frequency	Expected values
$A\ (x = 1)\quad A\ (y = 1)$	$p^2 + pqF$	$E(x) = p^2 + pqF + pq - pqF = p^2 + pq = p(p + q) = p$
$A\ (x = 1)\quad a\ (y = 0)$	$pq - pqF$	$E\ (x^2) = p^2 + pqF + pq - pqF = p^2 + pq = p(p + q) = p$
$a\ (x = 0)\quad A\ (y = 1)$	$pq - pqF$	$E\ (y) = p^2 + pqF + pq - pqF = p^2 + pq = p(p + q) = p$
$a\ (x = 0)\quad a\ (y = 0)$	$q^2 + pqF$	$E\ (y^2) = p^2 + pqF + pq - pqF = p^2 + pq = p(p + q) = p$
		$E\ (xy) = p^2 + pqF$
		$Cov(xy) = E(xy) - E(x)E(y) = p^2 + pqF - p \times p = pqF$
		$V(x) = E\ (x^2) - [E(x)]^2 = p - p^2 = p(1 - p) = pq$
		$V(y) = E\ (y^2) - [E(y)]^2 = p - p^2 = p(1 - p) = pq$

$$r_{UG} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{pqF}{pq} = F \tag{6.3}$$

Wright, to the end of his long and extraordinarily productive life (he died in 1988 at the age of 98), always preferred his own interpretation of F as a correlation, because under some exceptional circumstances r_{UG} can be negative, and in these cases the probability interpretations fails because a probability cannot be negative.

REDUCTION IN THE FREQUENCY OF HETEROZYGOUS GENOTYPES One of the main effects of inbreeding is that a group of inbred individuals has reduced frequency of heterozygous genotypes, relative to a group of noninbred individuals (see Equation 6.1b). To examine this effect quantitatively, let H_I denote the probability that a gene in an inbred individual is heterozygous, and let H_S denote the proportion of heterozygous genotypes expected with random mating in the subpopulation of which I is a member. With two alleles, Equation 6.1b implies that $H_I = 2pq - 2pqF$, and the Hardy-Weinberg principle implies that $H_S = 2pq$. The proportionate reduction in heterozygosity due to inbreeding, relative to the subpopulation as a whole, is symbolized as F_{IS} and given by the expression

$$F_{IS} = \frac{H_S - H_I}{H_S} = \frac{2pq - (2pq - 2pqF)}{2pq} = F \tag{6.4}$$

As we shall see Section 6.2, this formulation is particularly useful in thinking about subdivided populations, when both inbreeding and random genetic drift contribute to the overall probability of identity by descent.

PROBLEM 6.1 Plants able to undergo self-fertilization are said to be *self-compatible*. In a population of self-compatible plants, if each plant undergoes self-fertilization a fraction s of the time and otherwise mates randomly, then it can be shown (Crow and Kimura 1970; Hedrick and Cockerham 1986) that F very quickly attains the value $F = s/(2 - s)$. *Phlox cuspidata* is self-compatible, and for this species the amount of self-fertilization is estimated at approximately $s = 0.78$ (Levin 1978). From s we can predict the inbreeding coefficient as $F = 0.78/(2 - 0.78) = 0.64$. In a sample of 35 plants from a Texas population of *P. cuspidata*, two alleles of the phosphoglucomutase-2 gene were observed, which we will designate as the *A* and *a* alleles. The sample included were 15 *AA*, 6 *Aa*, and 14 *aa* genotypes (Levin 1978). Are these numbers consistent with the estimate $F = 0.64$? (*Note:* The χ^2 in this case has one degree of freedom because only the allele frequency is estimated from the data; if F also were estimated from the data, rather than being calculated independently from the degree of self-fertilization, then there would be zero degrees of freedom and no goodness-of-fit test would be possible.)

ANSWER The allele frequencies of *A* and *a* are estimated as $(30 + 6)/70 = 0.514$ and $1 - 0.514 = 0.486$, respectively. The hypothesis is that $F = 0.64$, and so $1 - F = 0.36$. The expected numbers of the genotypes *AA*, *Aa*, and *aa* are, respectively, $[(0.514)^2(0.36) + (0.514)(0.64)](35) = 14.8$, $[2(0.514)(0.486)(0.36)](35) = 6.3$, and $[(0.486)^2(0.36) + (0.486)(0.64)](35) = 13.9$. With these expectations, the $\chi^2 = 0.02$ with one degree of freedom, and the associated probability is about 0.96. The fit to the inbreeding model is excellent.

Genetic Effects of Inbreeding

In **outcrossing** species (those that regularly avoid mating between relatives), close inbreeding is generally harmful. The effects are seen most dramatically when inbreeding is complete or nearly complete. In most species of animals, complete autozygosity requires many generations of brother-sister mating. But in *Drosophila melanogaster*, autozygosity of entire chromosomes can be achieved in just a few generations because of the absence of crossing over in the male and the ready availability of genetically marked chromosomes with multiple inversions to prevent crossing over in the female. One widely used inversion chromosome is marked with the dominant mutation *Cy* (for Curly wings), and the critical experimental cross is of the form $Cy/+_i \times Cy/+_i$, where is the *i*th member of a sample of wildtype chromosomes isolated from a natural population, and the $+_i$ chromosomes are identical by descent. Homozygous *Cy* genotypes do not survive, and so the theoretically expected progeny are $Cy/+_i$ (with curly wings) and $+_i/+_i$ (with straight wings) in a ratio of $\frac{2}{3}:\frac{1}{3}$. If the wildtype chromosome carries one or more mutations that decrease survivorship, then there will be fewer than $\frac{1}{3}$ straight-wing flies,

and if the chromosome carries a recessive lethal mutation, then straight-wing flies will be absent. Control crosses are of the form $Cy/+_i \times Cy/+_j$, where the $+_i$ and $+_j$ chromosomes are not identical by descent. For either type of mating, an estimate $\hat{v}$ of the viability (survivorship) of the +/+ genotype, relative to that of the $Cy/+$ genotype, is given by

$$\hat{v} = \frac{2n_{+/+}}{1+n_{Cy/+}} \tag{6.5}$$

where $n_{+/+}$ and $n_{Cy/+}$ are the counts of wildtype and *Curly* offspring, respectively (Haldane 1956). The addition of 1 to the denominator makes the estimate of v almost unbiased. When the total number of offspring is large, $\hat{v}$ is essentially equal to two times the number of wildtype offspring divided by the number of *Curly* offspring.

Results of such experiment to estimate the viabilities of autozygous (homozygous) and allozygous (heterozygous) wildtype chromosomes from a natural population are shown in Figure 6.4. It is evident that the homozygous genotypes (shaded histogram) are relatively poor in viability. In fact, about 37% of the homozygous genotypes are lethal. Moreover, among the homozygous genotypes that have viabilities within the normal range of the heterozygous genotypes (open histogram), virtually all can be shown to have reduced fertility (Sved 1975; Simmons and Crow 1977). Inbreeding so close as to make entire chromosomes homozygous is rare in outcrossing species, except in the kind of experiment in Figure 6.4, but the effects are clearly very harmful and provide a new dimension of genetic diversity. In the case of single nucleotide polymorphisms (SNPs), genetic diversity results from common alleles that do not perceptibly impair viability or fertility when homozygous. In the case of inbreeding, the effects are mainly due to rare alleles that are severely detrimental when homozygous. (The fact that the alleles are rare is shown by the small proportion of lethal or near-lethal heterozygous combinations.) Figure 6.4 shows that natural populations of *Drosophila* contain considerable hidden genetic variation in the form of rare deleterious recessive alleles.

Detrimental effects of inbreeding, called **inbreeding depression**, are found in virtually all outcrossing species, and the more intense the inbreeding, the more harmful the effects. Inbreeding in human beings is also generally harmful, but the effect is difficult to measure because the degree of inbreeding is less than that in experimental organisms, and the effects may also vary from population to population. Nevertheless, children of first-cousin matings are, on the average, less capable than noninbred children in any number of ways (for example, higher rate of mortality, lower IQ scores). It should be emphasized, however, that many such children are within the normal range of abilities, and some are quite gifted. Sewall Wright, the celebrated population geneticist, was the child of a first-cousin marriage.

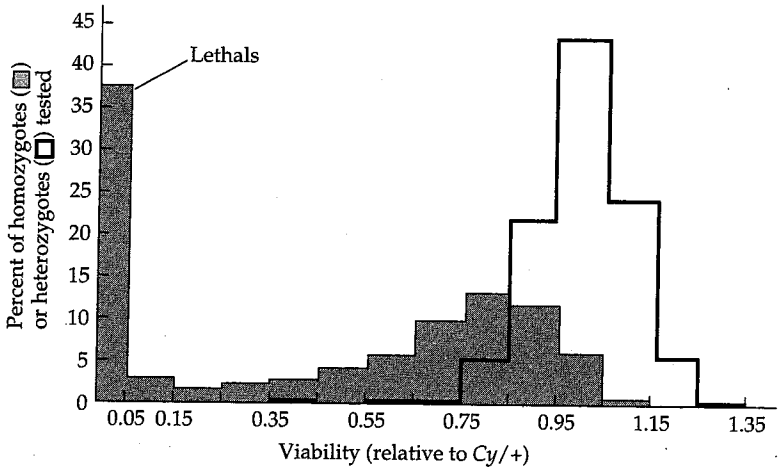


FIGURE 6.4 Viability distributions of wildtype homozygotes (shaded area) and wildtype heterozygotes (black outline) of second chromosomes extracted from *Drosophila melanogaster*. The histograms depict results of testing 691 homozygous combinations and 688 heterozygous combinations. Note that, in this sample, nearly 37% of the wildtype chromosomes are lethal when homozygous, and many more have viabilities substantially below normal. (Data from Mukai et al. 1974.)

In human populations, as in most organisms, the deleterious effects of inbreeding are largely due to the increased homozygosity of rare recessive alleles, and so inbreeding effects in human beings are seen most dramatically in the increased frequency of genetic abnormalities due to harmful recessive alleles among the children of first-cousin matings. The increased frequency of such conditions can be deduced from the genotype frequency given in Equation 6.1c. For the offspring of a first-cousin mating, $F = \frac{1}{16}$, as will be shown in the next section. Suppose that a is a rare deleterious recessive allele with an allele frequency of q. Then, among the children of first-cousin matings, the frequency of aa is expected to be $q^2 + pq(\frac{1}{16})$. On the other hand, among the offspring of matings that take place at random, the frequency of homozygous recessives equals q^2.

Now if c is the proportion of first-cousin matings in a population, then the expected proportion of homozygous aa offspring in the population as a whole that result from the first-cousin matings is given by

$$\frac{c(q^2 + pq/16)}{q^2(1-c) + c(q^2 + pq/16)} = \frac{c(1+15q)}{c+16q-cq} \qquad [6.6]$$

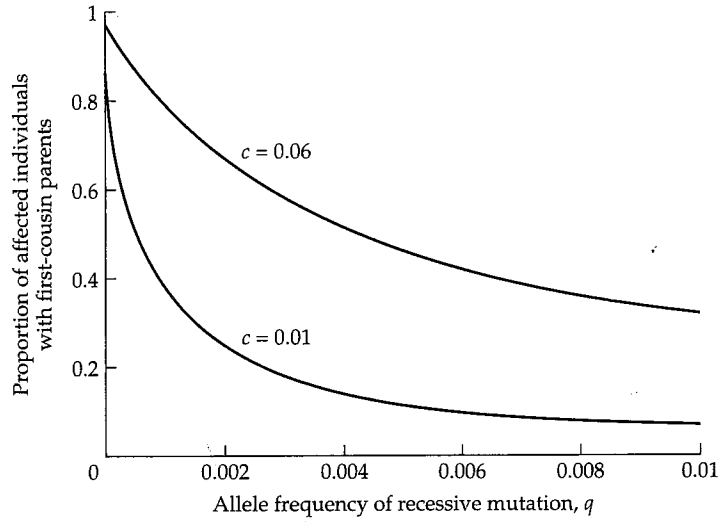


FIGURE 6.5 For rare recessive alleles, even low levels of inbreeding can account for a disproportionate share of homozygous recessive offspring. The curves correspond to overall proportions of first-cousin mating of 1% and 6%, and show that when a recessive allele is rare, a large proportion of homozygous genotypes result from the small proportion of first-cousin matings. The reason is that the offspring of first cousins have a $\frac{1}{16}$ chance of carrying alleles that are identical by descent.

Figure 6.5 shows plots of this proportion for $c = 0.01$ and $c = 0.06$, a range that includes most human populations. Note that, as the recessive allele becomes more rare, the first-cousin matings account for an increasing proportion of all affected children. Consider albinism as an example, which is due to a rare recessive mutation. Although the allele frequency differs among human subpopulations, we will take $q = 0.005$ as typical, which predicts a frequency among the children of nonrelatives of $q^2 = 0.0025\%$, or about one in 40,000. The curves in Figure 6.5 imply that, when the frequency of first cousin matings equals 1% (approximately the value in the Unites States), then the proportion of albino children whose parents are first cousins is 12%. In a population with $c = 0.06$, although first-cousin matings account for 6% of all matings, they account for 46% of matings with albino children.

Calculation of the Inbreeding Coefficient from Pedigrees

Calculation of F from a pedigree is simplified by drawing the pedigree in the form shown in Figure 6.6A, where the lines represent gametes contributed by parents to their offspring. The same pedigree is shown in conventional form in Figure 6.6B. The organisms in gray in part B are not represented in part A

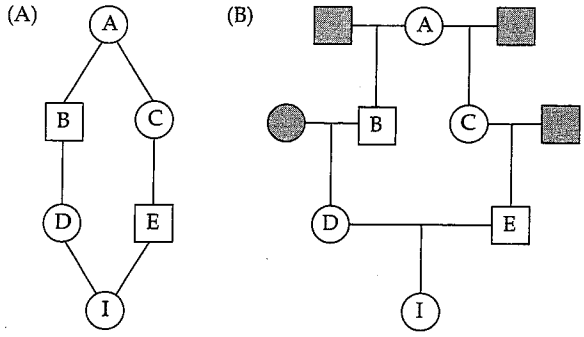


FIGURE 6.6 (A) Convenient way to represent pedigrees for calculation of the inbreeding coefficient. In this case, the pedigree shows a mating between half-first cousins. (B) Conventional representation of the same pedigree as in part A. Squares represent males, circles represent females, and the shaded organisms in part B are not depicted in part A because they do not contribute to the inbreeding of the inbred organism designated I.

because they have no ancestors in common and therefore do not contribute to the inbreeding of the organism denoted I. The inbreeding coefficient F_I of I is the probability that I is autozygous for the alleles of an autosomal gene under consideration. The first step in calculating F_I is to locate all the common ancestors in the pedigree, because an allele could become autozygous in I only if it were inherited through both of I's parents from a common ancestor; in this case, there is only one common ancestor, namely, A. The next step in calculating F_I, which is carried out for each common ancestor in turn, is to trace all the paths of gametes that lead from one of I's parents back to the common ancestor and then down again to the other parent of I. These paths are the paths along which an allele in a common ancestor could become autozygous in I. In Figure 6.6A, there is only one such path: DB<u>A</u>CE, in which the common ancestor is underlined for bookkeeping purposes, an especially useful procedure in complex pedigrees.

The third step in calculating F_I is to calculate the probability of autozygosity in I due to each of the paths in turn. For the path DB<u>A</u>CE, the reasoning is illustrated in Figure 6.7. Here the black dots represent alleles transmitted along the gametic paths, and the number associated with each step is the probability of identity by descent of the alleles indicated. For all steps except that around the common ancestor, the probability is $\frac{1}{2}$ because, with Mendelian segregation, the probability that a particular allele present in a parent is transmitted to a specified offspring is $\frac{1}{2}$. To understand why $(\frac{1}{2})(1 + F_A)$ is the probability associated with the loop around the common ancestor, denote the alleles in the common ancestor as α_1 and α_2. These symbols are

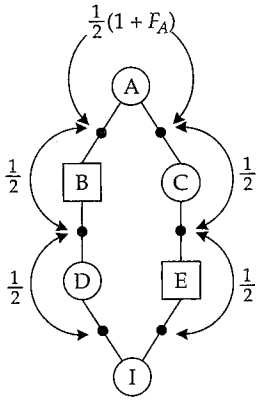


FIGURE 6.7 Loops for the pedigree in Figure 6.6A, showing probabilities that designated alleles (solid dots) are identical by descent. Each loop is independent of the others, so their probabilities multiply. Thus, the inbreeding coefficient of organism I is $F_I = (\frac{1}{2})^5(1 + F_A)$, where F_A represents the inbreeding coefficient of the common ancestor.

used to avoid confusion with conventional allele symbols designating functional types of alleles, such as *A* for dominant and *a* for recessive. The pair of gametes contributed by A could contain $\alpha_1\alpha_1$, $\alpha_2\alpha_2$, $\alpha_1\alpha_2$, or $\alpha_2\alpha_1$, each with a probability of $\frac{1}{4}$ because of Mendelian segregation. In the first two cases, the alleles are clearly identical by descent; in the second two cases, the alleles are identical by descent only if α_1 and α_2 are already identical by descent, which means that A is autozygous. The probability that A is autozygous is, by definition, the inbreeding coefficient of A, F_A. Hence, the probability for the step around the common ancestor A is $(\frac{1}{4}) + (\frac{1}{4}) + (\frac{1}{4})\,F_A + (\frac{1}{4})\,F_A = (\frac{1}{2}) + (\frac{1}{2})\,F_A = (\frac{1}{2})(1 + F_A)$. Because each of the steps in Figure 6.7 is independent of the others, the total probability of autozygosity in I due to the path through A is $(\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2})(1 + F_A) \times (\frac{1}{2}) \times (\frac{1}{2}) = (\frac{1}{2})^5(1 + F_A)$. Make special note that the exponent on the $(\frac{1}{2})$ is simply the total number of ancestors in the path. In general, if a path through a common ancestor A contains *i* individuals, the probability of autozygosity due to that path is

$$(\tfrac{1}{2})^i(1 + F_A)$$

Thus, the inbreeding coefficient of I in Figure 6.6A is $(\frac{1}{2})^5(1 + F_A)$. Assuming that A itself is not inbred ($F_A = 0$), the inbreeding coefficient of I reduces to $F_I = (\frac{1}{2})^5 = \frac{1}{32}$.

In pedigrees of greater complexity, there is more than one common ancestor and there may be more than one path through any of the common ancestors. The paths are mutually exclusive because autozygosity due to an allele inherited along one path excludes autozygosity due to an allele inherited along a different path. Thus, the total inbreeding coefficient is the sum of the probabilities of autozygosity due to each path considered separately. The whole procedure for calculating *F* is summarized in an example of a first-cousin mating in Figure 6.8. In a first-cousin mating, there are two common ancestors (A and B) and two paths (one each through A and B). The total inbreeding coefficient of I is the sum of the two separate contributions shown in Figure 6.8. If A and B are both noninbred, then $F_A = F_B = 0$, and so $F_I = (\frac{1}{2})^5 + (\frac{1}{2})^5 = \frac{1}{16}$. The result $F_I = \frac{1}{16}$ is the probability that I is autozygous at the specified locus; alternatively, F_I can be interpreted as the average proportion of all genes in I in which the alleles present are autozygous.

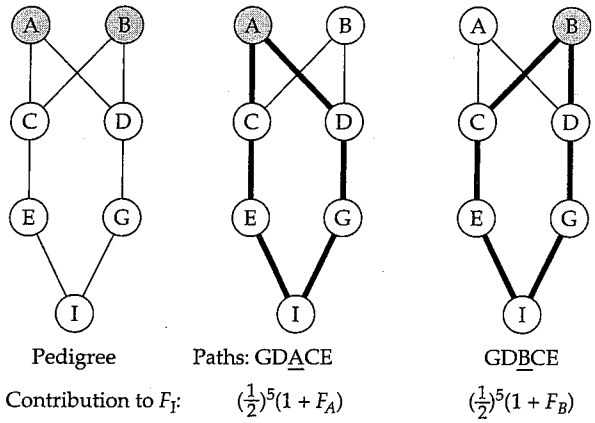


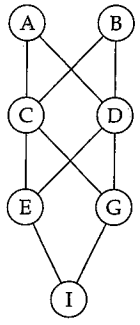
FIGURE 6.8 On the left is a pedigree of individual I, the offspring of a first-cousin mating. On the right are the two paths through common ancestors (heavy lines) used in calculating the inbreeding coefficient of I. Below each path is the contribution to F_I due to that path, calculated as in Figure 6.7. Each path is mutually exclusive of the others, and so their probabilities add. Thus, the total inbreeding coefficient of I is the sum of the two separate contributions. If $F_A = F_B = 0$, then $F_I = \frac{1}{16}$.

In general, for any autosomal gene, the formula for calculating the inbreeding coefficient F_I of an inbred organism I is

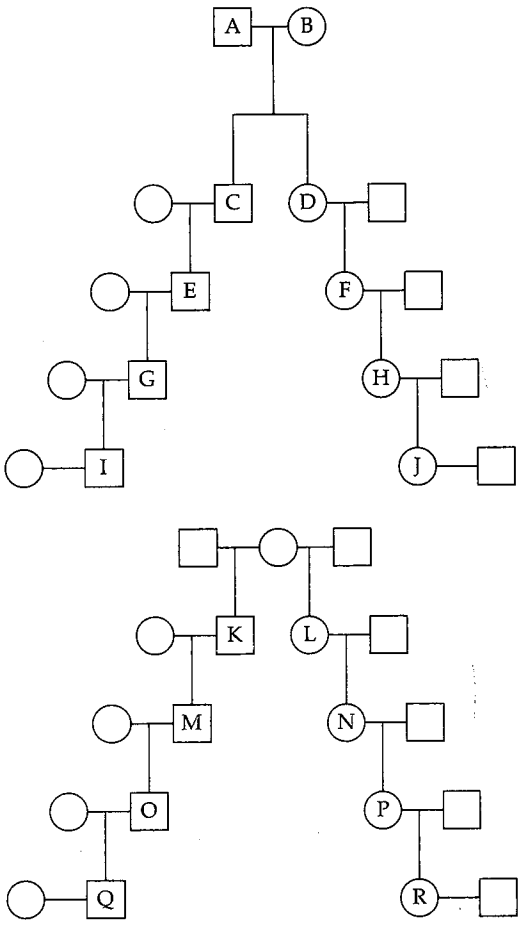
$$F_I = \sum_A \left(\frac{1}{2}\right)^i (1 + F_A) \qquad (6.7)$$

in which the summation over A means summation over all possible paths through all common ancestors, i is the number of organisms in each path, and A is the common ancestor in each path. Figure 6.9 gives the inbreeding coefficient of an offspring produced by mating between any of several common types of relatives in human pedigrees.

PROBLEM 6.2 The accompanying pedigree depicts two generations of brother-sister mating. Calculate the inbreeding coefficient of I, assuming that none of the common ancestors is inbred. (Altogether, there are four common ancestors and six paths.)



ANSWER $F_I = (\frac{1}{2})^3(1 + F_C) + (\frac{1}{2})^3(1 + F_D) + (\frac{1}{2})^5(1 + F_A) + (\frac{1}{2})^5(1 + F_B) + (\frac{1}{2})^5(1 + F_A) + (\frac{1}{2})^5(1 + F_B)$. When the common ancestors are assumed to be noninbred, then $F_A = F_B = F_C = F_D = 0$, and so $F_I = \frac{3}{8}$.



Individuals	Relationship	*F* of hypothetical offspring
AB	Unrelated	0
AD	Parent-offspring	$\frac{1}{4}$
CD	Full siblings	$\frac{1}{4}$
CF	Uncle-niece	$\frac{1}{8}$
EF	First cousins	$\frac{1}{16}$
EH	First cousins once removed	$\frac{1}{32}$
GH	Second cousins	$\frac{1}{64}$
GJ	Second cousins once removed	$\frac{1}{128}$
IJ	Third cousins	$\frac{1}{256}$
KL	Half siblings	$\frac{1}{8}$
MN	Half first cousins	$\frac{1}{32}$
MP	Half first cousins once removed	$\frac{1}{64}$
OP	Half second cousins	$\frac{1}{128}$
OR	Half second cousins once removed	$\frac{1}{256}$
QR	Half third cousins	$\frac{1}{512}$

FIGURE 6.9 Inbreeding coefficient of the offspring of various types of consanguineous mating.

Regular Systems of Mating

In plant and animal breeding, it is often important to know how rapidly the inbreeding coefficient increases when a strain is propagated by a **regular system of mating**, a systematic and repeated pattern of inbreeding, such as self-fertilization, sib mating, or backcrossing to a standard strain. The reasoning involved in calculating the inbreeding coefficient for any generation is illustrated in Figure 6.10 for repeated self-fertilization. In this figure, the labels $t-1$ and t refer to the inbred organisms after $t-1$ and t generations of self-fertilization. The loop around the ancestor in generation $t-1$ designates the probability that the two indicated alleles are identical by descent. Here the

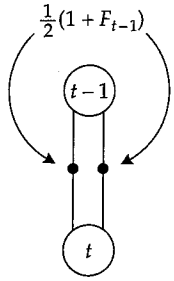


FIGURE 6.10 Increase in F resulting from continued self-fertilization. The organism in generation t is the offspring of self-fertilization of the organism in generation $t - 1$. The loop shows that $F_t = (\frac{1}{2})(1 + F_{t-1})$.

formula in Equation 6.7 applies with only one path and only one ancestor in the path, and so $F_t = (\frac{1}{2})^1(1 + F_{t-1})$, where F_t is the inbreeding coefficient in generation t. This equation is easy to solve in terms of the quantity $1 - F_t$, which is often called the *panmictic index* (*panmixia* is an old-fashioned word for *random mating*). Multiplying both sides of the equation for F_t by -1 and then adding $+1$ to each side leads to $1 - F_t = 1 - (\frac{1}{2})(1 + F_{t-1}) = 1 - (\frac{1}{2}) - (\frac{1}{2})F_{t-1} = (\frac{1}{2})(1 - F_{t-1})$, or

$$1 - F_t = (\tfrac{1}{2})^t(1 - F_0) \qquad [6.8]$$

where F_0 is the inbreeding coefficient in the initial generation when the repeated self-fertilization begins. Self-fertilization therefore leads to an extremely rapid increase in the inbreeding coefficient. When $F_0 = 0$, then $F_1 = \frac{1}{2}$, $F_2 = \frac{3}{4}$, $F_3 = \frac{7}{8}$, $F_4 = \frac{15}{16}$, and so on. The increase in F under self-fertilization and several other regular systems of mating is shown in Figure 6.11. No mat-

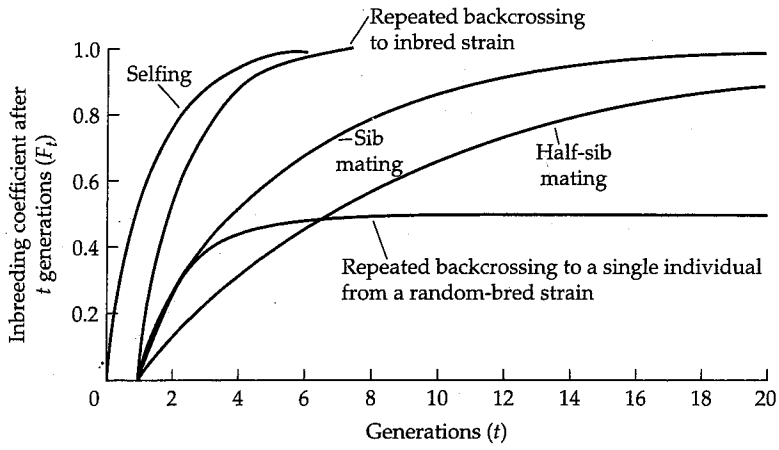


FIGURE 6.11 Theoretical increase in the inbreeding coefficient F for regular systems of mating: selfing, sib mating, half-sib mating, and repeated backcrossing to a single organism from a random-bred strain. In each case, the initial value of F is assumed to be $F_0 = 0$.

ter how much inbreeding has taken place in a population, a single generation of random mating completely erases the effects, and the genotype frequencies return to the Hardy-Weinberg proportions.

Many plants reproduce predominantly by self-fertilization, including crop plants such as soybeans, sorghum, barley, and wheat. As expected of highly self-fertilizing species, each plant is highly homozygous for alleles. Yet in comparing different populations, the proportion of polymorphic genes is comparable to that found in outcrossing species. Polymorphisms are found because self-fertilization does not eliminate genetic variation; it simply reorganizes genetic variation into homozygous genotypes. On the other hand, self-fertilizing species do contain fewer deleterious recessives than do outcrossing species, presumably because the increased frequency of homozygous recessive genotypes permits deleterious mutations to be eliminated from the population by natural selection. The high frequency of homozygous genotypes in naturally self-fertilizing species also impedes recombination producing new gametic types not already present in the parent. Therefore, a predominance of self-fertilization has the effect of slowing the approach to linkage equilibrium because the approach to linkage equilibrium is through recombination in double heterozygotes (*A B/a b* and *A b/a B* in the case of two alleles at each locus); with extreme inbreeding, such doubly heterozygous genotypes are rare. Indeed, the most extreme examples of linkage disequilibrium have been found in predominantly self-fertilizing species such as barley (*Hordeum vulgare*) and wild oats (*Avena barbata*).

Barley, which regularly undergoes more than 99% self-fertilization, provides an extreme example of linkage disequilibrium between two unlinked esterase genes (Clegg et al. 1972). A population that had originated as a complex cross was maintained for 26 generations under normal agricultural conditions without conscious selection. The population was polymorphic for two alleles of an *Esterase-B* gene, which we will designate as alleles *A* and *a*, and also polymorphic for two alleles of an *Esterase-D* gene, which we will designate as alleles *B* and *b*. The gametic types were found in the following proportions. For all practical purposes, these numbers also refer to homozygous genotypes because there is such close inbreeding.

AB	1501	(1642.6)
Ab	754	(613.7)
aB	720	(577.1)
ab	74	(215.6)

(The numbers in parentheses are the expected numbers based on the assumption of linkage equilibrium, calculated as in Chapter 2.) The χ^2 value in this case is 172.7 with one degree of freedom. The associated probability is much less than 0.0001, and so there is undoubtedly linkage disequilibrium. For the above data, the linkage disequilibrium parameter (see Equation 2.13)

is $D = -0.046$, which is about 66% of its theoretical minimum. On the other hand, in spite of the small amount of outcrossing in natural populations of barley, the DNA sequences of most genes show evidence for recombination (Morrell et al. 2003).

One of the dramatic successes of plant breeding has come from the crossing of inbred lines to produce high-yielding hybrid corn. Yield of a genetically heterogeneous, outcrossing variety of corn can be improved by selecting the plants with the highest yields in each generation to be the progenitors of the next generation; such artificial selection results in only gradual improvement, however (see Chapter 9). If a large number of self-fertilized lines are established from a heterogeneous population, each line declines in yield as inbreeding proceeds, because of the forced homozygosity of deleterious recessives. Many lines become so inferior that they have to be discontinued. Self-fertilized lines are not likely to become homozygous for exactly the same set of deleterious recessives, however, and when different lines are crossed to produce a hybrid, the hybrid becomes heterozygous for these genes. Alleles favoring high yield in corn are generally dominant, and there may also be genes in which the heterozygous genotypes have a more favorable effect on yield than do the homozygous genotypes; in any case, the hybrid has a much higher yield than either inbred parent. The phenomenon of enhanced hybrid performance is called **hybrid vigor** or **heterosis**. In practice, inbred lines are crossed in many combinations to identify those that produce the best hybrids. Yields of hybrid corn are typically 15–35% greater than yields of outcrossing varieties, and the successful introduction of hybrid corn has been remarkable. Virtually all corn acreage in the United States today is planted with hybrids, as compared to 0.4% of the acreage in 1933 (Sprague 1978).

6.2 POPULATION SUBDIVISION

Most populations are grouped into smaller subpopulations within which mating usually takes place. Such grouping is called **population structure** or **population subdivision**, and it is almost universal among organisms. Many organisms naturally form subpopulations in the form of herds, flocks, schools, colonies, or other types of aggregations. When there is population subdivision, there is almost inevitably some genetic differentiation among the subpopulations. By **genetic differentiation** we mean that the allele frequencies among the subpopulations become different. Genetic differentiation may result from natural selection favoring different genotypes in different subpopulations, but it may also result from random processes in the transmission of alleles from one generation to the next or from chance differences in allele frequency among the initial founders of the subpopulations. The effects of random genetic drift in increasing the variance in allele frequency among subpopulations have already been examined in Chapter 3.

When the subpopulations are completely isolated from migration, then all matings must take place between individuals within each subpopulation. The intra-population mating implies that the individuals within each subpopulation will share some common ancestors, and hence even matings that take place "at random" in the subpopulation are matings that unite individuals who have common ancestors. These common ancestors transmit alleles that are identical by descent that can come together in the progeny of the mating, and a nonzero probability of identity by descent constitutes inbreeding. In other words, population subdivision, in and of itself, results in inbreeding because the individuals in the subpopulation share remote ancestors, even in situations in which the members of each subpopulation choose their mates at random. The relationship between population structure and inbreeding is subtle, but it has profound consequences in population genetics.

Many populations have a **hierarchical population structure**, which means that the subpopulations can be grouped into progressively inclusive levels in which, at each grouping, the next lower levels are included ("nested") within the next higher ones. To consider a concrete example, imagine we were interested in the population structure of a widespread species of freshwater fish. The lowest population level consists of a local interbreeding population of animals within a stream. A stream might contain more than one such local population. The next-higher level in the hierarchy could be the organization of streams into groups feeding the same river. Another higher level could be rivers within watersheds. An even higher level of organization might be watersheds within continents. The aggregation of subpopulations into progressively more inclusive groups can continue for as many levels as is convenient and informative. It is inevitably somewhat arbitrary how the groups at each level are combined to form the next higher level in the hierarchy. The choice of classification is pragmatic: One tries to group the subpopulations in such a way as to highlight the genetic similarities and differences among them. If there were so much migration of fish among subpopulations that all members of the species constituted essentially a single, random-mating population, then there would be no need to define a hierarchical population structure because it would be uninformative. However, most organisms do have significant population structure.

Reduction in Heterozygosity Due to Population Subdivision

One of the important consequences of population structure is a reduction in the average proportion of heterozygous genotypes relative to that expected under random mating. The reason for the reduction in heterozygosity may be understood by considering the somewhat whimsical example in Figure 6.12. The outline is the floor plan of a large barn. The organisms of interest are the mice concentrated primarily into two subpopulations of equal size at the west and east ends of the barn. The movement of mice between the sub-

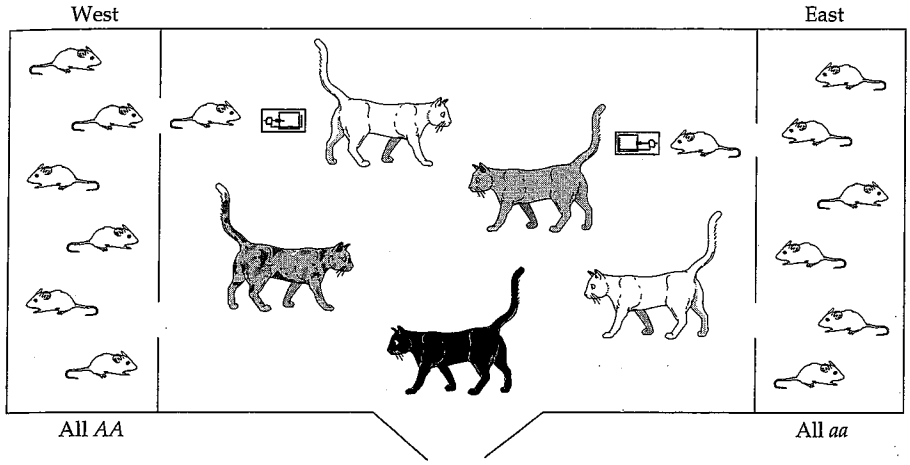


FIGURE 6.12 An extreme example of the general principle that a difference in allele frequency among subpopulations results in a deficiency of heterozygotes. The floor plan is that of a hypothetical barn. The mouse subpopulations in the east and west enclaves are completely isolated because of the cats in the middle. The west subpopulation is fixed for the *A* allele and the east subpopulation for the *a* allele. Trapping mice at random in the area patrolled by the cats would yield an overall allele frequency of $\frac{1}{2}$, but no heterozygous genotypes.

populations is prevented by a large population of hungry and vigilant cats in the central area. The occasional mouse that comes out of its refuge is quickly eaten. (These hypothetical mice have not been endowed with the ingenuity to find alternative routes between the west and east ends of the barn, like sneaking along the rafters.) Because of chance effects in the founding of the subpopulations, the west and east subpopulations are completely homozygous for alternative alleles of a gene. All the mice in the west subpopulation are *AA*, and all those in the east subpopulation are *aa*. In technical terms, the west subpopulation is fixed for the *A* allele (its allele frequency equals 1), and the east subpopulation is fixed for the *a* allele. The genotype frequencies of *AA*, *Aa*, and *aa* in the west subpopulation are 1, 0, and 0, respectively, and those in the east subpopulation are 0, 0, and 1, respectively. Within each subpopulation there is random mating, and the genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. In particular, the frequencies of *AA*, *Aa*, and *aa* within each subpopulation are given by p^2, $2pq$, and q^2, where $p = 0$ in the east subpopulation, and $p = 1$ in the west subpopulation. Therefore, within any one of the subpopulations in Figure 6.12, the frequency of heterozygotes equals the frequency expected with HWE.

The situation regarding the total mouse population in Figure 6.12 is very different, however, as there is an overall deficiency of heterozygotes. By "total population" in this context, we mean the aggregate of all mice without regard to the population subdivision. Suppose we were unaware of the population structure in the barn. We might then suppose that the barn contained a single randomly mating population. To study the total population of the barn, we trap mice at random in the center area, catching the occasional escapee from the cats. Because the subpopulations are fixed for either A or a, half the time we would trap an AA homozygote and half the time an aa homozygote. Consequently, we estimate the allele frequency of A as $\hat{p} = \frac{1}{2}$. Assuming random mating and Hardy-Weinberg genotype frequencies in the total population, the expected genotype frequencies of AA, Aa, and aa are given by the HWE as $\hat{p}^2, 2\hat{p}\hat{q}$, and $\hat{q}^2$. Because the overall allele frequency of A among the trapped animals is $\frac{1}{2}$, we would naively expect a fraction $2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$ of the animals to be heterozygous. In fact, we would have caught no heterozygotes at all!

This rather paradoxical result—that there is a deficiency of heterozygotes in the total population even though random mating takes place within each subpopulation—is a consequence of the difference in allele frequency among the subpopulations. Were the allele frequencies in both subpopulations the same, it would not matter whether we sampled from the west subpopulation, the east subpopulation, or from the area in between. We would recover genotypes in Hardy-Weinberg proportions because both subpopulations are genotypically identical and in HWE. In an organism with hierarchically structured subpopulations, there is an analogous deficiency of heterozygotes at each level in the hierarchy. The following section examines the heterozygosities in more detail.

Average Heterozygosity

In the Mojave desert, local populations of the annual plant *Linanthus parryae* are polymorphic for white versus blue flowers. The plant is diminutive, averaging just 1 cm in height, and when the plant is in bloom, the ground cover of white flowers justifies the popular name "desert snow." Blue flowers result from homozygosity for a recessive allele. The geographical distribution of the frequency q of the recessive allele across a region of the Mojave desert is illustrated in Figure 6.13. Each allele frequency is based on an examination of approximately 4000 plants over an area of about 30 square miles (Epling and Dobzhansky 1942).

Judging from the allele-frequency map in Figure 6.13, the highest frequencies of the blue-flower allele are largely concentrated at the west and east ends of the region in question. The unequal allele frequencies across the range imply a decrease in average heterozygosity relative to HWE, analogous to the mouse example in Figure 6.12, though not as extreme. Figure 6.13 shows the estimated allele frequency in each of 30 subpopulations. Suppose

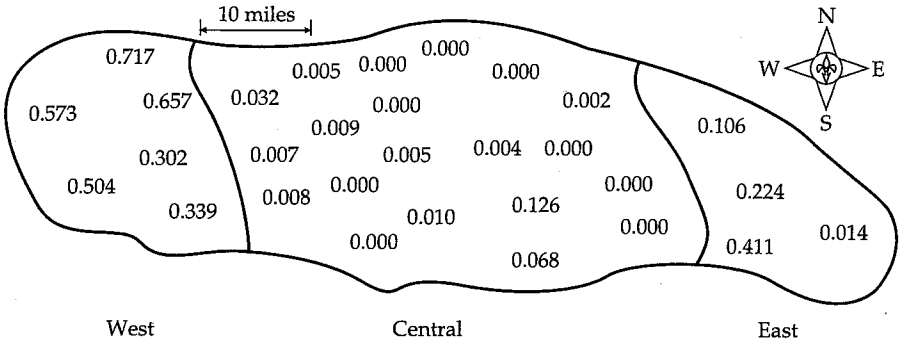


FIGURE 6.13 Estimated frequency of a recessive allele for blue flower color in populations of *Linanthus parryae* in an area of approximately 900 square miles in the Mohave desert. Each allele frequency is based on an examination of approximately 4000 plants over an area of about 30 square miles. (After Wright 1943a.)

each of the subpopulations is regarded as a random-mating unit in HWE for the flower-color alleles. The average heterozygosity among the subpopulations can be denoted as H_S, where the subscript indicates subpopulation. The calculations are shown in the third column in Table 6.3; the heterozygosity in each subpopulation is calculated as $2pq$, where p and q are the estimated frequencies of the alleles for white versus blue flower color, respectively, in each subpopulation. The H_S tabulated at the bottom is the average of all the subpopulation heterozygosities (counting the value 0.000 a total of nine times because of the nine different subpopulations in which $q = 0.000$).

A second hierarchical level of population structure is that of region—west (W), central (C), or east (E). To calculate the heterozygosity expected from HWE in each region, we first estimate the average allele frequency in the region by taking the mean allele frequency across all subpopulations in the region. For example, the average allele frequency q in region E is $(0.106 + 0.224 + 0.411 + 0.014)/4 = 0.1888$. In each region, the heterozygosity expected from HWE is calculated as $2pq$, where p and q are the average allele frequencies in the region. In region E, therefore, the regional heterozygosity equals $2 \times (1 - 0.1888) \times 0.1888 = 0.3062$. The average heterozygosity within regions at the bottom of column 5 is denoted H_R; it is the weighted average of the regional heterozygosities, where each regional heterozygosity is weighted by the number of subpopulations in the region. In this example, $H_R = (6 \times 0.4995 + 20 \times 0.0272 + 4 \times 0.3062)/30 = 0.1589$.

Yet another hierarchical level of population structure in Figure 6.13 is the total population—the aggregate population obtained by conceptually uniting all subpopulations to form a single random mating unit. The average

TABLE 6.3 Hierarchical Structure of *Linanthus parryae*

	Subpopulations		Regions		Total	
Region	Allele frequency	Heterozygosity	Average allele frequency	Heterozygosity	Average allele frequency	Heterozygosity
W	0.573	0.4893				
	0.717	0.4058				
	0.504	0.5000				
	0.657	0.4507				
	0.302	0.4216				
	0.339	0.4482	0.5153	0.4995		
C	9 × 0.000	0.0000				
	0.032	0.0620				
	0.007	0.0139				
	0.008	0.0159				
	0.005	0.0100				
	0.009	0.0178				
	0.005	0.0100				
	0.010	0.0198				
	0.068	0.1268				
	0.002	0.0040				
	0.004	0.0080				
	0.126	0.2202	0.0138	0.0272		
E	0.106	0.1895				
	0.224	0.3476				
	0.411	0.4842				
	0.014	0.0276	0.1888	0.3062	0.1374	0.2371
Average heterozygosity		$H_S = 0.1424$		$H_R = 0.1589$		$H_T = 0.2371$

Source: Data from Wright 1943a.

allele frequency is the mean allele frequency across all subpopulations, and $q = 0.1374$. Then H_T is calculated as $2pq = 2 \times 0.8626 \times 0.1374 = 0.2371$.

To sum up:

- H_S is the average heterozygosity assuming HWE among organisms within random-mating subpopulations.
- H_R is the average heterozygosity assuming HWE among organisms within regions.
- H_T is the average heterozygosity assuming HWE among organisms within the total area.

The concepts of hierarchical population structure and the various levels of heterozygosity were originally developed by Wright (1943a,b), in his theory of **isolation by distance**, to quantify genetic differences among subgroups at the various levels. The motivation for developing such a method was summarized in the following passage from Wright (1943b):

> Study of statistical differences among local populations is an important line of attack on the evolutionary problem. While such differences can only rarely represent first steps toward speciation in the sense of the splitting of the species, they are important for the evolution of the species as a whole. They provide a possible basis for intergroup selection of genetic systems, a process that provides a more effective mechanism for adaptive advance of the species as a whole than does the mass selection which is all that can occur under panmixia.

Furthermore, the reduction in heterozygosity resulting from population subdivision is intimately related to the reduction in heterozygosity caused by inbreeding due to mating between relatives. As explained earlier, the relation of population structure to inbreeding can be understood by interpreting each subpopulation as a sort of "extended family" or set of interconnected pedigrees. Organisms in the same subpopulation will often share one or more recent or remote common ancestors, and so a mating between organisms in the same subpopulation may result in offspring whose alleles at a locus are identical by descent (autozygous). The larger the subpopulation and the more recently it has been isolated, the smaller the probability of autozygosity, but in any finite subpopulation the probability of autozygosity increases through time.

Wright's F *Statistics*

To quantify the inbreeding effect of population subdivision, Wright (1921) defined what has come to be called the **fixation index**. This index equals the reduction in heterozygosity expected with random mating at any one level of a population hierarchy relative to another, more inclusive level of the hierarchy. The fixation index is a useful index of genetic differentiation because it allows an objective comparison of the overall effect of population structure among different organisms without getting into details of allele frequencies, observed levels of heterozygosity, and so forth. The genetic symbol for a fixation index is F embellished with subscripts denoting the levels of the hierarchy being compared. For example, F_{SR} is the fixation index of the subpopulations relative to the regional aggregates:

$$F_{SR} = \frac{H_R - H_S}{H_R} \tag{6.9}$$

In words, Equation 6.9 defines F_{SR} as the decrease of heterozygosity among subpopulations within regions $(H_R - H_S)$, relative to the heterozygos-

ity among regions (H_R). For the *Linanthus* example in Table 6.3, F_{SR} = (0.1589 − 0.1424)/0.1589 = 0.1036.

At the next level of the hierarchy, we may define the fixation index F_{RT} as the proportionate reduction in heterozygosity of the regional aggregates relative to the total combined population:

$$F_{RT} = \frac{H_T - H_R}{H_T} \tag{6.10}$$

The data in Table 6.3 indicate that F_{RT} = (0.2371 − 0.1589)/0.2371 = 0.3299. Comparison of this value with F_{SR} above already makes it clear that there is substantially more variation among regions (as measured by F_{RT}) than there is among subpopulations within regions (as measured by F_{SR}). The comparison of the fixation indices at the two levels gives quantitative expression to the regional differences apparent in Figure 6.13.

The fixation index F_{ST} compares the least inclusive to the most inclusive levels of the population hierarchy and measures all effects of population structure combined:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{6.11}$$

From Table 6.3, F_{ST} = (0.2371 − 0.1424)/0.2371 = 0.3993. The overall reduction in average heterozygosity is therefore close to 40% of the total heterozygosity—a very substantial effect.

The **hierarchical *F*-statistics** defined in Equations 6.9 through 6.11 are all types of fixation indices, but they differ in the reference populations: F_{SR} is concerned with subpopulations (S) relative to the regional aggregates (R), F_{RT} is concerned with the regional groupings relative to the total population (T), and F_{ST} is concerned with the subpopulations relative to the total population. The index F_{ST} is the most inclusive measure of population subdivision.

The mathematical relation between the three types of *F* statistics is demonstrated in the following problem.

PROBLEM 6.3 Show that F_{SR}, F_{RT}, and F_{ST} are related by the equation

$$1 - F_{ST} = (1 - F_{SR})(1 - F_{RT})$$

ANSWER From Equation 6.9, $F_{SR} = 1 - (H_S/H_R)$, or $1 - F_{SR} = H_S/H_R$. Equation 6.10 implies that $F_{RT} = 1 - (H_R/H_T)$, or $1 - F_{RT} = H_R/H_T$. Finally, Equation 6.11 implies that $F_{ST} = 1 - (H_S/H_T)$, or $1 - F_{ST} = H_S/H_T$. Now multiply the expressions for $1 - F_{SR}$ and $1 - F_{RT}$ together to obtain $(1 - F_{SR}) \times (1 - F_{RT}) = (H_S/H_R) \times (H_R/H_T) = H_S/H_T = (1 - F_{ST})$.

For examining the overall level of genetic divergence among subpopulations, F_{ST} is the informative statistic, and the concept has been extended to multiple alleles (Nei 1973). Although F_{ST} has a theoretical minimum of 0 (indicating no genetic divergence) and a theoretical maximum of 1 (indicating fixation for alternative alleles in different subpopulations), the observed maximum is usually much less than 1. Wright (1978) has suggested the following qualitative guidelines for the interpretation of F_{ST}:

- The range 0 to 0.05 may be considered as indicating *little* genetic differentiation.
- The range 0.05 to 0.15 indicates *moderate* genetic differentiation.
- The range 0.15 to 0.25 indicates *great* genetic differentiation.
- Values of F_{ST} above 0.25 indicate *very great* genetic differentiation.

On the other hand, Wright also notes that, among subpopulations, "differentiation is by no means negligible if F_{ST} is as small as 0.05 or even less." Difficulties in interpreting F_{ST} are alleviated somewhat by the use of a standardized version in which F_{ST} is expressed as the proportion of the maximum differentiation possible for the observed level of subpopulation homozygosity (Hedrick 2005).

PROBLEM 6.4 One of the limitations of F_{ST} is that it is does not capture the full range of possibilities that can be found in natural populations. To see this for yourself, consider two subpopulations with two alleles each, A_1 and A_2; in one subpopulation the allele frequencies are $(3 + \sqrt{3})/6 = 0.788675$ and $(3 - \sqrt{3})/6 = 0.211325$, and in the other subpopulation the allele frequencies are reversed. (The choice of these allele frequencies may seem strange, but the rationale for the choice will become clear when you work the problem.) Now consider the same gene in two different subpopulations; one of these subpopulations has alleles A_1 and A_2 at frequencies $\frac{1}{2}$ and $\frac{1}{2}$, and the other has alleles A_3 and A_4 at frequencies $\frac{1}{2}$ and $\frac{1}{2}$. Use Equation 6.11 to calculate F_{ST} for both pairs of subpopulations, and explain why the result seems paradoxical.

ANSWER In the first case, the heterozygosity in each subpopulation is $2 \times (3 + \sqrt{3})/6 \times (3 - \sqrt{3})/6 = \frac{1}{3}$, and hence the average subpopulation heterozygosity is $H_S = \frac{1}{3}$. The average allele frequency for each allele is $\frac{1}{2}$, and hence the total heterozygosity is $H_T = \frac{1}{2}$. In this case, $F_{ST} = [(\frac{1}{2}) - (\frac{1}{3})]/(\frac{1}{2}) = \frac{1}{3}$. In the second case, the heterozygosity in each subpopulation is $2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$, and so $H_S = \frac{1}{2}$. The average allele frequencies are $\frac{1}{4}$ for each of the four alleles, and so $H_T = 1 - (\frac{1}{4})^2 = \frac{3}{4}$. In this case, $F_{ST} = [(\frac{3}{4}) - (\frac{1}{2})]/(\frac{3}{4}) = \frac{1}{3}$, exactly the same as before. The paradox is that the subpopulations have the same value of F_{ST}, when the first two subpopulations differ only in allele frequencies, whereas the second two are so different that they have no alleles in common.

PROBLEM 6.5 Some subpopulations of *Drosophila melanogaster* show an altitudinal gradient in the allozymes of alcohol dehydrogenase in which the frequency of the *Adh-F* allele increases with altitude. The data in the accompanying table are estimates of the allele frequency of *Adh-F* in seven samples of adult flies captured either in the mountains, in the foothills, or on the plains of the Caucasus Mountains of the former Soviet Union. Each allele frequency is based on electrophoresis of approximately 300 adult flies (Grossman et al. 1970). Calculate the F statistics F_{SE} (subpopulations within elevations), F_{ET} (elevations within the total), and F_{ST} (subpopulations relative to the total). What do the magnitudes of the F statistics suggest regarding genetic differentiation among subpopulations in the frequency of *Adh-F* with respect to altitude?

Elevation	Allele frequency	Elevation	Allele frequency	Elevation	Allele frequency
Mountain	0.321	Foothill	0.131	Plain	0.082
Mountain	0.226	Foothill	0.109	Plain	0.088
				Plain	0.035

ANSWER Let p represent the allele frequency of *Adh-F*. For each subpopulation, the HWE heterozygosity equals $2p(1 - p)$, which for the seven samples are 0.4359 and 0.3498 (mountain), 0.2277 and 0.1942 (foothill), and 0.1506, 0.1605, and 0.0676 (plain). The average of these values is H_S, which equals 0.2266. At each of the elevations, the average allele frequency is the mean across the subpopulations sampled at that elevation. For mountain, foothill, and plain, these means equal 0.274, 0.120, and 0.068, respectively, yielding the elevation HWE heterozygosities 0.3974, 0.2112, and 0.1273, respectively. (Results may differ slightly according to the number of significant digits carried along.) The average of the elevation heterozygosities equals the mean elevation heterozygosity (H_E), and it is the weighted average $(2 \times 0.3974 + 2 \times 0.2112 + 3 \times 0.1273)/7 = 0.2285$. Finally, the allele frequency for the total heterozygosity is equal to the mean allele frequency across subpopulations, which is 0.142, yielding a total HWE heterozygosity (H_T) of 0.2433. The F statistics are $F_{SE} = (H_E - H_S)/H_E = 0.0081$, $F_{ET} = (H_T - H_E)/H_T = 0.0609$, and $F_{ST} = (H_T - H_S)/H_T = 0.0684$. [As a check, note that $(1 - F_{SE}) \times (1 - F_{ET}) = 1 - F_{ST}$.] Judging from the magnitudes of the F statistics, we can see that most of the differentiation among subpopulations is correlated with altitude; there is very little genetic differentiation among subpopulations at each elevation.

The method of estimating the F statistics by replacing the parameters in Equations 6.9 through 6.11 with their observed or estimated values is not necessarily the best, particularly with small samples. Ideally, estimates of the F statistics should correct for the effects of sampling a limited number of subpopulations as well as for the effects of sampling a limited number of organisms in each subpopulation. Methods for making these corrections have been suggested but are quite complex and raise additional issues. For an excellent discussion, see Weir and Cockerham (1984) and Weir (1996). Important issues

are also addressed in Wright (1978, pp. 86–89), Curie-Cohen (1982), Nei and Chesser (1983), and Nei (1986).

Linanthus Revisited: Evidence for Selection Associated with Flower Color

Studies of subpopulation differentiation in *L. parryae* have been carried out for over 60 years, and the history is thoroughly documented in Schemske and Bierzychudek (2001). The primary evolutionary forces at work have been in much dispute. The pioneering study is that of Epling and Dobzhansky (1942), who obtained the data summarized in Figure 6.13 and pointed out that the distribution of allele frequencies resembled what might be expected from Wright's (1931) theory of random genetic drift. Wright himself carried out an independent analysis of the data (Wright 1943a,b). He estimated the effective population size as 14–25 individuals per subpopulation, and concluded that the subpopulation differences did result primarily from random genetic drift.

But Epling was not so sure. He went on to demonstrate that the seeds of *L. parryae* can survive in the soil and germinate for as long as at least seven years (Epling et al. 1960), suggesting much larger effective population sizes than Wright had estimated. His group also examined a set of subpopulations every year from 1944–1958, and found substantial geographical variation in the frequencies of blue and white flowers, but not much variation from year to year (Epling et al. 1960). This finding also argued against random genetic drift. But Wright remained unconvinced, and he was not about to give up. He had made *Linanthus* the observational cornerstone of his theory of isolation by distance (Wright 1943b), and again he carried out an independent analysis of these and other data. Once again he concluded that random genetic drift played a key role at the level of the subpopulation, but conceded that over larger spatial scales there might be some modest selective differences between the color morphs (Wright 1978).

Studies of *L. parryae* in the Mojave desert were taken up again in 1988 by Schemske and Bierzychudek (2001), who found evidence for quite strong selection. Based on studies of three polymorphic populations over 11 years, they observed that in years of harsh weather, when overall seed production is low, plants with blue flowers produce more seeds than plants with white flowers; but in years when the weather is mild and overall seed production is high, plants with white flowers produce more seeds. The differences in relative fitness were sometimes very large, with selection coefficients on the order of 0.60. Figure 6.14 tells the tale. It shows the ratio of the average number of flowers on blue-flowered plants to the average number on white-colored plants, as a function of the average number of flowers on both types of plants in any given year. The y-axis is an index of relative fitness, because the number of seeds per flower is nearly the same for the blue and white morphs (Schemske and Bierzychudek 2001). The slope of the line is significant, indi-

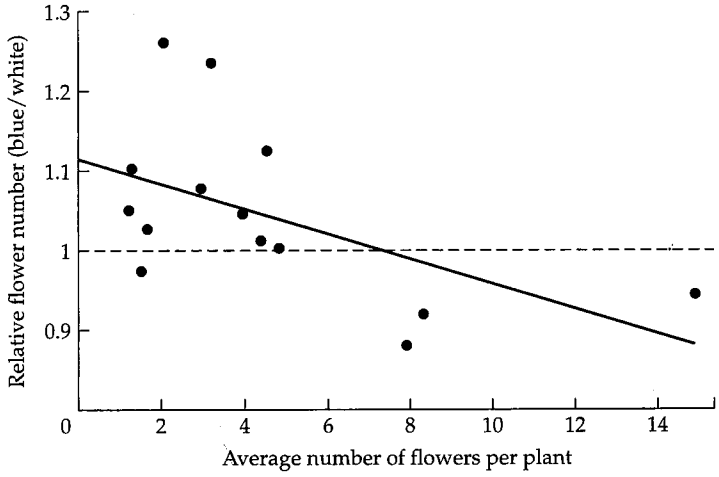


FIGURE 6.14 Ratio of average number of blue flowers per plant relative to the average number of white flowers per plant, as related to the average number of flowers for both types of plants together. The x-axis is an index of environmental quality. Poor conditions in which none of the plants do very well are at the left, and good conditions in which all plants do quite well are at the right. The y-axis is an index of relative fitness. Plants with blue flowers have a greater fitness than those with white flowers under poor conditions, but a smaller fitness than those with white flowers under good conditions. (Data from Schemske and Bierzychudek 2001.)

cating that, relative to plants with white flowers, plants with blue flowers have more numerous flowers per plant in bad years and fewer flowers per plant in good years.

Inference of Population Structure from Multilocus Genotype Data

Despite some limitations, the fixation index F_{ST} defined in Equation 6.11 has served as a convenient and widely used measure of genetic differences among subpopulations. The identification of the causes underlying a particular value of F_{ST} observed in a natural population is often difficult. Allele frequencies among subpopulations can become different because of random processes (random genetic drift) as well as by natural selection with complications from migration among the subpopulations. Difficulties in the assignment of cause do not, however, compromise the usefulness of F_{ST} as an index of genetic differentiation.

The levels of genetic divergence among human subpopulations and among subpopulations of several other species are presented in Table 6.4. The values of F_{ST} imply that genetic divergence between human subpopula-

TABLE 6.4 Total heterozygosity (H_T), average heterozygosity among subpopulations (H_S), and fixation index (F_{ST}) for various organisms

Organism	Number of populations	Number of loci	H_T	H_S	F_{ST}
Human (Africa, Europe, East Asia)	3	35	0.130	0.121	0.069
Human, Yanomama Indian villages	37	15	0.039	0.036	0.077
House mouse *(Mus musculus)*	4	40	0.097	0.086	0.113
Jumping rodent *(Dipodomys ordii)*	9	18	0.037	0.012	0.676
Drosophila equinoxialis	5	27	0.201	0.179	0.109
Horseshoe crab *(Limulus)*	4	25	0.066	0.061	0.076
Lycopod plant *(Lycopodium lucidulum)*	4	13	0.071	0.051	0.282

Source: Protein electrophoretic data from Nei 1975.

tions is quite small. Of the total genetic variation found in samples from three major geographical regions (Africa, Europe, and East Asia), only 7% (0.07) is ascribable to genetic differences among them. In other words, about 93% of the total genetic variation is found among individuals within any single geographical region. Similarly, of the total genetic variation found in the native Yanomama Indians of Venezuela and Brazil, only 7.7% (0.077) is due to differences in allele frequency among villages, which implies that 92.3% of the total genetic variation is found within any single village. Values of F_{ST} for other organisms are quite variable, presumably because F_{ST} is influenced by the size of the subpopulations—which is a major determinant of the magnitude of random changes in allele frequency—subpopulation size is in turn influenced by the amount and pattern of migration between subpopulations and by other factors, including natural selection.

The human data in Table 6.4 are based on protein polymorphisms, but the conclusions have held up remarkably well in studies of many more individuals with hundreds of genetic markers assayed using modern genotyping techniques. For example, Rosenberg et al. (2002) studied 377 microsatellite polymorphisms among 1056 individuals from 52 populations. They used a computer algorithm to group individuals into genetic clusters according to estimated shared ancestry among their genomes (Pritchard et al. 2000a,b; Rosenberg et al. 2005). They found that the individuals could be grouped

into six genetic clusters, five of which correspond to subpopulations within major geographical regions, namely, Africa, Europe, East Asia, Oceania, and America. Genetic differences among individuals within any one cluster accounted for 93–95% of the total genetic variation, with only 3–5% of the genetic variation ascribable to differences between the major clusters. Similar results were also obtained in a subsequent analysis of 993 microsatellite and insertion/deletion polymorphisms in a sample of 1048 individuals (Rosenberg et al. 2005).

On the other hand, the clustering algorithm requires the number of clusters to be specified in advance, although the effects of cluster number can be examined in different computer runs. An alternative method of analysis uses Markov chain Monte Carlo methods to implement a Bayesian analysis of a hierarchical population structure in which the number of genetic groups is not specified in advance (Corander et al. 2004). Application of this method to the data of Rosenberg et al. (2002) confirmed the major findings, but suggested that additional groups were needed to capture all of the genetic differences in the sample, especially in the Americas (Corander et al. 2004).

Although the genotypes of people can be clustered into major groups coinciding with large geographical regions, genetic differences among these groups are small and subtle. As we have noted, 93–95% of the total genetic variation occurs between individuals within any group, and only 3–5% between groups. In other words, the genetic differences between two randomly chosen individuals from different groups are only slightly greater than those between two unrelated individuals from the same group. Furthermore, among 4199 alleles represented more than once in the sample of Rosenberg et al. (2002), 46.7% of the alleles appeared in all major geographical regions, whereas only 7.4% were specific to a particular region.

6.3 THE WAHLUND PRINCIPLE

The flip side of the coin of heterozygosity is homozygosity, because a gene in a diploid organism that is not heterozygous must be homozygous. Mathematically, *homozygosity* = 1 – *heterozygosity*. Therefore, a corollary of the deficit in average heterozygosity that results from population subdivision is that there is an equal excess in average homozygosity. If the population subdivision were eliminated and the former subpopulations allowed to undergo random mating, the average homozygosity would decrease and the average heterozygosity increase by an equal amount. The phenomenon that the average homozygosity decreases when subpopulations join together is called **isolate breaking** or **the Wahlund principle**, after the Swedish statistician and human geneticist Sten Gösta William Wahlund (1901–1976) who first described the effect (Wahlund 1928).

The subpopulations of hypothetical mice in Figure 6.12 afford an illustration of the Wahlund principle. As long as the cats keep the subpopulations

separate, the homozygosity equals 1 because the west subpopulation is genotypically *AA* and the east subpopulation is genotypically *aa*. If the cats were to disappear and the subpopulations of mice came together and practiced random mating, the genotype frequencies would be $\frac{1}{4}$ *AA*, $\frac{1}{2}$ *Aa*, and $\frac{1}{4}$ *aa*. The homozygosity in the fused population is $(\frac{1}{4}) + (\frac{1}{4}) = \frac{1}{2}$, which is a substantial decrease over the average in the subpopulation prior to fusion and random mating. Not only is the total homozygosity reduced by population fusion, so is the average frequency of each homozygous genotype. Consider *aa*, for example. Prior to fusion, the average frequency of *aa* across both subpopulation equals $\frac{1}{2}$; after fusion and random mating, the frequency of *aa* equals $\frac{1}{4}$.

In human population genetics, the Wahlund principle is usually cited for its implication that fusion of subpopulations results in a decrease in the average frequency of children born with a genetic disease resulting from homozygosity for a rare recessive allele, particularly an allele with a relatively high frequency in one of the subpopulations. Examples of harmful recessive alleles at high frequency in some human subpopulations include the alleles for α_1-antitrypsin deficiency ($q \approx 0.024$) and cystic fibrosis ($q \approx 0.022$) in Europeans, sickle-cell anemia ($q \approx 0.05$ in African Americans, up to $q \approx 0.1$ in some African subpopulations), albinism ($q \approx 0.07$ in the Hopi and some other Southwest Native American subpopulations), and Tay-Sachs disease ($q \approx 0.013$ in Ashkenazi Jews).

The Wahlund principle for a recessive allele in two subpopulations is illustrated in Figure 6.15. On the left are two subpopulations, each undergoing random mating, in which the frequency of the recessive allele and the frequency of homozygous recessive genotypes are indicated. The average frequency of the homozygous recessive genotype across both subpopulations equals $(q_1^2 + q_2^2)/2$. The result of fusion and random mating of the subpopulations is shown on the right. Assuming that the subpopulations are equal in size, the allele frequency in the combined population is $\bar{q} = (q_1 + q_2)/2$, and the frequency of the homozygous recessive genotype equals $\bar{q}^2$. Therefore, subpopulation fusion and random mating reduces the average frequency of homozygous recessives by:

$$\begin{aligned} R_{separate} - R_{fused} &= \frac{q_1^2 + q_2^2}{2} - \bar{q}^2 \\ &= \frac{1}{2}(q_1 - \bar{q})^2 + \frac{1}{2}(q_2 - \bar{q})^2 \\ &= \sigma_q^2 \end{aligned} \tag{6.12}$$

We leave it as an exercise to verify that the expressions in q_1 and q_2 on the first and second lines are equal. The symbol σ_q^2 is the variance in allele frequency among the original subpopulations. Because the variance is always nonnegative, isolate breaking decreases the homozygosity and increases the heterozygosity, unless the allele frequencies are equal to begin with.

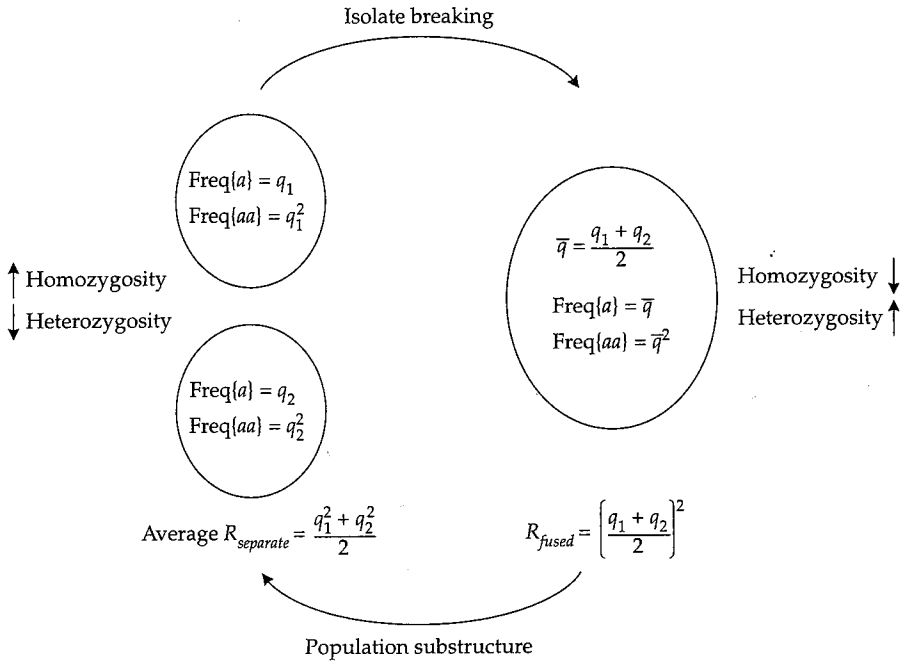


FIGURE 6.15 Illustration of the Wahlund principle. The frequency of homozygous recessives after population fusion and random mating is less than the average frequency before fusion. The difference in frequency of the homozygous recessives equals the variance in allele frequency among the subpopulations.

Furthermore, the result in Equation 6.12 is true for any number of subpopulations of equal or unequal size; in words, the Wahlund principle states that:

> Fusion and random mating of subpopulations decreases the average frequency of homozygous recessives by an amount equal to the variance in allele frequency among the original subpopulations.

To illustrate the effect of isolate breaking, imagine a subpopulation of gray squirrels that has a high frequency of autosomal-recessive albinism equal to 16%. In a nearby forest there is another subpopulation of equal size in which the albino mutation is absent, so that the allele frequency in this subpopulation is 0. Overall, the average frequency of albinos in the two populations is $(0.16 + 0)/2 = 8\%$. If the two subpopulations underwent fusion and random mating, the allele frequency of the albino mutation in the fused population would be $(0.4 + 0)/2 = 0.2$, and the frequency of the homozygous recessive would equal $(0.2)^2 = 4\%$. Note that the frequency of albinos in the

fused population is substantially smaller than the average frequency in the original subpopulations.

Wahlund's Principle and the Fixation Index

Equation 6.12 applies equally well to *AA* homozygotes as to *aa* homozygotes. Therefore, letting P represent the frequency of homozygous *AA* genotypes, we can write

$$P_{separate} - P_{fused} = \sigma_p^2 \tag{6.13}$$

When there are only two alleles, the total reduction in homozygosity must be the summation of Equations 6.12 and 6.13, which equals $\sigma_p^2 + \sigma_q^2$. Because there are only two alleles, it is also true that $\sigma_p^2 = \sigma_q^2$, which we will write as σ^2. Hence, the total reduction in homozygosity from the Wahlund effect upon population fusion and random mating can be expressed as follows:

$$\text{Reduction in total homozygosity} = 2\sigma^2$$

On the other hand, the reduction in total homozygosity with population fusion must also equal the increase in heterozygosity $H_T - H_S$, which Equation 6.11 says is the numerator of F_{ST}. Hence, $F_{ST} = (H_T - H_S)/H_T = 2\sigma^2/H_T$. However, H_T is the heterozygosity with random mating when the allele frequencies equal the average allele frequencies across subpopulations, $\overline{p}$ and $\overline{q}$. Therefore, the connection between the fixation index F_{ST} and the variance in allele frequency is given by

$$F_{ST} = \frac{\sigma^2}{\overline{p}\,\overline{q}} \tag{6.14}$$

Consequently, the F statistics at the various levels of a hierarchical population are related to the variances in allele frequencies among the subpopulations grouped together at the various levels. Equation 6.14 affords a convenient method of estimating F_{ST} from allele-frequency data. For example, among the subpopulations of *Linanthus* in Figure 6.13, the variance in allele frequency is 0.0473. Earlier we calculated the average allele frequencies as $\overline{p} = 0.8626$ and $\overline{q} = 0.1374$. Hence, $\sigma^2/(\overline{p} \times \overline{q}) = 0.3993$, which confirms the previous calculation that $F_{ST} = 0.3993$. (The values as stated may differ slightly from yours because they were calculated with more than four significant digits.)

Genotype Frequencies in Subdivided Populations

In many organisms in which the population structure is hierarchical, it is useful to be able to calculate directly the average genotype frequencies across all subpopulations. Equations 6.12 through 6.14 make it possible to deduce the average genotype frequencies. To do this, first note that R_{fused} in Equation 6.12 equals $\overline{q}^2$ and P_{fused} in Equation 6.13 equals $\overline{p}^2$. Hence Equation 6.12 implies

the average genotype frequency of *aa* among the subpopulations is given by $\sigma_q^2 + \bar{p}\,\bar{q}$, and Equation 6.13 implies the average genotype frequency of *AA* among the subpopulations is given by $\sigma_p^2 + \bar{p}\,\bar{q}$. Since there are only two alleles, $\sigma_q^2 = \sigma_p^2 = \sigma^2$, and Equation 6.14 says that $\sigma^2 = F_{ST} \times \bar{p} \times \bar{q}$. Putting all this together, the average genotype frequency of *AA* across subpopulations must equal $\bar{p}^2 + F_{ST}\bar{p}\,\bar{q}$, and the average genotype frequency of *aa* across subpopulations must equal $\bar{q}^2 + F_{ST}\bar{p}\,\bar{q}$.

Because every genotype that is not homozygous must be heterozygous, the average genotype frequency of heterozygotes across subpopulations is given by $1 - (\bar{p}^2 + F_{ST}\bar{p}\,\bar{q}) - (\bar{q}^2 + F_{ST}\bar{p}\,\bar{q})$. Note that $1 - \bar{p}^2 - \bar{q}^2 = 2\bar{p}\,\bar{q}$, and so the average frequency of heterozygotes simplifies to $2\bar{p}\,\bar{q} - 2\bar{p}\,\bar{q}F_{ST}$.

Consequently, the genotype frequencies averaged across subpopulations in a subdivided population can be written as:

$$\begin{aligned} aa &: \bar{q}^2 + \bar{p}\bar{q}F_{ST} \\ Aa &: 2\bar{p}\bar{q} - 2\bar{p}\bar{q}F_{ST} \\ aa &: \bar{q}^2 + \bar{p}\bar{q}F_{ST} \end{aligned} \tag{6.15}$$

These genotype frequencies depart from the Hardy-Weinberg principle in having an excess of homozygotes and a deficiency of heterozygotes. This result may seem somewhat paradoxical because, within any particular subpopulation, mating is random and the genotype frequencies do obey the Hardy-Weinberg principle. The reason for the departure from Hardy-Weinberg equilibrium in the population as a whole is that the subpopulations differ in allele frequency. Because the allele frequencies differ, random mating within each subpopulation is not equivalent to random mating among all the organisms in the entire population.

From the expressions in Equation 6.15, it is clear that the value of F_{ST} determines the degree of departure from Hardy-Weinberg equilibrium. If $F_{ST} = 0$, the second term in each expression vanishes, and the genotype frequencies reduce to the Hardy-Weinberg frequencies; on the other hand, $F_{ST} = 0$ means that there is no variation in allele frequency among the subpopulations for the gene in question. Because F_{ST} may vary from one gene to the next, other genes in the same subpopulations may have nonzero values of F_{ST}. The extreme case is $F_{ST} = 1$, which happens when two subpopulations are fixed for alternative alleles. In this case, the average allele frequencies are $\frac{1}{2}$ for each allele, and the average genotype frequencies of *AA*, *Aa*, and *aa* across subpopulations are $\frac{1}{2}$, 0, and $\frac{1}{2}$, respectively. This case is illustrated in Figure 6.12.

The genotype frequencies in Equation 6.15 may remind you of those in Equation 3.15, where population subdivision was examined from the stand-

point of random genetic drift. There we noted that F_t meant the value of F_{ST} in generation t, and in Equation 6.15 F_{ST} is also time dependent, although the time dependence is not explicit. The only other difference is that p_0 and q_0 in Equation 3.15 are replaced by $\bar{p}$ and $\bar{q}$ in Equation 6.15. In the special case when the alleles are selectively neutral, then the equations are identical, because if all subpopulations have the same initial frequency p_0, then $E(\bar{p}) = p_0$. On the other hand, if there is selection, then $E(\bar{p})$ will not in general equal p_0, and therein lies an important difference between the equations.

Relation between the Inbreeding Coefficient and the F *Statistics*

So far we have assumed that mating within each subpopulation of a subdivided population is random. What happens when mating between relatives takes place, in addition to the population subdivision? The answer to this question is implicit in Equation 6.4 and Equation 6.11. Equation 6.4 implies that $1 - F_{IS} = H_I/H_S$. where F_{IS} is the reduction in heterozygosity in an inbred individual, relative to the heterozygosity expected with random mating in the subpopulation to which the inbred individual belongs. The value of F_{IS} takes into account only the autozygosity due to immediate inbreeding of the individual, not the accumulated autozygosity due to population structure.

Equation 6.11 implies that $1 - F_{ST} = H_S/H_T$, where F_{ST} is the reduction in heterozygosity in the subpopulation, relative to that expected with random mating in the population as a whole. The value of F_{ST} takes into account only the autozygosity due to population subdivision.

To take both inbreeding and population structure into account, we can define another measure of autozygosity, denoted F_{IT}, which measures the probability of autozygosity of an inbred individual relative to the population as a whole, were all the subpopulations to fuse and undergo random mating. Equation 6.16 asserts that F_{IT} is equal to

$$F_{IT} = \frac{H_T - H_I}{H_T} \tag{6.16}$$

which implies that $1 - F_{IT} = H_I/H_T$. Considering this expression in light of $1 - F_{IS} = H_I/H_S$ and $1 - F_{ST} = H_S/H_T$, it follows that

$$(1 - F_{IS})(1 - F_{ST}) = 1 - F_{IT} \tag{6.17}$$

Hence, if we know both F_{IS} and F_{ST}, then we can obtain F_{IT} from Equation 6.17. The value of F_{ST} measures the autozygosity resulting from random genetic drift in a finite subpopulation, and the value of F_{IS} measures the autozygosity resulting from inbreeding above and beyond that accounted for in F_{ST}. The value of F_{IT} is the probability of autozygosity taking both processes into account.

PROBLEM 6.6 Suppose a large population is subdivided into smaller subpopulations within which mating takes place at random, and that random genetic drift takes place until the probability of autozygosity in the subpopulations is $F_{ST} = \frac{1}{16}$, which is the value expected from the mating of first cousins. Now suppose that a first-cousin mating takes place in one of the subpopulations, so that inbred progeny have an inbreeding coefficient, relative to the subpopulation, of $F_{IS} = \frac{1}{16}$. What is the overall probability of autozygosity in the inbred offspring, F_{IT}?

ANSWER Use Equation 6.17 with the values of F_{ST} and F_{IS} as given. The result is $F_{IT} = 1 - (\frac{15}{16}) \times (\frac{15}{16}) = \frac{31}{256} = 0.121$. Note that this is smaller than $F_{IS} + F_{ST}$ by an amount equal to $F_{IS} \times F_{ST}$, because the two sources of autozygosity are not mutually exclusive.

6.4 ASSORTATIVE MATING

Inbreeding affects all the genes in the organism, but one form of nonrandom mating affects only a subset of genes. In a type of mating known as **assortative mating**, individuals choose their mates according to their phenotype. Assortative mating affects only those genes that influence the phenotype affecting mate choice and genes linked with them in the chromosomes. Most assortative mating is *positive assortative mating*, which means that mating pairs have, on the average, more similar phenotypes than expected with random mating. There are also examples of *negative assortative mating*, sometimes called *disassortative mating*, in which mating pairs are more dissimilar than expected by chance, but we will focus on positive assortative mating.

In human populations, positive assortative mating has been reported for age, height, weight, skin color, facial appearance, IQ score, educational level, personality characteristics, tobacco smoking, alcohol consumption, religious affiliation, nationality, and other traits (Alvarez and Jaffe 2005). Many of these traits, such as religious affiliation and nationality, have no genetic component but are reflections of cultural preference or geographical proximity.

For traits that do have a genetic component, the consequences of positive assortative mating are complex. Hereditary traits associated with assortative mating are rarely determined by the alleles of a single gene. Rather they are multifactorial and often affected by environmental factors as well as by genes. For such traits the consequences of assortative mating depend on the strength of the genetic contribution to variation in the trait, the number of genes that influence the trait, the number of alleles of the genes, the number of different phenotypes, the sex performing the mate selection, and the criteria for mate selection. There is, however, one generally expected result of positive assortative mating. Because similar phenotypes tend to form mating pairs, assortative mating is expected to increase the frequency of homozygous genotypes in the population at the expense of heterozygous genotypes,

and thus there is an increase in the phenotypic variance of the trait in the population. The increase in the phenotypic variance is rather modest unless the assortative mating is very pronounced.

6.5 MIGRATION

In a subdivided population, random genetic drift results in genetic divergence among subpopulations. Subpopulations are rarely completely isolated, however. The process of **migration** refers to the movement of some organisms (or their gametes) among subpopulations. Migration results in **gene flow** between the subpopulations; gene flow acts as a sort of genetic glue that holds the gene pools of subpopulations together and that limits how much genetic divergence can take place. To understand the homogenizing effects of migration, it is useful to study migration in several simple models of population structure.

One-Way Migration

When migration takes place predominantly from one population into another, without an equal amount of migration in the reverse direction, then there is said to be **one-way migration**. An illustration of one-way migration between a large mainland population and a small island subpopulation is shown in Figure 6.16. For simplicity, we consider a gene with two alleles, A and a, with respective frequencies p^* and q^* on the mainland and p and q on the island. Suppose that, in any generation, a proportion m of zygotes in the island subpopulation originates as a random sample of organisms from the mainland. Then, if p and p' are the frequencies of A in the island subpopulation in two successive generations, it follows that

$$p' = (1 - m)p + mp^* \tag{6.18}$$

In Equation 6.18, m is called the **migration rate** between the mainland and the island. Subtracting p^* from both sides of Equation 6.18 and simplifying leads to the expression $p' - p^* = (1 - m)(p - p^*)$, and from this expression

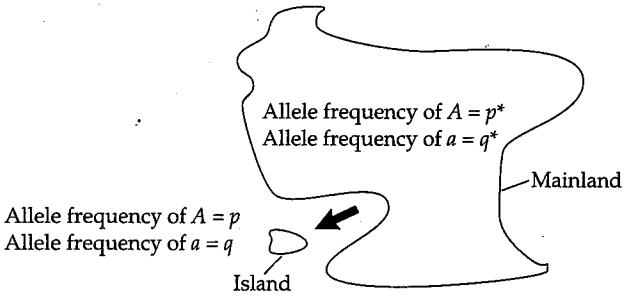


FIGURE 6.16 Model of one-way migration from a large land mass onto an island. The allele frequencies in the source population, p^* and q^*, are assumed to remain constant, whereas those in the recipient population, p_t and q_t, change with time.

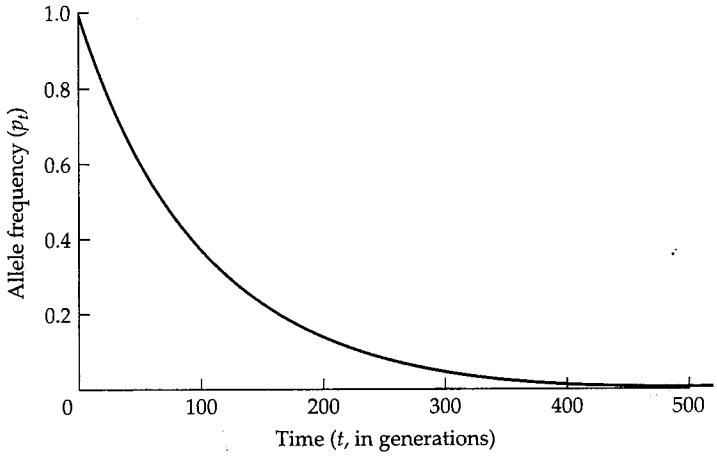


FIGURE 6.17 Change of allele frequency with one-way migration, assuming that an allele A is initially fixed in the recipient population and absent in the source population. The migration rate is $m = 0.01$.

it follows that $p_t - p^* = (1 - m)^t(p_0 - p^*)$, where p_t is the frequency of A in the island subpopulation in generation t. Hence,

$$p_t = p^* + (1 - m)^t(p_0 - p^*) \tag{6.19}$$

Equation 6.19 expresses mathematically what should be clear intuitively: With one-way migration, the allele frequency of A in the island subpopulation gradually approaches that of the mainland population, and the rate of approach is m per generation. As a check on Equation 6.19, note that, when $t = 0$, then $p_t = p_0$, as must be the case, and as t becomes large, p_t goes to p^*.

As an evolutionary process that brings potentially new alleles into a population, migration is qualitatively similar to mutation. The major difference is quantitative: Generally speaking, the rate of migration among subpopulations of a species is vastly greater than the rate of mutation of a gene. The contrast is illustrated in Figure 6.17 for the unrealistic case in which the A allele present in an island subpopulation is absent on the mainland. In this case, Equation 6.19 becomes $p_t = p_0(1 - m)^t$, which has the same form as Equation 4.1 for one-way mutation, except that m replaces μ. The identity in the shape of the curves is apparent, but the time axis in Figure 6.17 is compressed because, when $m = 0.01$, as in this example, compared with the value of $\mu = 0.0001$ in Figure 4.1, it requires only one generation of migration to change the allele frequency to the same extent as 100 generations of mutation.

Equation 6.19 holds more generally for one-way migration by letting p be the frequency of any allele in the population that receives the migrants

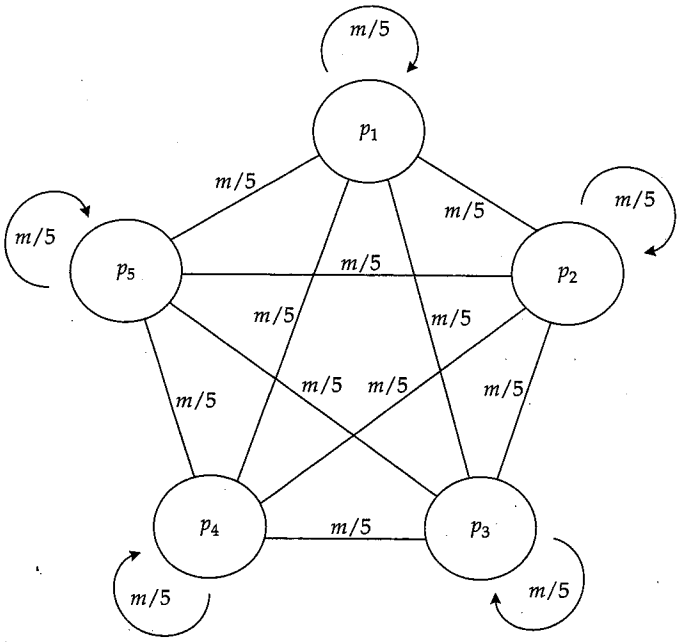


FIGURE 6.18 The island model of migration with five subpopulations. Migration is completely symmetrical. Each subpopulation contributes individuals or gametes to a pool of migrants, which then distribute themselves randomly among the subpopulations. In this model, a migrant can re-enter the same subpopulation it came from, indicated by the loops.

and p^* be the frequency of the same allele in the population that supplies the migrants.

The Island Model of Migration

In the **island model** of migration, a large population is split into many subpopulations dispersed geographically. Examples of island population structure might include fish in freshwater lakes or slugs in dispersed garden plots. Each subpopulation is assumed to be so large that random genetic drift can be neglected. Figure 6.18 shows an example with five island populations, where p_i denotes the allele frequency of A in the ith subpopulation. Each subpopulation receives an equal proportion of migrants from each subpopulation (including itself), and also contributes an equal proportion to each subpopulation (including itself). In essence, the model is one in which each subpopulation sends out a proportion of migrant individuals or gametes, and the migrants from all the subpopulations form a pool, the members of which then disperse randomly among the subpopulations. In this way, an

individual or gamete in the migrant pool can return to the subpopulation it came from originally. Since the total proportion of migrants coming into each subpopulation in every generation is m, the proportion coming from each of the five subpopulations is $m/5$.

Now let $\bar{p}$ be the average allele frequency of A among the subpopulations. Since the migrant individuals or gametes form a pool with an equal contribution from each subpopulation, the expected allele frequency among the migrants must equal the average allele frequency among the subpopulations. The parameter m is the probability that a randomly chosen allele in any subpopulation comes from a migrant. Let us consider a particular subpopulation with an allele frequency of A equal to p_t in generation t. For a randomly chosen allele in this subpopulation in generation t, the allele could have come from the same subpopulation in generation $t-1$ with probability $1-m$, in which case it is an A allele with probability p_{t-1}. Alternatively, the allele could have come from the migrant pool in generation $t-1$ with probability m, in which case it is an A allele with probability $\bar{p}$. Because all evolutionary processes other than migration are ignored, $\bar{p}$ stays the same in all generations. Altogether,

$$p_t = p_{t-1}(1-m) + \bar{p}m \tag{6.20}$$

Equation 6.20 is similar to Equation 4.2 for mutation, and its solution in terms of p_0 is

$$p_t = \bar{p} + (1-m)^t(p_0 - \bar{p}) \tag{6.21}$$

The similarity with Equation 6.19 is apparent: In fact, the equations are identical except that the role of p^* in one-way migration is replaced with $\bar{p}$ in the island model. Perhaps less obvious is the similarity with Equation 4.4 for reversible mutation, in which case $v/(\mu + v)$ plays the role of $\bar{p}$ and $\mu + v$ plays the role of m. The correspondence between the equations again emphasizes the similarity between the effects of migration and those of mutation. The processes result in similar mathematical expressions because both mutation and migration act linearly on allele frequency, which means that p_t is a linear function of p_{t-1}. Although Equation 6.21 for migration is mathematically similar to Equation 4.4 for mutation, the biological implications are quite different. Because rates of migration are typically much greater than rates of mutation, changes in allele frequency are generally much faster with migration.

As an example of the use of Equation 6.21, suppose there are only two populations with initial allele frequencies of A of 0.2 and 0.8, respectively, with $m = 0.10$. Thus 10 percent of the organisms in either subpopulation in any generation are migrants having an allele frequency of A of $\bar{p} = (0.2 + 0.8)/2 = 0.5$. What is the allele frequency of A in the two populations after 10 generations? For the population with initial allele frequency 0.2, we substitute $p_0 = 0.2$, $\bar{p} = 0.5$, and $m = 0.10$ into Equation 6.21 to obtain $p_{10} = 0.5 + (1 -$

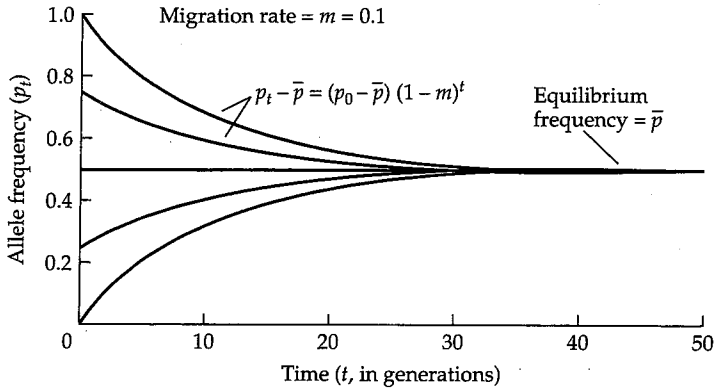


FIGURE 6.19 Change of allele frequency with time in five subpopulations exchanging migrants at the rate $m = 0.1$ per generation. Note the rapid convergence to a common equilibrium frequency.

$0.10)^{10}(0.2 - 0.5) = 0.395$; for the other population, we substitute $p_0 = 0.8$, $\bar{p} = 0.5$, and $m = 0.10$, and so $p_{10} = 0.5 + (1 - 0.10)^{10}(0.8 - 0.5) = 0.605$. Another example using Equation 6.21 is shown in Figure 6.19, where there are five subpopulations (initial frequencies 1, 0.75, 0.50, 0.25, and 0), again with $m = 0.10$. Note how rapidly the allele frequencies converge to the same value, in this case, 0.5.

How Migration Limits Genetic Divergence

It is remarkable how little migration is required to prevent significant genetic divergence among subpopulations as measured by, for example, the fixation index F_{ST}. To understand the homogenizing effect of migration, consider the model in Figure 4.5, in which two alleles drawn at random from a subpopulation in generation $t + 1$ are replicas of the same allele in generation t with probability $1/(2N)$ and replicas of different alleles in generation t with probability $1 - 1/(2N)$. In the first case, the alleles are necessarily identical by descent; in the second case, they are identical by descent with probability F_{t-1}, where F is shorthand for F_{ST}. In either case, the identity by descent is unbroken only if neither allele is replaced by an allele from a migrant, and so

$$F_t = \left(\frac{1}{2N}\right)(1-m)^2 + \left(1 - \frac{1}{2N}\right)(1-m)^2 F_{t-1} \qquad (6.22)$$

Illustrating again the analogy between migration and mutation, Equation 6.22 is identical to Equation 4.8 measuring the effect of mutation on the probability of identity by descent, except that m replaces μ. The equilibrium

value $\hat{F}$ of F can be found by setting $\hat{F} = F_t = F_{t-1}$; after expanding the squared terms on the right-hand side, and assuming that m is small enough and N large enough, and that terms in m^2 and m/N can be ignored, some rearrangement leads to

$$\hat{F} = \frac{1}{1+4Nm} \tag{6.23}$$

As might be expected, Equation 5.17 is identical in form to Equation 4.9 for mutation, but the biological implications are very different owing to the fact that the rate of migration is typically much greater than the rate of mutation.

The product Nm in Equation 6.23 has a straightforward biological interpretation. The total number of alleles in a subpopulation of size N diploid organisms is $2N$. In any generation, the proportion of alleles that are replaced by alleles from migrant organisms is m; hence the number of migrant alleles in any generation equals $2Nm$. However, $2Nm$ is also the total number of alleles in Nm diploid organisms, and so Nm can be interpreted as the absolute number of migrant organisms that come into each subpopulation in each generation.

Because the absolute number of migrants per generation equals Nm, Equation 6.23 implies that $\hat{F}$ decreases as the number of migrants increases. Indeed, the decrease in $\hat{F}$ with increasing Nm is extremely rapid, as shown in Figure 6.20. In the extreme case of complete genetic isolation between the subpopulations, $Nm = 0$ and $\hat{F} = 1$. The decrease is then so rapid that for:

- $Nm = 0.25$ (one migrant every fourth generation), $\hat{F} = 0.50$
- $Nm = 0.5$ (one migrant every second generation), $\hat{F} = 0.33$
- $Nm = 1$ (one migrant every generation), $\hat{F} = 0.20$
- $Nm = 2$ (two migrants every generation), $\hat{F} = 0.11$

The implication of Figure 6.20 is that migration is a potent force acting against genetic divergence among subpopulations. The effects are seen dramatically in Figure 6.21. Part A pertains to the moth *Biston betularia*, part B to the moth *Gonodontis bidentata*. Both species have evolved melanic (blackened)

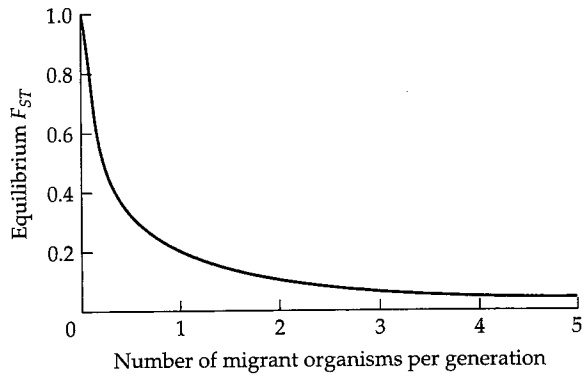


FIGURE 6.20 Decrease in the fixation index F_{ST} among subpopulations at equilibrium in the island model of migration. The curve is that in Equation 6.23, giving $\hat{F}$ as a function of Nm. In the island model, Nm is the number of migrant organisms that come into each subpopulation in each generation.

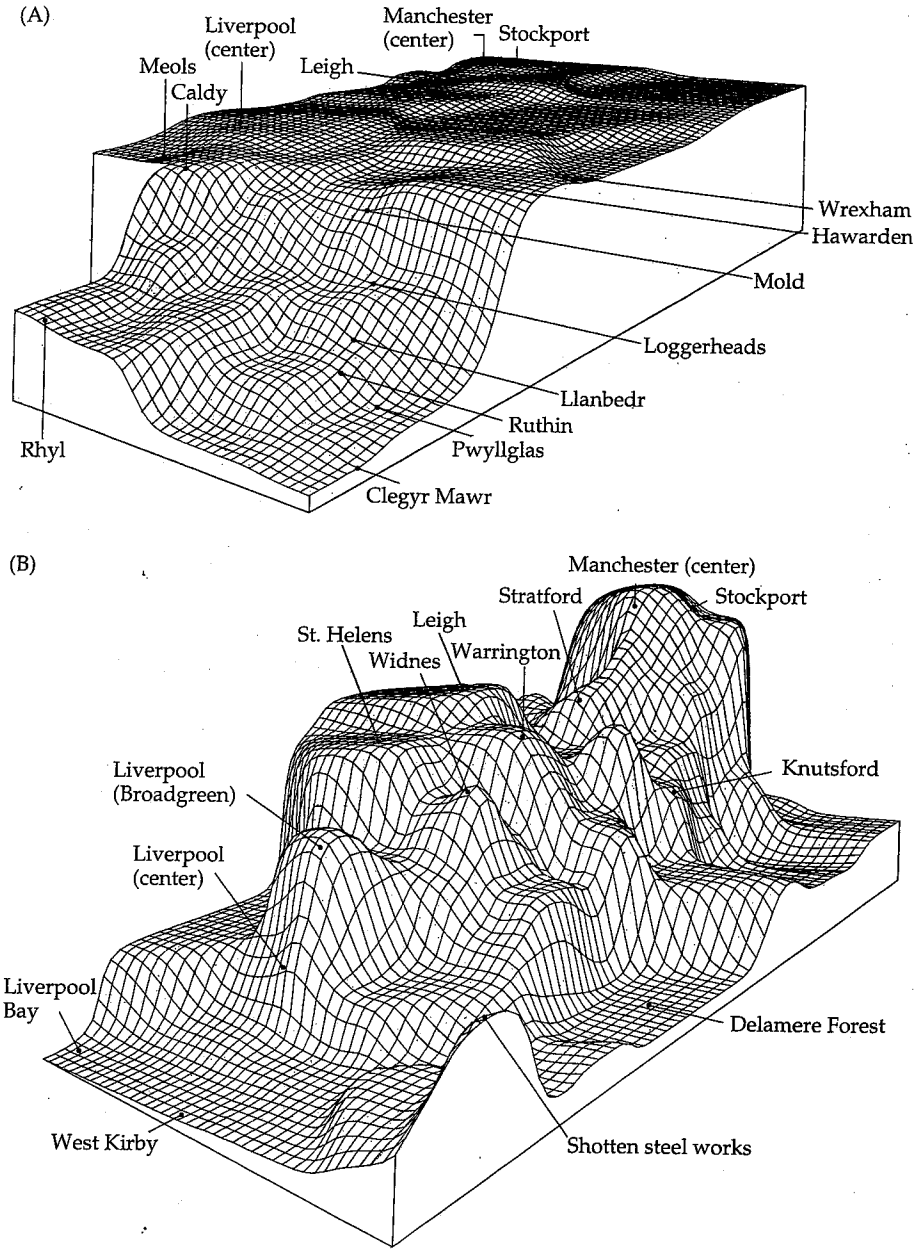


FIGURE 6.21 (A) Distribution of melanic moths of the species *Biston betularia* over an area including Liverpool and Manchester, as viewed from rural Wales. (B) Distribution of melanic moths of the species *Gonodontis bidentata* over a smaller area than that in (A) but viewed from the same perspective. (From Bishop and Cook 1975.)

forms in response to heavy air pollution (*industrial melanism*, see Chapter 2), and the graphs give the frequency of the melanic forms in the two species. The geographical area in A includes Liverpool and Manchester, as viewed from rural Wales. Note the fall-off in frequency of melanics in the nonindustrial areas toward the front of the graph. *Biston betularia* exists in low population densities and must fly relatively long distances to find a mate. The resulting high rate of migration hinders differentiation of populations, hence the smooth surface. In contrast, *Gonodontis bidentata* exists in high population densities and the migration rate is low; hence there is substantial genetic differentiation among populations, as evidenced by the bumpy surface of the graph in part B.

The homogenizing effects of migration should not be overestimated, however. The measure of genetic divergence in Figure 6.20 is F_{ST}, the value of which is determined by the variance in allele frequency among subpopulations (see Equation 6.14) and so is affected primarily by polymorphic alleles that are at intermediate frequencies. Rare alleles present in one subpopulation but absent in others have hardly any effect on F_{ST}. Because rare alleles are rare, they are unlikely to be included among migrant organisms unless the migration rate is very great, and so rare alleles will tend to remain present in only one or a few subpopulations in a local area until such time as their frequency may become

TABLE 6.5 Estimates of *Nm* and F_{ST}

Species	Type of organism	Estimated *Nm*	Estimated F_{ST}
Stephanomeria exigua	Annual plant	1.4	0.152
Mytilus edulis	Mullusc	42.0	0.006
Drosophila willistoni	Insect	9.9	0.025
Drosophila pseudoobscura	Insect	1.0	0.200
Chanos chanos	Fish	4.2	0.056
Hyla regilla	Frog	1.4	0.152
Plethodon ouachitae	Salamander	2.1	0.106
Plethodon cinereus	Salamander	0.22	0.532
Plethodon dorsalis	Salamander	0.10	0.714
Batrachoseps pacifica ssp. 1	Salamander	0.64	0.281
Batrachoseps pacifica ssp. 2	Salamander	0.20	0.556
Batrachoseps campi	Salamander	0.16	0.610
Lacerta melisellensis	Lizard	1.9	0.116
Peromyscus californicus	Mouse	2.2	0.102
Peromyscus polionotus	Mouse	0.31	0.446
Thomomys bottae	Gopher	0.86	0.225

Source: Data from Slatkin 1985.

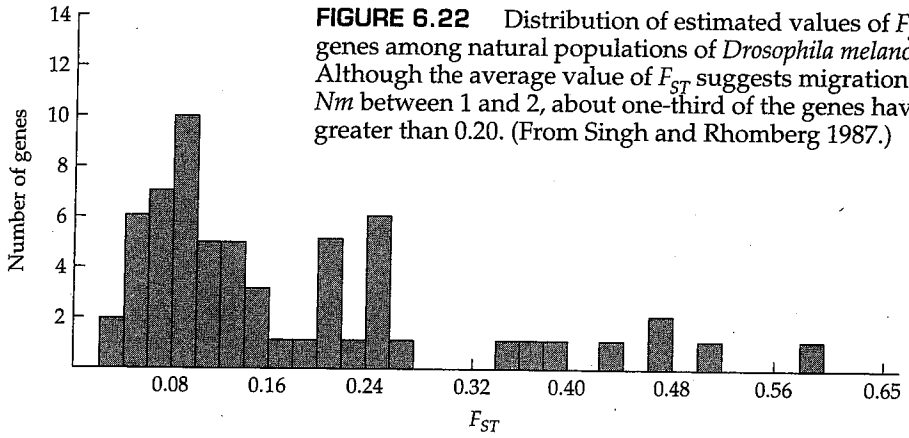


FIGURE 6.22 Distribution of estimated values of F_{ST} for 61 genes among natural populations of *Drosophila melanogaster*. Although the average value of F_{ST} suggests migration at a level of Nm between 1 and 2, about one-third of the genes have F_{ST} values greater than 0.20. (From Singh and Rhomberg 1987.)

great enough to be dispersed by migration. An allele found in only one subpopulation is called a **private allele**. Next we shall see that the rate of migration can be estimated by an examination of the frequency of private alleles.

Estimates of Migration Rates

One method of estimating genetic migration in natural populations relies on the finding that, in theoretical models, the logarithm of Nm decreases approximately as a linear function of the average frequency of private alleles in samples from the subpopulations (Slatkin 1985). Data on the average frequency of private alleles has been compiled and analyzed by Slatkin (1985), and the resulting estimates of Nm and equilibrium values of F_{ST} are summarized in Table 6.5. There is obviously considerable variation in Nm among organisms. However, many of the values of Nm are smaller than about 2, which means that there is still considerable opportunity for genetic divergence among subpopulations.

A second kind of approach to estimating Nm in natural populations is illustrated in Figure 6.22, which gives the distribution of estimated values of F_{ST} among 61 genes in natural populations of *Drosophila melanogaster* (Singh and Rhomberg 1987). The average of the estimated values is $F_{ST} = 0.16$, which, assuming equilibrium, is an estimate of $1 + 4Nm$ (Equation 6.23). The estimate is therefore $Nm = [(1/0.16) - 1]/4 = 1.3$. This estimate is within the range for other *Drosophila* species in Table 6.5. However, there are many genes in Figure 6.22 that have F_{ST} values greater than 0.30. An analogous method of estimating Nm from the F_{ST} values of polymorphic nucleotides within a gene is discussed in Hudson et al. (1992).

Coalescence-Based Estimates of Migration

An island model of migration of the sort depicted in Figure 6.18 assumes that the subpopulations all have the same population size and that migration

between the subpopulations is symmetric. Modern methods based on the coalescence allow these assumptions to be relaxed. For example, Beerli and Felsenstein (1999, 2001) have developed methods that can analyze data from an arbitrary number of subpopulations and estimate the effective population size of each subpopulation and the possibly very unequal rates of migration between any pair of populations. This approach confronts the complexities of migration in nature, where migrants often come primarily from nearby subpopulations. To the extent that neighboring subpopulations have similar allele frequencies, the effects of migration are smaller, and sometimes much smaller, than predicted by the island model. Migration rates are asymmetric because subpopulations may be strung out along one dimension, such as a river bank, or distributed more or less regularly in two dimensions, or there may be one large population with an internal genetic structure caused by the tendency for mating to take place between organisms born in the same region.

The approach of Beerli and Felsenstein (1999, 2001) yields maximum likelihood estimates of subpopulation sizes and migration rates using coalescence theory (see Chapter 3). In this context, as the lineages of alleles are traced back in time, a coalescent event can consist either of an event in which the ancestral lineages of two alleles within the same subpopulation merge into a common ancestral allele, or it can consist of a migration event in which the lineage of an allele switches from one subpopulation to another. The principle is illustrated in Figure 6.23, in which the merging events are depicted as solid lines and the migration events (in this case only one) are denoted by dashed lines.

In this formulation, we consider three kinds of objects: $\boldsymbol{D}$ is a set consisting of the data, $\boldsymbol{P}$ is a set of parameters in the model (in this case, effective population numbers and migration rates), and G is a genealogy of the ancestral history of alleles in the sample. The objective of the analysis is to maximize the likelihood of the data parameters $\boldsymbol{P}$ given the data $\boldsymbol{D}$, which is represented as $L(\boldsymbol{P} \mid \boldsymbol{D})$, through an analysis of all possible genealogies G. Rendered as an equation, the method seeks to find the maximum of

$$\frac{L(\boldsymbol{P} \mid \boldsymbol{D})}{L(\boldsymbol{P}_0 \mid \boldsymbol{D})} \cong \frac{1}{g} \sum_{i=1}^{g} \frac{\text{Prob}(G_i \mid \boldsymbol{P})}{\text{Prob}(G_i \mid \boldsymbol{P}_0)} \tag{6.24}$$

where the summation is across all possible genealogies, and $\boldsymbol{P}_0$ is the set of parameters used to generate the genealogies.

A method for generating genealogies with random topologies and branch lengths was discussed in Chapter 3. There are infinitely many such topologies, so inferences about the values of the parameters must be based on a sample of genealogies. Even so, the space of possible genealogies is so large that purely random genealogies are likely to be far from the region in which Equation 6.24 is maximized. What is needed is a method for systematically exploring the space of genealogies to find the region in which the likelihood

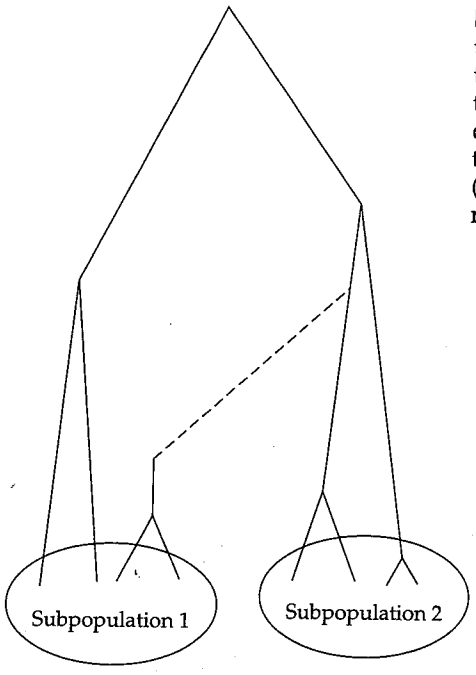


FIGURE 6.23 Coalescence when there is population subdivision. At each coalescence, the lineages of two alleles in the same subpopulation may come together in a common ancestral allele, or the lineage of an allele in one subpopulation may merge with the lineage of an allele from the other subpopulation (indicated here by the dashed line), representing a migration event.

ratio in Equation 6.24 is large, and then concentrating on the genealogies in this region. The most widely used method of this type is called *Metropolis–Hastings Markov chain Monte Carlo* (Gilks et al. 1996). Implementing, checking, and debugging programs that implement such algorithms is an art in itself, and fortunately most authors make their programs freely available through the Internet. The Beerli and Felsenstein (1999, 2001) program to maximize Equation 6.24 for multiple populations with asymmetric migration is called MIGRATE (see also Beerli 2006).

Results from an analysis of a highly variable region of human mitochondrial DNA in a sample of 225 individuals in the Nile valley are shown in Figure 6.24. The groups represented are from Egypt, ancient Nubia, and Sudan, and the number of individuals in each group is shown in parentheses. The authors caution that the groups are themselves assemblages of subpopulations, and that the effective population numbers and migration rates probably vary in time (Beerli and Felsenstein 2001). Because mitochondrial DNA is maternally transmitted, the estimates of effective population size and migration rate are those pertaining only to females. Figure 6.24 shows the estimated number of female immigrants per generation for each of the subpopulations. Gene flow among the groups is on the order of a few females per generation, except for migration into Nubia, which is substantially greater.

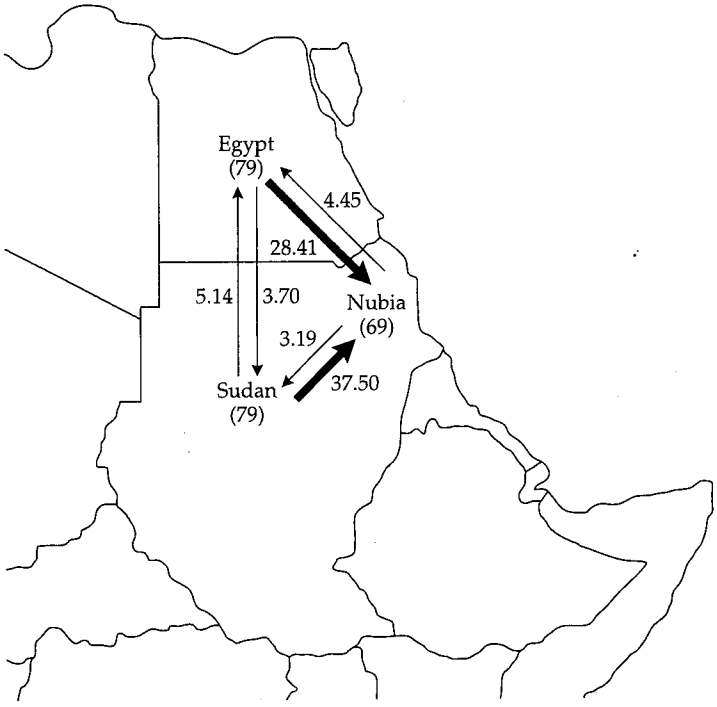


FIGURE 6.24 Estimated migration among subpopulations in Egypt, Nubia, and Sudan based on mitochondrial DNA sequences. The number of individuals sampled in each subpopulation is in parentheses. The number next to each arrow is the estimated number of female migrants along that pathway per generation. (Data from Beerli and Felsenstein 2001.)

Gene flow can also take place between closely related species prior to the time that reproductive isolation becomes complete. Migration resulting in genes flowing from one species into a related species is known as **introgression**. The principles of coalescence can be applied to this situation, too, using the model diagrammed in Figure 6.25. The model is called the *IM model*, where *IM* stands for isolation with migration (Nielsen and Wakeley 2001; Hey and Nielsen 2004). In Figure 6.25, the shaded area represents populations present at various times in the ancestry of two closely related species and their common ancestor. The time scale runs from the earliest time at the top to the present time at the bottom. Six parameters are of interest: the divergence time (t), represented by the horizontal dashed line; three values of $\theta = 4N\mu$, where θ and N are subscripted for the ancestral species A and descendant species 1 and 2; and two values of m, subscripted for introgression from species 1into species 2 (m_{12}) or species 2 into species 1 (m_{21}). Once again, the

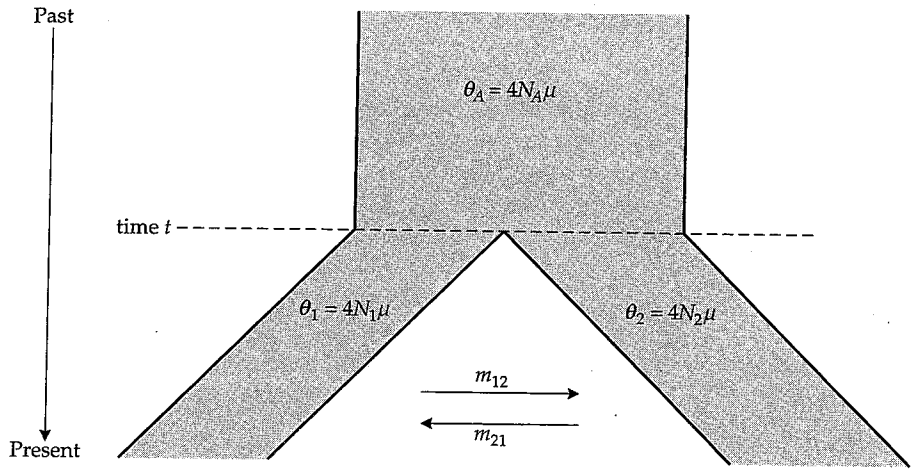


FIGURE 6.25 The isolation-migration model for estimating rates of gene flow between closely related species. The shaded region at the top represents an evolving population, which at time t splits into two species with incomplete reproductive isolation. The six parameters that characterize this situation are indicated, where m_{12} and m_{21} are the rates of migration resulting in gene flow between the species. (After Hey and Nielsen 2004.)

approach is to maximize the likelihood ratio in Equation 6.24, where $\boldsymbol{P}$ is a set of the six parameter values, $\boldsymbol{D}$ is the set of data, and G is a genealogy based on some particular set of parameters $\boldsymbol{P_0}$. Application of Metropolis–Hastings Markov chain Monte Carlo to this situation is described by Hey and Nielsen (2004). In their analysis they also explain how the parameters need to be scaled in some consistent manner using the mutation rate μ. To illustrate the method, they analyzed data from many loci in the closely related species *Drosophila pseudoobscura* and *D. persimilis*. They estimate the divergence time at ~600,000 years, very close to the conventional estimate of ~500,000 years, and find evidence for low levels of gene flow between the species (average Nm in the range 0.06–0.19), with great variation among loci and the direction of the introgression.

Migration-Selection Balance

Just as recurrent mutation to a deleterious allele can maintain the allele in a population in spite of selection against individuals that carry the allele, resulting in a mutation-selection balance (see Chapter 5), recurrent migration can maintain a deleterious allele in a state of **migration-selection balance**. This situation can arise when an allele is deleterious in one geographical

region but not deleterious, or less deleterious, in a neighboring geographical region. Migrants from the latter region continually replenish the deleterious allele in the former region, where selection acts against it.

A selection model similar to that used for mutation-selection balance (see Chapter 5) reveals the principal offsetting forces for migration-selection balance. Let *AA*, *Aa*, and *aa* be three genotypes at a locus, where *a* is a deleterious recessive or partially recessive allele. As in Chapter 5, we denote the relative fitnesses of *AA*, *Aa*, and *aa* as 1, $1 - hs$, and $1 - s$, respectively, where s is the selection coefficient against *aa* and h is the degree of dominance of *a*. When $h = 0$ the *a* allele is completely recessive, and when $h = \frac{1}{2}$ the relative fitness of *Aa* is the arithmetic mean of those of *AA* and *aa*, indicating additive effects of *A* and *a*. Let p and q be the allele frequencies of *A* and *a*, with $p + q = 1$, and suppose that the selection is sufficiently weak, or the recessive allele sufficiently rare, and that the three genotypes are approximately in the Hardy-Weinberg frequencies of p^2, $2pq$, and q^2.

This model was originally studied by Haldane (1930) and Wright (1931), who showed that the change Δq in the allele frequency of *a* in the region where it is deleterious is given by

$$\Delta q = \frac{-spq[q + h(p - q)]}{1 - sq(2hp + q)} + m_i q^* - m_o q \qquad (6.25)$$

where m_i is the rate of in-migration of individuals from outside the population, among whom the allele frequency is q^*, and m_o is the rate of out-migration of individuals who leave the population.

An ingenious application of Equation 6.25 is reported in Hoekstra et al. (2004). These authors studied rock-pocket mice (*Chaetodipus intermedius*) in southern Arizona, where there is a gradient in habitat color from the presence of dark-colored volcanic lava, whereas surrounding regions are light-colored granitic rocks. The mice inhabiting the volcanic rock have dark coats composed of uniformly melanic hair, whereas the mice inhabiting the light areas have light coats composed of hair with only a small band of melanin. The difference in phenotype is presumed to be an adaptation to reduce visibility and hence predation. The genetic basis of the difference is due to four amino acid replacements in the melanocortin-1 receptor protein, encoded in the gene *Mc1r*. Genotypes denoted *DD* and *Dd* have coats that are melanic and dark, whereas the genotype *dd* coat is nonmelanic and light. In the habitat consisting of dark-colored volcanic lava, the *d* allele is presumably a harmful recessive allele maintained by migration from the surrounding light areas.

To investigate this hypothesis in detail, Hoekstra et al. (2004) sequenced the *Mc1r* gene in 57 individuals trapped in dark areas and 118 individuals trapped in light areas at sites spread across a 35-km east-west transect. They also sequenced two genes in mitochondrial DNA having no relation to the coat-color polymorphism. As expected, they found a strong correlation

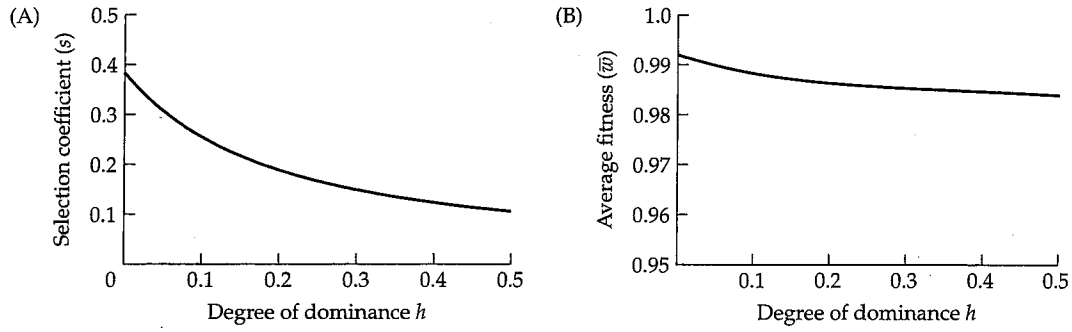


FIGURE 6.26 (A) Theoretical relation between the selection coefficient against the nonmelanic recessive allele (s) and its degree of dominance (h) in rock-pocket mice that inhabit dark-colored volcanic lava, assuming migration-selection balance. (B) Equilibrium average fitness in the population for various degrees of dominance, h. (Based on data in Hoekstra et al. 2004.)

between the dark habitat and the D allele frequency, but no correlation with the mitochondrial genes.

In the next step of the analysis, they used the coalescent methods of Beerli and Felsenstein (1999, 2001), discussed in the previous section, to estimate the migration rates m_i and m_o in Equation 6.25, based solely on the mitochondrial markers and an estimate of the effective population size. Their *Mc1r* sequences provided values for q, the frequency of the recessive nonmelanic allele d in dark habitats, and q^*, the frequency of the same allele in light habitats. Assuming migration-selection balance, Equation 6.25 gives the relation between the selection coefficient s and the degree of dominance h in terms of q, q^*, m_i, and m_o, all of which had been estimated.

Figure 6.26A shows the inferred relation between s and h for an effective population size of $N_e = 10{,}000$. As h runs from 0 (d a complete recessive) to $\frac{1}{2}$ (additive effects), s runs from 0.389 to 0.108. The decrease in s reaffirms a point made in Chapter 5, which is that partial dominance of a rare recessive allele has a large effect on the equilibrium frequency, because with even a small degree of dominance, the selection is spread across a much larger number of individuals. The effect of dominance on the average fitness in the equilibrium population is shown in Figure 6.26B, and it is very small, ranging from $\overline{w} = 0.992$ for $h = 0$ to $\overline{w} = 0.985$ for $h = \frac{1}{2}$. Hoekstra et al. (2004) suggest that h is unlikely to be greater than about 0.4, and point out that their largest estimates of the selection coefficient are on the same order of magnitude as estimates of the selection coefficient for melanic peppered moths (see Figure 6.21A).

SUMMARY

1. The inbreeding coefficient is the probability that two alleles in an inbred individual are identical by descent (autozygous) through DNA replication of a single allele in a common ancestor, relative to some arbitrary reference time in the past.
2. For a mating between relatives whose pedigree is known, the inbreeding coefficient can be calculated using elementary principles of probability.
3. The genotype frequencies among inbred individuals depart from Hardy-Weinberg frequencies in that the expected frequency of heterozygous genotypes is reduced and that of homozygous genotypes increase. In the extreme case of complete inbreeding, the frequency of heterozygous genotypes is 0.
4. In species that normally undergo outcrossing, inbreeding typically has harmful effects because of the increased frequency of genotypes that are homozygous for rare deleterious alleles.
5. In most regular systems of mating, in which generation after generation individuals with the same degree of genetic relationship are mated together, the inbreeding coefficient increases gradually through time. At any stage in the process, a single generation of random mating erases all of the accumulated inbreeding, and the population returns to Hardy-Weinberg genotype frequencies.
6. Population structure (population subdivision) increases the probability that two randomly chosen alleles in the same subpopulation are identical by descent because of random genetic drift among the subpopulations and the dispersion of allele frequencies. Even though each subpopulation may undergo random mating and its genotype frequencies conform to the Hardy-Weinberg proportions, across the population as a whole there is a deficiency of heterozygous genotypes and an excess of homozygous genotypes. The fixation index is one measure of the magnitude of the departure from Hardy-Weinberg proportions in the population as a whole.
7. A blue versus white polymorphism in flower color in the "desert snow" plant *Linanthus parryae* in the Mojave desert became the classic example of isolation by distance in an organism with a hierarchical population structure. Although the relative roles of random genetic drift and natural selection in causing the differences in flower-color frequencies among subpopulations have been argued for more than 60 years, the most recent evidence supports earlier studies in indicating that selection is a key factor.
8. Studies of hundreds of molecular polymorphisms in large samples of human individuals supports grouping the individuals by genotype into a few groups largely coinciding with major geographical regions. How-

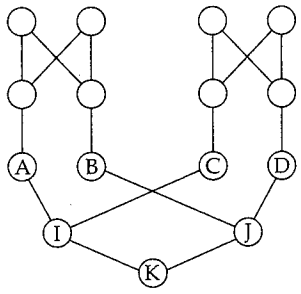
ever, the genetic differences between two randomly chosen individuals from different groups are only slightly greater than those between two unrelated individuals from the same group. In particular, genetic differences among individuals within any one group account for 93–95% of the total genetic variation, and only 3–5% of the genetic variation is ascribable to differences between the major groups.

9. The Wahlund principle refers to the reduction in the average frequency of homozygous genotypes that occurs when subpopulations fuse and form a larger random-mating population. The magnitude of the reduction is a function of the variance in allele frequency among the subpopulations.
10. Migration among subpopulations tends to counteract the dispersion of allele frequencies due to random genetic drift. In simple models, such as the island model of migration, even a few migrant individuals per generation are sufficient to maintain the fixation index of genetic differentiation among the subpopulations in the "little" to "moderate" range.
11. Application of coalescent theory to subdivided populations enables estimates to be made of effective population number and asymmetrical migration rates among subpopulations, as well as of divergence time and the magnitude of introgression among closely related species.
12. Deleterious alleles can be maintained in a population through migration from adjacent populations in which the allele is not as deleterious. A migration-selection balance analogous to mutation-selection balance results. An example is the selection for melanic rock-pocket mice that live on dark-colored volcanic lava, where the recessive nonmelanic allele is continually introduced by migration of mice from surrounding habitats consisting of light-colored granitic rocks.

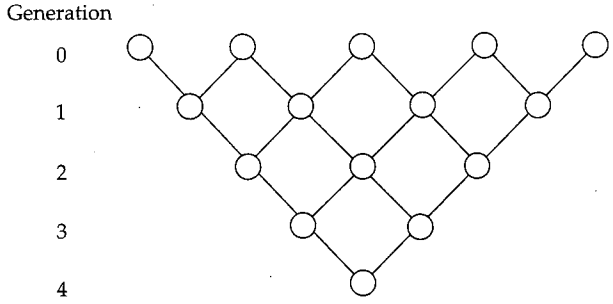
PROBLEMS

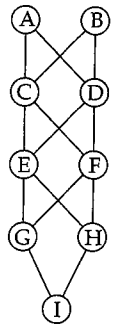
1. Explain why the concept of identity by descent (autozygosity) is fundamental to an understanding of the effects of inbreeding. Under what circumstances can the probability of autozygosity in a population increase without any increase in homozygosity?
2. The coefficient of consanguinity between two individuals is the probability that two alleles of a gene, one drawn at random from each of the individuals, are identical by descent. How is the coefficient of consanguinity between two individuals related to the inbreeding coefficient of a hypothetical offspring of these two individuals?
3. Consider two alleles A and a at frequencies $\frac{1}{2}$ and $\frac{1}{2}$ in a population in which the inbreeding coefficient equals F. What value of F results in genotype frequencies of $\frac{1}{3} : \frac{1}{3} : \frac{1}{3}$?

4. Show that $p^2(1 - F) + pF = p^2 + pqF = p - (1 - F)pq$, when $q = 1 - p$.
5. In a population of monoecious plants in Hardy-Weinberg proportions for two alleles with allele frequency p, what is the variance in allele frequency among plants? What is the variance if the population were completely inbred? If a random mating population were to undergo repeated self-fertilization, what would the variance be when the inbreeding coefficient equals F?
6. Equation 6.7 can also apply to the probability of autozygosity for X-linked genes provided that (a) males are regarded as having an inbreeding coefficient of 1, and (b) any path with two consecutive males is disregarded. Explain why these provisions are necessary.
7. What is the inbreeding coefficient of individual K in the accompanying pedigree, assuming that none of the individuals at the top of the pedigree is inbred.



8. The accompanying pedigree shows several generations of half-sib mating. Assuming that the individuals in generation 0 have $F_0 = 0$, what are the inbreeding coefficients of the individuals in generations 1, 2, 3, and 4?





9. Assuming $F_A = F_B = 0$, calculate the inbreeding coefficient for each of the individuals C–I in the accompanying pedigree (see figure at right).
10. Derive a recursion equation for F_t for repeated parent-offspring mating, and calculate F_t for $t = 0$ to 5.
11. For a gene with two alleles and $p = 0.3$, what are the expected genotype frequencies after five generations of sib mating? What are the expected genotype frequencies after one additional generation of random mating?
12. With two alleles and $p = \frac{1}{2}$, what are the expected genotype frequencies in a random-mating population and among the offspring of first cousins? How great is the decrease in heterozygosity in the inbred population relative to the random mating population?
13. If the frequency of an autosomal recessive disorder is $\frac{1}{1600}$ among unrelated parents, what is the expected frequency among the offspring of first cousins?
14. For a recessive allele at frequency q in a population in which one percent of the matings are between first cousins, but otherwise occur at random, the proportion of affected individuals having first-cousin parents is $(1 + 15q)/(1 + 1599q)$. Calculate this ratio for $q = 0.1, 0.05, 0.1, 0.005$, and 0.001. Interpret the result when $q = 1$.
15. Two-way hybrid corn is produced by crossing two different inbred lines; three-way hybrids are produced by crossing a two-way hybrid with an unrelated inbred; and four-way hybrids are produced by crossing two different two-way hybrids. What is the inbreeding coefficient of the offspring of randomly mated two-way, three-way, or four-way hybrids? (Hint: Consider the allele frequencies in gametes.)
16. If a population is maintained by self-fertilization in even-numbered generations and by random mating in odd-numbered generations, what happens to the inbreeding coefficient?
17. Consider a population of plants in which, in each generation, a fraction S of the population (a random sample of all individuals) undergoes self-fertilization, and the remaining fraction $1 - S$ undergoes outcrossing (random mating). Assuming that there is no heredity tendency for the plants to self-fertilize or outcross, show that the magnitude of the inbreeding coefficient F at equilibrium equals $S/(2 - S)$.
18. Two diploid random mating populations have allele frequencies $q + \varepsilon$ and $q - \varepsilon$ for a recessive allele of a gene. What are the frequencies of homozygous recessives before and after population fusion?
19. Show that $F_{IT} = F_{IS} + F_{ST} - F_{IS}F_{ST}$ and interpret the expression.

20. Calculate F_{ST} among the three random-mating populations below based on the specified allele frequencies. What is the maximum value of F_{ST} in this situation?

Population	Population 1	Population 2	Population 3
Allele 1	0.1	0.2	0.3
Allele 2	0.3	0.3	0.3
Allele 3	0.6	0.5	0.4

21. Calculate F_{IS}, F_{ST}, and F_{IT} for the populations with the genotype frequencies shown in the following table:

	Population 1	Population 2
Genotype *AA*	0.056	0.072
Aa	0.288	0.256
aa	0.656	0.672

22. What is the inbreeding coefficient in a population of size 50 that undergoes

    (a) exactly 47 generations of random mating followed by three generations of sib mating?

    (b) 50 generations of random mating?

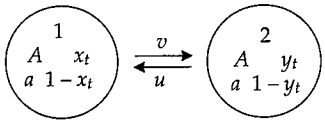
23. If a mainland population of snails has an allele frequency of 0.8 and an island population has a frequency of 0.2, how many generations are required for the island population to achieve an allele frequency of 0.5, given a migration rate of 0.01?

24. If four populations with allele frequencies 0.2, 0.4, 0.6, and 0.8 undergo migration according to the island model with $m = 0.05$, what are the expected allele frequencies after 10 generations?

25. In the island model of migration, how does the variance in allele frequency among populations at time t, σ_t^2, change as a function of m and t?

26. When random genetic drift is offset by migration among populations in the island model, what value of m is necessary to keep the equilibrium value of F smaller than 0.05?

27. Two island populations 1 and 2 are shown in the accompanying diagram. In population 1 the allele frequency of A in generation t is x_t, and in population 2 it is y_t. In every generation, a fraction $u > 0$ of the alleles in population 1 is removed and replaced with alleles from population 2, and a fraction $v > 0$ of the alleles in population 2 is removed and replaced with alleles from population 1.



The equations relating x_t to x_{t-1} and y_t to y_{t-1} are

$$x_t = x_{t-1}(1-u) + y_{t-1}u$$
$$y_t = y_{t-1}(1-v) + x_{t-1}v$$

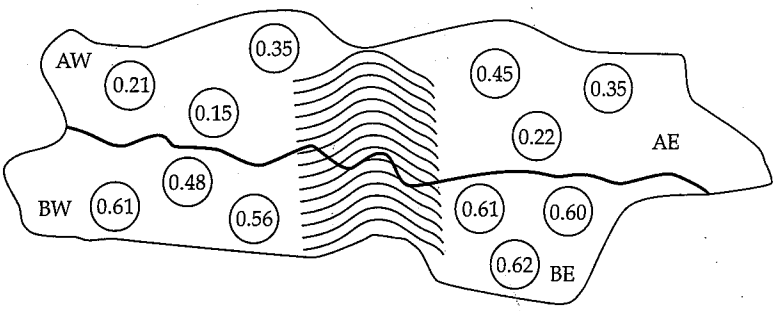
(a) Derive an expression for $x_t - y_t$ in terms of $x_0 - y_0$, and conclude that the equilibrium frequency $\hat{x}$ of A in population 1 equals the equilibrium frequency $\hat{y}$ of A in population 2.

(b) Derive an expression for $vx_t + uy_t$ in terms of x_0 and y_0, and conclude that the equilibrium frequencies are give by

$$\hat{x} = \hat{y} = \frac{vx_0 + uy_0}{u+v}$$

(c) Explain how the approach to equilibrium differs between the case $0 < u + v < 1$ and the case $1 < u + v < 2$.

28. In the Swabian Alps in Southern Germany, a verdant meadow serves as home to subpopulations of the incredible edible snail *Helix pomatia*. The subpopulations differ in the allele frequency of a mutation affecting shell coloration. A river meanders through the meadow from west to east, and a prominent knoll interrupts it from north to south. The accompanying crude diagram of the site shows the estimated allele frequency in a sample taken at each of 12 collection sites.



Assuming that the differences in allele frequency are due mainly to random genetic drift, does the river or the knoll seem to be the stronger isolating barrier between the snail subpopulations? To solve this problem, first consider the subpopulation as divided into regions above (A) and below (B) the river (R), or as divided into regions west (W) or east (E) of the knoll (K). You should calculate F_{SR}, F_{RT}, and F_{ST} for the division based on the river, and F_{SK}, F_{KT}, and F_{ST} for the division based on the knoll. The relative values of F_{SR} and F_{KT} should tell you whether the river or the knoll is the more significant barrier to genetic exchange.

7.1 The Neutral Theory and Molecular Evolution 318

Theoretical Principles of the Neutral Theory 318

7.2 Estimating Rates of Molecular Sequence Divergence 321

Rates of Amino Acid Replacement 322

Rates of Nucleotide Substitution 326

Statistical Fitting of Nucleotide Substitution Models 329

7.3 The Molecular Clock 329

Variation among Genes in the Rate of the Molecular Clock 332

Variation across Lineages in Clock Rate 334

The Generation-Time Effect 337

The Overdispersed Molecular Clock and the Neutral Theory 338

The Nearly Neutral Theory 339

7.4 Patterns of Nucleotide and Amino Acid Substitution 340

Calculating Synonymous and Nonsynonymous Substitution Rates 340

Codon Substitution Models 343

Observations of Synonymous and Nonsynonymous Substitution Rates 345

Polymorphism within Species 347

Implications of Codon Usage Bias 349

7.5 Polymorphism and Divergence in Nucleotide Sequence—The McDonald-Kreitman and HKA Tests 351

Polymorphism and Divergence in Noncoding Sequences 354

Impact of Local Recombination Rates 354

Substitution Models for Structural RNA Genes 356

7.6 Gene Genealogies 358

Hypothesis Testing Using Trees 358

7.7 Mitochondrial and Chloroplast DNA Evolution 362

Chloroplast DNA and Organelle Transmission in Plants 363

Maintenance of Variation in Organelle Genomes 364

Evidence for Selection in mtDNA 365

7.8 Molecular Phylogenetics 366

Algorithms for Phylogenetic Tree Reconstruction 366

Distance Methods versus Parsimony 369

Bootstrapping and Statistical Confidence in a Tree 370

Bayesian Methods 370

Trans-Species Polymorphism 371

7.9 Multigene Families 372

Concerted Evolution 374

Subfunctionalization 375

Birth-and-Death Process 376

CHAPTER 7

MOLECULAR POPULATION GENETICS

It has been relatively straightforward to convert models of classical population genetics from genes as chemically unspecified entities to genes as sequences of nucleotides in a molecule of DNA. All the forces studied in classical population genetics, such as mutation, migration, selection, and random drift have an impact on the pattern of variation seen in molecular sequences of genes. A primary focus of molecular population genetics is to make inferences about the contribution of each of these evolutionary forces to generate the patterns of molecular sequence variation we see today. Usually this process involves a close interplay between mathematical model building, statistical parameter estimation, and experimental observation. Having information about the complete DNA sequence of genes provides a much richer picture of variation than is connoted by the A and a of classical population genetics, and so the field of molecular population genetics requires more detailed parameterization of processes like mutation. Whenever unexpected patterns of sequence variation are observed, new opportunities for theoretical inquiry are opened. Often inferences about evolutionary forces transcend species boundaries by making use of data on both within-species polymorphism and between-species divergence. The genetic basis for species isolation is itself amenable to population genetic analysis. But first we begin with the basic theoretical principles that underlie molecular population genetics.

7.1 THE NEUTRAL THEORY AND MOLECULAR EVOLUTION

Motoo Kimura suggested that most polymorphisms observed at the molecular level are selectively neutral, so that their frequency dynamics in a population are determined primarily by random genetic drift (Kimura 1968). When this theory was first developed, the data in hand consisted of protein polymorphism data, but the model applies directly to nucleotide sequence data as well. Recently, the explosion in availability of DNA sequence variation has provided enough statistical power to test the correspondence between theories and data, and in many cases the strict neutral theory can be rejected. Nevertheless, the neutral theory has been of great importance in population genetics in stimulating the collection and analysis of data. The neutral theory has also led to one of the most complete and elegant mathematical theories in all of biology. Traditional tests of the concordance of sample data with the neutral theory are almost universally low in power, which means that large sets of data are needed before one has a reasonable chance of rejecting neutrality. Even when enough data exists so that the neutral theory can be rejected, there are many possible reasons for the rejection, including heterogeneity in the mutation process, mis-specification of the sampling process, hidden population structure and migration, and finally, natural selection. Regardless of the action of other forces shaping molecular sequence variation in populations, the force of random drift is always present to one extent or another, and for this reason the neutral theory remains useful in generating rigorous null hypotheses. The next section summarizes some of the theoretical implications of the neutral theory and some of the data bearing on it.

Theoretical Principles of the Neutral Theory

The original formulation of the neutral theory focused on mutations that are strictly selectively neutral so that their fate is determined purely through random genetic drift. Kimura knew, of course, that mutations could have effects on fitness, but he supposed that the mutations would either be deleterious and rapidly eliminated, or much more rarely, that they would be favorable and rapidly fixed. Because the deleterious mutations are removed quickly, the effect would be as though the neutral mutation rate were simply lower by an amount equal to the proportion of mutations that are deleterious. For now we will develop the principles of this black-and-white strictly neutral theory, and later we will see that there is an important shades-of-gray version called the nearly neutral model.

A variety of neutral mutation models have been considered, including the infinite-alleles and infinite-sites models discussed in Chapter 4, as well as finite-sites models. In all models, though, random drift occurs when N adult individuals produce an infinite pool of gametes from which $2N$ are chosen at random to create the N zygotes of the next generation. Much of the complexity of the mathematics of the neutral theory arises from the fact that the

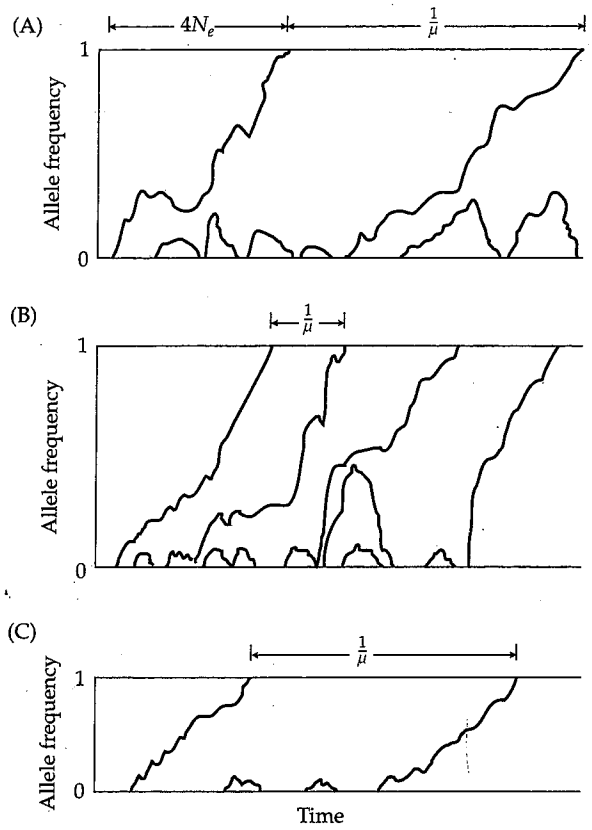


FIGURE 7.1 The trajectories of neutral alleles in an ideal population. New alleles enter the population by mutation and have an initial allele frequency of $1/(2N)$. Most alleles are lost, but those that go to fixation take an average of $4N$ generations. The time between successive fixations of neutral alleles is $1/\mu$ generations. (A) A moderate size population. (B) The same population size with a higher mutation rate has the same time to fixation but less time between fixations. (C) A smaller population has alleles that go to fixation more rapidly, but the time between fixations is still $1/\mu$. (After Kimura 1980.)

mutational histories of alleles are not independent, because they share an overlapping genealogical history. Before we get into the details of the predictions of the neutral theory, let us first review some of the theory's main principles (Kimura 1983).

1. If a population contains a neutral allele with allele frequency p_0, then the probability that the allele eventually becomes fixed equals p_0. In particular, a newly arising neutral mutation occurs in just one copy, so the initial allele frequency is $p_0 = 1/(2N)$, and the probability of eventual fixation of the mutation is therefore $1/(2N)$. Figure 7.1 shows that a mutant allele

arising in a small population has higher chance of fixation than if it occurred in a large population.

2. The steady-state rate at which neutral mutations are fixed in a population equals μ, where μ is the **neutral mutation rate**. It is noteworthy that the equilibrium rate of fixation does not involve the population size N. The reason is that the N cancels out: The overall rate is determined by the product of the probability of fixation of new neutral mutations $1/(2N)$ and the average number of new neutral mutations in each generation $(2N\mu)$; hence $1/(2N) \times (2N\mu) = \mu$.
3. The average time that occurs between consecutive neutral substitutions equals $1/\mu$. This principle follows directly from the one above. If the steady-state rate of fixation is μ per unit time, the average length of time between substitutions will be the reciprocal, or $1/\mu$. By way of analogy, if a Swiss clock cuckoos at the rate of 24 times per day, then the average length of time between cuckoos is $\frac{1}{24}$th of a day, or one hour. As Figure 7.1 shows, the time interval between fixations is independent of population size, and elevating the mutation rate decreases the time interval between fixations.
4. Analysis of the diffusion equation has shown that, among newly arising neutral alleles that are destined to be fixed, the average time to fixation is $4N_e$ generations (where N_e is the effective population size). This too is evident in Figure 7.1: Alleles that go to fixation do so in less time in the smaller population. Among newly arising neutral alleles destined to be lost, the average time to loss is $(2N_e/N)\ln(2N)$ generations. The average times required for fixation or loss apply to newly arising alleles, which are necessarily present in just one copy, so $p_0 = 1/(2N)$. The implication of these formulas is that, on average, neutral mutations that are destined to be fixed require a very long time for this to occur, but mutations destined to be lost are lost quite rapidly.
5. If each neutral mutation creates an allele that is different from all others existing in the population in which it occurs, then at equilibrium when the average number of new alleles gained through mutation is exactly offset by the average number lost through random genetic drift, the expected homozygosity equals $1/(4N_e\mu + 1)$, where μ is the neutral mutation rate. The model of mutation in which each new allele is novel is called the **infinite-alleles model** of mutation. The quantity $4N_e\mu$, which appears frequently in the neutral theory, is often denoted as θ. The equilibrium average homozygosity is therefore $1/(1 + \theta)$. Since the heterozygosity equals one minus the homozygosity, the average heterozygosity at equilibrium in the infinite-alleles model equals $\theta/(1 + \theta)$. Larger populations are expected to have a higher heterozygosity, as reflected in the greater number of alleles segregating at any one time in the larger populations in Figure 7.2.

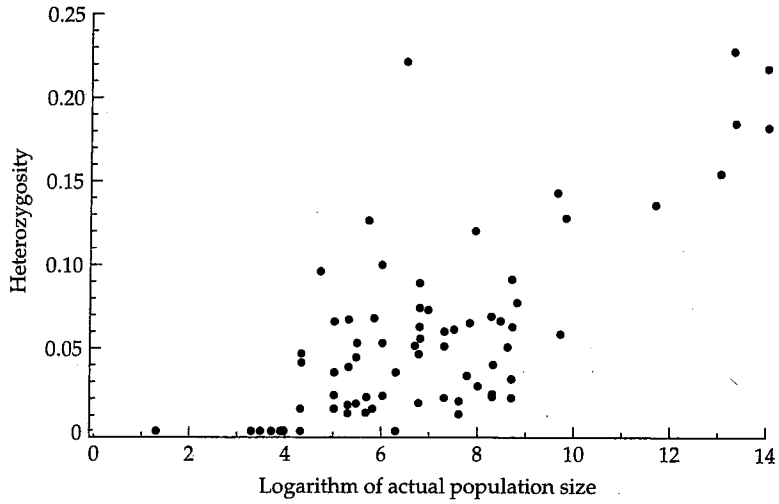


FIGURE 7.2 Given the enormous variation in population sizes, one would expect to see a wider range in variation in heterozygosity than is actually observed. The relation between population size and heterozygosity does not fit the simplest form of the neutral theory across a wide range of intermediate population sizes. (After Nei and Graur 1984.)

PROBLEM 7.1 Considering principles 3 and 4 governing the random genetic drift of neutral alleles, what magnitude of $4N_e\mu$ is needed to ensure that the trajectories of two independent neutral mutations destined to be fixed are unlikely to overlap?

ANSWER Since the average time to fixation is $4N_e$ generations and the average interval between mutations destined to be fixed equals $1/\mu$, the required condition is $1/\mu >> 4N_e$, or $4N_e\mu << 1$.

7.2 ESTIMATING RATES OF MOLECULAR SEQUENCE DIVERGENCE

One of the fundamental principles of biological evolution is that processes that produce genetic variation within populations also produce genetic differences between species. Genetic variation within populations is known as *polymorphism,* whereas genetic differences among species are known as **divergence.** In this section we consider patterns of sequence divergence from a quantitative point of view. When two species initially diverge their nucleotide sequences are virtually identical, but as time goes on the se-

quences accumulate differences until eventually, except for sequences that are constrained by selection, they become so different that they resemble each other no more than would be expected by chance.

Rates of Amino Acid Replacement

The initial impetus for the neutral theory came from observations on the rate of amino acid replacements in proteins. When extrapolated to the entire genome, the inferred rate of evolution was several nucleotide substitutions per year. This rate was regarded as much too high to result from natural selection, because the intensity of selection must be limited by the total amount of differential survival and reproduction that occurs in the organism. Direct DNA sequencing later revealed that rates of nucleotide substitution vary according to the function (or presumed absence of function) of the nucleotides. The type of data that must be analyzed is best illustrated by example. The first 18 amino acids present at the amino terminal end of the human and mouse insulin proteins constitute a signal peptide that is used in secretion of the molecules. The sequences are:

Human:	Met	Ala	Leu	Trp	Met	Arg	Leu	Leu	Pro	Leu	Leu	Ala	Leu	Leu	Ala	Leu	Trp	Gly
Mouse:	Met	Ala	Leu	Leu	Val	His	Phe	Leu	Pro	Leu	Leu	Ala	Leu	Leu	Ala	Leu	Trp	Glu

In order to calculate the proportion of amino acids that differ in the two signal sequences, we can simply count the number of sites that are the same and the number of sites that differ. Among the 18 amino acids there are 5 differences, so the proportion different is $\frac{5}{18} = 0.28$.

To interpret these data, let us suppose that amino acid replacements occur at the rate λ per unit time. Consider two independently evolving sequences, initially identical, which at time t are found to differ in the proportion D_t of their amino acids. After the next time interval, the proportion of differences D_{t+1} is given by

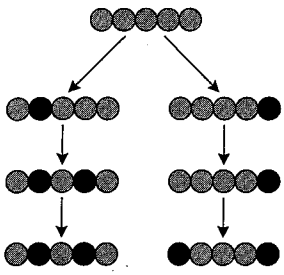
$$D_{t+1} = (1 - D_t)(2\lambda) + D_t \tag{7.1}$$

In this equation, $(1 - D_t)(2\lambda)$ is the proportion of sites, previously identical, in which one or the other underwent an amino acid replacement during the time interval in question. This proportion must be added to the already existing differences D_t in order to give the total. (The equation ignores the unlikely possibility of an amino acid replacement making two previously different amino acid sites identical.) The factor of 2 is present because the total time for evolution is $2t$ units (t units in each lineage after the split), which is illustrated in Figure 7.3. Equation 7.1 suggests the differential equation

$$dD/dt = D_{t+1} - D_t = 2\lambda - 2\lambda D_t \tag{7.2}$$

which has the solution

$$D_t = 1 - e^{-2\lambda t} \tag{7.3}$$



Elapsed time	Divergence time	Proportion of differences	Substitutions per site
		D	K
1	2	$\frac{2}{5}$	0.51
2	4	$\frac{3}{5}$	0.92
3	6	$\frac{4}{5}$	1.61

FIGURE 7.3 The ancestral sequence of a five amino acid–long peptide (top left) is depicted as changing over time in two derived species. In the first time interval ($t = 1$), one substitution occurred in species 1 and a different substitution occurred in species 2. The total time for the divergence between the two species is $2t$, because the mutations could occur independently along each lineage. Note also the difference between the observed proportion of differences in sequence and estimated number of substitutions per site.

An alternative argument can be used to derive Equation 7.3 without resorting to differential equations. If λ is the rate of amino acid replacement per unit time, then the probability that a particular site remains unsubstituted for t consecutive intervals along each of two independent lineages is $(1 - \lambda)^{2t}$, which is approximately equal to $e^{-2\lambda t}$, provided that λt is not too large. Thus, the probability D_t of one or more replacements occurring in t units of time after divergence is approximately $1 - e^{-2\lambda t}$, which is shown in Equation 7.3.

Since λ is the rate of amino acid replacement per unit time, the expected proportion of differences between two sequences at any time t is

$$K = 2\lambda t \tag{7.4}$$

where the factor of 2 is again present because the total time for evolution is $2t$ units (see Figure 7.3).

Substituting K from Equation 7.4 into Equation 7.3 and rearranging yields the following estimate $\hat{K}$ of K,

$$\hat{K} = -\ln(1 - \hat{D}) \tag{7.5}$$

where $\hat{D}$ is the observed proportion of sites in which two sequences differ. This is the equation used to obtain the right-hand column in Figure 7.3. If the sequences under comparison are L amino acids in length, then the estimated variance $Var(\hat{K})$ of $\hat{K}$ is estimated from the distribution of K implied by the substitution process and is approximately

$$Var(\hat{K}) = \hat{D}/[(1 - \hat{D})L] \tag{7.6}$$

The rate of evolution at the molecular level is given by the amount of sequence divergence that occurs per unit of time. Thus, as suggested by

Equations 7.4 and 7.5, if two sequences are compared, and these are known to have diverged from a common ancestral sequence an estimated $\hat{t}$ time units ago, then the rate of evolution λ may be estimated as

$$\hat{\lambda} = \hat{K}/(2\hat{t}) \qquad [7.7]$$

The units of $\hat{\lambda}$ are usually expressed as replacements per amino acid site (or substitutions per nucleotide site) per year.

The quantity $\hat{K}$ is used in preference to $\hat{D}$ in estimating the rate of molecular evolution because $\hat{K}$ takes multiple substitutions into account. Over long periods of evolutionary time, the amino acid present at a particular site may be replaced several times, first by one alternative, then by another, then still another, and perhaps, at some stage, even return to the amino acid originally present at the site. When comparing two sequences, only the sites that are different can be identified. Sites that are identical at the present time may include some that were different in the past, and sites that are different at the present time might have undergone more than one substitution. The quantity $\hat{D}$ is determined only by the proportion of differences between the sequences observed at the present time. The estimate $\hat{K}$ makes a correction for multiple substitutions, but at the cost of introducing assumptions that the substitutions occur independently and at the same rate through time.

For relatively short intervals of evolutionary time, during which multiple substitutions remain uncommon, the correction is minor, and the value of $\hat{K}$ is close to that of $\hat{D}$. This can be seen by the fact that the initial slope of the curve plotted in Figure 7.4 is 1. As the observed sequence divergence increases, it becomes more likely that multiple hits have happened, and the slope

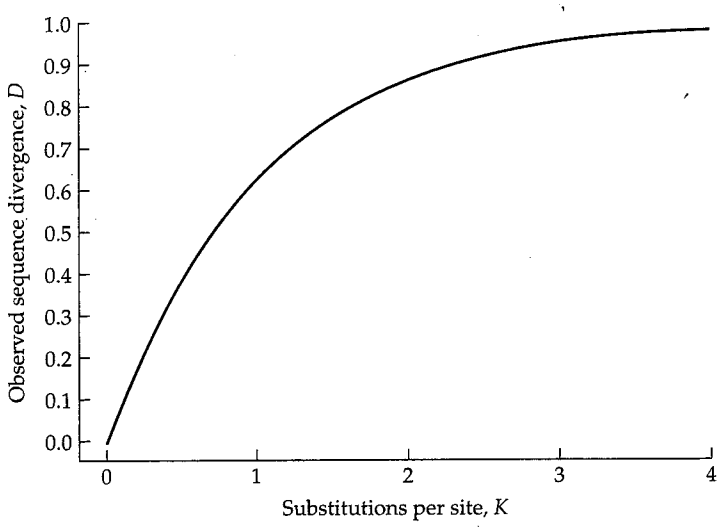


FIGURE 7.4 As sequences become more divergent over time, the number of substitutions per site (K) can continue to increase, but the proportion of sites that mismatch in the observed sequences (D) becomes saturated.

decreases. Over longer intervals, when many multiple substitutions have occurred, the correction is important and the assumptions on which it is based must be evaluated critically. Correction for multiple substitution events is even more important for nucleotides than it is for amino acids. With amino acids, the probability of a random replacement returning an amino acid site to its original identity is $\frac{1}{20}$ (assuming equal frequencies), whereas for nucleotides it is $\frac{1}{4}$.

PROBLEM 7.2 Use the data in the insulin example on page 322 to estimate the average rate of amino acid replacement in the signal peptide of insulin during the divergence of mice and humans. Based on fossil evidence, the separation of these species occurred approximately 80 million years ago.

ANSWER For the signal peptide, $\hat{D} = 0.28$ and $\hat{K} = -\ln(1 - 0.28) = 0.33$. The estimated rate of evolution is therefore $0.33/[2 \times (80 \times 10^6)] = 2.1 \times 10^{-9}$ amino acid replacements per amino acid site per year. The standard deviation of $\hat{K}$ is estimated as equal to $\sqrt{0.28/(0.72 \times 18)} = 0.15$. With such a small sample size, the estimates could ordinarily not be taken too literally. However, in this case, the average rate for the signal sequence is characteristically faster than the average rate for the molecule as a whole. For insulin, among 108 aligned amino acid sites there are 23 differences, giving $\hat{K} = 0.24 \pm 0.05$ and an average rate of 1.5×10^{-9} amino acid replacements per amino acid site per year.

Rates of amino acid replacement vary over a 500-fold range in different proteins (Li 1997). The rate of amino acid replacement among the most rapidly evolving mammalian proteins is about 5×10^{-9} per year, whereas the slowest evolving proteins it is about 0.01×10^{-9} per year. The average rate among a large number of proteins is very close to the rate found in hemoglobin, which is approximately 1×10^{-9} amino acid replacements per amino acid site per year.

To be concrete about the interpretation of the rate of amino acid replacement, consider a protein exactly 100 amino acids in length, in which the rate of amino acid replacement per amino acid site equals 1.0×10^{-9} per year. For the entire protein, the rate of replacement equals $100 \times 1.0 \times 10^{-9} = 1 \times 10^{-7}$ per year. In two different species, therefore, the protein would accumulate amino acid differences at the rate of one replacement every 5 million years since their divergence from a common ancestor [because $(5 \times 10^6) \times 2 \times (1 \times 10^{-7}) = 1.0$].

The simple model that we just examined makes an assumption that is contradicted by an abundance of data. We assumed that all amino acid replacements occur with equal likelihood. Besides the fact that real proteins violate this assumption, we might not have expected it to be true because some amino acid changes require a single underlying nucleotide change while others require two or even three changes. More sophisticated models

for amino acid sequence evolution account for these differences by weighting each amino acid change with its observed rate of change across a large number of proteins (Dayhoff 1972; Jones et al. 1992).

Rates of Nucleotide Substitution

Nucleotide sequences are analyzed in the same manner as amino acid sequences, but the analog of Equation 7.1 is slightly more complicated because it has to correct for cases in which a substitution makes two previously different nucleotide sites identical. The correction is significant for nucleotide sequences because an expected one-third of random substitutions will make two previously different nucleotides identical. The correction is usually unnecessary for proteins because only $\frac{1}{19}$ of random replacements make two previously different amino acids identical.

Several models of nucleotide substitution have been studied, which differ primarily in the assumptions about rates of mutation between pairs of nucleotides. The simplest model is the **Jukes-Cantor model** in which mutations occur at a constant rate, and each nucleotide is equally likely to mutate to any other (Jukes and Cantor 1969). If α is the rate of mutating from one nucleotide to a different nucleotide, then in any time interval, A mutates to C with probability α, A mutates to T with probability α, and A mutates to G with probability α. The probability that A does not change in this interval is therefore $1 - 3\alpha$. The probability that a particular site is A at time $t + 1$ is

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha(1 - P_{A(t)}) \tag{7.8}$$

because the first part of the equation gives the probability of having been A at time t and not mutating, and the second part is the probability of being any other nucleotide and mutating to A. From Equation 7.8 it follows that

$$P_{A(t+1)} - P_{A(t)} = dP_{A(t)}/dt = -4\alpha\, P_{A(t)} + \alpha \tag{7.9}$$

Solving this differential equation,

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{3}{4}\, e^{-4\alpha t} \tag{7.10}$$

assuming that the initial state was A. Equation 7.10 gives the transition probability from A to A in time t, which we can write as P_{AA}. If we observe two sequences that have been separated for time t, then the probability that they continue to carry the same nucleotide (N) at a particular site is

$$P_{NN} = \tfrac{1}{4} + \tfrac{3}{4}\, e^{-8\alpha t} \tag{7.11}$$

because $2t$ is the total duration of time along both lineages during which changes could occur. Let d be the proportion of nucleotide sites that differ between two sequences:

$$d = 1 - P_{NN} \tag{7.12}$$

and so

$$\hat{d} = \tfrac{3}{4}(1 - e^{-8\alpha t}) \tag{7.13}$$

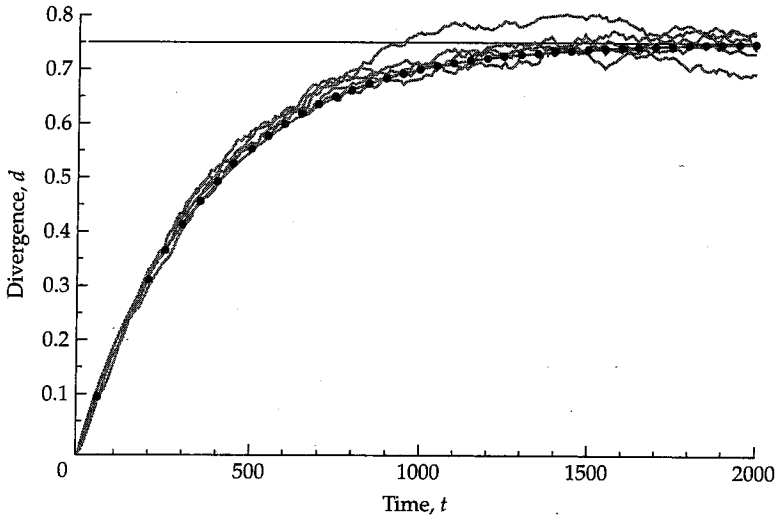


FIGURE 7.5 Simulations of the substitution process for nucleotide sequences show that the sequence divergence saturates at $d = 0.75$ assuming that A, G, C, and T are equally abundant. The jagged lines are numerical simulations of a sequence of length 1000, and the dots give the prediction under the Jukes-Cantor model.

Now let λ be the rate of mutation to a nucleotide different from the current nucleotide, so that $\lambda = 3\alpha$. This relation implies that $k = 2\lambda t = 2(3\alpha)t = 6\alpha t$. Taking logarithms of both sides of Equation 7.13, we deduce that

$$8\alpha t = -\ln(1 - 4\hat{d}/3) \tag{7.14}$$

and, since $k = \frac{3}{4}(8\alpha t)$,

$$\hat{k} = -\tfrac{3}{4}\ln(1 - 4\hat{d}/3) \tag{7.15}$$

where k is the expected proportion of nucleotide sites that differ between two sequences at a time t units after their evolutionary separation. By analogy with protein evolution, $\hat{d}$ is the observed proportion of L nucleotide sites in which the sequences differ. The variance $Var(k)$ of the estimate can be estimated as

$$Var(\hat{k}) = \hat{d}(1 - \hat{d})/[L(1 - 4\hat{d}/3)^2] \tag{7.16}$$

Figure 7.5 depicts the relationship between time and d, and shows that nucleotide sequences that follow the Jukes-Cantor pattern of mutation, in which all nucleotides are equally interchangeable, approach an asymptote with a divergence of $\frac{3}{4}$. This fraction makes intuitive sense because, after sufficient time, the common ancestry of the sequences has been erased, and $\frac{1}{4}$ of the sites will match by chance.

PROBLEM 7.3. The sequences below compare part of the coding region of the gene encoding glucose-6-phosphate dehydrogenase in the intestinal bacterium *Escherichia coli* (*Ec*) and the plague bacillus *Yersinia pestis* (*Yp*). Estimate the amount of amino acid divergence $\hat{K}$ and nucleotide divergence $\hat{k}$ and their standard deviations.

Ec	GCT	GAC	TGG	GAT	AAA	GCG	GCA	TAT	ACC	AAA	GTT	GTC	CGC	GAG	GCG	CTC	GAA	ACT	TTC	ATG
	Ala	Asp	Trp	Asp	Lys	Ala	Ala	Tyr	Thr	Lys	Val	Val	Arg	Glu	Ala	Leu	Glu	Thr	Phe	Met
Yp	GCT	GAG	TGG	GAT	AAA	GAT	GCG	TAC	ACC	GCT	GTG	GTA	AAG	GAA	GCC	CTC	GAT	ACC	TTT	ATG
	Ala	Glu	Trp	Asp	Lys	Asp	Ala	Tyr	Thr	Ala	Val	Val	Lys	Glu	Ala	Leu	Asp	Thr	Phe	Met

Answer For the amino acid sequences, $L = 20$ and $\hat{D} = \frac{5}{20} = 0.25$; thus $\hat{K} = -\ln(0.75) = 0.288$ with standard deviation 0.142. For the nucleotide sequences, $L = 60$ and $\hat{d} = \frac{18}{60} = 0.30$; thus $\hat{k} = -(\frac{3}{4})\ln(0.60) = 0.383$ with standard deviation 0.099.

The Jukes-Cantor model assumes that all possible nucleotide changes are equally probable. In fact, it is generally observed from sequence comparisons that **transitions**, or changes either from purine to purine (G ↔ A) or from pyrimidine to pyrimidine (C ↔ T) are more frequent than **transversions** (any other possible change). Kimura (1980a) sought to accommodate this observation by making a model with two mutation-rate parameters. Transitions occur with rate α and transversions occur with rate β. The rate matrix for the Kimura two-parameter model is:

		Ending base			
		A	C	G	T
Initial base	A	–	β	α	β
	C	β	–	β	α
	G	α	β	–	β
	T	β	α	β	–

As you might guess, other models can also be specified by adding parameters to this table. One of the more popular is the "HKY" model, which allows for four unequal rates of substitution (Hasegawa, Kishino, and Yano 1985). The general reversible model has six parameters and is the most general case for a model that is time-reversible (Tavaré 1986). It is even possible to fit the full 12-free-parameters model to data, although it is rare that this most complex model yields a significant improvement over simpler models. These mutation models can be fitted to the data in a variety of ways, such as with solutions like those we saw for the Jukes-Cantor model as well as with more complex numerical methods.

Statistical Fitting of Nucleotide Substitution Models

Equation 7.15 represents the simplest case of estimation of the nucleotide substitution rate from observed sequence data. Real data present many additional complications, including missing data, multiple species (with or without specified phylogenetic tree relations), or additional parameters allowing more classes of substitutions.

If the phylogeny of the organisms in the data set is known, it is possible to calculate the likelihood of the observed sequences, given the phylogeny and the parameters in the model (Felsenstein 1981). Many advances have been made in recent years in applying the method of maximum likelihood for estimating parameters of the substitution process in this context (Goldman 1993; Yang 1996a; Yang and Nielsen 1998, 2000, 2002). Software now exists that makes these estimates easily, including the packages PAML, DNAsp, HyPHY, and MEGA, all of which are freely available. Testing the goodness-of-fit of substitution models and comparison of fits across models are important problems also solved by several computer packages (Pond et al. 2005; Posada 2006).

In some cases, the data that are available may not be the direct nucleotide sequences but rather an indirect inference of divergence at the DNA level, such as estimates based on differences in the sizes of restriction fragments. So long as a statistical model can be devised that relates the magnitude of DNA sequence divergence to the difference in observed patterns of restriction fragments, it is possible to estimate nucleotide divergence (see Problem 1.4, for example). Other kinds of molecular data that allow estimation of nucleotide polymorphism and divergence include microsatellite variability (Goldstein et al. 1995), randomly amplified polymorphic DNA (RAPD) (Clark and Lanigan 1993), and VNTR loci (loci that are polymorphic due to variable numbers of tandem repeats) (Shriver et al. 1995). There are also methods for estimation of divergence from insertion-deletion data (Ogurtsov et al. 2004).

7.3 THE MOLECULAR CLOCK

Although the rate of nucleotide substitution and amino acid replacement varies among different genes, the average rate of molecular evolution can be rather uniform throughout long periods of evolutionary time. Such uniformity in the rate of amino acid replacement or nucleotide substitution, first noted by Zuckerkandl and Pauling (1962), is known as a **molecular clock**.

An example of the approximate uniformity in amino acid substitutions is illustrated in the evolution of the α-globin gene in a series of vertebrates summarized in Table 7.1. The numbers above the diagonal are the percent amino acid differences ($\hat{D} \times 100$) between the α-globin sequences. For example, the α-globin genes of dog and human differ in 16.3% of their amino acid sites; since mammalian α-globin contains 141 amino acids, this percentage corresponds to 23 sites in which the amino acids differ. The percentages

TABLE 7.1 Rate of Evolution in the α-Globin Gene

	Shark	Carp	Newt	Chicken	Echidna	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp	**0.90**		53.2	51.4	53.6	50.7	47.9	48.6
Newt	**0.95**	**0.76**		44.7	50.4	47.5	46.1	44.0
Chicken	**0.91**	**0.72**	**0.59**		34.0	29.1	31.2	24.8
Echidna	**0.93**	**0.77**	**0.70**	**0.42**		34.8	29.8	26.2
Kang	**0.81**	**0.71**	**0.64**	**0.34**	**0.43**		23.4	19.1
Dog	**0.84**	**0.65**	**0.62**	**0.37**	**0.35**	**0.27**		16.3
Human	**0.76**	**0.67**	**0.58**	**0.28**	**0.30**	**0.21**	**.018**	
Avg $\hat{K}$	0.87	0.71	0.63	0.35	0.36	0.24	0.18	
Time	450	410	360	290	225	135	80	

(Percentage data from Kimura 1983.)

Note: Values above the diagonal are the observed percent amino acid differences (D) between the α-globin sequences in the species, values in **boldface** are the expected amino acid differences per site [$\hat{K} = -\ln(1 - D)$]. Average values of $\hat{K}$ and the estimated times of divergence (in millions of years) are given at the bottom of the table. Abbreviation: Kang, kangaroo.

exclude differences that result from the insertion or deletion of amino acids, which are called **gaps** in sequence comparisons. For example, the comparison between human and shark α-globin is based on 139 amino acid sites that are homologous, and excludes gaps amounting to 11 additional amino acid sites. Missing from Table 7.1 are plants, which have leghemoglobin, a protein with significant homology to vertebrate globins that is essential for support of nitrogen-fixing bacteria (Ott et al. 2005).

Beneath the diagonal in Table 7.1 are the estimated proportions of differences per amino acid site, calculated from Equation 7.5 as $\hat{K} = -\ln(1 - \hat{D})$. The table also gives the average value of $\hat{K}$ in all comparisons with the shark, carp, newt, chicken, echidna, kangaroo, and dog, respectively, and the divergence times estimated by a variety of methods.

The average proportion of differences per site is plotted against divergence time in Figure 7.6. The very close fit to a straight line is evident. Since the divergence time is exactly half of the total time available for evolution, the rate of evolution λ can be estimated as one-half times the slope of the line in Figure 7.6. For these data, the slope is 1.8×10^{-9}, and therefore $\hat{K} = 0.9 \times 10^{-9}$ amino acid replacements per amino acid site per year. The good fit of the points to the straight line indicates that the actual rate of α-globin evolution has deviated little from the average for the past 450 million years.

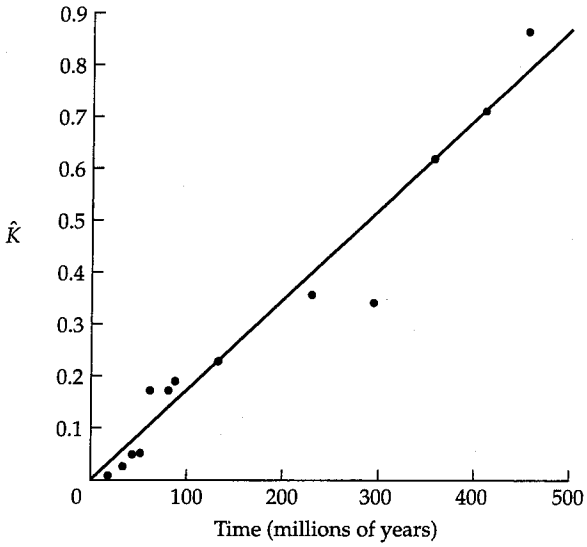


FIGURE 7.6 Relation between estimated number of amino acid substitutions in α-globin ($\hat{K}$) between pairs of the vertebrate species, against time since each pair diverged from a common ancestor. The straight line is expected based on a uniform rate of amino acid substitution during the entire period. (From Kimura 1983.)

PROBLEM 7.4 The β-globin molecule in primates contains 146 amino acids, and estimates of the number of amino acid differences among various primates are tabulated below (data from Kimura 1983). Calculate the average rate of evolution of β-globin molecule in primates. (*Hint:* First calculate $\hat{D}$ and $\hat{K}$ for each species pair, and then plot the points with time on the x-axis and $\hat{D}$ on the y-axis. Finally, find the slope of the linear regression to estimate the average rate of substitution. (*Note*: students unfamiliar with linear regression will find the equation for the slope in Equation 8.6.)

Time of divergence (millions of years)	Average number of amino acid differences
85	25.5
60	24.0
42	6.25
40	6.0
30	2.5
15	1.0

ANSWER $\hat{D}$ values are obtained by dividing each number of amino acid differences by 146, and average values of $\hat{K}$ are estimated as $-\ln(1 - \hat{D})$. The average $\hat{K}$ values, from top to bottom, are 0.192, 0.180, 0.044, 0.042, 0.018, 0.007, respectively. These are the y values in the linear regression, and the x values are the divergence times. Altogether there are $n = 6$ points. In this case, $\Sigma xy = 3.1263 \times 10^7$, $\Sigma x = 2.72 \times 10^8$, $\Sigma y = 0.482$, and $\Sigma x^2 = 1.5314 \times 10^{16}$. The slope of the regression is 3.15×10^{-9}, and the rate of evolution is half of this, or 1.58×10^{-9} amino acid replacements per amino acid site per year. This estimate is reasonably close to the value of 0.9×10^{-9} per year calculated for α-globin. (*Note:* Rather than calculate $\hat{K}$ from the average number of amino acid differences, it would be more accurate to calculate $\hat{K}$ for each species comparison and then take the average; however, in this example, it makes very little difference.)

Variation among Genes in the Rate of the Molecular Clock

If an organism has a particular rate of mutation in its genome, one might think at first that the rate at which the molecular clock runs would be the same for all genes. But the neutral theory predicts that the rate of molecular evolution should depend on the *neutral* mutation rate, which may be quite a bit lower than the overall mutation rate, and may vary widely across genes. Figure 7.7 shows that three different proteins in the same organisms have widely differing molecular clock rates. Nevertheless, within each gene, we observe reasonably uniform rates of change. The variation across genes

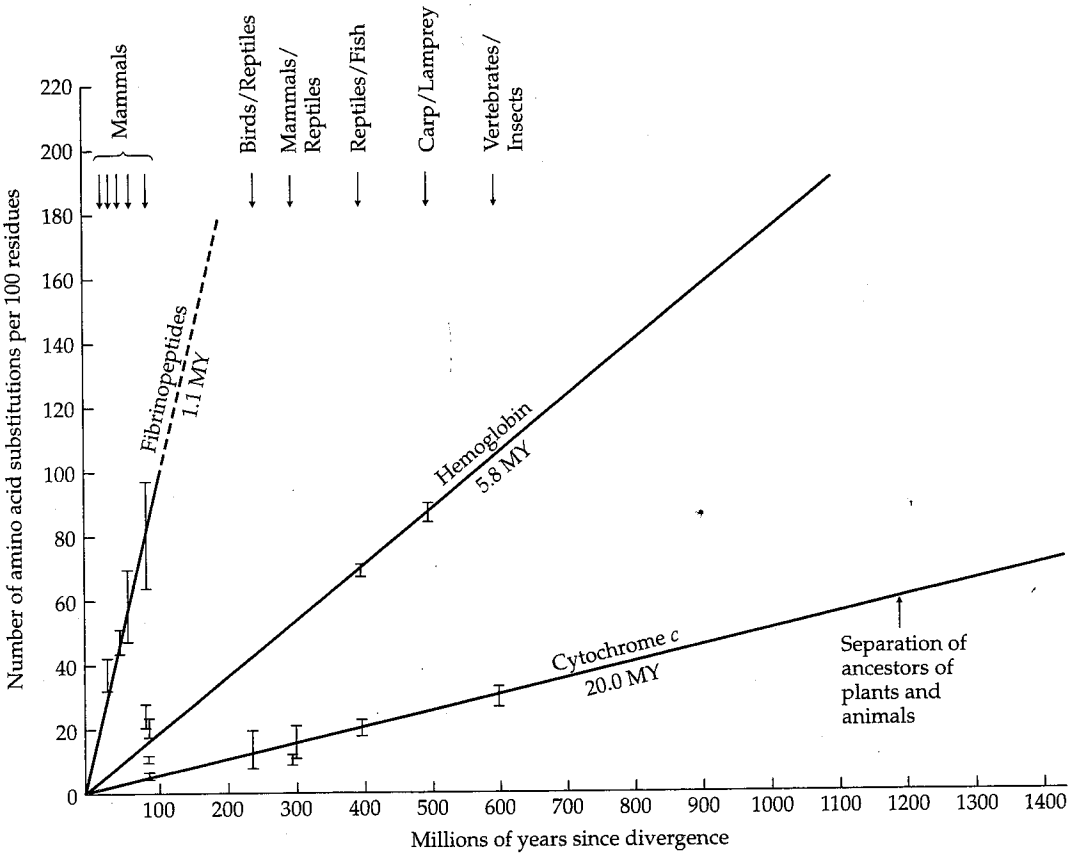


FIGURE 7.7 The molecular clock runs at different rates in different proteins. One reason is that the substitution rate differs among proteins. Fibrinogen appears to be relatively unconstrained and has a high substitution rate, while cytochrome *c* has a lower substitution rate and may be more constrained. Data are from a wide variety of organisms. (From Dickerson 1971.)

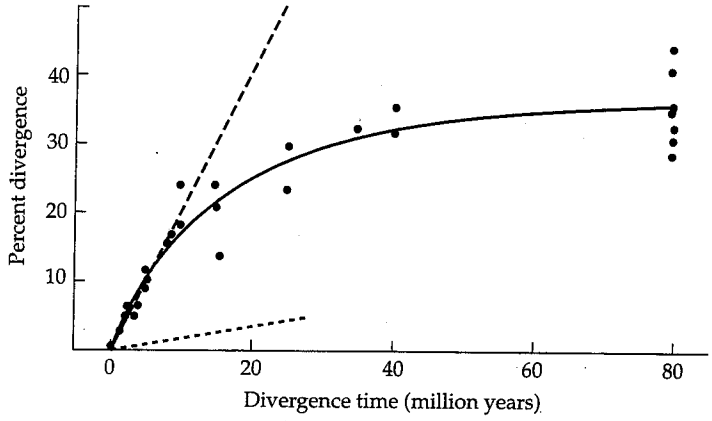


FIGURE 7.8 Relationship between percent of sequence divergence (100*d*) and divergence time for mtDNA restriction sites. The points represent estimates from pairwise comparisons of restriction endonuclease cleavage maps. The initial rate of mtDNA sequence is shown by the longer dashed line and the rate of divergence of single-copy nuclear DNA by the shorter dashed line. (From Brown et al. 1979.)

appears to be due to the fact that some proteins are highly tolerant of substitutions, whereas others suffer deleterious effects from even one or a few minor changes. Genes whose function is well buffered from the environment generally have a slower rate of substitution than genes whose products have a premium on variability. The extremes are represented by histone H4, at the low end, and γ-interferon, at the high end, with globin proteins near the middle of the spectrum. In short, the molecular clocks for different genes "tick" at different rates.

In addition to functional constraints affecting substitution rate, the pattern of hereditary transmission also affects substitution rate. Organelle genomes are replicated and transmitted in a manner distinct from nuclear genes, so it may not be surprising that they undergo substitutions with different dynamics. Mitochondrial DNA exhibits wide variation in substitution rates across its relatively tiny genome, but in animals the substitution rate is generally much higher than the substitution rate of chromosomal genes (Figure 7.8). In plants, on the other hand, comparisons among nucleotide substitution rates of nuclear DNA, chloroplasts, and mitochondria reveal clear differences, with mtDNA showing less than one-third the substitution rate of chloroplast DNA, which in turn has about half the substitution rate of nuclear genes (Wolfe et al. 1987; Chaw et al. 2004).

Not only do substitution rates vary from one gene to another, but they also vary widely across sites within each gene. If all sites did undergo sub-

stitution at the same rate, then the number of substitutions per site should have a Poisson distribution. Fitch and Margoliash (1967) noticed that the cytochrome *c* data did not fit this model unless invariant (unchanging) and hypervariable (rapidly changing) sites were excluded. Alternatively spliced exons whose splicing status is not conserved across species have a higher rate of nucleotide substitution (Plass and Eyras 2006). The models that we have developed so far assume that all sites evolve in the same way, so to accommodate the observed variability (and to test for how different the rates are) models that specifically incorporate variation in rates must be developed. One convenient model is to assume that the rates vary according to a gamma distribution (Golding 1983; Wakeley 1993). Yang (1996b) reviewed the estimates of the rate-variation parameter of the gamma distribution, and found that all 17 cases he examined show significant among-site variation in substitution rate.

A specific class of among-site variability occurs when the substitution rate at a site depends on the state of other sites. Several models for estimating substitution rates in the face of this context dependence have been devised, and they demonstrate clearly that nucleotides do not evolve independent of one another (Siepel and Haussler 2004b; Hwang and Green 2004). Perhaps the most obvious example is the nearly tenfold higher substitution rate for 5′-CpG-3′ dinucleotides in mammalian genomes, a rate increase caused by methylation and deamidation of methylated cytosines adjacent to guanines (Youssoufian et al. 1986).

In principle the molecular clock should apply to intergenic (noncoding) regions of the genome as well as to genes themselves. When noncoding regions of human and mouse chromosome 21 were aligned, it was immediately apparent that there are segments of noncoding sequence that have an exceptionally slow substitution rate (Dermitzakis et al. 2002) and that these slow rates persist even when multiple species are compared (Dermitzakis et al. 2003). A powerful test for exceptions to the molecular clock is to use the phylogeny to make inferences along each branch and to fit a hidden Markov model to scan along the sequence. Such phyloHMM models have identified many highly conserved segments, which almost certainly have some conserved function across groups of yeast, flies, and mammals (Siepel et al. 2005). In this application, the molecular clock was used to identify functionally conserved segments of the genome.

Variation across Lineages in Clock Rate

The neutral theory predicts that the rate of the molecular clock should run at different rates for different organisms having different neutral mutation rates. The range of mutation rates is impressive. Figure 7.9 shows the number of nucleotide differences observed in the influenza *NS* genes, plotted against the year of isolation of the virus containing them. The rate of gene substitution averages $\hat{\lambda} = 1.94 \pm 0.09 \times 10^{-3}$ nucleotide substitutions per nucleotide

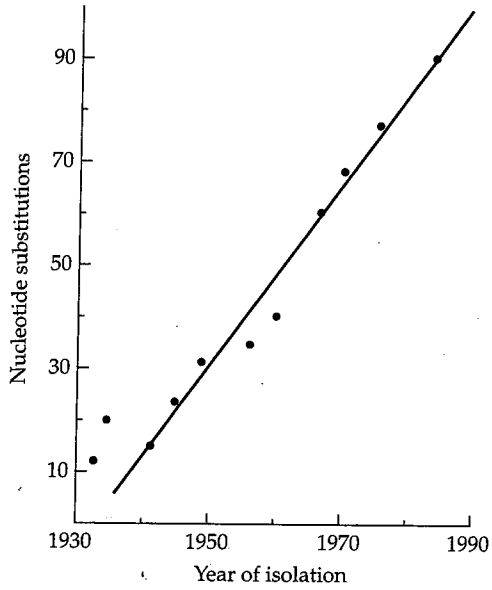


FIGURE 7.9 Molecular evolution in the *NS* genes of influenza virus determined from strains isolated and stored during the past 60 years. The total rate of evolution in the 890-nucleotide sequence averages 1.73 ± 0.08 nucleotide substitutions per year, and the rate is remarkably uniform. (From Buonagurio et al. 1986.)

site per year. Although the rate of gene substitution is about 10^6-fold faster than observed in germline genes in eukaryotes, it is nevertheless approximately constant during the period available for study. The extraordinary rate of evolution in influenza virus is thought to be related to a high rate of spontaneous mutation resulting from errors in replication (Gojobori et al. 1990). As in many other RNA-based viruses, the RNA replicase enzyme that replicates the influenza genome lacks a proofreading function. Rapid rates of gene substitution can be of immense medical significance. Yokoyama et al. (1988) estimated the rate of substitution in the *pol* gene of the human immunodeficiency virus as 0.5×10^{-3} per nucleotide site per year. The time of divergence between HIV1 and HIV2 was estimated at just 200 years ago, and the bulk of the genetic variability among recently isolated strains of HIV1 has been generated in the last 20 years. By following sequences of the Pol protein across nine transmissions from person to person, Lemey et al. (2005) identified a clear increased fixation rate of mutations in the HIV genome of a multidrug-resistant patient, demonstrating a remarkably fine-grained scale to the heterogeneity in substitution rates.

The rate of the molecular clock also varies among taxonomic groups. For example, the insulin gene evolved much more rapidly in the evolutionary line leading to the guinea pig than in other evolutionary lines (King and Jukes 1969), and the C-type viral sequences integrated into the primate genome evolved at twice the rate in Asian primates as in African primates (Benveniste 1985). Figure 7.10 illustrates another example of a retardation in

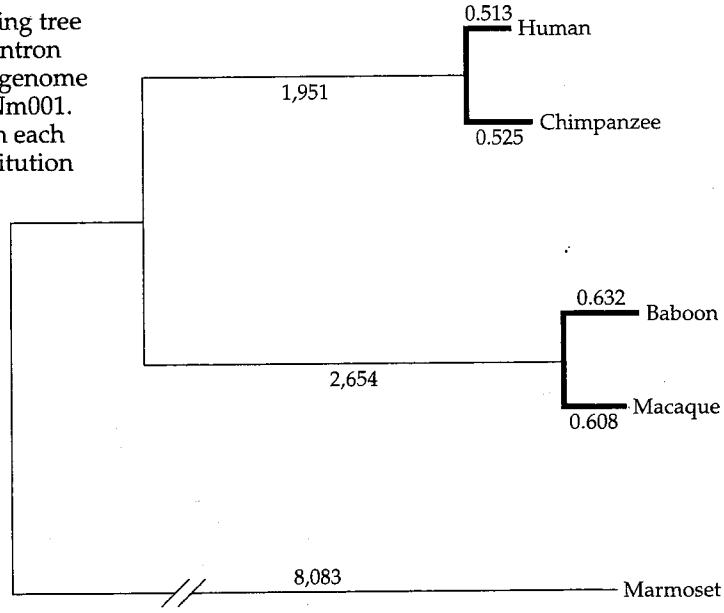


FIGURE 7.10 Neighbor-joining tree inferred from noncoding and intron sequences in a segment of the genome known as ENCODE region ENm001. Numbers of sites are shown on each long branch and rates of substitution per 100 sites on the shorter branches. Noncoding and intron regions show a clear slowing of the rate of substitution in humans and chimps. (From Kim et al. 2006.)

the clock in the hominid lineage. Through an analysis of 28 Mb of sequence from primates, Kim et al. (2006) showed that the molecular clock for sites in excluding CpG dinucleotides show strong generation-time dependency (36% slowdown in hominids compared to baboon), indicating that most of the associated mutations occur in meiosis, but the CpG sites themselves undergo substitutions in a more uniform fashion. Such departures from constancy of the clock pose a problem in using molecular divergence to date the times of existence of most recent common ancestors. Before this inference can be justified, one needs to know that the species have a uniform molecular clock.

PROBLEM 7.5 The simplest way to test whether substitutions have occurred at the same rate in different organisms is to consider a tree like that in Figure 7.11. We expect that the divergence between A and C should be the same as the divergence between B and C if the clock is uniform on all branches. Tests of this hypothesis are known as **relative rate tests**. Any site that underwent a substitution along the branch from X to C (but not on the other branches) will have the property that A = B ≠ C. Sites that underwent a substitution on the branch from X to B (but not the other branches) will show A = C ≠ B. Tajima (1993) showed that a simple and robust relative rate test could be performed by simply doing a chi-square test of the null hypothesis that the numbers of these two kinds of sites are equal. Suppose we observe sequences as follows:

A	ATG	CTA	GCA	TGC	ATG	CTA	GC
B	ATC	CTA	GCA	TCC	ATG	GTA	GT
C	ATG	CTA	TCA	TGC	TTG	GTA	GC

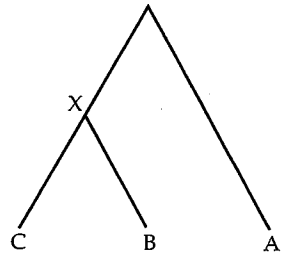


FIGURE 7.11 This tree with three sequences (A, B, C) illustrates the relative rate test of Tajima (1993).

Calculate the observed and expected numbers of sites in the two categories (A = B ≠ C and A = C ≠ B), and calculate the chi-square statistic to determine whether they are equal.

ANSWER The observed number of sites for which A = B ≠ C is 2, and for A = C ≠ B there are 3 sites. Sites where A = B = C or A ≠ B ≠ C are ignored in this test. The expected number of sites of the two types is each (2 + 3)/2, so the chi-square tests gives $(2 - 2.5)^2/2.5 + (3 - 2.5)^2/2.5 = 0.2$, which is clearly not significant. This example had insufficient data for an adequate test, but it provides an example in which there is no evidence for significant difference in rates. A more flexible but more involved test, based on maximum likelihood, can be found in Muse and Weir (1992).

The Generation-Time Effect

The theoretical derivation of the molecular clock makes clear that one expects a constant rate of substitutions when time is measured on a generation scale. If two organisms have a rate of neutral mutation of 10^{-6} per gene per generation, say, then the organism with a generation time of 5 years should have a molecular clock running four times as fast as the organism with a 20-year generation time. But the data show that the molecular clock rate is approximately constant over a time scale measured in *years*. There are exceptions; for example, rodents do have a faster clock than primates (Wu and Li 1985; Li and Wu 1987). Sequence data from 18 genes in 64 species showed clearly that the primate lineage has the slowest rate of nucleotide substitution among mammals (Murphy et al. 2001). Ohta (1994) examined this problem and found that synonymous sites had a clock that ran closer to the generation time scale, whereas nonsynonymous sites had a clock that ran closer to the per-year time scale. This is the observation that motivated her to ask what could cause this type of departure from strict neutrality and led her to develop the nearly neutral model.

Several other factors are relevant to the relatively weak dependence of the molecular clock rate on generation time. One reason is that organisms with short generation times tend to be small and to maintain large population sizes. In such organisms, the proportion of nearly neutral mutations will be

reduced because effective neutrality requires that $Ns << 1$, where s is the selection coefficient against the mutation. However, the smaller proportion of nearly neutral mutations in these organisms is offset against the occurrence of more mutations per unit time than in larger organisms, because the generation time is shorter. Thus, the effects of short generation time and larger population size act in opposite directions and tend to cancel out.

The Overdispersed Molecular Clock and the Neutral Theory

In a theoretically perfect molecular clock driven by a random process identical to that of radioactive decay (a Poisson process), the variance in the rate of ticking would be equal to the average rate of ticking. Tests based on the number of substitutions between pairs of species in three proteins showed that the variance was significantly larger than the mean (Ohta and Kimura 1971b). Langley and Fitch (1974) confirmed the larger variance by an analysis in which they estimated the number of substitutions on each branch of the phylogenetic tree and compared the mean and variance of these counts for each branch. Gillespie (1989) examined the ratio R of the variance to the mean number of substitutions in a set of four nuclear and five mitochondrial genes in mammals, and found that R ranged from 0.16 to 35.55. (The value of 35.55 is for cytochrome oxidase II, which shows 65 amino acid differences between human and mouse, 61 differences between human and cow, and only 21 differences between mouse and cow.) Gillespie argued that the large range of R (also known as the **dispersion index**) implies a sixfold difference among mammalian lineages in rates of nucleotide substitution. Even with the most heroic corrections for generation time, Ohta (1994) still found that, for both silent and replacement sites, the variance-to-mean ratio was greater than 5. This excess variance in substitution rate has been called an "overdispersed" or "episodic" clock, characterized by periods of stasis alternating with periods of rapid substitution.

Three factors can cause an inflation in the dispersion index in molecular clocks: fluctuations in rates that occur on a long time scale, advantageous or deleterious mutations, and interactions between mutations (Cutler 2000). Fluctuations in rates of the substitution process would result in a departure from a simple Poisson process. For example, the neutral mutation rate itself might change in a random or stochastic manner. Such a process, where the substitution rate for a Poisson process is itself stochastic, is called a **doubly stochastic process**, and it does indeed seem to fit the data better (Gillespie 1991). Such a compound Poisson process ought to show clusters of rapid change separated by periods of relative quiescence, a pattern that is generally supported by the data (Gingerich 1986; Gillespie 1989, 1991). Another means of causing variation in the substitution rate is natural selection in a stochastically varying environment, and such models can also fit the data satisfactorily (Gillespie 1986). Considering a constant environment, advantageous mutations would lower the dispersion index, assuming only some

constant probability that the next mutation will be favorable and sweep to fixation. Deleterious mutations increase the dispersion index, in part because normal fluctuations in the population size for alleles with $Ns \approx 1$ will inflate the variance: When the population size is low they will be effectively neutral, but when the population size is large, selection can act upon them as deleterious mutations. This effect on the index of dispersion works only over a fairly narrow range of Ns. Cutler (2000) showed that a plausible model of deleterious mutations is sufficient to explain the observed values of the dispersion index, assuming that most mutations are deleterious. But the models still require that $2 < Ns < 10$, which raises the question of why the index of dispersion appears to be reasonably constant across a huge range of population sizes. Viruses allow one to directly test the fit between known divergence times and molecular clock estimates, and it is clear that positive selection can be accommodated by simply identifying the subset of nucleotides showing non-clocklike behavior (Wrobel et al. 2006). Clearly, the strictly neutral model does poorly in predicting observations when compared with a model that can accommodate weak selection.

The Nearly Neutral Theory

The poor fit of some attributes of the data to the neutral theory is not a cause to reject it outright. There are actually two flavors of the theory, one which posits strictly neutral variation and the other of which allows for mildly deleterious mutations. The version allowing mildly deleterious mutations is called the **nearly neutral theory** and was developed largely by Tomoko Ohta. The nearly neutral theory considers the case where $|Ns| \approx 1$ and was discussed in Chapter 5. When $|Ns| \approx 1$, there is still an important role for random drift in affecting the allele frequency in the population; however, selection also plays a role. Because there may be a reasonably important proportion of mutations that fall in this class, they may influence the overall behavior of the molecular clock and also impact standing levels of variation in a population.

The effect of weak negative selection can produce an overdispersed clock, and so this is one way that the nearly neutral theory can explain what appears to be an empirical disagreement with the neutral theory. The fact that substitution rates seem to have a relatively weak dependence on generation time can also be partly explained by the nearly neutral theory. What about the apparent lack of dependence of heterozygosity on effective population size? Under strict neutrality, we expect $H = 4N\mu/(1 + 4N\mu)$, and so the strictly neutral expectation is a strong positive trend of increasing heterozygosity with increasing population size. The nearly neutral theory allows for a substantial fraction of deleterious mutations, but this fraction itself increases with population size because larger populations are successful at removing an ever larger portion of weakly deleterious alleles. The result is effectively a reduction in the neutral mutation rate as the population size gets larger. The

nearly neutral theory also has a positive correlation between heterozygosity and population size, but the heterozygosity reaches an asymptote at a level below 1 because of the reduced neutral mutation rate. Near neutrality therefore predicts a weaker correlation between population size and H, and so, again, the nearly neutral mutation model helps to shore up the neutral theory. One aspect we will not consider in detail is the fact that mutations may have fitness effects that depend on one another, and the role of epistatic interactions among weakly selected mutations is a problem of considerable current interest. Empirically it is clear that amino acid replacements within a protein have consequences on function that can be highly nonadditive (Lunzer et al. 2005; Weinreich et al. 2006).

7.4 PATTERNS OF NUCLEOTIDE AND AMINO ACID SUBSTITUTION

We have now seen several examples illustrating the general principle that nucleotide substitutions occur at a greater rate than amino acid replacements. The difference in rates, sometimes much greater than in these data, results from redundancy in the genetic code. As illustrated in Table 7.2, the codons for eight amino acids contain N (standing for any nucleotide) in their third position, seven terminate in Y (any pyrimidine, which means T or C), and five terminate in R (any purine, which means A or G). Coding sites containing an N are called **fourfold degenerate** sites because any of the four nucleotides will do, and those containing a Y or R are **twofold degenerate** sites (Li et al. 1985b). Because of degeneracies, nucleotides in a gene can change without affecting the amino acid sequence. These changes are called **synonymous** or **silent** nucleotide substitutions. Nucleotide substitutions that do change amino acids are **nonsynonymous substitutions**.

Calculating Synonymous and Nonsynonymous Substitution Rates

In calculations involving synonymous and nonsynonymous nucleotide sites, the total number of synonymous sites is calculated as the number of fourfold degenerate sites plus one-third of the number of twofold degenerate sites. The total number of nonsynonymous sites in a coding region is defined as the number of nondegenerate sites (nucleotides in which any change results in an amino acid substitution), plus two-thirds of the number of twofold degenerate sites (the latter because, with random mutation at twofold degenerate sites, two-thirds of the mutations are expected to result in amino acid changes). These conventions for counting degenerate sites are summarized in the note accompanying Table 7.2.

Estimates of synonymous and nonsynonymous substitution rates for a set of mammalian protein-coding are plotted in Figure 7.12. A striking observation is that the synonymous rates are generally much greater than the rates of

TABLE 7.2 Degeneracy in the Genetic Code

First nucleotide in codon	Second nucleotide in codon: *T*	*C*	*A*	*G*
T	TTY Phe TTR Leu	TCN Ser	TAY Tyr TAR Stop	TGY Cys TGA Stop TGG Trp
C	CTN Leu	CCN Pro	CAY His CAR Gln	CGN Arg
A	ATH Ile ATG Met	ACN Thr	AAY Asn AAR Lys	AGY Ser AGR Arg
G	GTN Val	GCN Ala	GAY Asp GAR Glu	GGN Gly

Note: In this representation of the standard genetic code, the symbol *N* stands for any nucleotide (T, C, A, or G), the symbol *Y* for any pyrimidine (T or C), and the symbol *R* for any purine (A or G). The *H* in the set of codons for isoleucine (Ile) stands for "not-G" (T, C, or A). Degeneracies are as follows: *N* represents a fourfold degenerate site, *Y* and *R* represent twofold degenerate sites. The *H* in the set of codons for isoleucine is considered as twofold degenerate, as are the first nucleotides in four leucine codons (TTA, TTG, CTA, and CTG) and four arginine codons (CGA, CGG, AGA, and AGG). All other nucleotides are nondegenerate.

substitution at nonsynonymous sites. These rates are scaled, so that if all mutations were equally likely to go to fixation, the rates would be equal. The depression in nonsynonymous substitution rate is interpreted as being caused by natural selection eliminating those changes that are deleterious. There also appears to be greater variability of nonsynonymous rates than there is in the synonymous rates, although even the latter vary by more than twofold. Figure 7.13 shows that, across a wide range of protein-coding genes in *Drosophila*, the two rates are correlated, suggesting that either the mutation rates vary from gene to gene or that the constraints on nonsynonymous sites are somehow correlated with those on synonymous sites. Figure 7.13 illustrates another important point. The relative rates of synonymous to nonsynonymous substitution differ widely among genes, and in some cases the nonsynonymous rate is the larger. Because natural selection must operate mostly on the nonsynonymous sites, the relative values of these two substi-

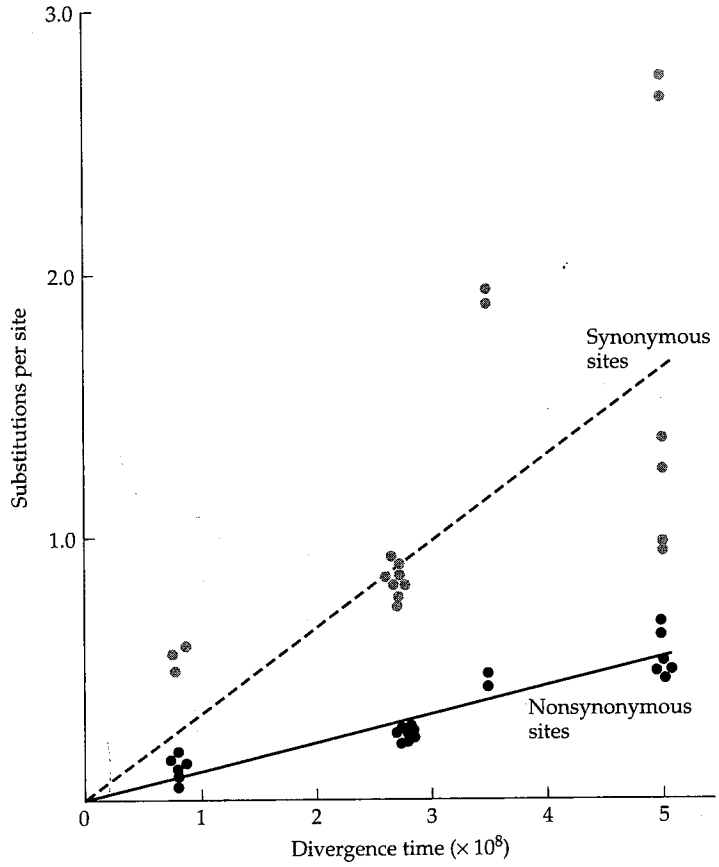


FIGURE 7.12 Synonymous sites and nonsynonymous sites in the β-globin gene undergo substitutions at different rates, but to a first approximation both may appear to exhibit a clocklike substitution process. (From Li et al. 1985a.)

tution rates may provide a key for making inferences about adaptive evolution of protein-coding genes.

One problem that may be apparent with the above method for counting synonymous and nonsynonymous sites is that the status of a particular site may change during evolution. The reason is that changes elsewhere in the codon may make a site that was formerly four-fold degenerate now become two-fold degenerate. In fact, the way the sites are tallied depends on the order in which they are considered. Another way to calculate nonsynonymous and synonymous substitution rates is to consider each codon and count the number of changes that occurred. For codons that changed at a single site, the change is scored as synonymous if there was no alteration in the resulting amino acid sequence, and nonsynonymous if there was an alteration. When there are two differences in a codon, then it is necessary to consider both orders of occurrence, and if we have no reason to assume one order is more likely, then both are considered equally likely (Nei and Gojo-

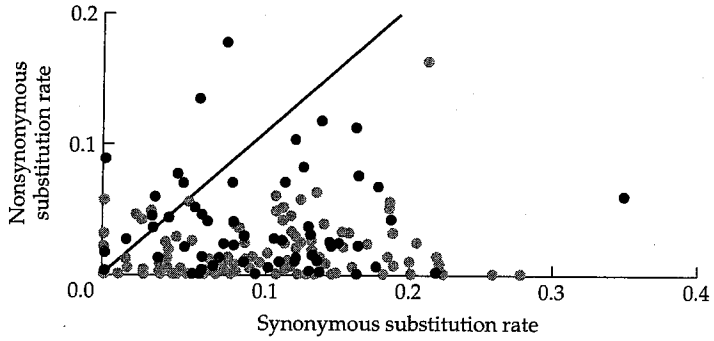


FIGURE 7.13 Rates of synonymous and nonsynonymous divergence between *D. melanogaster* and *D. simulans* for a set of genes expressed in the male accessory gland (black dots) and control genes (gray dots). (Data from Swanson et al. 2001.)

bori 1986). The two orders may have differing numbers of synonymous and nonsynonymous changes. For example, if a codon changes from CCG (proline) to AGG (arginine), it could have done so either through CCG → ACG (threonine) → AGG or through CCG → CGG(arginine) → AGG. The first possibility entails two nonsynonymous changes, whereas the second entails only one. If there are three changes in a codon, there are six possible orders in which they might have occurred. Counting the number of synonymous and nonsynonymous sites presents challenges over longer evolutionary distances, and the next section outlines the most statistically rigorous way to estimate these rates and to test hypothesis about the rates using sequence data from protein-coding genes.

Codon Substitution Models

Because the genetic code has a somewhat irregular structure of redundancy, there is no easy way to consider the status of individual sites without considering neighboring sites. A solution is to model the changes not in terms of single sites, but rather in terms of one codon substituting for another codon. Previously we considered a 4×4 matrix for the rates of the four nucleotides changing from one to another. With codons we need a 64×64 matrix to enumerate all possible changes. Because terminator codons are generally thought of as being inadmissible in paths of evolutionary change in a protein (apart from the one obligate terminator at the end), they are generally not included in the model, leaving a 61×61 matrix of nonterminating codons. Each cell in this matrix specifies the rate at which one codon changes to another. The codon substitution models make a key assumption that each mutation changes only a single nucleotide, and so any change like

AAA → AGC must result from two separate mutation and substitution events.

The matrix we are considering is the instantaneous rate matrix, in which any change involving more than 1 nucleotide has instantaneous rate of zero. All the remaining codon pairs either result in a change that retains the same amino acid and is hence synonymous, or it changes the amino acid and is hence nonsynonymous. The idea of modeling codons this way was first published nearly simultaneously by Muse and Gaut (1994) and by Goldman and Yang (1994). A particular parameterization by Yang and Nielsen (1998) has proven to be particularly useful. Let Q_{ij} be the instantaneous rate matrix specifying the rate of change from codon i to codon j:

$$\begin{aligned} Q_{ij} &= \mu\pi_j \text{ for a synonymous transversion} \\ &= \mu\kappa\pi_j \text{ for a synonymous transition} \\ &= \mu\omega\pi_j \text{ for a nonsynonymous transversion} \\ &= \mu\omega\kappa\pi_j \text{ for a nonsynonymous transition} \\ &= 0 \text{ otherwise.} \end{aligned}$$

where μ is a normalizing parameter, π_j is the equilibrium frequency of codon j, κ is the transition-transversion rate ratio, and ω is the ratio of rates of nonsynonymous to synonymous substitutions (d_N/d_S).

Given this parameterization of the model, the next step is to fit the model to an alignment of codons and estimate all the parameters. First note that the probability transition matrix $P(t)$ for codons can be obtained from the instantaneous rate matrix, Q, by letting $P(t) = e^{Qt}$. Given the alignment and the topology of the phylogenetic tree relating the different species in the alignment, we can write the likelihood of the data as a function of the parameters. Standard numerical methods can then be used to estimate the parameters given the data, and along the way this approach also delivers a likelihood score. A simple test for equality of synonymous versus nonsynonymous rates is to fit the model forcing $\omega = 1$ (the null model, having likelihood L_0) and to fit the model allowing ω to vary freely (the alternative hypothesis, with likelihood L_1), and then to calculate the ratio of the two likelihoods. Because the two models differ by one parameter, the statistic $-2[\log(L_1) - \log(L_0)]$ is asymptotically chi-square distributed. If this test statistic is greater than 3.84, the null hypothesis is rejected with probability 0.05. Codon substitution models have been extended in many ways, including allowance for radical and conservative amino acid changes, differences among synonymous codons, rate variation, context dependence, and so on. (Yang and Nielsen 2000, 2002; Yang et al. 2000, 2005; Nielsen and Yang 2003; J. Zhang et al. 2005). The calculations are rather involved, but freely available computer programs (e.g., PAML) can be employed. It is also possible to obtain posterior densities of all these model parameters by applying a full Bayesian framework (Huelsenbeck and Dyer 2004).

Observations of Synonymous and Nonsynonymous Substitution Rates

Paralleling the evolutionary rates for substitutions changing amino acids, the rates of nonsynonymous nucleotide substitution vary tremendously among different proteins. Among the slowest rates is that of histone H4, for which $\hat{k} = 0.004 \times 10^{-9}$ substitutions per nonsynonymous nucleotide site per year, and among the fastest is that of the anti-viral protein γ-interferon, for which $\hat{k} = 2.80 \times 10^{-9}$ substitutions per nonsynonymous nucleotide site per year. The average rate among a large number of proteins is very close to the rate found in hemoglobin, which is 0.87×10^{-9} substitutions per nonsynonymous nucleotide site per year (Figure 7.14). As in the examples given here, rates of nonsynonymous nucleotide substitution are usually quite similar to the rates of amino acid replacement in the same genes.

In contrast with the highly variable rates of nonsynonymous nucleotide substitutions among proteins, the rates of synonymous substitution are much more uniform. For example, in mammalian genes, the fastest rate of synonymous substitution is only 3 to 4 times greater than the slowest rate (see Figure 7.14). However, the average rate, $\hat{k} = 4.7 \times 10^{-9}$ substitutions per synonymous site per year, is not only greater than the average rate of nonsynonymous substitutions, but it is greater than the fastest known rate of nonsynonymous substitutions (for γ-interferon).

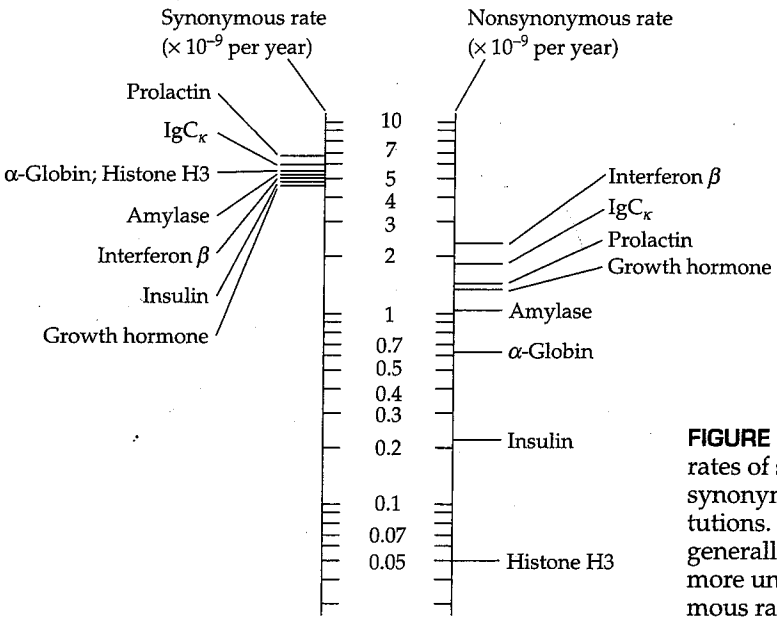


FIGURE 7.14 Comparison of rates of synonymous and nonsynonymous nucleotide substitutions. Synonymous rates are generally much faster and much more uniform than nonsynonymous rates. (From Kimura 1986.)

The great variability among proteins in the rate of nonsynonymous nucleotide substitution, when contrasted with the much smaller variability found in the rate of synonymous substitutions, is illustrated graphically in Figure 7.14. This disparity has been used as evidence in favor of the neutral theory. Interpreted according to the neutral theory, the variation in rates occurs because there are selective constraints on amino acid substitutions that do not operate as strongly on synonymous nucleotide substitutions. Not just any amino acid will serve at a particular position in a protein molecule, because each amino acid must participate in the chemical interactions that fold the molecule into its three-dimensional shape and give the molecule its specificity and ability to function. The need for proper chemical interactions and folding constrains the acceptable amino acids that can occupy each site. Although some amino acid replacements may be functionally equivalent or nearly equivalent, many more are expected to impair protein function to such an extent that they reduce the fitness of the organisms that contain them. Thus, the constraints on acceptable amino acids are **selective constraints** because unacceptable amino acid replacements are eliminated by selection.

If an amino acid replacement does occur, its effect on the function of the protein product will depend on many factors, but one of the most important determinants of protein conformation is the charge of the amino acid. Different amino acid replacements give different numbers of charge changes, and in most cases the smallest change in charge might be expected to result in the smallest conformational change. Peetz et al. (1986) examined the charge changes in the evolution of seven proteins, and found that hemoglobin α, hemoglobin β, myoglobin, and insulin all accumulated charge changes at a rate slower than expected by random substitution. This finding is consistent with constraints on the conformation of these proteins that limit permissible charge changes. On the other hand, cytochrome *c* and fibrinogens A and B accumulate charge changes at the expected neutral rate.

For comparison of rates, it would be useful to study rates of nucleotide substitution in stretches of DNA wholly devoid of function and therefore subject exclusively to the whims of mutation and random drift. A likely candidate is found in a class of genes called **pseudogenes**, which are DNA sequences that are homologous to known genes but that have undergone one or more mutations eliminating their ability to be expressed. Pseudogenes are thought to be completely nonfunctional relics of mutational inactivation, and, in fact, their extremely rapid rate of nucleotide substitution is offered in support of this view. The average rate of nucleotide substitution in pseudogenes is faster than the average rate found in intervening sequences, flanking regions, and fourfold degenerate (synonymous) sites. Pseudogenes evolve at the fastest rates known, which may correspond to rates of substitution when DNA is completely unconstrained by natural selection. The fact that fourfold degenerate sites evolve more slowly than pseudogenes may be a suggestion

fourfold degenerate sites

that these sites are not totally lacking in constraint, an idea we shall return to shortly.

Rates of nucleotide substitution also vary within protein molecules. Human insulin is a good illustration. The A and B polypeptide chains found in the mature insulin molecule are created by post-translational cleavage of a longer polypeptide known as preproinsulin. Preproinsulin contains a signal peptide for secretion and an internal C-peptide, neither of which are present in the active molecule. The relative rates of nucleotide substitution in these three regions are 0.16 for the A and B chains, 0.99 for the C peptide, and 1.16 for the signal peptide. [As in Li et al. (1985b), rates are expressed in terms of nonsynonymous nucleotide substitutions per nonsynonymous site per billion years.] In insulin, while there is a sevenfold difference between the maximum and minimum rates of nonsynonymous substitution in different regions of the molecule, the rates of synonymous substitution differ only twofold. Moreover, there is a negative correlation between functional importance and rate of nonsynonymous substitution within the insulin molecule. Many diverse amino acid sequences can serve as signal peptides provided they are hydrophobic, which suggests that selective constraints on signal peptides may be reduced in comparison with sequences in mature polypeptides. In insulin, as expected, the rate of nonsynonymous substitution is fastest in the signal peptide and slowest in the functional subunits of the mature molecule. This kind of negative correlation between selective constraint and substitution rate has also been observed in thousands of genes by several genome-wide analyses of their substitution rates (Clark et al. 2003; Nielsen et al. 2005b; Richards et al. 2005).

Polymorphism within Species

So far we have talked only about differences between nucleotide sequences of genes from distinct species. DNA sequence differences between alternative alleles of the same gene in a single species may also be synonymous or nonsynonymous, and it is instructive to compare levels of within-species polymorphism at synonymous versus nonsynonymous sites. In this case we do not generally talk about substitution rate, but rather quantify the variability with the nucleotide diversity. *Nucleotide diversity*, often symbolized with the Greek letter π, is the probability that a sample of a particular nucleotide site drawn from two individuals will differ. It is essentially the heterozygosity at the nucleotide level.

The first systematic study of DNA sequence variation within a population generated a set of 11 alcohol dehydrogenase allele sequences from *Drosophila melanogaster* (Kreitman 1983). Of the 2659 nucleotides sequenced, 52 were variable across the 11 alleles. The nucleotide diversity over the entire gene was 0.0065 ± 0.0017, meaning that 99.4% of the time, pairs of alleles matched at any particular site. The level of nucleotide diversity differs in different regions of genes. Figure 7.15 illustrates the estimates of nucleotide

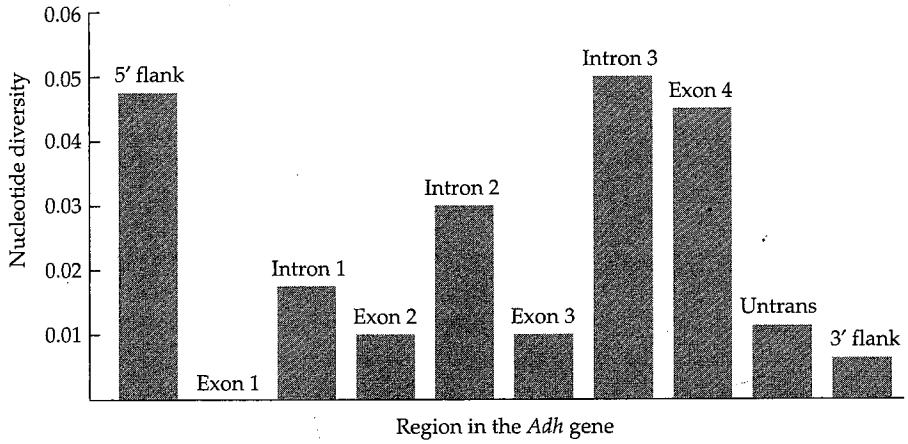


FIGURE 7.15 Nucleotide diversity for different functional components of the *Adh* gene in *Drosophila melanogaster.*

diversity found in different parts of the *Drosophila Adh* gene. The different parts are the 5′ (upstream) flanking region, the 5′ transcribed but untranslated region, the coding region (nonsynonymous substitutions only, with both the slowest and fastest rates shown), intervening sequences, the 3′ (downstream) transcribed but untranslated region, and the 3′ untranscribed region. On the average, the fastest rates of substitution occur in intervening sequences and the 3′ flanking regions, but the average rates in the 5′ flanking regions and the 3′ untranslated region are all substantially faster than 0.88×10^{-9}, which is the average rate of nonsynonymous substitution in coding regions (see Figure 7.15). Adherents to the neutral theory would argue that the high rates of substitution in noncoding regions and variation among different parts of the coding region result from varying degrees of selective constraints on different parts of the gene. It is to be emphasized that Figure 7.15 depicts the results for just one gene, and in individual instances, especially in comparisons of closely related species, there may be fewer substitutions observed in flanking sequences than in coding sequences, or fewer changes in synonymous sites than in nonsynonymous sites.

Comparison of nucleotide diversity in different functional regions of a single gene can reveal features of the gene's evolutionary history. For example, in 11 sequenced *Adh* genes of *Drosophila melanogaster*, among 14 substitutions that were observed in the coding region, 13 were silent substitutions (Kreitman 1983). Considering the genetic code and the codon usage in the *Adh* gene, we can calculate what portion of the substitutions would be silent if all substitutions occurred with equal frequency. This figure is about 30% in the case of the *Adh* gene in *Drosophila*, implying that about 70% of the substitutions would be expected to cause amino acid replacements. Since only one

out of 14 observed substitutions was an amino acid replacement, such substitutions are greatly underrepresented. This finding is consistent with the view that most amino acid replacements are eliminated from the population by purifying selection. The same logic can be extended to argue that sequences that are conserved are likely to be functionally important; this type of reasoning has been widely applied in algorithms that seek to identify genes from alignments of sequences.

The action of natural selection can sometimes be inferred from levels of synonymous and nonsynonymous polymorphism. For genes that determine surface antigens of pathogens or those that determine the major histocompatibility antigens of mammalian cells, the rates of nucleotide substitution can be quite high. One way to address whether the high rate of substitution is driven by selection is to examine the levels of synonymous and nonsynonymous diversity in these genes. For example, Hughes and Nei (1988) found that in the regions coding for the antigen recognition sites in the class I *MHC* (major histocompatibility complex) genes of humans and mice, the rate of nonsynonymous substitution exceeded the rate of synonymous substitution by a ratio of 3 : 1. This ratio is the reverse of that found in the usual situation and in other regions in the same genes, where silent substitutions are present in excess. The excess of amino acid replacements is consistent with a model in which mutations that generate diversity are often advantageous, and hence natural selection accelerates the substitution process. Just as the host defense molecules show excess nonsynonymous polymorphism, the merozoite surface protein-2 gene of the malarial parasite *Plasmodium falciparum* also shows a large excess of nonsynoymous polymorphism (Ferreira and Hartl 2006). With whole genome sequences of related species, several groups have performed comparisons of rates of substitution and synonymous and nonsynonymous sites specifically to find genes that appear to have faced positive selection (Clark et al. 2003; Nielsen et al. 2005b; see Chapter 9). Clustering of genes by functional categories shows a clear trend for immune defense genes and genes expressed in the testis to show signatures of positive selection. High rates of nonsynonymous substitution are also found in protein toxins called colicins that certain bacteria produce to kill potential competitors in their immediate vicinity (Riley and Wertz 2002).

Implications of Codon Usage Bias

Synonymous substitutions occur at a greater rate than nonsynonymous substitutions, implying that they face weaker selective constraints. But are synonymous changes completely neutral, or do they too face some form of constraint? One potential type of constraint occurs through codon preferences, which are correlated with the relative abundance of tRNA molecules that interact with and translate the codons. In bacteria and yeast, for example, highly abundant proteins tend to use codons for abundant tRNA molecules, whereas proteins produced in small amounts tend toward codons for

less abundant tRNA molecules (Ikemura 1985). A plot of the frequency of use of the synonymous codons that code for leucine shows that CUG is much more frequent than the others, corresponding to an increased abundance of this tRNA. A second potential constraint on synonymous substitutions occurs through possible secondary structures that the RNA might form, in which certain nucleotides must undergo base pairing in order to form stable double-stranded regions. Pre-messenger RNA secondary structure may influence the speed or accuracy of intron splicing, rate of transport, or stability. A third potential constraint on synonymous substitutions is related to the fact that, during translation, the probability of misincorporation of the wrong amino acid increases if there is a pause while the translation machinery waits to find a rare tRNA. Such translation errors are known to occur. (In fact, mistranslation of an mRNA that bears a frameshift mutation can yield an active protein.) A pause during translation may also be of importance to the folding of the protein into its proper three-dimensional structure.

If synonymous codons are neutral, then one would expect their frequencies of use to correspond to the product of the nucleotide frequencies. If all four bases were equally frequent, all synonymous codons should be used equally frequently. A more subtle way to test for departure from equal codon use is to count the incidence of polymorphisms and substitutions toward or away from the most abundant codon. If the most abundant codon became the most abundant by chance, then the substitutions toward and away from this codon should show no bias. But if the most abundant codon is "preferred," then there will be a deficit of substitutions away from this codon. Application of this kind of approach for codon bias in *E. coli* suggested an average selection coefficient against disfavored codons of about $s = 7.3 \times 10^{-9}$ (Hartl et al. 1994). Even *Drosophila*, whose effective size might be around 10^6, appears to exhibit significant codon preference (Akashi 1995, 1997), suggesting that selective constraints on synonymous codons must be greater than 10^{-6} in this organism (Figure 7.16).

The subject of codon usage bias is made more complex when considering species that may not be in mutation-drift-selection balance. In particular, if the GC content of the genome is in the process of changing, it becomes very difficult to assess the magnitude of the impact of the mutational bias in driving codon bias. Marais et al. (2004) examined 630 gene pairs from *D. melanogaster* and *D. yakuba* and found that genes with a higher nonsynonymous substitution rate have lower codon usage bias. This would be consistent with genes having variation in functional constraint, so that genes with higher nonsynonymous rates are less constrained. Singh et al. (2005) found that X-linked genes have significantly higher codon usage bias than autosomal genes in both *Drosophila* and *Caenorhabditis*. One possible explanation is that natural selection is more efficient on the X chromosome due to hemizygosity of the X chromosomes in males. Similarly, in the comparison of *Drosophila melanogaster*, *D. yakuba*, and *D. simulans*, data suggest that the

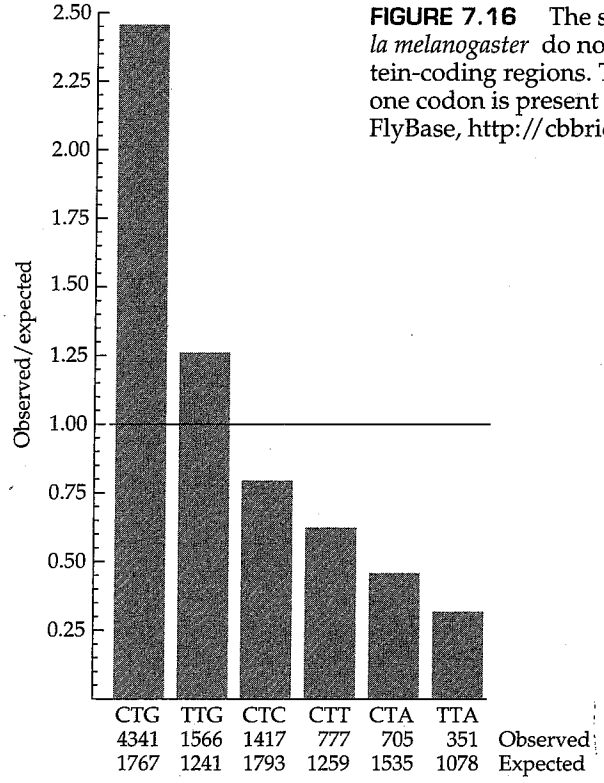


FIGURE 7.16 The six codons that encode leucine in *Drosophila melanogaster* do no appear with uniform frequency in protein-coding regions. This kind of codon-usage bias, in which one codon is present in excess, is often observed. (Data from FlyBase, http://cbbridges.harvard.edu:7081.)

effective size of *melanogaster* is lower. Tests of substitution rates toward and away from preferred codons show that in this species trio, *D. melanogaster* appears to have a relaxation in its codon usage bias (Nielsen et al. 2006).

7.5 POLYMORPHISM AND DIVERGENCE IN NUCLEOTIDE SEQUENCE—THE MCDONALD-KREITMAN AND HKA TESTS

Under the neutral theory, polymorphism levels depends only on the product of the neutral mutation rate and the effective population size, through the formula $H = \theta/(1 + \theta)$ that we encountered in Chapter 4. For strictly neutral genes, interspecific divergence does not depend on the population size, but instead follows the formula $k = 2\mu t$. Because both polymorphism and divergence are driven by the neutral mutation rate, these two attributes ought to co-vary across genes (Figure 7.17). If we compare two particular genes, the level of intraspecific polymorphism would let us estimate a θ value for each gene. Given the θ value for gene A and the observed interspecific divergence, the divergence time could be estimated. For gene B, we would also have a θ

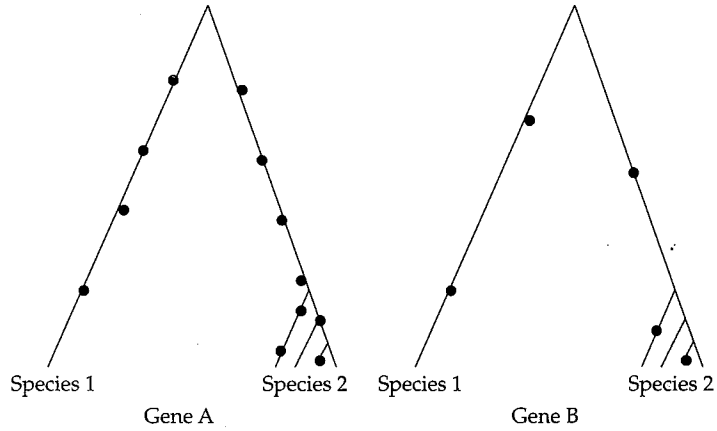


FIGURE 7.17 Under the neutral theory, levels of divergence and polymorphism are expected to be correlated. For gene *A*, there is a high neutral mutation rate, driving many sites of nucleotide divergence and high levels of polymorphism within species 2. For gene *B*, the neutral mutation rate is low, so species 1 and 2 have few sites of divergence, but also species 2 has fewer polymorphic sites. The HKA test is a goodness-of-fit test to the observed levels of intraspecific diversity and interspecific divergence under a model whose parameters are population sizes, neutral mutation rates, and times of divergence.

estimated from the level of polymorphism, and we could use the divergence time estimated from gene *A* to determine a predicted value of divergence in gene *B*.

The above reasoning has been formalized in a popular test of neutrality based on nucleotide sequence data within and among species (Hudson et al. 1987). Sequences of at least two genes from a number of individuals of each of two species are needed to apply the test. Define S_i^A and S_i^B as the number of polymorphic nucleotide sites in gene *i* in species A and B, respectively, and d_i as the number of differences in gene *i* between a pair of alleles sampled randomly, one from species A and one from species B. The expected values of these parameters are obtained from the infinite-sites neutral model, assuming that the two species diverged *t* generations ago, that the population sizes are 2*N* and 2*Nf*, and that each gene has an associated $\theta_i = 4N\mu_i$. Estimates of θ_i, *f*, and *t* are obtained by a least-squares method that gives the best fit of the expressions for the expected values and variances of S_i^A, S_i^B, and d_i to the data, and goodness-of-fit is tested with an appropriate chi-square test. Using data from the *Adh* coding and 5′ flanking regions in *D. melanogaster* and *D. sechellia*, Hudson et al. (1987) found that the observed values deviated significantly from the neutral model in a direction consistent with the operation of

balancing selection acting on the coding region of *Adh*. This finding is consistent with Kreitman's (1983) observation of an excess of silent substitutions in *Adh*, except that the test of Hudson et al. makes use of the genetic variation observed within and among species. This **HKA test** has seen many applications in molecular population genetics and many extensions of the basic idea of contrasting polymorphism and divergence have yielded useful results (Aguadé et al. 1992; Begun and Aquadro 1993; Gaut and Clegg 1993; Wright and Charlesworth 2004; Innan 2006). McDonald-Kreitman and HKA tests rely on the ability to identify synonymous and nonsynonymous variation and divergence, and if such extrinsic functional partitioning can be applied to noncoding sequence, such as inside versus outside a regulatory domain, then these tests can be applied to cis-regulatory regions of the genome as well (Hahn 2006).

PROBLEM 7.6 In a set of 12 *Adh* sequences in *Drosophila melanogaster*, McDonald and Kreitman (1991) observed 42 silent (synonymous) polymorphisms and two replacement (nonsynonymous) polymorphisms. As had been concluded by Kreitman (1983), this finding suggests that most replacement mutations are deleterious and are eliminated from the population. When they examined fixed differences between *D. melanogaster* and either *D. simulans* or *D. yakuba*, they found that seven of the fixed differences were replacements and seventeen were silent. What is the significance of this observation?

ANSWER A null hypothesis might be that the effects on fitness of a mutation would be the same whether within a species or at any time along the ancestral history of two species back to the common ancestor. If this is true, then we would expect the ratio of silent to replacement polymorphisms to be the same as the ratio of silent to replacement fixed differences. A simple test of this is to do a 2×2 contingency chi-square:

	Fixed	Polymorphic
Replacement	7	2
Silent	17	42

For this table we get $\chi^2 = 8.20$, and with one degree of freedom, $P < 0.01$. (A correction is often applied to the chi-square for tables with counts less than 5, but it does not make much difference in this case.) The low probability means we reject the null hypothesis and conclude that, within species, there is a tendency to avoid replacement polymorphisms; however, between-species replacement differences are much more likely to occur. McDonald and Kreitman (1991) argue that this pattern is consistent with adaptive fixation of amino acid replacements, since they are relatively more frequent in interspecific comparisons, and such adaptive polymorphisms would be less common than neutral polymorphisms because adaptive differences would not remain polymorphic for as long a duration. This simple test is useful in assessing the relative importance of neutral drift versus selection in interspecific differences.

One of the more practical applications of analysis of polymorphism and divergence rates arises in human genetics for prediction of the likely damaging effects of a polymorphism. If a site is invariant across multiple species, this may be a signal that the residue is constrained by natural selection so a polymorphism for a radical amino acid change at this site might be likely to be deleterious. Ng and Henikoff (2005) review the application of tests of this sort and show that simple rules like this actually do have good prediction ability for deleterious functions. When the impact of structural changes on the protein is also predicted, the ability to predict damaging mutations is even stronger (Sunyaev et al. 2001).

Polymorphism and Divergence in Noncoding Sequences

Only about 1.5% of the mammalian genome is protein coding, so it is of interest to understand evolutionary changes in noncoding regions as well. By calibrating the distribution of levels of divergence between human and mouse with the likely neutral divergence seen among inactive transposable elements, Waterston et al. (2002) inferred that about 5% of the mammalian genome is subjected to purifying selection. This implies that about two-thirds of the constrained sequence in our genome is noncoding. Haddrill et al. (2005) sought to characterize evolutionary forces on noncoding sequences in *Drosophila melanogaster* by sequencing 10 X-linked regions in three African and two non-African population samples. They found elevated linkage disequilibrium, excess high-frequency derived alleles, and wide heterogeneity across regions in patterns of polymorphism, all classic signatures of positive selection. But they also noted that relatively simple demographic models with population bottlenecks could explain most of the data.

It seems that the strongest inference can be made by contrasting patterns of polymorphism and divergence. By quantifying the low level of interspecific divergence, Andolfatto (2005) estimated that 40–70% of nucleotides in intergenic regions, untranslated regions of mature mRNAs, and most intronic DNA are more evolutionarily constrained than are synonymous sites. He also applied a McDonald-Kreitman test (partitioning sites into noncoding versus synonymous classes) and found a large fraction of the nucleotide divergence in noncoding regions was driven to fixation by positive selection. It seems that the potential for natural selection to act on noncoding DNA is considerable.

Impact of Local Recombination Rates

Recall from Chapter 4 that the level of polymorphism in *Drosophila* shows a striking correlation to the local rate of recombination. Regions of low recombination rate are nearly devoid of variation, whereas regions with high rates of recombination are highly polymorphic. The comparison of polymorphism and divergence makes this pattern even more striking and allows us to eliminate a possible cause. One possible reason for the correlation is that recom-

bination itself is mutagenic, or that somehow the two processes are related mechanistically. (That is, perhaps when recombination occurs, the DNA configuration is altered to increase the chance of mutation.) If this were the case, then the regions of low recombination rate should also have a low mutation rate, and hence lower interspecific divergence. In fact, the level of interspecific divergence is independent of local recombination rates. The conclusion is that the correlation between recombination rates and levels of polymorphism observed by Aquadro et al. (1994) must be due to more rapid elimination of the variation in regions of low recombination.

Careful consideration of levels of polymorphism and divergence across regions of low and high recombination can be highly informative. Using polymorphism data from 98 genes in *Drosophila melanogaster*, and rooting the changes with sequences from *D. simulans* and *D. yakuba*, Presgraves (2005) showed that regions of high recombination had considerably higher ratios of replacement to silent divergence compared to regions of low recombination, consistent with greater efficacy of selection to detect positive effects in regions of high recombination (Figure 7.18). On the other hand, the elevated ratio of replacement to silent polymorphism in low-recombination regions suggests that natural selection against mildly deleterious replacement muta-

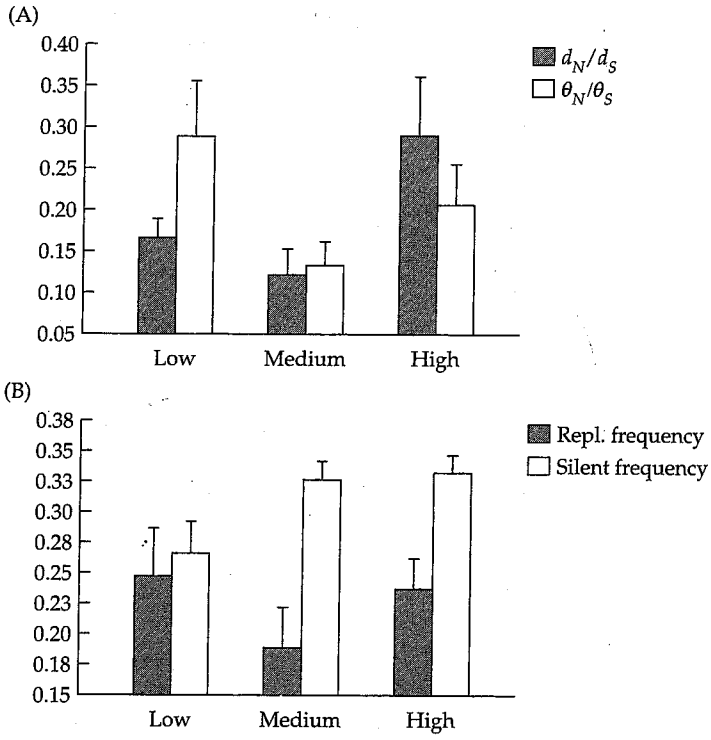


FIGURE 7.18 The effect of low, medium, and high levels of local recombination on polymorphism and divergence. (A) The ratio of replacement to silent nucleotide polymorphism and ratio of replacement to synonymous substitution rates (d_N/d_S) for genes differing in recombination. (B) The frequencies of derived replacement and silent alleles are significantly different in medium- and high-recombination regions. Bars indicate means plus or minus standard error. (From Presgraves 2005.)

tions is compromised in these regions. Furthermore, the frequencies of derived mutations at replacement and silent sites were roughly equal for regions of low recombination, but nonsynonymous derived alleles were maintained at much lower frequency in regions of high recombination, again consistent with the greater ability of selection to reduce the frequency of deleterious variants.

Two known mechanisms that remove variation faster in regions of low recombination are selective sweeps and background selection. Background selection is thought to be the primary mechanism for the reduced variation (discussed in Chapter 5 in the section on linkage and recombination), with sweeps not as common. *Selective sweeps* occur when a favorable mutation takes place, and selection rapidly increases its frequency. Such sweeps can have a dramatic effect on levels of variation in the selected gene and the region around it. The size of the "swept" region depends on the rate of recombination and is larger for regions of low recombination. This means that the chance that a particular site has been swept free of variation is greater in regions of low recombination, assuming the density of selective sweeps is uniform across the genome. An example of a selective sweep driving a global spread of a novel mutation causing pesticide resistance is the *Ester-2* allele in the mosquito *Culex pipiens* (Figure 7.19). The inference that the resistant allele has undergone a nearly global sweep was made based on the worldwide sequence identity of the *Ester-2* allele (Labbe et al. 2005). We do not know how frequent such sweeps are, but one possible means of identifying them is to score many highly polymorphic markers in many populations and look for regions of reduced variation. In Chapters 9 and 10 methods for identifying selective sweeps from genome-wide scans will be discussed.

Substitution Models for Structural RNA Genes

Specialized RNA molecules such as microRNAs, transfer RNA, and ribosomal RNA derive their biochemical properties from the secondary structure into which they fold. We are still learning the chemical rules by which such macromolecules attain their final folded configuration, but one thing that is very clear is that complementary base pairing is important. The stems of tRNAs are critical to maintaining the tightly folded structure of these essential molecules. Substitutions that occur in stems will weaken the stability of the stem unless there is a compensatory change on the other strand that maintains base pairing. Kimura (1985) realized that one could obtain evidence for such compensatory changes. Such complementary changes have also been demonstrated in an intron, demonstrating that the folding structure of introns may also be important to regulating gene expression (Chen and Stephan 2003).

One attribute of secondary structure is measured as the difference in free energy attributable to complementary base pairing in the folded versus

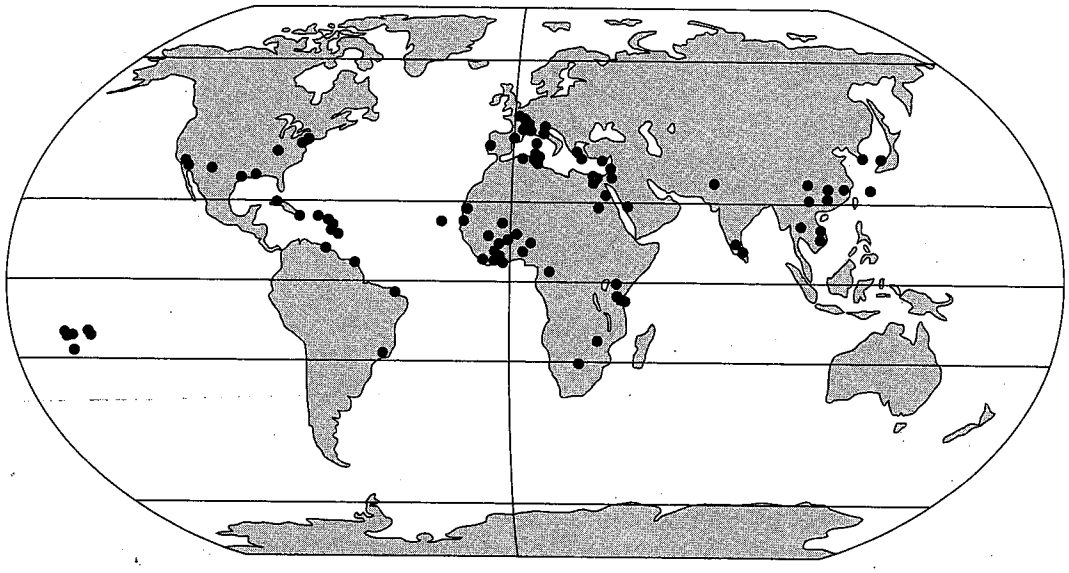


FIGURE 7.19 Global presence of the *Ester-2* esterase mutation conferring insecticide resistance in *Culex* mosquitoes. (Labbe et al. 2005.)

unfolded state. Computer predictions of the best folding structure of the rRNA pseudogenes suggested that the difference in free energy decreases as the sequences accumulate substitutions. Tests of randomly permuted sequences showed that the functional rRNA sequences are significantly more stable than would be obtained by chance, whereas predicted pseudogene RNAs are not. Some introns have a significantly open secondary structure, such that random substitutions in their sequences result in more stable structures (Leicht et al. 1995). The reason some introns retain an open structure may be for access to regulatory proteins. This possibility has been indirectly demonstrated by showing that stable stems inserted into introns in yeast can disrupt normal splicing.

In order to test whether natural selection has an impact on a sequence in a manner consistent with constraints on secondary structure, models must be devised that specifically target the structural effects. For example, there are physical-chemical models that allow one to predict whether a DNA sequence will kink at a particular position. Wong and Nielsen (2004) used this type of model to posit a substitution model for noncoding DNA, allowing different rates of substitution for sites that did or did not introduce a kink. A challenge in these models is that sites do not evolve independently, but rather depend on nearby sites with which they interact. Siepel and Haussler (2004b) devised

a general approach for analysis of context-dependent substitutions on a set of aligned sequences. Methods like this found highly conserved short RNA-encoding genes with a strong secondary structure predicted. When otherwise constrained genes also show rapid evolution within one lineage, it provides an interesting possibility of a species-specific adaptive change. Just such a pattern was observed for the gene *HAR1*, which not only shows rapid human-specific change, but is also specifically expressed in the developing brain (Pollard et al. 2006).

7.6 GENE GENEALOGIES

There is an important distinction between the construction of trees from sequences of genes from different species and from sequences of alleles from a single species. The former yields a customary phylogenetic **gene tree**, while the latter produces what is called a **gene genealogy**. The relationships among species result from macroevolutionary processes, whereas allelic differences result from a number of microevolutionary processes, including aspects of genetic transmission. Once the nucleotide sequences of alleles are known, the different alleles can be treated like genes in different species in applying standard methods for inferring a phylogenetic tree. However, great care is needed in constructing gene genealogies, because recombination among the sequences results in a gross violation of the assumptions of most tree-building methods. Provided the rate of recombination is not too high, localized blocks of sequence can be identified in which there appears to have been no recombination in the ancestral history of the sampled alleles. With this caveat, gene genealogies can be of great use in inferring the evolutionary history of a polymorphism. For example, they can reveal which of a group of alleles is older, or which alleles are more closely related to each other. Figure 7.20 shows the gene genealogy from Kreitman's (1983) *Adh* sequence data, and the higher diversity of the *S* allele clearly makes it appear to be older.

Hypothesis Testing Using Trees

Beyond the descriptive approach to showing relationships among alleles, gene genealogies can be used to test fundamental forces of population genetics, including natural selection. For example, consider a phylogenetic tree based purely on neutral variation. As illustrated in Figure 7.21A, when the substitution rate is μ, the expected time to coalescence to a common ancestor for a randomly chosen pair of alleles is $4N$ generations (see Chapter 3). Under a model like Ohta's (1973), where many mutations are slightly deleterious, the tree is not changed very much because the alleles included in a sample are the subset of mutations that occurred that were nearly neutral.

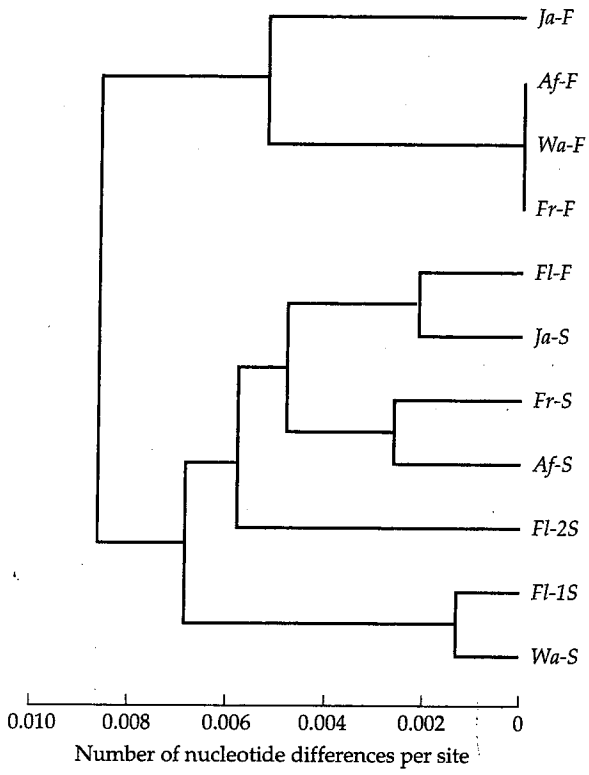
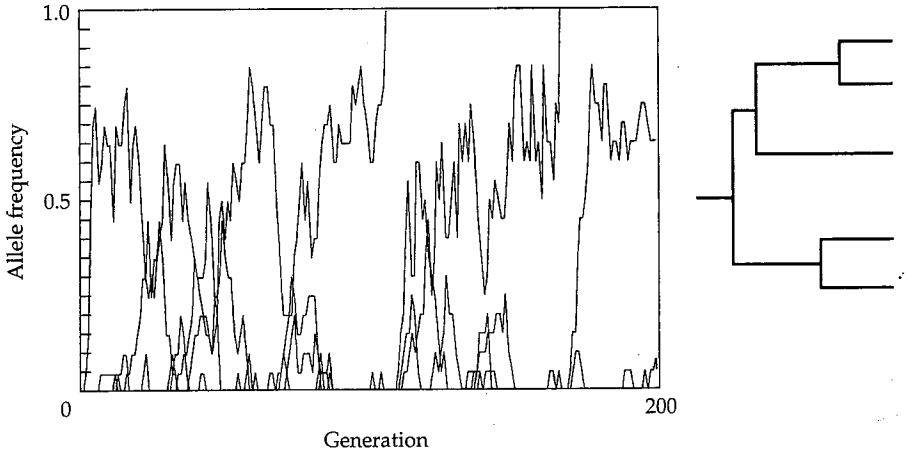


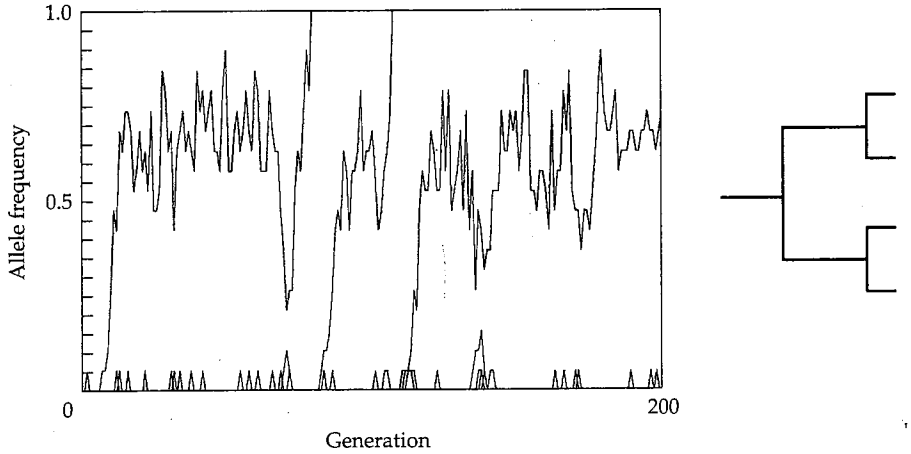
FIGURE 7.20 A phylogenetic tree for 11 *Adh* alleles of *Drosophila melanogaster* based on 43 nucleotide differences. The scale is the number of nucleotide differences per site. *Ja*: Japan; *Af*: Africa; *Wa*: Seattle, Washington; *Fr*: France; *Fl*: Southern Florida; *S* and *F* refer to the slow and fast electrophoretic forms. (Data from Kreitman 1983.)

On the other hand, in the case of adaptive mutations, the rate of fixation would be much faster than with neutrality, so that sites of adaptive mutation would have shorter coalescence times than flanking neutral sites (Figure 7.21B). Finally, with balancing selection, polymorphisms would be maintained for a longer time than under the pure drift model (Figure 7.21C). The number of statistical methods for inference of population genetic forces from gene genealogies is increasing rapidly, and there is ample opportunity for exciting progress in this area.

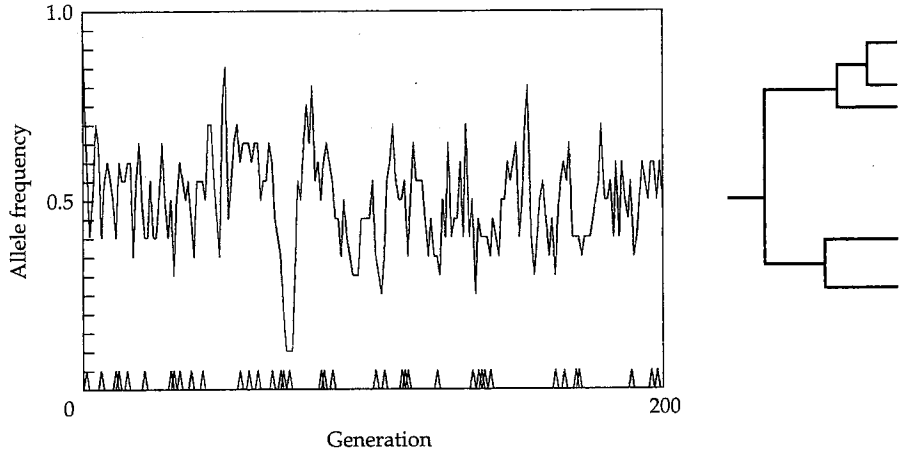
(A) No selection



(B) Purifying selection



(C) Stabilizing selection



PROBLEM 7.7 A study of variation in the gene encoding superoxide dismutase in *Drosophila melanogaster* (Hudson et al. 1994) revealed 63 polymorphic sites in three *slow* alleles and 22 *fast* alleles (where fast and slow refer to the mobility of the protein product in an electrophoretic gel). An additional 16 *slow* alleles were separately scored, giving a total of 19 *slow* alleles that were found to be identical in nucleotide sequence. The *fast* allele broke into 10 distinct haplotypes, and the most common was *FastA* with nine copies. The partial table of pairwise counts of numbers of sites that differ between alleles is:

	FastA	*FastH*	*FastB*	*FastJ*	*FastK*
Slow	1	3	4	9	16
FastA		2	3	8	15
FastH			3	10	17
FastB				11	18
FastJ					7

How would you address the question of whether this sample is typical of one with only neutral alleles?

ANSWER The aspect of the pattern of variation that is unusual is that the *fast* alleles appear to be quite variable, whereas all 19 *slow* alleles are identical. A gene genealogy of the *fast* alleles would look like a typical neutral tree with roughly exponentially distributed branch lengths, but the complete tree would then have 19 identical *slow* alleles placed one substitution away from *FastA*. The suspicion is that the *slow* allele must have arisen recently and is being pulled to high frequency by selection. An observed increasing trend in the *slow* allele frequency supported this conjecture. To make a formal test out of this observation, Hudson et al. (1994) used the coalescent procedure described in Chapter 4 to generate simulated data sets with a sample size of 25 and having 63 polymorphic sites. For each of the 10,000 simulated samples, they asked, how often is there a set of 12 alleles that differ by 0 or 1 substitutions? (The 9 *FastA* alleles and 3 slow alleles in the original observed sample differ at just 1 site.) The answer was 81 of the 10,000 cases, giving a probability of 0.0081. The observed sample is not a likely occurrence under neutrality.

◀ **FIGURE 7.21** Computer simulations of the infinite-allele model of molecular evolution. (A) With strict neutrality, the expected time from mutation to fixation of alleles that will go to fixation is $4N_e$ generations. (B) Purifying selection (in this case with half of the mutations having a fitness of 0.5) results in less polymorphism at any given time. (C) Stabilizing selection (overdominance or frequency-dependent selection) can retain alleles in a polymorphic state for much longer times. Representative trees are plotted to the right of each panel.

It is instructive to note that the data in Problem 7.7 were consistent with neutrality by the Fu and Li (1993) test, the Tajima (1989) test, and the HKA test (Hudson et al. 1987), demonstrating that even strong departures from neutrality may be missed by these standard tests. This problem illustrates a common principle in molecular population genetic analysis, which is that *ad hoc* approaches tailored to particular observations often are necessary.

7.7 MITOCHONDRIAL AND CHLOROPLAST DNA EVOLUTION

The mitochondrial genome attracted attention early on because it could be surveyed easily for polymorphism by restriction digestion. Some of the current advantages of using DNA sequence variation from this organelle genome include its small size and ease of manipulation, the fact that it is present in multiple copies per cell and so is robust to degradation, its lack apparent of recombination, and its rapid rate of molecular sequence evolution. A primary problem with mtDNA includes the fact that the lack of recombination means that the gene tree constructed from any mitochondrial DNA gene will reflect just a single realization of the genealogical process. As such, the data will not be as informative about species or population trees as, say, a dozen nuclear genes.

In animals, mitochondria are usually inherited through the egg cytoplasm (maternal inheritance) and are usually genetically uniform within an individual. The mitochondrial genome consists of a single circular DNA molecule, denoted mtDNA, the size of which varies over a remarkably narrow range in different species of vertebrates (15.7–19.5 kb), averaging about 16 kb. Human mtDNA is fairly typical, at 16568 bp, and contains a control region for the initiation of DNA replication, genes for two ribosomal RNA molecules, 22 transfer RNA molecules, and 13 proteins (see www.mitomap.org). Twelve of the proteins are subunits of enzyme complexes that carry out electron transport and ATP synthesis. The genetic code of mammalian mitochondria differs from the standard code in that ATA codes for Met, TGA codes for Trp, and AGR codes for End (termination of protein synthesis); thus, every codon in the mitochondrial code can be written as either NNY or NNR. Animal mitochondria also contain several hundred enzymes used in metabolic functions, but these are coded for by nuclear genes, and the enzymes are transported into the mitochondria.

At the nucleotide level, the rates of substitution in mammalian mtDNA are typically 5 to 10 times greater than occur in single-copy nuclear genes, averaging approximately 10×10^{-9} substitutions per nucleotide site per year. The reason for the high rate of substitution is thought to be either a high rate of nucleotide misincorporation or a low efficiency of repair of the DNA polymerase. Support for the latter view comes from the observation that, unlike the nuclear DNA polymerase, the mitochondrial DNA polymerase lacks the proofreading function. In protein-coding mitochondrial genes, the rate of

synonymous substitution is about five times greater than the rate of nonsynonymous substitution, which is comparable with the ratio found in nuclear genes. Mitochondrial tRNA genes are the fastest evolving genes in the mitochondrial genome of humans, with a substitution rate about 100 times their nuclear counterparts (Kivisild et al. 2006). One result of this faster rate of nucleotide substitution is that the divergence between two sequences saturates relatively quickly, so that the linearity of divergence over time (the molecular clock) is an accurate approximation only for species that have diverged less than about 10 million years (see Figure 7.8). Exceptions to the elevated rate of mtDNA divergence have been found, notably in *Drosophila* (Ballard 2000).

PROBLEM 7.8 The mitochondrial DNA of 21 humans of diverse geographic and racial origin were digested with 18 restriction enzymes, 11 of which exhibited one or more fragments in which size polymorphism occurred (Brown 1980). All restriction site polymorphisms could be explained by single-nucleotide differences; thus there was no evidence for insertions, deletions, or other mtDNA rearrangements. Altogether, 868 nucleotide sites were assayed for differences among individuals, and the average number of differences per nucleotide site per individual was estimated at 0.0018. Assuming that mammalian DNA undergoes sequence divergence at the rate of 5 to 10×10^{-9} nucleotide substitutions per site per year, and that the rate is uniform in time, calculate how long it has been since all of the 21 contemporary mtDNA molecules last shared a common ancestor. Calculate the effective size of the population from the level of mtDNA variability.

ANSWER Given an average number of differences per nucleotide site per individual of 0.0018 and an average rate of divergence of 5 to 10×10^{-9} per site per year, the time of the most recent common ancestor would be between $0.0018/(10 \times 10^{-9})$ and $0.0018/(5 \times 10^{-9})$ or 180,000 to 360,000 years. Assuming a generation time of 20 years, this means that all mtDNA in the diverse sample could have been from a single female in the population between 9,000 to 18,000 generations ago. To estimate the long-term effective size of the population, recall that the expected time to fixation of a newly arisen neutral mutation is $4N_e$ generations. This result applies to an autosomal gene in a diploid species. Mitochondrial genes, however, are transmitted by females only and they are effectively haploid, so the corresponding fixation time for mtDNA is just N_e generations. The statement that the one mtDNA type went to fixation in 9,000 to 18,000 generations is equivalent to saying that the long-term population size has been N_e = 9,000 to 18,000. Although this sounds like a low number, most modern anthropologists find it reasonable, given the population structure of ancient humans and their rapid, nearly starburst-like growth since the adoption of agricultural methods.

Chloroplast DNA and Organelle Transmission in Plants

Chloroplasts are cellular organelles that also have their own genome and also are transmitted in a non-Mendelian fashion. Chloroplast DNA (cpDNA)

TABLE 7.3 Rates of Sequence and Structural Evolution in Organelle DNA

Genome	Rate of nucleotide substitution	Rate of structural evolution
Angiosperm cp DNA	Slow	Slow
Angiosperm mtDNA	Slow	Rapid
Mammalian mtDNA	Rapid	Slow
Fungal mtDNA	Rapid	Rapid

ranges in size from 135 to 160 kb and occurs in multiple copies in each chloroplast. Its structural organization is conserved in higher plants, and the rate of synonymous nucleotide substitution is approximately 1×10^{-9} substitutions per site per year. Thus, the evolution of cpDNA is conservative in regard to both sequence and structure (Table 7.3). The opposite extreme, with a very fast rate of evolution, is found in the mtDNA of fungi, which changes rapidly in both sequence and structure.

The mtDNA of angiosperm plants has the opposite pattern of evolution from that found in animal mtDNA. In sequence evolution the rate in angiosperms is slow, but in structural evolution it is fast. In plants, the mtDNA genome is large and highly complex. In some instances, a single molecule can resolve itself into smaller circles and even linear molecules. For example, in the turnip (*Brassica campestris*), a 218-kb molecule undergoes an internal recombination event that produces smaller circles of 135 kb and 85 kb. Maize mtDNA contains six pairs of repeated sequences that can undergo recombination and create a variety of structural derivatives. The *Arabidopsis* mtDNA genome spans 366 kb, but nearly all the increase in size compared to mammalian mtDNA is noncoding (Unseld et al. 1997). Many plant mitochondria also contain autonomously replicating plasmid DNA molecules, and mtDNA is also capable of incorporating segments of cpDNA. Why plant mtDNA genomes are so large, complex, and variable in size is not understood.

Maintenance of Variation in Organelle Genomes

Organelle genomes have unusual population genetics because of their (typically) uniparental transmission and because many copies are passed from the mother to the progeny through the egg. Uniparental transmission has important implications in the operation of natural selection, since it is equivalent to a haploid clonal population structure, and pure selection models can maintain polymorphism in such populations only if the fitnesses are frequency-dependent. From the outset, then, uniparental transmission makes it less likely for polymorphisms to be maintained by natural selection, even if epistatic

effects with the nuclear genome are allowed (Clark 1984). The widespread polymorphisms observed in mtDNA must then be attributed largely to high mutation rates, just as the rapid substitution rate was attributed to a high mutation rate. Polymorphisms can also be maintained by interspecific hybridization, and it is possible to obtain estimates of rates and directions of interspecific matings from nuclear and mtDNA data (Sites et al. 1996). Unusual forms of transmission, such as the doubly uniparental transmission of the mussel *Mytilus edulis*, results in separate male and female lineages, which are highly divergent (Skibinski et al. 1994; Stewart et al. 1995; Burzynski et al. 2006), despite the fact that they can occasionally recombine (Rawson 2005).

The theory of random genetic drift for organelles is more complex than that for nuclear genes because individual cells have many organelles that are apportioned among daughter cells; thus there is an additional level of sampling when heteroplasmic cells divide. Models of the dual sampling process have been examined in some detail (Takahata 1983, 1984; Bergstrom and Pritchard 1998). These models predict some level of **heteroplasmy**, or presence of two or more genetically distinct organelle genomes, and although early empirical studies did not detect heteroplasmy, it now seems to be nearly universal, even if it occurs at a low level in general. In humans, any suspicion of a mitochondrial pathology results in getting the full mtDNA genome of the patient sequenced, and now well over 1200 complete genome sequences have been obtained. PCR-based methods afford considerable sensitivity to detect heteroplasmy, and in humans it is relatively common. Distributions of levels of heteroplasmy appear in general to be consistent with a model of mutation-selection balance, with smaller genomes favored by selection (Korpelainen 2004).

Evidence for Selection in mtDNA

There are several clear examples of non-neutrality of mtDNA mutations. For example, many forms of cytoplasmic male sterility in plants are caused by defects in mtDNA (Levings 1983). Similarly, cytoplasmically transmitted drug-resistance genes have been shown to be associated with the mitochondrial genome of yeast. The potential importance of mtDNA variation in human health was revealed in the implication of mitochondrial DNA defects in the muscle diseases known as **mitochondrial myopathies**. Effects of natural selection also have left their mark on extant patterns of mtDNA sequence variation, as revealed by the discordance between levels of polymorphism and divergence in synonymous versus nonsynonymous sites (Ballard and Kreitman 1994; Rand and Kann 1996; Rand et al. 2000). The strongly skewed distribution of frequencies of segregating sites also suggests that human mtDNA has faced selection pressure (Mishmar et al. 2003). Direct tests of oxidative phosphlorylation function demonstrate physiological effects of mtDNA polymorphisms that interact with the nuclear genome (Sackton et al. 2003; Rand et al. 2004).

If a cytoplasmically related factor of any sort is associated with a particular mtDNA type, then the mtDNA will "hitchhike" along with the other cytoplasmic factor. A striking example of this mode of evolution in action was caught by Turelli et al. (1992) when they noticed that a cytoplasmically transmitted *Wolbachia* infection in *Drosophila simulans* was rapidly spreading north in California and, as it did so, it propelled a single mtDNA type to high frequency. While the mtDNA genome may seem small, its uniparental transmission makes it susceptible to any cytoplasmic factor that may carry a particular cytoplasmic type to fixation. However, most populations have fairly high levels of mtDNA variation, suggesting that such sweep events are not very common.

7.8 MOLECULAR PHYLOGENETICS

The use of techniques of molecular biology, particularly those for determining amino acid or nucleotide sequences, has added a new dimension to phylogenetic inference. For example, the analysis of 5S RNA sequences in a broad variety of microorganisms has led to a reclassification at the deepest of phylogenetic levels, resulting in a new kingdom, the Archaea (Woese 1981). In addition to fostering an understanding of the history of relationships among living things, the application of comparative molecular analysis to infer robust and accurate phylogenetic relationships has spawned interest in the application of those phylogenetic trees for testing hypotheses about evolutionary mechanisms. The problem of inferring the correct branching topology for a tree that relates a set of organisms is a challenge in part because of the enormous number of possible bifurcating trees. If there are n species to be placed, there are $(2n-3)!/2^{n-2}(n-2)!$ rooted trees that describe possible ancestral histories. For five species this number is 105, and for 10 species it is 34,459,425. For many data sets of 30 or more species, the number of possible trees is so enormous that it is not possible to examine all topologies and assess the fit of the data to each tree, even with the very fastest computers. Fortunately, the trees are not all independent of one another, and the key to many of the algorithms that try to find the best-fitting tree is to eliminate whole classes of trees based on the observed data. Below we will consider a few of these tree-building methods, and for the interested reader a definitive source is Felsenstein (2003).

Algorithms for Phylogenetic Tree Reconstruction

If a gene in a pair of species or populations evolves in clocklike fashion, and if the degree of divergence between two genes implies that they have been diverging for t generations, then we can infer that the genes separated from a common ancestor $t/2$ generations ago. This reasoning provides a group of methods of tree construction based on measures of genetic distance. One such method is the unweighted pair-group method with arithmetic mean

(**UPGMA**) or average distance method. This method requires that all sequences evolve at the same rate, an assumption that other methods can relax to some degree, but the ease of understanding UPGMA still gives it heuristic appeal. With a matrix of all pairwise distances, a tree is built up by first grouping the two species with the smallest distance. A new distance matrix is then constructed, with the grouped species now considered as one unit. If the grouped species were indexed i and j, then for all $k \neq i, j$, the distance from k to the group $\{i,j\}$ is $d_{k(ij)} = \frac{1}{2}(d_{ik} + d_{ij})$. In words, the distance from each other species k to the group $\{i,j\}$ is the average of the distances from species k to each of species i and j in the group. The new distance matrix is again searched for the smallest element, and the appropriate grouping again occurs. This process is repeated until all species are clustered into a tree.

Tree-building methods can not only produce a tree topology, but they generally also give estimates of branch lengths of the tree. An example of one method for branch-length estimation is the method of Fitch and Margoliash (1967). Suppose the number of substitutions distinguishing sequence i and j is d_{ij}. If the tree relating sequences 1, 2, and 3 has branch lengths A, B, and C (Figure 7.22), then the branch lengths can be estimated from

$$\begin{aligned} A &= \tfrac{1}{2}(d_{12} + d_{13} - d_{23}) \\ B &= \tfrac{1}{2}(d_{12} + d_{23} - d_{13}) \\ C &= \tfrac{1}{2}(d_{13} + d_{23} - d_{12}) \end{aligned} \qquad (7.17)$$

These relations were found by solving the equations $d_{12} = A + B$, $d_{13} = A + C$, and $d_{23} = B + C$. With more than three sequences, the tree is built up by considering three units at a time, beginning with the two most closely related sequences and grouping the remaining sequences. If sequences 1 and 2 are the most similar, then the distances from sequence 1 to the remaining group

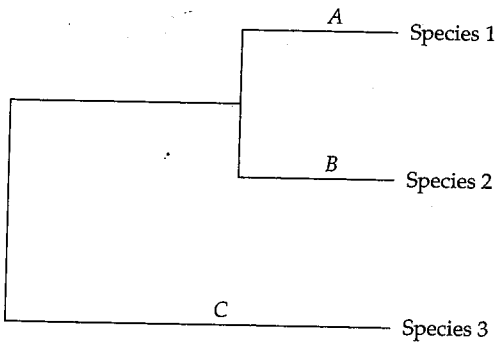


FIGURE 7.22 A phylogenetic tree of three species. A, B, and C represent branch lengths from the most recent common ancestor.

	B	C	D	E
A	0.53	0.99	1.02	0.82
B		0.80	0.93	0.73
C			0.65	0.81
D				0.94

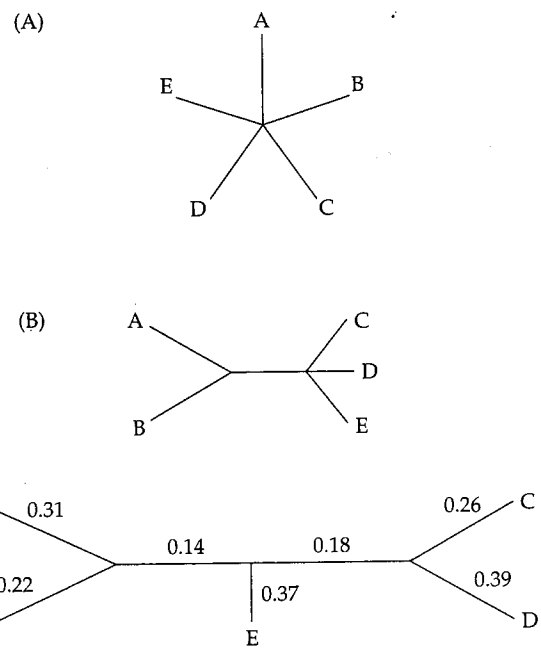


FIGURE 7.23 Illustration of the neighbor-joining method for phylogeny reconstruction. Given a distance matrix (top), one starts with a star phylogeny (A) and tests all trees having different pairs separated from the rest. The tree with A–B joined is the shortest such tree (B). The process of testing all pairs of "neighbors," where a neighbor may be either a single allele or a cluster of alleles, is repeated until no more joining can be done (C). (See Saitou and Nei 1987.)

is the average of the distances from sequence 1 to each member of the group. In this way, only three distances are considered at a time, and Equations 7.17 allow branch lengths to be estimated. This method is known as **least squares**, and it turns out that Equations 7.17 minimize the sum of squared deviations from the model, much like linear regression discussed in Chapter 8.

Another algorithm for tree construction is particularly well suited to the situation in which one does not know whether rates of substitution are constant across clades of the tree. This method is known as **neighbor joining**, because it groups species having the property "neighbors" (Saitou and Nei 1987). Begin by assuming that the sequences are all related to one another by a star phylogeny like tree (A) in Figure 7.23. For a star phylogeny with N sequences, the sum of the branch lengths is

$$S_0 = \sum_{i \neq j} d_{ij} / (N-1)$$

(It may help to draw a star phylogeny to see that each branch gets counted $N - 1$ times.) Next we begin a procedure that groups certain sequences together. For each possible pair of sequences, a tree like that in tree (B) in Figure 7.23 is constructed. Branch lengths for this tree are estimated by least squares, and the sum of the branch lengths for the entire tree (S_{ij}) is calculated. We consider as neighbors that pair of sequences i and j that give the minimum of the S_{ij}'s. After the first pair of neighbors is found, that pair is considered as a single entity (joined neighbors), and the process of considering all possible pairings is repeated. The distance from any one sequence k to this pair of neighbors (i and j) is the average of the two distances, or $\frac{1}{2}(d_{ik} + d_{jk})$. The process ends when there are just three neighbors left, and at this point we have a finished neighbor-joining tree complete with branch lengths. The criterion for neighbor-joining is to minimize the sum of branch lengths. Sometimes it is possible to find tree topologies that are even shorter by using a method called **minimum evolution** trees (Rzhetsky and Nei 1992).

PROBLEM 7.9 Consider a sample of one allele drawn from each of three species. Suppose that the tree that one gets from these alleles may be represented ((A,B),C), implying that A and B are most closely related, and C is the outgroup. What are the possible relationships among the species bearing these alleles?

ANSWER This problem bears on an important issue in phylogeny reconstruction, namely that any one gene tree does not necessarily reflect the true pattern of splitting of species. The easiest way to see this is to consider ancestral populations as being polymorphic, in which case the speciation process may sort out the alleles in various ways. It turns out that the possible species trees include ((A,B),C), ((A,C),B), and ((B,C),A). In other words, the gene tree does not eliminate the possibility of *any* of the species trees.

Distance Methods versus Parsimony

There is no general theory that provides a single optimal way to construct phylogenetic trees, and as basic as the distance matrix seems, it is not required by all methods. Another method, known as **maximum parsimony**, uses the smallest number of mutational events necessary to account for the evolution of a set of sequences from a common ancestor to construct the trees. There are a number of such parsimony methods based on trees with the smallest number of substitutions, but none guarantee that the most parsimonious tree is the correct tree. For example, when rates of substitution differ in different branches of the tree, the parsimony method often fails to give the correct topology (Felsenstein 1978). Methods for constructing phylogenetic trees have been reviewed by Felsenstein (2003).

Bootstrapping and Statistical Confidence in a Tree

Because there are so many possible tree topologies, it is important to assess how much statistical confidence one can place in a particular tree. One cannot assign a numerical standard error to a tree; by its geometrical nature a tree is actually a complicated estimate of phylogenetic relationships; we might have high confidence in some branches and low confidence in others. A widely used method of assessing confidence in the nodes of a tree is the **bootstrap test** (Felsenstein 1985). The basic idea is quite simple: A subset of the original data is drawn with replacement and from this new data set a tree is drawn. For each node in the original tree, we ask whether the new tree has the same cluster of sequences. The whole operation of resampling the data, drawing a tree, and tallying up nodes that are in the original tree is repeated perhaps one thousand times. The final result is displayed graphically as a number next to each node indicating the percentage of time that cluster is present among the resampled trees. If that fraction is high, then one gains confidence that the given cluster actually belongs together.

Another means of testing the statistical confidence in a tree is to test the null hypothesis that each interior branch has length zero. From distance methods, we often obtain estimates of all branch lengths in the tree, along with their standard errors. If we fail to reject the null hypothesis of zero length for an interior branch, then we lose confidence in the nodes surrounding that branch.

The challenge of testing confidence in tree topology grows with the size of the tree, and the scale of DNA sequence data both in terms of numbers of genes and numbers of distinct organisms challenges the largest computers. Often the matrix of genes × species is sparsely populated, so that some genes are sequenced in only a subset of species, yet there is at least some sequence for a large collection of related species. A phylogeny that is assembled from this kind of piecemeal data is called a **supertree** (Sanderson and Driskell 2003). Methods for testing confidence in supertrees by bootstrapping have also been developed (Burleigh et al. 2006).

Bayesian Methods

An altogether different way to think about phylogenetic inference comes from casting the whole problem in a Bayesian framework. This approach allows one to construct a posterior probability density conditional on the observed data for a set of phylogenetic relations. Let τ_i be the topology of the ith tree, and let X represent the DNA sequence data. Bayes' rule states:

$$\Pr(\tau_i \mid X) = \frac{\Pr(X \mid \tau_i)\Pr(\tau_i)}{\sum_{j=1}^{jmax} \Pr(X \mid \tau_j)\Pr(\tau_j)}$$

where *jmax* is the set of admissible tree topologies. $\Pr(\tau_i)$ is the prior probability of topology i and is not always easy to justify. Huelsenbeck et al. (2001, 2002) provide an excellent summary of the advantages and disadvantages of Bayesian inference for evolutionary problems. The primary advantage is unusual for computational Bayesian methods, namely that it is so efficient. Rather than having to iterate through millions of tree topologies, Bayesian methods sample the parameter and topology space with surprising speed. Bayesian methods also can yield inferences about model parameters such as branch lengths. Offsetting these advantages are several factors that are not all easy to overcome. For example, the influence of prior information (priors) on the outcome can be nonintuitive, and the simple approach of trying different priors erodes that initial gain in efficiency. Then again, as with all Bayesian methods, it is not always clear that the sampling process has continued long enough to assure adequate mixing. Also, interpreting the meaning of posterior densities of the parameters in the language of bootstrap estimates is not direct. One can, however, compare the goodness-of-fit of a pair of evolutionary models by calculating a Bayesian factor, defined as the ratio of the probability of the topology under model 1 to the probability of the topology under model 2 (analogous to a likelihood ratio). Ronquist and Huelsenbeck (2003) distribute a computer package to perform the intricate calculations.

Trans-Species Polymorphism

One might intuitively expect that all the alleles of a species should cluster together on a gene tree, implying that the common ancestor of all the alleles is an ancestral allele within the same species. A few gene trees have been found to have the unexpected property that alleles in two or more species appear to be interdigitated on the tree. This pattern, known as **shared polymorphism** or **trans-species polymorphism**, has been observed in major histocompatibility alleles in primates (Loisel et al. 2006), in self-incompatibility alleles of plants (Ioerger et al. 1991), and in several genes in the *melanogaster* species subgroup of *Drosophila* (Charlesworth et al. 2005). Figure 7.24 shows the probable means by which shared polymorphism arises, namely, that the ancestral species was polymorphic, and two or more alleles remain in the descendant species ever since the time of the common ancestor. Recall that the expected fixation time for a new mutation, given that it goes to fixation, is $4N_e$ generations. Consequently, observation of shared polymorphism between species implies that either strong selection is retaining the alleles in the population or that the species have diverged relatively recently. On the other hand, some apparent shared polymorphism, such as the ABO blood groups of humans and chimpanzees, arise by parallel mutations and are not really shared polymorphisms in the sense of Figure 7.24 (Sumiyama et al. 2000). Also, horizontal gene transfer, a phenomenon that is well-documented in plants (Richardson and Palmer 2006), is sometimes responsible for generating shared polymorphism between species.

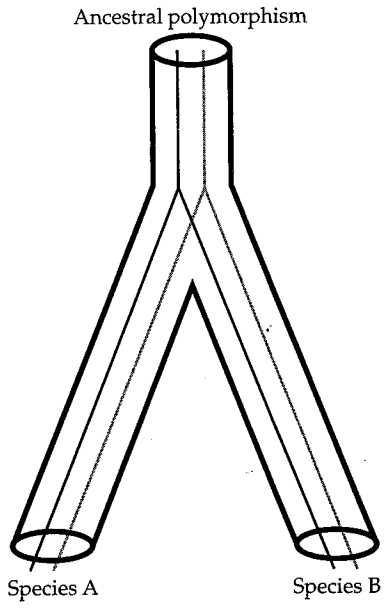


FIGURE 7.24 Trans-species or shared polymorphism may occur if the ancestor was polymorphic for two or more alleles and if alleles persist in both species.

7.9 MULTIGENE FAMILIES

Genes increase in number through duplication. Several successive rounds of duplication result in a family of homologous genes with related functions, a **multigene family**, the members of which are often arrayed in tandem along the chromosome. Among genes that normally exist in tandemly arrayed multigene families are the rRNA genes and the histone genes. Analysis of the sequences of members of multigene families has led to some surprises. Figure 7.25 shows a scenario in which a gene underwent a duplication that ultimately became fixed in the population either through drift or selection. Subsequently, sufficient sequence divergence occurred that the two genes could be distinguished. Later a speciation event produced two different species sharing this pair of genes. Figure 7.26 shows the gene genealogies at two time points in the evolution of this gene family. At Time 1, the *A* duplicates in species 1 and 2 have a more recent common ancestor than do genes *A* and *B* within species 1. At Time 2 the pairs of genes present in the same species are more similar. This is the pattern that is observed in some multigene families. The close resemblance of A_1 with B_1, and of A_2 with B_2, seems paradoxical, since both species have the duplication, and Figure 7.26 makes it appear that genes A_1 and A_2 in the two species have a more recent common ancestor than do genes A_1 and B_1. Genes A_1 and B_1, as well as A_2 and B_2, may have more similar sequences because the genes evolve together, in concert, under the influence of mechanisms that operate to homogenize their sequences through such processes as gene conversion and unequal crossing-over, discussed in the next section. This tendency toward homogenization is known as **concerted evolution**.

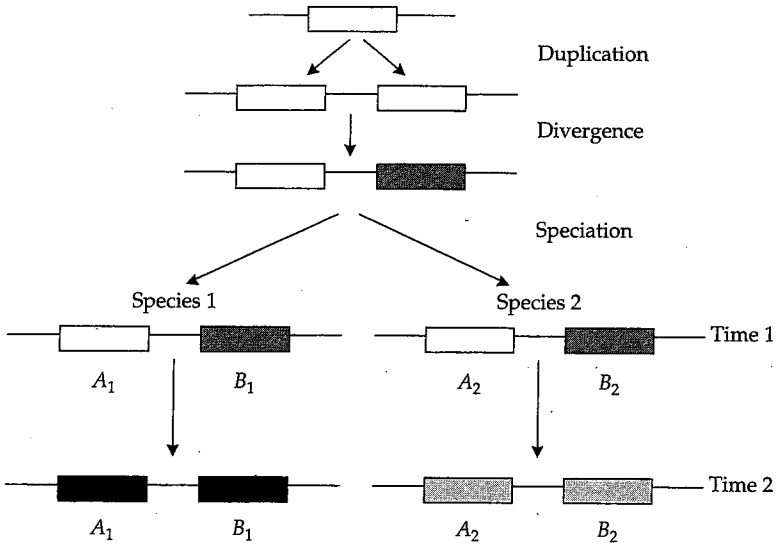


FIGURE 7.25 Multigene families originate by a process of gene duplication. After the duplication, the genes may retain very similar functions (like rRNA genes), or they may diverge (like globin genes). If the species splits into two species, then Time 1 and Time 2 depict the relationship between the genes shortly after speciation and at a much longer time after speciation.

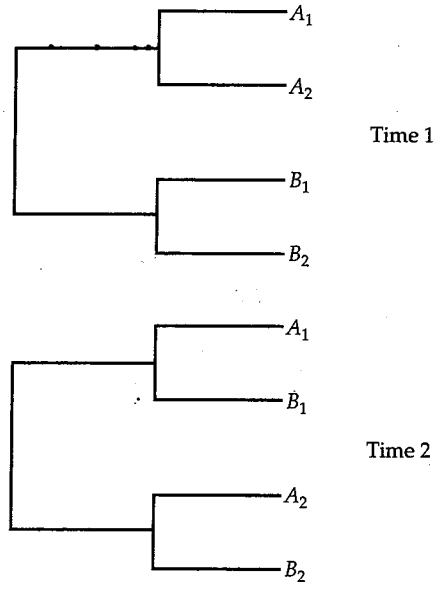


FIGURE 7.26 Referring to Figure 7.25, at Time 1, genes A_1 and A_2 in the two species are more similar to each other than either is to gene B, and likewise B_1 and B_2 are closest neighbors. This tree depicts the fact that the common ancestor of A_1 and A_2 is more recent than that of A_1 and B_1. If at another Time 2 a tree like that in the lower panel is observed, then sequences of A_1 and B_1 have become more similar, possibly by the process of gene conversion. The bottom tree illustrates the phenomenon known as *concerted evolution*.

Concerted Evolution

As noted, two important mechanisms of concerted evolution are gene conversion and unequal crossing-over. **Gene conversion** is a process in which a double-strand break in the DNA is repaired by nucleotide pairing between two sufficiently similar genes (not necessarily the homolog) is accompanied by resynthesis of a portion of the template sequence in a way that results in replacement of sequence at the original site by the invading template. Formally, the result is that the sequence in one gene "converts" the sequence in the other gene to be exactly like itself. In **unequal crossing-over**, meiotic pairing between the tandem repeats in homologous chromosomes is out of register, and crossing-over results in an increase in the number of copies in one chromosome and a corresponding decrease in the number of copies in the other chromosome. Repeated rounds of unequal crossing-over can result in the disproportionate representation of certain sequences among members of the multigene family, a result that is formally equivalent to gene conversion.

A theoretical model of concerted evolution has been studied by Ohta (1982). In this model, a tandemly arranged multigene family consists of a fixed number of n members, and λ is the probability that a particular member of the gene family becomes converted by another member in any one generation. (Equivalently, λ is the probability of completion of a cycle of unequal crossing-over resulting in the replacement of one sequence in the family by another.) The mutation rate per copy is μ, and the population number is N.

In a tandemly arrayed multigene family, there are three distinct ways in which alleles can be identical by descent (IBD) among the gene copies (Figure 7.27):

1. Genes at different positions in the same chromosome may be IBD (probability c_1).
2. Genes at different positions in different chromosomes may be IBD (probability c_2).

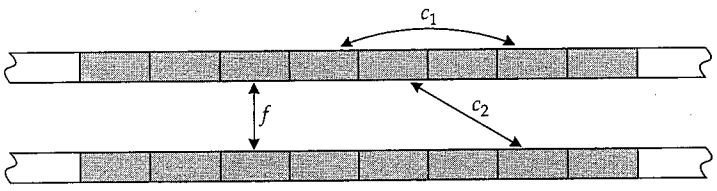


FIGURE 7.27 Three types of identity by descent in multigene families. They are the identity between genes at homologous sites (probability f), between genes at nonhomologous sites in the same chromosome (probability c_1), and between gene at nonhomologous sites in different chromosomes (probability c_2). (After Ohta 1982.)

3. Genes at the same position in different chromosomes may be IBD (probability f).

Complex formulas for the equilibrium values of c_1, c_2, and f have been derived by Ohta (1982), but they are greatly simplified when recombination within the gene cluster is ignored. In such a case, the equilibrium values are approximately

$$c_1 = c_2 = \frac{\lambda}{\lambda + (n-1)\mu}$$
$$f = \frac{4N\lambda c_2 + 1}{4N\lambda + 4N\mu + 1} \qquad (7.18)$$

In Equations 7.18, the quantity $(n-1)\mu$ is very nearly equal to $n\mu$ if n is reasonably large. Because n is the number of copies of the gene in each tandem array, $n\mu$ is the total rate of mutation in the multigene family, summed across all copies. Thus, the implication of Equation 7.18 is that there is a delicate balance between the rate of gene conversion λ and the total mutation rate $n\mu$. If the rate of gene conversion is much greater than the total mutation rate, then the probability of identity by descent of genes at different positions within the family (c_1 and c_2) is close to 1.0. On the other hand, if λ is much smaller than the total mutation rate, then the probability of identity by descent of genes at different positions within the family is close to zero.

Concerted evolution does not homogenize all multigene families. Depending on the balance of the forces of mutation, gene conversion, and unequal crossing-over, the pair of genes may remain active and very similar, or they may diverge in function (such as different tissue-specific forms of amylase or lactate dehydrogenase), or one gene may lose function and become a pseudogene.

Subfunctionalization

The study of multigene families has been reinvigorated by sequencing of multiple complete genomes. Chapters 9 and 10 deal specifically with genome-scale data and how population genetics can provide many useful approaches to analysis. But the issues of duplicated genes and multigene families warrants special attention. Duplicate genes can evolve in separate ways under the influence of natural selection, mutation, and random genetic drift. If natural selection drives the divergence between a pair of recently duplicated genes, then one might expect an excess ratio of nonsynonymous to synonymous differences between such duplicated genes. In many cases, this is exactly what is observed (Thornton and Long 2005).

In time, some members of a multigene family may diverge to a greater or lesser degree in their function. This process of duplication and divergence is thought to be the major mechanism by which genes with novel functions are created. Many members of gene families have tissue-specific isoforms.

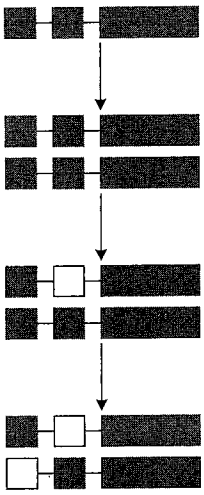


FIGURE 7.28 Consider a gene that is expressed in two tissues because of tissue-specific enhancer sequences located in promoter elements directing gene expression. When this gene undergoes duplication, one copy can suffer a loss-of-function mutation of one tissue-specific enhancer, and the other gene copy can lose expression driven by the other enhancer. The result is called *subfunctionalization*, a process in which one gene is expressed in one tissue and the other gene is expressed in the other tissue.

Enzyme-coding genes in particular tend to have a gene copy for each of several tissues: a liver-specific isoform, a heart-specific isoform, and so on. Rather than being viewed as a strictly positive, gain-of-function process, the divergence in tissue specificity is thought to come about through loss of enhancer elements that drive expression in all tissues, through a process called **subfunctionalization** (Lynch and Force 2000). If the ancestral gene had promoter elements driving expression in tissues A, and B (Figure 7.28), then after duplication, one gene copy can lose the promoter element directing tissue B expression and the other gene copy can lose the other enhancer. The result is that the organism retains expression in both tissues A and B, but from different (**paralogous**) copies of the original gene. Such loss-of-function mutations are much more likely to occur than *de novo* acquisition of expression in new tissues, and subfunctionalization is seen as a likely explanation for many instances of divergence of duplicate gene function.

Birth-and-Death Process

Some multigene families retain a tandemly arrayed structure and similarity in function among members despite the fact that the differences between individual members is of functional significance. This pattern is particularly true of genes in the immune system, including immunoglobulin genes and major histocompatibility genes. Interspecific comparisons of genes in families of this sort exhibit some genes that are clearly homologous, and others that are more distantly related. In addition, the rate of duplication, loss of function through becoming pseudogenes, and loss by deletion may be fairly high. This kind of pattern of multigene family evolution is different from concerted evolution, because the differences between the genes can be high enough that intergenic conversion is very rare. Figure 7.29 illustrates the distinctness of this pattern of gene evolution, called a **birth-and-death process** by Ota and Nei (1994). Many examples of multigene family evolution appear to show the evolutionary patterns expected of this process (Nei and Rooney 2005). The elimination of genes that are no longer essential is evident in the genomes of parasites and symbionts because they get a "free ride" using their host's metabolic machinery (Moran and Degnan 2006).

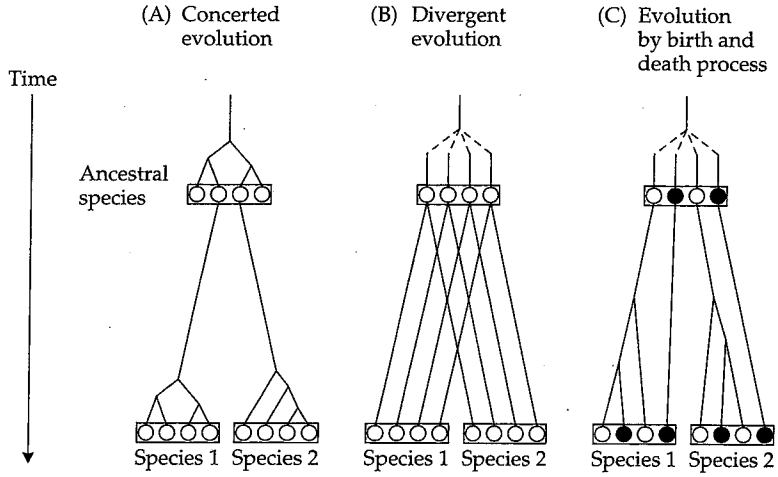


FIGURE 7.29 In addition to concerted evolution (A) and divergent evolution (B), multigene families frequently exhibit the phenomenon of genes being added or lost by a "birth and death process" (C). In this process, some members of the gene family may proliferate owing to successive gene duplications, whereas other members of the gene family may be lost by deletion. (From Ota and Nei 1994.)

Unequal crossing-over in multigene families can result in a decrease in the number of genes as well as an increase. For example, globin genes exist as two tandem arrays in the genomes of most mammals, and many mammals are polymorphic in globin gene copy number. In humans, complete deletion of the β-gene results in death when homozygous, but a partial β deletion and other mutations that decrease the abundance of the β-hemoglobin chain are relatively common in the Mediterranean Sea basin where malaria was common. For this reason, the decreased-β-chain diseases are called β-thalassemias (literally translated as "sea-anemias"). The well-established link between sickle-cell anemia and malaria, along with the geographical correlation between the β-thalassemias and malaria (see Chapter 5), provides a strong circumstantial case for malarial parasites being an important selective agent. Deletion of one or more of the α-globin genes results in another form of anemia called α-thalassemia, whose frequency in populations is also correlated with the incidence of malaria.

Red-green colorblindness is a common X-linked disorder with a frequency of about 5% in males of European ancestry. The genes for the red and green visual pigments match at 98% of their nucleotides, indicating that they arose by a relatively recent duplication. Individuals with normal color vision have one copy of the red pigment gene and varying numbers of

copies of the green pigment gene. When genomic DNA from colorblind males was analyzed by Southern blotting, those defective in green vision were lacking fragments of the green pigment gene. Further analysis showed that 24 of 25 colorblind individuals had lost one or the other pigment gene through gene rearrangements that were due either to unequal crossing-over or gene conversion. In this example, the high sequence similarity of the red and green pigments works to our disadvantage by greatly increasing the likelihood of exchange events that lead to loss of color vision (Nathans et al. 1986). The relationship between the molecular basis of light absorption and perception was made particularly clear when it was found that a normal polymorphism in red pigments, which confers a difference in the absorption peak of the protein product, also confers a measurable difference in the perception of color balance (Merbs and Nathans 1992). The red opsin gene (*OPN1LW*) displays zero amino acid fixed differences between humans and chimpanzees, and yet in a sample of 236 humans there was abundant polymorphism (Verrelli et al. 2004). Moreover, the human population exhibits a significant excess of high-frequency of amino acid replacements in *OPN1LW* in a pattern consistent with inter-genic gene conversion and selection. Evidence for selection on visual pigment genes is strong, including the finding that many vertebrates have vision into the ultraviolet range of the spectrum (Yokoyama et al. 2006).

One of the most abundant gene families in the genomes of many organisms is that for olfactory receptors (ORs). In primates, there are more than one thousand copies of these genes, whose function is to bind to odorants and play a role in recognition of identity of odorant molecules. The exact way that the odorant receptors work is not fully known, but it is clearly more complicated than a one-for-one matching of odorant molecules and their respective receptors, and instead involves some kind of combinatorial code. By a series of specialized PCR amplifications, it is possible to recover a substantial fraction of the OR genes and to characterize their subfamily relationships whether or not they are still active genes (Gilad et al. 2005). By sampling OR genes and assessing the portion that are pseudogenes across 19 primates, an intriguing pattern emerged: humans and Old World monkeys have significantly fewer olfactory receptors than do New World monkeys, with one exception. The exception is a New World monkey with a high proportion of pseudogenes, and interestingly it shares with humans and Old World monkeys a full trichromatic color perception. The hypothesis is that the ORs become less necessary once color vision had developed fully (Gilad et al. 2004). Looking across a broader range of vertebrate phylogeny, mammals have lost several classes of ORs that are present in fishes, another striking example of the fluctuation of gene content of multigene families that occurs by the birth and death process (Niimura and Nei 2006).

SUMMARY

1. Estimating substitution rates requires accounting for mutations recurring at the same site.
2. Even the simplest nucleotide substitution model of Jukes and Cantor has subtle complications. More realistic models account for differences in rates of transition and transversion substitutions as well as other features of mutations.
3. Many sequences appear to diverge at an approximately clocklike rate. The molecular-clock concept should be interpreted somewhat loosely. Some lineages appear to have accelerated or decelerated clock rates. One cause of rate variation is a change in generation time (for example, from rodents to primates).
4. Synonymous and nonsynonymous substitutions have different effects on the protein product, so estimating the rates of these two kinds of substitution independently can be informative about the causes of evolutionary change.
5. The neutral theory makes predictions about the shape of gene trees; there have been important advances in testing hypotheses about evolutionary forces based on inferred gene genealogies.
6. Organelle genome evolution occupies an important position in the development of molecular population genetics, in part because of the numerous studies of mtDNA and cpDNA variation.
7. Molecular phylogenetics seeks to reconstruct the ancestral history of extant organisms, and shares many analytical procedures with molecular population genetics. There are several widely used algorithms for reconstructing trees from sequence data, including the UPGMA method, least-squares, neighbor-joining, parsimony methods, and Bayesian methods.
8. When multiple copies of similar genes exist in the genome, they can exchange sequences through unequal recombination and gene conversion. Such exchanges can result in concerted evolution, a process whereby genes in a multigene family are very similar to one another within a species, even though the duplication events that gave rise to the family occurred in the distant past.
9. Different sublineages of multigene families can undergo differential proliferation or dimunition owing to gene duplications and deletions constituting a birth and death process.

PROBLEMS

1. What inference would you make regarding the selective constraints on a region of DNA in which the rate of evolution was 5×10^{-9} nucleotide substitutions per site per year?

2. Among the 61 codons of the genetic code that specify an amino acid, how many have third-position nucleotides that are nondegenerate? Twofold degenerate? Threefold degenerate? Fourfold degenerate? Assuming equal usage of all codons. what is the average degeneracy at the third position?
3. In Equation 7.15 a value of $d = 0.5523$ results in $k = 1$.
   (a) What is the biological meaning of $k = 1$?
   (b) Assuming that the number of nucleotide substitutions at a site follows a Poisson distribution with mean k, then the probability that a site has undergone exactly i nucleotide substitution equals $(k^i/i!)e^{-k}$. For a sequence of nucleotides evolving independently, at a time when $k = 1$ what proportion of the sites will have undergone no substitutions? Exactly one substitution? Exactly two substitutions? More than two substitutions?
4. In Equation 7.15, what is the largest value of d that makes biological sense, and what does it mean?
5. What might one infer about the evolutionary forces affecting a coding region in which the rate of amino acid replacement was greater than the rate of synonymous nucleotide substitution?
6. Ribosomal RNA forms a complex folded structure in which many regions of the molecule are double-stranded because they contain complementary nucleotide sequences. What pattern of nucleotide sequence evolution might be expected in these paired regions?
7. If the rate of nucleotide evolution along an evolutionary lineage is 0.5% per million years, what is the rate of substitution per nucleotide per year? What is the total rate of divergence of two lineages?
8. Consider sequences of gene A and gene B from each of two species. The fraction of sites that differ in gene A is 0.05 and the fraction of nucleotide sites that differ in gene B is 0.70. Apply the Jukes-Cantor correction for multiple hits to obtain the estimate of the number of substitutions per site for each gene. Which gene do you think would have a smaller estimate of variance of substitution rate? Why?
9. The genome of the malaria parasite *Plasmodium falciparum* is extraordinarily rich in A–T base pairs, which is thought to result from an extreme mutational bias toward A–T base pairs. Across the genome as a whole the percentage of A–T base pairs is approximately 85%, but in some noncoding regions it approaches 100%. Consider an aligned region of the genome consisting exclusively of A–T base pairs in two related species, and suppose that only mutations of A → T or T → A take place and that these mutations occur at equal frequency. What equation should be used in place of Equation 7.15 to correct for multiple mutations at the same nucleotide site?

10. A marine biologist happens upon two related species of deep sea creatures that have very unusual DNA containing six possible base pairs instead of four. The base pairs A–T and G–C are found as in usual DNA duplexes, but base pairs X–Z are also found containing the novel bases X and Z. Suppose that homologous DNA sequences from the two species differ in 20% of their base pairs. Derive an expression analogous to the Jukes-Cantor correction (see Equation 7.15) for these species, and use it to estimate the number of substitutions per site that have occurred since the common ancestor of the two species.

11. The following is a small portion of the gene coding for 6-phosphogluconate dehydrogenase in two natural isolates of *Escherichia coli*.

```
CTGACCAAAATCGCCGCCGTAGCTGAAGACGGTGAACCATGCGTTACCTATATTGGTGCC
CTGAAGCAGATCGCGGCGGTTGCTGAAGACGGTGAGCCGTGTGTGACTTATATAGGTGCC
```

Infer the correct translational reading frame of the sequences and estimate

(a) the number of amino acid differences per amino acid site.

(b) the number of nucleotide differences per nucleotide site

(c) the number of nonsynonymous substitutions per nonsynonymous site (counting the number of nonsynonymous sites as equal to the number of nondegenerate sites plus $\frac{2}{3}$ the number of twofold degenerate sites plus $\frac{1}{3}$ the number of threefold degenerate sites).

(d) the number of synonymous substitutions per synonymous site (counting the number of nonsynonymous sites as equal to the number of fourfold degenerate sites plus $\frac{2}{3}$ the number of threefold degenerate sites plus $\frac{1}{3}$ the number of twofold degenerate sites).

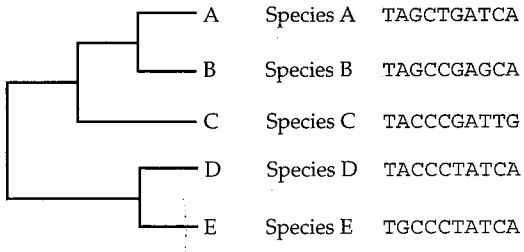
12. The mountain grasshopper *Podisma pedestris* has a genome size more than 100 times larger than that of the fruit fly *Drosophila melanogaster* (1815 Mb vs 176 Mb). Part of the reason for the difference may be that the rate of spontaneous deletion of nonessential DNA in *P. pedestris* is much less than that in *D. melanogaster*. Bensasson et al. (2001) have estimated the rates of DNA loss in the two species at 0.0063 per base pair per million years in *P. pedestris* and at 0.0486 per nucleotide per million years in *D. melanogaster*. Use these estimates to calculate the half-life of a nucleotide present in a nonessential DNA sequence in the two species. (The half-life is the time at which the probability of loss of the nucleotide by deletion equals $\frac{1}{2}$.)

13. The human immunodeficiency virus HIV is the cause of acquired immune deficiency syndrome (AIDS). In this virus the rate of nucleotide evolution has been estimated at about 0.01 substitutions per synonymous site per year. Two viruses isolated in 1983 in Zaire and San Francisco differ in approximately one-third of their synonymous sites. Estimate the year in which the viruses last shared a common ancestor.

14. The data below give the proportion of nucleotide sites that differ in a gene in four RNA viruses. HIV1 and HIV2 are two rather distinct types of human immunodeficiency viruses, VISNA is a lentivirus, and MMLV is a mouse cancer-causing virus. Estimate the number of nucleotide substitutions per site using these data. What do the numbers imply about the evolutionary relationships among the viruses?

	HIV2	VISNA	MMLV
HIV1	0.34	0.54	0.62
HIV2		0.52	0.63
VISNA			0.63

15. The phylogeny of five species A–E is shown here along with a set of aligned nucleotide sequences sampled from each species. Number the nucleotide sites from left to right as 1–10, and for each variable site identify the most likely branch in the phylogeny along which a mutation took place at that site.



16. For a single copy of new neutral mutation destined to be fixed in an ideal diploid population, the average time to fixation is approximately $4N$ generations.

    (a) What is the average number of copies of the neutral mutation that will ever exist from the time of its origin by mutation to the time of its fixation by random genetic drift?

    (b) Suppose that $N = 10^6$ and that the mutation rate per nucleotide per generation is approximately 5×10^{-8}. How many mutations are expected to occur at each nucleotide site in a gene containing a neutral mutation from the time of its origin by mutation to the time of its fixation by drift?

17. For the model of gene conversion with gene identities given in Equation 7.18, what value of λ makes the organization of the gene family irrelevant in the sense that $f = c_1 = c_2$? What is the value of f in this case? The equations assume that $4N\mu << 1$, and λ is the probability that a particular member of the gene family becomes converted in any one generation.

18. For the model of gene conversion with gene identities given in Equation 7.18, what are the values of f and $c_1 = c_2$ when $\lambda = \mu$? (The equations assume that $4N\mu << 1$.)

19. Consider a gene that is expressed in two distinct tissue types A and B owing to the presence of distinct A-specific and B-specific enhancers of expression in the upstream regulatory region. In the process of subfunctionalization, such a gene undergoes duplication followed by the fixation of mutations in such a way that the A-type enhancer is knocked out in one copy and the B-type enhancer is knocked out in the other. Assume that such a gene has undergone duplication and that f_A, f_B, and f_C are the relative probabilities of fixation of knockout mutations in the type-A enhancer, the type-B enhancer, and the coding sequence, respectively. In how many ways can exactly two knockouts be fixed in the two genes? What is the probability that exactly two knockouts will result in subfunctionalization?

8.1 Types of Quantitative Traits 387

8.2 Resemblance between Relatives and the Concept of Heritability 388

8.3 Artificial Selection and Realized Heritability 394

Contribution of New Mutations to the Response to Selection 396

Prediction Equation for Individual Selection 397

Limits to Selection 400

8.4 Genetic Models for Quantitative Traits 403

Change in Allele Frequency 412

Change in Mean Phenotype 413

Linearity of Response 415

8.5 Components of Phenotypic Variance 416

Genetic and Environmental Sources of Variation 416

Components of Genotypic Variation 422

8.6 Covariance among Relatives 426

Twin Studies and Inferences of Heritability in Humans 430

Estimation of Genetic Variance Components in Natural Populations 432

8.7 Norm of Reaction, Threshold Traits, and Genetic Correlation 433

Norm of Reaction and Phenotypic Plasticity 433

Threshold Traits: Genes as Risk Factors in Disease 437

Genetic Correlation and Correlated Response 439

8.8 Evolutionary Quantitative Genetics 443

Inference of Selection from Phenotypic Data 443

Evolution of Multiple Intercorrelated Traits 444

Random Genetic Drift and Phenotypic Evolution 446

Mutational Variance and Mutation-Accumulation Experiments 448

Mutation-Selection Balance for Quantitative Traits 449

8.9 Genes that Affect Quantitative Traits 452

The Number of Genes Affecting Quantitative Traits 453

Methods for Mapping QTLs 455

CHAPTER 8

EVOLUTIONARY QUANTITATIVE GENETICS

Many important problems in evolutionary biology focus on phenotypic variation. Characteristic differences in phenotype distinguish one species from another, and within a species phenotypic variation distinguishes one individual from the next. Darwin formulated his ideas about evolution by means of natural selection based on observations of phenotypic variation, both in natural populations and in populations of domesticated pigeons, rabbits, cattle, sheep, and many other organisms. For many years he struggled to explain the causes of the phenotypic variation, but he was unaware of Mendelian genetics and ultimately proposed an incorrect theory of inheritance. Darwin did, however, appreciate the importance of the commonplace observation that offspring resemble their parents. Most continuously varying traits, like height and weight, are **multifactorial traits**, meaning that they are influenced by multiple genetic factors, multiple environmental factors, and the interactions between them. They are also known as **quantitative traits** because the phenotype variation is quantitative in the sense that the phenotypes form a graded series from one extreme to the other. The study of quantitative traits constitutes **quantitative genetics**.

Genetic experiments demonstrate the multifactorial nature of quantitative traits, because the offspring do not group into discrete classes with familiar Mendelian ratios. Instead, what is observed among the graded phenotypes is a statistical resemblance between parents and offspring, with the magnitude of the resemblance suggesting the relative importance of

genetic factors compared with environmental factors in accounting for phenotypic variation in the trait. Multifactorial traits are also called **complex traits** to emphasize that their inheritance is more complex than that of simple Mendelian inheritance.

At one time, the study of complex traits was exclusively statistical. A quantitative trait in a population was characterized by its mean, variance, covariance between parents and offspring, the phenotypic correlations among individuals with various degrees of genetic relationship, and so forth. Although these statistical attributes could be formulated in terms of alleles, allele effects, allele frequencies, dominance, gene interactions, and other genetic concepts, there was no means of determining how many genes contributed to the trait, where they were located in the genome, what their functions were, and how they interacted with each other and with the environment. This situation has changed dramatically for several reasons, chief among them being the ability to identify large numbers of polymorphic molecular markers and to carry out large-scale genotyping of individuals. These capabilities enable polymorphisms that are genetically linked with genes affecting any quantitative trait to be identified. A gene affecting a quantitative trait is known as a **quantitative trait locus (QTL)**, and from the genetic linkage of a QTL with known marker genes, further studies can be used to identify the QTL itself. In many organisms of interest, the complete genomic DNA sequence is known, and **candidate genes** can be identified whose functions themselves suggests that these genes may affect the quantitative trait.

In the study of quantitative traits, variation among individuals within populations is paramount. Many genes may be critical for the development or physiology of a trait and yet contribute little or nothing to phenotypic variation in the trait. For example, a number of genes are known that are essential for the embryonic development of the human eye, but these genes are monomorphic and so do not contribute to variation in visual acuity. In other words, in quantitative genetics we are primarily interested in polymorphic genes whose alleles account for the normal range of phenotypic variation in a trait, and less interested in rare mutant alleles that may cause variation outside the normal range.

Quantitative genetics is also essential to understanding phenotypic evolution in natural populations. Natural selection operates on variation in phenotype, and because most traits that show variation in phenotype in natural populations are multifactorial, the principles of quantitative genetics prevail. For natural selection, fitness is the preeminent trait, and it is manifestly a quantitative trait because it is affected by the alleles of very many genes and their interactions. One of the great challenges for evolutionary genetics is to unify the principles governing evolution at the phenotypic level with those governing evolution at the molecular level, and in this regard the applications of genomic resources and methods give reason for optimism.

8.1 TYPES OF QUANTITATIVE TRAITS

Quantitative traits are of utmost importance to plant and animal breeders, because agriculturally important characteristics such as yield of grain, egg production, milk production, efficiency of food utilization, and meat quality are all quantitative traits. Even as modern molecular methods are applied to animal and plant improvement, quantitative genetics continues to play an important role because commercially desirable traits result from complex interactions among many genes. In addition to being essential to plant and animal improvement programs, the principles of quantitative genetics, appropriately modified and interpreted, can be applied to the analysis of quantitative traits in humans and natural populations of plants and animals.

It is convenient to distinguish three types of quantitative traits:

1. Traits for which there is a continuum of possible phenotypes are called **continuous traits** or **metric traits**; examples include height, weight, milk yield, and growth rate. The distinguishing feature of continuous traits is that the phenotype can take on any one of a continuous range of values. In theory, there are infinitely many possible phenotypes, among which discrimination is limited only by the precision of the instrument used for measurement. However, in practice, similar phenotypes are often grouped together for purposes of analysis.
2. Traits for which the phenotype is expressed in discrete, integral classes are called **categorical traits**; examples include number of offspring, number of ears on a stalk of corn, number of petals on a flower, and number of bristles on a fruit fly. The distinguishing feature of a categorical trait is that the phenotype of an individual is an integer value determined by counting. For example, a popular categorical trait in *Drosophila* is the number of bristles that occur on the abdominal segments or sternites. Normally there are 14–24 bristles per sternite. A male with 19 bristles on the fifth abdominal sternite therefore has a phenotype of 19. The distribution of numbers of abdominal bristles in a sample of *Drosophila* appears in Figure 8.1. When the number of possible phenotypes of a categorical trait is large (as it is with abdominal bristle number) then the line between continuous traits and categorical traits becomes indistinct.
3. The third category of quantitative traits consists of **dichotomous traits**, which are either present or absent in any one individual. Dichotomous traits are also called **threshold traits** because, while the visible phenotype is either present or absent, presence of the trait is determined by an underlying risk or **liability** toward the trait, a continuous variable that is affected by multiple genetic and environmental factors and their interactions. The liability values are not directly observable, but an individual with a liability value greater than some **threshold**, or triggering level, will actually express the trait. In this way an underlying continuous trait (liability) becomes manifested as a discrete trait (presence or absence),

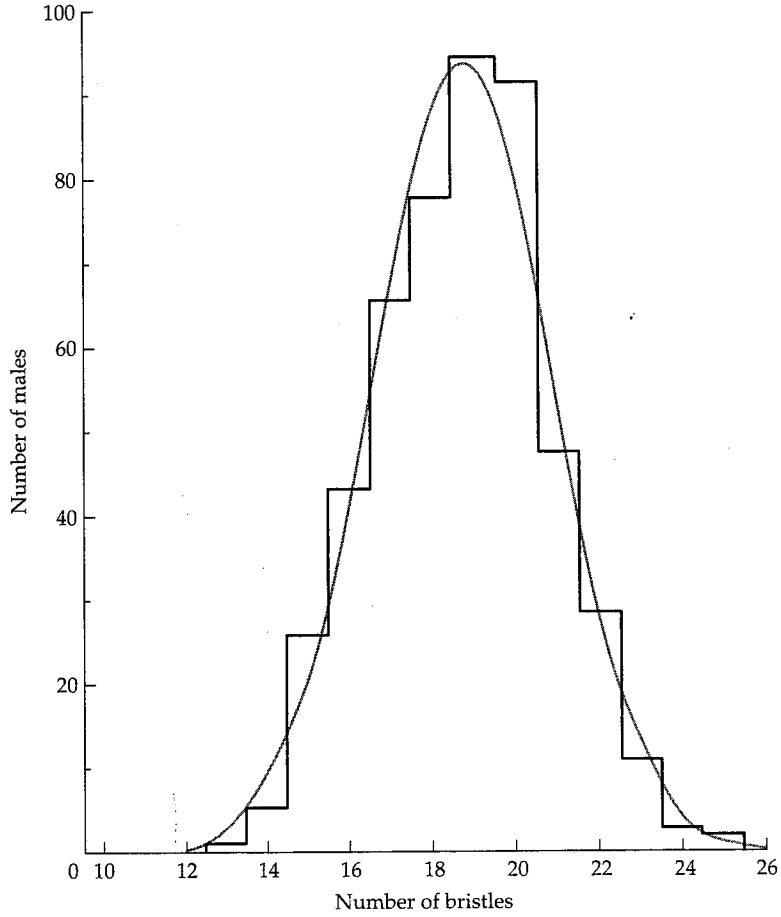


FIGURE 8.1 Number of bristles on the fifth abdominal segment (sternite) in males of a strain of *Drosophila melanogaster*. The smooth curve is that of a normal distribution with mean 18.7 and standard deviation 2.1. (Data courtesy of Trudy F. C. Mackay.)

owing to the threshold of liability. Threshold traits include many complex human diseases, such as schizophrenia and the most common form of diabetes. With threshold traits, studies of affected individuals and their relatives permit inferences to be made about the underlying values of liability. These methods are discussed later in this chapter.

8.2 RESEMBLANCE BETWEEN RELATIVES AND THE CONCEPT OF HERITABILITY

That offspring tend to resemble their parents is an everyday observation. Even before the rediscovery of Mendel's work, Francis Galton (a half-cousin

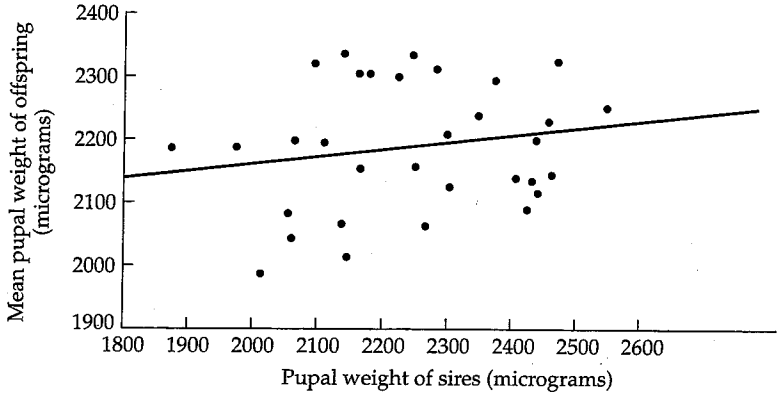


FIGURE 8.2 Mean weight of male pupae of the flour beetle *Tribolium castaneum*, against pupal weight of father (sire). Each point is the mean of about eight male offspring. The regression coefficient of male offspring weight on sire's weight is $b = 0.11$, and h^2 is estimated as $2b = 0.22$. (Data courtesy of F. D. Enfield.)

of Charles Darwin) was collecting detailed statistical data on resemblance between parents and offspring (Galton 1889). Galton's observations revealed important aspects of the transmission of quantitative traits, but an interpretation of his empirical rules in terms of Mendelian genetics came almost three decades later (Fisher 1918). Notoriously difficult, Fisher's 1918 paper was of great historical importance to population genetics because it provided the first demonstration that multiple Mendelian genes could account for the observed patterns of transmission of quantitative traits.

Galton was a pioneer in the application of statistics to biology; to illustrate some of his key ideas we will use Figure 8.2, which shows a plot of the mean of male offspring for a quantitative trait (y values) against the phenotypic value of the father (x values), displayed in a manner that Galton devised. The plot is called a *scattergram* because it shows the scatter among the data points. The magnitude of the scatter for either variable, for example x, is measured by the variance in x, which is defined as $\sigma_x^2 = E[x - E(x)]^2$, where as usual the symbol E stands for expectation (the mean or average). In words, the variance among the x values equals the average of the squared deviation from the mean. We use the symbol σ_x^2 to draw attention to an important distinction between parameters and estimates. The **parameters** of a distribution are those constants that characterize the distribution in an entire population, such as the mean (usually denoted μ) and the variance (usually denoted σ^2). The values of the parameters are almost always unknown, but inferences about the likely values of the parameters can be

made from samples taken from the population. The numerical inferences from samples are **estimates** of the parameters, and they will differ from one sample to the next solely because of chance. An *unbiased estimator* is one whose expected value equals the value of the parameter. For example, the sample mean $\bar{x}$ is an unbiased estimator of the parameter μ because $E(\bar{x}) = \mu$. Because $E[x - E(x)]^2 = E(x^2) - [E(x)]^2$, an obvious estimator for the variance is given by

$$Var(x) = \overline{x^2} - \bar{x}^2 \tag{8.1}$$

where $\overline{x^2}$ means the mean of the squared values of x in the sample and $\bar{x}^2$ is the square of the mean in the sample. Actually, $Var(x)$ in Equation 8.1 is a biased estimator of σ^2 because $E[Var(x)] = (n-1)\sigma^2/n$, where n is the sample size. An unbiased estimator of σ^2 for small samples is therefore $[n/(n-1)]Var(x)$, but this adjustment is negligible unless the sample size is smaller than about 50.

What caught Galton's attention in scattergrams like that in Figure 8.2 is that there is a systematic relationship between the values of x and y, indicated by the straight line. To describe this situation Galton invented the concept of **regression**, an idea that is so pervasive in quantitative genetics that we need to describe it in some detail. For plots like that in Figure 8.2, the best-fitting straight line is called the **regression line**, and in this case it is the regression line of offspring phenotype on that of the father. For reasons that will become clear in a moment, we are interested in the slope of the regression line. The slope is most easily expressed in terms of the **covariance** of x and y, defined as $\sigma_{xy} = E(xy) - E(x)E(y)$, where again E means expectation and the Greek letter signifies that the covariance is a parameter of the joint distribution of x and y. An estimate of the covariance can be obtained by substituting sample means for the expectations, hence

$$Cov(x, y) = \overline{xy} - \bar{x}\,\bar{y} \tag{8.2}$$

where $\overline{xy}$ is the mean of the product of x and y in the sample and $\bar{x}\,\bar{y}$ is the product of the means. There is again a slight bias because the expected value of $Cov(x, y)$ in Equation 8.2 is $E[Cov(x, y)] = (n-1)\sigma_{xy}/n$, where n is the sample size, and so an unbiased estimator of σ_{xy} for small samples is $[n/(n-1)]Cov(x, y)$.

The covariance is a convenient measure of the degree of association between x and y. If x and y are independent, then σ_{xy} equals zero. Since the covariance between any two variables measures their degree of association, the covariance may be positive or negative. Positive covariance means that values of x and y tend to increase or decrease together; negative covariance means that, as one variable increases, the other tends to decrease. The limiting values of the covariance are $-\sigma_x\sigma_y$ on the negative side, and $+\sigma_x\sigma_y$ on the positive side. These limits are achieved only when the variables demonstrate a perfect linear relationship with each other.

Galton realized that the regression line is important because it contains information about the inheritance of the trait, and in particular it indicates how closely offspring resemble their parents. In a scattergram like that in Figure 8.2, the best-fitting straight line is a line that minimizes the sum of the squared deviation of each point from the line. To be specific, let the best-fitting straight line be given by $a + bx$. Then the distance between any value of y and the corresponding point on the line is equal to $y - (a + bx)$. We therefore seek to find values of a and b that minimize the expected sum of the squared deviations (SS), defined as

$$SS = E[y - (a + bx)]^2 \tag{8.3}$$

Students who have studied calculus may recall that the way to find the values of a and b is to differentiate Equation 8.3 separately with respect to a and b, set the resulting equations equal to 0, and then solve for a and b. The derivatives, when set equal to 0, are

$$\frac{\partial SS}{\partial a} = -2E(y - a - bx) = -2[E(y) - a - bE(x)] = 0 \tag{8.4a}$$

$$\frac{\partial SS}{\partial b} = -2E[x(y - a - bx)] = -2[E(xy) - aE(x) - bE(x^2)] = 0 \tag{8.4b}$$

Equation 8.4a implies that $a = E(y) - bE(x)$, and when this is substituted into Equation 8.4b the resulting equation for b is

$$[E(xy) - E(x)E(y)] - b[E(x^2) - [E(x)]^2] = 0 \tag{8.5}$$

Note that the first term in square brackets equals the covariance σ_{xy} and that the second term in square brackets equals σ_x^2. Hence, the slope of the best-fitting straight line is $b = \sigma_{xy}/\sigma_x^2$, and this is estimated from Equations 8.1 and 8.2 as

$$\hat{b} = \frac{Cov(x, y)}{Var(x)} \tag{8.6}$$

where the circumflex is a widely used symbol for an estimated quantity. The slope in Equation 8.6 is called the **regression coefficient** of offspring on one parent. A related quantity that also arises in quantitative genetics is the product-moment correlation coefficient, often simply referred to as the **correlation coefficient**, which is usually symbolized as r and estimated as $\hat{r} = Cov(x, y) / \sqrt{[Var(x)\ Var(y)]}$. Note that, while the covariance, the regression coefficient, and the correlation coefficient all measure aspects of association between x and y, the concepts are quite different. For example, the covariance and the regression coefficient are unbounded, whereas the correlation coefficient must be a value between –1 and +1.

A graphical interpretation of regression is illustrated in Figure 8.3, which shows the distribution, in two dimensions, of the variables x and y. The variables may represent, for example, the phenotypic values of parents (x) and offspring (y). When there is no association between x and y, the distribution

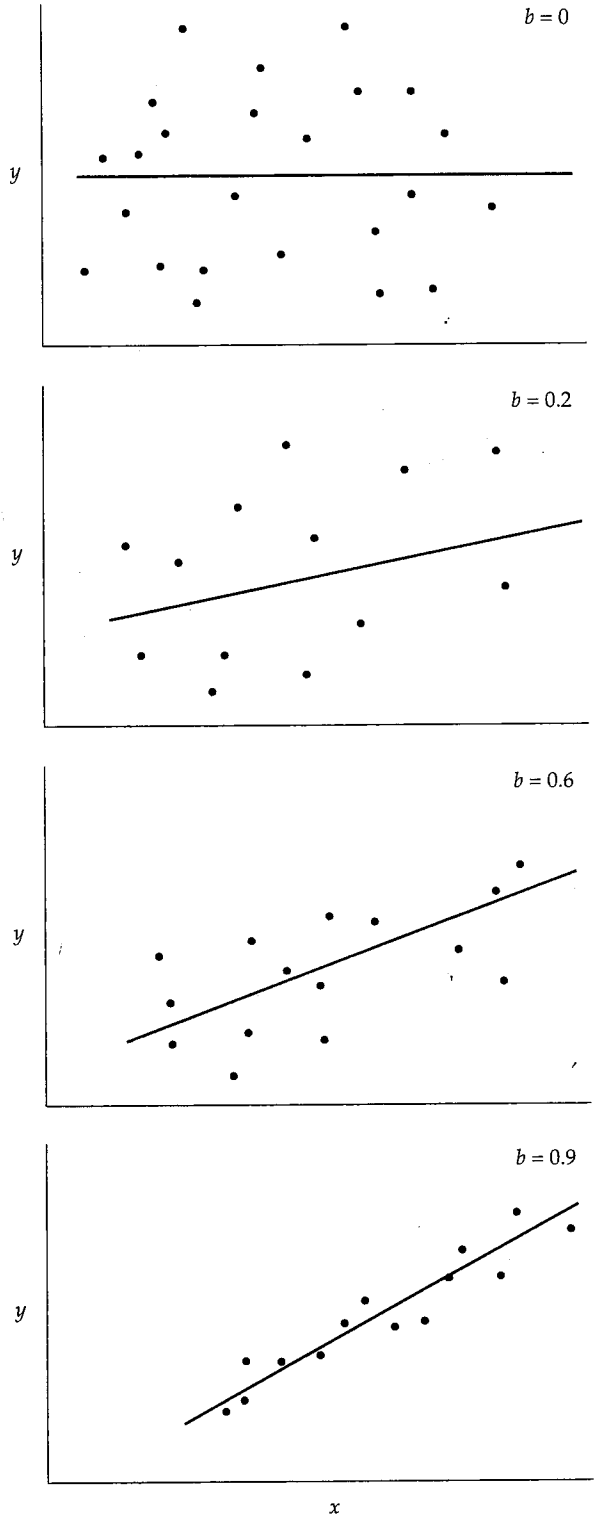


FIGURE 8.3 Plots of random scatters of points having the same variance on the x axis but a range of covariances. With zero covariance (top), the regression coefficient is zero. A stronger linear trend results in a higher regression coefficient.

is a random scatter of points, and any line through the points fits equally badly. Figure 8.3 shows the appearance of the scatter of points for different values of association between the two variables.

Two extreme examples may help clarify parent-offspring regression. At one extreme, if there were no genetic contribution to the trait, then the scattergram might appear as a random scatter as in the top panel of Figure 8.3 with no tendency to follow a line. In such a case, knowing the phenotype of the parents would not help to predict that of the offspring, because there would be no parent-offspring resemblance. On the other hand, a substantial tendency to follow a line is no guarantee that the trait is determined by genes. To see why this is so, consider genetically identical families living in different environments. In favorable environments with plenty of food and resources, parents and offspring might all be big and strong, while in unfavorable environments, parents and offspring might be small and sickly. A parent-offspring plot would show that big strong parents have big strong offspring, while small sickly parents have small sickly offspring, even though there is absolutely no genetic basis for the difference. The tendency of points to follow a line in a parent-offspring scattergram tells us nothing about the genetic basis of phenotypic variation, unless we are willing to make some claims (which hopefully could be tested experimentally) about the environmental covariance (the tendency of parents and offspring to resemble one another due to shared environments). Only if there is no environmental covariance will the parent-offspring regression indicate a degree of genetic influence on the resemblance. The possibility of environmental covariance is absolutely critical in human quantitative genetics (see Chapter 10), where the influence of shared environments can be very strong but difficult to measure.

Assuming now that the environmental covariance is zero, the regression coefficient b of offspring on one parent can be calculated for any random-mating population, and it indicates the degree to which the phenotypic variance in the trait is determined by genetic variation. It is for this reason that the regression coefficient is related to an important quantity in quantitative genetics called **heritability**. There are two types of heritability that will be distinguished later, but for now we note that the "narrow-sense" heritability (h^2) can be estimated from the relationship

$$b = (\tfrac{1}{2})\, h^2 \qquad (8.7)$$

The factor of $\frac{1}{2}$ occurs in Equation 8.7 because the regression involves only a single parent (the father, in the case of Figure 8.2), and only half of the genes from any one parent are passed on to the offspring. In Figure 8.2, $\hat{b} = 0.11$, and so $\hat{h}^2 = 0.22$. Notice the considerable scatter among the points in Figure 8.2, which represents data from 32 families. Because this sort of scatter is typical, heritability estimates tend to be quite imprecise unless based on data from several hundred families. On the other hand, with even an enormous sample there would be no less scatter: the larger sample would merely provide a more accurate measure of how much scatter there is. One further point about

Figure 8.2: In organisms such as mammals, the regression is better performed on the father's phenotype, rather than on the mother's, in order to avoid potential bias in the estimate of heritability caused by such maternal effects as intrauterine environment. In organisms where nurturing does not impart significant maternal effects, scattergrams can be constructed with the x-axis being the average of the two parents (the **midparent**) and the y-axis the offspring phenotypes. From this sort of plot the regression coefficient is equal to the heritability: In symbols, when the x-axis is the midparent, then $b = h^2$.

PROBLEM 8.1 The data below are values of shell breadth in 119 sibships of the snail *Arianta arbustorum*, grouped into six categories for computational convenience (Cook 1965). Estimate the narrow-sense heritability of shell breadth from these data by means of parent-offspring regression.

Number of sibships	Midparent value (mm)	Offspring mean (mm)
22	16.25	17.73
31	18.75	19.15
48	21.25	20.73
11	23.75	22.84
4	26.25	23.75
3	28.75	25.42

ANSWER The sample values needed are $\bar{x} = 20.26$, $\bar{y} = 20.18$, $\overline{x^2} = 418.68$, and $\overline{xy} = 414.01$.

Then $Cov(x, y)$ is estimated as $\overline{xy} - \bar{x}\bar{y} = 5.14$ and $Var(x)$ is estimated as $\overline{x^2} - (\bar{x})^2 = 8.11$. The regression coefficient of offspring on midparent is $\hat{b} = Cov(x, y)/Var(x) = 0.63$, and for the midparent regression, $\hat{h}^2 = \hat{b} = 0.63$. Your values may differ slightly from these due to round-off error. In actual practice, we might not want to group the data into categories because there is some loss of accuracy from grouping. The regression coefficient for the ungrouped data is $\hat{b} = 0.70$. In addition, it should be noted that there is substantial assortative mating for shell breadth, and so the heritability estimate is artificially large.

8.3 ARTIFICIAL SELECTION AND REALIZED HERITABILITY

The deliberate choice of a select group of individuals to be used for breeding constitutes **artificial selection**. The most common type of artificial selection is **directional selection**, in which phenotypically superior animals or plants are chosen for breeding. Although artificial selection has been practiced successfully for thousands of years (for example, in the body size of domesticated dogs), only in the twentieth century did the genetic principles underlying its successes become clear. Understanding the genetic principles of artificial selection permits prediction of the rapidity and amount by which a population can be altered through artificial selection in any particular generation or small number of generations. The theory of artificial selection is also strongly motivated by

the idea that natural selection may operate in a similar way. For example, if only those individuals with greater than a certain amount of body fat survive, or only those individuals with less than a critical rate of evaporative water loss survive, then natural selection acts on the distribution of phenotypes in much the same way that breeders select characters of agricultural importance.

Artificial selection in outcrossing, genetically heterogeneous populations is usually successful in that the mean phenotype of the population changes over generations in the direction of selection, provided the population has not previously been subjected to long-term artificial selection for the same trait. In experimental animals, the mean of almost any quantitative trait can be altered in whatever direction desired by artificial selection. For example, in *Drosophila*, body size, wing size, bristle number, growth rate, egg production, insecticide resistance, and many other traits can be increased or decreased by selection. In domesticated animals and plants, birth weight, growth rate, milk production, egg production, grain yield, and countless other traits respond to selection. Figure 8.4 shows the results of a long-term

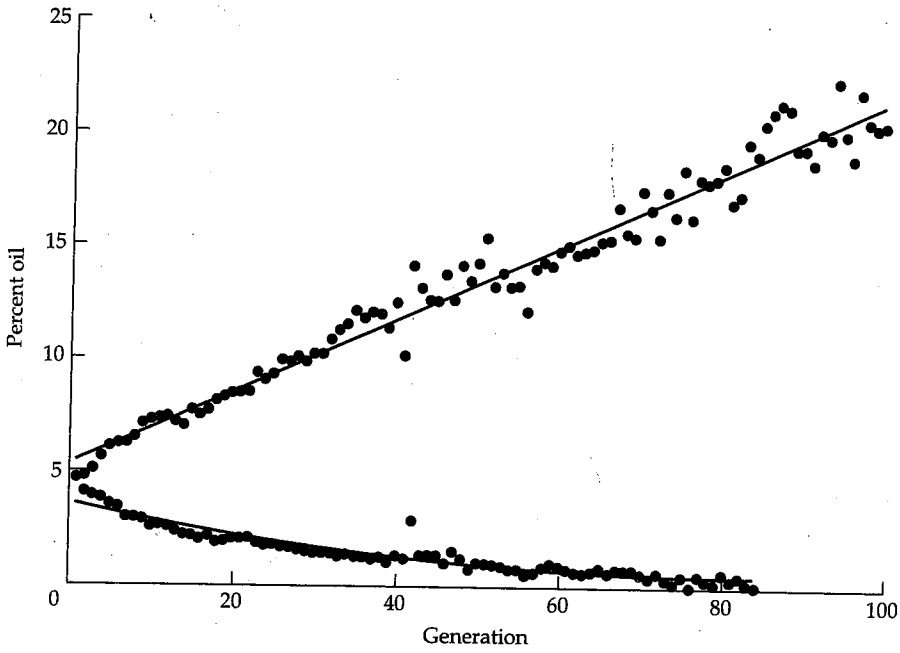


FIGURE 8.4 Results of a famous long-term experiment selecting for high and low oil content in corn seeds. Begun in 1896, the experiment has the longest duration of any on record and still continues at the University of Illinois. Note the steady, linear rise in oil content shown in the up-selected line. The curve is the best linear fit. The down-selected line started on a roughly linear path and continued so for about ten generations, but then the response tapered off, presumably because zero percent oil is an absolute lower limit for the trait. The curve here is the best quadratic fit. (Data courtesy of J. W. Dudley.)

selection program for oil content in corn, *Zea mays*. Begun in 1896, separate populations of maize were selected for either high or low oil content (Dudley and Lambert 2004; Laurie et al. 2004; Hill 2005). In each generation, about 60 ears in each population were assayed for oil content, and seeds from the 12 highest (high line) or lowest (low line) were planted. The initial populations began with a mean of 4.6% oil and a standard deviation of 0.4%. Selection in the high line continued at a steady pace for 100 generations (the experiment is still going on), and in generation 100 the mean was 20.4%, which is much greater than levels found in cultivated maize. But the high line is a Johnny one-note selected only for high oil content, and its yield and other agronomic characters are poor. The low line also responded to selection until the selection was finally terminated when the oil content became so low that it could no longer be measured accurately.

Contribution of New Mutations to the Response to Selection

The general effectiveness of artificial selection in outcrossing species indicates that a wealth of genetic variation affecting quantitative traits exists. By contrast, in a genetically uniform population, such as an inbred line, the mean phenotype of the population usually changes slowly if at all through artificial selection because genetic variation is required to change a trait, and in an inbred line this means that progress under selection is limited by the occurrence of rare mutations. For example, in classical selection experiments in homozygous lines of the Princess bean, *Psophocarpus tetragonolobus*, Johanssen (1909) found that artificial selection had little or no effect on seed weight.

On the other hand, progress under selection can occur in inbred lines if the number of QTLs is large enough or the mutation rate of the QTLs is high enough. The magnitude of phenotypic variation generated through new mutations is discussed in detail in Lynch and Walsh (1998), Mackay (2001), and Barton and Keightley (2002), and we shall have more to say about it later in this chapter. Here we will note that Walsh (2004) has argued that part of the maize selection response in Figure 8.4 must have been due to new mutations occurring during the course of the experiment. More generally, Barton and Keightley (2002) have emphasized that strong directional selection will fix most alleles of large effect in roughly 20 generations, and that a sustained response to selection must therefore be based either on segregating alleles with small effects or else on new mutations. In the experiment in Figure 8.4, studies of 440 single-nucleotide polymorphisms in the up-selected population indicated that the response involved at least 50 genes (Laurie et al. 2004).

In some experimental systems, the mutational effects of transposable elements are large enough to produce detectable genetic variation in quantitative traits. *Transposable elements*, discussed in Chapter 9, are DNA sequences that can insert themselves into the genome and move from one location to another. In *Drosophila*, insertions of a transposable element called the *P element* have been shown to produce significant genetic variation affecting

intermediary metabolism (Clark et al. 1995), number of sensory bristles (Lyman et al. 1996), and the ability to detect odors (Anholt et al. 1996). Some individual *P*-element insertions with significant effects on sensory bristles or olfaction map genetically to chromosomal regions containing genes known to affect these traits (Lyman et al. 1996; Anholt et al. 1996). Similarly, in laboratory mice, strains with multiple insertions exhibit an increased genetic variation for body weight relative to otherwise identical strains lacking the insertions (Keightley et al. 1993).

Although new mutations can contribute to the response to selection in some cases, the sluggishness of the response indicates that most of the response observed in genetically heterogeneous populations result from preexisting mutations that are already polymorphic in these populations. In practice, the magnitude of the response to selection might be used as a measure of the extent of genetic variation affecting the trait. The notion that selection response reflects genetic variation is formalized in the next section.

Prediction Equation for Individual Selection

When individuals are selected for breeding based solely on their own individual phenotypic values, the type of artificial selection is called **individual selection.** Figure 8.5 illustrates a variety of individual selection called **truncation selection.** The curve in panel A represents the normal distribution of a

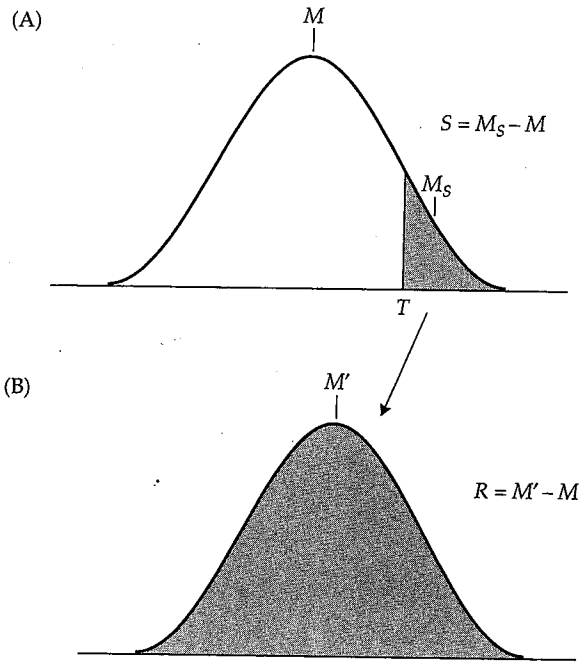


FIGURE 8.5 Diagram of truncation selection. (A) Distribution of phenotypes in the parental population with mean M. Individuals with phenotypes above the truncation point (T) are saved for breeding the next generation. The selected parents are denoted by the shading, and their mean phenotype is denoted M_S. (B) The mean of the distribution of phenotypes in the progeny is denoted M'. Note that M' is greater than M but less than M_S. The quantity S is called the selection differential, and R is called the response to selection.

quantitative trait in a population, and the shaded part of the distribution to the right of the phenotypic value denoted T indicates those individuals selected for breeding. The value T is called the **truncation point**. The mean phenotype in the entire population is denoted M, and that of the selected parents is denoted M_S. When the selected parents are mated at random, their offspring have the phenotypic distribution shown in panel B, where the mean phenotype is denoted M'.

An example of truncation selection for seed weight in the common bean, *Phaseolus vulgaris*, is shown in Figure 8.6. In this example, $T = 650$ mg, $M = 403.5$ mg, $M_S = 691.7$ mg, and $M' = 609.1$ mg. In this case—as is typical of truncation selection—the offspring mean M' is greater than the previous population mean M but less than the parental mean M_S. The reason M' is greater than M is that some of the selected parents have favorable genotypes and therefore pass favorable genes on to their offspring. At the same time, M' is generally less than M_S for two reasons:

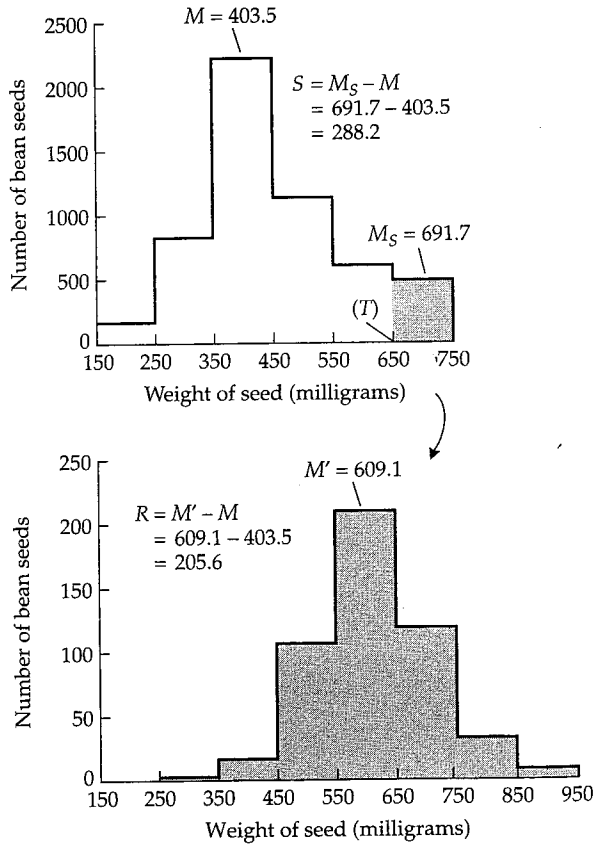


FIGURE 8.6 Truncation selection experiment for seed weight in edible beans of the genus *Phaseolus*, laid out as in Figure 8.5. The truncation point (T) is 650 mg. The selection differential S is the difference in mean between the selected parents and the whole population. The response R is the difference in mean between the progeny generation and the entire population in the previous generation. The quantity R/S is the realized heritability. (Data from Johannsen 1903.)

1. Because some of the selected parents do not have favorable genotypes; rather, their exceptional phenotypes result from chance exposure to exceptionally favorable environments.
2. Because alleles, not genotypes, are transmitted to the offspring, and exceptionally favorable genotypes are disrupted by Mendelian segregation and recombination.

The difference in mean phenotype between the selected parents and the entire parental population is the **selection differential** and is designated S. In symbols,

$$S = M_S - M \tag{8.8}$$

The difference in mean phenotype between the progeny generation and the previous generation is the **response to selection** and is designated R. Symbolically,

$$R = M' - M \tag{8.9}$$

In quantitative genetics, any equation that defines the relationship between the selection differential S and the response to selection R is known as a **prediction equation**. Since selection can be applied to a population in many different ways (for example, based on the mean of the family as a group, or based on the mean of the individual and its siblings), the prediction equation may differ corresponding to the different modes of selection. A general prediction equation that applies to many forms of selection, including truncation selection (the type of selection illustrated in Figure 8.5), is

$$R = h^2 S \tag{8.10}$$

where h^2 is the **realized heritability**. (For historical reasons, the square in h^2 is an integral part of the symbol for heritability, and so the heritability equals h^2, not h.) Later in this chapter, we will show that the realized heritability is identical to the narrow-sense heritability defined by regression, provided the phenotypes and the magnitudes of genetic effects follow a normal, bell-shaped (Gaussian) distribution. These assumptions are necessary in order to apply regression to the problem. This equivalence emphasizes that heritability can be understood at several different levels. Equation 8.10 implies that the realized heritability of a trait can be interpreted as a mere description of what happens when artificial selection is practiced. In Figure 8.6, for example, $S = 288.2$ and $R = 205.6$, so $h^2 = R/S = 205.6/288.2 = 71.3\%$. When estimated like this from empirical data, h^2 is the realized heritability, and it simply summarizes the observed result.

PROBLEM 8.2 Below are data on the number i of abdominal bristles in samples from two consecutive generations G_1 and G_2 of an experiment in directional selection for increased bristle number. In the G_1 generation, individuals with 22 or more bristles (enclosed in brackets) were mated together at random to form the G_2 generation. Estimate the realized heritability of the number of abdominal bristle in this experiment. (Data kindly provided by Trudy Mackay. In order to make the sexes comparable, the value of 2 has been added to the bristle number in males.)

i	G_1	G_2	i	G_1	G_2	i	G_1	G_2
15	0	2	20	20	13	25	[1]	3
16	21	4	21	12	14	26	0	2
17	5	7	22	[13]	12	27	0	0
18	18	16	23	[3]	6	28	0	2
19	17	17	24	[5]	3			

ANSWER The means are $M = 2220/15 = 19.3$, $M_S = 22.7$, and $M' = 2035/11 = 20.1$. The selection differential $S = 22.7 - 19.3 = 3.4$ (see Equation 8.8) and the response $R = 20.1 - 19.3 = 0.8$ (see Equation 8.9). The realized heritability estimated from Equation 8.10 is $h^2 = 0.8/3.4 = 0.235$.

Limits to Selection

Unless the number of QTLs is extremely large, or the mutation rate of QTLs extremely high, progress under artificial selection cannot continue forever. Any population must eventually reach a **selection limit**, or **plateau**, after which it no longer responds to selection. One of the reasons that a population eventually reaches a plateau is exhaustion of genetic variance, when all alleles affecting the selected trait have become fixed, lost, or are otherwise unavailable for selection. With no genetic variance, no progress under individual selection can be achieved. However, many experimental populations that have reached a selection limit readily respond to **reverse selection** (selection in the reverse direction of that originally applied), so genetic variance affecting the trait is still present. Indeed, in such populations, the phenotype may change in the direction of its original value if continuing artificial selection is simply suspended (**relaxed selection**).

One frequent reason for the occurrence of selection limits in populations with considerable genetic variation is that artificial selection is opposed by natural selection. In mice, for example, response to selection for small body size ultimately ceases because small animals are less fertile than larger ones, and the smallest animals are sterile (Falconer and Mackay 1996). Selection for small body size gradually becomes less effective due to the opposing effects

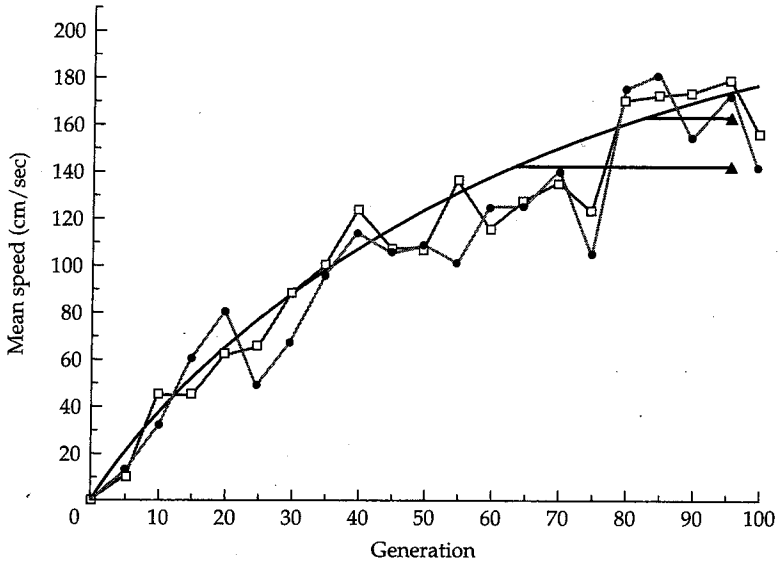


FIGURE 8.7 Response of *Drosophila melanogaster* to selection for flight speed in a wind tunnel. The horizontal lines terminating in triangles indicate subpopulations in which selection was relaxed in generations 65 or 85. In neither case was there a reduction in flight performance with relaxed selection. (After Weber 1996.)

of natural selection until, eventually, no further progress is possible. When selection is relaxed, the natural selection is unopposed and results in a retrogression in the artificially selected trait. Some backward slippage with relaxed selection also results from diminution in the linkage disequilibrium that usually builds up during the course of long-term artificial selection. If natural selection opposes the artificial selection, then when artificial selection is relaxed, natural selection results in at least a partial return to the initial phenotypic mean. On the other hand, relaxed selection does not always result in backward slippage of the trait, as demonstrated by the example of selection for flight speed in *Drosophila* in Figure 8.7. The horizontal lines starting at generations 65 and 85 that terminate in triangles show the mean flight speed in subpopulations in which selection had been relaxed for 30 and 10 generations, respectively. There is no decrease in the mean flight speed from that in the selected population at the time that each subline was initiated, indicating either that most of the alleles affecting flight speed had been fixed, or that the increase in flight speed had a negligible effect on fitness.

In most genetically heterogeneous populations, artificial selection can change the phenotype well beyond the range of variation found in the original population. Pertinent data for populations of mice are presented in Table

TABLE 8.1 Selection Limits and Duration of Response for Various Traits in Laboratory Mice

Character selected	Direction of selection	Total response[a]	Half-life of response[b]
Weight (in strain N)	Up	3.4 σ_p	0.6N
	Down	5.6 σ_p	0.6N
Weight (in strain Q)	Up	3.9 σ_p	0.2N
	Down	3.6 σ_p	0.4N
Growth rate	Up	2.0 σ_p	0.3N
	Down	4.5 σ_p	0.5N
Litter size	Up	1.2 σ_p	0.5N
	Down	0.5 σ_p	0.5N

Source: From Falconer 1977.

[a] Total response is expressed as a multiple of the initial phenotypic standard deviation, σ_p.

[b] Half-life of response is the number of generations taken to progress halfway to the selection limit; here the half-life is expressed in multiples of effective population number (N).

8.1. As can be seen, a total selection response of three to five times the original phenotypic standard deviation is not unusual, and for selection to change a population of effective size N halfway to its selection limit typically requires about $1/(2N)$ generations.

In some cases the total response to artificial selection is very large. For example, in a long-term selection experiment for pupal weight in the flour beetle *Tribolium castaneum*, in which the base population consisted of the progeny of a cross between two inbred lines, 100 generations of selection resulted in a population in which the mean pupal weight in the selected population was 17 standard deviation units greater than the mean in the base population (Enfield 1980). The ability to select a population in which virtually every phenotype is greater than the maximum in the original population strikes many students as paradoxical. It does seem plausible to argue that, if all of the alleles eventually selected are already present in the original population, then all possible favorable genotypes should be present also, though perhaps at low frequency. The fallacy in the argument is that real populations subjected to artificial selection are relatively small in size, consisting of at most a few hundred organisms. Therefore, if the favorable alleles are initially rare, then the frequency of the favored genotypes may be so small that the expected number of such genotypes will be much smaller than one, and so the superior genotypes, while theoretically possible, do not actually exist in the original population.

Some traits consistently fail to respond to artificial selection, suggesting a lack of suitable genetic variation. Bilateral symmetry is an example of a trait

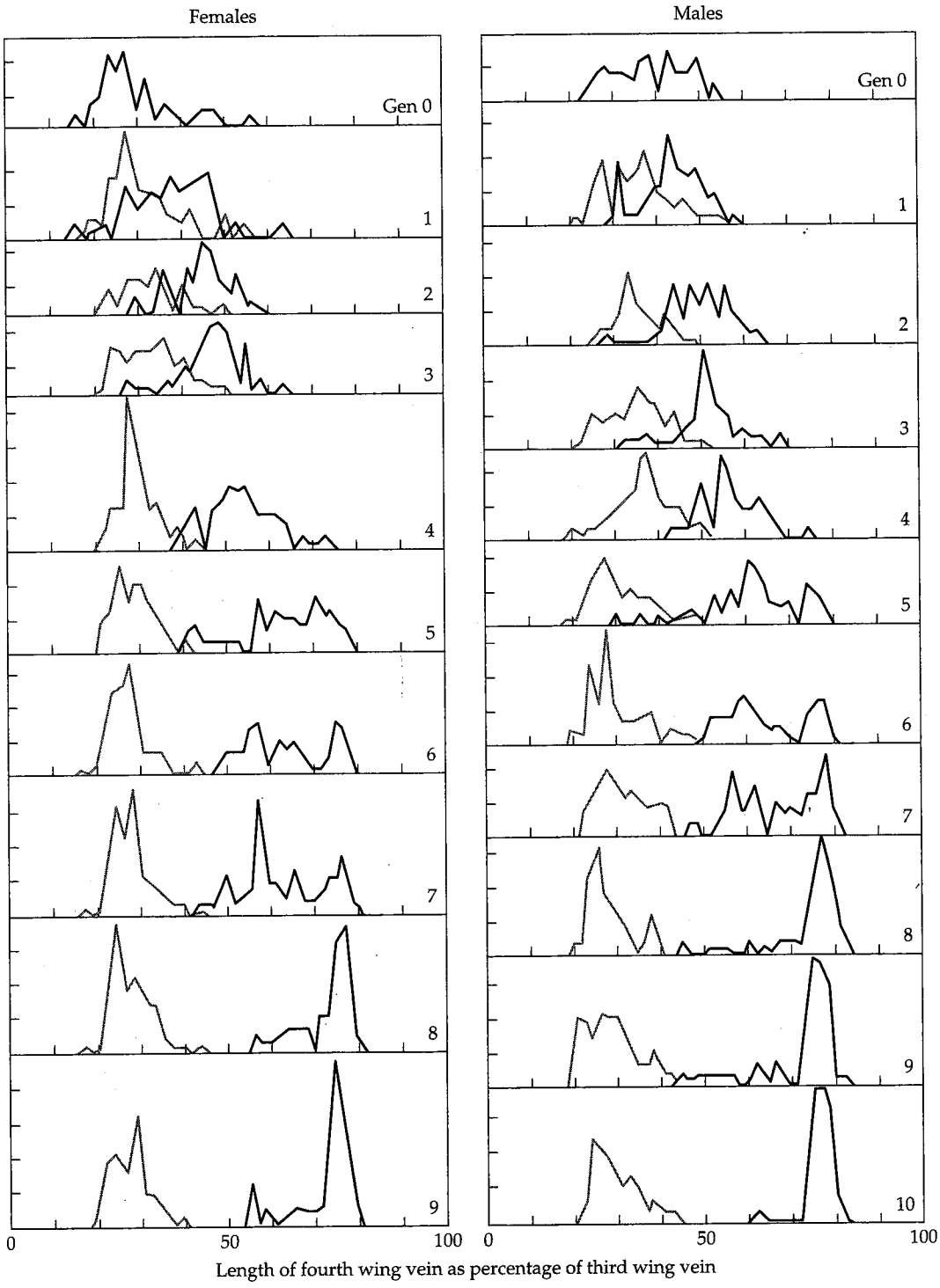
that has not been amenable to change by artificial selection. The failure of Maynard-Smith and Sondhi (1961) to create bilateral asymmetry in *Drosophila* by selecting for an excess of dorsal bristles on the left side is typical. The apparent lack of genetic variation determining bilateral asymmetry is of interest in regard to embryonic development, for it implies that the genetic control of development of symmetrical structures specifies patterns that are common to the left and the right sides of the body. That is, rather than left-bristle genes and right-bristle genes, there appear to be generic bristle genes whose spatial expression is determined symmetrically. Of course asymmetrical structures do exist (such as the vertebrate heart) and recently inroads have been made in understanding the molecular genetic basis for this asymmetry (Casey 2001).

Not all traits with heritable variation obey the prediction equation and show a simple linear change in the mean. Sometimes a trait responds to directional selection for a few generations, then ceases to respond, but later responds again as selection is continued. One possible mechanism for this stop-and-start response is that the population at a plateau is in linkage disequilibrium, and that it takes time for recombination to break up the allelic associations and release the latent genetic variation. This phenomenon was observed in a long-term study of the quantitative genetics of wing veins in *Drosophila* (Scharloo 1987). In this case a bimodal phenotypic distribution was also generated during selection (Figure 8.8), which was proposed to reflect a nonlinear mapping from genetic and environmental factors to the determination of phenotype.

As we have seen, heritability can be interpreted in purely statistical terms with no genetic content. However, if we postulate that there are Mendelian genes underlying the phenotypes, then the genetic underpinning allows us to do more than merely describe statistical relations among individuals. By bringing Mendelian genetics into the picture, we will see why the response to any kind of artificial selection is determined by the magnitude of the heritability. In particular, the genetic basis of response to artificial selection comes from changes in gene frequencies and sometimes also to changes in linkage disequilibrium.

8.4 GENETIC MODELS FOR QUANTITATIVE TRAITS

When h^2 is interpreted as realized heritability, then Equation 8.10 is hardly a "prediction equation" inasmuch as it merely describes what has already happened in one generation of selection. Of course, the equation could be used to predict the result of the next generation of selection, but artificial selection is impossible in many natural populations and is time-consuming and expensive in many domesticated plants and animals. It would therefore be useful if one could estimate heritability without actually performing any artificial selection. If the heritability h^2 could be estimated in such a manner, then


Females
Males
Gen 0
1
2
3
4
5
6
7
8
9
10
0
50
100
Length of fourth wing vein as percentage of third wing vein

◀ **FIGURE 8.8** Frequency distributions in females (left) and males (right) of a line of *Drosophila melanogaster* selected for length of the fourth wing vein. The gray lines represent selection for a short vein, and black lines represent selection for a long vein. In the line selected for long veins, both sexes displayed a bimodal frequency distribution when the relative vein length was approximately 60–80%. (From Scharloo 1987.)

Equation 8.10 would be a true prediction equation in the sense that the response R could be predicted for any selection differential S, based on the estimated value of h^2. Such an estimate of h^2 is indeed possible, but it involves an understanding of heritability at a level that includes the underlying genetic basis of quantitative traits.

An understanding of the genetics behind Equation 8.10 requires three items: (1) a concept of how alternative alleles of a gene affect a quantitative trait; (2) a determination of how selection changes the allele frequencies; and (3) a calculation of how much the mean of the trait increases as a result of the change in allele frequency. Some detail is required to establish these three items, but the detail is necessary in order to understand the genetic meaning of heritability.

Nilsson-Ehle (1909) was the first to show that a trait with a nearly continuous distribution of phenotypes could result from the joint effects of several genes. The trait of interest was the intensity of red pigment in the glume of wheat *Triticum vulgare*, which Nilsson-Ehle found to result from three unlinked genes, each with two alleles. The situation is exceptionally simple for a quantitative trait; the environment has a negligible effect on phenotype, because the alleles of each gene are **additive** (i.e., heterozygous genotypes have a phenotype that is exactly intermediate between the phenotypes of the homozygous genotypes), and because the genetic effects are also additive across genes (i.e., the total genetic effect of any three-locus genotype is just the sum of the separate effects of each gene). To simplify matters, consider just two of the genes, and let their alleles be denoted A, a and B, b. Figure 8.9 shows the nine possible two-locus genotypes, their frequencies with random mating when the allele frequencies of A and B are both $\frac{1}{2}$, and two examples differing in the level of dominance. In both cases the phenotype is determined entirely by genes, but in the case of additive gene action (panel A), $h^2 = 1.0$, whereas when A and B are both dominant with additivity across loci, $h^2 = 0.81$.

To understand why dominance affects the heritability, consider first the situation in Figure 8.9A. With additivity within and across genes, we may assume that the genotype *aa bb* has a color score of 0 (white) and that each A or B allele in the genotype contributes one unit of red pigment. The maximum pigment score is therefore 4. In this population, the mean color phenotype is 2, and when the allele frequencies of A and B are both p, then the mean of a population with random mating can be shown to equal $4p$. To con-

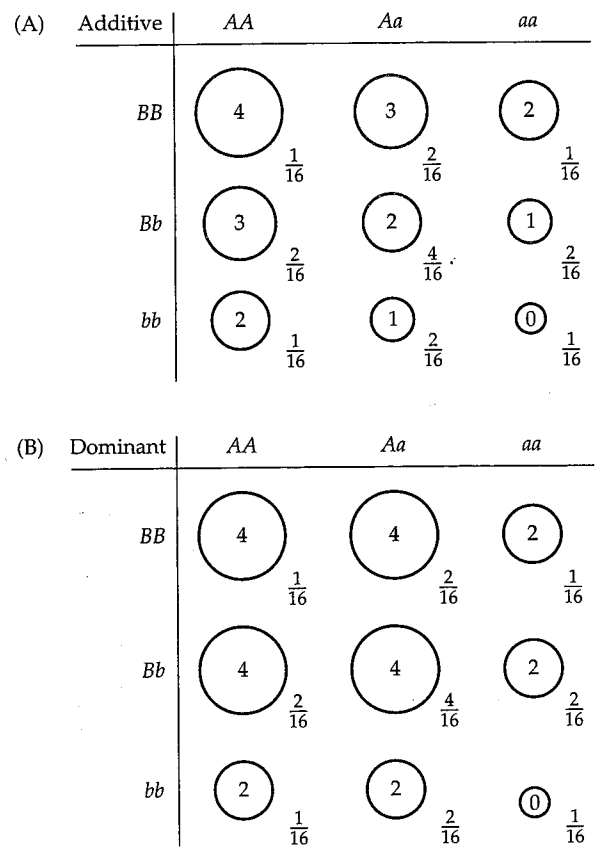


FIGURE 8.9 Frequencies of two-locus genotypes (positioned outside of the circles) and respective phenotypes (positioned within the circles) in a population with allele frequency $\frac{1}{2}$ for each locus. Panel A illustrates the case of additivity of effects at each locus and across loci. In panel B, the alleles A and B are each dominant to alleles a and b, respectively, but the effects of the two loci are additive.

nect this trait with the prediction Equation 8.10, suppose that the two lowest phenotypic classes (i.e., 0 and 1) are selected as parents of the next generation. The mean color phenotype among the selected individuals is therefore $M_S = [(\frac{1}{16})(0) + (\frac{2}{16})(1) + (\frac{2}{16})(1)]/(\frac{5}{16}) = \frac{4}{5}$. Among the selected parents, the mean allele frequency p of A and B is given by $p = [(\frac{1}{16})(0) + (\frac{2}{16})(0) + (\frac{2}{16})(\frac{1}{2})]/(\frac{5}{16}) = \frac{1}{5}$. With random mating among the selected individuals, the mean color phenotype M' among the progeny is $4p$, as noted earlier, which in this case is $M' = \frac{4}{5}$. Since the mean M of the original population is $M = 2$, the selection differential $S = M_S - M = \frac{4}{5} - 2 = -\frac{6}{5}$, and the selection response $R = M' - M = \frac{4}{5} - 2 = -\frac{6}{5}$ also, and hence the narrow-sense heritability is given by Equation 8.10 as $h^2 = R/S = 1.0$. As demonstrated in the next paragraph, this high heritability is due to the additivity within and across genes and not merely to the fact that environmental effects are negligible.

Figure 8.9B refers to a hypothetical situation in which the A and B alleles are dominant but still additive across genes. To keep the same minimum and

maximum phenotypes as in panel A, we assume that each locus with at least one dominant allele adds 2 units to the phenotype, and so the maximum color phenotype is 4. In this case, it can be shown that the mean of a random-mating population with allele frequencies of A and B both equal to p is given by $4p(1 + q)$, where again $q = 1 - p$. If the two lowest phenotypic classes (in this case, 0 and 2) are selected as parents of the next generation, then $M_S = \frac{12}{7}$. The allele frequency of A and B among the selected parents can be calculated using the same logic as used in the additive case, with the result that $p = \frac{2}{7}$, and hence the mean phenotype among the progeny M' is $\frac{96}{49}$. These results imply that $S = -\frac{9}{7}$, $R = -\frac{51}{49}$, and hence that $h^2 = R/S = \frac{17}{21} = 0.81$. Although environmental effects on seed color are still negligible in the example with dominance, the heritability is reduced to a value smaller than 1.0. This perhaps surprising result occurs because certain genetic effects (such as those resulting from dominance or, in other examples, nonadditivity across genes) are not useful in changing a population by means of the type of individual selection discussed here.

To see how an underlying genetic model can be formulated for continuous characters, refer to Figure 8.10, which shows the normal distribution of a trait in a hypothetical random mating population. In truncation selection, all individuals with phenotypes above the truncation point T are saved for breeding, and the shaded area B of the distribution represents the proportion of the population selected. (The total area under any normal density equals 1.) The height of the normal density at the point T is denoted Z, and, as

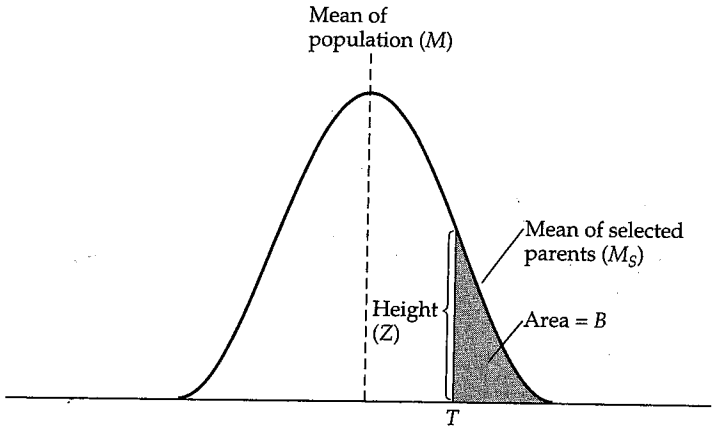


FIGURE 8.10 Normal distribution of a quantitative trait in a hypothetical population, showing some important symbols used in quantitative genetics. Here M is the mean of the population, T the truncation point, Z the height (ordinate) of the normal density at the point T, B the shaded area under the normal curve to the right of T, and M_S is the mean among selected parents.

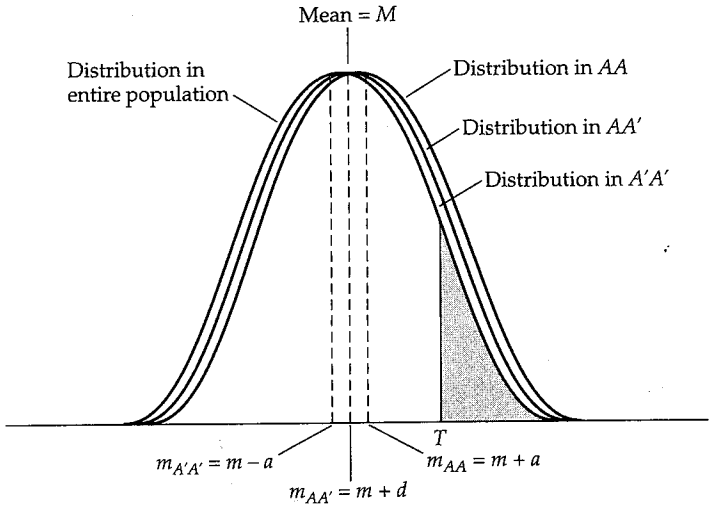


FIGURE 8.11 Same distribution as in Figure 8.10, showing the slightly different distribution of phenotypic value among the three genotypes (AA, AA', and $A'A'$) for a gene with two alleles that contributes to the quantitative trait. The means of the distributions of $A'A'$, AA', and AA are symbolized $m - a$, $m + d$, and $m + a$, respectively.

before, the mean phenotype among the selected individuals is called M_S. One of the special properties of the normal distribution to be used below is that

$$(M_S - M)/\sigma^2 = Z/B \tag{8.11}$$

To determine the amount of increase in mean phenotype in a population resulting from one generation of truncation selection, we first imagine a gene that affects the trait in question and that has alleles A and A' with respective allele frequencies p and q. (Here we use A and A' as symbols for the alleles because in quantitative genetics, as we shall see shortly, the symbol a is traditionally reserved for another purpose.) Because of random mating, genotypes AA, AA', and $A'A'$ are present in the population with frequencies p^2, $2pq$, and q^2, respectively, but the individual genotypes cannot be identified through their phenotypic values because of the variation in phenotype caused by environmental factors and genetic differences in other genes. If the genotypes could be identified, their individual distributions of phenotypic value might appear as shown in Figure 8.11. Each distribution is normal and has the same variance, but the means are very slightly different. The mean phenotypes of AA, AA', and $A'A'$ are denoted as m_{AA}, $m_{AA'}$, and $m_{A'A'}$, respectively, and we could do all of the subsequent calculations using these quanti-

ties. But there is a more natural way to write the phenotypes by letting $m_{AA} = m + a$, $m_{AA'} = m + d$, and $m_{A'A'} = m - a$, where

$$m = \frac{m_{AA} + m_{A'A'}}{2} \tag{8.12a}$$

$$a = \frac{m_{AA} - m_{A'A'}}{2} \tag{8.12b}$$

$$d = m_{AA'} - \left(\frac{m_{AA} + m_{A'A'}}{2}\right) \tag{8.12c}$$

The phenotypes are written in this seemingly arbitrary way for two reasons. First, a and d have a natural biological interpretation in terms of the regression coefficient of phenotype on genotype (see below); and second, use of a and d in this manner greatly simplifies some of the subsequent equations. In any case, the symbols a and d serve as convenient representations of the effects of the alleles A and A' on the quantitative trait. The difference between the means of the homozygous genotypes is $(m + a) - (m - a) = 2a$, and d/a serves as a measure of dominance. The relationship $d = a$ means that A is dominant, $d = 0$ implies additivity (the phenotype of the heterozygous genotype is the exact average of those of the homozygous genotypes), and $d = -a$ means that A' is dominant.

Calculation of a and d for an actual example involving two alleles that affect coat coloration in guinea pigs is illustrated in Table 8.2. In this case, $a = 0.127$, $d = -0.016$ (the negative sign on d means that the c^d allele is partially dominant), and $d/a = -0.126$.

A biologically meaningful interpretation of a and d is shown for the coat-color example in Figure 8.12, which shows a plot of the mean phenotype for each genotype in Table 8.2 as a function of the number of A alleles in the

TABLE 8.2 Calculation of m^*, a, and d for Alleles at a Locus Affecting Coat Coloration in Guinea Pigs[a]

Genotype	Amount of black coloration[b]	Values of a and d
$c^r c^r$ (AA)	$m_{AA} = 1.202$	$a = 1.202 - 1.075 = 0.127$
$c^r c^d$ (AA')	$m_{AA'} = 1.059$	$d = 1.059 - 1.075 = -0.016$
$c^d c^d$ ($A'A'$)	$m_{A'A'} = 0.948$	$-a = 0.958 - 1.075 = -0.127$
	$m = (1.202 + 0.948)/2 = 1.075$	

Source: Data from Wright 1968.

[a] The calculations to be carried out first are those beneath the data; then the right-hand column is completed.

[b] Here the amount of black coloration is measured as arcsin $(\sqrt{x})$, where x is the proportion of black coloration on the animal. For $c^r c^r$, $c^r c^d$, and $c^d c^d$ genotypes, the corresponding x values are 0.87, .0.76, and 0.66, respectively.

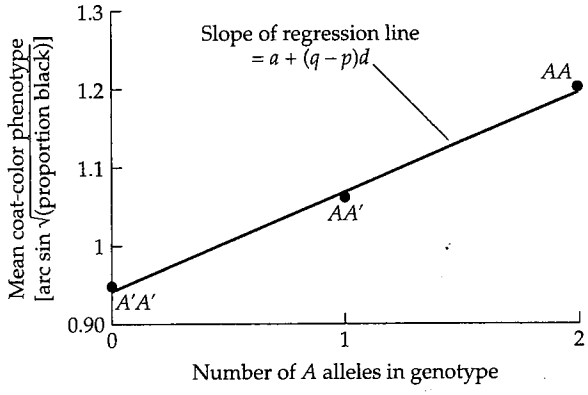


FIGURE 8.12 Regression of mean coat-color phenotype on genotype. The genotypic values of $A'A'$, AA', and AA have arbitrarily been coded as 0, 1, and 2, respectively. The proportion of black fur on the animals has been transformed using the formula shown to make the distribution more nearly normal. The regression coefficient in a population with random mating is $a + (q - p)d$, where a is the main (additive) effect of the allele A and d is a measure of dominance.

genotype (0 for $A'A'$, 1 for AA', and 2 for AA). Using the principles of regression in Equations 8.1, 8.2, and 8.6, it is straightforward to show that the regression of mean phenotype on the number of A alleles in the genotype is given by $a + (q - p)d$. Note first that, with Hardy-Weinberg genotype frequencies, the mean phenotype in the entire population is given by

$$\begin{aligned} M &= p^2(m + a) + 2pq(m + d) + q^2(m - a) \\ &= m + (p - q)a + 2pqd \end{aligned} \tag{8.13}$$

Also, the mean number of A alleles per genotype in the population is given by $2p^2 + 2pq = 2p(p + q) = 2p$, and the mean of the product of the phenotype and the number of A alleles equals $2p^2(m + a) + 2pq(m + d) = 2pm + 2p^2a + 2pqd$. From Equation 8.2, therefore, the covariance between phenotype and number of A alleles in Figure 8.12 is

$$\begin{aligned} &Cov(\text{phenotype, number of } A \text{ alleles}) \\ &= 2pm + 2p^2a + 2pqd - [m + (p - q)a + 2pqd](2p) \\ &= 2pqa + 2pq(q - p)d \end{aligned} \tag{8.14}$$

Similarly, Equation 8.1 implies that the variance in the number of A alleles among the genotypes is

$$Var(\text{number of } A \text{ alleles}) = 4p^2 + 2pq - (2p)^2 = 2pq \tag{8.15}$$

Hence Equation 8.6 gives the regression coefficient of the phenotype on the number of A alleles in the genotype as the ratio of Equation 8.14 to Equation 8.15, or

$$b = \frac{2pqa + 2pq(q - p)d}{2pq} = a + (q - p)d \tag{8.16}$$

In the following sections we shall see that the expression on the right-hand side of Equation 8.16 plays a prominent role in defining the heritability in terms of allele affects and allele frequencies.

Note that the coat-color phenotypes in Table 8.2 and Figure 8.12 are not given in terms of the proportion of black coloration, which seems the natural scale of measurement. The reason is that the theory assumes that the phenotypes are normally distributed, and when the data are in terms of proportions, the transformation arc sin[√(proportion black)] makes the distribution more normal. Such a transformation of scale is called a **normalizing transformation**. Another example is given in the following problem.

PROBLEM 8.3 Crosses between the inbred Danmark (P_1) and inbred Red Currant (P_2) tomato gave the following mean fruit weights and their log transforms. P_1 and P_2 are the parental means, F_1 and F_2 are the first and second hybrid generation, and B_1 and B_2 are the progeny of the backcross of $F_1 \times P_1$ and $F_1 \times P_2$, respectively.

	Expected mean	Mean fruit weight	Log (weight)
P_1	$m + a$	10.36 ± 0.581	0.98 ± 0.03
P_2	$m - a$	0.45 ± 0.017	−0.36 ± 0.02
F_1	$m + d$	2.33 ± 0.130	0.33 ± 0.03
F_2	$m + (\frac{1}{2})d$	2.12 ± 0.105	0.27 ± 0.01
B_1	$m + (\frac{1}{2})(a + d)$	4.82 ± 0.253	0.64 ± 0.02
B_2	$m + (\frac{1}{2})(d - a)$	0.97 ± 0.045	−0.05 ± 0.01

Use this information to calculate m, a, and d for both the fruit weights and the log transformed weights. Do the simple weights or the log transformed weights fit the model better? (Data from Powers 1951.)

ANSWER The difference between the two parental means is $2a$, and so $a = (10.36 - 0.45)/2 = 4.96$. This gives $m = 5.4$. The F_1 has a mean $m + d = 2.33$, and so $d = 2.33 - 5.4 = -3.07$. The F_2 should have mean $(1/4)(m + a) + (\frac{1}{2})(m + d) + (\frac{1}{4})(m - a) = m + (\frac{1}{2})d = 5.4 + (\frac{1}{2})(-3.07) = 3.86$. The B_1 refers to backcrosses of the F_1 to P_1, which should yield one-half of genotypes like P_1 and one-half of genotypes like the F_1, so the mean should be $(\frac{1}{2})(m + a) + (\frac{1}{2})(m + d) = m + (\frac{1}{2})(a + d) = 6.34$. Similar reasoning gives an expected mean for B_2 of 1.38. The estimates for the means of the F_2, B_1, and B_2 do not fit very well at all. Trying again with the log transformed data, we get $a = 0.67$, $m = 0.31$, and $d = 0.02$. The expected means for the F_2, B_1, and B_2 are then $0.31 + (\frac{1}{2})(0.02) = 0.32$, $0.31 + (\frac{1}{2})(0.67 + 0.02) = 0.65$, and $0.31 + (\frac{1}{2})(0.02 - 0.67) = -0.01$. The log transformed data clearly fit much better, suggesting that the better scale to use for the quantitative genetic models is the log transformed scale. In actual practice, the entire set of data is used to estimate m, a, and d by a least squares method analogous to that in Equation 8.4, and then goodness of fit of the model to the data is tested by a chi-square test.

Effects of the scale of measurement are known as **scaling effects**. Since estimates of the additive (a) and dominance (d) values of alleles depend on scaling, so does the heritability. An important point is that the equivalence

between the heritability defined by parent-offspring regression and by realized heritability depends on the correct choice of scaling. Only one scaling provides a normal Gaussian distribution of phenotypes and of genetic effects, and this is the appropriate scaling that yields the prediction Equation 8.10.

Change in Allele Frequency

Suppose for the moment that we were practicing artificial selection for increased amount of black coat coloration in the guinea pigs in Table 8.2. Selection for black-coat coloration in a population containing both the c^r (i.e., A) and c^d (i.e., A') alleles would be successful in increasing the allele frequency of A, and the average amount of black coloration among individuals of the next generation would increase. Therefore, in order to calculate the expected increase in black coloration in one generation of selection, we must first calculate the corresponding change in the allele frequency of A. An equation for change in allele frequency with natural selection was derived in Chapter 5, which remains valid for artificial selection if we agree to interpret the "fitness" of an individual as the probability that the individual is included among the group selected as parents of the next generation. With this interpretation of fitness, differences in fitness (i.e., reproductive success) of AA, AA', and $A'A'$ genotypes correspond to the differences in area to the right of the truncation point in Figure 8.11, because only those individuals in the shaded area are allowed to reproduce. The differences in area are easy to calculate if you shift or slide each curve horizontally until its mean coincides with m. The $A'A'$ curve must slide a units to the right, and the AA' and AA curves must slide d and a units to the left, respectively. This shifting brings the distributions into coincidence, but it slides the truncation points slightly out of register, as shown in Figure 8.13. The difference in "fitness" between AA and AA', denoted $w_{11} - w_{12}$ (as in Chapter 5), is equal to the small area indicated in Figure 8.13, as is the difference in fitness between AA' and $A'A'$, denoted $w_{12} - w_{22}$. The areas corresponding to $w_{11} - w_{12}$ and $w_{12} - w_{22}$ are approximately rectangles, and the area of a rectangle is the product of the base and the height. The approximation is most accurate when the effect of this one locus on the phenotype is small. Therefore, since Z represents the height of the normal distribution at the point T, we can make the following approximations

$$w_{11} - w_{12} \approx Z[(T - d) - (T - a)] = Z(a - d) \qquad (8.17a)$$

$$w_{12} - w_{22} \approx Z[(T + a) - (T - d)] = Z(a + d) \qquad (8.17b)$$

The average fitness $\bar{w}$ of the entire population simply equals B, because B is the proportion of the population saved for breeding. From Chapter 5 we know that

$$\Delta p = pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})] / \bar{w}$$

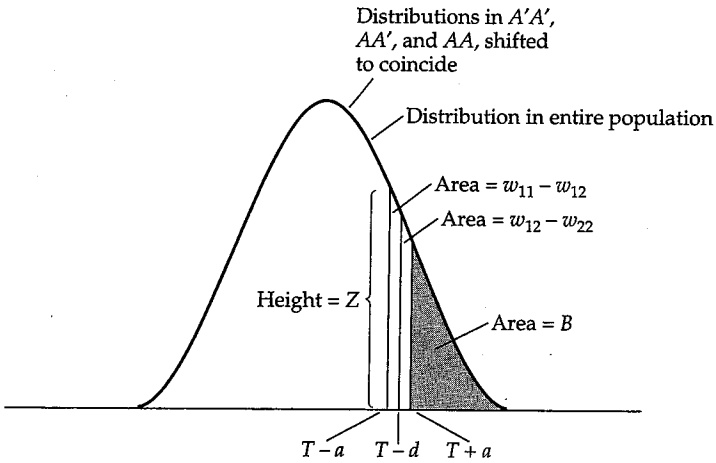


FIGURE 8.13 Same distribution as in Figures 8.10 and 8.11, but with the distributions of AA, AA', and $A'A'$ shifted laterally to coincide. Shifting the distributions slides the truncation points slightly out of register, so the truncation points for AA, AA', and $A'A'$ become $T - a$, $T - d$, and $T + a$, respectively. The small area that is denoted $w_{11} - w_{12}$ is the difference between the proportions of AA and AA' genotypes that are included among the selected parents, and the area $w_{12} - w_{22}$ is the difference in the proportion of AA' and $A'A'$ genotypes included among the selected parents.

where Δp is the expected change in frequency of the allele A in one generation of selection. Substituting from Equation 8.17 and using $\bar{w} = B$ leads to

$$\Delta p = pq[pZ(a - d) + qZ(a + d)]/B \tag{8.18}$$

or, since $p + q = 1$,

$$\Delta p = (Z/B)pq[a + (q - p)d] \tag{8.19}$$

The biological meaning of Equation 8.19 is clear from Figure 8.12. The expected change in the frequency of a favored allele under selection is a function of the intensity of selection (measured by Z/B), the allele frequencies, and the regression coefficient of phenotype on genotype (see Equation 8.16). An equation corresponding to 8.19 could be obtained for any gene affecting the trait, but the values of p, a, and d would differ for each gene. A generalization that accounts for nonrandom mating is found in Falconer (1985).

Change in Mean Phenotype

Equation 8.19 provides an expression for Δp that can be used to calculate the mean phenotypic value of coat color after one generation of selection. In the next generation, the allele frequencies of A and A' are $p + \Delta p$ and $q - \Delta p$,

respectively. With random mating, the mean phenotype in this generation is given by Equation 8.13 as

$$M' = (p + \Delta p)^2(m + a) + 2(p + \Delta p)(q - \Delta p)(m + d) + (q - \Delta p)^2(m - a) \qquad (8.20)$$

When the right-hand side of Equation 8.20 is multiplied out and terms in $(\Delta p)^2$ are ignored because Δp is usually small, then M' is found to be approximately

$$\begin{aligned} M' &\approx p^2(m + a) + 2pq(m + d) + q^2(m - a) + 2[a + (q - p)d]\Delta p \\ &= M + 2[a + (q - p)d]\Delta p \end{aligned} \qquad (8.21)$$

where the substitution of M comes from Equation 8.13. The approximation in Equation 8.21 is rather good even for relatively large values of Δp.

Equation 8.21 warrants a little more development since it yields the prediction equation $R = h^2S$ (Equation 8.10) and also provides an expression for h^2 in terms of the parameters a, d, and p that can be interpreted genetically. First, rewrite Equation 8.21 as

$$M' - M = 2[a + (q - p)d]\Delta p \qquad (8.22)$$

Then substitute for Δp from Equation 8.19, which yields

$$M' - M = (Z/B)2pq[a + (q - p)d]^2 \qquad (8.23)$$

Now use the expression for Z/B given in Equation 8.11 to obtain

$$M' - M = (M_S - M)2pq[a + (q - p)d]^2/\sigma^2 \qquad (8.24)$$

where σ^2 equals the variance in phenotype in the population. Finally, substitute from Equations 8.8 and 8.9 for the selection differential S and the response R, yielding

$$R = (S)2pq[a + (q - p)d]^2/\sigma^2 \qquad (8.25)$$

However, $R = h^2S$ also (Equation 8.10), and so

$$h^2 = 2pq[a + (q - p)d]^2/\sigma^2 \qquad (8.26)$$

Equation 8.26 for h^2 is the one we were after, as it defines the heritability in terms of p, q, a, and d, each of which has a genetic meaning. Furthermore, the right-hand side of Equation 8.26 has a wonderfully simple biological interpretation. Although we will not get into the details, it can be shown that the right-hand side of Equation 8.26 is the proportion of the total variance in phenotype that can be explained by the regression of the phenotype on the genotype of the alleles under consideration, where the regression is as illustrated in Figure 8.12.

Equation 8.26 can be generalized to include any number of genes affecting a quantitative trait, in which case the right-hand side of the equation must be replaced by a summation of similar terms, one for each gene. That is, for many genes, $R = h^2S$ where

$$h^2 = \frac{1}{\sigma^2}\sum_i 2p_iq_i[a_i + (q_i - p_i)d_i]^2 \tag{8.27}$$

where the summation is over all genes that affect the trait, and where p_i, q_i, a_i, and d_i are the values of p, q, a, and d for the ith gene. As will be discussed in more detail later, the summation in Equation 8.27 is often symbolized as σ_a^2 and called the **additive genetic variance** of the trait:

$$\sigma_a^2 = \sum_i 2p_iq_i[a_i + (q_i - p_i)d_i]^2 \tag{8.28}$$

Although the individual components in the additive genetic variance are difficult to identify except in special cases like the one involving coat color in guinea pigs, the collective effects (represented by the summation) can be estimated from the correlation between relatives, as we shall see later.

Linearity of Response

Equation 8.27 implies that heritability values are determined in part by allele frequencies. Because the allele frequencies change during the course of selection, the heritability is also expected to change. In practice, however, the heritability changes sufficiently slowly that over the course of a few generations, it can be regarded as approximately constant. The approximate constancy of heritability has a twofold cause: (1) if a particular gene accounts for only a small proportion of the total phenotypic variance in a quantitative trait, then the gene frequency does not change very rapidly, and (2) the values of a and d remain nearly constant provided that the environment does not change drastically from one generation to the next. Thus, at least for the first 10 generations or so, heritability usually remains approximately constant and can be used as a constant in the prediction equation (see Equation 8.10). To be precise, suppose h^2 is constant and let M_t and S_t represent the mean of the population and the selection differential in the tth generation. Then, over the length of time during which h^2 is approximately constant,

$$M_t - M_0 = h^2(S_0 + S_1 + \cdots + S_{t-1}) \tag{8.29}$$

The quantity $M_t - M_0$ is the *total response* to selection, and $S_0 + S_1 + \cdots + S_{t-1}$ is called the **cumulative selection differential**. During the time in which h^2 is approximately constant, therefore, a plot of M_t against cumulative selection differential is expected to yield a straight line with slope equal to h^2, as illustrated for the case in mice in Figure 8.14.

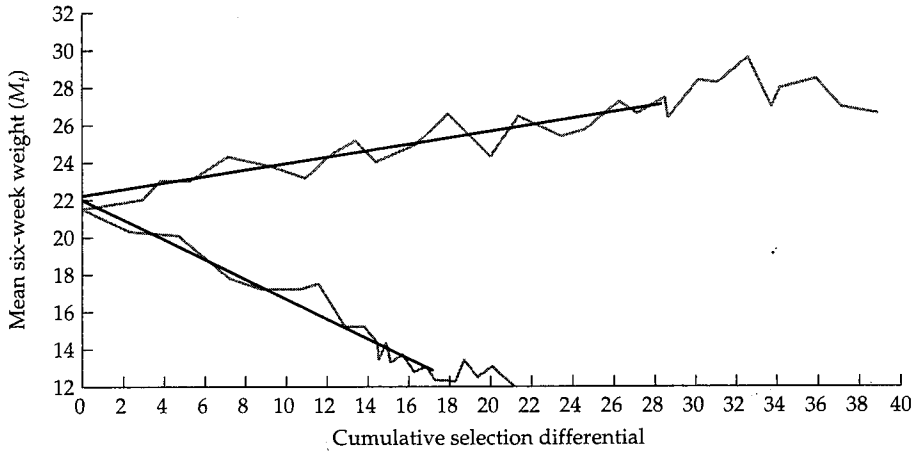


FIGURE 8.14 Linearity in response against the cumulative selection differential for body weight in mice at age six weeks. Linearity in the up (high-weight) direction continues for about twice as long as it does in the down (low-weight) direction. (After Falconer 1955.)

8.5 COMPONENTS OF PHENOTYPIC VARIANCE

As Equation 8.28 suggests, the variance of a quantitative trait can be split into various components representing different causes of variation. Resemblance between relatives is conveniently expressed in terms of the variance components, but variance partitioning is also of interest in its own right. Since the rate of change of a trait under selection depends on the amount of genetic variation affecting the trait, if there is no genetic variation, there is obviously no response to selection. What is not so obvious is that some components of genetic variation cannot be acted upon by some kinds of selection. In other words, certain populations have ample genetic variation, yet fail to respond to selection. The part of the genetic variation amenable to selection is clarified by partitioning the variance.

Genetic and Environmental Sources of Variation

As shown in Table 8.3, the phenotypic value of any individual can be represented as a sum of three components: (1) the mean of the entire population, for which we use the conventional symbol μ; (2) a deviation from the population mean due to the specific genotype of the individual in question, symbolized as G_1, G_2, and G_3 for AA, AA', and $A'A'$ genotypes, respectively; and (3) a deviation from the population mean due to the specific microenvironment of the individual in question; these environmental deviations are unique to each individual and are represented as $E_1, E_2, \ldots E_9$. The microen-

TABLE 8.3 Phenotypes of Various Genotypes as the Sum of μ, G, and E[a]

Genotype	Phenotypic value
AA	$\mu + G_1 + E_1$
AA	$\mu + G_1 + E_2$
AA	$\mu + G_1 + E_3$
AA'	$\mu + G_2 + E_4$
AA'	$\mu + G_2 + E_5$
AA'	$\mu + G_2 + E_6$
$A'A'$	$\mu + G_3 + E_7$
$A'A'$	$\mu + G_3 + E_8$
$A'A'$	$\mu + G_3 + E_9$

[a] μ is the population mean. G is a contribution due to genotype, different for each genotype. E is a contribution due to environment, different for each individual.

vironmental effects might be due to random differences in nutrition, temperature, or other external factors, or they might be seen even in an absolutely uniform external environment due to the vagaries of embryonic development. It is important to note that the Gs and Es are not directly observable. Nevertheless, as we shall see, the total variance in phenotypic value can be partitioned into a component due to variation among the Gs and another component due to variation among the Es. The model can be summarized by writing

$$P = \mu + G + E \tag{8.30}$$

where P represents the phenotypic value of any individual and G and E are the genotypic and environmental deviations pertaining to that individual.

To connect the above symbols with actual numbers, we may use Table 8.2 and assume an allele frequency of A of $p = 0.2$. Equation 8.13 then implies that the mean of the population (here denoted μ rather than M) is $\mu = 0.994$. Thus, the respective G_1, G_2, and G_3 deviations for AA, AA', and $A'A'$ genotypes are

$$G_1 = 1.202 - 0.994 = 0.208$$
$$G_2 = 1.059 - 0.994 = 0.065$$
$$G_3 = 0.948 - 0.994 = -0.046$$

For a particular animal of genotype AA whose actual coat color score is, for example, 1.312, the corresponding value of E for the animal would be calculated using Equation 8.30 from the expression $1.312 = 0.994 + 0.208 + E$; thus,

for this animal, $E = 0.11$. Similarly, a particular animal of genotype AA' with an actual phenotype of $P = 1.009$ would have a value of E given by $1.009 = 0.994 + 0.065 + E$, or $E = -0.05$. Because the E values are defined as deviations from their mean, the average of Es for any genotype is 0. Likewise, since the Gs are defined as deviations from their mean, the mean of the Gs is 0. This result can be verified in the guinea pig example because

$$(0.2)^2G_1 + 2(0.2)(0.8)G_2 + (0.8)^2G_3 = 0$$

Equation 8.30 is appropriate when the effects of genotype and environment are *additive*—that is, when the deviation of the phenotype of any particular individual from the population mean $(P - \mu)$ can be written as the sum of an effect resulting from the genotype of that individual and a separate effect resulting from the environment of that individual.

PROBLEM 8.4 In Problem 8.3 the values of μ, a, and d were found to be 0.31, 0.67, and 0.02, respectively, for the logarithms of tomato weight. Calculate the additive genetic variance in the F_2 population, the B_1 population, and the B_2 population.

ANSWER In the F_2 population the allele frequency is $p = q = \frac{1}{2}$, and so the formula for the additive genetic variance (see Equation 8.28) is $\sigma_a^2 = 2pqa^2 = (\frac{1}{2})_a^2 = 0.224$. In the backcross 1 population B_1, the allele frequencies are $p = \frac{3}{4}$ and $q = \frac{1}{4}$, so, applying Equation 8.28, we get $\sigma_a^2 = 0.173$. The backcross 2 population B_2 has allele frequencies $p = \frac{1}{4}$ and $q = \frac{3}{4}$, hence $\sigma_a^2 = 0.163$. When the dominance parameter is so small, the additive variance is at a maximum when the allele frequencies are both $\frac{1}{2}$, and the graph of additive variance against allele frequency is symmetric.

Until this point the discussion has been restricted to a particular population in a single environment, and the sources of variation have been due to genetic and microenvironmental differences among individuals. A change in environment is easiest to see in an experimental setting, where, for example, in environment 1 all of the guinea pigs get twice as much food as in environment 2. Additivity of genetic and environmental effects is true whenever the ratio of $G_1 : G_2 : G_3$ is the same in each of the relevant environments. For the genotypes in Figure 8.15, for example, if the actual range of environments is the range designated E_1, then the genetic and environmental effects are additive because the ratio $G_1 : G_2 : G_3$ is the same for any particular environment in E_1. For the same reason, the genetic and environmental effects are additive if the actual range of environments is E_2.

However, if the actual range of environments includes both E_1 and E_2, then the ratio $G_1 : G_2 : G_3$ depends on the particular environment, and therefore the genetic and environmental effects may not be additive. Nonadditivity of genetic and environmental effects is called **genotype-by-environment interaction**, and in Equation 8.30, it appears that there is an assumption that

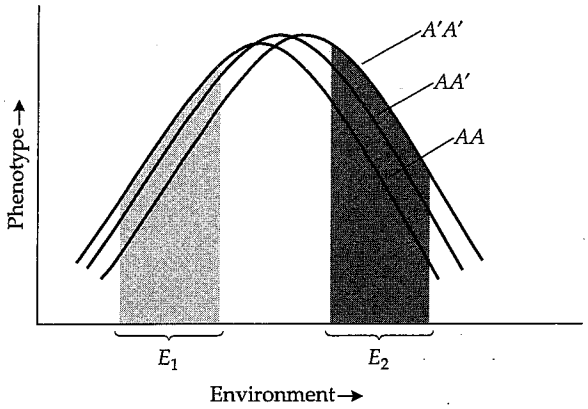


FIGURE 8.15 The *norm of reaction* is the relation between the phenotype and the environment, and this relation is known to vary from genotype to genotype. Hypothetical norms of reaction for genotypes *AA*, *AA'* and *A'A'* are shown here. In the range of environments denoted E_1, *A* is very nearly dominant to *A'* (that is, *AA* and *AA'* have nearly the same phenotype). However, in the range E_2, *A* and *A'* are very nearly additive (no dominance). The heritability of the trait resulting from this gene differs according to whether the population is reared in E_1 environments or E_2 environments.

there is no genotype-environment interaction. In experimentally tractable organisms, it is not necessary to assume that there is no genotype-environment interaction, because we can explicitly examine the phenotypes when reared in different environments and estimate the magnitude of the interaction directly. Alternatively, we can arbitrarily define the environmental variance as also including the effects of genotype-by-environment interaction.

One special type of genotype-by-environment interaction that needs to be singled out for emphasis is called **genotype-by-sex interaction**, which refers to situations in which the effect of QTLs on a trait differ according to the sex of the individual. Genotype-by-sex interaction is widespread among quantitative traits (Mackay 2004; Mosher et al. 2005), and it is often necessary to analyze data for each sex separately. A familiar example is human height, for which the distribution of height among females is shifted to the left relative to males, so that females have a smaller mean height. The segregating alleles that account for genetic variation in height may be the same in females and males, but in females they act in the context of a developmental plan that results in a smaller adult height. Genotype-by-sex interactions can result from biological differences between the sexes, for example, through differences in sex hormones, or they can result from environmental differences between females and males, for example, in differing social opportunities.

When Equation 8.30 is valid, the total phenotypic variance σ_p^2 in the population equals the mean of $(P - \mu)^2$. However, Equation 8.30 implies that $(P - \mu)^2$ equals $(\mu + G + E - \mu)^2$, which is

$$\sigma_p^2 = (G + E)^2 = G^2 + 2GE + E^2 \tag{8.31}$$

Because G and E are already deviations from their means, the mean of G^2 is the phenotypic variance in the population resulting from differences in genotype, and the mean of E^2 is the phenotypic variance resulting from differences in environment. The mean of G^2 is called the **genotypic variance** and is denoted σ_g^2. The mean of E^2 is called the **environmental variance** and is denoted σ_e^2. The remaining term—the mean of $2GE$—is two times the *genotype-by-environment covariance*. If the genotypic and environmental deviations are uncorrelated—that is, if there is no systematic association between genotype and environment—then there is said to be no **genotype-by-environment association**, and the mean of $2GE$ equals zero. When there is no genotype-environment association, therefore,

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 \tag{8.32}$$

Equation 8.32 is the theoretical foundation for partitioning the variance into genetic and environmental effects. The assumption that genotype-by-environment association can be ignored is frequently a valid assumption in animal and plant breeding where, because breeders have a degree of control not available to human geneticists, for example, experiments can be intentionally designed in such a way as to minimize genotype-by-environment association. However, genotype-by-environment association can occur even in animal and plant breeding. For example, dairy farmers routinely provide more feed supplements to cows that produce more milk; because milk-producing ability is partly due to genotype, this feed regimen will provide superior environments (better feed) to cows that have superior genotypes to begin with, so there will be a genotype-by-environment association. Similarly the fastest race horses get the best trainers and the children of highly educated parents often go to the best schools. If one is not careful to correct for such associations, genotype-by-environment association can inflate the apparent σ_g^2 and possibly give spurious overestimates of heritability.

The biological meaning of Equation 8.32 is shown for the alleles of one gene in Figure 8.16. The solid curves represent the phenotypic distributions in the genotypes AA, AA', and $A'A'$ with means denoted G_1, G_2, and G_3, and the dashed curve represents the phenotypic distribution in the entire population. The total phenotypic variance σ_p^2 is the variance of the dashed distribution; the genotypic variance σ_g^2 is the variance among the Gs (i.e., $\sigma_g^2 = p^2G_1^2 + 2pqG_2^2 + q^2G_3^2$, where p is the allele frequency of A); and the environmental variance σ_e^2 is obtained by subtraction: $\sigma_e^2 = \sigma_p^2 - \sigma_g^2$. Although the Gs are not generally known, σ_g^2 must equal zero in a genetically uniform population. The observed variance of a randomly bred population, therefore, provides an

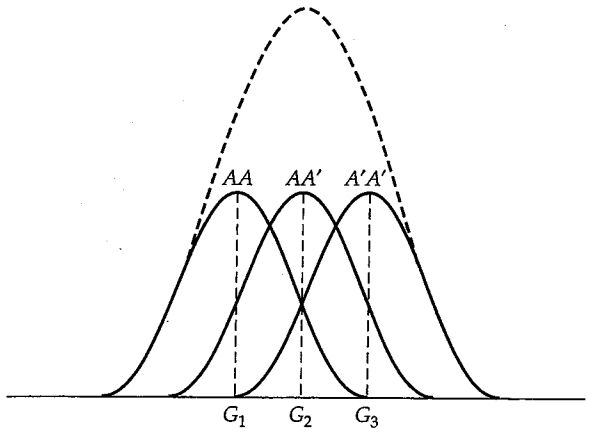


FIGURE 8.16 Phenotypic distribution (dashed curve) of a quantitative trait in a hypothetical population, showing distributions (solid curves) of three constituent genotypes for two alleles of a gene. The means of *AA*, *AA′*, and *A′A′* genotypes are denoted G_1, G_2, and G_3, respectively.

estimate of $\sigma_g^2 + \sigma_e^2$, whereas the average observed variance of genetically uniform populations provides an estimate of σ_e^2. The estimate of σ_g^2 is obtained by subtraction, as shown in an example using thorax length in *Drosophila* (Table 8.4). In this case, genetic variation among individuals in the randomly bred population accounts for about 0.180/0.366 = 49.2% of the phenotypic variance. Genetically uniform populations, such as inbred lines or crosses between inbreds, are not available in human populations, but identical twins are often used instead because of the identical genotypes of the twins, as discussed later in this chapter.

An example of a naturally occurring organism that exhibits remarkably low levels of genetic variability is the African cheetah, *Acinonyx jubatus*

TABLE 8.4 Calculation of Genotypic Variance (σ_g^2) and Environmental Variance (σ_e^2)[a]

	Populations	
Variance	**Random-bred**	**Uniform**
Theoretical	$\sigma_g^2 + \sigma_e^2$	σ_e^2
Observed	0.366	0.186

$\sigma_e^2 = 0.186$

$\sigma_g^2 = (\sigma_g^2 + \sigma_e^2) - \sigma_e^2 = 0.366 - 0.186 = 0.180$

Source: Data from Robertson 1957.

[a] Trait is length of thorax in *Drosophila melanogaster* (in units of 10^{-2} mm).

TABLE 8.5 Expressions for Population Mean and Genotypic Deviations

Genotype	Frequency	Mean phenotype	Genotypic deviation from population mean (G)
AA	p^2	$m + a$	$G_1 = m + a - \mu = 2q[a + (q - p)d] - 2q^2d$
AA'	$2pq$	$m + d$	$G_2 = m + d - \mu = (q - p)[a + (q - p)d] + 2pqd$
$A'A'$	q^2	$m - a$	$G_3 = m - a - \mu = -2p[a + (q - p)d] - 2q^2d$

$$\begin{aligned}\text{Population mean } \mu &= p^2(m + a) + 2pq(m + d) + q^2(m - a)\\ &= (p^2 + 2pq + q^2)m + (p^2 - q^2)a + 2pqd\\ &= (p + q)^2m + (p - q)(p + q)a + 2pqd\\ &= m + (p - q)a + 2pqd\end{aligned}$$

(O'Brien et al. 1983, 1987; May 1995). One might suppose that limited genetic variation would also result in depressed phenotypic variability, however a study of cranial measures by Wayne et al. (1986) revealed that the amount of variability was not appreciably less than that in three other large cats. In fact, there was a significant increase in the amount of **fluctuating asymmetry** (that is, the difference in measurements from the left and the right side of the body). The fluctuating asymmetry is consistent with the notion that genetic homozygosity results in reduced developmental stability—an idea that has considerable empirical support (Phelan and Austad 1994; Leamy and Klingenberg 2005), but so far no good explanation in molecular terms. In any event, reduction of genetic variance and concomitant high homozygosity can result in an increase in phenotypic variance due to developmental instability. Since extreme homozygosity may result in phenotypes that are very sensitive to environmental fluctuations, the paradoxical increase in phenotypic variance results from genotype-by-environment interaction.

Components of Genotypic Variation

So far, the phenotypic variance has been partitioned into the genotypic variance and the environmental variance according to the Equation 8.32. The genotypic variance can be partitioned further into terms that are particularly important for interpreting the resemblance between relatives. The appropriate model is shown in Table 8.5, where the phenotypic means of AA, AA', and $A'A'$ are denoted by $m + a$, $m + d$, and $m - a$, as they were earlier in Figure 8.11. To obtain the G values, the mean of each genotype must be expressed as a deviation from the population mean, which is $\mu = m + (p - q)a + 2pqd$, and the deviations are shown in the last column of Table 8.5. The genotypic variance σ_g^2 is calculated as

$$\begin{aligned}\sigma_g^2 &= p^2G_1^2 + 2pqG_2^2 + q^2G_3^2\\ &= 2pq[a + (q - p)d]^2 + (2pqd)^2\end{aligned} \tag{8.33}$$

The first term on the right-hand side in Equation 8.33 is the additive genetic variance σ_a^2 encountered earlier in Equation 8.28. The second term is a new quantity called the **dominance variance**, which is symbolized σ_d^2. From Equation 8.33, therefore,

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 \tag{8.34}$$

which allows us to express the total phenotypic variance as the sum of three terms, namely

$$\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2 \tag{8.35}$$

When Equation 8.26 for heritability is written in terms of variance components rather than p, q, a and d, the equation implies that

$$h^2 = \sigma_a^2/\sigma_p^2 \tag{8.36}$$

Equation 8.36 is an important result because it states that the heritability depends only on the additive genetic variance and not on the dominance variance. Therefore, if all the genetic variance in a population results from dominance variance (i.e., $\sigma_a^2 = 0$), then the population cannot respond to individual selection because in this case h^2 equals zero. To restate the same idea in another way, the dominance variance σ_d^2 represents that portion of the genetic variance that is not acted upon by individual selection.

Equation 8.36 means that the heritability of a trait is the ratio of the additive genetic variance to the total phenotypic variance. Sometimes the word *heritability* is used in reference to a different variance ratio, namely the ratio of the total genotypic variance to the total phenotypic variance (i.e., σ_g^2/σ_p^2). To avoid confusion, quantitative geneticists distinguish the two types of heritability as follows:

1. The ratio σ_a^2/σ_p^2 is called **heritability in the narrow sense** (This is the variance ratio we have been using all along.)
2. The ratio σ_g^2/σ_p^2 is called **heritability in the broad sense** (often denoted H^2).

Generally speaking, narrow-sense heritability is the more important with individual selection (or any mode of selection that capitalizes primarily on the additive genetic variance), whereas broad-sense heritability is the more important when selection is practiced among clones (a clone is a group of genetically identical individuals), inbred lines, or varieties. We use the term *heritability* to mean narrow-sense heritability unless otherwise stated.

The additive and dominance components of genetic variation σ_a^2 and σ_d^2 are statistical attributes of a quantitative trait in a population that depend on the allele frequencies p and q as well as the particular values of a and d (Figure 8.17). Estimates of narrow-sense heritability also depend on these parameters (Figure 8.18). With many genes that act together, σ_a^2 is defined as a summation of the values of $2pq[a + (q - p)d]^2$ for each gene affecting the trait, and σ_d^2 represents a summation of the values of $(2pqd)^2$ for each gene. Fur-

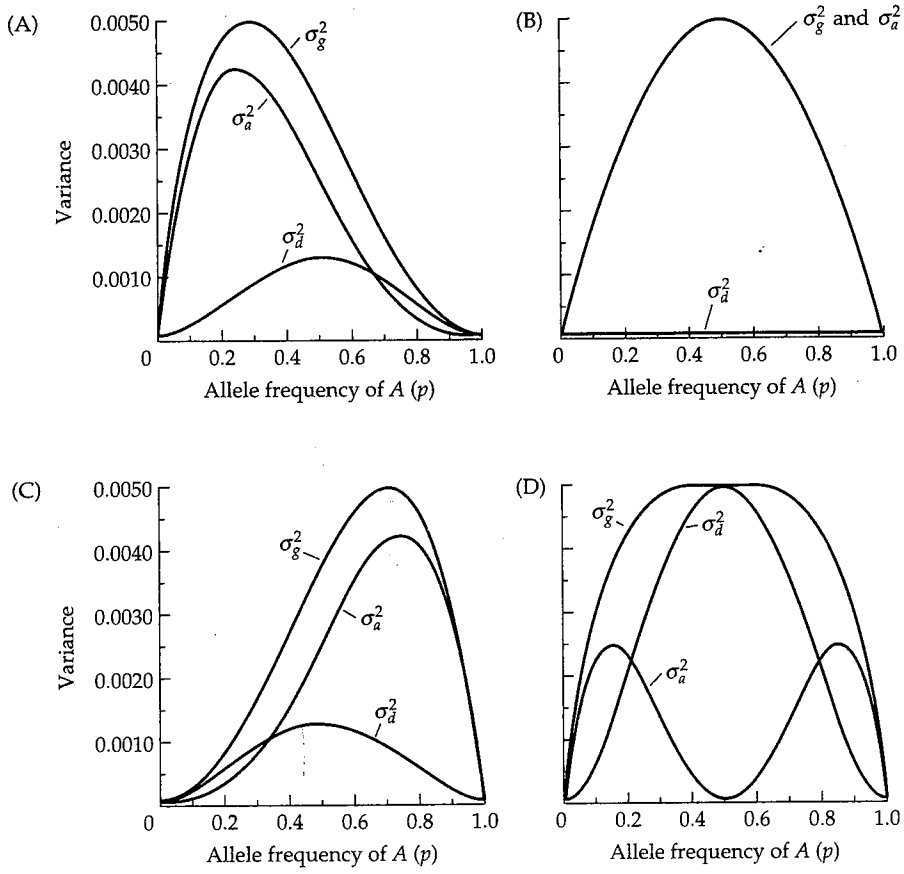


FIGURE 8.17 Total genetic variance (σ_g^2), additive genetic variance (σ_a^2), and dominance variance (σ_d^2) for a locus with two alleles (A and A') plotted against the frequency of allele A (p). The mean phenotypes of AA, AA' and $A'A'$ are denoted $m + a$, $m + d$, and $m - a$, respectively. In all cases, the variances are $\sigma_a^2 = 2pq[a + (q - p)d]^2$, $\sigma_d^2 = (2pqd)^2$, and $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$. (A) $a = d = 0.0701$ (A dominant to A'); (B) $a = 0.1$, $d = 0$ (no dominance), (C) $d = -a = 0.0707$ (A' dominant to A); (D) $a = 0$, $d = 0.141$ (overdominance). For ease of comparison, the values of a and d have been chosen to make the maximum of σ_g^2 equal to 0.005 in each case.

thermore, when the trait is affected by multiple genes, the formula for σ_g^2 in Equation 8.34 must be extended to include an additional term that pertains to interaction among the genes. This interaction term is called the **interaction variance** or the **epistatic variance** and is symbolized σ_i^2. With the interaction variance included, Equation 8.34 becomes

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2 \qquad (8.37)$$

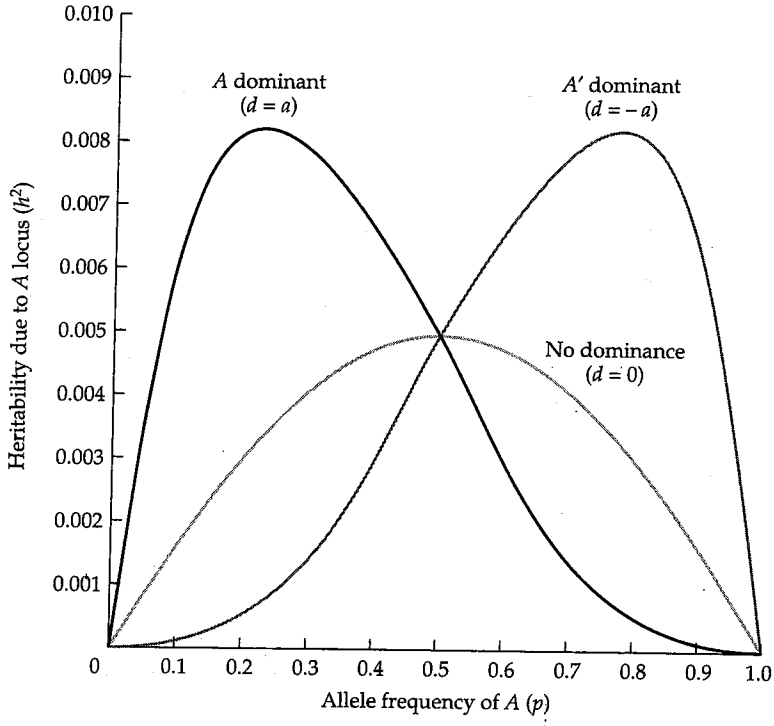


FIGURE 8.18 Narrow-sense heritability due to a single locus with two alleles (A and A') as a function of p, the allele frequency of A. In general, for one locus, $h^2 = 2pq[a + (q - p)d]^2/\sigma_p^2$, where σ_p^2 is the total phenotypic variance. The curves correspond to $a = 0.1$, and $d = 0.1$ (A dominant), $d = 0$ (no dominance), and $d = -0.1$ (A' dominant).

The important point to remember about the components of genotypic variance is that they represent the cumulative, statistical effects of all genes affecting the trait. Few inferences about the actual mode of inheritance of the trait are possible from the variance components, particularly concerning the number of genes involved and their individual effects.

The term **epistasis** requires a little more discussion, since it is used by geneticists in different, and sometimes contradictory, ways (Moore 2005). The term was originally introduced by William Bateson to describe situations in which the phenotypic effect of one gene interfered with the expression of another gene (Bateson 1909). A classic example is a gene in white Leghorn chickens in which the effects of a gene for feather color is concealed by a dominant allele at another locus that prevents any pigment from being deposited. Through the years molecular geneticists broadened the definition of epistasis to include any sort of interaction between the alleles of different genes. Quantitative geneticists, however, began using the term in a different

way to refer to those specific types of gene interaction that contribute to the epistatic variance in the sense of Equation 8.37. Much confusion has arisen from the fact that the products of genes can have very strong interactions that may profoundly affect the physiology of the organism, yet make little or no contribution to the epistatic variance of a quantitative character. A way to reconcile these concepts has been suggested by Cheverud and Routman (1995), who provide a quantitative definition of epistasis that clarifies the distinction between interaction effects on physiology and interaction effects on the epistatic variance.

PROBLEM 8.5 By definition, a simple Mendelian trait is one that is determined entirely by genotype in the prevailing environment. Therefore, $\sigma_e^2 = 0$ in Equation 8.32, and the broad-sense heritability $\sigma_g^2/\sigma_p^2 = 1$. Show that, for a simple Mendelian recessive, the narrow-sense heritability equals $2q/(1 + q)$, where q is the recessive allele frequency, and that for a trait due to a rare recessive allele the narrow-sense heritability is approximately equal to the frequency of heterozygous genotypes.

ANSWER Let the phenotypes of AA, AA', and $A'A'$ be assigned phenotypic values 0, 0, and 1, respectively, so that the A' allele is recessive. In this case, $m = \frac{1}{2}$, $a = -\frac{1}{2}$, and $d = -\frac{1}{2}$ (see Equation 8.12). The numerator of Equation 8.36 is the additive genetic variance, which equals $2pq^3$. Then the mean phenotype equals q^2, and the variance in phenotypic value is $q^2 - (q^2)^2 = q^2(1 - q^2) = q^2(1 + q)(1 - q) = pq^2(1 + q)$. The heritability is the additive variance divided by the phenotypic variance, namely $2pq^3/[pq^2(1 + q)] = 2q/(1 + q)$. When the autosomal recessive trait is rare, $q \approx 0$ and the heritability is approximately $2q$. Since $p \approx 0$, the frequency of heterozygous carriers is $2pq \approx 2q$ also.

8.6 COVARIANCE AMONG RELATIVES

Components of genetic variation are important because they may be used to express the phenotypic covariance between relatives. Since the distribution of offspring from a given parental genotype depends on the distribution of potential mates, the variance components and estimates of heritability depend not only on allele frequencies but also on the distribution of genotype frequencies. To simplify things, we assume that the trait is determined by one gene with two alleles, and that the population is in Hardy-Weinberg proportions. However, the same results are also true for many genes when the trait is determined by summing the individual allelic effects provided that the population is in multilocus linkage equilibrium.

Table 8.6 displays three genotypes of parents, their genotypic value, and the mean genotypic value of the offspring with random mating. As an example of where these numbers come from, consider the parental genotype AA, which has a frequency in the population of p^2. We have been usually been writing the genotypic value of AA as $m + a$, but here it is more convenient to

TABLE 8.6 Derivation of Parent–Offspring Covariance

Parent's genotype	Frequency	Genotypic value[a]	Offspring mean genotypic value[a]
AA	p^2	$2q(a - pd)$	$aq + dq(q - p)$
AA'	$2pq$	$a(q - p) + d(1 - 2pq)$	$\frac{1}{2}a(q - p) + \frac{1}{2}d(q - p)^2$
$A'A'$	q^2	$-2p(a + qd)$	$-ap - dp(q - p)$

[a] The values in these columns are expressed as deviations from the population mean.

express the genotypic value of each genotype as a deviation from the population mean, which equals $m + (p - q)a + 2pqd$ (see Table 8.5). For parental genotype AA, therefore, the genotypic value expressed as a deviation from the population mean equals $m + a - [m + (p - q)a + 2pqd] = 2q(a - pd)$, which is the entry in the third column. With random mating, an AA parent has an AA offspring with probability p and an AA' offspring with probability q. The mean of the genotypic values of the offspring of an AA parent is therefore $p[2q(a - pd)] + q[a(q - p) + d(1 - 2pq)] = aq + dq(q - p)$, which is the entry in the fourth column. A similar calculation yields the corresponding values for each of the other possible parental genotypes. Since the genotypic values are already expressed as deviations from the population mean, the covariance of the offspring and one parent is calculated by summing the product of the last three columns of Table 8.6. After tedious algebra, the covariance of offspring and parent turns out to be:

$$Cov(offspring, parent) = pq[a + (q - p)d]^2 = (\tfrac{1}{2})\sigma_a^2 \qquad (8.38)$$

This is a remarkably simple result because it says that the covariance in phenotype of parents and offspring is one-half the additive genetic variance. No component of dominance inflates the covariance in this case. Since environmental effects are assumed to be random with respect to genotypic values (there is no genotype-by-environment association), environmental effects also play no role in parent-offspring covariance. We must, however, assume that the environments of parents and offspring are uncorrelated for Equation 8.38 to be valid. In order to see the relation between narrow-sense heritability and regression, recall that the regression coefficient of y on x is defined as $Cov(x, y)/Var(x)$ (see Equation 8.6), and so the regression of offspring on one parent is, from Equation 8.27:

$$\begin{aligned} b_{OP} &= Cov(offspring, parent)/\sigma_p^2 \\ &= \left(\frac{1}{2}\right)\left(\frac{\sigma_a^2}{\sigma_p^2}\right) \\ &= \left(\frac{1}{2}\right)h^2 \end{aligned} \qquad (8.39)$$

TABLE 8.7 Theoretical Covariance in Phenotype between Relatives[a]

Degree of relationship	Covariance
Offspring and one parent	$\sigma_a^2/2$
Offspring and average of parents (midparent)	$\sigma_a^2/2$
Half siblings	$\sigma_a^2/4$
Full siblings	$(\sigma_a^2/2) + (\sigma_d^2/4)$
Monozygotic twins	$\sigma_a^2 + \sigma_d^2$
Nephew and uncle	$\sigma_a^2/4$
First cousins[b]	$\sigma_a^2/8$
Double first cousins	$(\sigma_a^2/4) + (\sigma_d^2/16)$

[a] Variance terms due to interaction between loci (epistasis) have been ignored.

[b] First cousins are the offspring of matings between siblings and unrelated individuals; double first cousins are the offspring of matings between siblings from two different families.

In the regression of offspring on the **midparent** (that is, the average value of the parents), the denominator in Equation 8.39 becomes $\sigma_p^2/2$ instead of σ_p^2 because $\sigma_p^2/2$ is the variance of the mean of the two parents, assuming random mating. Hence the regression coefficient on the midparent equals $(\sigma_a^2/2)/(\sigma_p^2/2) = \sigma_a^2/\sigma_p^2$, or to say the same thing in another way, the regression coefficient of offspring on midparent equals the narrow-sense heritability.

The same reasoning can be followed to obtain covariances between other pairs of relatives, as summarized in Table 8.7. As can be seen, the additive genetic variance can be estimated directly either from parent-offspring covariance or from half-sib covariance. However, full-sib covariance includes a term resulting from dominance. The expressions in Table 8.7 are correct as long as there are no complications such as genotype-by-environment associations or other nonrandom environmental effects such as full sibs sharing environmental factors common to the whole family but not shared by other families. Since the total variance in phenotypic value σ_p^2 can be estimated directly, once σ_a^2 is estimated from the covariance between relatives, the narrow-sense heritability can be estimated from Equation 8.36. The first three relationships in Table 8.7 are the most useful in quantitative genetics and are commonly used in animal and plant breeding. The other relationships are used mainly in human quantitative genetics.

Figure 8.19 presents the narrow-sense heritabilities of diverse quantitative traits in farm animals and one important crop plant as estimated from the correlation between relatives. The data are presented merely to show the values of heritability with which breeders typically must deal. It is important to keep in mind that the heritabilities in Figure 8.19 pertain to one population

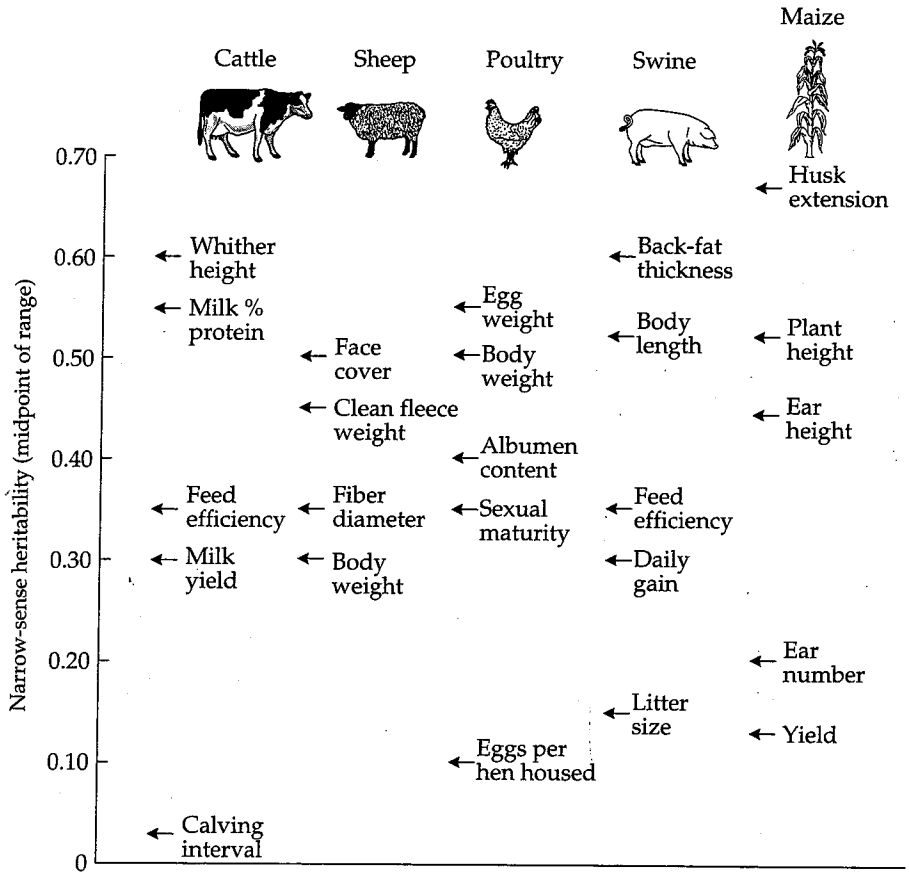


FIGURE 8.19 Narrow-sense heritabilities for representative traits in plants and animals. Traits closely related to fitness (calving interval, eggs per hen, litter size of swine, yield and ear number of corn) tend to have rather low heritabilities. (Animal data from Pirchner 1969, who gives the range of heritabilities in various studies. The midpoint of the range is plotted here. Corn data from Robinson et al. 1949.)

in one type of environment at one particular time. The same trait in a different population or in a different environment might well have a different heritability. Generally speaking, traits that are closely related to fitness (such as calving interval in cattle or eggs per hen in poultry) tend to have rather low heritabilities. Ignoring complications such as antagonistic pleiotropy (discussed later), long-term natural selection is expected to gradually reduce the additive genetic variance until the effect is balanced against the input of new mutations.

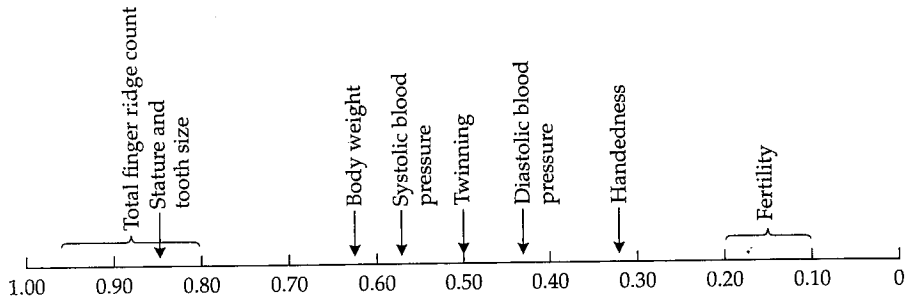


FIGURE 8.20 Broad-sense heritabilities and ranges of heritabilities of various traits in humans. Uncertainties about the correlation between environments of relatives make such estimates in humans very tentative. (Data from Smith 1975.)

For purposes of comparison, Figure 8.20 shows estimated broad-sense heritabilities of a number of quantitative traits in humans. Broad-sense heritabilities vary widely for different traits, as they do in other species. Note the low heritability of fertility, a trait that is obviously closely related to fitness. At the other end of the scale is total fingerprint ridge count, which is apparently not a major component of fitness considering its relatively high broad-sense heritability.

The distinction between estimation of variance components and knowledge of genetic causes of human differences has particularly important implications for social applicability of human quantitative genetics. Some of the problems are forcefully conveyed by Lewontin (1974a) and Feldman and Lewontin (1975). In particular, estimates of heritability within a population, even if they are sound, tell us nothing about the degree to which genetic differences account for the differences in phenotypes between populations. Experimentalists working with organisms that can be manipulated in the field or laboratory can be more rigorous in the assessment of genetic parameters by examining the traits in several environments, and by doing studies that combine classical and quantitative genetic analysis.

Twin Studies and Inferences of Heritability in Humans

Because identical twins are genetically identical, phenotypic differences between identical twins would seem to be a straightforward measure of how much phenotypic variance is caused by environment. Twin studies raise their own unique problems, however, and the results must be interpreted with caution. Before discussing the use of twins in quantitative genetics, we should back up a few steps and first discuss the phenomenon of twinning itself.

Twins are relatively frequent among human births, though the rate of twinning varies from population to population. Among Caucasians in the

United States, for example, about one in 88 births results in twins; among Japanese in Japan, the rate is about one in 145 births (Bulmer 1970). Two kinds of twins actually occur. **Identical twins**, often called **monozygotic twins** or **one-egg twins**, arise from a single zygote that very early in embryonic development splits into two distinct clumps of cells, each clump thereafter undergoing its own embryonic development. Because they arise from a single zygote, identical twins are necessarily genetically identical. The other kind of twins are called **fraternal twins**, **dizygotic twins**, or **two-egg twins**. Fraternal twins arise from a double ovulation in the mother, each egg being fertilized by a different sperm. Because of their mode of origin, fraternal twins are related genetically as siblings. Most of the variation in twinning rates in humans is due to variation in the rate of dizygotic twinning. In the absence of hormonal treatment for infertility, the rates of monozygotic twinning among Caucasians in the United States and among Japanese in Japan are one in 256 and one in 238, respectively, whereas the respective rates of dizygotic twinning are one in 135 and one in 370 (Bulmer 1970). With hormonal treatment for infertility, the rates of both dizygotic and monozygotic twinning increases dramatically.

For studies in quantitative genetics, identical twins are often compared with same-sex fraternal twins in order to discount the effects of common intrauterine environments. Such an approach is only partially successful, as identical twins often share embryonic membranes *in utero* (the amnion and chorion) that are not usually shared by fraternal twins. Moreover, because identical twins often have astonishingly similar facial features, they may be treated more similarly by parents, teachers, and peers than are fraternal twins. Some of these problems can be overcome by studying twins that are raised apart (in different households), but data of this sort are usually limited (Shields 1962; Bouchard and McGue 2003), making estimates of heritability highly imprecise. Even when twins are reared apart, the environments into which they are adopted are generally similar. This effect of correlated environments has the effect of inflating the apparent degree to which traits are genetically determined. In any case, if r_{MZ} and r_{DZ} represent the correlation coefficients of a quantitative trait among monozygotic and dizygotic twins, then $2(r_{MZ} - r_{DZ})$ provides a rough estimate of the broad-sense heritability of the trait. To see where this formula comes from, first look at Table 8.7. The covariance of monozygotic twins in the absence of environmental correlation is $\sigma_a^2 + \sigma_d^2$, and so the correlation between monozygotic twins is this covariance divided by the phenotypic variance, or the broad-sense heritability. If monozygotic and dizygotic twins have the same degree of environmental correlation, then subtracting the correlation of one from the other should remove the environmental correlation. The correlation between dizygotic twins is in theory the same as that of full sibs, or $(\frac{1}{2})\sigma_a^2 + (\frac{1}{4})\sigma_d^2$. Assuming that the phenotypic variance is the same in both types of twins, the expression $2(r_{MZ} - r_{DZ})$ is equal to $[\sigma_a^2 + (\frac{3}{2})\sigma_d^2]/\sigma_p^2$, which is not exactly equal

to the broad-sense heritability, but it is an approximation (Smith 1975). Even when the mathematically precise estimators are used, the problem of shared environments does not go away.

Estimating genetic variance components is a very different thing from identifying the causes of differences among human populations (Lewontin 1974a; Feldman and Lewontin 1975; Ehrlich and Feldman 2003). The key point is that heritability is a ratio of variances based on differences among individuals *within* a population. Even when technically impeccable, estimates of heritability within a population tell us nothing about whether genetic variation contributes to differences in average phenotype *among* populations. Particularly in studies of behavioral or achievement differences among human populations, misunderstanding the concept of heritability and failure to recognize its limitations have led to a great deal of mischief. Heritability is an important concept, but one easily open to misinterpretation.

Estimation of Genetic Variance Components in Natural Populations

Field biologists face special challenges in estimating the variance components of quantitative traits in natural populations because of the need to identify and keep track of familial relationships. In practice, because the natural environment is so variable, it is often preferable to do analyses in a controlled laboratory environment, but then the environmental variance is that of the laboratory and not that in nature. The special approaches and statistical methods applicable to estimating heritability and other components of variation in natural populations are discussed in Lynch and Walsh (1998) and Conner and Hartl (2004).

The computational prowess of modern computers has stimulated the development of statistical methods that are more powerful than traditional analyses such as parent-offspring regression. Two such methods for estimating the variance components are maximum likelihood (ML) and restricted maximum likelihood (REML) (Lynch and Walsh 1998). These approaches are based on analysis of a likelihood function that describes the likelihood of obtaining the observed data given the family structure and a set of unknown parameters to be estimated. The unknown parameters are the magnitudes of the various genetic and environmental variance components. The method then finds the values of the unknown parameters that maximize the likelihood function. In practice, the unknown parameters form a variance-covariance matrix whose elements are the variances and covariances of pairs of phenotypic traits, and therefore the computer algorithms entail extensive matrix manipulations. The computer output consists of estimates of the parameters, including the heritability of each trait, that take into account all relationships that are informative, for example, parent-offspring, full-sib, half-sib, first cousin, and so forth.

These methods gain their power by making use of the data in all informative relationships and can handle the large and often unbalanced datasets collected from wild populations (Knott et al. 1995; Kruuk 2004; Charmantier and Garant 2005). The models also allow different components of the environmental variance to be estimated, such as the variance due to maternal effects, and they are relatively robust to departures from normality of the phenotypic distributions. On the other hand, although multivariate normality of the character data is not necessary to obtain the estimates, the estimates have their conventional meaning, and the prediction equation is reliable only if the phenotypes are normally distributed. Testing the statistical significance of the estimates does requires normality. Shaw (1987) provides a thorough review of the merits of maximum likelihood methods in quantitative genetics.

8.7 NORM OF REACTION, THRESHOLD TRAITS, AND GENETIC CORRELATION

In this section we consider a few more advanced features of quantitative traits. First we look at the way in which quantitative traits respond to changes in the environment and how these responses can profoundly affect components of genetic variance and the heritability. Then we examine categorical traits (all-or-none traits) and show how multiple genetic factors affecting categorical traits can be understood from the perspective of quantitative genetics by interpreting risk as an unobservable continuous variable. Finally, we discuss why selection for one trait often results in response of a different trait, and we specify a prediction equation for such correlated responses.

Norm of Reaction and Phenotypic Plasticity

In considering the influence of environment on the determination of phenotypes, it is tempting to think of environmental effects as random noise added to traits that are basically genetically determined. This view is often misleading. One simple way to analyze the effect of environment on phenotype is to examine the phenotype produced by a single genotype across a range of environments. This is particularly easy to do in model organisms that can be reared under different laboratory conditions. The phenotypic response of an organism to a change in environment is called **phenotypic plasticity**, and the range of phenotypes that results from a given genotype is known as the **norm of reaction**, a term coined by the Russian evolutionary biologist Ivan Schmalhausen (1949).

Figure 8.21 shows the norm of reaction of the level of transcription of a gene involved in cell wall maintenance in budding yeast (Landry et al. 2006). Six diploid strains were studied across a range of environments representing different levels of stress (principally nitrogen starvation). Five of the strains

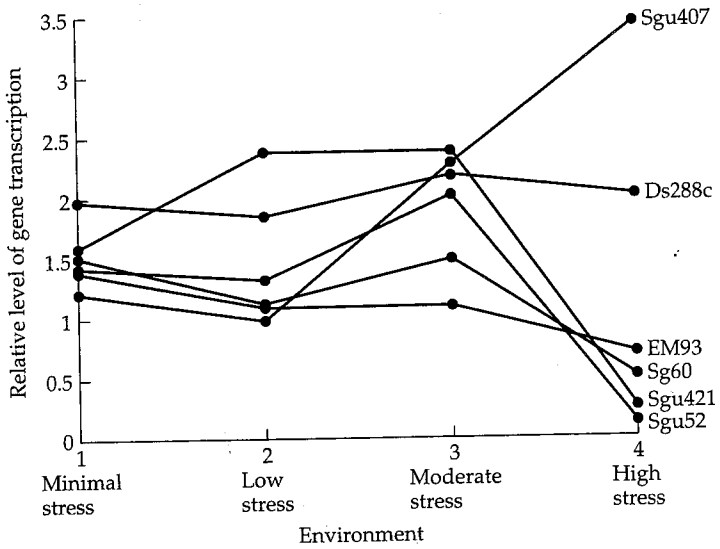


FIGURE 8.21 Norm of reaction for level of transcription of the gene *CRH1* in budding yeast *Saccharomyces cerevisiae*. The gene product is a glycosidase that plays a role in maintaining cell-wall architecture. The stress conditions relate to the source and abundance of nitrogen in the medium. The variance in gene expression among strains is much greater under high stress, and the strain whose *CRH1* gene is expressed most abundantly under high stress is expressed the least under minimal stress. (Data from Landry et al. 2006.)

are natural isolates of *S. cerevisiae*, whereas Ds288c is a diploid derivative of the standard haploid laboratory strain S288c. The rank ordering of the strains differs according to the stress level, and an astounding 50% of pairwise comparisons of strains are reversed in ranking when one stress condition is compared to another. Figure 8.21 shows a greater variation among lines at high stress than at minimal stress. Therefore, heritability estimates carried out under conditions of minimal stress are expected to be greater than those carried out under conditions of high stress. This finding is in accord with experiments in wild populations of animals, in which morphometric traits not closely related to fitness generally show increased heritability under more favorable conditions (Charmantier and Garant 2005). The difference underscores the fact that heritability is a measure defined in one environment, and it shows that the methods of quantitative genetics cannot separate the causes of phenotypic differences among populations living in different environments. In order to detect environmental effects, the norm of reaction must be ascertained by examining phenotypes in a variety of environments.

In addition to direct effects on the environmental variance, the environment can also affect heritability because the values of a and d depend on the environment. Figure 8.15 may be used as an example; it shows the norms of reaction of AA, AA', and $A'A'$ genotypes. If we were dealing with the range of environments denoted E_1 in Figure 8.15, A would be the favored allele and A would be nearly dominant to A'. On the other hand, if we were dealing with the range of environments denoted E_2, A' would be the favored allele and there would be essentially no dominance. Thus, switching a population from E_1 to E_2 would change the values of a and d and substantially alter the heritability of the trait, even though the total phenotypic variance of the population might remain the same.

PROBLEM 8.6 The accompanying table shows the relative viability of homozygous chromosomes A and B extracted from natural populations of *Drosophila pseudoobscura*, relative to the heterozygous genotype, estimated at each of two temperatures (Dobzhansky and Spassky 1944).

	Genotype		
Temperature (°C)	***A/A***	***A/B***	***B/B***
16.5	0.92	1.00	0.71
25.5	0.32	1.00	0.75

From these data, estimate m, a, and d for the AA, AB, and BB genotypes in populations maintained at 16.5 °C and 25.5 °C. Then, letting p represent the frequency of the A chromosome and q that of the B chromosome, assume $p = 0.3$ and estimate the additive genetic variance of viability resulting from these genotypes at both temperatures.

ANSWER From Equation 8.12, at 16.5 °C, $m = (0.92 + 0.71)/2 = 0.815$, $a = 0.92 - 0.815 = 0.105$, and $d = 1.00 - 0.815 = 0.185$. At 25 °C, $m = 0.535$, $a = -0.215$, and $d = 0.465$. The additive genetic variance for these genotypes is given by Equation 8.28, where $p = 0.3$, and $q = 0.7$. At 16.5 °C, $\sigma_a^2 = 0.01346$. At 25.5 °C, $\sigma_a^2 = 0.00035$. Note that the additive genetic variance has been decreased by a factor of almost 40, yet all we did was raise the temperature!

The norm of reaction is important in evolutionary genetics because the fate of genetic variation in a population depends on the fitness of the organism, which in turn depends on the environment. In turn, the norm of reaction is itself a property that may be under genetic control and be subject to adaptive evolution by means of natural selection (Schlichting and Pigliucci 1994; Via et al. 1995; Scheiner 2002; de Jong 2005). A useful way to model the evolution of adaptive phenotypic plasticity is to consider the phenotype in different environments as different traits that may be genetically correlated

(Falconer and Mackay 1996; Via and Lande 1985; Scheiner 2002). Further discussion of this model will be deferred until after we have discussed selection of more than one character.

One biological factor that affects the correspondence between genotypes and phenotypes is the chronological age of an organism. A phenotype, such as body weight, often changes with age, and different genotypes can have different age-related growth curves or other developmental profiles (Meyer and Kirkpatrick 2005). Consequently, the heritability of traits depends on the age at which individuals are tested. Figure 8.22 shows components of variance of mouse body weight at different ages. In this study, 2700 mice from 700 full-sib families were measured from days 14 to 70 (Riska et al. 1984). Variance of all components was maximal at an age of about 20 days (which happens to be the time of maximal growth), and then declined to fairly stable values at age 40 days. Despite the additive genetic variance decreasing markedly after 20 days of age, the heritability itself actually increased with age, with values of 22%, 27%, 31%, and 37% at days 12, 20, 50, and 70, respectively. Although there is an assumed increase in environmental vari-

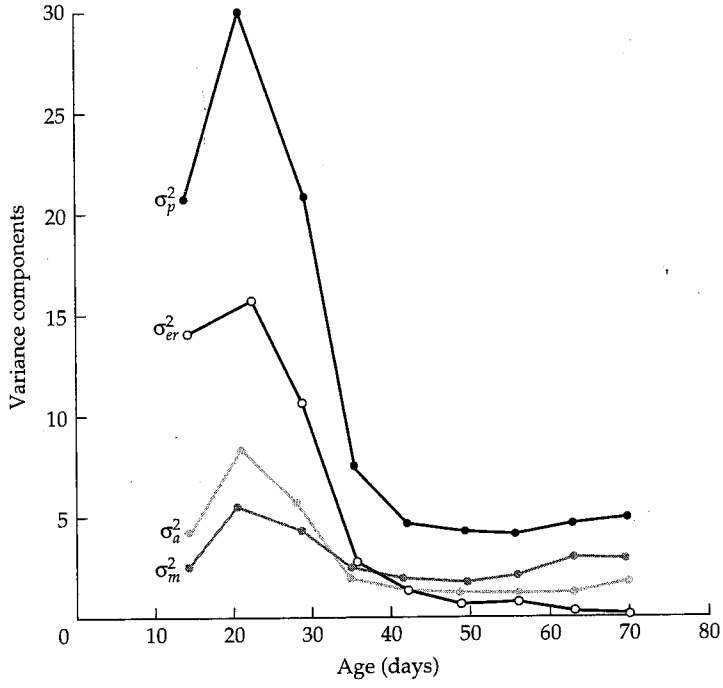


FIGURE 8.22 Variance components for the logarithm of body weight in a random-bred (genetically variable) strain of mice plotted as a function of age. σ_m^2 is the variance due to maternal effects, and $\sigma_{er}^2 = \sigma_e^2 - \sigma_m^2$. (From Riska et al. 1984.)

ance in many traits in humans, there is no marked tendency for heritability to decrease with age.

Threshold Traits: Genes as Risk Factors in Disease

Some multifactorial traits do not exhibit continuous variation. Although the variation is discontinuous (individuals either express the trait or not), the trait is nevertheless influenced by multiple genetic factors and also by environment. Such traits are called **threshold traits**. A human example is diabetes, an abnormality in sugar metabolism that affects one or two percent of the Caucasian population. In a sense, diabetes is a continuous trait because the severity of the disease varies from nearly undetectable to extremely severe. On the other hand, diabetes can also be considered a threshold trait because all individuals may be classified according to whether or not they are so severely affected that clinical treatment is required. With such a classification, there are only two phenotypes, "affected" and "not affected," even though there is phenotypic variation within each category. The genetic influence on the trait is shown by the enhanced risk of diabetes in relatives of affected individuals. However, environmental factors such as diet are also important in determining whether high-risk genotypes actually develop the disease. At one time, many threshold traits were "explained" by postulating a simple genetic mechanism (such as a single recessive allele in the case of diabetes) and invoking "incomplete penetrance" to account for the poor fit of pedigree data to a simple Mendelian hypothesis. Now it is preferred, and probably more realistic, to consider threshold traits as true polygenic traits and to calculate heritabilities as for any other quantitative trait. The use of single-nucleotide polymorphisms to map the relevant QTLs will help to identify the genes affecting such traits and the magnitudes of their individual effects (see Chapter 10).

The basic idea behind the threshold model is illustrated in Figure 8.23. The normal curve in panel (A) represents the (unobservable) distribution of a hypothetical **liability** (or **risk**) toward the threshold trait, measured on a scale such that the mean value is 0 and the variance is 1. It is assumed that individuals whose liability is above a certain threshold (T) actually express the trait. Thus, the shaded area in Figure 8.23A delimits the proportion of individuals in the population who are affected (B_p), and the mean liability among affected individuals is denoted M_S. Figure 8.23B gives the (again unobservable) distribution of liability among the offspring of affected individuals. The offspring mean is denoted M', and the proportion of offspring above the threshold is denoted B_o. The setup here is like that in the earlier section of this chapter in which we calculated the regression coefficient of offspring on one parent. In this model it can be shown that the regression coefficient b is given by $b = M'/M_S$, and the appropriate estimate of the heritability of liability is obtained from the relation

$$h^2 = 2b = 2M'/M_S \tag{8.40}$$

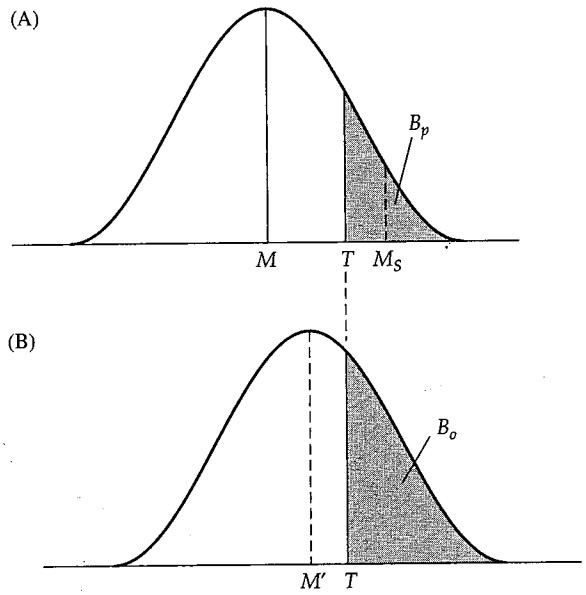


FIGURE 8.23 (A) Distribution of liability assumed for a threshold trait in a hypothetical population. The shaded area denotes individuals having liability above a critical threshold (T) and consequently affected with the trait. B_p is the frequency of affected individuals in the entire population, M is the mean liability of individuals in the entire population, and M_S, the mean liability of affected individuals. (B) Distribution of liability among offspring who have one parent affected with the trait. M' denotes the mean liability among offspring, and B_o is the proportion of affected offspring.

One method for calculating heritability of threshold traits is illustrated in Problem 8.7.

PROBLEM 8.7 For pyloric stenosis, the incidence among males in the general population is B_p = 0.005, and the incidence among sons of affected males is B_o = 0.05. If liability follows a normal distribution, and 0.005 is the frequency of affected ir als, this represents 2.89 standard de above the mean. From these two nu infer M_S and M' in order to estimate t itability of liability.

ANSWER The mean liability of the fathers is M_S = 2.89. B_p can be obtained from a table of the normal distribution as follows. The tables generally give the value on the x-axis in standard deviation units for an observed area in the two tails of the distribution. If the fraction affected is 0.05, and this is the area in one tail, then the area in both tails is 0.10. A probability area of greater than 0.10 is obtained if an observation is more than 2.58 standard deviations from the mean. Thus, T = 2.58 is the threshold. Using the same ing for the sons, $2B_o$ = 0.10, and th appears in the tail of the normal distrib an observation is 1.64 standard de from the mean. This means that T 1.64. We know that T = 2.58, so M' = = 0.94. From Equation 8.40 we g $2M'/M_S$ = 2(0.94)/(2.89) = 0.65. Thi estimate of the narrow-sense herita liability to pyloric stenosis.

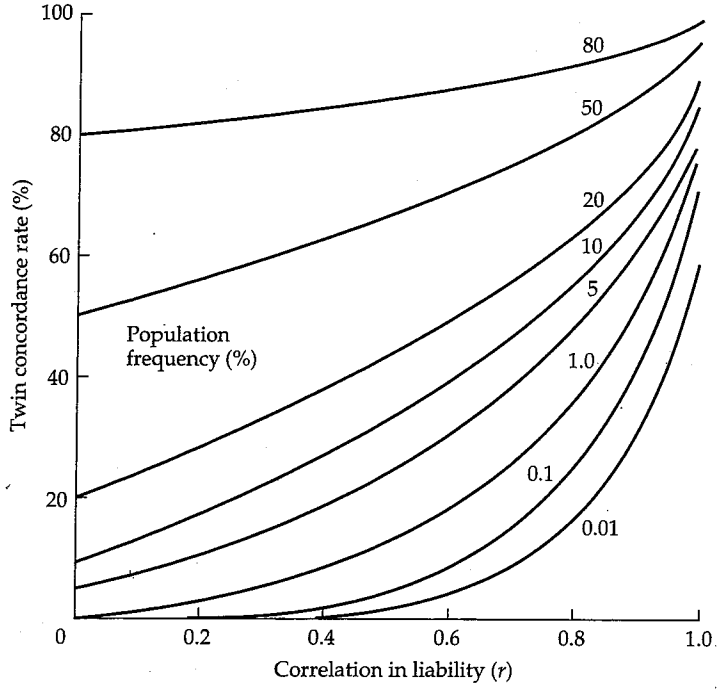


FIGURE 8.24 Expected concordance rates for threshold traits in monozygotic twins, plotted against the correlation in liability toward the trait and the population incidence of the trait. (From Smith 1975.)

Twins are frequently used in human quantitative genetics for the study of threshold traits, but twin data are best expressed in terms of **concordance**. The concordance of a trait in a population of twins is the proportion of affected twins that have affected co-twins. For example, suppose that 100 affected individuals are found to be twins and that in 35 cases the co-twin is also affected. The concordance rate is then $\frac{35}{100} = 35\%$. From the concordance rates for monozygotic and dizygotic twins and the incidence of the trait in the population, the correlations in liability between monozygotic twins (r_{MZ}) and dizygotic twins (r_{DZ}) can be calculated (Figure 8.24). The broad-sense heritability is then estimated by $2(r_{MZ} - r_{DZ})$, as discussed earlier. As with quantitative traits in general, twin data are more reliable if the twins are reared apart, but it is seldom possible in practice to obtain a sufficient number of such twin pairs.

Genetic Correlation and Correlated Response

Every gene potentially affects every trait in the organism, either as a primary effect or as a secondary, indirect effect known as a **pleiotropic effect**. Therefore, the alleles that are favorable for one quantitative trait may have unfavorable effects on another quantitative trait, and as these alleles are increased in frequency by artificial selection (thereby improving the pheno-

typic value with respect to the selected quantitative trait), the very same alleles may bring about a deterioration of some other aspect of performance. Pleiotropy is one cause of correlated response—a change in phenotypic value of one trait that accompanies response to selection of a different trait. A second possible cause of correlated responses is linkage disequilibrium (see Chapter 2)—a favorable allele for one trait that increases in frequency under selection may drag along with it an allele of another, tightly linked gene that has a detrimental effect on an unselected trait.

Correlated responses are quite common in artificial selection and often, but not always, result in a deterioration in reproductive performance. A well-documented example is the selection in Leghorn chickens, in which 12 generations of selection for increased shank length reduced the egg hatchability by nearly half (Lerner 1958). Another well-documented example is the selection in turkeys during the period 1944–1964, in which selection for growth rate, body conformation, and body size, also produced a steady decline in some aspects of reproductive fitness such as fertility, egg production, and egg hatchability (Nordskog and Giesbrecht 1964). On the other hand, correlated responses can sometimes be useful. For example, selection for larger mature body size often increases litter size in mice and swine. If a trait has a low heritability or is difficult to measure, it is sometimes possible to practice selection for another, correlated trait, obtaining progress in the trait of interest by correlated response. Theoretically, the maximum response to artificial selection occurs when the criterion for selection is determined by a **selection index** (a weighted average value across several traits), which takes genetic correlations into account. However, the theoretical advantage of index selection is often overridden by practical difficulties in estimating the components of the index and implementing the selection procedure.

From a theoretical point of view, the covariance between two quantitative traits can be partitioned in a manner analogous to the partitioning of the variance for one trait outlined earlier. The covariance can therefore be partitioned into an additive covariance, a dominance covariance, an environmental covariance, and so on. The most important theoretical result is that the magnitude of correlated response with individual selection depends only on the additive covariance, much as the direct response to individual selection depends only on the additive variance. The components of covariance between traits can be estimated from the resemblance between relatives, but often it is preferable to estimate the correlated response by direct observation in a manner analogous to the determination of realized heritability (Falconer and Mackay 1996).

The phenotypic correlation is the correlation one would obtain by measuring two traits, say X and Y, and calculating the correlation coefficient directly. In symbolic terms, the phenotypic correlation is

$$r_p = \frac{Cov_p(X,Y)}{\sqrt{\sigma^2_{pX}\sigma^2_{pY}}} \tag{8.41}$$

where $Cov_p(X,Y)$ is the phenotypic covariance and σ^2_{pX} and σ^2_{pY} are the phenotypic variances of traits X and Y. Although correlations do not partition in exactly the same way as variances, the phenotypic covariance can be expressed as the sum

$$Cov_p(X,Y) = Cov_a(X,Y) + Cov_e(X,Y) \tag{8.42}$$

where $Cov_a(X, Y)$ is the **additive genetic covariance**, and $Cov_e(X,Y)$ is the **environmental covariance**. The **genetic correlation** (which is essentially the correlation of additive genetic effects) is defined as

$$r_a = \frac{Cov_a(X,Y)}{\sqrt{\sigma^2_{aX}\sigma^2_{aY}}} \tag{8.43}$$

where $Cov_a(X, Y)$ is the additive genetic covariance, and σ^2_{aX} and σ^2_{aY} are the additive genetic variances of the two traits. The genetic covariance is estimated in much the same manner as the additive genetic variance. For example, the additive genetic covariance of two traits X and Y among half sibs equals $(\frac{1}{4})Cov_a(X, Y)$.

Correlated response (CR) occurs when characters other than directly selected traits respond. The magnitude of correlated responses to selection is related to the genetic correlation between the selected and correlated characters. The expected response is expressed by the equation

$$CR_y = i\sqrt{h^2_X}\sqrt{h^2_Y}\ r_a\sigma_{pY} \tag{8.44a}$$

$$= \frac{Cov_a(X,Y)}{\sigma^2_{pX}} S_X \tag{8.44b}$$

where X is the directly selected character with narrow-sense heritability h^2_X, i is the intensity of selection on character X, the correlated character is Y with heritability h^2_Y, r_a is the genetic correlation between X and Y, and σ_{pX} and σ_{pY} are the phenotypic standard deviations of the characters X and Y, respectively (Falconer and Mackay 1996; Kathuria et al. 1996). The **intensity of selection** (i) is defined as the selection differential expressed as a multiple of the phenotypic standard deviation, hence $i = S_X/\sigma_{pX}$, where S_X is the selection differential on character X. As before, $Cov_a(X,Y)$ is the additive genetic covariance between characters X and Y. That Equation 8.44a implies Equation 8.44b can be shown by substituting S_X/σ_{pX} for i and using the definitions of genetic correlation and heritability. Equation 8.44b is informative because it shows the correlated-response analog of the prediction equation $R = h^2S$. The implication of the expressions in Equation 8.44 is that, all else being equal, a doubling in the genetic correlation (that is, a doubling of the additive genetic covariance) will double the magnitude of the correlated response.

PROBLEM 8.8 A herd of dairy cattle yields milk whose distribution of fat content has a mean of 3.4% and a standard deviation of 0.65%, and whose distribution of protein content has a mean of 3.3% and a standard deviation of 0.45%. The narrow-sense heritabilities of these traits are 0.60 and 0.70, respectively, and the genetic correlation is 0.55. If selection is practiced for percent protein with a selection intensity of $i = 1.5$, what increase in percent protein and percent fat would be expected? What intensity of selection would produce the same increase in percent fat by direct selection?

ANSWER Because $i = S_X/\sigma_{pX}$, Equation 8.10 can be written as $R_X = ih_X^2\sigma_{pX}$ where i is the intensity of selection. For percent protein, $R_X = (1.5)(0.7)(0.45) = 0.47$, so the expected percent protein equals 3.3% + 0.47% = 3.8%. For a correlated response, use Equation 8.44a to get $CR = (1.5) \times \sqrt{(0.6)}\ \sqrt{(0.70)} \times (0.55)(0.65) = 0.35$, so the expected percent fat equals 3.4% + 0.35% = 3.75%. The fat increase corresponds to a direct selection for fat content of $i = 0.35/(0.60)(0.65) = 0.90$.

The absence of correlated response can sometimes be very informative. Artificial selection on spacing between vein junctions in the *Drosophila* wing might be expected to result in correlated changes all over the wing. When Weber (1992) did such an artificial selection for a tiny (100 cells) region of the wing, he obtained a strong direct response with no significant correlated response in other wing measures. This suggests that the control of morphological development in the wing involves many genes, and that independent selection for minute aspects of wing morphology are possible.

As mentioned above, the two primary mechanisms that produce genetic correlation are pleiotropy and linkage. With pleiotropy, genes that affect one character also affect others. These effects may be direct, as when a gene product has two distinct functions, or they may be indirect in the sense that more than one physiological step may connect the two phenotypes. In both cases, replacement of an allele at the relevant gene will affect both phenotypes. Alternatively, linkage disequilibrium can result in genetic correlation even though the genes affecting the two traits are entirely distinct. If there is linkage disequilibrium, then selection of one gene will affect allele frequencies of nearby genes, and the end result is a correlation of changes in two phenotypes. The effects of pleiotropy versus linkage disequilibrium may be distinguished experimentally by attempting to control or eliminate linkage disequilibrium. By following molecular markers such as single nucleotide polymorphisms in crosses or selection experiments, it is much easier to distinguish these two causes of genetic correlation, as described later.

8.8 EVOLUTIONARY QUANTITATIVE GENETICS

Evolutionary quantitative genetics is an application of population genetics to detect natural selection, random genetic drift, and other genetic processes in natural populations, and to explain the patterns of genetic variation that are observed within and among populations. This is a challenging field both experimentally and theoretically, and it is an area of very active research. The theory deals with the phenotypic and genetic variances and covariances among a set of correlated traits, and changes in the frequencies of the underlying alleles are used to infer changes in these variances and covariances. The genetic models are intrinsically multilocus models; they can be very complex except when simplifying assumptions are made. In this section some of the basic principles and conclusions of evolutionary quantitative genetics will be considered.

Inference of Selection from Phenotypic Data

The results of models for selection on quantitative traits can be used to infer the operation of natural selection, but beyond considering whether the assumptions of the models are valid, some serious statistical problems must also be overcome. Either by sampling a population at two different times, or by doing a cross-sectional study, changes in the phenotypic distribution resulting from differential mortality of the different phenotypes can be detected (Lande and Arnold 1983; Arnold and Wade 1984; Shine et al. 2001; Gimenez et al. 2006). Changes in phenotypic distributions reflect the operation of natural selection, despite the lack of any proven changes in allele frequency. Human birth weight is easily shown to be a character that is under the influence of natural selection, because the mortality rate of very light and very heavy babies is higher than the mortality of babies whose weight is close to the population average (Karn and Penrose 1951). Individuals at the extremes of body weight also have the highest mortality in elderly humans (Harris et al. 1988). Phenotypic selection will result in changes in gene frequency to the extent that the selected phenotypes are heritable, but selection on traits with low heritability will have minimal immediate effects on the genetic composition of the population.

The correspondence between phenotypes and relative fitness can be estimated from the mortality that takes place from one population census to another. This approach was pioneered by the naturalist Hermon C. Bumpus (1899), who studied 136 house sparrows (*Passer domesticus*) that had been incapacitated by a severe winter storm in Rhode Island. Eight measurements were taken of every bird, and about half of the birds subsequently revived. The phenotypic measures and the fraction of birds that survived provide a way to examine the relation between the phenotypic measures and the fitness.

The Bumpus data have been analyzed and reanalyzed at intervals since 1911 (Pugesek and Tomer 1996). A regression analysis carried out by Lande and Arnold (1983) revealed that birds of both sexes with smaller weight, and males with longer total body length, were significantly more likely to survive. Pugesek and Tomer (1996) examined seven morphological measurements (length of each of four long bones plus length of wing, head, and skull) using a method called *structural equation modeling* in which the measurements are combined into a set of factors that have the maximum explanatory power with respect to mortality. They found that a general size factor composed of a weighted average of all seven measurements, as well as wing length (independent of general size), were significantly correlated with survival. Putting these results together with those of Lande and Arnold (1983), one would infer that long, skinny birds survived the snow and ice storm better than short, plump birds.

Similar methods have been used to study a species of Darwin's finches, *Geospiza fortis*, on the Galápagos Island of Daphne Major (Gibbs and Grant 1987), although with fewer measurements of body size independent of weight. During dry years, birds with a larger body size are more likely to survive than others, evidently because the most abundant food consists of large, hard seeds. After a prolonged disruption in Pacific Ocean currents known as El Niño, there occurred a year with 10 times the normal rainfall, and by the time the rains stopped both the seed availability and the density of young birds was very high. Under these conditions a reversal in the direction of selection took place, favoring birds with smaller body weight and, in one year, increased wing length.

The Bumpus method for inferring phenotypic selection has faced criticism on a number of grounds. Mitchell-Olds and Shaw (1987) pointed out that, as in other applications of multiple regression, the traits can be correlated among themselves, and in some cases the estimators may not be *consistent* (that is, they may not converge to true values with increasing sample size). Lande and Arnold (1983) discussed the additional problem of a strongly selected character that is not among those studied. The limitations of regression analysis have encouraged the development of other approaches for analyzing such data (Pugesek and Tomer 1996; Gimenez et al. 2006), but the strongest evidence for selection is confirmation by direct experimental manipulation whenever feasible (Mitchell-Olds and Shaw 1987).

Evolution of Multiple Intercorrelated Traits

The fitness of an organism is a composite of many traits that are correlated with fitness as well as with one another, and these traits are influenced by both genetic and environmental factors. A theoretical analysis of such a complex situation might at first seem hopeless, but in fact there is a relatively simple analog of the prediction equation $R = h^2S$ that can be applied. The multivariate analog of the prediction equation asserts that

$$\Delta\bar{z} = G\beta \tag{8.45}$$

where **z** is a vector of the mean values $z_1, z_2, \ldots, z_n$ of a set of n traits, and $\Delta\bar{z}$ is the expected change in trait means in one generation of selection (Lande 1979; Turelli 1988). The G represents a square matrix called the **genetic variance-covariance matrix** (or simply the G **matrix**), whose diagonal elements are the additive genetic variances of each of the n traits (see Equation 8.28) and each of whose off-diagonal elements is the additive genetic covariance between the corresponding pair of traits (see Equation 8.42). The symbol β represents a vector called the **selection gradient**, each of whose elements is the regression coefficient of relative fitness on the phenotypic value of one of the traits. Recalling that $h^2 = \sigma_a^2/\sigma_p^2$ (see Equation 8.36), we can write $R = h^2S$ as $R = (\sigma_a^2/\sigma_p^2)S$. By analogy with this formulation of the prediction equation, the selection gradient β plays the role of S/σ_p^2, and G plays the role of σ_a^2. A major difference, however, is that here we are not talking about artificial selection, but rather about changes in a quantitative trait due to natural selection that occur as an indirect effect of the trait's relation to fitness.

If the G matrix remains constant, the Equation 8.45 implies that

$$\bar{z}_t - \bar{z}_0 = G\sum_0^t \beta_t \tag{8.46}$$

where the summation on the right-hand side is the **net selection gradient** and is analogous to the cumulative selection differential in Equation 8.29 divided by σ_p^2. We emphasize again, however, that the changes in phenotype described in Equation 8.46 result from natural selection acting on fitness and not artificial selection acting on the traits themselves.

Given a difference in mean phenotypes for a set of traits in two species, and a constant genetic variance-covariance matrix G, Equation 8.46 allows the net selection gradient to be estimated (Price and Grant 1985; Lofsvold 1986; Begin and Roff 2003; Game and Caley 2006; McGuigan 2006). The problem is that G may not remain constant through time (Turelli 1988). Without direct information on the stability of genetic variance-covariance matrix G, it is difficult to assess how much confidence there should be estimates of the cumulative selection gradient. Changes in G have been documented in artificial selection experiments with *Drosophila* (Wilkinson et al. 1990; Shaw et al. 1995), among inbred strains of mice (De Brito et al. 2005), and in natural subpopulations of the frog *Rana temporaria* (Cano et al. 2004). Theoretical studies indicate that random drift as well as selection can have a significant affect on G (Whitlock et al. 2002; Jones et al. 2003), but computer simulations suggest that some suites of intercorrelated traits may have much more stable G matrixes than others (Jones et al. 2003). On the other hand, the key issue may not be whether G changes, but in what manner it changes and how rapidly (Steppan et al. 2002).

Random Genetic Drift and Phenotypic Evolution

Genetic variation in a finite population is subject to the vagaries of random genetic drift, and if the genetic variation affects a quantitative character, then changes in allele frequency due to random drift will alter the genetic variance of the character. In a finite population of effective size N_e, the genetic variance changes over successive generations because the heterozygosity changes as the allele frequencies change. To make this quantitative, recall that Equation 3.14 says that the expected heterozygosity at each locus ($H = 2pq$) decreases by the fraction $1/(2N_e)$ in each generation, and Equation 8.28 shows that the additive genetic variance at each locus σ_a^2 is proportional to the heterozygosity ($2pq$). Since fixation or loss is the ultimate fate of every allele in a finite population, the additive genetic variance would ultimately go to 0 except for the effect of new mutations. Let us suppose that the increment in additive genetic variance added in each generation due to new mutation is σ_m^2. Then we can write

$$E[\sigma_a^2(t)] = \left(1 - \frac{1}{2N_e}\right) E[\sigma_a^2(t-1)] + \sigma_m^2 \tag{8.47}$$

where $E[\sigma_a^2(t)]$ is the expected value of the additive genetic variance in generation t (Clayton and Robertson 1955, 1957; Lande 1979, 1980; Turelli et al. 1988). When the loss of additive genetic variance due to random genetic drift in each generation is exactly offset by the gain of additive genetic variance due to new mutations, then the population arrives at a mutation-drift equilibrium. The expected genetic variance at the mutation-drift equilibrium is

$$E(\sigma_a^2) = 2N_e\sigma_m^2 \tag{8.48}$$

The expected time for a population to reach mutation-drift equilibrium is $4N_e$ generations, the same as with random genetic drift for the alleles of a single locus. For two populations undergoing random genetic drift independently for t generations, the expected difference between the mean phenotypes is $2t\sigma_m^2$. The implication is that the rate of phenotypic divergence due to selectively neutral alleles depends only on the rate of neutral mutations affecting the phenotype (Lynch and Hill 1986), just as in the neutral theory the rate of gene substitution is independent of population size and depends only on the rate of neutral mutations. Lynch and Hill (1986) also showed that the rate of divergence is dependent only on the rate of purely additive mutations and is independent of the effects of dominance and epistasis.

Figure 8.25 shows the increase in within-population genetic variance among initially homogeneous populations, until the time at which the populations reach a steady-state balance of mutation and drift. Simulations also verify the nearly linear increase in variance between populations with time. Extending neutral divergence to the case of multiple characters, Lynch and

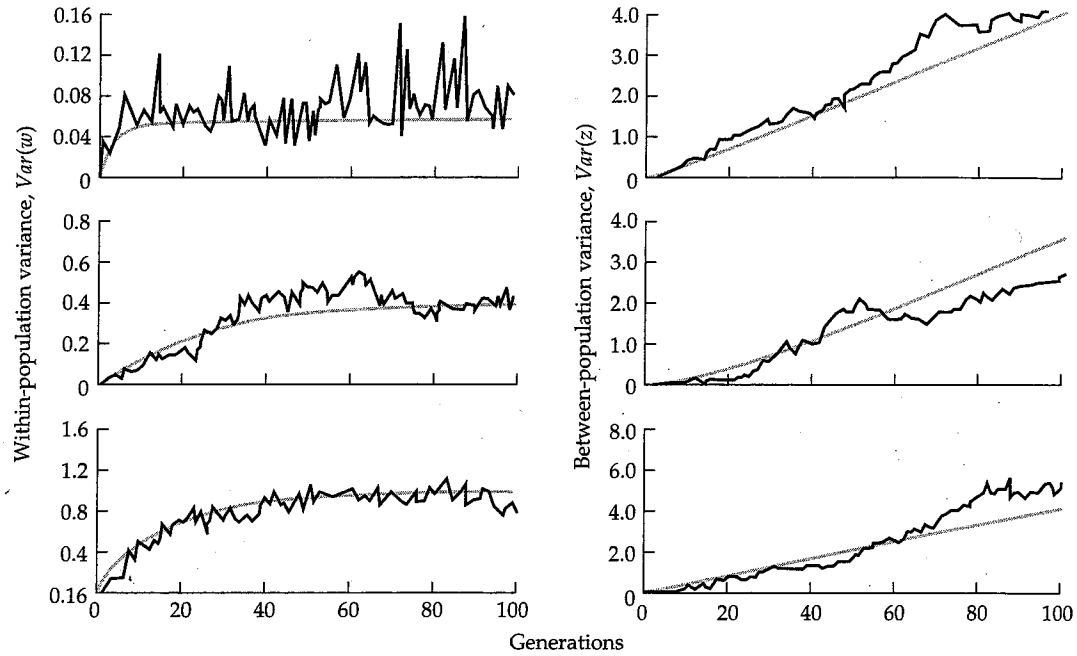


FIGURE 8.25 Simulations of a model with mutation of genes underlying a quantitative character and random genetic drift in a finite, subdivided population. The jagged lines represent the simulations, whereas the smooth curves are based on an analytical model. One hundred populations were run in each sample for each case. A mutation rate of 0.001 per locus per generation was used in all runs. (Top) Effective population size $N_e = 2$, number of loci $n = 50$. (Middle) $N_e = 10$ with $n = 10$ loci. (Bottom) $N_e = 10$ with $n = 50$ loci. The steady-state within-population variance increases with population size, but the rate of increase of variance between populations is twice the mutational variance.

Hill (1986) showed that the neutral divergence of the variance-covariance matrix depends only on the mutational variance-covariance matrix.

Turelli et al. (1988) used the mutation-drift equilibrium as a null hypothesis for devising a statistical test. Evolution has been too rapid to be explained by the neutral model (with 95% confidence) if

$$\frac{\sigma_m^2}{\sigma_p^2} < \frac{(\Delta z / \sigma_p)^2}{2t(1.96)^2} \qquad (8.49)$$

where σ_m^2 is the mutational variance, σ_p^2 is the phenotypic variance, and Δz is the change in the mean phenotype of a particular trait in a time period of t generations. In other words, the mutation-drift equilibrium places a lower bound on the ratio of mutational variance to the total phenotypic variance.

If the observed σ_m^2/σ_p^2 is smaller than the critical value given in Equation 8.49, then mutation does not introduce enough variance to account for the observed divergence. The test based on Equation 8.49 is useful only on a time scale sufficiently long that the diverged populations are likely to have reached a mutation-drift equilibrium.

Mutational Variance and Mutation-Accumulation Experiments

The magnitude of the additive genetic variance in a quantitative trait that is introduced by mutation each generation, σ_m^2, is a quantity of considerable interest and evolutionary importance. The mutational variance can be estimated by several methods, including estimates of the increase in additive variance among initially identical subpopulations as mutations accumulate over generations, or by the response to selection in a population that is initially in mutation-drift equilibrium or that lacks additive genetic variation (Lynch 1994). It might be supposed that an artificial selection experiment in a population lacking genetic variation would be doomed to failure, but as discussed in Section 8.3, populations can accumulate sufficient mutational variation to produce a response in a few generations (Fry et al. 1995; Lynch and Walsh 1998; Mackay 2001; Barton and Keightley 2002). In both methods of estimation, the action of natural selection will bias the estimates of mutational variance, so steps must be taken to minimize the effects of natural selection. In mutation accumulation experiments in *Drosophila*, this has traditionally been done with balancer chromosomes that prevent recombination and minimize selection by maintaining chromosomes in a heterozygous state, or by sib mating. For a wide range of traits in a variety of organisms, experimental estimates of σ_m^2/σ_e^2 generally fall in the range of 10^{-2} to 10^{-3} (Houle et al. 1996; Lynch and Walsh 1998). These estimates imply that, in the absence of selection or random genetic drift, lost heritability would be restored by mutation in approximately 100–1000 generations. Interestingly, the ratio of the genetic variance to the mutational variance, σ_g^2/σ_m^2, is expected to equal the median persistence time of mutant alleles (Houle et al. 1996). The estimated persistence times for mutations affecting life history traits averaged around 50 generations, whereas those for mutations affecting morphological traits were around 100 generations (Houle et al. 1996).

Estimates of the mutational variance allow Equation 8.49 to be used in practice. One interesting application is to genome-wide levels of gene expression. One study compared gene expression levels across a range of time scales, from pairs of closely related strains of laboratory mice to pairs of species in *Drosophila*, mouse, and higher primates (Lemos et al. 2005). The main finding was that differences in the levels of gene expression were far smaller than predicted from Equation 8.49, suggesting that the level of expression of most genes is subject to stabilizing selection. There were, however, significant differences among different functional classes of genes and among genes associated with reproductive processes.

The rate of occurrence of new mutations and the distribution of their homozygous effects, degree of dominance, and interactions are fundamental not only to understanding the maintenance of genetic variation but also the response to long-term selection, the decrease in average population fitness due to mutation, and even the evolution of sex. Hence a great deal of effort has gone into estimating the rate and properties of new mutations, especially those affecting quantitative traits and fitness. The original idea of studying mutations accumulated over a large number of generations in a set of originally identical and genetically homogeneous lines is due to Muller (1928), but it was first put into experimental practice by Mukai (1964), whose experiments on estimating the rate and effects of deleterious mutations in *Drosophila* were very influential. One of the key parameters to be estimated is often designated U, the genome-wide rate of mutation to deleterious alleles. In his early experiments, Mukai and collaborators had found a decrease of about 1% per generation in the average homozygous viability of nonlethal chromosomes, and estimated U as 0.35–0.47 per haploid genome per generation for nonlethal mutations affecting viability. This value is far greater than the genomic mutation rate of 0.01 estimated for recessive lethals (Crow and Simmons 1983), and it suggested that a great deal of natural selection must take place merely to eliminate the new deleterious mutations that occur in each generation. Furthermore, the Haldane-Muller principle (Chapter 5) implies that the equilibrium average viability of an outbred population should be approximately e^{-2U}, which for Mukai's estimates equals 0.39–0.50. This is much lower than observed for populations of *Drosophila*.

Mutation-accumulation experiments performed later yielded far smaller estimates of U for nonlethal mutations, ranging from 0.004 in *Caenorhabditis elegans* to 0.05 in *D. melanogaster* (Keightley and Eyre-Walker 1999). The lowest estimate is $U = 0.00017$ in *Escherichia coli* (Kibota and Lynch 1996) and the highest is in mammals with U approaching if not greater than 1 (Kondrashov 2001). There is obviously great deal of variation from one organism to the next. What to make of Mukai's earlier estimates is still not clear, but it has been suggested that they were unduly affected by the mutational effects of transposable elements that may have been mobilized in the lines he studied (Keightley and Eyre-Walker 1999).

Mutation-Selection Balance for Quantitative Traits

As emphasized at the beginning of this chapter, genetic variation for quantitative traits is ubiquitous in natural populations, and directional selection in an outbreeding population can change the mean of almost any trait. The phenotypic variation maintained for many metric traits is thought to reflect the action of stabilizing selection, which favors individuals in the middle of the range over those at the extremes. Stabilizing selection, like directional selection, tends to reduce genetic variation over time, because equal numbers of plus and minus alleles become fixed. Yet the heritabilities of many such traits

are in the range 25–50%. How such levels of genetic variation are maintained is an open question. One possibility is that selection in nature is weaker then generally supposed. Although selection in natural populations has been documented for many metric traits in many species (Endler 1986), a review of 63 published studies reporting 2500 estimates revealed that the intensity of selection was actually rather weak, with a median selection gradient in the range 0.1–0.2 (Kingsolver et al. 2001).

But even with relatively strong stabilizing selection, it seems intuitively reasonable to suppose that the observed levels of additive genetic variation could result from a balance between stabilizing selection, which tends to reduce genetic variation, and new mutations, which tend to increase it. Although this hypothesis is deceptively simple when stated verbally, mathematical versions of this hypothesis are complex and include many parameters whose values are unknown or only crudely estimated (Lande 1975; Turelli 1984; Keightley and Hill 1988, 1990; Barton 1990; Kondrashov and Turelli 1992; Caballero and Keightley 1994; Bürger 2000; Zhang and Hill 2002, 2003; X.-S. Zhang et al. 2002, 2004). The models are complex because they have to deal with complications such as the number of genes, the type of action of mutant alleles and their interactions, the type and intensity of selection, linkage between loci, and the influence of selection through pleiotropic effects on correlated traits. Pleiotropy is reflected in the widespread tendency of mutant genes to affect several traits simultaneously (Estes et al. 2005). The effect usually results from the fact that complex phenotypic traits are determined by the interactions of the products of many genes during development. The number of genes affecting a trait is relevant to mutation-selection balance because, for a given genetic variance, the selection intensity per locus decreases as the number of loci increases. If the total mutation rate is fixed, the per-locus mutation rate must decrease as the number of loci increases. The challenge is to develop a model that predicts a substantial additive genetic variance at equilibrium in a population in which stabilizing selection is balanced against mutation.

Such models have been surprisingly difficult to devise (X.-S. Zhang et al. 2004). Part of the problem is that the situation has two levels of selection: stabilizing selection that acts on mutations that affect only the trait itself, and selection that acts on pleiotropic mutations that affect both the trait and fitness. Among the latter, both the distribution of main effects and the degrees of dominance may be different between fitness and the metric trait. Mutations affecting fitness have a broad range of effects, but those that are most deleterious are also nearly recessive. If we write the fitnesses of the genotypes AA, AA' and $A'A'$ as $1 : 1 - hs : 1 - s$, for example, then for recessive lethals ($s = 1$) the degree of dominance h may average about $h = 0.01$. In regard to effects on a metric trait, however, the effects are more sharply peaked and have a fat tail. Such a distribution is said to **leptokurtic**. Figure 8.26 shows the approximate distribution of effects of single transposable *P*-

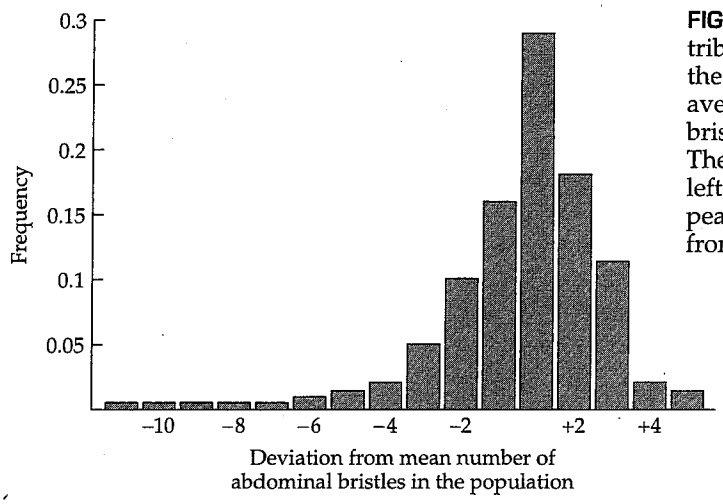


FIGURE 8.26 Approximate distribution of effects of insertions of the transposable *P*-element on average number of abdominal bristles in *Drosophila melanogaster*. The distribution is skewed to the left and leptokurtic (narrowly peaked and with fat tails). (Data from Lyman et al. 1996.)

element insertions on the number of abdominal bristles in *Drosophila* (Mackay et al. 1992). The distribution is strongly skewed to the left, meaning that mutations with large effects on bristle number tend to reduce the number of bristles, and it is leptokurtic, which means that the observations are more clustered near the mode than in a normal distribution, and the tails are fatter. Just as the variance of a distribution depends on deviations from the mean raised to the second power, the kurtosis of a distribution depends on the deviations from the mean raised to the fourth power. Leptokurtosis is an issue because the strength of stabilizing selection depends on the variance of the squared deviations of a metric trait from its optimum, and hence involves the fourth power of the individual deviations. Moreover, pleiotropic effects on a metric trait are expected to show a greater degree of dominance that the dominance of the effect on fitness. If we let genotypes AA, AA' and $A'A'$ have values for a metric trait of $m + a : m + d : m - a$ (as in Equation 8.12), then the dominance can be measured as the absolute value of d/a. For *P*-element insertions affecting abdominal bristles in *Drosophila*, the mean value of $|d/a|$ was 0.75 (Mackay et al. 1992).

These considerations about direct and pleiotropic selection have been combined into a model for selection-mutation balance on a metric trait subject to strong stabilizing selection (X.-S. Zhang et al. 2004). The model is complex, but simulations imply that when mutations are more recessive for fitness than for the metric trait, and when pleiotropic effects on the metric trait are leptokurtic, then high levels of genetic variation in the quantitative trait can be maintained even in the face of strong apparent stabilizing selection. Most of the genetic variation comes from alleles that are nearly neutral for fitness in heterozygous genotypes, whereas most of the apparent stabiliz-

ing selection comes from alleles that have large effects on the quantitative trait. The degree of leptokurtosis that is needed in the model decreases with the strength of apparent stabilizing selection.

To see how the pleiotropy side of the model works, consider a deleterious allele with a fitness of $1 - hs$ in heterozygous genotypes that has a pleiotropic effect on the metric trait that is additive ($d = 0$). Then at mutation-selection balance, the equilibrium allele frequency of the mutant allele is approximately $q = \mu / hs$ (see Chapter 5), and the mutant allele contributes an amount of additive genetic variance to the quantitative trait equal to $2pqa^2$ (Equation 8.28), which equals approximately $(2\mu / hs)a^2$. Hence, a deleterious allele with $hs = 0.01$ contributes ten times more to the additive genetic variance in the metric trait than one with $hs = 0.1$, because the equilibrium frequency of the former is ten times larger. Taking this a step further, if there were n deleterious alleles affecting the metric trait, each with $hs = 0.01$, then the equilibrium additive genetic variance for the metric trait would be $\sigma_a^2 = (2n\mu / 0.01)a^2$, or $s_a^2 = a^2$ when $n\mu = 0.005$. For *P*-element insertions affecting abdominal bristles, the average value of a^2 equals $0.30\sigma_p^2$ (Lyman et al. 1996) and so for these values of the parameters the narrow-sense heritability for abdominal bristles would be $\sigma_a^2/\sigma_p^2 = 0.30\sigma_p^2/\sigma_p^2 = 30\%$. This example shows that mutation-selection balance can account for high heritability in a quantitative trait, but consistency with observed values does not prove that the model is necessarily correct. Fortunately, the model makes testable predictions about the pattern of change of the additive variance during inbreeding, as well as about the duration and repeatability of response to selection (X.-S. Zhang et al. 2004).

8.9 GENES THAT AFFECT QUANTITATIVE TRAITS

In this section we consider various aspects of the synthesis of quantitative genetics with genomics. The knowledge and information provided by genomics has transformed the field of quantitative genetics and given hope of identifying the genetic factors underlying phenotypic variation. Genomic sequences that are annotated with gene locations and gene functions can supply a set of **candidate genes** whose functions suggest that alleles of these genes might contribute to genetic variation. In *Drosophila*, for example, genes that function in the development of abdominal bristles are obvious candidates for affecting variation in bristle number in natural populations (Mackay 2001). Each candidate gene must be tested individually, because many do not contribute to variation, and some that do contribute do so to only a small extent. One limitation of candidate genes is that in most genomic sequences only about half the genes have putative functions, and many of the genes that are assigned putative functions may also participate in other pathways or processes.

Genomics has also allowed the identification of genomic-wide molecular polymorphisms such as single nucleotide polymorphisms (SNPs) or

microsatellites. These molecular polymorphisms serve as genetic markers in crosses, and allow the genetic mapping of loci whose alleles affect a quantitative trait. A gene identified in this way is known as a **quantitative trait locus** or **QTL**. If a QTL affects gene expression, it is known as an *expression QTL* (eQTL), and if the effect of a QTL can be ascribed to a single polymorphic nucleotide site, then the site is called a *quantitative trait nucleotide* (QTN).

The Number of Genes Affecting Quantitative Traits

As discussed in connection with Figure 8.4 showing selection for oil content in maize, the number of genes affecting a quantitative trait is a key parameter in the total response to selection and in the extent to which new mutations contribute to the response. Traditional methods for estimating the number of genes contributing to phenotypic variation make unrealistic assumptions, but newer methods based on genetic mapping with SNPs seem very promising.

A traditional estimate of the number of genes, first suggested by Wright (in Castle 1921), is

$$n = \frac{D^2}{8\sigma_g^2} \tag{8.50}$$

where D is the difference in phenotype between two inbred lines in which n plus alleles are fixed in one line and n minus alleles in the other, and σ_g^2 is the genetic variance estimated by, for example, subtracting the phenotypic variance of the F_1 generation of a cross between the inbred lines (σ_e^2) from the phenotypic variance of the F_2 generation ($\sigma_g^2 + \sigma_e^2$). Equation 8.50 assumes that all genes have equal effects on the trait, that there is no dominance, and that all the genes are unlinked When these assumptions are violated (which they usually are), application of the equation results in estimates of gene number that are smaller than the actual numbers, and sometimes much smaller (Lande 1981; Zeng 1992).

The rationale for Equation 8.50 is straightforward. With additive genes, the total genetic variance equals the additive genetic variance, and because in the F_2 generation of a cross between inbred lines, $p = q = \frac{1}{2}$, then $\sigma_g^2 = \sigma_a^2 = na^2/2$ for n genes (see Equation 8.28), where for each gene the phenotypic values of AA, AA' and $A'A'$ are $m + a$, m, and $m - a$ (see Equation 8.12). The difference in phenotype between the inbred lines D equals $2na$, and so $D^2 = 4n^2a^2$. Hence $D^2/(8\sigma_g^2) = n$, the number of additive, unlinked genes.

The main problem with Equation 8.50 is that it gives misleadingly low estimates of n. In the first place, it is unrealistic to suppose that the genes have equal effects. It is much more plausible to assume that the effects are exponentially distributed, with many small values and fewer large ones. (The exponential distribution is positively skewed and leptokurtic.) Figure 8.27 shows the result of simulations when the gene effects are distributed exponentially. Each point is based on 100 replicates for each given number of

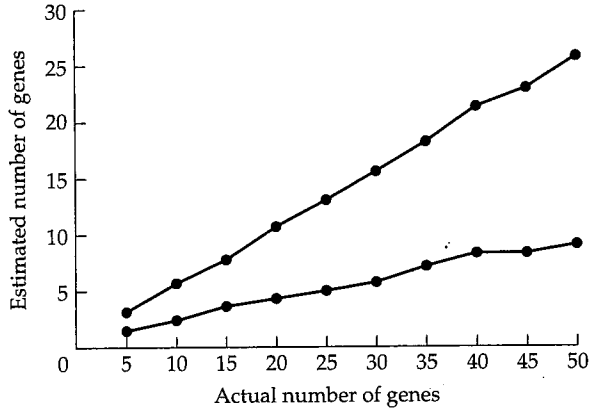


FIGURE 8.27 Estimates of gene number based on the equation $n = D^2/(8\sigma_g^2)$ when the distribution of allele effects is not equal, as assumed in the model, but rather exponential. The simulations in the top curve assume that, for all the genes in which the original inbred lines differ, one of the inbred lines is fixed for all the plus alleles and the other fixed for all the minus alleles. The simulations in the bottom curve assume that each inbred line has 20% of its fixed alleles having the opposite sign from the others.

genes, and the estimated number is that given by Equation 8.50. The estimated numbers are only about half the true values. Dominance and linkage would make matters worse. Furthermore, it is unrealistic to assume that all plus alleles are fixed in one inbred line and all minus alleles in the other. Random genetic drift and hitchhiking due to linkage will inevitably fix some minus alleles in the plus line and some plus alleles in the minus line. In the maize experiment in Figure 8.4, for example, about 20% of the alleles identified by SNP mapping had effects of the opposite sign (Laurie et al. 2004). These alleles reduce the value of D in Equation 8.50, and so cause an even greater underestimate of gene number. In Figure 8.27, the bottom line shows results of the simulations when 20% of the alleles in each of the inbred lines have the wrong sign.

PROBLEM 8.9 At generation 40, the "up" and "down" maize lines in Figure 8.4 were inbred and then crossed to produce an F_2 generation. For oil content, a transformation of scale that makes the allele effects approximately additive is log[(percent oil in kernels) + 1.87] (Lande 1981). With this scale of measurement, the means of two inbred lines were found to be 1.122 and 0.513, corresponding to a percent oil content of 11.5% and 1.4%, respectively. On the transformed scale, the phenotypic variances of the F_1 and F_2 generations were 0.00030 and 0.00303, respectively. Use Equation 8.50 to estimate the number of genes.

ANSWER In this case the relevant variances are $\sigma_e^2 = 0.00030$ and $\sigma_g^2 + \sigma_e^2 = 0.00303$, hence $\sigma_g^2 = 0.00303 - 0.00030 = 0.00273$. Also $D = 1.122 - 0.513 = 0.609$, and so $D^2 = 0.37088$. Therefore $n = 0.37088/(8 \times 0.00273) = 17$.

The estimate of 17 genes in Problem 8.9 seems unreasonably small. Starting with selected lines in generation 70, Laurie et al. (2004) carried out genetic mapping with 440 SNPs and identified the locations of about 50 QTLs accounting for about 50% of the genetic variation. Since genetic mapping identifies the QTLs with the largest effects and often misses those with small effects, the number of genes could be substantially larger than 50. If we assume that the gene effects are exponentially distributed and that in generation 40 about 20% of the alleles that were fixed in each inbred line had a sign opposite of that expected, then simulations like those in Figure 8.27 suggest that the true number of segregating genes affecting the trait in generation 40 was probably about 90.

The idea that quantitative genetic variation results from the segregation of alleles of a large number of genes, many with very small effects, is known as the **infinitesimal model** (Fisher 1918). Taken literally, the infinitesimal model might be cause for pessimism, because if the effects of the alleles are all very small, then QTLs will be difficult to detect. But there are many reasons for optimism even if the infinitesimal model is true. First, the distribution of allelic effects is likely to be exponential, and therefore while many alleles may have small effects, some alleles will have much larger effects and so be detectable as QTLs. For example, in computer simulations of a quantitative trait affected by 100 genes whose effects are exponentially distributed, the 10 genes with the largest effects contribute an average of 88% of the total genetic variation. A second reason for optimism is that the effects of a gene on the genetic variance depend on its allele frequencies, and those with allele frequencies in an intermediate range ($p = 0.2$–0.8) contribute the most, all else being equal. Finally, traits that are affected by many genes have a larger mutational variance than those effected by few genes, and so directional selection, whether natural or artificial, can produce larger differences among populations.

Methods for Mapping QTLs

Interest in genetic mapping of QTLs has a long history, beginning with Thoday's (1961, 1979) studies of bristle traits in *Drosophila*. The essence of the experiments on *Drosophila* was to construct lines that differed with respect to a quantitative character (such as "up" and "down" selected lines for bristle number), and to construct a random set of recombinants among them, using multiply marked chromosomes, each bearing several recessive mutations whose phenotype can be readily scored, to infer which chromosomal regions the recombinants carried. The end result of such experiments is not a genetic map in the classical sense, identifying locations of a series of specific genes that determine the quantitative character. The result is rather a statistical description of the character that indicates what fraction of the genetic variance in the parental lines is contributed by each region identified by the linked genetic markers. When this method is applied on a sufficiently fine

	AS $\frac{1-r}{2}$	AS' $\frac{r}{2}$	$A'S$ $\frac{r}{2}$	$A'S'$ $\frac{1-r}{2}$
AS $\frac{1-r}{2}$	$AA\ SS$ $\left(\frac{1-r}{2}\right)^2$	$AA\ S'S$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$AA'\ SS$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$AA'\ S'S$ $\left(\frac{1-r}{2}\right)^2$
AS' $\frac{r}{2}$	$AA\ S'S$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$AA\ S'S'$ $\left(\frac{r}{2}\right)^2$	$AA'\ S'S$ $\left(\frac{r}{2}\right)^2$	$AA'\ S'S'$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$
$A'S$ $\frac{r}{2}$	$AA'\ SS$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$AA'\ S'S$ $\left(\frac{r}{2}\right)^2$	$A'A'\ SS$ $\left(\frac{r}{2}\right)^2$	$A'A'\ S'S$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$
$A'S'$ $\frac{1-r}{2}$	$AA'\ S'S$ $\left(\frac{1-r}{2}\right)^2$	$AA'\ S'S'$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$A'A'\ S'S$ $\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)$	$A'A'\ S'S'$ $\left(\frac{1-r}{2}\right)^2$

FIGURE 8.28 Expected genotypic composition in the F_2 generation of a cross segregating for a QTL (alleles A and A') and a linked marker gene (alleles S and S'), when the frequency of recombination between the QTL and the genetic marker is r.

scale, it is sometimes possible to identify individual genes that affect a trait, and the application of dense sets of SNPs and other molecular markers has motivated a resurgence of interest in QTL mapping (Tanksley 1993; Doerge 2002; Paterson 2002; Mackay 2004, Slate 2005; Price 2006).

The simplest means for mapping QTLs begins with a cross between two inbred parental lines that differ in phenotype, and we will consider this case in some detail to illustrate the principles. If the parental lines are fixed for alternate alleles at many loci, the F_1 hybrids will be heterozygous at these loci. Intercrossing the F_1 then produces a highly variable F_2 population, both in terms of the underlying genotypes and in the resulting distributions of phenotypes. The F_2 generation is evaluated for the phenotypes of interest and also for a series of molecular markers that differ between the original parental lines.

Figure 8.28 shows the distribution of genotypes expected in the F_2 generation for the alleles A and A' of a QTL affecting the trait of interest and for the alleles S and S' of a single nucleotide polymorphism genetically linked to the QTL with a frequency of recombination of r. As before, we will represent the phenotypes of the AA, AA', and $A'A'$ genotypes by $m + a$, $m + d$, and $m - a$, respectively (see Equation 8.12). Note in Figure 8.28 that the genotype fre-

quencies of AA, AA' and $A'A'$ are $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$, and so the mean phenotype in the F_2 generation equals $m + d/2$. The next steps are made easier if we rewrite the mean phenotype of each genotype as a deviation from the F_2 population mean, and so as deviations the phenotypic values of AA, AA', and $A'A'$ are $a - d/2$, $d/2$, and $-a - d/2$, respectively.

Now we are going to do two types of regressions, one to access the additive effect a of the QTL and the other to access the dominance effect d. The first regression is exactly like that in Figure 8.12, and here we will regress the phenotype of the individuals on the genotype of the SNP polymorphism, coding $S'S'$ as 0, SS' as 1, and $S'S'$ as 2. Since the frequencies of the SNP genotypes are $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$, the mean value of the SNP genotype, call it s, is given by $s = (\frac{1}{4})(0) + (\frac{1}{2})(1) + (\frac{1}{4})(2) = 1$. Again it will be convenient to express the genotypic values of the SNP genotypes as deviations from the mean, and for $S'S'$, SS' and SS these are -1, 0, and $+1$, respectively.

The covariance between the phenotype (p) and the SNP genotype (g) is the mean of the products of the phenotypic value and genotypic value of genotype in Figure 8.28, and hence

$$\begin{aligned} Cov(p,g) &= \left(\frac{1-r}{2}\right)^2 [(a-d/2)(1)+2(d/2)(0)+(-a-d/2)(-1)] \\ &+\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right)[2(a-d/2)(0)+2(d/2)(1)+2(d/2)(-1)+2(-a-d/2)(0)] \\ &+\left(\frac{r}{2}\right)^2 [(a-d/2)(-1)+2(d/2)(0)+(-a-d/2)(1)] \qquad (8.51) \\ &= a(1-2r)/2 \end{aligned}$$

Furthermore, the variance $Var(g)$ of the SNP genotypes equals $(\frac{1}{4})(-1)^2+(\frac{1}{2})(0)+(\frac{1}{4})(1)^2 = \frac{1}{2}$. Since the regression coefficient of p on g is $Cov(p, g)/Var(g)$, then the covariance of phenotype on SNP genotype in the F_2 generation is given by

$$b_{pg} = a(1 - 2r) \qquad (8.52)$$

The association between the phenotype and the SNP genotype will therefore be significant if $a(1 - 2r)$ is large enough, and this regression allows the estimation of a if r is known. Furthermore, the fraction of the total phenotypic variance accounted for by the SNP association is given by $a^2(1 - 2r)^2/(2\sigma_p^2)$.

Information about the degree of dominance affecting the metric trait can be obtained by regressing the phenotype on the heterozygosity. For this purpose, we will code the SNP genotypes $S'S'$, SS', and SS as 0, 1, and 0, and so the mean heterozygosity equals $\frac{1}{2}$. Expressed as deviations from the mean, therefore, the genotypic values of $S'S'$, SS', and SS become $-\frac{1}{2}$, $\frac{1}{2}$, and $-\frac{1}{2}$. The covariance between phenotype p and heterozygosity h can be obtained from

Equation 8.51 by replacing (1) with $(-\frac{1}{2})$, (0) with $(\frac{1}{2})$, and (–1) with $(-\frac{1}{2})$. The result is

$$Cov(p, h) = (\tfrac{1}{4})d(1 - 2r)^2 \tag{8.53}$$

The variance in heterozygosity in the F_2 generation equals $(\frac{1}{4})(\frac{1}{2})^2 + (\frac{1}{2})(\frac{1}{2})^2 + (\frac{1}{4})(-\frac{1}{2})^2 = \frac{1}{4}$, and so the regression coefficient of phenotype p on heterozygosity h is given by

$$b_{ph} = d(1 - 2r)^2 \tag{8.54}$$

Equations 8.52 and 8.54 are of special interest in the case $r = 0$, which means biologically that the QTL and the SNP locus are one and the same gene, or at least are inseparable by recombination. When $r = 0$, the regression coefficient in Equation 8.52 estimates a, whereas the regression coefficient in Equation 8.54 estimates d. This special case brings us full circle to classical quantitative genetics, because a and d were originally defined through conceptual regressions of phenotype on number of favorable QTL alleles or heterozygosity (Fisher 1918; Falconer and Mackay 1996), long before the beginnings of modern molecular genetics. The dense SNP maps available to modern researchers allow those traditional conceptual regressions to be replaced with actual regressions, and enable direct estimates of the parameters of greatest interest.

For any QTL and SNP locus that are associated, the values of a, d, and r could be estimated from the mean phenotypes of the SS, SS', and $S'S'$ genotypes. This approach is somewhat precarious because the three means are used to estimate three parameters, and there is no opportunity for an independent test of goodness of fit. A more favorable strategy is to have each QTL *flanked* by marker SNPs, which is called **interval mapping** (Lander and Botstein 1989). For the flanking markers there are nine genotypes that can be used to estimate a, d, and the frequencies of recombination r_1 and r_2 between the QTL and the flanking SNPs, with estimation and significance testing done by maximum-likelihood methods. For data from natural populations it is also necessary to estimate the allele frequencies of the QTL alleles.

A still more comprehensive approach is **composite interval mapping**, in which multiple regression is carried out on all the marker loci simultaneously (Zeng 1994; Jansen and Stam 1994; Zeng et al. 1999). The analysis is carried out stepwise, identifying first the strongest effect and subtracting this out of the data, then identifying the next strongest effect, and so forth. These methods have been extended to handle various types of traits and data (Doerge 2002; Sen et al. 2005; Wang and Zeng 2006; Li et al. 2006). Bayesian methods for QTL mapping have also been developed (Yi et al. 2005; M. Zhang et al. 2005). These readily incorporate the *a priori* assumption that most intervals will have a negligible effect on the trait, and at least in some cases they produce posterior probabilities of allelic effects that appear to be less biased than QTL mapping methods based on likelihood.

QTL mapping has proven successful in a wide variety of model organisms and in virtually every aspect of agriculture, aquaculture, agronomy, horticulture, and other applied fields of biology (Price 2006). There is. for example, an Animal QTL database that lists all publicly available QTL data on livestock animal species. It includes over 1000 QTLs affecting various traits in pigs and more than 600 QTLs each in cattle and chickens. There is also a Gramene QTL database that includes more than 10,000 QTLs affecting numerous agronomic traits in rice, maize, barley, oats, pearl millet, foxtail millet, and wild rice.

SIGNIFICANCE TESTING OF QTLS A serious problem with QTL mapping methods is how to decide when a likelihood ratio is significant. In the usual way one does a hypothesis test in statistics, one gets a test statistic whose distribution under the null hypothesis one knows. The null hypothesis is rejected if it has a probability less than 5%. This means one expects to reject the null hypothesis 5% of the time, even if it is true. With QTL mapping, one essentially tests the null hypothesis for thousands of potential locations of the QTL, and so hundreds of false positives would result. The problem is made even worse by the fact that we do not know precisely what the null distribution of the likelihood ratio test is, and the many tests that are done are not all independent of one another.

Fortunately there is a way out of this morass. Rather than relying on asymptotic theory to get an expected null distribution, we can build an empirical null distribution for the likelihood ratio by randomly permuting (shuffling) the association of marker genotypes and phenotypes (Churchill and Doerge 1994; Doerge and Churchill 1996). When the observed likelihood ratios are tested against this empirical null, then many of the problems of multiple testing are minimized. The methods for controlling for false positives and limiting the false discovery rate are virtually the same as those discussed in Chapter 2 in connection with multiple tests for Hardy-Weinberg equilibrium.

MARKER-ASSISTED SELECTION One might think that by mapping QTLs, geneticists would immediately learn where the genes are that affect a character and then be able to identify those genes and quickly isolate them. This has not been the case. Most QTL mapping projects have not been followed to the point of identifying the individual genes, in part because the isolation and proof of a QTL effect are laborious. Nevertheless, merely identifying genetic markers linked to QTLs can be of value in plant and animal improvement, because any molecular marker can be genotyped essentially without error, and so be used to aid in selecting on the trait affected by the QTL to which the marker is linked (Dekkers and Hospital 2002; Collard et al. 2005; Varshney et al. 2005).

The use of molecular markers to aid in artificial selection is called **marker-assisted selection**, and it is useful when a molecular marker is tightly linked to the QTL, when the trait of interest is expressed in only one sex (such as milk production or egg production), when the heritability is low, or when the trait is difficult or expensive to assay. As the financial cost of genotyping has decreased, the attractiveness of marker-assisted selection has increased, because large-scale genotyping is far less expensive than cultivating acres of crops or rearing large herds of animals.

To show how marker-assisted selection works we can use the genotypes in the F_2 generation depicted in Figure 8.28. Suppose we were to ignore the phenotype with regard to the quantitative trait and simply select the 25% of the population that is homozygous *SS*. In the resulting population, the genotype frequencies of *AA*, *AA′*, and *A′A′* will be $(1-r)^2$, $2r(1-r)$, and r^2, respectively, and so the allele frequency of *A* in the next generation will be $p' = (1-r)^2 + r(1-r) = 1-r$. In the F_2 generation the allele frequency of *A* was $p = \frac{1}{2}$, and so the change in gene frequency due to marker-assisted selection is $\Delta p = p' - p = (\frac{1}{2}) - r$.

This value can be compared with that in Equation 8.19 for selection based solely on phenotype. When 25% of the population is saved for breeding, then in this example it can be shown that the change in allele frequency equals $\Delta p = \sqrt{h^2}/2$ where h^2 is the narrow-sense heritability of the trait. Therefore, for the type of marker-assisted selection in Figure 8.28, selection based on the marker will produce greater gains than selection based on the phenotype itself if $(\frac{1}{2}) - r > \sqrt{h^2}/2$, or

$$h^2 < (1-2r)^2 \tag{8.55}$$

The relevant value of r is plotted against h^2 in Figure 8.29. For high heritabilities, marker-assisted selection offers little improvement over phenotypic

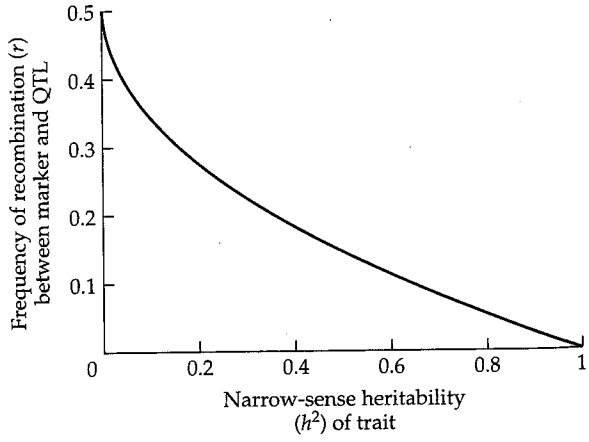


FIGURE 8.29 Marker-assisted selection improves on phenotypic selection for all values of the recombination fraction underneath the curve. The specific values are for an ideal case in which 25% of the F_2 generation is selected. When the narrow-sense heritability is low, even loose linkage is helpful, but when the narrow-sense heritability is high, marker assistance helps only when linkage is very tight.

selection unless the linkage is very tight, but for low heritabilities even fairly loosely linked molecular markers can help substantially.

Note that Figure 8.29 pertains to the special case of selection in the F_2 generation when 25% of the population is selected for breeding. In genetically heterogeneous populations, the efficacy of marker-assisted selection depends not only on the frequency of recombination and the fraction of the phenotypic variance accounted for by the QTL, but also on the allele frequencies of the QTL alleles, those of the marker alleles, and the magnitude of linkage disequilibrium between the QTL and the marker.

SUMMARY

1. Complex traits, often called quantitative traits, are determined by the joint effects of the alleles of multiple genes that interact among themselves and with multiple environmental factors. The traits may be expressed as continuous (metric) traits, categorical traits, or even as dichotomous traits in cases when the multiple genetic and environmental factors determine the underlying liability or risk of the trait.
2. Francis Galton, a pioneer in the application of statistics to biology, invented the concept of regression to describe the tendency for the phenotypes of offspring to be closer to the mean of the population as a whole than to the mean of their own parents.
3. The least-squares regression line of a variable y onto an explanatory variable x has a slope equal to the covariance of x and y divided by the variance of x. In the case of regression of offspring phenotype on the average phenotype of the parents (the midparent value), the regression coefficient equals the narrow-sense heritability h^2.
4. Artificial selection is the deliberate choice of a select group of individuals to be used for breeding. In individual selection, each individual is chosen or rejected according to whether its own individual phenotype exceeds some threshold value. Directional selection often results in large long-term gains, in which the mean of the selected population is many phenotypic standard deviations above the mean of the original population. Long-term progress under directional selection may depend on the input of new mutations affecting the trait.
5. With individual selection, the narrow-sense heritability h^2 predicts the expected deviation of the offspring mean from the population mean as a multiple of the deviation of the parental mean from the population mean. Values of heritability depend on the main (additive) affects of alleles, the dominance effects, the interactions among alleles of different genes, and the allele frequencies. Nevertheless, the heritability of a trait can be estimated by the correlation in phenotype among relatives.
6. Complications in the estimation of genetic and environmental components of variation include nonrandom associations between genotypes and environments (genotype-by-environment association), nonlinear

effects of environment on phenotype (genotype-by-environment interaction), and differential genetic effects in the sexes (genotype-by-sex interaction). When these complications can be ignored, or controlled through a suitable experimental design, then the total phenotypic variance can be expressed as a sum of the genotypic variance and the environmental variance. The genotypic variance, in turn, can be expressed as a sum of the additive genetic variance, the dominance variance, and the epistatic variance. The narrow-sense heritability h^2 is the ratio of the additive genetic variance to the total phenotypic variance, whereas the broad-sense heritability H^2 is the ratio of the genotypic variance to the total phenotypic variance.

7. Analysis of the genetic basis of quantitative traits in natural populations is complicated by the fact that an organism's fitness is affected by many phenotypic traits that are correlated with one another. Nevertheless the direction and magnitude of selection can sometimes be inferred from phenotypic data, and theoretical models for the evolution of multiple interdependent traits have been developed.
8. As in experimental populations, the genetic diversity affecting quantitative traits in natural populations is affected by the offsetting effects of mutation and random genetic drift. Experiments suggest that the input of genetic variance due to new mutations in each generation is usually on the order of 10^{-2} to 10^{-3} times the phenotypic variance, but experimental estimates of genomic mutation rates and the average effects of mutant alleles vary widely from one experiment to the next.
9. Offsetting mutation and selection also contributes to maintaining the genetic variance of quantitative traits in natural populations, but it has been difficult to devise models that account for the typically high narrow-sense heritabilities ($h^2 \approx 25$–50%) of quantitative traits. In one model that can account for high levels of genetic variation, deleterious mutations are assumed to be more recessive for fitness than for the metric trait, and the pleiotropic effects on the metric trait are assumed to be leptokurtic. In this case most of the genetic variation comes from alleles that are nearly neutral for fitness in heterozygous genotypes.
10. Genomic sequencing, dense maps of polymorphic genetic markers such as SNPs, and high-throughput genotyping have greatly stimulated the search for the underlying genes affecting quantitative traits, both through QTL (quantitative traits locus) mapping and studies of candidate genes. Methods for QTL mapping are highly developed for a great variety of types of traits and experimental designs.
11. Marker-assisted selection uses genetic polymorphisms linked to QTLs to improve the efficiency of artificial selection. This procedure is particularly valuable with tight linkage, when the trait of interest is expressed in only one sex, when the heritability is low, or when the trait is difficult or expensive to assay.

PROBLEMS

1. Samples from a distribution are said to be *standardized* if each value is expressed as a deviation from the mean in multiples of the standard deviation. In symbols, the standardized value of a sample observation x is given by $x' = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the distribution from which the sample was drawn. Usually μ and σ are not known but are estimated as the mean and standard deviation in the sample. Shown below are six values from a much larger sample in which the overall sample mean and standard deviation are 100 and 15, respectively. What is the standardized value of each of the six samples? Which two are the most deviant?

Sample values (x)	78	121	141	116	89	118

   Standardized samples from a normal distribution are often called *normal deviates*. What proportion of normal deviates are expected to be negative and what proportion are expected to be positive? What are the expected mean and variance among a set of normal deviates?

2. Consider two X-linked alleles X and X' that that affect a quantitative trait. Suppose that the phenotypic values of XX, XX', and $X'X'$ females are a, d, and $-a$, respectively, and those of XY and $X'Y$ males are a and $-a$, respectively. If the population is in Hardy-Weinberg proportions for these alleles, with p the frequency of X and q the frequency of X' ($p + q = 1$), show that the regression coefficient of the phenotypic value of sons on the genotype of fathers equals 0. Explain why the mechanism of X-linked inheritance causes this result. How would the result be affected by assortative mating?

3. In a study of abdominal bristle number in *Drosophila melanogaster*, the mean bristle number in each of a set of parents was compared to that of the mean bristle number among four of their offspring. The values were adjusted for differences in bristle number between females and males. In the pairs of numbers (x, y) shown here, x is the average bristle number of the parents (called the *midparent* value) and y is the average bristle number among four of their offspring.

   (18.5, 20.19) (18.0, 19.45) (23.0, 20.87) (20.0, 20.71) (21.0, 20.84) (19.5, 18.14)

   (23.0, 20.84) (18.5, 17.92) (19.0, 18.73) (21.0, 18.52) (23.5, 22.69) (17.0, 19.05)

   From these data, estimate the regression coefficient of offspring on midparent, and estimate the narrow-sense heritability of bristle number in this population.

4. In the pairs of numbers (x, y) shown here, x is the adult height of a father in centimeters and y is the adult height of his son.

(172, 178)	(194, 190)	(171, 176)	(178, 172)
(178, 180)	(185, 153)	(181, 186)	(179, 178)
(177, 187)	(174, 183)	(186, 181)	(173, 159)

From these data, estimate the regression coefficient of son's height on father's height, and estimate the narrow-sense heritability of adult stature in this population.

5. Estimate the correlation coefficient in litter size between the first and second litters using the following data from 10 female mice:

First litter	8	9	9	10	10	10	11	11	13	13
Second litter	6	8	12	10	12	12	9	10	12	12

6. A quantitative trait has a mean value of 23.8 in AA genotypes, 25.2 in AA' genotypes, and 19.4 in $A'A'$ genotypes.
   (a) What are the values of a and d?
   (b) What allele frequency of A would maximize the mean value of the trait in the entire population?

7. Two inbred varieties of tobacco are crossed and give a variance in leaf number in the F_1 generation of 1.5. The variance in the F_2 generation is 6.0. What are the genotypic and environmental variance components and the broad-sense heritability?

8. In a population of the flour beetle *Tribolium castaneum* the mean weight of pupae is 2000 mg. The phenotypic variance is 40,000 mg^2, and the additive genetic variance is 10,000 mg^2. If individuals with a mean pupa weight two phenotypic standard deviations above the mean are selected, what is the expected average pupa weight among the progeny?

9. If a population of *Drosophila melanogaster* has a mean number of abdominal bristles of 20 with a narrow-sense heritability of 30%, what is the expected bristle number after one generation when the selection differential is four bristles? What is the expected number after 10 generations of equally intense selection?

10. Five generations of selection for decreased plasma cholesterol level in mice decreased the mean from 2.16 mg/dl to 2.01 mg/dl. (One deciliter, abbreviated dl, equals 100 milliliters.) The average selection differential was 0.07 mg/dl. What is the realized heritability?

11. Two strains of rats were selected for increased or decreased pigmentation on the head and back. After 10 generations the high strain had a rating of 3.73 and the low strain a rating of –2.01. The strains were crossed,

and the standard deviations in the F_1 and F_2 generations were 0.60 and 0.87, respectively. Estimate the effective number of factors affecting the trait in these strains.

12. How many generations of selection with a selection differential of 10 would be required to increase the average number of eggs laid per hen per year from 180 to 220, given a narrow-sense heritability of 20%?

13. If the selection differential differs in males and females, show that the proper value to use for M_S in Equation 8.8 is the mean of males and females.

14. Show that a simple Mendelian dominant gene has a narrow sense heritability of $2(1-q)/(2-q)$, where q is the frequency of the dominant allele.

15. For an overdominant locus with two alleles, show that the additive genetic variance at equilibrium equals 0.

16. A herd of dairy cattle yields milk with a fat content of 3.4% ± 0.65% (mean ± standard deviation) and a protein content of 3.3% ± 0.45%. The narrow-sense heritabilities of these traits are 60 and 70%, respectively, and the genetic correlation is 0.55. If selection is practiced for percent protein with a selection intensity of $i = 1.5$, what increase in percent protein and percent fat would be expected? What intensity of selection would produce the same increase in percent fat by direct selection?

17. The intensity of selection i is the mean of the selected parents in a standard normal distribution when B is the proportion saved. It equals the selection differential in units of standard deviation, so $i = S/\sigma = (M_S - M)/\sigma$. Over the range $B = 0.05$ to $B = 0.005$, i is given approximately by $i = 0.8 + 0.41 \ln[(1/B) - 1]$ (Simmonds 1977). Calculate i for $B = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and $\frac{1}{32}$.

18. Consider a locus with genotypes AA, AA', $A'A'$ whose contribution to a quantitative trait has $a = 0.6$ and $d = 0.2$; another locus with genotypes BB, BB', $B'B'$ contributes to the same trait with $a = 0.4$ and $d = 0$. If the loci are unlinked and additive and the allele frequencies of A and B are 0.5 and 0.7, respectively, calculate the narrow-sense and broad-sense heritability of the trait when the total phenotypic variance is 1.0.

19. Imagine a haploid organism with genotypes A and A' at frequencies p and q with phenotypic values a and $-a$, respectively, with $p + q = 1$.
    (a) Show that the regression coefficient of offspring (O) on parent (P) equals 1. Now suppose that, in each generation, A mutates to A' with probability u, and that A' mutates to A with probability v.
    (b) Show that, in this situation, $Cov(O, P) = 4pqa^2(1 - u - v)$ and $Var(P) = 4pqa^2$, so that the regression coefficient of offspring on parent equals $1 - u - v$.

20. The malaria parasite *Plasmodium falciparum* is a sexual organism that spends most of its life stages as a haploid, including all of the stages that occur inside the human host. Suppose that K is a drug-resistance allele and k its sensitive counterpart, and that M and m are the alleles of a linked single-nucleotide polymorphism. Suppose that the gametic frequencies of KM, Km, kM, and km are u, v, w, and x, respectively ($u + v + w + x = 1$), with linkage disequilibrium parameter $D = ux - vw$. The allele frequencies of the SNP marker are therefore $u + w = p$ and $v + x = q$, where $p + q = 1$. Suppose that the phenotypes of the resistant and sensitive phenotypes are coded as $+a$ and $-a$, respectively. Show that the regression coefficient of the resistance phenotype on the SNP genotype is given by $b = 2aD/pq$. (In working the problem, your calculations will coincide with ours if you code gametes with the M allele as 1 and those with the m allele as 0.)

9.1 Evolution of Genome Size and Composition 470

Organismic Complexity and the C-Value Paradox 470

Base Composition of Genomic DNA 471

9.2 Genome-Wide Patterns of Polymorphism 473

Excess Polymorphism in Subtelomeric Regions 473

Polymorphism and Rates of Recombination 475

Hitchhiking versus Background Selection 477

Linkage Disequilibrium and Haplotype Structures 481

Decline of Linkage Disequilibrium with Genetic Distance 483

9.3 Differences between Species 484

Comparison of Nonsynonymous and Synonymous Divergence 484

Positive Selection 486

Exploiting a Phylogenetic Signal 487

Polymorphism and Divergence 487

Compensated Pathogenic Deviations 489

Structure-Function Analysis 491

9.4 Sexual Selection and the Sex Chromosomes 492

Faster-Male Molecular Evolution 493

Molecular Evolution of Genes in the X Chromosome 495

Haldane's Rule 495

Demasculinization of the X Chromosome 497

9.5 Transposable Elements 498

Diverse Types of Transposable Elements 498

Factors Controlling the Population Dynamics of Transposable Elements 501

Insertion Sequences and Composite Transposons in Bacteria 502

Transposable Elements in Eukaryotes 505

Population Dynamics of Transposable Elements 505

Nonuniformity of Transposition Rates 506

Horizontal Transmission of Transposable Elements 511

CHAPTER 9

POPULATION GENOMICS

The term **genomics** refers to the comprehensive study of all genes in the genome of an organism. The approaches include large-scale DNA sequencing, assays for transcript abundance, functional studies through systematic elimination or inhibition of each gene's activity, and computational methods for the management, comparison, and interpretation of such data. Lacking the comprehensive tools of genomics, traditional population genetics necessarily focused on relatively small samples of genes present in natural populations. Genome-wide approaches expand the purview of population genetics to such an extent that the term **population genomics** is gaining prominence to denote the application of population genetics on a genomic scale.

Genomics has vastly expanded molecular evolution, too, and so it might be worthwhile to explain briefly how molecular evolution differs from population genomics. The main objective of molecular evolution is to analyze the patterns of change in macromolecular sequences over evolutionary time, first to infer what evolutionary forces act on organisms to promote changes in the sequences and also to substantiate principles of macromolecular structure and function. The goals of population genomics are similar, but there is less emphasis on differences in macromolecular sequences among species and more emphasis on differences within populations. Hence in population genomics, inferences about the demographic history of a population and the evolutionary forces acting on the individuals in a population are based on

such observations as which genes are polymorphic, where these genes are located in the chromosomes, the distribution of allele frequencies of polymorphic genes, the extent of linkage disequilibrium along the chromosome, and so forth. The ability to make such inferences requires substantially more information than the DNA sequence of a single reference genome, supposedly representing an entire species. A reference sequence, by definition, is a single sequence without variation. In population genomics, the goal is to understand differences among genomes, and so the information needed is either complete sequence data from multiple individuals or the genotypes of *single nucleotide polymorphisms* (SNPs) scattered throughout the genome and scored in a multiple individuals.

9.1 EVOLUTION OF GENOME SIZE AND COMPOSITION

The living world presents an enormous variety in genome size and organization. Among free-living organisms, one of the smallest genomes is that of the marine bacterium *Brevundimonas diminuta* at 1.6 Mb. Among multicellular eukaryotes, one of the largest is the genome of the Australian lungfish *Neoceratodus forsteri* at greater than 50 Gb [1 gigabase (Gb) = 10^9 base pairs]. Similarly, the nucleotide composition of the genome varies among species. In bacteria, the percentage of nucleotide pairs that are GC (called the **G+C content**) ranges from about 20% to about 70%. In this section we consider some of the evolutionary considerations related to genome size and composition.

Organismic Complexity and the C-Value Paradox

In contrast to the genome sizes of bacteria and archaea, both of which are almost always smaller than 10 Mb, the genomes of eukaryotic organisms can be very large. One of the largest genomes is that of the free-living amoeba *Amoeba dubia*, whose genome size is 670 Gb. This gigantic genome is more than 200 times the size of the human genome, but an amoeba is manifestly no more complex than a human being in regard to evolution, development, or behavior. Amidst the huge diversity of eukaryotic organisms, there is no discernable correlation between genome size and organismic complexity. The seemingly arbitrary differences in genome size constitute what is known as the **C-value paradox**, because classical cytologists used the term *C-value* to refer to the characteristic amount of DNA per cell found in an organism (Hartl 2000b).

Eukaryotes with large genomes have no more genes than those with small genomes. The extra DNA consists largely of various types of repeated DNA sequences, such as transposable elements, longer introns, and noncoding sequences between genes. Lynch and Conery (2003) have argued that the increase in genome size in eukaryotes resulted from a decrease in effective population number, which caused a reduction in the efficiency of selection to eliminate gene duplications and insertions of transposable elements.

In their model, the increase in genome size was not in itself adaptive, but the extra DNA afforded many new opportunities for the origin of genetic novelty through specialization of function of duplicate genes, evolution of regulatory sequences in large introns, and so forth. On the other hand, while the effective population number is certainly smaller in large-bodied multicellular organisms than in bacteria and archaea, the effective population number of unicellular eukaryotes is unknown, and it is among unicellular eukaryotes that some of the very largest genomes are found.

Base Composition of Genomic DNA

One hypothesis for the wide range of G+C content among organisms is **mutation bias** (Suoeka 1988), which means the preferential change of GC to AT, or the other way around. A mutational bias toward GC will gradually push the genome toward a higher G+C content, and a bias toward AT will gradually push the genome toward a lower G+C content. In coding regions, the extent of the change is constrained by the nature of the genetic code, because some amino acids require codons rich in A+T and others require codons rich in G+C. The third position of many codons is more flexible, and different base compositions among organisms are usually reflected most dramatically in differences in the G+C content of the third codon position, which is commonly symbolized as GC_3.

In mammalian and bird genomes, the nucleotide composition along the chromosomes is not homogeneous, but varies significantly in G+C content in regions ranging in size from hundreds of kilobases to megabases (Filipski et al. 1973; Bernardi et al. 1985). Figure 9.1 shows an example across the MHC region of human chromosome 6, a region which includes many genes

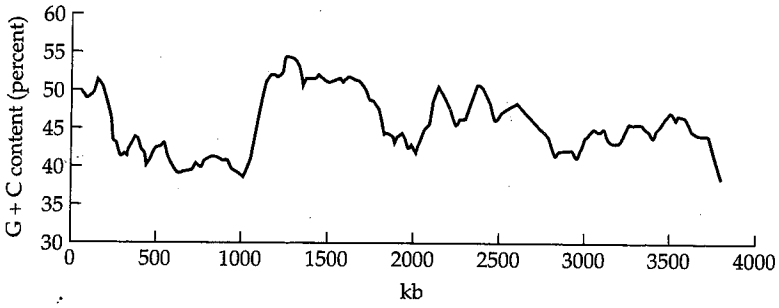


FIGURE 9.1 Variation in average G+C content in the human major histocompatibility complex. The graph shows the percent G+C in a window of 100 kb moved in increments of 10 kb across 4 Mb. Note regions of relatively low G+C that rapidly transition into regions of relatively high G+C. The regions of relatively homogeneous base composition constitute isochores. (From Eyre-Walker and Hurst 2001.)

involved in immune recognition and response (Eyre-Walker and Hurst 2001). Here the G+C content was averaged across a moving window of 100 kb, advanced in steps of 10 kb across the region. The variance in percent G+C is very much greater than would be expected from a binomial distribution. (For a G+C content of 40%, the binomial standard deviation in a random 100-kb sample would be 0.5%.) In Figure 9.1, the percent G+C drops from above 50% to about 40% in a very small region between 1500–2000 kb. Regions of high local similarity in base composition are known as **isochores** (reviewed in Eyre-Walker and Hurst 2001, Bernardi 2004). In the genomes of mammals and birds, the average isochore is longer than 300 kb, and the base composition of an isochore can range from 35% G+C to 60% G+C.

Isochores have been shown to differ in important genetic properties, such as gene density, time of replication in the cell cycle, and local frequency of recombination (Bernardi 2004). Detailed sequence analysis (Costantini et al. 2006) indicates that they are also responsible for the characteristic banding patterns revealed when human chromosomes are treated with suitable reagents.

Although isochores certainly exist, there is much less certainty about how they were formed in the first place, how they are maintained, and whether they are stable across evolutionary time (Bernardi 2004). Three principal hypotheses have been discussed: mutation bias (Wolfe et al. 1989), selection (Bernardi and Bernardi 1986), and biased gene conversion (Eyre-Walker 1993, 1999). The phenomenon of **biased gene conversion**, illustrated in Figure 9.2, can take place when similar DNA sequences interact in the process of recom-

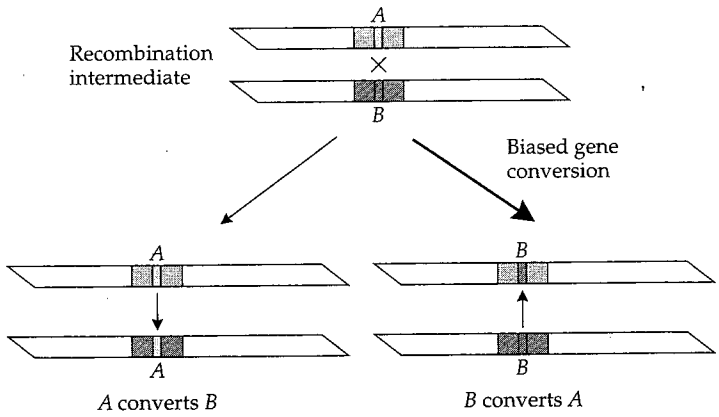


FIGURE 9.2 Biased gene conversion. During recombination, the individual strands of two double-stranded DNA molecules of sufficient sequence similarity exchange pairing partners, and mismatch repair enzymes fix any mispaired base by replacing the nucleotide in one of the strands. The nucleotide replaced can be a random choice, or the choice can be biased. In this example, the choice is biased toward the *B* sequence.

bination. During recombination, the DNA strands in a small region of one double helix come apart and form Watson-Crick base pairs with one of the strands of another double helix in the same nucleus. Although the pairing partner is often the corresponding part of the DNA molecule constituting the homologous gene, pairing can take place with any DNA strand that contains a region of sufficient sequence similarity. Within the paired region there are sometimes nucleotide mismatches, and these are resolved by **mismatch repair**, in which the mismatched nucleotide is removed and replaced with a correct Watson-Crick pairing partner. In unbiased gene conversion, which of the mismatched bases prevails is random, but in biased gene conversion there is a preference for either a G–C base pair or an A–T base pair. The hypothesis of biased gene conversion asserts that in G+C-rich regions the former prevails, and in G+C-poor regions the latter prevails. The main evidence for biased gene conversion as contributing to the maintenance of isochores is that there is a direct relation between the G+C content of an isochore and the local frequency of recombination (Fullerton et al. 2001; Meunier and Duret 2004).

The hypotheses of mutation bias, selection, and biased gene conversion are not mutually exclusive, and so there is no necessary reason why one hypothesis must be correct to the exclusion of the others. It is possible that all three processes contribute to the maintenance of isochores, with perhaps different relative weights at different times in vertebrate evolution or in different isochores. The various pros and cons of the hypotheses are discussed incisively by Eyre-Walker and Hurst (2001) and Bernardi (2004).

9.2 GENOME-WIDE PATTERNS OF POLYMORPHISM

In many species, polymorphisms are not distributed at random along the chromosomes. In some cases, excess polymorphisms result from the presence of similar DNA sequences that can exchange genetic information. In other cases, the distribution of polymorphisms reflects regional differences in rates of recombination. In still other cases, low levels of polymorphism and high levels of linkage disequilibrium are best explained as resulting from a recent episode of strong natural selection favoring a particular gene in the region. In this section we consider examples of each of these situations.

Excess Polymorphism in Subtelomeric Regions

Some species show a dramatic excess of polymorphism in the subtelomeric regions near the chromosome tips (*telomeres*). These regions can be rich in repetitive sequences that have similar, but not identical, sequences, such as copies of multigene families or transposable elements. Figure 9.3 illustrates two processes by which such repetitive sequences can interact to promote polymorphism. Part A shows **unequal crossing-over**, in which the repetitive sequences undergo recombination while misaligned (lightly shaded region paired with darkly shaded region). The result is variation in copy number (in

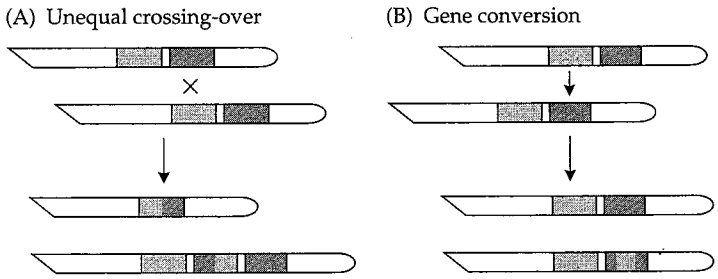


FIGURE 9.3 Two processes that increase genetic variation in subtelomeric regions containing copies of multigene families. (A) Unequal crossing over, in which recombination between mispaired copies results in products with either extra copies or fewer copies. (B) Gene conversion, in which mispaired copies can exchange genetic information through mismatch repair in a recombinational intermediate. The gene conversion may be unbiased or biased.

this example, a chromosome with one copy and another with three copies) as well as increased polymorphism due to the chimeric light/dark and dark/light sequences. Part B shows **gene conversion**, which as noted earlier typically results from DNA breakage and repair. During this process, individual strands from two different DNA duplexes can be paired temporarily, allowing nucleotides from one strand to direct the repair synthesis of the other strand. The result is that part of the sequence of one gene is replaced with that from a homologous gene, increasing the level of polymorphism.

Examples of both copy-number and excess sequence polymorphism are found in the subtelomeric regions of budding yeast, *Saccharomyces cerevisiae*, and the malaria parasite, *Plasmodium falciparum*. In yeast, most subtelomeric regions include one or more families of sequences that are shared with subtelomeric regions of other chromosomes, and these sequences can undergo unequal crossing-over and generate differences in gene copy number (Louis et al. 1994). Interestingly, the subtelomeric regions of yeast are enriched for genes with known functions in transport, transport facilitation, fermentation, and C-compound metabolism (Liti and Louis 2005), which are all functions that are important for growth in fruit juices. The subtelomeric genes include those encoding maltases, alcohol dehydrogenases, and sodium-phosphate transporters. Almost half of the 15 hexose transporter genes known in yeast are located in the subtelomeric regions.

The highly polymorphic subtelomeric regions of the malaria parasite *P. falciparum* also contain copies of repetitive elements and multigene families. The gene families include the *var* genes, which encode variants of the protein PfEMP1 that projects from the surface of infected red blood cells and promotes adherence to the blood-vessel walls, and also the *rifin* and *stevor* genes

whose functions are still obscure. As a measure of the level of polymorphism in subtelomeric regions, one study of mismatches in 25-mer oligonucleotide probes across a set of reference strains revealed that, whereas the subtelomeric regions accounted for only 22% of the total chromosome length, they accounted for 76% of the polymorphisms (Volkman et al. 2002).

Polymorphism and Rates of Recombination

Because recombination between adjacent nucleotides is infrequent, nearby nucleotide sites tend to evolve together. Owing to genetic linkage, forces that tend to maintain genetic diversity or that tend to reduce genetic diversity will act regionally. Therefore, the level of polymorphism found in any region of the genome is expected to be correlated with the level of polymorphism in a closely linked region. Evolutionary forces thus leave their mark on the level and type of genetic variation found within closely linked regions of the genome.

The chromosomes of the fruit fly *Drosophila melanogaster* are illustrated in Figure 9.4. The black bars indicate regions in which the level of nucleotide polymorphism is reduced relative to the genome as a whole. Reduced polymorphism characterizes the entire Y chromomosome (Zurovcova and Eanes 1999; Bachtrog 2004) as well as chromosome 4 (Berry et al. 1991). Because a female and a male together have 4 copies of the autosomes, 3 copies of the X

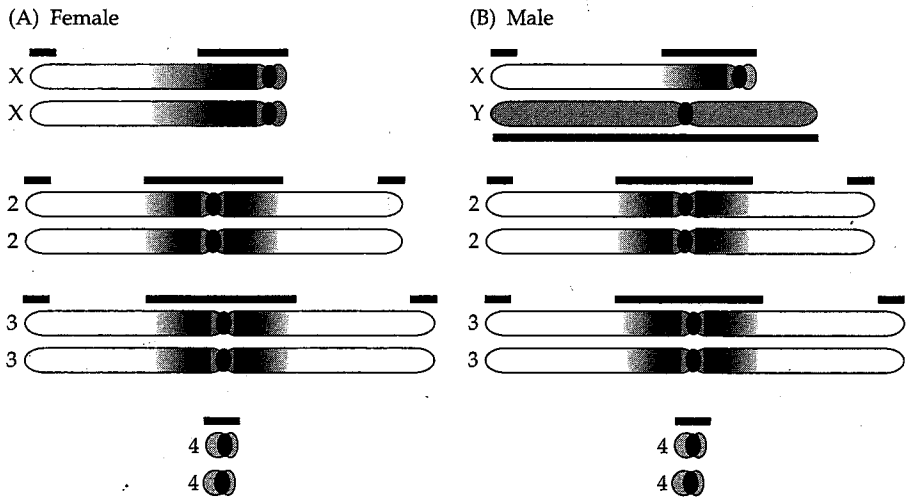


FIGURE 9.4 Diagram of the chromosomes in somatic cells of *D. melanogaster*. Black dots represent centromeres, and shaded regions around the centromeres and the Y chromosome represent repetitive DNA. The black overbars approximate the locations of regions where the frequency of recombination is reduced relative to regions of normal recombination.

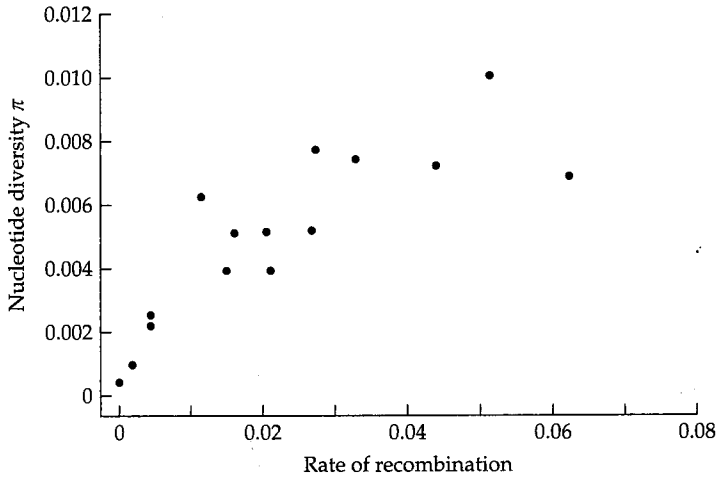


FIGURE 9.5 Observed relation between the level of nucleotide polymorphism and the rate of recombination in *Drosophila*. (From Aquadro et al. 1994.)

chromosome, and 1 copy of the Y chromosome, it follows that the effective population size for genes in these chromosomes should be in the ratio of 4 : 3 : 1. However, the reduction in level of polymorphism in the Y chromosome is far smaller than would be expected merely from the reduced effective population size.

Note in Figure 9.4 that the regions of reduced recombination include the subtelomeric regions and the regions surrounding the centromere. These are all regions in which the rate of recombination is reduced. For example, the centromeric region of chromosome 2, which is rich in highly repetitive DNA sequences and transposable elements but relatively poor in genes, accounts for about 25 percent of the length of the chromosome but only 3 percent of the amount of recombination. Chromosome 4 also has a much reduced rate of recombination relative to the rest of the genome, and the Y chromosome undergoes no recombination because it has virtually no homology with the X chromosome.

A pattern of reduced genetic polymorphism associated with less recombination in *Drosophila* is illustrated in Figure 9.5 (Begun and Aquadro 1992; Aquadro et al. 1994). In this plot, the level of polymorphism is expressed as the nucleotide diversity π, which is the average proportion of nucleotide mismatches per pair of sequences (see Chapter 4). For the regions plotted, π ranges over more than a factor of 10, so there is clearly an important effect of tight linkage in reducing the level of polymorphism.

One possible explanation for this result is that mutations rates are correlated with recombination rates, so that regions of reduced recombination

simply have less mutation. But this explanation can be ruled out by the observation that the amount of nucleotide divergence between *Drosophila* species is about the same in all regions of the genome, irrespective of the level of recombination (Jensen et al. 2002; Lea et al. 2003). If mutation were responsible for the pattern in Figure 9.5, then the level of interspecific divergence should also be affected. (Although *Drosophila* shows no relation between the frequency of recombination and the magnitude of divergence, we shall see in Chapter 10 that there is such a relation observed in primates.)

Hitchhiking versus Background Selection

If reduced recombination is responsible for the pattern in Figure 9.5, there are still two possible mechanisms—and they are diametrically opposed. In one case, the reduction results from the fixation of favorable mutations, and in the other case the reduction results from the elimination of harmful mutations. Both processes are consequences of the negative linkage disequilibrium associated with the *Hill-Robertson effect* (discussed in Section 4.6 of Chapter 4). The explanations do, however, have somewhat different implications for the pattern of polymorphism in regions of tight linkage, and so they can be distinguished experimentally.

Consider first the consequences of fixation of a favorable mutation. On its way to fixation, any new favorable mutation may carry along a small surrounding region of the genome and render the region monomorphic. The monomorphism will not usually be complete. Some degree of polymorphism may remain in the region, either because new mutations happen in the process of fixation or because of rare recombination events that take place. The process in which a favorable mutation becomes fixed in a population is called a **selective sweep**. During a selective sweep of a favorable allele, any neutral alleles sufficiently tightly linked go along for the ride and are said to be **hitchhiking**. The main effect of hitchhiking is that a small region around the favored allele will be overrepresented in the population. In other words, there will be an apparent deficiency of rare genetic variants owing to the overrepresentation of the region that profited from the hitchhiking.

The theoretical expectations of the allele frequencies in a sample are shown in Figure 9.6. The distribution of allele frequencies is often called the **allele frequency spectrum**. In Figure 9.6, the example is one in which the sample contains exactly 10 alleles. The histogram on the left shows the allele frequency spectrum that would be expected from an equilibrium between neutral mutation and random genetic drift in a population maintained at constant size, based on the theory developed in Chapter 4. The histogram on the right shows the pattern that would be expected of neutral alleles hitchhiking with a selective sweep. There are too many common alleles, and conversely too few rare alleles. On the other hand, the detection of selection through its perturbation of the allele frequency spectrum is problematical, in part because factors other than selection can cause such patterns (for exam-

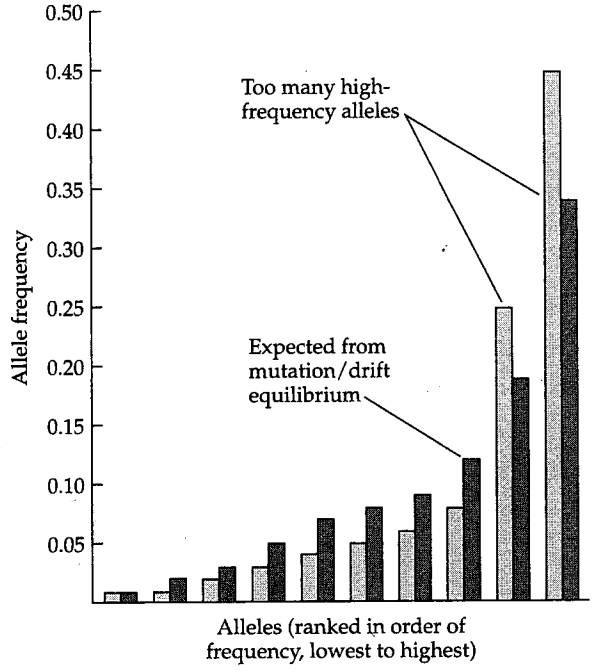


FIGURE 9.6 Comparison of the expected allele-frequency spectrum for 10 alleles at equilibrium between neutral mutation and random genetic drift in a population of constant size (dark bars), and that of neutral alleles hitchhiking with a selective sweep (light bars). In an extreme case of hitchhiking, there would be only one allele at very high frequency, with all other alleles being rare.

ple, rapid population growth), and in part because the patterns are transient in that the excess of alleles with high frequency disappears when the beneficial mutation in the region is fixed (Kim and Stephan 2000, 2002).

Next consider the consequences of a harmful mutation in a region of low recombination. For concreteness, consider the genetic map diagrammed in Figure 9.7A, in which the short vertical lines indicate adjacent nucleotide sites. One site that can undergo neutral mutation is embedded in the middle surrounded by sites that can undergo harmful mutations only. The rate of harmful mutation per site per generation is denoted μ, and the rate of recombination between adjacent sites is denoted r.

Suppose further that each mutation, even when heterozygous, is sufficiently harmful that any chromosome in which a mutation is present is ultimately doomed. In the absence of recombination, the fate of a chromosome depends on whether it is free of harmful mutations because, under our assumptions, no chromosome can persist for long unless it is free of mutations. The effect of harmful mutation, which in this context is called **background selection**, is to reduce the number of chromosomes that can contribute to the ancestry of remote generations. Indeed, the effect of background selection is identical to that of a reduction in population size except that the reduction applies, not to the genome as a whole, but to a tightly linked region (Charlesworth et al. 1993, 1995). Background selection there-

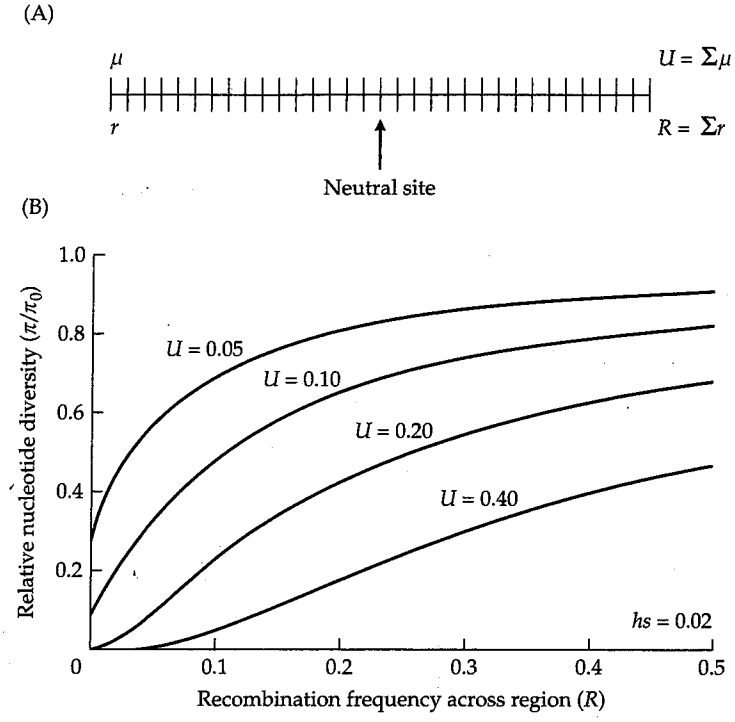


FIGURE 9.7 Effects of background selection on nucleotide polymorphism. (A) A region of a chromosome containing a set of genes (tick marks) that can mutate to detrimental alleles; within this set of genes is a single neutral site. The mutation rate per locus is μ and the rate of recombination between adjacent loci is r. (B) Relative nucleotide diversity as a function of U, the total mutation rate, and R, the total recombination rate, across the chromosomal region. Note the positive correlation between level of nucleotide polymorphism and rate of recombination.

fore reduces the level of genetic polymorphism. Looser linkage means that a linked neutral mutation can escape the fate of a harmful neighboring mutation by recombination with a mutation-free chromosome. Hence, the tighter the linkage, the greater the reduction in polymorphism due to background selection. Although there is a reduction in the level of polymorphism, background selection does not skew the distribution of rare polymorphisms because, for all practical purposes, the harmful allele merely causes one chromosome to drop out of the population, much as if it were to go extinct by chance (Braverman et al. 1995). Hence the expected allele frequency spectrum is that shown on the left in Figure 9.6.

The effect of background selection on the level of genetic variation is shown graphically in Figure 9.7B for the genetic map diagrammed in part A. The curves are plotted from the formula

$$\frac{\pi}{\pi_0} = \exp[-U / (2hs + R)] \tag{9.1}$$

(Hudson and Kaplan 1995). The symbol π is the *nucleotide diversity,* defined in Chapter 4 as the average proportion of nucleotide differences between all possible pairs of sequences; π_0 is the value of π in the absence of background selection. U and R refer to the diagram in part A. U is the total mutation rate per diploid genome, summed across all genes in the region; and R is the total rate of recombination across the region, summed over each of the intervals between genes. The quantity hs measures the degree of harmfulness of each deleterious mutation in a heterozygous genotype; the extremes are $hs = 0$, when there is no effect in the heterozygote, and $hs = 1$, when the heterozygote is lethal. The model on which Equation 9.1 is based includes the assumption that hs is small but not 0.

The curves in Figure 9.7B are for the specific value $hs = 0.02$, which means that a genotype that is heterozygous for one deleterious mutation has a 2% reduction in survival compared with a homozygous nonmutant. For each curve, the relative nucleotide diversity (π/π_0) decreases as the total recombination rate R decreases. This result means that, with tighter linkage, each detrimental mutation that is eliminated takes with it a larger surrounding region of chromosome. The relative nucleotide diversity also decreases as the total mutation rate increases; that is, greater background selection eliminates a greater number of chromosomes. Together, tight linkage and a moderate or high total mutation rate can result in a very substantial decrease in relative nucleotide diversity, reducing it to a level of 20% or less of that expected in the absence of background selection.

Controversy roils over hitchhiking versus background selection (Andolfatto 2001; Storz 2005). Part of the problem is that the allele frequency spectra are difficult to distinguish (Innan and Stephan 2003; Charlesworth et al. 2003). Hence the analysis of different samples can easily yield contradictory results as to whether there is (Andolfatto and Przeworski 2001) or is not (Braverman et al. 1995) a skewed distribution toward rare variants. Computer simulations suggest, perhaps counterintuitively, that nucleotide diversity in regions of very low recombination would be determined primarily by hitchhiking, and that in regions of somewhat greater recombination primarily by background selection (Kim and Stephan 2000). This observation brings us back to the low level of diversity in the *Drosophila* Y chromosome (Zurovcova and Eanes 1999). In this case, because the *Drosophila* Y chromosome encodes no more than a handful of genes, background selection seems less likely to explain the situation than hitchhiking (Bachtrog 2004). On the other hand, background selection does impede gene fixation, as evidenced by the finding in *Drosophila* that the average rate of amino acid replacement

is reduced in regions of low recombination (Betancourt and Presgraves 2002).

Linkage Disequilibrium and Haplotype Structures

Genetic linkage can help to overcome some of the problems in detecting selection based on tests of the allele- frequency spectrum. Although most new mutations are lost, even those that are favorable (the probability of ultimate fixation of a favorable allele with additive effects and selection coefficient s is approximately $2s$), those lucky few favorable mutations that escape loss in the first few generations have a good chance of becoming fixed. As a favorable allele gradually increases in frequency in the population owing to selection, the neighboring sequences in the chromosome are also given a boost in frequency. Selection thereby creates a state of linkage disequilibrium (a correlation between alleles of different genes) between the selected allele and linked nucleotide sites in the genome. Recombination tends to restore linkage equilibrium (random association of alleles in gametes), and so the linkage disequilibrium generated by hitchhiking persists for a relatively short time, on the order of $0.4N$ generations for realistic recombination rates (Przeworski 2002). Nevertheless, if selection is sufficiently strong and recent, and linkage sufficiently tight, the signature of selection can sometimes be detected.

When there is linkage disequilibrium, the set of nucleotide or allelic states that are associated in a chromosome is called a **haplotype**. Examples of haplotypes in a region of the genome of *D. simulans* are shown in Figure 9.8. Only nucleotide sites that are polymorphic in the sample are shown. These are scattered across a region of about 1300 bp in the gene *RpL32* for a ribosomal protein (Rozas et al. 2001). Almost all of the polymorphisms are silent (synonymous sites or noncoding sites), and a few small length polymorphisms have been ignored. The haplotypes prefixed with the letter S are from a site in Spain, those prefixed M are from a site in Mozambique.

The nonrandomness in the haplotypes in Figure 9.8 is apparent. The shaded haplotypes S8–S10 and M7–M14 are clearly related to one another in spite of showing a considerable amount of polymorphism. The unshaded haplotypes (S1–S6 and M1–M5) are identical to one another except for a single polymorphic nucleotide in M1. The low level of polymorphism in this haplotype implies a relatively recent common ancestor, which is consistent with a model of selection with hitchhiking. Two haplotypes, S7 and M6, are clearly recombinants.

Other than selection, how can such data be explained? One possibility is that the haplotype structure is due purely to chance. This hypothesis seems unlikely because the genomic region in which *RpL32* is located shows a relatively high level of recombination. But the hypothesis can be tested explicitly by coalescent (ancestry) computer simulations of neutral mutation and random drift in a chromosome undergoing recombination, as explained in

```
S1    TTTTGTTAAGAATCAAGTTATCGGGGCGTCCCAGGGTGCATTTCCAGGGGGGT
S2    ....................................................
S3    ....................................................
S4    ....................................................
S5    ....................................................
S6    ....................................................
S7    ....TCCGC..T...CCCA...AACT..........................
S8    ....TCCGC.TT...CCCA..TA.C..C...............G..AAAA.
S9    ....TCCGC..TCTGC.C.....A.C...............ACA..GA..AAA.
S10   ....TCCGC..T.T.C......AACT....TA........GCA..G......G

M1    .C..................................................
M2    ....................................................
M3    ....................................................
M4    ....................................................
M5    ....................................................
M6    ....TCCGC..TCTGC.C....A.CT..C.......................
M7    .G..TCCGCA.T...C......AACT...T...A...........G.A.AAA.
M8    ....TCCGG..TCTGC.C....A.C...............ACA..GA..AAA.
M9    ....TCCGCA.T...C......AACT...T...A...........G.A.AAA.
M10   ...CTCCGC..T...CCCAG....C.T.......A....GGCA..G...AAA.
M11   ...CTCCGC..T...CCCAG....C.T.......A....GGCA..G...AAA.
M12   ....TCCGCA.T...C......AACT...T...A...........G.A.AAA.
M13   ....TCCGC..T...CCCAGC.AACT..C...T.....T.GCA..G...AAA.
M14   A...TCCGC.TT...CCCA..TA.C..C...............G..AAAA.
```

FIGURE 9.8 Polymorphic nucleotides observed in haplotypes of *D. simulans* across a region of about 1.3 kb near the gene *RpL32*. The shaded and unshaded regions are clearly more similar among themselves than between the groups. (Data from Rozas et al. 2001).

Chapter 3. As expected, the observed haplotype structure is extremely unlikely to be due to chance (Rozas et al. 2001).

A second possibility for any observed haplotype structure is to invoke historical changes in population size or structure (Charlesworth et al. 2003; Haddrill et al. 2005). One such demographic effect is a **founder effect**, because a small number of individuals who initiate a subpopulation will contribute disproportionately to its gene pool. Another is **population admixture**, because a population formed from a mixture of individuals from subpopulations differing in haplotype frequencies will exhibit linkage disequilibrium until recombination can randomize the alleles in gametes (see Chapter 6). In some cases, rapid population growth (*expansion*) or shrinkage (a *bottleneck*) can also result in linkage disequilibrium. Because the demographic history of most natural populations is unknown, it is difficult to exclude such explanations. In principle, however, demographic effects should leave a similar signature throughout the genome, whereas effects due to hitchhiking should be localized. Application of this principle is limited by the fact that demographic effects, as they play out through recombination and random genetic

drift, can have a large variance from one position in the genome to the next. However, the haplotype structure in Figure 9.8 is so extreme for *Drosophila* that its presence in two geographically distinct populations provides a strong case for positive selection of some gene or genes in the region accompanied by hitchhiking (Rozas et al. 2001). Additional studies of the region around *RpL32* have yielded additional evidence for this interpretaton (Quesada et al. 2003; Meiklejohn et al. 2003, 2004).

Decline of Linkage Disequilibrium with Genetic Distance

Linkage disequilibrium is expected to decrease more rapidly with higher rates of recombination. In a finite population, the effect of recombination is offset to some extent by random genetic drift of the haplotype frequencies. For two alleles at each of two loci in an ideal population of size N, the expected linkage disequilibrium r^2 at steady state in a finite population is neatly captured in an equation due to Hill and Robertson (1968) and Ohta and Kimura (1971a):

$$E(r^2) = \frac{1}{1+4Nc} \tag{9.2}$$

where $r^2 = D^2/(p_1p_2q_1q_2)$ is a measure of linkage disequilibrium (see Equation 2.14 in Chapter 2). In perhaps more familiar terms, the square root of r^2 ($\sqrt{r^2}$) is the correlation coefficient between alleles in gametes. In Equation 9.2, c is the frequency of recombination between the genes. Typically we have represented the frequency of recombination as r, but in this case we switched to c to avoid confusion with the symbol for the linkage disequilibrium. The principal assumptions underlying Equation 9.2 are the absence of mutation and selection, a relatively large population, and random mating. See Gaut and Long (2003) for a good discussion of what happens when these assumptions are violated.

The implications of Equation 9.2 for linkage disequilibrium in the human genome have been examined in Chapter 3 (see Figure 3.18). They are shown in a somewhat different format in Figure 9.9 for various values of $4Nc$. The curve labeled 0.025 corresponds approximately to the parameters for *Drosophila*. Linkage disequilibrium may be considered nominal if $r^2 \leq 0.2$ (Gaut and Long 2003). For *Drosophila*, the corresponding distance is about 200 bp. There is, however, very large variation in this number among organisms and from one region of the genome to another. In humans, for example, the corresponding value is about 1 kb but can be as high as 50 kb (Long and Langley 1999). Data for six genes in maize yielded values for $r^2 \leq 0.2$ of 200 bp (three genes), 600 bp (one gene), and 1400 bp (two genes) (Remington et al. 2001). In addition to large region-to-region variation in linkage disequilibrium, there is a very large stochastic variation between different pairs of nucleotides in any given region, owing to large variation in the ancestral history of recombination and drift between any pair of nucleotides. It is not

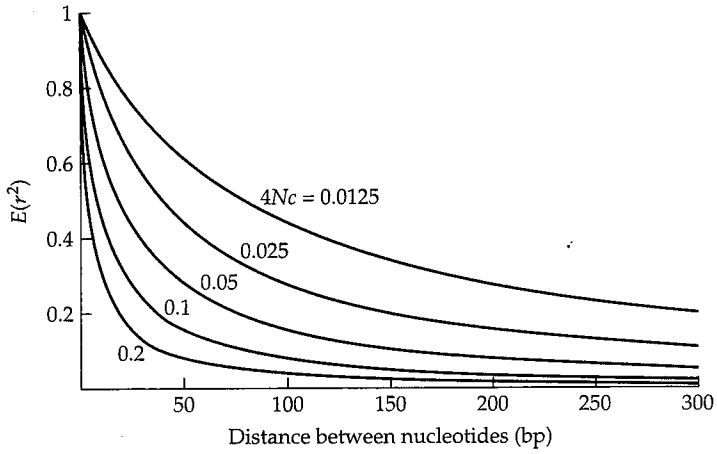


FIGURE 9.9 Theoretical relation at steady state comparing linkage disequilibrium between single nucleotide polymorphisms with distance in nucleotides, for different values of $4Nc$, where N is the effective population number and c is the distance in frequency of recombination between the nucleotides.

uncommon, for example, to find that two nucleotide sites are in linkage equilibrium with each other but for both sites to be in linkage disequilibrium with a site in the middle!

9.3 DIFFERENCES BETWEEN SPECIES

Although reduced polymorphism or linkage disequilibrium sometimes result from selection, these signatures are relatively weak unless the selection is relatively recent and relatively strong. Comparisons among species take advantage of a much longer time scale; hence more subtle patterns of selection can be detected. In this section we consider some of the approaches that are used in interspecific comparisons.

Comparison of Nonsynonymous and Synonymous Divergence

Coding sequences give abundant evidence that most mutations that change amino acids are deleterious. By chance they may persist in a population for a few generations, but eventually they are eliminated by selection; in this context the process is called **purifying selection**. The evidence for this assertion is that amino acid differences usually accumulate between species at a much lower rate than would be expected. As discussed in Chapter 7, the proper way to compare two aligned coding sequences, for example, sequences from different species, is to consider them site by site, taking into account which of all possible nucleotide substitutions at each site would be nonsynonymous and which would be synonymous, and then to estimate the number of nonsynonymous differences per nonsynonymous site (usually denoted as either d_N or K_A) relative to the number of synonymous differences per synonymous site (usually denoted either d_S or K_S). Nonsynonymous and synonymous sites are

not fixed entities, but can change through time. For example, if an AUG (methionine) codon is replaced with a GUG (valine) codon, the third position changes from a nonsynonymous site to a fourfold degenerate synonymous site. Hence it is best to think of nonsynonymous and synonymous sites as "mutational opportunities," rather than any particular subset of nucleotide sites.

Details aside, the ratio d_N/d_S tells us the rate at which amino acids have changed in a protein relative to the rate at which silent sites have changed. If there are no selective constraints, as in a pseudogene, then the expected value of $d_N/d_S = 1$, but if there is purifying selection, then the expectation is $d_N/d_S < 1$. Hence, as noted in Chapter 7, the quantity d_N/d_S serves as an index for inference of positive selection (allele fixation driven by selection), but a test of significance for any individual gene may be lacking in statistical power.

The utility of d_N/d_S comparisons increases greatly when carried out on a large scale, ideally across the entire genome. One informative example is shown in Figure 9.10 (Sharp 1997). It is a histogram of d_N/d_S for 363 protein-coding genes compared in mouse and rat. The median value is 0.08, and more than $\frac{3}{4}$ of the genes have a d_N/d_S value smaller than 0.2. This evidence shows the preponderance of purifying selection. Most nonsynonymous mutations must be sufficiently deleterious that they have little or no chance of every becoming fixed.

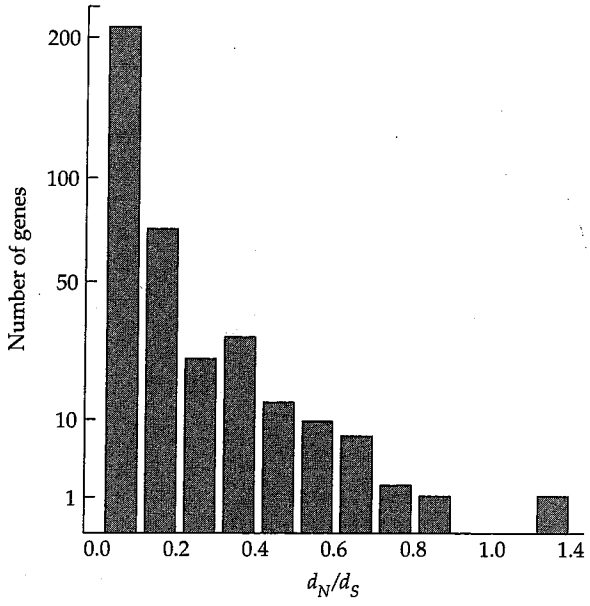


FIGURE 9.10 Observed distribution of d_N/d_S among 363 protein-coding genes in mouse versus rat. (From Sharp 1997.)

Positive Selection

What about the single gene in Figure 9.10 with $d_N/d_S > 1$? This gene evidently undergoes more amino acid replacements than expected by chance, as if the changes were driven by positive selection. The gene encodes interleukin-3, which stimulates the development of mast cells and bone-marrow cells of the immune system. It is only one of many example of genes associated with the immune system that appear to be undergoing rapid adaptive evolution at the protein level (reviewed in Nei 2005). Undoubtedly, other genes in Figure 9.10 have also had one or more amino acid replacements driven by positive selection, but the overwhelming preponderance of purifying selection drowns out any signal from positive selection.

Even though many proteins subject to positive selection will remain unidentified, the ratio of nonsynonymous to synonymous substitutions can still be used to identify the most rapidly evolving proteins. The approach is simply to compare the d_N/d_S ratio across many proteins and focus on those with values greater than 1. Sequences derived from messenger RNA molecules, known as *expressed sequence tags* (ESTs), are useful for such comparisons. In one study of 176 independent ESTs from *D. melanogaster* and *D. simulans*, 19 showed evidence of strong positive selection, especially genes encoding male-specific reproductive proteins (Swanson et al. 2001). Similarly, a study of 304 ESTs between *Arabidopsis thaliana* and *A. lyrata* identified 14 candidates for genes under positive selection (Barrier et al. 2003).

Heterogeneity is an issue that arises when trying to infer positive selection from d_N/d_S with genome-wide data. Genomes are heterogeneous in may ways, with wide variation in local rates of recombination, G+C content, and even codon usage. The earliest methods for estimating d_N and d_S had the property of being sensitive to base composition (Nei and Gojobori 1986). Newer methods have several advantages for scaling up to genome-wide tests. These methods invoke a codon-substitution model that uses the full 61 $\times$ 61 matrix of nontermination codons, and the methods estimate the rates of substitution among all single-step changes (Goldman and Yang 1994; Muse and Gaut 1996). These models estimate a parameter ω, defined as the relative rate of nonsynonymous to synonymous substitution, which is identical to d_N/d_S when there is no change in nucleotide frequencies. Using these methods, Nielsen et al. (2005b) studied 13,731 annotated genes between chimpanzees and humans, and they found that many of the genes showing evidence for positive selection are involved in sensory perception, immune defenses, tumor suppression, apoptosis (programmed cell death), or spermatogenesis. X-linked genes had a somewhat greater likelihood of showing evidence for positive selection, but there was no evidence for greater positive selection of genes maximally expressed in the brain.

Exploiting a Phylogenetic Signal

The power of the d_N/d_S ratio, which is called ω in the context of codon-substitution models, can be enhanced by comparison of orthologous sequences among related species because each codon can be tracked separately to estimate differing selection intensity among amino acid sites. An extension of the codon-substitution model pioneered by Nielsen and Yang (1998) features amino acids sites that are either neutral or subject to purifying selection (for which $\omega \leq 1$), and those under positive selection (for which $\omega > 1$). Whether any amino acid sites fall into the latter category is first tested by maximum likelihood, and then candidate sites for positive selection are classified. Another version of this model allows rate variation among lineages as well as among sites (Yang and Nielsen 2002; Yang et al. 2005). This approach was used to examine 7645 orthologous sequences from the genomes of human, chimpanzee, and mouse to identify genes whose rate of amino acid replacement had increased in the human lineage. Accelerated evolution was found in a number of functional categories including genes associated with olfaction, hearing, skeletal development, and neurogenesis (Clark et al. 2003).

Other comparative methods that exploit phylogenetic signal have also been developed. One widely used approach, which examines the evolution of individual codons through each branch of a gene tree (Suzuki and Gojobori 1999), is in principle very powerful if the gene tree can be estimated accurately and the sequences at each node of the tree correctly inferred. Another approach allows for uncertainties in the tree and the substitution process through the use of a Bayesian model that averages across all possible combinations of trees, branch lengths, transition bias, and codon usage (Huelsenbeck and Dyer 2004). This approach is relatively computation intensive, but it has the advantage of yielding an estimate of the full posterior probability distribution that any particular site has undergone positive selection. Sometimes these distributions are highly skewed. For example, for one site in vertebrate beta-globin whose average posterior probability of being under positive selection was 0.71, the 95 percent credible interval for the posterior probability was 0.006 to 0.993 (Huelsenbeck and Dyer 2004). All of these methods improve in power and phylogenetic specificity as the number of species is increased, with a considerable gain for each species added until about six species are included altogether.

Polymorphism and Divergence

Recall from Chapter 7 that the signature of positive selection is reflected in a relatively greater magnitude of divergence between species as compared with polymorphism within species. The between-species versus within-species comparisons are conveniently carried out with χ^2 tests in 2×2 tables,

where the synonymous sites in a coding sequence serve as a control for comparison with the nonsynonymous sites (McDonald and Kreitman 1991; Sawyer and Hartl 1992). This test is indifferent to the allele-frequency spectrum of the polymorphisms, and estimates of the selection coefficient are remarkably robust to population structure (Wakeley 2003). While lacking in statistical power on a gene-by-gene basis, this approach gains statistical traction if applied to large samples of genes from any pair of species with a suitable degree of evolutionary divergence. As noted in Chapter 7, many genes in *Drosophila* give evidence for positive selection based on such comparisons, and from such data Smith and Eyre-Walker (2002) estimated that about 45 percent of amino acid replacements between *D. simulans* and *D. yakuba* had been driven by positive selection.

One possibility for the apparent excess of amino acid replacements is that it is an artifact of recent population expansion (Fay et al. 2002; Haddrill et al 2005). The basis of the argument is that, if the effective population size is sufficiently small, slightly deleterious mutations with $|Ns| < 1$ will accumulate as polymorphisms, and some of them will become fixed. If at some point the population size increases dramatically, then the increase in N may make $|Ns| > 1$, and so many deleterious polymorphisms will be eliminated from the population. But the already fixed differences will remain fixed, and so the demographic change will result in an excess of amino acid replacements, mimicking the effect of positive selection.

The demographic explanation has been examined using data from 45 genes in *D. melanogaster* and *D. simulans* (Fay et al. 2002). The authors conclude that population expansion is unlikely to explain the excess of amino acid fixations. Instead they emphasize that most of the deviation from neutral expectation is due to the high amino acid divergence observed in about one-third of the genes, many of which encode proteins implicated in male reproductive functions.

A more formal approach to the analysis of polymorphism and divergence is based on explicit models of selection and estimation of the relevant parameters from the data (Bustamante et al. 2002, Sawyer et al. 2003). This approach allows simultaneous analysis of polymorphism and divergence across an unlimited number of genes, an approach that has the advantage that estimates of common parameters (such as the time of species divergence) are based on all of the data. One explicit model assumes that, for each gene, the relative fitness of a new nonsynonymous mutation is given by $\exp(Ns)$ and that the values of Ns are drawn at random from a normal distribution whose mean and variance are parameters to be estimated (Sawyer et al. 2003). Application of this model to polymorphism and divergence of 56 genes in *D. simulans* and *D. melanogaster* supports a number of intuitive notions. The first is that purifying selection is predominant. Although the estimated mean Ns differs among the genes, the overall average is -7.3, reinforcing the idea that most new mutations are harmful. The second observa-

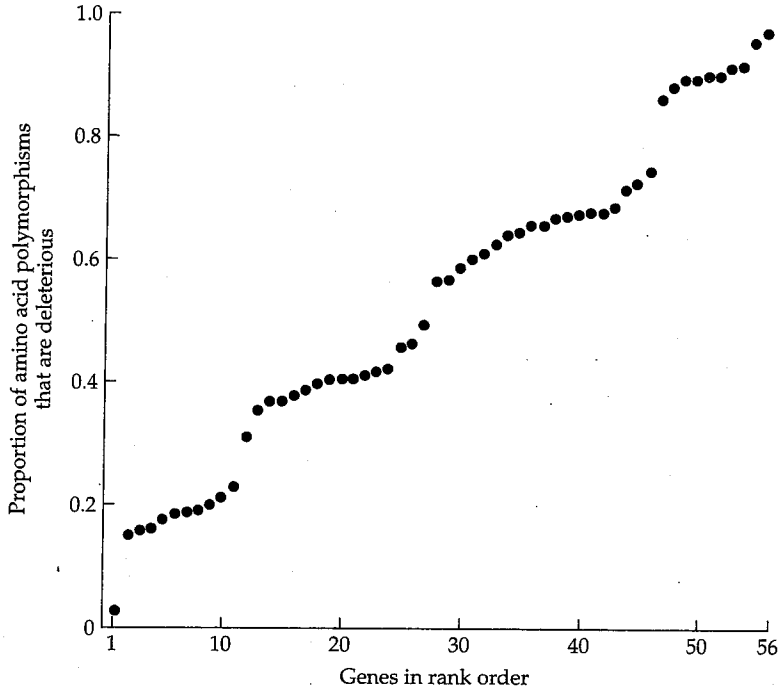


FIGURE 9.11 Proportion of segregating alleles for amino acid polymorphisms that are deleterious, estimated from analysis of polymorphism and divergence of 56 genes in *D. simulans* and *D. melanogaster*. (Data from Sawyer et al. 2003.)

tion is that a substantial proportion of amino acid polymorphisms are deleterious. In this sample of genes, the average proportion of deleterious polymorphisms is 53% (Figure 9.11). The analysis delivers unexpected results in its estimate of the proportion of amino acid replacements that are positively selected, a number that averages 93% across the genes. One limitation of this analysis is that it assumes that each nonsynonymous mutation in a gene has an independent and additive effect on fitness, and hence individually deleterious but mutually compensatory mutations would appear in the model as positively selected. The average value of Ns of fixed mutations is modest in any case, with an estimated mean in this model of $Ns = 5$ (Sawyer et al. 2003).

Compensated Pathogenic Deviations

That compensatory amino acid replacements do take place in the course of evolution is demonstrated by a class of replacements called **compensated pathogenic deviations**. These are amino acid replacements that are known to

be severely deleterious (pathogenic) in one species, but that appear as the wildtype amino acid (whose deleterious effects must therefore be compensated) in the orthologous protein in a related species. Such compensated replacements can be identified in only a few organisms because they require a database of mutations known to be deleterious.

Compensated pathogenic deviations were first described by Kondrashov et al. (2002) in a study of 32 human proteins, each with at least 50 different pathogenic mutations. They compared the human sequences with orthologous proteins from other species and discovered that, among the amino acid sites that were not completely conserved across all species, about 10 percent of the pathogenic amino acids in humans were present as the wildtype amino acid in at least one other species. Notably, the 10 percent figure seemed to be independent of the phylogenic distance from the human lineage, as if the pathogenic deviation and the compensatory mutation (or mutations) were fixed almost simultaneously. This inference suggests positive selection of combinations of mutations that cancel (or more than cancel) each other's possibly deleterious effects.

Another organism with an extensive database of pathogenic deviations is *D. melanogaster*, the fruit of almost 100 years of intensive genetic research. Many pathogenic amino acid replacements have been identified that cause gross morphological defects, sterility, or lethality. The proteins containing pathogenic deviations were compared with orthologous molecules in *D. pseudoobscura*, which diverged from *D. melanogaster* approximately 40–50 MYA, and the mosquito *Anopheles gambiae*, which diverged approximately 250 MYA. Once again, among amino acid sites that were not completely conserved, the proportion of compensated pathogenic deviations was approximately 10 percent (Kulathinal et al. 2004). This proportion was found for both *D. pseudoobscura* and *A. gambiae*, which again suggests that the compensatory changes occur nearly simultaneously in evolutionary time.

Compensatory changes in RNA sequences are well known. In ribosomal RNA sequences, they result from complementary changes in both members of a pair of nucleotides in the stem of a stem-loop structure, so that, for example, a G–C nucleotide pair in a stem structure observed in one species is found as an A–U nucleotide pair in another species (Kimura 1990; Dixon and Hillis 1993). Compensatory changes driven by selection for RNA secondary structure have also been observed in mRNA molecules (Stephan 1996; Innan and Stephan 2001).

There is no reason to think that compensatory intramolecular interactions are restricted to RNA. Most proteins, for example, are only marginally stable, and the free-energy difference between being too unstable to fold properly and being too stable to function properly is often on the order of one or two hydrogen bonds (DePristo et al. 2005). At the same time, the free energy of folding results from the difference between a large number of stabilizing interactions (such as hydrogen bonding, burial of hydrophobic groups, and

disulfide bonds) and a large number of destabilizing interactions (such as conformational entropy, burial of peptide groups, and burial of polar groups), each of individually small effect. The large number of interactions means that almost any amino acid replacement will have some effect on protein folding and stability; these can be compensated by a different amino acid replacement somewhere else in the molecule. Such compensatory changes would allow a protein to meander through sequence space maintaining approximately the same folding and stability in different lineages, with no change in function (DePristo et al. 2005).

Structure-Function Analysis

Identifying candidate sites for positive selection is only one step in understanding the physiological basis of adaptation. Going beyond candidate sites requires a detailed analysis of molecular structure and an understanding of how changes in structure relate to changes in function. There are still relatively few examples in which molecular evolutionary changes have been linked to physiological adaptations in the organism. Many of them emerge from studies of hemoglobin or of the light-sensitive proteins in the eye. We will consider one example from each system.

Some species of geese are adapted to the low levels of oxygen present at high altitudes. One is the bar-headed goose *Anser indicus*, which lives and reproduces in high Tibetan lakes 13,000–20,000 feet (4000–6000 meters) above sea level, and then in August flies over the summit of Mount Everest to overwinter in the plains of India. Another is the Andean goose *Chloephaga melanoptera*, which lives in South America at altitudes of 13,000–20,000 feet (5000–6000 meters). In both of these species, their hemoglobin has a much higher affinity for oxygen than does that of species that live near sea level, such as the greylag goose *Anser anser*. Specifically, the partial pressure of oxygen for 50 percent saturation of hemoglobin is 29.7 mmHg in the bar-headed goose, 33.9 mmHg in the Andean goose, and 39.5 mmHg in the greylag goose. It turns out that, in adapting to life at low oxygen, the geese in the Himayalas and those in the Andeas underwent independent—but complementary—amino acid replacements in the alpha and beta chains of the hemoglobin molecule.

The hemoglobin molecules of the greylag goose (sea level) and the bar-headed goose (Himalayas) are identical in the beta chain but differ at four sites in the alpha chain. Three of the four differences are near the surface, but one difference (alpha Pro119Ala) affects a position at which the alpha chain normally touches the beta chain at position beta Leu55. In the goose that lives at sea level, the proline does touch beta Leu55, but in the Himalayan goose the alanine leaves a gap that relaxes the structure and is expected to raise the affinity of the molecule for oxygen. Confirming the prediction, when the Pro119Ala replacement was introduced into human hemoglobin, the oxygen affinity increased by 43 percent (Jessen et al. 1991).

The hemoglobins of the greylag goose (sea level) and the Andean goose differ at 9 sites in the alpha chain and 7 sites in the beta chain. Suspicion immediately fell on one difference in the beta chain, Leu55Ser, because it affects precisely the same alpha-beta contact as Pro119Ala in the previous comparison. In the goose that lives at sea level, the beta Leu55 touches alpha Pro119, but in the Andean goose the serine again leaves a gap predicted to increase the oxygen affinity. When this amino acid replacement was introduced into human beta-globin, the oxygen affinity again increased, in this case by 41 percent (Jessen et al. 1991).

Human hemoglobin differs from goose hemoglobin at about 30 percent of the amino acid sites, yet the replacements to alpha 119Ala and beta 55Ser show the increase in oxygen affinity expected from the evolved changes in the Himalayan and Andean geese. This does not imply that the other amino acid sites in hemoglobin do nothing. They may affect folding, stability, resistance to aggregation, or other properties of the molecule, without having a major effect on oxygen affinity. Analyses like those in goose hemoglobin are possible only when the functional anatomy of the protein molecule and its relation to organismic physiology are exceptionally well understood.

Human color vision is mediated by three opsin proteins, which absorbs maximally at wavelengths of about 420 nm (blue opsin), 530 nm (green opsin), and 560 nm (red opsin). The genes for the red and green opsins are tandemly arrayed near the tip of the long arm of the X chromosome, and originated as a duplication of an ancestral red opsin gene about 40 MYA, at about the time that Old World primates and New World primates diverged. Although the human red and green opsin proteins differ at 15 amino acid sites among 360 sites in the molecule, their difference in spectral sensitivity (pattern of wavelength absorption) resides primarily in differences at five sites—Ser180Ala, His197Tyr, Tyr277Phe, The285Ala, Ala308Ser (red opsin sequence given first)—with a strong interaction between sites 180 and 197 (Yokoyama and Radlwimmer 2001).

The situation is similar in the opsins of the blind cave fish *Astyanax fasciatus*. It, too, has red and green visual pigments, which arose from an ancestral opsin gene independently of those in the lineage of Old World primates. The red and green Astayanax opsins differ at 71 amino acid sites, yet the difference in spectral sensitivity appears to reside in the same five key residues as distinguish the red and green human opsins (Yokoyama and Yokoyama 1990). It warrants emphasis again that the other 65 amino acid differences may play a role in protein folding, stability, resistance to aggregation, packing into the membrane, or other attributes of the molecule.

9.4 SEXUAL SELECTION AND THE SEX-CHROMOSOMES

In addition to providing a grand scale for analysis of natural selection, genome-wide approaches present excellent opportunities for making infer-

ences about sexual selection and the processes involved in the evolution of reproductive isolation between species. Recall that **sexual selection** refers to the selective advantage an individual gains over others of the same sex in the struggle to acquire mates and produce offspring. To the extent that genes relating to these traits can be identified, genome-wide approaches can be very powerful.

At the molecular and developmental level, male reproduction seems profligately uneconomical. Some species of *Drosophila* produce sperm whose tails are 60 mm long (ten times the length of the males themselves!) that require more than two weeks to become mature (Pitnick et al. 1995). *Drosophila* males produce about 85 different proteins in the accessory gland that are secreted and transferred to females during copulation. These proteins promote sperm storage, displacement of sperm from other males, egg production and ovulation, and reduced female receptivity to further mating (Swanson et al. 2001; Wolfner 2002; Fiumera et al. 2005). In mammals, many genes are transcribed at much higher levels than in somatic cells, but then are translated with low efficiency (Kleene 2005). Other genes produce transcripts from alternative promoters that skip parts of the protein-coding region or that contain signals in the 5′ untranslated region or upstream reading frame that inhibit translation (Kleene 2001). The germ-line nucleotide mutation rate in male primates is more than five times larger than that in females (Li et al. 2002).

Faster-Male Molecular Evolution

Studies in population genomics have demonstrated that genes involved in reproductive processes evolve at a rapid rate, especially in males (Swanson and Vacquier 2002). This pattern was first noticed by Civetta and Singh (1995) in their electrophoretic study of interspecific protein differences in *Drosophila*. Since then many examples have accumulated in various animal species (Wolfe and Li 2003; Oliver 2003; Singh and Kulathinal 2005). Many of the most rapidly evolving genes function in male reproduction, such as the protamine proteins that replace histones during spermiogenesis in humans and other mammals (Wyckoff et al. 2000). In the *Drosophila* genome, genes encoding the secreted male accessory gland proteins are among the most rapidly evolving. Table 9.1 shows d_N/d_S for a sample of 121 ESTs encoding proteins in the accessory glands of *D. simulans* males, along with 32 nonreproductive proteins (Swanson et al. 2001). Among the 44 accessory gland ESTs encoding proteins containing characteristic protein motifs associated with secretion, 13 (22.8 percent) showed a significant $d_N/d_S > 1$. The percentage of putative proteins with $d_N/d_S > 1$ was significantly smaller for nonsecreted accessory gland proteins as well as nonreproductive proteins. Secreted accessory gland proteins also have an elevated index of dispersion of amino acid replacements (Kern et al. 2004). This means that the ratio of the variance to the mean rate of amino acid replacement is large, which generally indicates a pattern of episodic bursts of protein evolution.

TABLE 9.1 Accelerated Evolution in *Drosophila* Male Reproductive Proteins

Gene product	$d_N/d_S \leq 1$	$d_N/d_S > 1$	Percent with $d_N/d_S > 1$
Secreted accessory gland protein	44	13	22.8
Nonsecreted accessory gland protein	77	6	7.2
Nonreproductive protein	32	0	0

Source: Data from Swanson et al. 2001.

Rapid evolution of genes involved in male reproduction is also observed at the level of transcription. In *Drosophila*, about 50 percent of the genes show a sex-biased in expression in adults, with greater expression in either males or females, and these sex-biased genes are primarily those associated with the reproductive tissues (Parisi et al. 2003). Microarray studies have shown that male-biased genes show greater levels of expression variation within species than do other classes of genes (Meiklejohn et al. 2003), and they also show more rapid divergence in expression between species (Ranz et al. 2003). Moreover, compared to other classes of genes, male-biased genes show elevated ratios of interspecific expression divergence to intraspecific expression variation, and these ratios suggest a process driven at least in part by sexual selection.

Male-biased genes also show significantly faster rates of amino acid replacement than other classes of genes as assayed by their d_N/d_S ratios (Z. Zhang et al. 2004). The acceleration is due primarily to greater values of d_N, and it is found both within the melanogaster species subgroup of *Drosophila* as well as across more distant phylogenetic comparisons.

Overall, the existing data show quite convincingly that genes involved in male reproduction evolve rapidly in both amino acid sequence and expression level. This pattern is often attributed to sexual selection, currently the most plausible hypothesis. On the other hand, more than sexual selection might be involved. For example, some authors have suggested that mutations affecting spermatogenesis that result in sex-ratio segregation distortion (nonmendelian segregation, also called meiotic drive) might evolve very rapidly (Frank 1991; Hurst and Pomiankowski 1991). There is ample precedent for such mutations (Lyttle 1991; Jutier et al. 2004), and repeated episodes of sex-ratio segregation distortion that are compensated by suppressor mutations could result in rapid evolution for progressive molecular remodeling of male gametogenesis. Such processes may underlie the uncovering of sex-ratio segregation distortion in certain interspecific hybrids (Tao et al. 2001). In

any event, the hypothesis that sexual selection drives faster male evolution certainly accounts for some of the cases, such as the secreted accessory gland proteins in *Drosophila*, but perhaps there is also more to the story.

Molecular Evolution of Genes in the X Chromosome

To the extent that beneficial mutations might be recessive, it is reasonable to expect that genes in the X chromosome might evolve more rapidly under selection than those in autosomes, because beneficial recessive genes would be expressed in males. Indeed, many genes in the X chromosome do show evidence of positive selection. Comparisons of human and chimpanzee genomic sequences show about a 30 percent excess of nonsynonymous substitutions in X-linked genes as compared with autosomal genes (Lu and Wu 2005; Nielsen et al. 2005b). The X chromosome differs from the autosomes in several other ways, too. For example, the effective population sizes of the X chromosome and autosomes are in the ratio 3 : 4, but this would effect d_N and d_S to approximately the same extent and so cannot account for the 30 percent difference. Then too, the rate of nucleotide mutation in the germ line of male primates is larger than that in females (Li et al. 2002; Makova and Li 2002). However, in any generation this would affect only one-third of the X chromosomes but half the autosomes.

X-linked genes associated with male reproduction also have increased rates of evolution. For example, comparison of X-linked, testes-expressed transcription factors in the homeobox family between humans and mice showed substantially higher rates of amino acid replacement than found in other genes (Wang and Zhang 2004). Adding to the evolutionary action, the X chromosome in mammals is also a preferred source and target for sequences that are moved from one location to another via a process in which transcripts are reverse-transcribed into double-stranded DNA and inserted into the genome (Khil et al. 2005).

Haldane's Rule

The full array of forces driving the evolution of genes in the X chromosome are far from understood, however. Paramount among generalizations about speciation is **Haldane's rule** that "when in the offspring of two different animal [species] one sex is absent, rare, or sterile, that sex is the [heterogametic] sex" (Haldane 1922). In animals with XX–XY sex determination, males are the heterogamatic sex, whereas in birds, moths and butterflies, where the sex chromosomes are reversed, females are the heterogametic sex. For a masterly review of Haldane's rule and other aspects of the genetics of speciation, see Coyne and Orr (2004).

Haldane's rule is exceptionally general in its applicability. It accounts for 95 percent of 255 cases tabulated by Coyne and Orr (2004). The genetic basis of the rule is therefore of great interest. Four major hypotheses have been proposed:

1. The *dominance theory*, which posits that mutations affecting hybrid sterility or inviability are recessive or partially recessive, so that the fitness of the heterogametic sex is more severely reduced than that of the homogametic sex (Orr and Turelli 1995). From an extensive set of experiments in which discrete regions of the genome of *Drosophila mauritiana* were introduced (*introgressed*) into the genome of its sibling species *D. simulans*, the degree of dominance of mutations affecting hybrid sterility was estimated as approximately 0.25 and that of mutations affecting hybrid lethality as approximately 0.35 (Tao and Hartl 2003). These are measured on a scale in which a completely dominant mutation would have a value of 1 and a completely recessive mutation a value of 0; hence for both hybrid sterility and hybrid lethality, the mutational effects are partially recessive.
2. The *faster-male theory*, which holds that rapid evolution of genes involved in male gametogenesis and sexual selection will result in the more rapid accumulation of factors affecting male fertility and viability (Wu and Davis 1993). We have already summarized some of the evidence for rapid evolution of genes involved in male reproduction. The hypothesis is also reinforced by *Drosophila* introgression experiments, which indicate that factors causing hybrid male sterility accumulate at a rate about seven times faster than those affecting hybrid female sterility (Tao and Hartl 2003). On the other hand, while there is some support for this hypothesis in species in which the male is heterogametic, it seems less applicable to species in which the female is heterogametic, unless of course heterogamy in and of itself promotes rapid evolution.
3. The *large-X effect*, which asserts that mutations in X-linked genes have a greater effect on hybrid fitness than mutations in autosomal genes (Charlesworth et al. 1987). This hypothesis also receives some support from *Drosophila* introgression experiments, because factors affecting hybrid male sterility appear to be more than twice as dense in the X chromosome as in the autosomes (Tao and Hartl 2003). On the other hand, experiments designed to estimate X versus autosomal effects on hybrid sterility have yielded mixed results (Coyne and Orr 2004).
4. The theory of *meiotic drive*, which holds that repeated cycles of sex-ratio segregation distortion and suppression can remodel meiosis and spermatogenesis so rapidly that incompatibilities cause hybrid male sterility (Frank 1991; Hurst and Pomiankowski 1991). This theory is perhaps too easily dismissed because cases of dramatic sex-ratio distortion are not often observed. But polymorphisms for sex-ratio distorters will usually be transient, because they automatically favor suppressors that tend to restore an equal sex ratio. Hence a species could undergo repeated cycles of sex-ratio distortion and suppression, yet with only brief periods of polymorphism. Nevertheless, there are many reported cases of sex-ratio meiotic drive in *Drosophila*, some observed as within-species polymorphisms, others revealed in interspecific crosses (Mercot et al. 1995;

Jaenike 1996; Presgraves et al. 1997; Tao et al. 2001; Yang et al. 2004; Orr and Irving 2005; Reed et al. 2005, and still more examples in Lyttle 1991 and Jutier et al. 2004). Since *Drosophila* is by far the most extensively studied organism in this regard, it is not known how far these findings can be generalized to other animal species. Nevertheless, the currently available data does support a role for sex-ratio meiotic drive in heterogametic hybrid sterility.

Demasculinization of the X Chromosome

In *Drosophila melanogaster*, as well as the nematode worm *Caenorhabditis elegans*, the X chromosome carries significantly fewer genes affecting male reproductive functions than would be expected based on the density of such genes in the autosomes (Jiang et al. 2001; Kelly et al. 2002; Meiklejohn et al. 2003; Parisi et al. 2003; Ranz et al. 2003). This has been called the *demasculinization* of the X chromosome, and the forces promoting it are unclear (Wu and Xu 2003; Oliver and Parisi 2004).

One possible force promoting demasculinization is selection based on *dosage compensation*, which refers to the mechanism by which the activity of genes in the X chromosome is adjusted to be equal in the sexes. The mechanism is quite different between flies and worms: In flies, the X chromosome in males is transcriptionally twice as active as either X chromosome in females, whereas in worms, each X chromosome in a female is transcriptionally half as active as the X chromosome in males. Both of these mechanisms differ from the mechanism of dosage compensation in mammals, in which only one X chromosome in each somatic cell of a female is transcriptionally fully active.

Other hypotheses for demasculinization of the X chromosome in flies and worms envisage a role for **sexually antagonistic selection**, which means a fitness trade-off between the sexes. In this model, genes that are beneficial for females are deleterious for males, and the other way around, so that genes favorable for female functions might accumulate in the X chromosomes and those favorable for male functions in the autosomes.

Yet another hypothesis for demasculinization relates it to X-chromosome inactivation in male gametogenesis (Wu and Xu 2003). The inactivation takes place during the early stages of spermatogenesis, and it has evolved independently in many lineages. To the extent that genes acting late in spermatogenesis are more sexually antagonistic than other genes, there should be selection promoting their transposition to the autosomes. In time, as ever more of these late acting genes go to the autosomes, the X chromosome could evolve to shut down owing to its excess of genes favoring female functions.

Transposition is likely the mechanism by which genes are moved from one chromosome to another, usually through the action of reverse transcriptase on RNA transcripts of the genes. In *Drosophila*, the X chromosome serves as a source of transposed genes about twice as often as expected. Moreover,

most of genes moved from the X chromosome to an autosome show expression in testes, perhaps reflecting their escape from X inactivation during spermatogenesis (Betran et al. 2002).

A paradox of X-chromosome demasculinization is that it is not observed in mammals (Wang et al. 2001), so whatever mechanisms are invoked for flies and worms cannot generally be true. Among the major differences between the sex chromosomes of mammals and those of flies and worms are the mechanisms of dosage compensation, discussed earlier, and the evolutionary origins of the sex chromosomes themselves. In mammals, it is quite clear that the X and Y chromosomes are derived from an autosomal pair of chromosomes that diverged in sequence (Lahn and Page 1999). In *Drosophila*, the Y chromosome appears to have evolved independently in different lineages from a dispensable extra (supernumerary) chromosome that acquired male reproductive genes and the ability to undergo meiotic pairing with the X chromosome (Carvalho and Clark 2005). The situation in *C. elegans* is different: there is no Y chromosome—females are XX, males X (often designated XO to emphasize the absence of a Y chromosome).

9.5 TRANSPOSABLE ELEMENTS

A DNA sequence that can change its location within the genome is called a **transposable element** or **transposon** (Berg and Howe 1989; Craig et al. 2002). In being able to create novel genome rearrangements, transposable elements are agents of genetic variation. A transposable element may insert into a coding region and inactivate a gene or insert into a regulatory region and change the pattern of expression of the gene. Also, pairs of transposable elements may undergo recombination and create novel chromosome rearrangements.

Diverse Types of Transposable Elements

The process of transposition requires specialized proteins, which are often encoded within the sequence of the transposable element itself. There are many distinct families of transposable elements, but most of them can be assigned to either of two broad classes.

- *Class I transposons* undergo transposition through a replicative process via an RNA intermediate that is then reverse-transcribed into double-stranded DNA by means of an enzyme, **reverse transcriptase**, prior to insertion (Figure 9.12A). These elements are often called *retrotransposons*. They include two large groups designated LTR (long terminal repeat) elements and non-LTR elements. The LTR elements include, at their ends, directly repeated DNA sequences of 100–5000 bp. The non-LTR elements lack these repeats. Non-LTR elements include elements designated LINEs (long interspersed elements) and SINES (short interspersed elements). SINE elements do not encode their own reverse transcriptase but can use that produced by other retrotransposons. The human genome contains

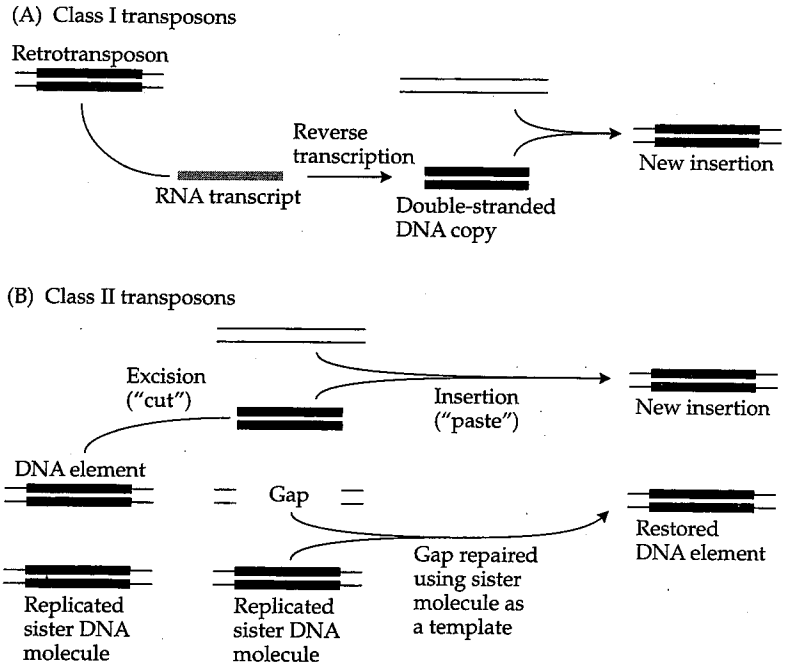


FIGURE 9.12 Two major classes of transposable elements are mobilized by different mechanisms. (A) Class I transposons move by means of an RNA intermediate, which is converted into double-stranded DNA by the enzyme reverse transcriptase. The parental element retains its original position in the genome. (B) Class II transposons move by means of a cut-and-paste mechanism, whereby the double-stranded DNA of a parental element is cleaved from its original location and inserted elsewhere in the genome. The gap in the parental molecule is often filled by template-directed gap repair using DNA in the sister chromatid formed by DNA replication in the cell cycle as a template, so the transposon in the original position is restored.

roughly half a million copies of LTR retrotransposons (accounting for about 10 percent of the genome), roughly one million copies of LINE elements (about 20 percent of the genome), and roughly 1.5 million copies of SINE elements (about 15 percent of the genome).

- *Class II transposons* use a **transposase** protein to move by a cut-and-paste process, in which a copy of the element is cleaved from one position in the genome and inserted into another (Figure 9.12B). Specificity of the excision is typically achieved through transposase recognition of inverted-repeat sequences at the ends of the transposon. Class II elements are often called *DNA transposons* because they never pass through an RNA intermediate. When the cleavage of a class II transposon takes place in a DNA molecule after replication, then the double-stranded break resulting

from the cleavage can be repaired using the sister DNA molecule as a template (see Figure 9.12B). In this way, the element that is "cut" is actually replaced by DNA-repair using the replicated copy as a template, so the copy number of the transposon can increase. In some genomes, many copies of class II elements have transposase mutations or internal deletions that inactivate the transposase. If the defective copies retain the inverted repeats, they can still be mobilized by functional transposase produced by intact elements in the same genome. The human genome contains roughly 300,000 copies of DNA transposons, accounting for about 3 percent of the genome.

Because of replicative transposition (in the case of class I transposons) or template-directed DNA repair (in the case of class II transposons), transposable elements can increase in copy number in the genome. In principle, their numbers would increase indefinitely until checked by deleterious fitness effects on the host organism. But in most cases the rate of transposition is regulated before the fitness cost becomes excessive. The rate of transposition can be regulated by properties of the transposase itself (Lohe and Hartl 1996; Simmons et al. 2004), or by factors intrinsic to the host including methylation (Bender 2004) or siRNA (small interfering RNA) (Sijen and Plasterk 2003; Blumenstiel and Hartl 2005). In addition, transposon copies can be deleted by any of a number of mechanisms.

Some transposons are highly specialized. For example, the non-LTR retrotransposons *R1* and *R2* (which are very dissimilar despite their names) insert only into specific target sites that occur at frequent intervals in the clustered 28S ribosomal RNA genes in arthropods. *R1* and *R2* have persisted in arthropod lineages for at least 600 million years (Eickbush 2002). Interestingly, *R1* elements transpose predominantly in the male germ line, but *R2* elements transpose in the germ lines of both sexes (Perez-Gonzalez et al. 2003).

Specialized transposons in *Drosophila* also include two non-LTR retrotransposons known as *HeT-A* and *TART* (Pardue and DeBaryshe 2003). These elements are found only in tandem arrays at the ends of chromosomes, and in fact they substitute for the telomeres usually composed of tandem arrays of short DNA sequences found at the chromosome ends in most eukaryotes. The evolutionary switch from short tandem telomeric arrays to *HeT-A* and *TART* arrays seems to have occurred more than 60 million years ago (Casacuberta and Pardue 2003).

HeT-A and *TART* are examples of transposons that have been recruited to carry out important functions of the host genome, an evolutionary process called **molecular domestication** (Miller 1999). Sometimes only part of a transposon is co-opted for host functions. For example, the immune system in all jawed vertebrates generates antibody diversity in part by a process of recombination between three gene segments (V, D, and J) encoding different regions of an antibody protein. These segments are flanked by recombination signal sequences, and the recombination is catalyzed by the products of two

genes, *RAG1* and *RAG2* (recombination activating genes 1 and 2). The RAG1 protein contains a catalytic core region of approximately 600 amino acids that is similar to the transposase protein of a family of 3–4 kb DNA transposons known as *Transib* elements, which are found in many vertebrates as well as some insects, sea urchins, lancelets, hydra, sea anemones, and plants (Kapitonov and Jurka 2005). In *Drosophila*, most transposon insertions are at low frequency, and a screen for putatively adaptive insertions present at high frequency revealed a recent insertion disrupting a choline kinase gene and conferring increased resistance to an organophosphate pesticide (Aminetzach et al. 2005).

Factors Controlling the Population Dynamics of Transposable Elements

Transposable elements were originally discovered in maize as the cause of certain genetically unstable mutations. They are now known to be ubiquitous among prokaryotes and eukaryotes (Berg and Howe 1989; Craig et al. 2002). The ability of transposable elements to increase in copy number and create novel chromosomal rearrangements reveals a dynamic aspect of genome structure and evolution not previously recognized. Some transposable elements have become widely disseminated among organisms because of their ability to undergo horizontal transmission between reproductively isolated genomes. Often referred to as *selfish DNA* because transposition alone may be sufficient for persistence in the genome of a species, transposable elements also may occasionally be domesticated and become agents of adaptive evolution.

Models for the population dynamics of transposable elements usually incorporate several features:

- A rate of infection, in which genomes previously lacking the transposable element become infected with it.
- A rate of transposition, which determines how rapidly the copy number increases; the effects of regulation are taken into account by assuming that the rate of transposition is a decreasing function of copy number.
- A mechanism, or combination of mechanisms, for eliminating elements from the population; otherwise, the copy number would increase indefinitely. The usual assumption is that the presence of transposable elements in the genome decreases the ability of an organism to survive and reproduce, resulting in the elimination of some elements by means of natural selection, or that elements can be eliminated from the genome by means of genetic deletion.

Through the study of such models, the diversity and novel attributes of transposable elements have been incorporated into the concepts of population genetics; see, for example, Langley et al. (1983), Montgomery and Langley (1983), Kaplan and Brookfield (1983), Sawyer et al. (1987), Hartl and Sawyer (1988), Ajioka and Hartl (1989), Charlesworth et al. (1994), Brookfield and Badge (1997).

Insertion Sequences and Composite Transposons in Bacteria

Bacteria contain several types of transposable elements. Among the simplest are **insertion sequences**, which are typically about 1000–2000 nucleotides in length and contain at least one long translational open reading-frame coding for the transposase protein. The transposase recognizes a short nucleotide sequence, inverted in orientation, present at each end of the insertion sequence, and so the element moves as an intact unit. The bacterium *Escherichia coli* contains several types of insertion sequences, each different but all sharing the same sequence organization with inverted repeats and at least one open reading frame. The factors controlling the population dynamics of insertion sequences can be deduced from the distribution of numbers of each element present among a sample of bacterial strains isolated from natural sources (Sawyer et al. 1987).

Population models of transposable elements in *E. coli* are greatly simplified because the organism has asexual reproduction, a low rate of recombination among strains, and a low rate of deletion of insertion sequences. The "state" of a bacterial strain with respect to a particular insertion sequence may be defined as the number of copies n of the element that are present. Among the factors that control the population dynamics are:

- The rate u at which uninfected cells become infected; u is the probability, per generation, that a cell initially in state $n = 0$ ends up in the state $n = 1$.
- The rate T of transposition in infected strains; T is the probability, per generation, that a cell in state $n > 0$ goes to state $n + 1$. This formulation embraces a multiplicity of biological mechanisms by which the activity of insertion sequences can be regulated (Nagy and Chandler 2004).
- The rate S at which reproduction of infected cells is less than that of uninfected cells. In terms of exponential growth, if r_0 is the intrinsic rate of increase of uninfected cells and r_0' is that of infected cells, then $S = r_0 - r_0'$.

The most general models of this type allow for T and S to be functions of n, but here we will assume that they are constant. Note, however, that the assumption that T is a constant implicitly defines a type of regulation because, if the probability of transition from state n to state $n + 1$ is independent of n, then the probability of transposition *per element* present in a strain must equal T/n and this fraction is a decreasing function of n.

Given constant values of u, T, and S, then it can be shown that a population of bacterial cells attains an equilibrium distribution of numbers of transposable elements in which the probability p_i that a cell contains exactly i copies of the transposable element is equal to

$$p_0 = \alpha \tag{9.3a}$$

$$p_i = (1 - \alpha)(1 - \phi)\phi^{i-1} \ (i \geq 1) \tag{9.3b}$$

where $\alpha = 1 - (u/S)$ and $\phi = T/(T + S - u)$ (Sawyer and Hartl 1986; Sawyer et al. 1987).

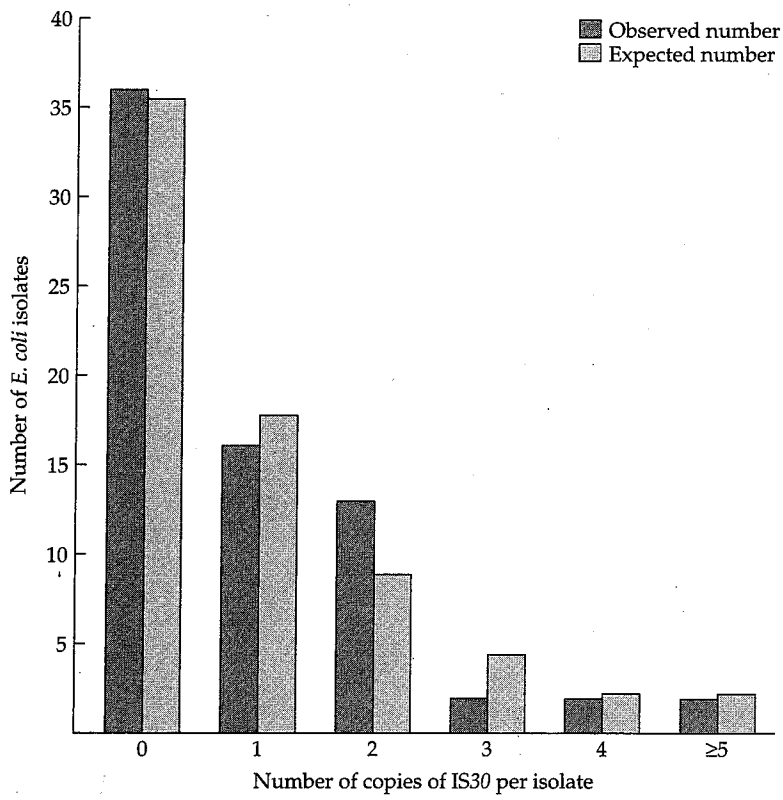


FIGURE 9.13 Observed and expected number of IS*30* elements in the genome of each of 71 isolates of *E. coli*. The expected numbers are from Equation 9.3, based on a model in Sawyer et al. (1987).

Equation 9.3 can be applied to the concrete case of insertion sequence IS*30* in *E. coli*, in which the distribution of numbers among 71 strains fits a model with $\alpha = \frac{1}{2}$ and $\phi = \frac{1}{2}$. With these parameters, the distribution simplifies to the remarkably simple formula $p_i = (\frac{1}{2})^i$ for $i \geq 0$. Among 71 strains, therefore, the observed and expected numbers of strains containing i elements are as indicated in Figure 9.13. The strains with five or more elements have been grouped in order to carry out a χ^2 test of goodness of fit. This χ^2 test has three degrees of freedom because α and ϕ were estimated from the data. The value of χ^2 equals 3.48, which has an associated probability level of about 0.32. Thus, the simple model for IS*30* fits the observed data very well. Although the χ^2 test cannot be completely trusted in this case because of the small expected numbers in some of the categories, the conclusion is supported by a more exact statistical test (Sawyer et al. 1987). The following problem deals with the distribution of three other insertion sequences in *E. coli*.

PROBLEM 9.1 The distribution of IS*1* fits Equation 9.3 with $\alpha = \frac{1}{5}$ and $\phi = \frac{5}{6}$; IS*2* fits the equation with $\alpha = \frac{2}{5}$ and $\phi = \frac{2}{3}$; and IS*4* fits with $\alpha = \frac{2}{3}$ and $\phi = \frac{3}{4}$. Calculate the expected numbers for 71 strains and carry out a χ^2 test. (The observed numbers are from Sawyer et al. 1987.)

	No. copies					
	0	1	2	3	4	≥ 5
IS*1*	11	14	8	6	7	25
IS*2*	28	8	12	5	5	13
IS*4*	43	5	5	3	5	10

ANSWER For IS*1*, the expected distribution is given by $p_0 = \frac{1}{5}$, $p_i = (\frac{4}{25})(\frac{5}{6})^i$ for $1 \le i \le 4$, and $p_{\ge 5} = 1 - (p_0 + p_1 + p_2 + p_3 + p_4)$. For IS*2*, the expected distribution is $p_0 = \frac{2}{5}$ and $p_i = (\frac{3}{10})(\frac{2}{3})^i$ $(1 \le i \le 4)$. For IS*4*, the expected distribution is $p_0 = \frac{2}{3}$ and $p_i = (\frac{1}{9})(\frac{3}{4})^i$ $(1 \le i \le 4)$. Expected numbers, χ^2 values, and associated probabilities are:

	No. copies							
	0	1	2	3	4	≥ 5	χ^2	*P* value
IS*1*	14.2	9.5	7.9	6.6	5.5	27.4	3.58	0.31
IS*2*	28.4	14.2	9.5	6.3	4.2	8.4	6.31	0.10
IS*4*	47.3	5.9	4.4	3.3	2.5	7.5	4.00	0.26

As in the case of IS*30*, more exact statistical tests confirm the conclusion that the model fits. However, the distribution of IS*1* has a very long tail, with nine strains containing from 15 to 20 copies and six strains containing from 21 to 30 copies; this distribution is approximated even more closely by a model in which the regulation of transposition decreases more gradually than *T/n* (Sawyer et al. 1987).

Apart from their own evolutionary dynamics, insertion sequences are important because they can mobilize other sequences in the genome. When two copies of an insertion sequence are on flanking sides of an unrelated sequence, the inverted repeats used in transposition are preferentially those at the extreme ends. This kind of insertion-sequence sandwich constitutes a **composite transposon**, which transposes as a single unit. In a composite transposon, the central sequence can include one or more genes that confer a selective advantage on the host cell, such as a gene for resistance to an antibiotic; hence, the possession of the composite transposon would be favored in an environment containing the antibiotic.

Mobilization of genes for antibiotic resistance, heavy-metal resistance, and other functions is one of the principal evolutionary implications of trans-

posable elements in bacteria. Transposable elements enable the piecewise assembly of specialized, infectious molecules called **plasmids**. Plasmids are autonomously replicating, usually circular molecules of DNA that exist within bacterial cells. Many plasmids contain genes that promote their transfer between different organisms. They may also contain genes, such as those for antibiotic resistance, that are highly advantageous to their hosts in certain environments. These genes are often contained in transposons, and they undoubtedly entered the plasmid through transposition from a different plasmid or from the genome of a previous host. Infectious plasmids containing multiple antibiotic-resistance genes are called **resistance transfer factors**, and they are a major source of multiple drug resistance in pathogenic bacteria.

Transposable Elements in Eukaryotes

Transposable elements can have important genetic consequences as mutagenic agents by the creation of novel genes, by alteration of the expression of genes in their vicinity, and in the genesis of major genomic rearrangements. Transposable elements also have important implications in population genetics and evolution. As we have described, several major classes of transposable elements have been identified that differ in the molecular mechanisms of transposition, such as LTR retrotransposons, non-LTR retrotransposons, and cut-and-paste DNA elements. Within each class of elements, the members can also differ in DNA sequence. Based on similarity in DNA sequence, transposable elements typically can be grouped hierarchically into "subfamilies," in which the elements resemble each other quite closely; "families," in which they differ from one another somewhat more; and "superfamilies," in which the differences are relatively great. Transposable elements are widespread in both animals and plants. For example, *Drosophila melanogaster* contains 5–150 copies of each of about 100 different families of transposable elements (Petrov et al. 2003). Although few of these elements have been studied in detail from the standpoint of population genetics, indirect evidence suggests that most of the elements, like insertion sequences in bacteria, are mildly harmful to the host (Golding et al. 1986; Lohe et al. 1995; Deininger and Batzer 1999; Petrov et al. 2003; Callinan et al. 2005).

Population Dynamics of Transposable Elements

For eukaryotic species that reproduce through sexual reproduction, segregation and recombination imply that, in each generation, the transposable elements in the genome become reassorted in such a way that a copy inserted at any particular position becomes nearly independent of copies inserted at other positions, except for those positions that are genetically tightly linked. This means that individual genomes need not be tracked, as was necessary in considering bacterial insertion sequences, but rather that the population dynamics can be encapsulated in the distribution of copy number among

individuals in the population. In one formulation of this approach, when the logarithm of the relative fitness of an individual containing n copies of the transposable element decreases linearly as n increases, then there is a stable equilibrium of copy number at which the average fitness in the population is given by

$$\overline{w} = \exp[-(u - v)\overline{n}] \tag{9.4}$$

where $\overline{n}$ is the average number of copies per individual and u and v are, respectively, the rates of transposition and excision per copy (Charlesworth and Charlesworth 1983).

To appreciate what Equation 9.4 implies, some specific examples may be instructive. To simplify matters, we will assume that $v = 0$ so that all copies that are eliminated are eliminated because of selection, not excision. Now consider a transposable element in *Drosophila* that has an average copy number of 50 per genome. For a rate of transposition per copy as high as $u = 10^{-4}$, then $\overline{w} = 0.995$, and for $u = 10^{-5}$, then $\overline{w} = 0.9995$. The implication of these calculations is that the effect of maintaining a transposable element on average population fitness is relatively modest. What about exceptional cases, such as the 300-bp SINE element *Alu* that is present in the human genome at about a million copies? *Alu* elements transpose at a rate of about one new insertion every 200 births (Deininger and Batzer 1999), which is to say one new insertion per 400 gametes. Each gamete represents one million copies of *Alu* that have replicated, hence the rate of transposition per *Alu* element can be estimated as $u = 1/(1{,}000{,}000 \times 400) = 2.5 \times 10^{-9}$. Assuming that one million is the equilibrium copy number, Equation 9.4 implies that $\overline{w} = 0.9975$. Hence, under the assumptions of Equation 9.4, even such a huge number of *Alu* elements has but a modest effect on the average fitness in the human population.

PROBLEM 9.2 The LINE element *LINE1* is present in the human genome in about 500,000 copies. Under the assumptions of Equation 9.4 with $v = 0$, what rate of transposition per element per generation would result in an average fitness in the population of 0.99? Of 0.999?

ANSWER With $v = 0$, Equation 9.4 implies that, $u = -\ln(\overline{w})/\overline{n}$ hence for $\overline{w} = 0.99$ and $\overline{n} = 500{,}000$, $u = 2 \times 10^{-8}$. For $\overline{w} = 0.999$, $u = 2 \times 10^{-9}$.

Nonuniformity of Transposition Rates

It is an oversimplification to think that rates of transposition are the same through time, or even that they are the same for each copy of a transposable

element. The *Alu* family of transposable elements exemplifies both types of inhomogeneity. As these elements evolve, different copies accumulate different mutations, and as each copy undergoes retrotransposition, any characteristic mutations in the parental copy are inherited by the daughter copy. In this way, various subfamilies of *Alu* elements evolve that can be distinguished from each other by their characteristic differences in sequence. Mutations that are more recent will be present in a minority of elements and perhaps in only part of a subfamily, but mutations that are older will be more widespread and even present in two or more subfamilies. The sequence similarities and differences between *Alu* elements allow them to be grouped into a hierarchical tree structure, much as orthologous gene sequences from various species can be grouped in a gene tree.

When *Alu* sequences are grouped in this fashion, it becomes clear that certain subfamilies are highly overrepresented relative to other subfamilies that contain approximately the same number of mutations and so are of approximately the same age. The generally accepted explanation of this overrepresentation is that *Alu* transposition occurs through a small number of persistent "master" copies (Han et al. 2005). On the other hand, some of the currently most active *Alu* elements in the human genome contain mutations that occurred long ago, and the close sequence similarity among copies in these subfamilies implies that the recent expansion is taking place after a long period of quiescence.

A case in point is the *AluYb* subfamily (Han et al. 2005). Accounting for about 40 percent of human *Alu* elements, this subfamily is one of the largest and most active in the human genome, with approximately 20 percent of *AluYb* insertions being polymorphic in contrast to less active elements for which most insertions are fixed. Most copies presently in the human genome became integrated with the past 3–4 million years, and the total copy number is about 2000 (Carter et al. 2004). The *AluYb* subfamily can be subdivided still further into subgroups such as *AluYb7*, *AluYb8*, and *AluYb9* based on diagnostic mutations present in each subgroup. The history of the most active of these subgroups, *AluYb8*, has been examined in a number of primate species.

The key results for *AluYb8* are shown in Figure 9.14 (Han et al. 2005). The founding element of the *AluYb* subfamily arose 18–25 million years ago, based on the absence of this subfamily in green monkeys but its presence in hominoids. However, the *AluYb* lineage remained quiescent for about 20 million years, but finally the *AluYb8* subgroup began a modest expansion in the common ancestor of the great apes. About 3–4 million years ago, the *AluYb8* elements "woke up" and underwent a major expansion in the human lineage, resulting in the large number of copies we see today. The mechanisms by which particular *Alu* subfamilies undergo periods of quiescence or active retrotransposition are very poorly understood.

The tempo of change in one type of transposable element can be very different from that of another kind. This point is illustrated in Figure 9.15 for

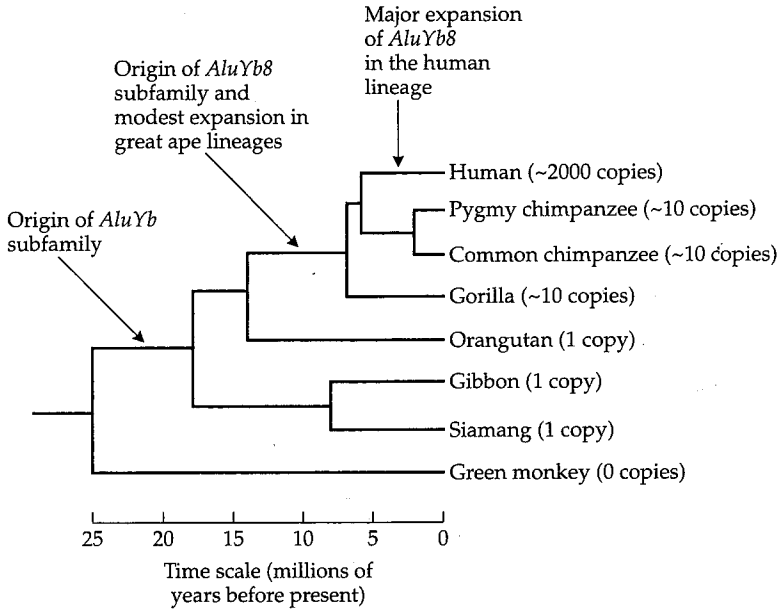


FIGURE 9.14 Origin and proliferation of the *AluYb8* subfamily in primates. Note the very large increase in copy number specifically in the human lineage. (From Han et al. 2005.)

Alu-type SINE elements and *L1*-type LINE elements in the human genome (International Human Genome Sequencing Consortium 2001). The *Alu*-type SINE elements show a peak divergence from the consensus sequence of about 7%. Since the rate of nucleotide substitution in these elements is roughly 0.3%–0.4% per million years (Li 1997), the peak at 7% divergence corresponds to a peak rate of transposition 18–23 million years ago. (The recent burst of transposition of *AluYb8* shown in Figure 9.14 is not apparent here, because the *AluYb8* subfamily accounts for only a small proportion of all *Alu*-type SINE elements.) The tempo of the *L1*-type LINE elements is very different, with peaks of divergence at about 4% and 21%, which correspond to times of approximately 10–13 million and 52–70 million years, respectively. Both types of transposable elements were abundant in the genomes of our evolutionary ancestors long before our species came into being.

Episodic bursts of transposition are also observed in *Drosophila*. There are a number of dramatic examples known as *hybrid dysgenesis*, in which particular transposable elements are mobilized in the progeny of crosses between males that carry the transposable element and females that do not carry it (Jensen et al. 1999, Castro and Carareto 2004, Blumenstiel and Hartl 2005).

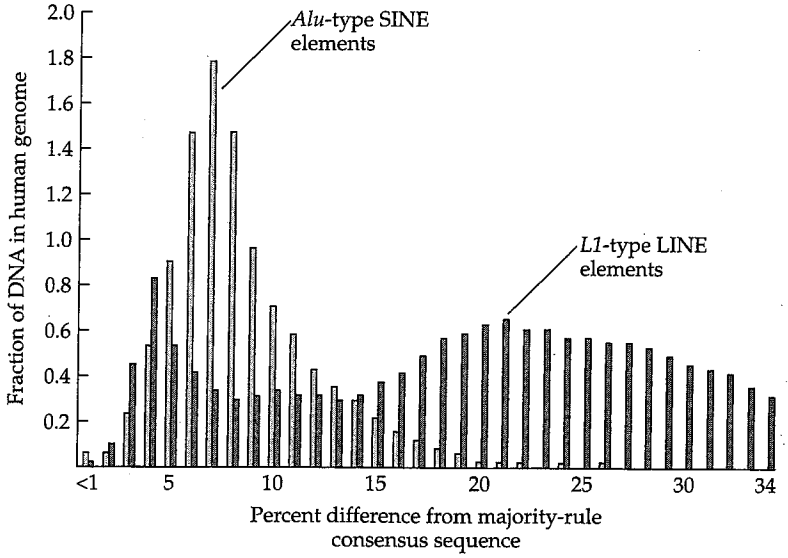


FIGURE 9.15 Evolutionary history of proliferation of *Alu*-type SINE elements and *L1*-type LINE elements in the human genome. In principle, the majority-rule consensus sequence approximates the sequence of each transposon at the time of its original insertion, and hence differences from the consensus sequence represent mutations that have become fixed since the time of insertion. On average, SINE and LINE elements undergo substitutions at the rate of about 0.3%–0.4% per million years, but this average may not have been constant throughout evolutionary time or in all transposon subfamilies. (Data from International Human Genome Sequencing Consortium 2001.)

Hybrid dysgenesis appears to result from lack of maternally transmitted factors that normally maintain the transposon in a quiescent state.

But sequence analysis of the sort described for *Alu* elements also implies changes in rates of transposition through time. Figure 9.16 shows data for 83 genomic copies of a non-LTR retrotransposon designated *jockey*, classified according to the estimates number of nucleotide substitutions since each element was inserted into its current location (Blumenstiel et al. 2002). More than 80 percent of the copies have fewer than 0.02 substitutions per base pair, whereas the others are much more divergent (0.08–0.16 substitutions/bp), and therefore much older. This bimodal distribution of ages is most easily explained by a burst of retrotransposition in the distant past (after which most of the insertions were eliminated by selection), then a period of quiescence, and then again a very recent burst of retrotransposition. Taking as the rate of nucleotide substitution in *Drosophila* as 1.5×10^{-8} substitutions/bp per

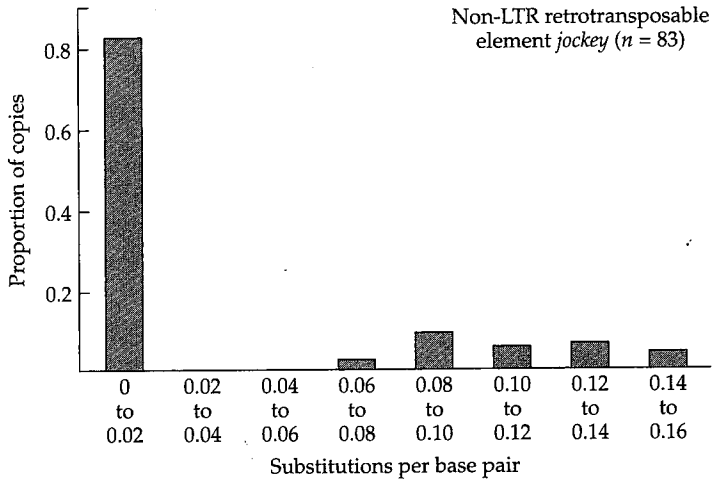


FIGURE 9.16 Recent burst of transposition of the transposable element *jockey* (a LINE element) in the genome of *D. melanogaster*. Remnants of an earlier burst (or bursts) of transposition are evident in the copies that show greater divergence from the consensus sequence. (Data from Blumenstiel et al. 2002.)

year (Rowan and Hunt 1991), old insertions with a divergence of, say, 0.15 substitutions/bp inserted approximately $0.15/1.5 \times 10^{-8} = 10$ million years ago, whereas newer insertions with a divergence of, say, 0.015 substitutions/bp were inserted approximately $0.015/1.5 \times 10^{-8} = 1$ million years ago. The former number is long before *D. melanogaster* diverged as a separate species (and so the burst of retrotransposition actually occurred in an ancestral species), whereas the latter postdates the emerge of *D. melanogaster* and so indicates a recent burst within the species itself (Blumenstiel et al. 2002).

Although it is not shown in Figure 9.16, there is another important difference between the ancient insertions and the recent insertions that reflects a pattern observed widely among transposable elements in *Drosophila* (Blumenstiel et al. 2002). The great majority of the recently inserted elements are polymorphic in the population and are present at sites in euchromatin (the unshaded regions of the chromosomes in Figure 9.4), whereas the great majority of the ancient insertions are fixed in the population and are present at sites in heterochromatin (the shaded regions around the centromeres in Figure 9.4). The main differences between euchromatin and heterochromatin are that the heterochromatin contains large amounts of highly repetitive DNA sequences, has much reduced levels of recombination, and contains very few functional genes. The likely causes of the discordance between recently and anciently inserted elements are that the recent elements insert nearly randomly in the genome, and therefore predominantly into euchromatin, and that most of these insertions are deleterious, either because of

effects of their own or because they can create chromosomal abnormalities by recombination with other copies elsewhere in the euchromatin (Petrov et al. 2003). The much smaller number of insertions into heterochromatin have a substantially greater chance of becoming fixed—first because of the low gene density and the unlikelihood of inserting in or near a functional gene, and second because of the enhanced probability of fixation due to background selection or selective sweeps (Blumenstiel et al. 2002).

Horizontal Transmission of Transposable Elements

Among the most widespread families of transposable elements is that of the *mariner*-like elements (MLEs), typified by the transposable element *mariner*. The molecular organization of the *mariner* element is illustrated in Figure 9.17A. The element is flanked by short (28 base pair) inverted repeats (IR) and includes a long open reading frame coding for the transposase protein

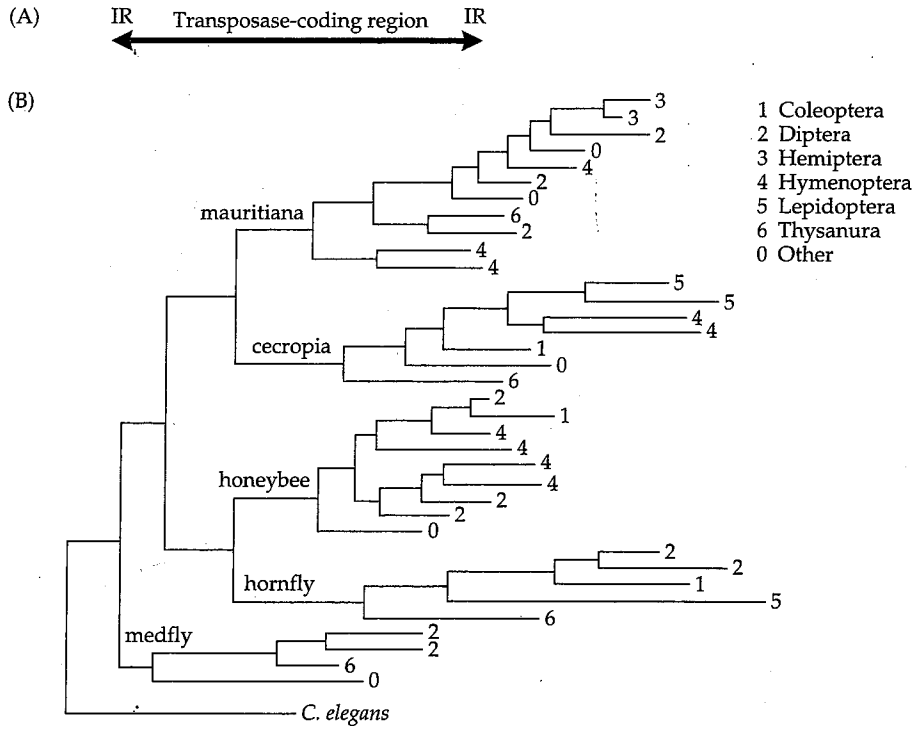


FIGURE 9.17 (A) The molecular organization of the transposable element *mariner* showing the inverted repeats flanking the transposase-coding region. (B) Distribution of MLEs among species representing major insect orders (numbered). Note that the MLEs can be grouped into subfamilies of elements (mauritiana, cecropia, and so forth) based on their similarity in sequence. *C. elegans* is the soil nematode *Caenorhabditis elegans*. (Data for B from Robertson 1993.)

(Hartl 2001). Insertion of the element is invariably adjacent to a 5′–TA–3′ dinucleotide in the host genome and is accompanied by a duplication of the dinucleotide, so that the inserted *mariner* is flanked by 5′–TA–3′. The target sequence and dinucleotide, as well as features of the amino acid sequence of the transposase protein, identify a transposable element as an MLE.

MLEs are widely distributed among insects and other invertebrates (Robertson 1993; Robertson and MacLeod 1993). Figure 9.17B shows the distribution among species in the major insect orders (Coleoptera, Diptera, and so forth). The number of copies of an MLE per genome varies widely among species, ranging from a few copies to many thousands. The MLEs in Figure 9.17B have been grouped according to similarity in nucleotide sequence and arranged in the form of a tree with the root to the left and the tips of the branches to the right. There are several subfamilies of insect MLEs, denoted mauritiana, cecropia, honeybee, and so forth. MLEs in different subfamilies are typically 40–50% identical in nucleotide sequence, and those within the same subfamily are usually 60% or more identical. All of the insect MLEs are more closely related to each other than they are to an MLE found in the soil nematode *Caenorhabditis elegans*.

Although MLEs are widespread, their distribution is "spotty," which means that, among closely related species, a particular subfamily of MLE may be found in some species but not in others. Furthermore:

- Any species may contain MLEs from two or more different subfamilies.
- Closely related MLEs are often found in distantly related species.

An example of the second principle is an MLE found in *Drosophila erecta*, a close relative of *D. melanogaster*, which is 97% identical in nucleotide sequence with an MLE found in the cat flea *Ctenocephalides felis* (Lohe et al. 1995). For comparison, a gene coding for a subunit of the cellular sodium pump sequenced in both species shows only 39% nucleotide identity at third codon positions.

What process can account for the virtual identity between MLEs in species as distantly related as a fruit fly and a cat flea? One possibility is that the MLE was present in the common ancestor of the species a few hundred million years ago and then virtually stopped evolving, so that the sequences remain almost identical today. Unless the nucleotide sequence is very highly constrained, including third codon positions, this possibility is very unlikely. Furthermore, if MLE sequences are so constrained, then why is there so much sequence variability within and among subfamilies? More likely than the idea that evolution stopped dead in its tracks for several hundred million years is the hypothesis of **horizontal transmission**, or the ability of an MLE to be transferred from a host species into the germline of a different, reproductively isolated species. (In bacteria, horizonal transmission is often called *lateral gene transfer*.) To account for the *D. erecta*–*C. felis* case by horizontal transmission, an MLE would have to have been transmitted from a *D. erecta*

ancestor to a *C. felis* ancestor (or the other way around) approximately 3–10 million years ago. Many additional examples of horizontal transmission of MLEs and other eukaryotic transposable elements have been discovered (Silva et al. 2004; Le Rouzic and Capy 2005). Although the process of horizontal transmission certainly takes place, the rate at which it happens and the vectors and mechanisms are as yet unknown.

Once introduced into a genome, MLEs can persist through multiple speciation events (Maruyama and Hartl 1991; Hartl et al. 1997). A lineage can, however, lose an MLE, as evidenced by *D. melanogaster*, which has lost an MLE (the *mariner* element itself) present in all its closest relatives. Two processes appear to contribute to loss of an MLE: (1) mutational inactivation, which may destroy the protein-coding function of an MLE or impair its ability to transpose; and (2) stochastic loss, by which we mean the elimination of an MLE from the genome as a result of random genetic drift (Figure 9.18). There might possibly also be a contribution from natural selection, depending on the extent to which presence of the MLE itself is deleterious. From the standpoint of the host species, an inactivating mutation in an MLE may be selectively neutral, or perhaps even favorable, inasmuch as natural selection may act to minimize the harmful mutagenic effects of transposition. Subsequent mutations in an already inactivated MLE are presumably selectively neutral and the element ultimately lost by chance. The role of mutational inactivation and stochastic loss in the evolutionary dynamics of MLEs is supported by the spotty distribution of MLEs among closely related species.

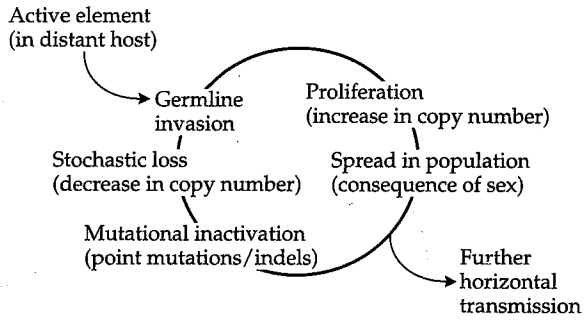


FIGURE 9.18 Evolutionary natural history of *mariner*-like elements. Rare horizontal transmission introduces the element into the genome of a species, where it may be lost by chance or proliferate in copy number. Sexual reproduction enables the element to spread among all members of the species. The increase in copy number is offset by mutations that inactivate the transposase or its recognition sites, and ultimately the copy number may decrease and the element be lost. By that time the element may have been transferred horizontally to yet another species, where the cycle starts anew. The term *indel* refers to small insertions or deletions. (After Hartl et al. 1997.)

SUMMARY

1. Population genomics applies the principles of population genetics on a genome-wide scale to understand the evolutionary forces that shape genome organization and the genes themselves.
2. Genomes differ greatly in size. In eukaryotes, the C-value paradox refers to the lack of a clear relation between the size of the genome and the metabolic, developmental, or behavioral complexity of the organism. Genomes also differ greatly in base composition. The genomes of mammals and birds are characterized by long (> 300 kb) stretches of relatively G+C-rich or G+C-poor regions known as isochores.
3. Many factors affect levels of polymorphism in the genome. Some species have abundant polymorphisms in the subtelomeric regions, where they have copies of multigene families that can expand by unequal crossing over and diversify by gene conversion.
4. In some organisms, the level of polymorphism increases with the local frequency of recombination. In *Drosophila*, this association appears to result from either background selection or recurrent selective sweeps, rather than to differences in the rate of mutation.
5. A selective sweep, if strong and recent enough, can be detected through its effect on the allele frequency spectrum and the haplotype structure in the population. Selectively favored haplotypes are expected to be more frequent and less polymorphic than expected from neutral mutation and random genetic drift.
6. Statistical tests for selection gain substantially in power when applied to large numbers of genes across the genome. Such tests include comparisons of nonsynomymous and synonymous divergence (d_N/d_S ratios) and analysis of polymorphism and divergence among related species. These analyses reveal that many amino acid replacements are deleterious, and that many genes that differ in sequence among species show evidence for positive selection. Compensatory changes in amino acid sequence have also been documented.
7. In some cases the molecular basis of selection has been identified by correlating differences in protein structure and function with changes in sequence. Examples include the oxygen carrying protein hemoglobin and the visual protein opsin.
8. Population genomics has opened up new opportunities for molecular evolutionary studies of sexual selection and the identification of the genes involved in the origin of reproductive isolation between species. Genes that function in male reproduction evolve rapidly, as do many X-linked genes. The genetic basis of Haldane's rule concerning reproductive incompatibilities in the heterogamatic sex has also been elucidated.
9. Transposable elements are present in most genomes, and in some they are particularly abundant. Almost 50 percent of the human genome con-

sists of transposable elements, many of which have lost the ability to be mobilized.

10. The population dynamics of transposable elements are governed by the rate of transposition, which is often regulated either by the transposon itself or by the host genome, balanced against a usually modest fitness cost resulting from the transposon. In bacteria, genes conferring resistance to antibiotics are often located in transposons present in transmissible plasmids; this location enables them to spread widely and quickly in response to selection.
11. Many transposable elements do not exhibit a constant rate of transposition through time but exhibit bursts of transposition during which they are exceptionally active. The occurrence and timing of these bursts can be identified by sequence comparisons among copies of the transposon present in a single genome.
12. The *mariner* family of transposable elements (*mariner*-like elements, or MLEs) represents a group of transposons that move via a cut-and-paste mechanism. MLEs are highly diverse, and members of multiple subfamilies are present in a significant proportion of animal genomes, including the human genome. Distantly related genomes often contain closely related MLEs, providing strong evidence for horizontal transmission.

PROBLEMS

1. What is meant by the term C-value paradox? Among animal genomes, the smallest known is that of the placozoan *Trichoplax adhaerence* at 0.04 Gb, and the largest known is that of the marbled lungfish *Protopterus aethiopicus* at 133 Gb. (A gigabase, abbreviated Gb, equals a billion nucleotide pairs.) What is the ratio in size of the largest to the smallest animal genome? The human genome is 3 Gb. Is its size closest to the arithmetic, geometric, or harmonic mean of the other two? (The geometric mean is the square root of the product; the harmonic mean is the reciprocal of the mean of the reciprocals.)
2. The genome of a species of bacteria undergoes GC to AT mutations at a rate of $\mu = 1 \times 10^{-9}$ per generation, and AT to GC mutations at a rate of $\nu = 2 \times 10^{-9}$ per generation. For DNA sequences that are not constrained by selection, what is the percent GC expected at equilibrium? (The evolutionary implications of reversible mutation are discussed in Chapter 4.)
3. What are isochores and why was their discovery unexpected?
4. In a random sequence of n independent trials with probability of success equal to p in each trial, the expected proportion of successes is p and the standard deviation in the proportion of successes equals $\sqrt{[p(1-p)/n]}$. Across the human genome, the GC content averages $p = 38.4\%$. If an iso-

chore of 300 kb has a GC content of 35%, by how many standard deviations does this isochore deviate from the value expected if adjacent nucleotides evolved independently? Answer the same question for an isochore of 300 kb with a GC content of 60%.

5. What is background selection? Why does background selection reduce the genetic diversity in a region of the genome to a level smaller than that expected under the neutral theory?

6. Equation 9.1 asserts that the nucleotide diversity π across a region of the genome, relative to that expected with neutrality π_0, is approximately equal to $\exp[-U/(2hs + R)]$. Explain what each of the parameters in this expression refers to, and describe what effect their magnitudes have on the level of nucleotide diversity.

7. What is Haldane's rule in regard to hybrid incompatibility? Why would recessive mutations associated with hybrid sterility or lethality contribute to Haldane's rule?

8. Recall from Chapter 2 that, in testing for linkage disequilibrium between pairs of alleles at two loci, the value of chi-square is numerically equal to r^2n, where n is the total number of chromosomes in the sample. This chi-square has one degree of freedom, and for significance at the 5% level a value of 3.84 or greater is needed. Equation 9.2 gives the expected value of r^2 in a population of effective size N, where c is the frequency of recombination between the genes. For samples of size 20, 40, and 80, what minimum value of r^2 is needed for significance at the 5% level, and what frequencies of recombination c do these values correspond to when $N = 10^4$? What distances in nucleotide pairs do these values of c correspond to in the human genome, in which there is roughly 1% recombination per megabase of DNA?

9. Consider a population of animals in which there are two types of Y chromosomes, Y and Y′. Males of genotype XY′ exhibit meiotic drive and produce X-bearing and Y-bearing gametes in the proportions $1 - k : k$ (with $k > \frac{1}{2}$), whereas males of genotype XY produce X-bearing and Y-bearing gametes in the normal proportions of 1 : 1. If XY′ and XY males have relative frequencies p and q, respectively, where $p + q = 1$, what is the expected frequency of XY′ males in the next generation? What is the equilibrium frequency of XY′ males, and the sex ratio in the population at equilibrium?

10. Meiotic drive (segregation distortion) is a potent force for changing allele frequencies. This principle can be readily appreciated by considering the change in allele frequency when an allele favored by meiotic drive is rare. For this purpose, consider a random-mating population with genotype frequencies p^2, $2pq$, and q^2 for AA, Aa, and aa, respectively, with $p + q = 1$. If the heterozygous Aa genotype produces a ratio of $A : a$ gametes

of $k : 1 - k$, what is the change in allele frequency Δp in one generation? Compare this to the change in allele frequency expected with Mendelian segregation when the relative fitnesses of *AA*, *Aa*, and *aa* are given by $1 + 2s$, $1 + s$, and 1. What is the equivalent value of s in terms of k when $p \approx 0$, and what are the specific values of s when $k = \frac{2}{3}$, $\frac{3}{4}$, and $\frac{4}{5}$?

11. If a bacterial insertion sequence has the population dynamics described in Equations 9.3 with $\alpha = \frac{1}{3}$ and $\phi = \frac{1}{2}$, what are the expected proportions of bacterial cells containing 0, 1, 2, 3, 4, and 5 IS elements?

12. The parameters in Equations 9.3 that best account for the copy numbers of the bacterial insertion sequences IS3 and IS5 among natural isolates of *E. coli* are $\alpha = \frac{1}{3}$ and $\phi = \frac{3}{5}$ for IS3, and $\alpha = \phi = \frac{2}{3}$ for IS5. Among 71 natural isolates, the observed copy numbers were as follows

	Copy number					
	0	1	2	3	4	≥ 5
IS3	23	10	19	10	6	3
IS5	46	12	3	2	2	6

(Sawyer et al. 1987). Use the equations to calculate the expected numbers in each category, and carry out a chi-square test for goodness of fit. Each chi-square test has 3 degrees of freedom because two parameters (α and ϕ) were estimated from the data.

13. For a bacterial insertion sequence whose population dynamics are as described in Equations 9.3, for any given value of α, what value of ϕ is required to make the expected proportion of cells with no copies equal to the expected proportion of cells with exactly one copy?

14. This problem explores some of the implications of Equation 9.4 that the average fitness of an individual with n copies of a transposable element has a relative fitness $w(n)$ equal to $w(n) = \exp[-(u - v)n]$. Consider a transposable element in *Drosophila* for which $(u - v) = 2 \times 10^{-4}$ that has an average number of copies per diploid genome of 49. What is the relative fitness of an individual whose genome contains the average number of copies?

15. Using the information in the previous problem, suppose that the distribution of copy number among diploid genomes is a Poisson distribution in which the variance is equal to the mean. Using the equation $w(n) = \exp[-(u - v)n]$ for $(u - v) = 2 \times 10^{-4}$, what is the predicted average fitness of an individual with a copy number equal to the mean plus 5 standard deviations? What is the predicted average fitness of an individual with a copy number equal to the mean minus 5 standard deviations? Relative to the average fitness of an individual with the mean number of copies, what is the percent difference in fitness between the plus-5 and minus-5 genotypes?

10.1 Human Polymorphism 520

Public SNP Resources and the HapMap Project 521

10.2 Population Genetic Inferences from Human SNPs 524

Ascertainment Bias of SNP Genotypes 524

Departures from Hardy-Weinberg Frequencies 527

Site Frequency Spectrum and Human Population Growth 527

Rooting Human Polymorphism 529

Inference of Inhomogeneities in the Mutation Process 530

Inferences about Male and Female Mutation Rates 531

10.3 Linkage Disequilibrium across the Human Genome 532

The Landscape of Human Linkage Disequilibrium 532

Inferences about Local Rates of Recombination 537

10.4 Population Structure Inferred from Human Polymorphism 539

Multilocus Methods of Inference of Stratification 540

Heterogeneity in Linkage Disequilibrium across Human Populations 542

Linkage Disequilibrium in Admixed Populations: Admixture Mapping 544

Inbred Populations and Homozygosity Mapping 544

10.5 Mendelian Disease and Population Genetics 547

Mutation-Selection Balance 548

Dating the Origin of Mutant Alleles 548

10.6 Genetic Basis for Variation in Risk of Complex Disease 550

Mapping Methods Based on Linkage 550

Linkage Disequilibrium Mapping 552

Genome-Wide Association Studies 553

10.7 Seeking Signatures of Human-Specific Genetic Adaptations 554

Interspecific Divergence 554

McDonald-Kreitman and Poisson Random Field Tests 555

Local Distortions in Linkage Disequilibrium 556

F_{ST} Tests 557

Genome Scans for Selection—Skewed Site Frequency Spectrum 557

10.8 Human Origins 558

Neanderthal Genome Sequence 562

CHAPTER 10

HUMAN POPULATION GENETICS

Human population genetics has been a lively and engaging field since the time when the only genetic markers were blood groups and a few protein polymorphisms. Back then it was intriguing to see how patterns of polymorphisms were distributed across human cultures, often reflecting major continental divisions and past waves of migration. There was a confidence that increasing amounts of genetic data would make these inferences more robust and accurate, but nobody guessed that we would be submerged in the oceans of data available today. Human population genetics has undergone one of the most dramatic expansions in all of science in the last five years, initially spurred by the Human Genome Project and its attendant volume of data about human genes and their polymorphism. But a major stimulus was also the idea that genes responsible for many genetic diseases might be mapped simply by their observed statistical association with genotypes at anonymous polymorphisms across the genome. (An **anonymous polymorphism** is one with no known function.) This idea of mapping by association gave rise to an effort to detect, map, and characterize human polymorphisms on a scale that dwarfs those in model organisms. Today more than 10 million single nucleotide polymorphisms in the human genome have been identified, and methods have been developed to score the genotypes of one million SNPs in an individual in less than a day. The sheer volume of data has given rise to a series of questions that initially were unique to human

population genetics, but soon presented themselves in other organisms as well. A small sampling of such questions includes:

- Why are there regions of low and regions of high SNP density?
- Are the patterns of nucleotide substitutions homogeneous across the genome?
- How should one interpret the *P*-values from 500,000 tests of fit to Hardy-Weinberg proportions?
- Why do human polymorphism data appear to be so far from a steady-state balance between mutation and random genetic drift?
- In nonequilibrium situations, how can population demographics be incorporated into standard population genetic tests, such as inference of departures from neutrality?
- What are the consequences of inhomogeneities in recombination rate?

10.1 HUMAN POLYMORPHISM

Initial efforts to characterize human polymorphism at the DNA level suggested that a pair of human alleles differ at a level ranging from one base in 500 to one base in 10,000 (Figure 10.1). Pooling estimates and additional sequence data, Li and Sadler (1991) came to an estimate of average nucleotide diversity of 0.11% for four-fold degenerate sites, a figure that is an order of magnitude lower than that for *Drosophila*. (Recall that a *fourfold degenerate site* is a site at the third position in a codon at which all nucleotide substitutions are synonymous.) Early on it was recognized that sequence diversity varied widely among human genes. At first, complete DNA sequencing of a gene or region from multiple individuals by PCR amplification and sequencing became the method of choice to characterize polymorphism in a region. The beta globin gene was one of the first targets, and data from this region displayed a level of nucleotide diversity of 0.14%, as well as considerable intragenic linkage disequilibrium (Harding et al. 1997). The gene for lipoprotein lipase (LPL) presented another facet of the complexity of the variation in human genes, yielding 88 segregating sites in just 10 kb across 71 individuals (Clark et al. 1998). The pattern of linkage disequilibrium did not immediately suggest large blocks of unrecombining haplotypes, but even this early study showed a signature suggestive of a recombination hotspot. This type of resequencing has been carried out on a large scale, and dense resequence data are available for more than 250 genes (http://droog.mbt.washington.edu/).

To understand polymorphism on a genomic scale would require thousands of such resequencing projects. A different strategy is therefore needed to identify SNPs across the whole genome. At first this was possible only in a very small sample. In the first whole-genome shotgun sequence of the human genome, the sequence came from five different people (Venter et al.

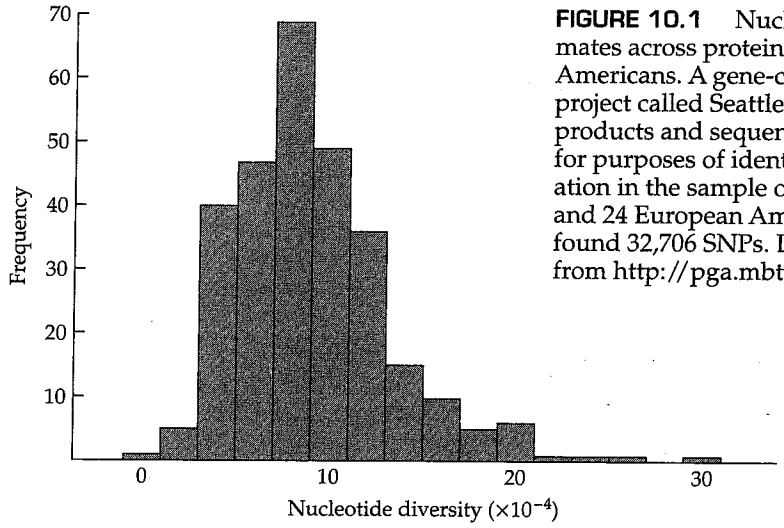


FIGURE 10.1 Nucleotide diversity estimates across protein-coding genes in African Americans. A gene-centered resequencing project called Seattle SNPs has obtained PCR products and sequenced them in 286 genes for purposes of identifying all sequence variation in the sample of 24 African Americans and 24 European Americans. So far they have found 32,706 SNPs. Data are freely available from http://pga.mbt.washington.edu/.

2001). When the total of 28 million sequence reads was aligned and assembled, the genetic differences among the five individuals became clear. It was a remarkably efficient way to discover SNPs, yielding about 4 million SNPs during genome assembly at no extra cost! Those data yielded a genome-wide average nucleotide diversity of 0.0008, and they also showed remarkable heterogeneity in SNP density across the genome (Figure 10.2).

Public SNP Resources and the HapMap Project

Before the human genome sequence was completed, it was already clear that standard methods for studying genetic linkage based on pedigrees could never accumulate sample sizes sufficient to map with a resolution finer than about 10 cM, which is physically equivalent to about 10 Mb (10 million base pairs of DNA). Because 10 Mb typically includes many genes, a mapping method with finer resolution was needed. If allelic variants that cause disease are in linkage disequilibrium with nearby anonymous SNPs, then simply by testing the statistical association between SNP alleles and health status, one could in principle map to the resolution determined by the extent of local linkage disequilibrium (Risch and Merikangas 1996). Furthermore, some studies were showing that human SNPs not only exhibited local patterns of linkage disequilibrium but also a tendency for certain haplotypes with multiple SNPs to have a relatively high frequency. These findings motivated a need for a survey of the human genome for its patterns of SNP variation, particularly with regard to linkage disequilibrium.

The International HapMap project was designed to identify SNP genotypes at sufficient density to determine the extent of local linkage disequilib-

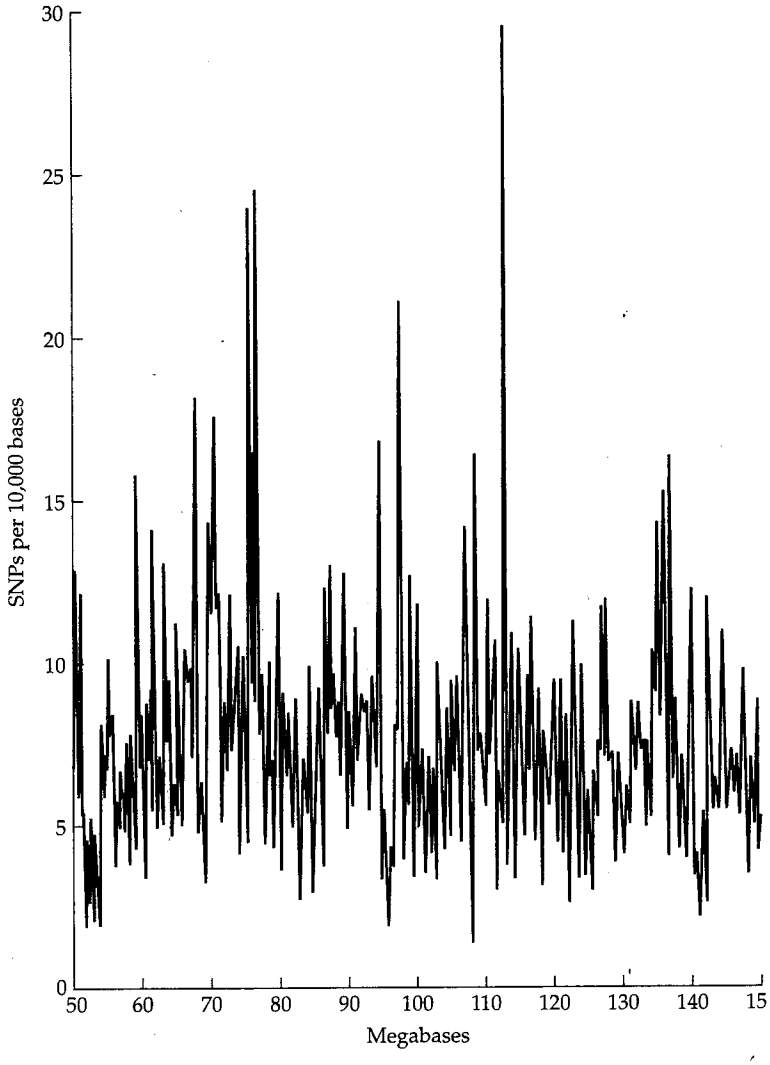


FIGURE 10.2 This plot shows the density of SNPs discovered by resequencing across a 100-Mb region of chromosome 22. Notice that some regions show much greater SNP density than others. (From Patil et al. 2001.)

rium in three different human populations. The samples came from an African population from Nigeria (Yoruban), a mostly Utah (U.S.) population of European ancestry analyzed by the Centre d'Étude du Polymorphism Humaine (CEPH) for genetic linkage, and a sample drawn from the Japanese and Han Chinese populations. The DNA samples from 90 Yorubans, 90 CEPH, 45 Chinese and 45 Japanese were distributed among 11 centers for the genotyping. There was a long and careful process of quality control on sever-

al technologies (called "platforms") for SNP genotyping. One byproduct of the project was to drive down the cost of highly accurate SNP genotyping. At first the number of known SNPs was insufficient to complete the project, and it became necessary to perform additional shotgun sequencing from different individuals to identify additional SNPs. With impressive coordination, the various centers delivered SNP genotypes on schedule (The International HapMap Consortium, 2005). The initial analysis of the 1.1 million SNP genotypes in the 270 human samples produced a wealth of information about the patterns of linkage disequilibrium in human populations (see Section 10.3). As the project wound down it became clear that the cost of SNP genotyping had fallen so quickly that an additional 4 million SNPs could be produced at little extra cost. So a new goal was set, and in short order an addition 4 million SNPs were genotyped. Altogether the HapMap project has identified the polymorphic nucleotides at 6.1 million SNPs. This information is presented in various ways, and the entire data set is available from the website www.HapMap.org. This represents an example of the extraordinary power of international scientific cooperation and collaboration; and the data have already established the foundation for modern human population genetics.

In addition to the HapMap database, the public dbSNP database (www.dbsnp.org) is a repository of essentially all known human SNPs, and it also includes information on insertion/deletion polymorphisms as well as population attributes such as SNP frequencies.

The SNPs used in the HapMap project were discovered in a variety of ways and by using a range of sample sizes during the discovery process. To detect whether such variation may seriously bias the analysis, and to ascertain the added value of an even higher density of SNPs, 10 regions of the genome were fully resequenced in 16 individuals from each of the 3 populations. The genomic regions were chosen from the ENCODE project, an effort funded by the National Institutes of Health (NIH) to analyze in detail a 1% portion of the human genome in order to identify all functional elements in the targeted regions. The ENCODE resequencing that was done for confirmation of the HapMap project also served as a source of additional human SNPs.

More recently it has become clear that insertion/deletion (indel) polymorphisms are underrepresented in the HapMap project because, for technical reasons, most SNP genotyping platforms do not reliably detect indels, but with clever analysis indels can be inferred (McCarroll et al. 2006). Several studies had already demonstrated the medical importance of indels (Sharp et al. 2006), and because resquencing projects had typically identified on the order of one indel for every 10 SNPs, there emerged a need for systematic identification of indels and other types of structural polymorphism (Redon et al. 2006). The discovery and cataloging of human genetic polymorphisms is of ongoing importance. To a population geneticist, it is reassuring to see how rapidly the research community came to realize that after obtaining the complete human genome sequence, the remaining big challenges included many

problems that center on population genetics and understanding the genetic basis of differences among individuals.

10.2 POPULATION GENETIC INFERENCES FROM HUMAN SNPS

Most of the theory developed in molecular population genetics assumes that the data consist of the complete sequence of each allele sampled. Human genetic data are almost never in this form. At best, such as when the DNA sequence of PCR products is accessible, one can determine the genotypes at positions that are polymorphic, but the linkage phases of sets of such polymorphic sites are ambiguous. That is, it is usually not known whether a double heterozygote has the genotype AB/ab or Ab/aB. We refer to such data as unknown **linkage phase**. If one is interested only in features of each individual SNP, such as its fit to Hardy-Weinberg proportions, such unphased data are adequate. However, often one wants to know about **haplotypes**, or runs of SNP alleles present at multiple sites along a chromosome. As we will see below, statistical methods exist to estimate the linkage phase, but obtaining solid molecular data for linkage phase is difficult. One approach is allele-specific PCR, another makes use of hydatidiform moles, and still another is based on human-rodent hybrid cells bearing only a single human chromosome. (A *hydatidiform mole* is growth of an abnormal fertilized egg that contains only paternal chromosomes, and hence has the genotype of a single human gamete.)

Ascertainment Bias of SNP Genotypes

Human genotypes are not usually obtained from complete DNA sequences, but from genotyping a collection of SNPs that are already known to be polymorphic in the human population. Focusing on DNA positions already known to be polymorphic is orders of magnitude faster and less expensive than full sequencing. But as mentioned above, SNP genotypes are obtained in two stages. In stage one, DNA from more than one individual is sequenced to discover SNPs. Depending on the criteria, generally some subset of the SNPs are moved forward into the second phase which is the actual genotyping. For the HapMap Project, the criteria for moving a SNP forward changed several times in midstream. For most of the project, the criterion was to move forward only "double hit" SNPs, or those whose least frequent allele was seen at least twice in the sample. This obviously biases the sample away from rare SNPs. SNPs that are rare in the population will be underrepresented, whereas common SNPs will have a high chance of staying in the study and being included in the genotyping platform.

The ascertainment bias of SNP genotypes is crucial because virtually all statistical tests in molecular population genetics rely on data obtained from complete resequencing of the targeted region or regions in all individuals. SNP genotypes violate this assumption, and the magnitude of the departure

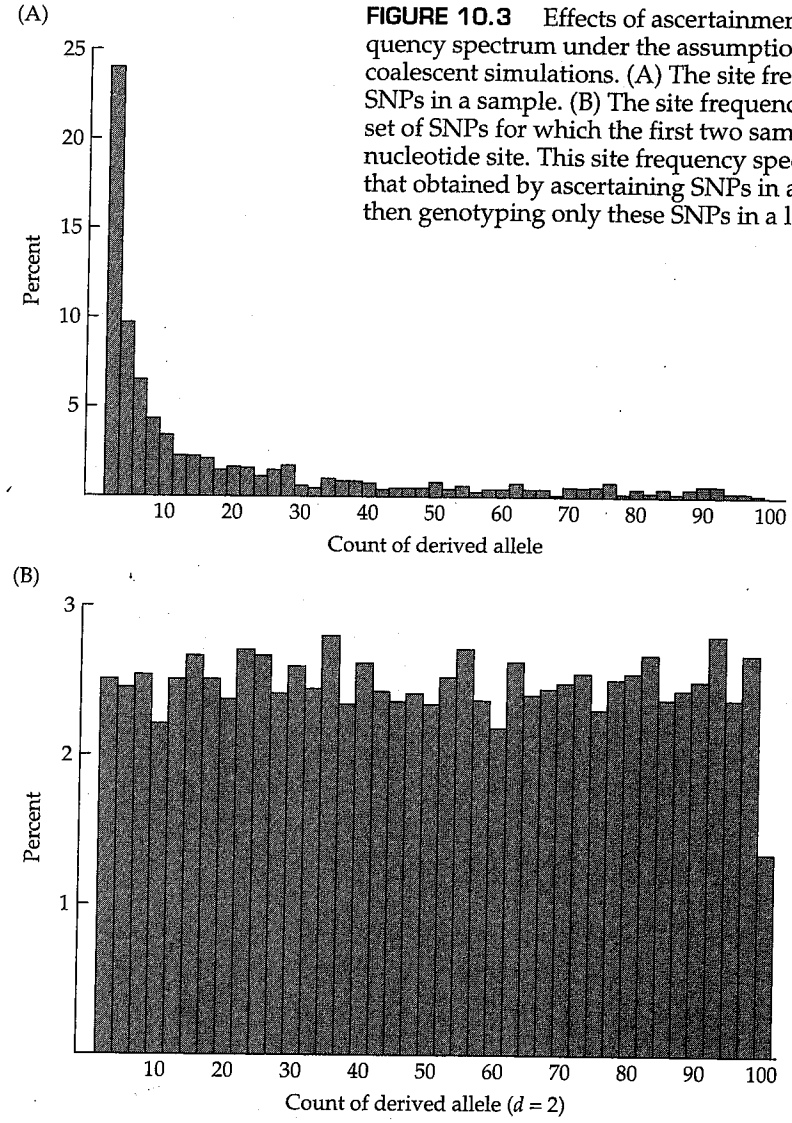


FIGURE 10.3 Effects of ascertainment bias on the site frequency spectrum under the assumption of neutrality based on coalescent simulations. (A) The site frequency spectrum for all SNPs in a sample. (B) The site frequency spectrum of the subset of SNPs for which the first two samples differ at the nucleotide site. This site frequency spectrum is equivalent to that obtained by ascertaining SNPs in a sample of size 2 and then genotyping only these SNPs in a larger sample.

depends on the degree of ascertainment bias. Perhaps the easiest way to see the impact of the ascertainment bias is to consider the **site frequency spectrum** (the distribution of allele frequencies in a sample). If a SNP has a true frequency of p in the population, and we sample only two copies in the SNP discovery phase, then the chance of discovering the SNP is the chance of drawing both distinct alleles in this sample of two, or simply $2pq$. Figure 10.3A shows the site frequency spectrum expected for a neutral gene in the

absence of ascertainment bias. In panel B the site frequency spectrum is plotted, assuming a discovery panel of two in the SNP discovery phase, with only those SNPs genotyped in a larger sample.

Notice that the site frequency spectrum in panel A has the expected appearance for a gene in mutation-drift equilibrium, whereas the site frequency spectrum in panel B has a severe under-sampling of rare alleles and over-sampling of common alleles. The fact that the final distribution appears to be uniform when the discovery sample has just two gene copies arises because the expected site frequency for a SNP whose allele frequency is j copies of the derived allele decreases as $1/j$ (based on the infinite-sites model). The probability that such a SNP is discovered in the discovery sample is also $1/j$, and so the factors precisely cancel!

If one knows the process by which SNPs were discovered, then it is possible to correct for the ascertainment bias. The correction procedures are beyond the scope of this book, but the basic idea is that each site frequency class gets corrected by a weighting equal to the probability of discovery of a SNP in that class (Nielsen et al. 2005a). While ascertainment bias may cause serious problems for some inferences from SNP genotype data, for estimates of other parameters such as the population recombination rate, $\rho = 4Nc$, the impact of ascertainment bias may be small. Most worrisome is the use of SNP genotype data to make inferences about population structure, since the SNP discovery process was not carried out in all human populations.

Nevertheless there is great interest in applying SNP genotyping methods to a wide range of human populations. One solution is to accompany these studies by at least some resequencing to assess the effects of the SNP ascertainment bias.

PROBLEM 10.1 Consider a study that began by discovering 100 SNPs from an initial discovery panel that consisted of a sample of 2 chromosomes. This discovery process identified 100 SNPs which were then genotyped in 1000 individuals. The genotype data reveals 10 SNPs, with a minor allele frequency between 0.001 and 0.05, and 10 SNPs whose minor allele frequency is between 0.45 and 0.5. This sample makes it appear that there was an equal number of SNPs with a rare minor allele (< 0.05) as there were SNPs with a common minor allele (0.45–0.5). Use a simple ascertainment correction to estimate the true ratio of rare to common SNPs in the population.

ANSWER The probability of ascertaining a SNP whose frequency is 0.05 in a sample of two is simply the chance of drawing two alleles that are different. This is $2pq = 2(0.05)(0.95) = 0.095$. For the common allele, the probability of ascertainment is $2pq = 2(0.5)(0.5) = 0.5$. In other words, we are $0.5/0.095 = 5.26$ times as likely to discover the common SNP as the rare one. Because the observed count of SNPs in these two classes is equal, there must have been 5.26 times as many rare SNPs as common ones in the population.

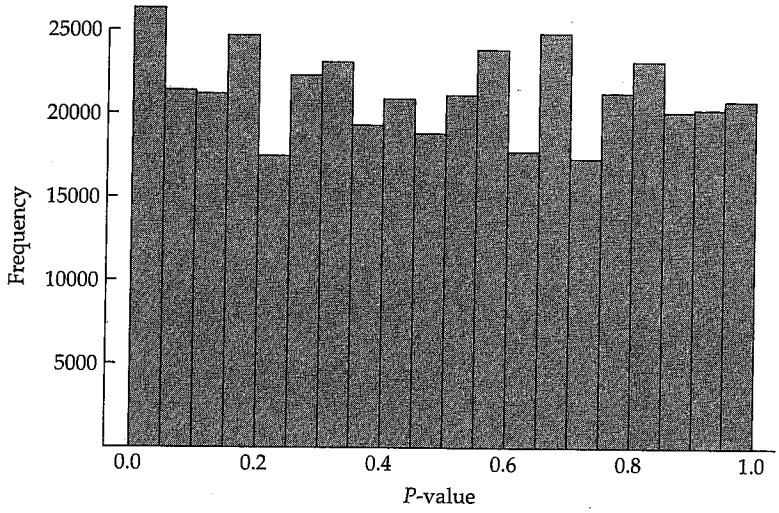


FIGURE 10.4 The international HapMap Project website (www.hapmap.org) provides genotype data for more than 4 million SNPs as well as tests of Hardy-Weinberg equilibrium. The plot shows a histogram of the *P*-values for the test of goodness of fit to Hardy-Weinberg expectations, indicating a good fit to the expected uniform distribution.

Departures from Hardy-Weinberg Frequencies

Suppose you carried out a Fisher exact test of genotype counts for 500,000 SNPs to test for goodness-of-fit to Hardy Weinberg proportions. Even if every SNP conformed to the assumptions of the Hardy-Weinberg principle, you would find that about 25,000 SNPs yield a *P*-value less than 0.05. Even with a stricter limit of *P* less than 0.0001, there would be 50 SNPs that fail to satisfy Hardy Weinberg proportions with what looks like high confidence. If there was one SNP that truly violated Hardy-Weinberg proportions, and its *P*-value was 0.001, there would be 500 false positives with a lower *P*-value than this true positive. This problem of multiple testing appears frequently in population genomics, and some methods for dealing with it, such as the false discovery rate, were already discussed in Chapter 2. Graphical methods are often very useful in understanding the behavior of statistical tests with such large samples (Figure 10.4).

Site Frequency Spectrum and Human Population Growth

Figure 10.5 shows a histogram of the expected site frequency spectrum under the standard neutral model, and it also shows the site frequency spectrum from the fully sequenced data of a SNP project. A clear departure is evident in the rarest minor allele frequencies (so called singletons and doubletons).

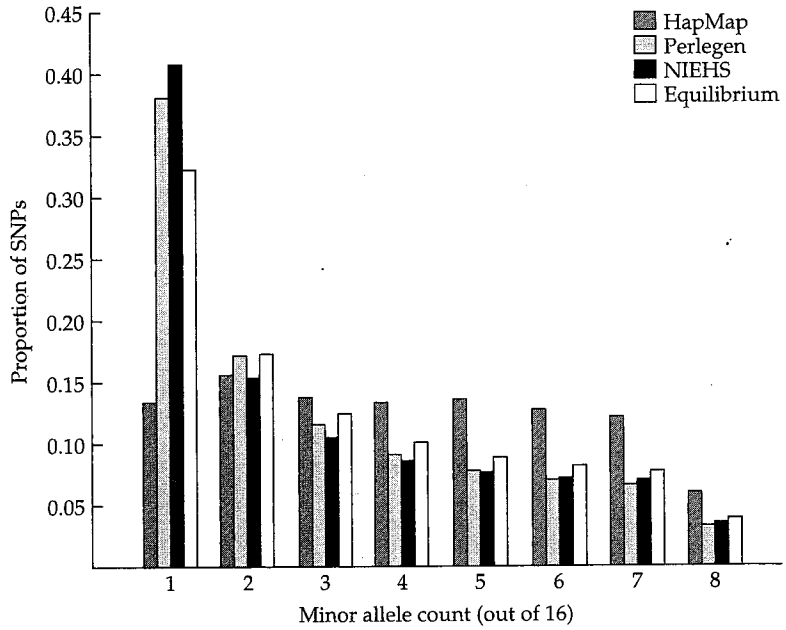


FIGURE 10.5 Site frequency spectra are compared to the equilibrium neutral model, including data from Perlegen Biosciences (resequencing by hybridization), the HapMap Project Phase I (genotyping of previously ascertained SNPs) and a resequencing project of the National Institute of Environmental Health Sciences (NIEHS). The strong departure of the HapMap Project is indicative of the ascertainment bias illustrated in Figure 10.3. Subsequent phases of HapMap data largely avoided this bias. (Courtesy Scott Williamson.)

The human data have too many rare alleles. In terms of commonly used population genetic statistics, Tajima's D is negative. This is true for almost any portion of the genome and almost any class of SNPs. Given the near-universal nature of this pattern of excess rare SNPs, a systematic bias is the most likely explanation. One obvious departure from the assumptions of the standard neutral model is that the human population has not been at a steady state in size for the past $4N_e$ generations. The rapid, recent growth of the population is expected to produce coalescent trees in which the alleles at the tips of the tree come from a very large population, and so the initial rate of coalescence is low, but as one progresses back in time the population shrinks, making coalescences occur faster relative to a population with a constant size. Population growth therefore results in an excess of mutations in the tips compared to populations with a constant size, and mutations that occur on the terminal branches will be singletons.

Given the ubiquity of the signature of human population growth in the site frequency spectrum, it seems that the standard neutral coalescent model

should be upgraded to a model that incorporates population growth. As it happens, including additional parameters for human demography yields a much better fit to the site frequency spectrum than does the standard model. Marth et al. (1999) fitted a series of different demographic models to fourfold degenerate sites, and found that African populations fit a population growth model reasonably well, but that non-African populations show a clear signature of a population bottleneck at about the time of the emergence of modern humans from Africa. This finding is important because it implies that many tests that rely on the site frequency spectrum should be corrected for the genome-wide distortions that past demography has imposed (Nielsen et al. 2005b, 2006; Williamson et al. 2005).

Rooting Human Polymorphism

For some kinds of analyses, it is important to infer which allele at a SNP was the ancestral allele and which was derived by mutation. The most common way to do this is by parsimony, that is, assuming the smallest number of mutations consistent with the data. For example, if a chimpanzee sample is homozygous GG at a SNP, and humans have an A/G polymorphism, then it is assumed that the common ancestor of human and chimp was G, and that the A is a derived mutation that arose in the human lineage. If the ancestral population were polymorphic at this SNP, then this inference could be wrong. Figure 10.6 shows the phenomenon known as **lineage sorting** in which an ancestral polymorphism has sorted in such a way as to suggest that species A and B are the most closely related, whereas species B and C are actually the closest relatives.

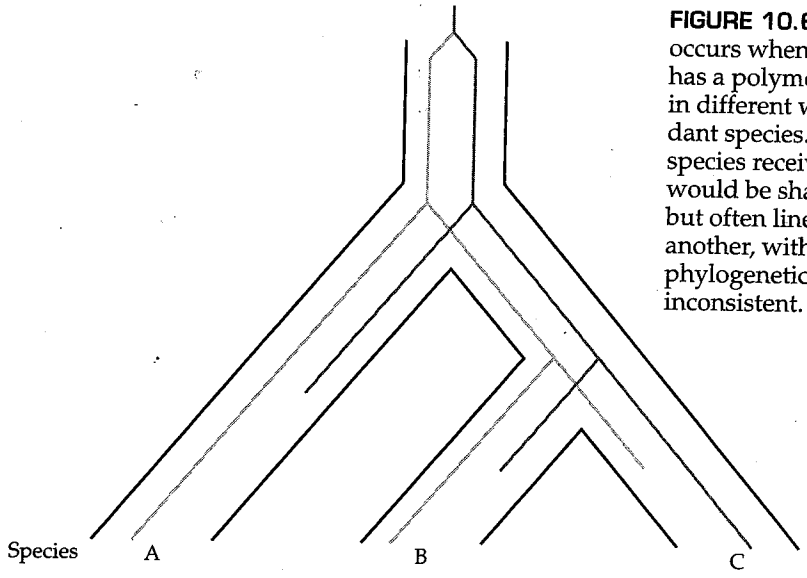


FIGURE 10.6 Lineage sorting occurs when an ancestral species has a polymorphism that sorts itself in different ways among descendant species. If both descendant species receive both alleles, there would be shared polymorphism, but often lineages lose one allele or another, with the result that the phylogenetic pattern is confusing or inconsistent.

The example using a single chimpanzee as an outgroup species to infer the ancestral state for human SNPs is of particular interest because the probability of incorrect ancestral-derived inferences can be estimated. We will see later that attempts to infer the action of natural selection on human SNPs often rely on the inference of the ancestral state, and so errors in inference result in inappropriately concluding that the high frequency allele is the one that arose since the split with chimpanzee and was driven to high frequency by selection. An excess of this kind of SNP site is an important signal of positive selection, and so correct assignment of the ancestral state is critical. Given the effective population size of humans and chimpanzees, and the effective population size of our common ancestral population, what is the chance that the parsimony approach provides incorrect assignment of ancestral state? Simulations show that this error is made for about 2% of the SNPs, which is large enough to produce some spurious signals of selection.

Inference of Inhomogeneities in the Mutation Process

The case has been made that roughly 5% of the mammalian genome has a rate of nucleotide substitution that is not compatible with neutrality (Waterston et al. 2002) but is instead consistent with a model in which these sites have a reduced rate of nucleotide substitution due to selective constraint. Since only about 1.5% of the mammalian genome is protein coding, this implies that about $\frac{2}{3}$ of the constrained or functionally important sites are noncoding. This conclusion of selective constraint comes from a model that makes certain assumptions about the way that mutations occur in the genome. If mutations occur uniformly and homogeneously across the genome, then differences in rates of substitution could be inferred to be due to differences in selection. But unless the mutation process can be shown to be homogeneous, there remains the possibility that the reduced rate of substitution might be due to local regions of reduced mutation. How do we test these competing hypotheses? In Waterston et al. (2002), the genome-wide human-mouse alignment yielded 14.3 million aligned segments of length 50 bp having 45 or more bases aligned. These alignments spanned about 25% of the human genome. The distribution of differences in these alignments was fitted to a mixture of two distributions, one due to neutral divergence and one due to constrained (negative selection). Roughly 21% of the probability mass fell into the selected portion. From this the authors concluded that $(0.21)(0.25) = 5\%$ of the human genome is under purifying selection. Although there are some gaps in the logic, this estimate has not been strongly contested as yet. Examination of alignments and conservation scores around genes in a genome browser makes the figure seem quite plausible: There clearly is a lot of conserved sequence outside of coding regions. The sequence conservation implies that mutations in these regions may be associated with loss of function (i.e., disease), and that approaches that examine only coding regions are likely to miss much that is important.

Inferences about Male and Female Mutation Rates

One of the more striking heterogeneities in patterns of variation along the human genome is the deficit of variation on the X chromosome. The average nucleotide diversity for autosomes is 0.8% while the average for the X chromosome is 0.5%. The X chromosome is hemizygous in males, exposing genes in the single X chromosome in males to the deleterious consequences of recessive mutations. Hence one possible explanation for reduced variation in the X chromosome is that deleterious alleles are more readily removed from the population by selection than those on the autosomes. One problem with this explanation is that the X chromosome also shows less variation in intergenic regions, fourfold synonymous sites, and practically everywhere else. This suggests that there is an explanation not involving gene function. The X chromosome is present at $\frac{3}{4}$ the effective population size of autosomes, and the lower effective size would reduce the levels of standing variation; but the effect is not large enough to explain the magnitude of reduced variability on the X. Because on average the X chromosome is present in females about twice as often as it is in males (we say "about" because the ratio depends on the distribution of offspring numbers per male and per female), it stands to reason that other differences between males and females may be manifested on the X chromosome. In particular, if the mutation rate were higher in males than in females, the X would have a lower mutation rate than autosomes because they are in females about $\frac{2}{3}$ of the time.

Rates of mutation in males and females have been estimated by Makova and Li (2002) who sequenced the Y-linked *DAZ* locus as well as the chromosome 3 paralog in male and female humans and four additional primates. Gene trees were constructed separately for the Y chromosomal sequences and for the chromosome 3 sequences, and the ratio of lengths of internal branches of the Y-chromosome : Autosome tree was 1.68. Assuming that the mutation rate of the Y chromosome is the same as that of the autosomes in males, this implies that the male mutation rate is more than 5 times that of females, presumably because of the greater number of cell divisions in the male germ line.

PROBLEM 10.2 A method for estimation of the ratio of mutation rates in males and females from the divergence of Y and autosomal genes was derived by Miyata et al. (1987). If Y is the rate of Y sequence divergence, and A is the rate of autosomal divergence, and α is the ratio of male to female mutation rates ($\alpha = \mu_m / \mu_f$), then $\frac{Y}{A} = \frac{2\alpha}{1+\alpha}$. Show that this relationship is valid.

ANSWER Y reflects only the male mutation rate, so $Y = \mu_m$. A reflects the divergence of genes that are in males half the time and in females half the time, so $A = (\mu_m + \mu_f)/2$. Substituting, we get:

$$\frac{2\alpha}{1+\alpha} = \frac{\frac{2\mu_m}{\mu_f}}{1+\frac{\mu_m}{\mu_f}} = \frac{2\mu_m}{\mu_m + \mu_f} = \frac{\mu_m}{\frac{\mu_m + \mu_f}{2}} = \frac{Y}{A}$$

10.3 LINKAGE DISEQUILIBRIUM ACROSS THE HUMAN GENOME

The study of linkage disequilibrium jumped from the attention of relatively few academic scientists to front-page news in the *Wall Street Journal* and other major news media propelled by the International HapMap Project. The simple idea motivating this project was that anonymous SNPs would, by virtue of their linkage disequilibrium with other SNPs in their vicinity, serve as proxies to tag variants that were genetic risk factors for disease. The rationale was that with the appropriate study design, it would be possible to identify relatively small genomic regions harboring alleles that increased risk. In order to assess the power of this strategy, we need first to take a careful look at the structure of linkage disequilibrium in the human genome.

The Landscape of Human Linkage Disequilibrium

The first thing to consider is how linkage disequilibrium declines with increasing physical separation of SNPs. Recall from Chapter 2 that one measure of linkage disequilibrium D, declines by a factor of $1 - c$ each generation, where c is the rate of recombination (crossing over) between the pair of loci being considered. In a sample from a population in mutation-drift-recombination balance, there will also be a relationship between recombination rate and linkage disequilibrium. Another measure of linkage disequilibrium is the squared correlation coefficient between alleles in the same gamete, denoted r^2. Ohta and Kimura (1971a) derived the expression $E(r^2) = 1/(1 + 4Nc)$. Before we compare empirical data to this theoretical prediction, we need to note some complications in estimating linkage disequilibrium from human data. First, human data are generally obtained by PCR amplification of SNPs from blood samples. When each SNP is genotyped in this way, the linkage phase relationships are unknown. Because the double heterozygotes (AB/ab and Ab/aB) are indistinguishable in the genotyping platform, one cannot simply count the haplotype frequencies. Therefore, the standard formulas for linkage disequilibrium cannot be applied. Provided the SNPs are in Hardy Weinberg proportions, there is a very good maximum likelihood procedure that provides estimates of the haplotype frequencies. When there is uncertainty about the fit to Hardy-Weinberg, a better estimate for linkage disequilibrium is called the *composite linkage disequilibrium* estimator (Weir 1996; Schaid 2004). In practice, if the sample is close to Hardy-Weinberg proportions, which is true of nearly all SNPs in most genotyping platforms, the composite and maximum likelihood estimators are very close to one another.

Figure 10.7 shows estimates of human linkage disequilibrium for the HLA region of the genome, a large region containing many genes important in the immune response. Notice that SNPs that are very close together show a continuous spread of values of linkage disequilibrium, including many pairs that have very little linkage disequilibrium and other SNP pairs that

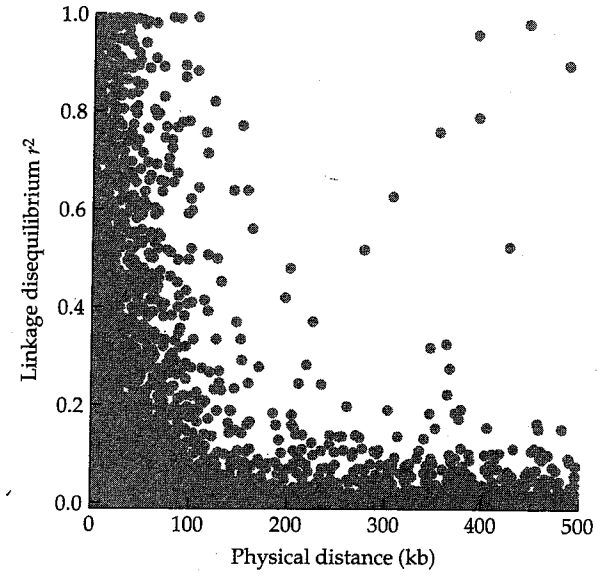


FIGURE 10.7 Linkage disequilibrium r^2 as a function of distance between SNPs in the human HLA region. SNPs that are far apart almost never display an elevated r^2; hence a large value for r^2 typically indicates close proximity between the SNPs. (Data from www.hapmap.org.)

have quite strong linkage disequilibrium. On the other hand, once the distance between SNPs exceeds 100 kb, it is relatively rare for SNP pairs to display high levels of linkage disequilibrium. Note that the pictures for D' and for r^2 are rather different. We saw in Chapter 2 that these two metrics depend on allele frequency in different ways, and they capture different attributes of the data. The r^2 is proportional to the chi-square statistic, and so it provides a metric that can be thought of as being monotonic with the P-value for a hypothesis test. D', on the other hand, has the property that $D' = 1$ whenever any of the four possible haplotypes is missing in the sample. $D' = 1$ may appear to be very strong linkage disequilibrium, but when $D' = 1$, the magnitude of r^2 can take on the full range of values, and hence there may be a complete lack of statistical significance of the linkage disequilibrium even when $D' = 1$.

Another important point about linkage disequilibrium is that it not only reflects the history of past recombination events, but also reflects the point of origin of mutations arising in the genealogy. In Figure 10.8, the genealogy of a set of haplotypes can be seen to produce linkage disequilibrium when pairs of mutations, generating derived SNP alleles, occur on the same branch of the genealogy. Until recombination breaks this haplotype apart, the process of random drift may produce linkage disequilibrium purely by virtue of this haplotype increasing in frequency in the population.

Note also that the processes of mutation and random drift on a genealogy may produce pairs of SNPs in very close proximity to one another along the chromosome, but these pairs may have very low or zero linkage disequilibri-

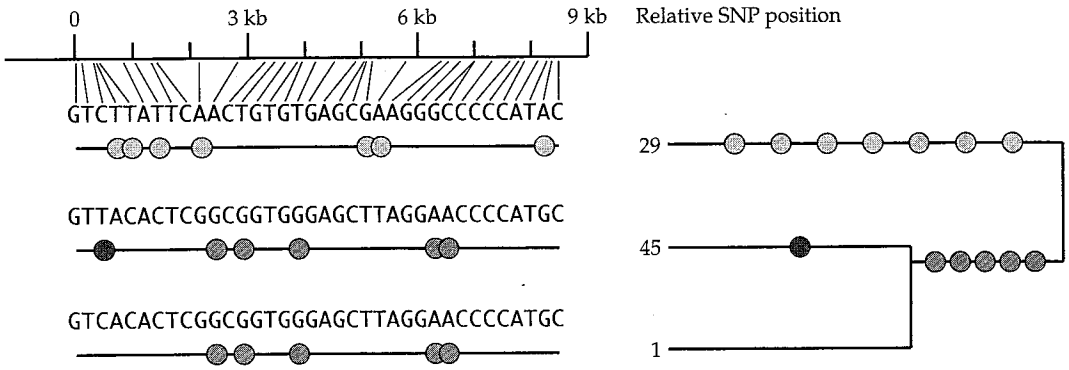


FIGURE 10.8 Some regions of the genome present a pattern in which the underlying genealogy of the haplotypes is clear, as shown in this example of 75 chromosomes. Such a clear pattern emerges when there is no evidence for recombination, back mutation, or recurrent mutation of the SNPs. (Redrawn from International HapMap Consortium 2005.)

um. In fact, even in the absence of recombination, as in the mitochondrial genome, there are SNP pairs that exhibit low linkage disequilibrium.

Linkage disequilibrium among sets of sites can be displayed conveniently in a matrix that shows the value of whatever metric is desired (Aquadro et al. 1992). This graphical approach is used in the web tool Haploview, available for browsing the HapMap data. After large quantities of human data had become available, one of the early observations from a graphical display, was that the pattern of linkage disequilibrium appeared to depart from homogeneity. Some blocks along the sequence showed high levels across nearly all SNP pairs, with gaps between blocks showing relatively lower levels of linkage disequilibrium (Figure 10.9).

Initially there was considerable discussion that these haplotype blocks indicated that the organization of variation in the human genome was much simpler than previous theory had suggested (Gabriel et al. 2002). If one restricts attention to common variants and ignores rare haplotypic combinations of alleles, the blocky structure becomes even more evident. In Daly et al. (2001), the blocky pattern was especially emphasized, and observations like this helped to motivate the HapMap Project. With an apparent block-like organization of high levels of linkage disequilibrium, the human genome really did seem to provide an opportunity to predict the allelic state of many flanking SNPs even when only a subset of SNPs was actually genotyped (Figure 10.10).

The blockiness of the haplotype structure has subsequently received less attention, in part because a rather blocky structure of linkage disequilibrium

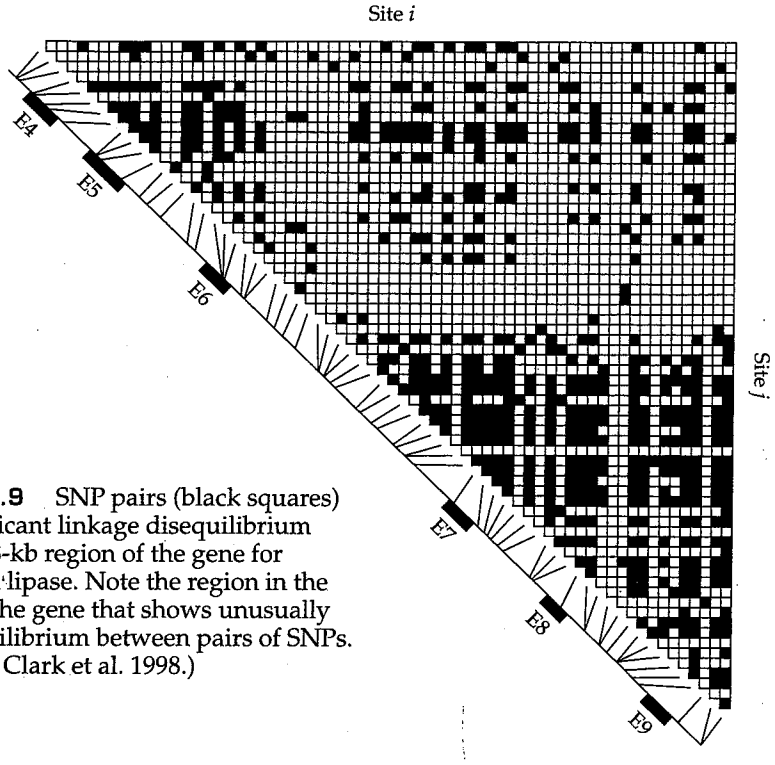


FIGURE 10.9 SNP pairs (black squares) with significant linkage disequilibrium across a 9.8-kb region of the gene for lipoprotein lipase. Note the region in the middle of the gene that shows unusually low disequilibrium between pairs of SNPs. (Data from Clark et al. 1998.)

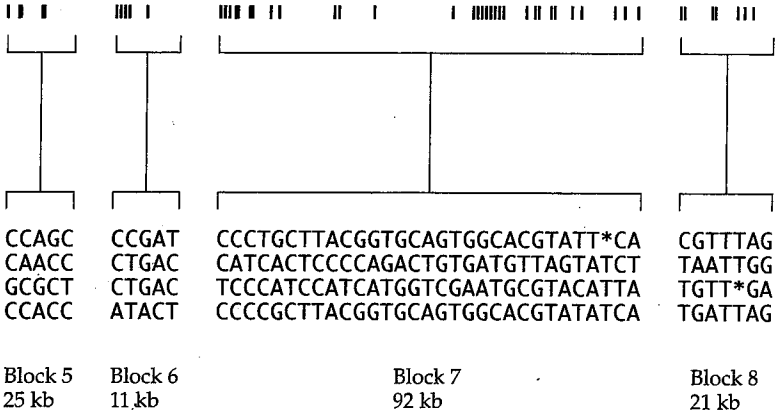


FIGURE 10.10 Haplotypes across a region in the long arm of human chromosome 5 showing a pronounced structure of linkage disequilibrium. Within certain regions there is high linkage disequilibrium with relatively few distinct haplotypes, but these are interspersed with others at sites that appear to have undergone abundant recombination. (Data from Daly et al. 2001.)

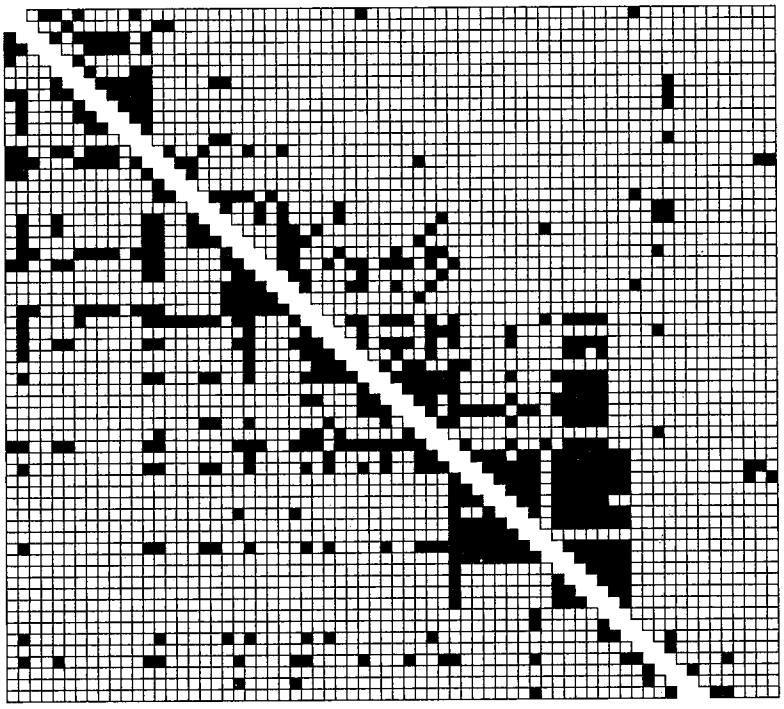


FIGURE 10.11 Simulations of the neutral coalescent can produce the appearance of block-like structures of haplotypes even when recombination is uniform across a region. This simulation mimics the roughly 400-kb region spanning the *TCRB* gene family. (From Subrahmanyan et al. 2001.)

is expected purely because of the sampling variance of the genealogical process (Figure 10.11). Furthermore, hidden SNPs can be predicted more reliably from analysis of all pairwise linkage disequilibrium measures rather than from such artificially identified haplotype blocks.

One attribute of the decay of linkage disequilibrium in the human genome remains at odds with the population genetic theory of mutation-drift-recombination balance. As mentioned before, under a model at equilibrium between these three processes, linkage disequilibrium as measured by r^2 decays according to $E(r^2) = 1/(4Nc + 1)$. Careful analysis of the data shows that relative to the predictions of this equation, the decay of linkage disequilibrium is generally too fast for very closely linked sites and then slows down at greater distances. Ardlie et al. (2001) and Przeworski and Wall (2001) make a good case for gene conversion being an explanation for this departure from expectation. Gene conversion results in what appear as nearby double-

recombination events, with the effect that the distantly flanking markers are not exchanged. That multiple genetic exchanges can influence nearby linkage disequilibrium more at nearer than at distant sites is not taken into account in the model, but with gene conversion included, the estimated rates of gene conversion provide a compatible fit.

Inferences about Local Rates of Recombination

The block-like organization of haplotypes interrupted by local regions of lower linkage disequilibrium led to the hypothesis that recombination rates were not homogeneous, but instead that there were local sites with elevated recombination. Inference of rates of recombination at this scale is technically challenging. Until very recently, the method for estimating human recombination rates was to apply likelihood methods to pedigrees (Ott 1985) to estimate rates of recombination between polymorphic microsatellites or SNPs. This approach has been successful in producing a genetic map across a large Icelandic pedigree (Kong et al. 2002), and in a unified genetic map that includes the CEPH families (Kong et al. 2004). One limitation of pedigree studies is that the number of meiotic exchanges is always limited, and this means that the mapping resolution is good only to a few percent recombination, which corresponds to a few megabase pairs of DNA sequence. Resolution for genetic mapping has been greatly improved for small regions of the genome by genotyping sperm (Jeffreys et al. 2001; Tiemann-Boege et al. 2006). This method requires PCR amplification of single sperm cells to reveal the genetic composition of single products of male meiosis. Generally the interest has been to find recombination events between tightly linked markers, and various tricks are used to favor PCR amplification of these rare recombination events. These methods have produced a surprising result, namely that there are local regions of the human genome where the rate of recombination is up to 1000 times that of the surrounding region (Figure 10.12). Further analysis suggests that as much as 80% of recombination in humans occurs in such hot spots.

The HapMap Project yielded a very dense set of SNP genotypes spanning the human genome and an ability to examine linkage disequilibrium across all chromosomes. We saw earlier that there is a relationship between the local recombination rate and the rate of decay of linkage disequilibrium, and methods were soon developed to estimate $4Nc$ from genotype data, where N is the effective population size. These methods showed surprisingly good correspondence between recombination rates estimated from pedigrees and this indirect estimate of population recombination rate. The agreement was surprising because the population recombination rate includes a term involving the effective population size, and many demographic and selective factors impact effective size. But apparently the distortions in effective size occur on a scale different enough from the effects of recombination that a clear signal remains.

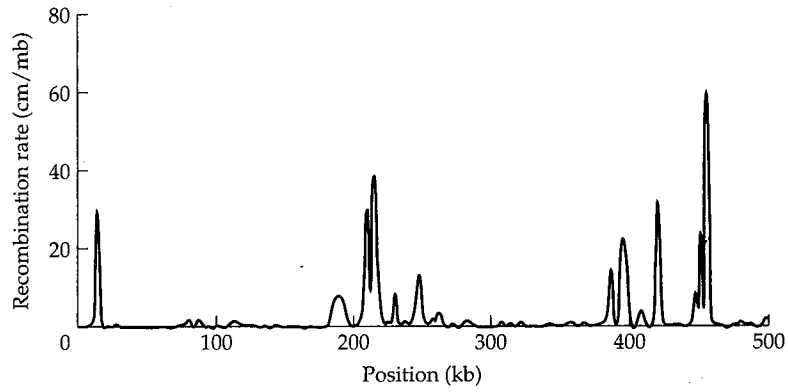


FIGURE 10.12 Hot spots of recombination across a 500-kb region of the human genome identified by Jeffries et al. (2000). The fist evidence for such hot spots came from the clustered structure of linkage disequilibrium discovered in the human HapMap Project. (International HapMap Consortium 2005.)

Population recombination rates have been estimated in segments across the entire human genome, and there is wide variation in local recombination rates (Myers et al. 2005). Given a collection of genotypes of adjacent SNPs in a local region of the genome and a proposed value of $4Nc$, a composite likelihood of these observed data is calculated (Hudson 2001). For the analysis of such data, McVean et al. (2004) devised a reversible jump Markov chain Monte Carlo algorithm that allows several kinds of updates to the current estimates of parameters: (1) increase (or decrease) the local recombination rate r, (2) extend the size of the block of homogeneous recombination rate to the 5′ direction (or 3′ direction), (3) split the current region at an arbitrary point, or (4) merge adjacent regions into one homogeneous block.

For each proposed change, the ratio of composite likelihoods is calculated, and this ratio is then used to decide whether to accept or reject the proposed update. When this algorithm is run (after considerable effort to tune the prior information about likely distributions of sizes of regions of homogeneous recombination) it eventually converges to a stable cloud of solutions giving a posterior distribution of recombination rates for each local region of the genome. Extensive simulations were done to show that the algorithm recovers estimates of local recombination rates that correspond well to the rates that were input in the simulation. A primary result of this analysis is the finding that the null hypothesis of homogeneous recombination rate across the genome is soundly rejected in favor of the alternative that recombination rates vary. In fact, the method predicts that 80% of all the recombination in the human genome occurs in 10–20% of the length of the chromosome. This local concentration of recombination would be expected to produce local

regions of unusually high linkage disequilibrium, a feature that can be used to advantage in mapping genes associated with genetic disease.

To the extent that variation across the genome in estimates of $4Nc$ are driven by variation in recombination rate and not by effective population size, it should be possible to use the estimates of $4Nc$ as an update of the linkage map at a very fine spatial scale. In particular, these maps identify many local sites of very high rates of recombination. The correspondence is not always perfect between the single-sperm typing approaches that give experimental identification of recombination hot-spot positions, and the inferential method that finds spikes in $4Nc$. In some cases there is a region where sperm typing has identified what is truly a recombination hot spot, and yet linkage disequilibrium remains high across the region. One explanation is that perhaps these recombination hot spots have arisen so recently in the population that linkage disequilibrium has not yet decayed, and so there is no signature in $4Nc$ (Jeffreys et al. 2005). The rapid evolution of hot spots is further supported by the observation that there is very poor correspondence between hot spots inferred in the human genome and the chimpanzee genome, despite the fact that they are only 1.2% diverged in sequence (Ptak et al. 2004, 2005). Many questions about recombination hotspots remain active areas of research, including their mechanistic basis, motifs of sequence that identify their likely locations, their longevity in the population, and the reasons for their low level of interspecific conservation.

10.4 POPULATION STRUCTURE INFERRED FROM HUMAN POLYMORPHISM

There is a rich history of analysis of population substructure in humans, perhaps best exemplified by the beautiful compendium of human migration history inferred from genetic studies (Cavalli-Sforza and Cavalli-Sforza 1995). Humans from different parts of the world differ in many ways. Genetic differences between human populations are very subtle, however. For example, with enough data tests of the null hypothesis that two samples of humans are drawn from a single panmictic population can sometimes be rejected, even if the individuals are from the same ethnic and racial background but are living in different cities. This implies that genetic variation can provide remarkably fine-grained discrimination of the birthplaces of people. But the classic F_{ST} statistic puts the amount of differentiation into comparative perspective. Despite the fact that European Americans from Birmingham, Chicago, Minneapolis, and Oakland can be identified as significantly heterogeneous, the F_{ST} for this set of samples is only 0.001, implying that less than 0.1% of the genetic variation differentiates these samples. Such is the power of modern genetics that it can identify such small differences as significant. We will see later in this section that F_{ST} among even the most diverse collections of human groups is only about 0.05, meaning that about 95% of the variation is within each group. This degree of geographic subdivision is

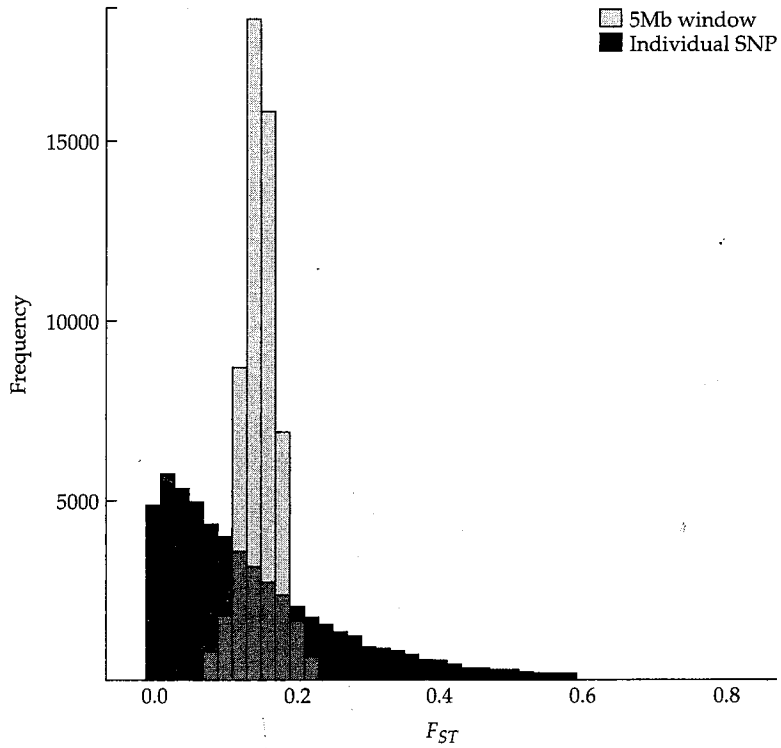


FIGURE 10.13 Distribution of F_{ST} values for human SNPs among Africans, Europeans, and Asians. The mean is approximately 0.10 but the range is very wide. When SNPs are combined across 5-Mb segments, the mean F_{ST} is somewhat greater than 0.10 but the variance is much reduced. (After Weir et al. 2005.)

much less than that observed in most other organisms that have extensive geographic ranges.

Note that these estimates of F_{ST} are averages across many SNPs. Figure 10.13 shows a histogram of F_{ST} values of individual SNPs estimated from the HapMap data. Although the mean value of F_{ST} is very small, there are many individual SNPs in the genome with F_{ST} values greater than 0.25. Some of these have such high values of F_{ST} only because of the vagaries of chance in the form of founder effects and random drift. Other SNPs may have acquired their large differences in allele frequency among populations due to population-specific natural selection. We will consider this issue in greater detail in Section 10.7.

Multilocus Methods of Inference of Stratification

When presented with genotype data from individuals sampled within a region, we must determine whether the sample came from a single panmictic

population or whether the sample represents a mixture of two or more subpopulations. This kind of hidden stratification in a population sample is extremely important for testing association between SNP variation and disease risk, because the presence of such stratification can result in a spurious conclusion of associations. A mixed sample from two populations might even be expected to depart from Hardy-Weinberg proportions, but the departure is often not statistically significant unless the sample is large and the subpopulations are very different in allele frequency. Methods that make use of many loci at once are much more powerful (Pritchard et al. 2000; Corander et al. 2003; Tang et al. 2006), and so widespread application of these methods can be expected as genome-wide association testing becomes more widely applied.

The first model of Pritchard et al. (2000a,b), implemented in the widely available computer program STRUCTURE, assumes that there are k populations that have been mixed without intermating, and that the samples from each of the k subpopulations are in Hardy-Weinberg proportions with all markers in linkage equilibrium.

Let us use the symbol **G** to refer to the vector of genotype frequencies, **P** to refer to the vector of population identity, and **A** to refer to the vector of allele frequencies. If we know the populations of origin and the allele frequencies in each population, then $Pr\{\mathbf{G} \mid \mathbf{P}, \mathbf{A}\}$ is simply the multinomial sampling property for a collection of genotypes under Hardy-Weinberg equilibrium. By a Markov chain Monte Carlo procedure the posterior density of $Pr\{\mathbf{P}, \mathbf{A} \mid \mathbf{G}\}$ can be obtained by using the Bayesian relation

$$Pr\{\mathbf{P}, \mathbf{A} \mid \mathbf{G}\} = Pr\{\mathbf{G}\}Pr\{\mathbf{A}\}Pr\{\mathbf{G} \mid \mathbf{P}, \mathbf{A}\}.$$

An iterative scheme produces a Markov chain whose stationary distribution gives this probability density, and from it we can ascertain whether the sample fits a single-population model or whether it appears to be a mixture of genotypes from disparate populations. The method requires one to specify the number k of subpopulations that might be in the mix, and it suffers the usual problems of MCMC methods including long compute times, occasional difficulties in convergence, and input of information about prior distributions of parameters. These challenges have not kept STRUCTURE from being widely and successfully used in many contexts where inference of hidden substructure was important.

The same model can also be used to estimate admixture levels in individuals. This is an especially important goal for association tests, and in fact there are methods that explicitly use the degree of admixture for mapping genes (next section). More recently, Tang et al. (2005) developed maximum likelihood methods that give fast and accurate estimates of individual admixture levels.

To study the global structuring of human genetic variation, Rosenberg et al. (2002, 2005) genotyped a sample of 1056 humans from a broad, global geographic distribution spanning 52 populations at 377 microsatellites. These genotype data were then run through STRUCTURE specifying a range of $k = 2$ through 6 subpopulations in order to see whether the genetic data corre-

sponded in any meaningful way to the known geographic origin of the samples. As discussed in Chapter 6, the best partitioning occurred in a model with six different subgroups (Rosenberg et al. 2002). The genetic data do provide a surprisingly good indication of the geographic structuring of the human population, and cases where there are some ambiguities generally have a good historical explanation.

Other methods exist for identification of population stratification, and with high-density SNP genotyping methods now available, the computational demands of the MCMC approaches will limit their usefulness. Instead, methods like Principal Components Analysis (PCA) will likely once again be widely applied. PCA can be readily applied to more than 500,000 SNP genotypes scored in thousands of individuals (Price et al. 2006). One advantage of this approach is that it can be applied to identify homogeneous subgroups, and then association testing can be done within each subgroup. The effects of hidden stratification on association tests can be subtle and complex, and so this is an area of intense investigation for improving methodology.

Heterogeneity in Linkage Disequilibrium across Human Populations

In addition to testing whether allele frequencies are homogeneous across populations, it is important to test whether the levels of linkage disequilibrium differ among populations. One of the most robust findings in human variation is the observation that populations of African descent have patterns of genetic variation consistent with a larger long-term effective population size than inferred for populations of non-African ancestry. The large effective population size is reflected in their elevated levels of diversity, elevated haplotype diversity, and reduced levels of linkage disequilibrium. Reich et al. (2001) examined linkage disequilibrium across predefined local regions of the genome and quantified the faster rate of decay of linkage disequilibrium in the Yoruban population of sub-Saharan Africa (Figure 10.14). This result implies that more SNPs will be required to adequately cover the genome for genome-wide association mapping in African populations, but that if the mapping is successful, the genetic resolution will be finer.

Subsequent to the HapMap Project, an important problem arose as to whether landmark SNPs that are identified as being highly predictive of flanking SNPs (because they have high linkage disequilibrium) will have a similar level of informativeness for the flanking SNPs in other human populations. This is another way of asking whether the flanking SNPs will display the same pattern of linkage disequilibrium, but it is a bit more specific because of the ability to predict missing genotypic data from multiple pairwise linkage disequilibrium values. The test of association among SNPs involves sampling genotypes from other populations, and then determining values of linkage disequilibrium and accuracy of prediction. De Bakker et al. (2005) report one such study, and the result demonstrates how much this problem can be seen as bad-news good-news situation. The bad news is that

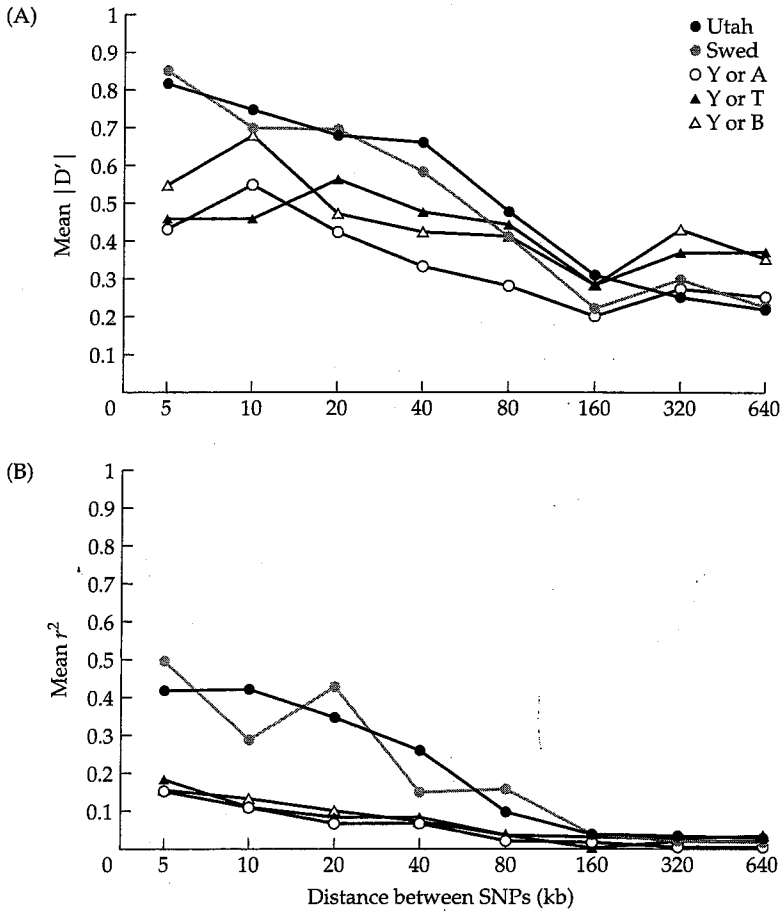


FIGURE 10.14 Decay of linkage disequilibrium with distance averaged across the human genome. African populations (Y, T, B) clearly have less linkage disequilibrium than the other populations (Utah, Sweden), a finding that is consistent with African populations having a larger long-term effective population size. (After Reich et al. 2001.)

there are measurable differences in the levels of linkage disequilibrium across human populations. But despite this overall pattern, the good news is that, by and large, the same SNPs are predictive in multiple populations; the average accuracy of prediction is eroded only a little by the heterogeneity among populations. Despite this result, it is nevertheless possible that a particular disease-causing SNP will be totally missed as a result of the across-population heterogeneity. For mapping genetic variation in human disease susceptibility, population-specific variation needs to be considered.

Linkage Disequilibrium in Admixed Populations: Admixture Mapping

If there are differences in the haplotype frequencies between two groups, then one would expect that a population representing an admixture of the two populations ought to have greater linkage disequilibrium than either parental population. In African Americans, for example, one might expect greater linkage disequilibrium than in Africans, as theory suggests. But it turns out that the decay of linkage disequilibrium in African Americans is very much like that of Africans, namely the decay is faster than that in Europeans. How can this be so? First, the level of admixture is on the order of 10–20%, so the European contribution is relatively small. In addition, the magnitude of allele frequency difference between European and African populations is not high, with F_{ST} less than 0.15 on average. In addition, a number of generations have passed since African populations entered North America, and so several rounds of recombination have occurred. The balance of these factors produces an African American population with high levels of genetic diversity and relatively less linkage disequilibrium than their European-derived neighbors.

If a gene has different allele frequencies between two populations, then one ought also to be able to make a probabilistic statement about the likely population of origin of a region of the genome. Tang et al. (2006) formalized this idea to identify population ancestry in local blocks across the entire genome. The idea of **admixture mapping** is to do this for each individual in an admixed population, and then correlate the disease risk with the inferences of ancestry. Any region along a chromosome that contains a genetic risk factor for disease is expected to show a greater shared ancestry from the population with the greater overall risk. Figure 10.15 shows how this works.

To understand the basic idea behind likelihood methods for admixture mapping, consider the estimation of the degree of admixture in a population. Suppose that an allele has frequency p_1 in population 1 and p_2 in population 2. In the admixed population, the frequency is p_A. If f is the fraction of the genome in the admixed population that came from population 1 (and $1-f$ came from population 2), then $p_A = fp_1 + (1-f)p_2$. Rearranging we get:

$$f = \frac{p_A - p_2}{p_1 - p_2}$$

This is also the maximum likelihood estimator for f. The linkage disequilibrium from admixed populations can also be used for genetic mapping; this approach offers considerable promise (McKeigue et al. 2000; Tang et al. 2006).

Inbred Populations and Homozygosity Mapping

Some human populations have been small in size or they have practiced marriage customs that have promoted consanguinity for sufficiently long

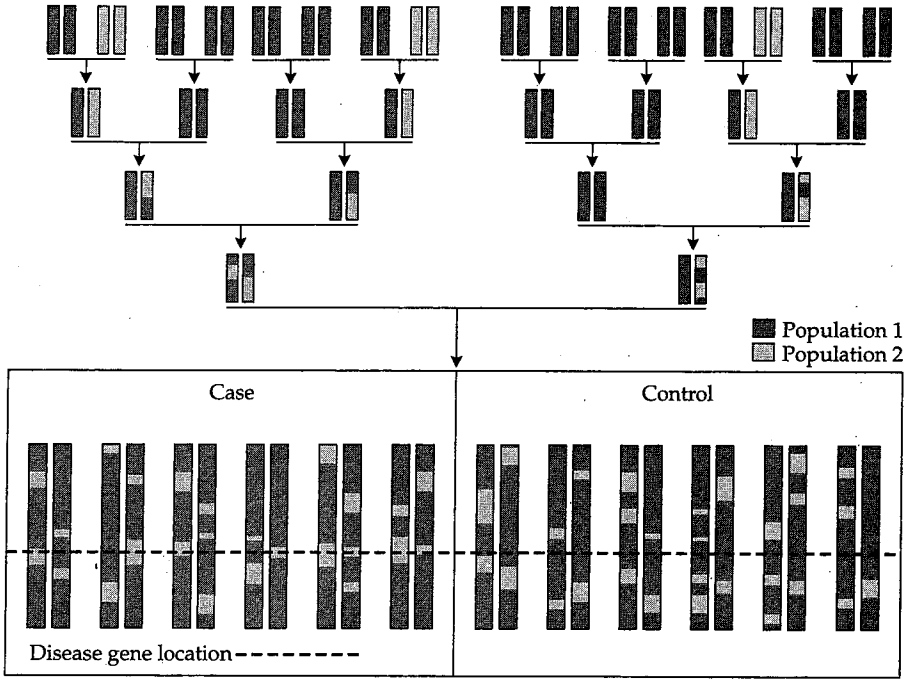


FIGURE 10.15 Admixture mapping makes use of cases and controls from an admixed population whose genotypes at a large number of SNPs are scored. The SNP data allow identification of regions of the genome derived from one ancestral population or the other. If a region of the genome shows a statistical association between disease status and SNP genotype, it suggests that a genetic risk factor for the disease is located in the region. (Redrawn from Darvasi and Shifman 2001.)

periods of time that the populations are markedly inbred. The genome is therefore expected to be structured with longer stretches that are homozygous than is found in other populations. Among inbred individuals affected with a particular disorder, it may be possible to identify regions of homozygosity that are shared among the affected individuals. This approach, called **homozygosity mapping**, works in outbred populations as well, provided that the cases are caused by the same mutation and that enough markers are genotyped (Lander and Botstein 1987). The efficacy of homozygosity mapping derives from the expectation that highly inbred populations, or populations with small founding sizes would have relatively large blocks of homozygosity (Figure 10.16), making it feasible to scan the entire genome with fewer markers.

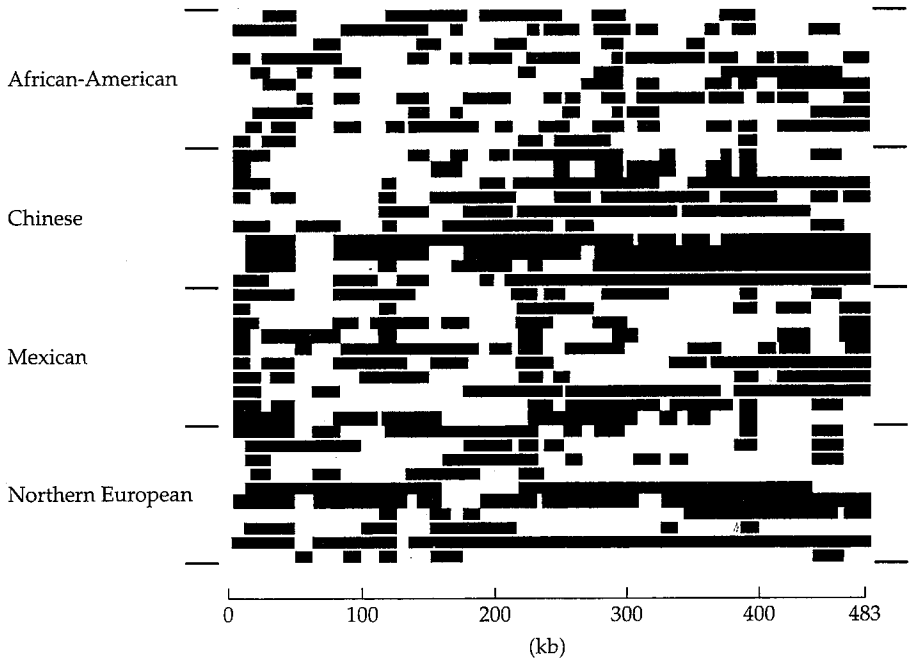
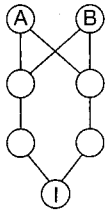


FIGURE 10.16 One signature of haplotype structure in a population is the distribution of runs of homozygosity. In the human gene encoding the T-cell receptor beta protein, different populations show very different patterns of run lengths of homozygosity. (From Subrahmanyan et al. 2001.)

PROBLEM 10.3 Consider two single nucleotide polymorphisms separated at a distance along the DNA such that the probability of recombination between the SNPs is c per generation. In the pedigree of first-cousin mating shown here, assume that neither of the common ancestors A and B is inbred. In the inbred individual I, what is the probability that the SNP alleles present in any of the homologous chromosomes in either A or B are transmitted down the pedigree together, with the result that the individual I is identical by descent and therefore homozygous for a region large enough to include both of the SNP alleles? What is this probability for a 1-Mb region of the human genome across which $c = 0.01$?



ANSWER This is a traditional type of inbreeding problem (see Chapter 6) except that it focuses not on an individual allele but on a region of chromosome. For each loop in the pedigree, therefore, the probability needed is not merely the probability that a particular allele present in two gametes is identical by descent, but it is the probability that the region of interest is identical by descent in the gametes *and* has not undergone recombination. Because the common ancestors are not themselves inbred, the desired probability is therefore $(\frac{1}{2})^5(1-c)^6 + (\frac{1}{2})^5(1-c)^6 = (1-c)^6/16$. For a region in which $c = 0.01$, the probability of identity by descent of the entire region equals $(1-0.01)^6/16 = 0.9415/16 = 0.0588$. In other words, the probability of identity by descent of a 1-Mb region is about 94% of the probability of identity by descent of a SNP located exactly in the center of the region.

10.5 MENDELIAN DISEASE AND POPULATION GENETICS

Several of the common severe Mendelian disorders have sufficient research importance that for each new patient the mutant gene sequence is determined in order to identify the molecular lesion. This effort has led to compendia of allelic sequences that serve as textbooks of gene dysfunction, the largest of which is the Human Gene Mutation Database (www.hgmd.cf.ac.uk). Among the defective alleles for phenylalanine hydroxylase (*PAH*), the gene whose defects result in phenylketonuria (PKU), there are mutations that result in premature translation termination, failed intron splicing, amino acid substitutions that destroy activity, gene deletions or insertions that change the reading frame, and alterations in the promoter that abolish normal expression. Altogether more than 500 different alleles for *PAH* have been characterized (Figure 10.17).

PAH is not alone in exhibiting this level of mutational diversity. The gene that causes cystic fibrosis, *CFTR*, has more than 900 different defective alleles characterized. Similarly the genes involved in Duchene muscular dystrophy (*DMD*) and familial breast cancer (*BRCA1*) each have hundreds of alleles described. Not only do many Mendelian disorders display huge numbers of defective alleles, but they also show a complex pattern of genotype-phenotype relations, with considerable variation in penetrance, in degrees of interallelic complementation, and different manifestations of pleiotropic effects. The mutational complexity is emphasized in the adage "monogenic disorders are not simple" (Scriver and Waters 1999).

The great mutational diversity in Mendelian disorders prompts the question whether the pattern departs from what is expected under reasonable models of mutation, random drift, and selection. If the pattern is what is expected based on realistic estimates of mutation and plausible strengths of selection, then the mutational complexity suggests that genes that increase risk of complex disease are also unlikely to have a single "good" allele and a single "disease" allele. Genes affecting complex traits may well have a myriad of rare

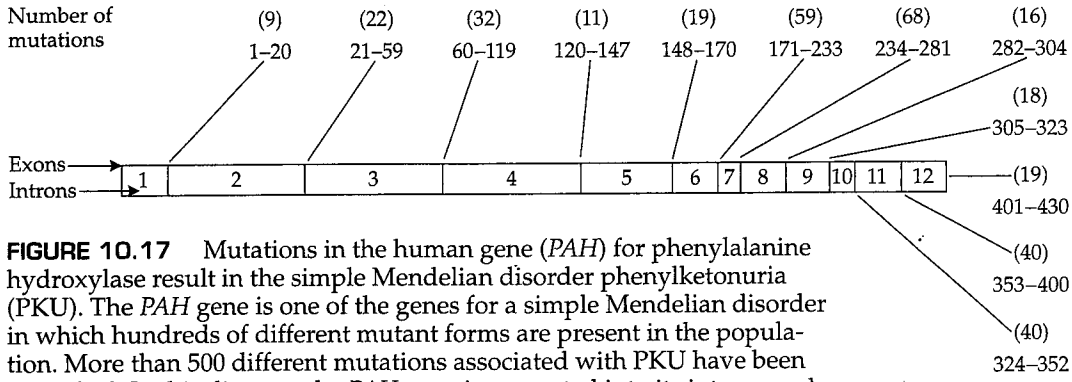


FIGURE 10.17 Mutations in the human gene (*PAH*) for phenylalanine hydroxylase result in the simple Mendelian disorder phenylketonuria (PKU). The *PAH* gene is one of the genes for a simple Mendelian disorder in which hundreds of different mutant forms are present in the population. More than 500 different mutations associated with PKU have been identified. In this diagram the *PAH* gene is separated into its introns and protein-coding exons (vertical bars), and the number of amino acids encoded in each exon is shown along with the number (in parentheses) of different mutations that have been identified in that exon. (Data from http://www.pahdb.mcgill.ca)

alleles. But if every gene for every complex disorder were to have the allelic complexity of *CFTR* or *PAH*, then genetic mapping of these genes would be an enormous challenge. The feature that allowed *CFTR* to be mapped was the high relative frequency of a particular allele, denoted *ΔF508*, in people of European ancestry. This example illustrates that a sufficient skew in the haplotype frequencies may provide the basis for successful genetic mapping.

Mutation-Selection Balance

Under strict neutrality, the allele frequency spectrum in a sample is known for both the infinite-allele and the infinite-sites model (see Chapter 4), and we have seen that human variation tends to fit the expectations reasonably well, except that human genes generally show an excess of rare alleles. One reason why simple Mendelian disorders often display such high numbers of distinct alleles is that there is such efficient identification of even exceptionally rare alleles identified through the disease itself. Although it may appear that the sample size is the number of sequences obtained, it is actually much larger because so many distinct alleles are ascertained through the presence of the disorder. Therefore vast numbers of nonmutant alleles are not analyzed because they do not cause disease. Depending on the level of dominance and the degree of interallelic complementation, surprisingly high equilibrium frequencies of defective alleles can be expected under mutation-selection balance (Crow and Kimura 1970; Hartl and Campbell 1982; Clark 1998).

Dating the Origin of Mutant Alleles

Several methods exist to place bounds on the estimated age of a particular allele (Slatkin and Rannala 2000). For example, one can use the allele fre-

quency alone, relying on diffusion theory (see Chapter 3) to produce estimates of the allele's age. When sequence data are available, genealogical approaches generally provide more precise estimates.

On the other hand, not all Mendelian diseases have a plethora of rare alleles. In some cases there appears to be a single major allele responsible for a large proportion of the cases. When this is observed, it may indicate that the allele has not been in the population for very long, and therefore the low allelic diversity stems from the brief time needed to trace the common ancestry of all extant members of the lineage. The significance of the depression in diversity of a clade of haplotypes was considered by Hudson et al. (1994). (A **clade** of haplotypes is a group of haplotypes that derive from a single common ancestor.) This analysis gave rise to the study of the **structured coalescent** which focuses on the properties of gene genealogies conditional on having a clade of a certain size. It is straightforward to generate gene genealogies in large numbers, and so one can rapidly simulate samples of gene genealogies that satisfy a criterion such as "the tree must have 100 tips with a single monophyletic clade having 23 tips." By studying the properties of these random trees, one can determine the bounds of the estimated age of a clade in a neutral coalescent.

As an example of this approach we may consider alleles encoding the red blood cell enzyme glucose-6-phosphate dehydrogenase (G6PD). Low activity alleles of *G6PD* attain high frequencies in human populations only in areas where falciparum malaria is present, and there is direct evidence for the protective effect of these "deficiency" alleles. By scoring RFLP and microsatellite variation in numerous isolates of *G6PD*, Tishkoff et al. (2001) fitted models to estimate the age of the low-activity variants. The 95% confidence interval for the origin of the *A*– variant was 3840 to 11,760 years ago, supporting the hypothesis that malaria has had a major impact on humans in the relatively recent past.

As another example of a young allele at high frequency, Bersaglieri et al. (2004) typed 101 single-nucleotide polymorphisms covering 3.2 Mb around the gene for intestinal lactase. This is the gene responsible for adult persistence of lactase, and it may have been important in allowing pastoralists to gain nutritional benefit from cow's milk even after adulthood. In northern European-derived populations, a common haplotype extends largely undisrupted for more than 1 Mb. This pattern is highly unlikely to exist by random drift and recombination, and computer simulations suggest that strong selection favored the alleles within the last 5000–10,000 years, a time frame consistent with the spread of dairy farming.

As a final example, consider the Δ32 allele of the chemokine CCR5. This allele is a loss-of-function mutation and yet has an appreciable frequency across Eurasia. Furthermore, this allele confers resistance to the progression of HIV to AIDS. Stephens et al. (1998) found strong linkage disequilibrium between two microsatellite loci and the *CCR5* Δ32 allele in a sample of 192

European chromosomes. Similar coalescent model simulations produced an estimate of just 700 years for this allelic lineage. Because it is highly unlikely that an allele could drift to the observed allele frequency (5–14% in Europe) in such a short time by chance alone, the data appear to be consistent with strong natural selection. However, Sabeti et al. (2005) reevaluated the case for selection at *CCR5* using much denser genetic maps and additional control data. Taking into account the level of genetic variation data across the human genome, these authors conclude that the pattern of variation at *CCR5* was not exceptional but is consistent with the Δ32 alleles sharing an ancestor more than 5000 years ago. Such a result does not prove that the Δ32 allele was not favored by selection, but it does demonstrate how rapidly data about human genetic variation has accumulated, and how the tests of selection have gained in precision.

10.6 GENETIC BASIS FOR VARIATION IN RISK OF COMPLEX DISEASE

Any statement about the structure of genetic variation in risk of complex disease must start with the **genetic architecture** of the disease, which means knowledge of how many genes are involved, what are the magnitudes of the effects of allelic differences on risk of disease, how much the alleles at different SNPs interact in determining risk, and to what degree genotype-by-environment interactions are important. A substantial part of the research effort in human genetics is currently devoted to answering these questions for at least the major public health threats (diabetes, heart disease, cancer, depression). The challenge is not only to describe the genetic architecture, but also to identify individual genes and the role they play in mediating risk. Given the myriad of genes that might be involved in some disorders, the task is formidable, but the hope is that at least some disorders will have a substantial portion of their risk attributable to just a handful of genes.

Studies of complex traits usually begin with a scheme for identifying patients and recruiting them into the study. In the **case-control design**, which is good for rare disorders, individual cases are first identified in clinics, and then other control individuals matched for age, sex, body weight, and other attributes are also recruited. Many complex traits have a high enough incidence that a random population sample will include sufficiently many cases. Such a **random cohort design** has advantages because a random sample is expected to include a cross section of the population, and therefore has a built-in control.

Mapping Methods Based on Linkage

One approach to genetic mapping of genetic risk factors is to test whether particular genetic variants are associated with increased disease risk because they are shared by affected individuals within families. The classical

approach was to study extended pedigrees. Methods for calculating the likelihood of a set of genetic markers given a model for disease transmission and the topology of the pedigree have been well established. This approach is still used with great efficiency on large pedigrees (e.g., Grant et al. 2006). Because it is not always easy to identify large, multigeneration pedigrees, the approach of using pairs of affected sibs has come to have a strong following, and likelihood methods for testing regions of the genome that are identical by descent in affected sib pairs have provided evidence for genetic linkage (Risch 1990). Another method that makes use of transmission within families is the **transmission disequilibrium test** (TDT) (Spielman et al. 1993). This test has a remarkably simple design: Affected children are ascertained and they and their parents are genotyped. Only parents that are heterozygous *Mm* at any marker to be tested are considered. Define a as the number of times in the sample that a heterozygous parent transmits allele M to the affected offspring, and b as the number of times the allele m is transmitted. The TDT test statistic, $(a-b)^2/(a+b)$, has a chi-squared distribution with one degree of freedom, provided the numbers are reasonably large. This test attained considerable popularity because it tests only the transmission pattern within each family, and so it is quite robust to population stratification.

PROBLEM 10.4 Let D denote the less frequent (minority) allele of a SNP that is actually the causative nucleotide (called a *quantitative trait nucleotide* or *QTN*) in a genetic risk factor for a complex disease, and let d denote the much more frequent alternative nucleotide of the same SNP. Let x be the probability that an individual of genotype *Dd* is affected with the trait, and let y be the probability that an individual of genotype *dd* is affected with the trait. Considering only affected offspring from matings of the form $Dd \times dd$, what is the probability that the heterozygous parent transmitted the D allele, and what is the probability that the heterozygous parent transmitted the d allele? Suppose that a study includes exactly 100 affected children from such matings, and that the observed number a of D transmissions and b of d transmissions equals their expected values. (Note: The "observed numbers" in this problem need not be integers.) What magnitude of x relative to that of y is required to yield a resulting chi-square value with one degree of freedom equal to 3.84, and therefore to be significant at the 5% level?

ANSWER The mating $Dd \times dd$ yields offspring of genotypes $\frac{1}{2}$ *Dd* and $\frac{1}{2}$ *dd*, Among the *Dd* offspring, a proportion x is affected, and among the *dd* offspring, a proportion y is affected. Therefore, among affected offspring, the probability of D transmission is $x/(x+y)$ and the probability of d transmission is $y/(x+y)$. Among 100 offspring, the "observed numbers," if they match their expectations perfectly, will be $a = 100x/(x+y)$ and $b = 100y/(x+y)$. The critical chi-square value specified is $(a-b)^2/(a+b) = 3.84$, which in terms of x and y is $100[(x-y)/(x+y)]^2 = 3.84$. It follows that the requirement for statistical significance at the 5% level (or greater) is $x = 1.4874y$.

Transmission-based mapping does not take advantage of the principles of population genetics. It also relies on meiotic crossover events that took place within the history of the sample to find evidence for linkage. An essay by Risch and Merikangas (1996) galvanized the human genetics community by explaining why these methods had such low power and low resolution, and laid out the argument for direct association testing making use of linkage disequilibrium within population samples. It took some time for this idea to catch on, but soon a large part of the human genetics community realized that they could use population genetics to advantage. Although it is gratifying to population geneticists to see the utility of population genetics recognized, there still remain strong advantages to linkage methods, especially in cases of allelic heterogeneity where increased risk is due to any of multiple different defective alleles at a locus. Linkage methods continue to work well in this setting, whereas methods that rely on association may be misleading.

Linkage Disequilibrium Mapping

Imagine an attempt to map by linkage disequilibrium in the following way. One collects a number of cases of a particular disease, and a set of age, sex, and ethnically matched controls. All samples are genotyped at a sufficiently large number of sites to saturate a region of interest (including possibly the whole genome). "Saturate" in this context means that enough SNPs are genotyped so that any other remaining SNP variation in this region of the genome will nevertheless be in linkage disequilibrium with the SNP markers that are genotyped. The need to determine directly how many markers (and which markers) would approximate saturation was one of the motivations for the HapMap Project. Once the genotype data are available, the very simplest test one can do is to perform a chi-square test, either based on alleles [(*A* vs. *a*) by (case vs. control)], which yields a 2×2 table or based on genotypes [(*AA*, *Aa*, or *aa*) by (case vs. control)], which yields a 3×2 table. Figure 10.18 shows an example of this simple approach, plotting $-\log_{10}$ of the *P*-value on the *y*-axis for SNPs whose location along the genome is indicated by position on the *x*-axis.

Because the SNPs are close enough together to be in linkage disequilibrium among themselves, if there is an unscored SNP in the region there are likely to be several other SNPs that have been genotyped that do produce a significant result. Some of the early attempts to formalize methods for mapping by linkage disequilibrium (e.g., Hill and Weir 1994) emphasized the large sampling variance of linkage disequilibrium and that the large sampling variance would make the mapping somewhat imprecise. But in the context of a saturation of SNPs, the issue is not so much whether precise inference of gene location can be obtained, but whether a significant signal of association can be obtained in a genome-wide scan. Statistical methods for performing association testing are steadily improving. Beyond the one-SNP-at-a-time approach, other methods use the haplotype structure in windows (overlapping regions) of SNPs (Schaid et al. 2002), information about the

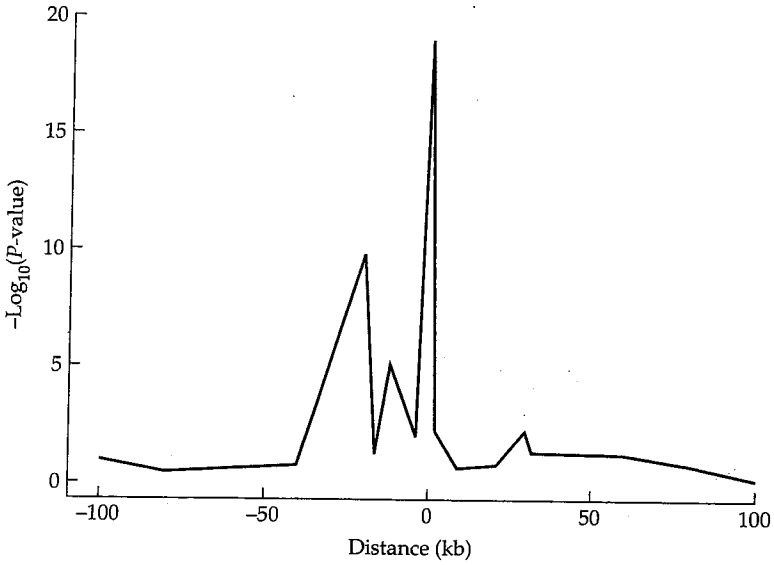


FIGURE 10.18 An example of association mapping to identify SNPs that serve as markers of chromosomal regions containing genetic risk factors. In this sample of 200 cases of familial Alzheimer's disease and 220 unaffected controls, 60 SNPs around the region of the genome encoding apolipoprotein E (ApoE) were genotyped. For each SNP, a chi-square test of association was done, and the $-\log_{10}$ of the *P*-value was plotted as shown here. Some *ApoE* genotypes are known to have an elevated risk of Alzheimer's disease, and this study demonstrated the efficacy of association mapping when the genetic effects are sufficiently large. (Redrawn from Martin et al. 2000.)

decline of linkage disequilibrium with distance (Lazzeroni 1998), Bayesian inference of haplotype patterns (Wang and Rannala 2005), or departures from the neutral coalescent for clusters of SNPs (Zollner and Pritchard 2005).

Genome-Wide Association Studies

Any of the tests mentioned above for linkage disequilibrium mapping can be applied to genome-wide data. Typically, one starts by calculating a test statistic for the probability that a disease-causing mutation exists in the population at each site or window of sites along the genome. Even though the statistical properties of the test might be satisfactory in the setting of a single test, severe problems can be encountered when doing hundreds of thousands of tests, many of which are not independent of one another. If so many tests each gave a reliable estimate of a *P*-value, then one could then simply rank those *P*-values and focus on the markers with the lowest *P*-values. The problem is that in a million tests, 20,000 are expected to be significant at the 5%

level by chance alone. It is doubtful whether any method could meaningfully discriminate between tests that are significant at $P < 10^{-4}$ or at $P < 10^{-6}$, even with 1000 cases and 1000 controls. Methods such as the **false discovery rate** (Storey and Tibshirani 2003) provide an excellent way to conceptualize the meaning of results when so many thousands of tests have been done (see Chapter 2). Nevertheless, actually identifying which gene or genes have a the true biological effect is a tremendous challenge.

While there is clearly room for development of novel statistical procedures that better accommodate multiple testing, it is widely believed that the best recourse is to verify the main results in an independent sample. The strategy of independent validation gives rise to various two-tier and even three-tier strategies (see, for example, Maraganore et al. 2005). The idea that a second test with a sample taken from another geographical location provides a truly independent replication is itself problematic, since no two samples are completely independent of each other, and neither are they replicates because there are differences in both genotype frequencies and environmental impacts. Many genome-wide association tests are being carried out, and there is reason for cautious optimism that important discoveries will emerge. The abundance of data will also make it possible to incorporate reliable population genetic inference to assess effects due to the underlying patterns of genetic variation.

10.7 SEEKING SIGNATURES OF HUMAN-SPECIFIC GENETIC ADAPTATIONS

Based on genome-wide genetic data from humans, chimpanzee, and additional outgroup species, several statistical approaches allow tests for the past action of natural selection on the human lineage that is distinct from that in other lineages. Such studies constitute the beginnings of a systematic discovery of human-specific adaptive changes. In other words, these studies seek to identify genes that have responded to the unique selection pressures acting in past human populations that ultimately made us who we are today. As exciting as this agenda may seem, it offers no definitive proof that any particular genes changed in a way crucial to the evolution of human-specific attributes, mostly because the statistical tests cannot be followed up by a direct experimental confirmation. Instead, the tests produce a set of candidate genes for further inquiry into the connection between genotypic and phenotypic variation. We have already been introduced to some of the key tests that have been applied; however, some have required adjusting to unique features of human populations, for example, past demographic history.

Interspecific Divergence

Although genome scans for interspecific divergence do not rely on population genetics, they do rely on models of molecular evolution, and so we will

discuss them briefly here. With an alignment of sequences from only one individual from each of several species, tests of selection can be based on a model that posits that the rates of substitution are the same under neutrality with different classes of sites possibly having different rates. In Chapters 7 and 9 we encountered tests that specifically contrasted rates of substitution at synonymous and nonsynonymous sites in protein-coding genes. The codon-substitution model is the most highly developed of these tests, and it has been applied to the detection of human-specific selection on genome-wide sets of alignments (Clark et al. 2003; Nielsen et al. 2005b).

When the contrast between synonymous and nonsynonymous sites is unreliable, but the genome sequences map onto a phylogeny, then branch-specific rates of divergence can be estimated and tests of local rates of acceleration or deceleration carried out. By combining the models of substitution in a phylogeny through time with hidden Markov-chain models that scan the sequence to ascertain whether contiguous sites along the genome evolve similarly, Siepel and Haussler (2004a) developed a powerful approach to identify genomic regions that showed signatures consistent with natural selection acting either to conserve or to accelerate substitutions in specific genomic regions.

The Siepel and Haussler phyloHMM tool was applied to a multi-alignment of mammalian genomes with the specific goal of identifying segments that displayed exceptional acceleration in humans and relative stasis among other mammals (Pollard et al. 2006). An RNA-encoding gene denoted *HAR1* was identified that displayed a conserved secondary structure across mammals. *In situ* hybridizations showed that the gene is expressed in the fetal brain, and there are suggestions that it plays a role in early development. From the observations of sequence alone, an accelerated rate of mutation in the human lineage cannot be ruled out. The inference of selection driving the accelerated evolution depends on the assumption of a constant rate of mutation. To investigate the mutation rate, the authors examined resequence data from the region to determine whether the pattern of polymorphism was any different from other regions; they found that hyper-mutability of the region could be excluded. The analysis of polymorphism data, particularly in the contrast of polymorphism to divergence, is a powerful approach for disentangling mutational and selective effects. As we shall see in the next sections, several approaches make effective use of this contrast.

McDonald-Kreitman and Poisson Random Field Tests

The McDonald-Kreitman test (see Chapter 7) is easy to implement and hence can readily be carried out on a genome-wide scale. The necessary data consist of genome-wide polymorphism identified from at least one species and a sequence from at least one outgroup. The test requires SNPs in two functional categories, usually synonymous vs. nonsynonymous SNPs, but other functional categories can be used as well. By contrasting the counts of poly-

morphic SNPs that are synonymous or nonsynonymous with counts of fixed divergent sites that are synonymous or nonsynonymous, a Fisher exact test (or, for large samples, a 2 × 2 chi-square test) reveals whether nucleotide polymorphisms and nucleotide substitutions exhibit the same relative proportions of synonymous and nonsynonymous sites. Assuming that the silent sites are neutral or nearly so, the neutral theory says that the number of nucleotide substitutions is expected to be $2\mu t$, where μ is the neutral mutation rate and t is the time back to the common ancestor of the species being compared. The count of neutral polymorphic sites comes from the infinite-sites model, which says that the expected number of segregating neutral sites is

$$\theta \sum_{i=1}^{n-1} \frac{1}{t}$$

Since some parameters this model can be estimated jointly across many genes, an application of the Poisson Random Field (Sawyer and Hartl 1992) has particular advantages. In this approach, every SNP is considered as independent, and so counts in each cell of the McDonald-Kreitman table follow a Poisson process whose parameters are to be estimated. The assumption of independence greatly simplifies parameter estimation and hypothesis testing. One of the key parameters is the strength of selection, written as $\gamma = 2Ns$. Bustamante et al. (2005) applied this analysis to data based on the sequences of PCR products from all known exons in 39 humans and a single chimpanzee. The data showed an overwhelming deficit of nonsynonymous fixations (i.e., an excess of nonsynonymous polymorphism) consistent with a strong bias toward purifying selection acting on protein-coding variants. That is, the human population contains a considerable burden of segregating variation that is deleterious. Aside from this signature of negative selection, a number of genes, particularly genes involved in functions like immunity and male fertility, displayed a signature of positive selection.

Local Distortions in Linkage Disequilibrium

If a mutation occurs with a strongly beneficial effect, the resulting selective sweep will produce a distortion in the local linkage disequilibrium (Maynard Smith and Haigh 1974; Przeworski 2002). Tests that contrast the local pattern of linkage disequilibrium and/or haplotype structure have been instructive in identifying particular alleles of genes that appear to be in the midst of a sweep. Hudson et al. (1994) showed that an apparently advantageous allele encoding superoxide dismutase (SOD) in *Drosophila* was at too high a frequency relative to the low diversity of haplotypes containing that allele. Andolfatto et al. (1999) also found an example of reduced haplotype complexity. The identification of positions of selective sweeps based on patterns of local diversity (Kim and Stephan 2002) was extended to include the evidence of distorted local linkage disequilibrium (Kim and Nielsen 2004). An alternative method of analysis, based on an examination of haplotypic vari-

ation surrounding a gene, was initially applied to the genes encoding G6PD and the CD40 ligand (Sabeti et al. 2002). The **extended haplotype homozygosity** (EHH) test identifies core haplotypes that extend a greater distance than expected from neutral coalescent simulations . The extended haplotype test was generalized to a genome-wide scale by Voight et al. (2006) and revealed numerous locations in the genome whose haplotype structure was not consistent with the standard neutral model.

F_{ST} *Tests*

If humans in Africa and humans in Europe faced different environmental stresses, then in principle there must be selection pressures driving some of the differential changes in allele frequency between these groups. This idea was initially formalized by Lewontin and Krakauer (1973). With genomic-scale data, the idea of calculating F_{ST} for each SNP, gene, or genomic window has an intuitive appeal based on the logic that regions of the genome with highly inflated values of F_{ST} might have attained that state through region-specific differences in selection pressures. Determination of the *P*-values is a problem because the full demographic history is unknown, but Akey et al. (2002) reasoned that the full null distribution of F_{ST} across genes would itself provide a sort of empirical null distribution, and so F_{ST} values in the extreme tail of the distribution could be judged as significant. While this test does appeal because of its simplicity, and can be shown to work when all the assumptions in the model are true, many seemingly minor violations of the assumptions seriously distort the test (Teshima et al. 2006). For example, if one of the populations (e.g., Europeans) went through a population bottleneck, then the random drift that takes place during the bottleneck results in a huge inflation in the variance of F_{ST} across sites.

Genome Scans for Selection—Skewed Site Frequency Spectrum

We saw in Chapters 7 and 9 that natural selection is expected to distort the spectrum of frequencies of variants within genes. Purifying selection drives variants to lower frequency, and positive selection produces an excess of high-frequency derived alleles. Simple tests of the correspondence of the site frequency spectrum to that expected under neutrality, such as Tajima's *D*, have been applied to human polymorphism data, but these tests suffer the same problem that the F_{ST} tests do, namely that the complexity of past human demographic history can seriously distort the results. One solution that seems very promising is to use putatively neutral sites to fit a model of demographic history, and to then assume that the genetic variation across the entire genome experienced this same demographic history, thereby enabling potentially non-neutral sites to be tested for their departure from neutrality conditional on the demographic model (Williamson et al. 2005, Figure 10.19). A key finding is that nonsynonymous variation shows a marked skew

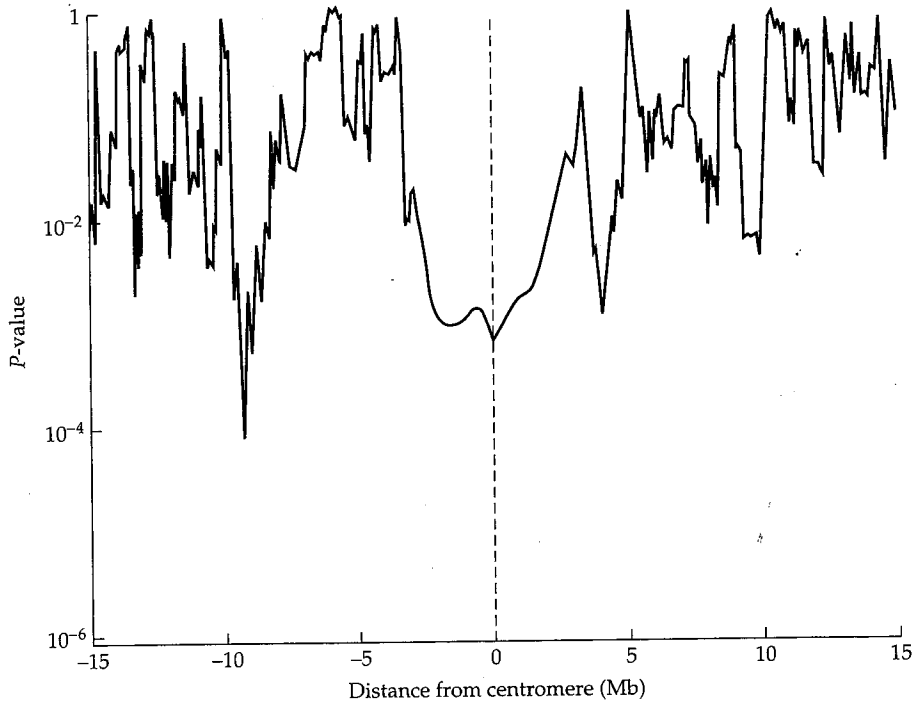


FIGURE 10.19 A genome scan for departures from selective neutrality of SNPs from a European population based on fitting the site frequency spectrum to a neutral demographic model. This scan includes a 30-Mb region of chromosome 8 centered on the centromere. (Data from Perlegen Biosciences, courtesy Scott Williamson.)

toward more rare variants than expected from demography alone, consistent with negative or purifying selection.

10.8 HUMAN ORIGINS

What might population genetics contribute to understanding of early human origins? Because these events occurred so long ago, how could extant variation in modern human populations be informative about such ancient events? If the important events in human history were too distant in the past, then studies of present-day genetic variation would not be informative, but in fact, major events in human evolutionary history appear to have occurred within the time horizon within which population genetic methods still allow reliable inferences. Recall from our discussion of the neutral theory that the expected time to fixation of a new mutation destined to be fixed is $4N_e$ gener-

ations (see Chapter 3). Various lines of evidence converge on estimates of the effective size of the human population at about 10,000 individuals. At 20 years per generation with $N_e = 10{,}000$, the value of $4N_e$ equals 800,000 years. Although *Homo erectus* fossils dating back to more than 1 million years are found in Africa, Asia, and Europe, the lineage that apparently gave rise to modern humans left no fossils outside of Africa until around 100,000 years ago. Furthermore, compelling population genetic evidence implies that non-African human populations arose from individuals who migrated out of Africa.

Consider a scenario in which modern humans originated in Africa and then subsequently spread from Africa. What pattern of genetic variation might be expected to be consistent with this migration history? Provided that the fraction of the population that left Africa was not too large, and that the subsequent effective population sizes in Africa and outside of Africa were not too different, this migration scenario should give a signature in which the most recent common ancestral haplotypes would be more likely to fall in Africa, and African population should also exhibit greater genetic diversity and less linkage disequilibrium. The first evidence supporting these expectations emerged from mitochondrial DNA in a classic paper by Cann et al. (1987), who showed convincingly that patterns of restriction-site polymorphisms among 147 people produced an mtDNA genealogy whose deepest lineages all led back to Africa (Figure 10.20). Furthermore, the time of common ancestry of all of the mitochondrial RFLP variants was only about 200,000 years ago, using the chimpanzee as an outgroup to calibrate the molecular clock.

In the popular press, the inferred ancestor of mtDNA was promptly and unfortunately called the mitochondrial "Eve," when in reality the human population at that time numbered perhaps in the hundreds of thousands. The mitochondrial "Eve" was not unique, but lucky: Her particular mtDNA became the common ancestor to all present today, as coalescent theory implies must always be true for some lucky sequence when there is no recombination. Subsequent studies, including analysis of multiple complete sequences of the mitochondrial genome gave a consistent result in implying that all extant humans have a mtDNA genome that derived from a single common ancestor that lived as recently as just 120,000 years ago. The convergence of inference led to renewed consideration of the possibilities for the modern origin of humans. Even though mtDNA variants do not appear to be strictly neutral (Mishmar et al. 2003; Kivisild et al. 2006), the coalescence date of somewhat more than 100,000 years appears to hold.

The human Y chromosome, another non-recombining unit but in this case transmitted in a male-male fashion, also has its deepest root in Africa (Figure 10.21). For the Y chromosome, the time to most recent common ancestor has a confidence interval of 35,000–89,000 years (Underhill et al. 2000, Figure 10.21). Many other studies of genetic variation in the autosomes and the X

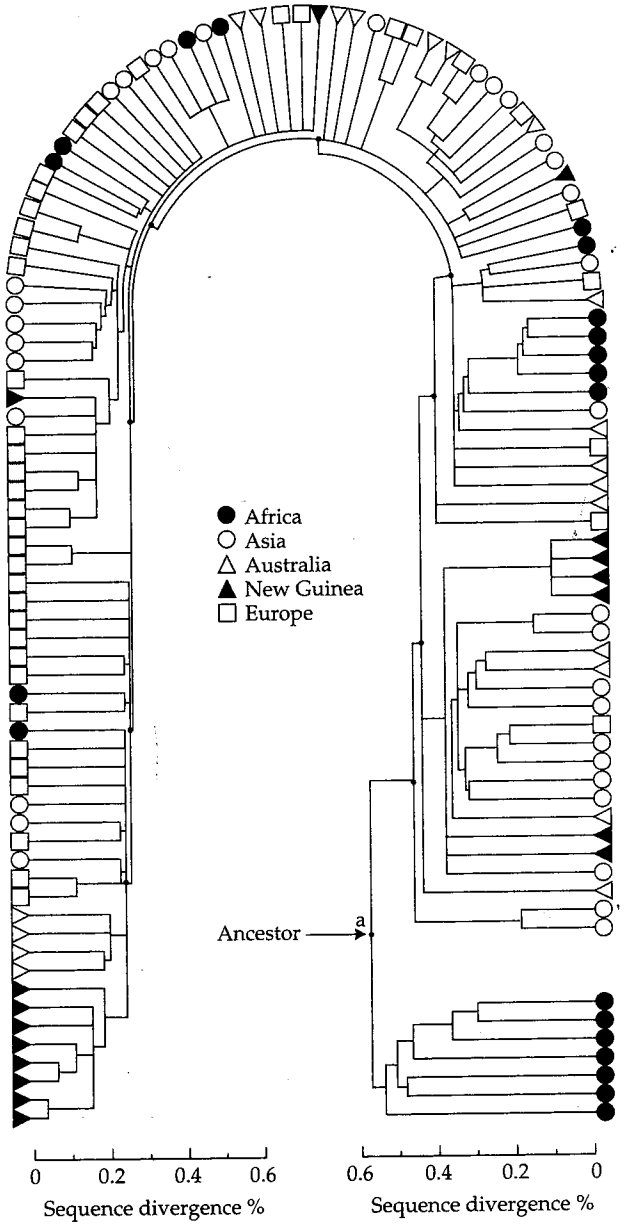


FIGURE 10.20 Maximum parsimony tree of human mitochondrial DNA based on a set of restriction sites. The key result is that the deepest node (labeled a) has a branch that is composed entirely of Africans, which identifies the continent of origin of all human mitochondrial DNA. This observation provided the first clear molecular evidence that an African population is ancestral to all other human populations. (From Cann et al. 1987.)

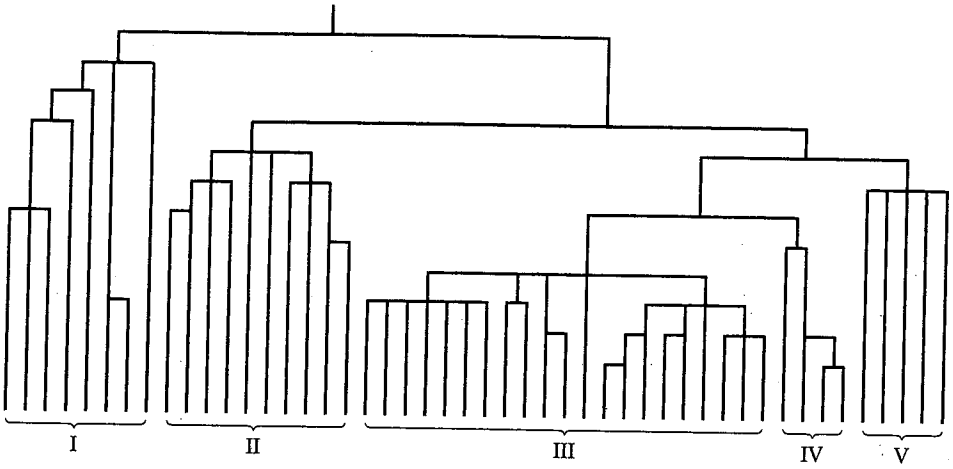


FIGURE 10.21 Bifurcating tree of a subset of polymorphic markers on the Y chromosome. The full phylogenetic tree of the sequences indicates a clear ancestral origin in Africa. (Data from Underhill et al. 2000.)

chromosome have also shown that African populations harbor greater genetic diversity than do non-African populations, and that much of the genetic diversity that appears outside of Africa is consistent with a founder effect from an out-of-Africa migration. When individual genes are examined, some show a pattern in which the African haplotypes are not the oldest, but this finding is expected occasionally as a result of the sampling process (Takahata et al. 2001). In any event, many and varied studies now support the out-of-Africa hypothesis, and at this stage in human genetics it appears that only subtleties about the timing of some migration events remain unknown. If you go back far enough, "we are all Africans," as Charles Rotimi has aptly stated.

Despite consistent evidence in favor of a relatively recent African origin, it remains possible that at least some small segments of the human genome may have an ancestry that extends deeper in time. Although the genetic data show greatest human genetic diversity in Africa, sometimes the deepest root for particular genes are non-African, and for much of the autosomal genome, the time to the most recent common ancestor exceeds one million years (Garrigan and Hammer 2006). On the other hand, even if the out-of-Africa hypothesis were true, some segments of the genome would be expected to have such a deep coalescence. A more direct approach to assessing whether all extant variation in humans derived from a population that emerged from Africa 120,000 years ago presented itself when mtDNA from a Neanderthal specimen was amplified by PCR and sequenced (Krings et al. 1999). In this

study, the time to common ancestry of Neanderthal and contemporary human mtDNA was estimated to have a 95% confidence interval of 317,000–741,000 years. This estimate was taken as evidence that Neanderthal mtDNA and that of modern humans evolved independently from a time long preceding the out-of-Africa migration. Nordborg (1998) has however pointed out that even though the results do fail to find evidence for common ancestral lineages, the statistical tests have very low power to conclude that admixed ancestry of some genes could not be possible. More recent Neanderthal samples do continue to suggest a lack of common ancestry with modern humans, but still fail to exclude the possibility of some admixture (Serre et al. 2004).

PROBLEM 10.5 Suppose that modern humans actually had 1% of their genome admixed from Neanderthal contributions. Further suppose that there is one region of the genome where the Neanderthal and modern human haplotypes can be unambiguously distinguished. How many modern humans would have to be sampled to have a 95% probability of detecting the admixture?

ANSWER To fail to detect the admixture, the modern human haplotype, whose frequency is 0.99 would have to be drawn exclusively. For a sample size of k alleles, this probability is $(0.99)^k$. We want only a 5% probability for failing to detect the admixture, so $0.05 = (0.99)^k$. Taking logs and solving, we get $k = 298$. It is not so easy to have high confidence that there was zero admixture.

Neanderthal Genome Sequence

One of the most exciting promises of novel high-throughput genome sequencing technologies based on simultaneous sequencing of many short oligonucleotides is the ability to sequence degraded DNA such as that recovered from ancient and archival samples. With these technologies, even the DNA amplified from Neanderthal bone specimens is of sufficient quality to obtain useful sequence (Green et al. 2006; Noonan et al. 2006). Fortunately, the Neanderthal genome is sufficiently divergent from modern humans that it is usually easy to identify contamination by modern human samples. The possibility that a complete Neanderthal genome sequence may serve as a starting point to identification of genetic differences responsible for unique features of modern humans serves as a strong motivation to complete the task. A new era in human population genetics and the understanding of human evolutionary history is now beginning to emerge.

SUMMARY

1. More than ten million SNPs have been mapped on the human genome, and population frequencies of more than 6 million are known, thanks to the HapMap Project.
2. The average level of nucleotide diversity in humans is $\pi = 0.0008$.
3. Analysis of SNP data, as opposed to full DNA sequence data, requires some caution, as there may be strong ascertainment bias in the initial discovery of the SNPs.
4. The human site frequency spectrum for SNPs shows a clear excess of rare variants, consistent with both population growth and with selection against deleterious mutations.
5. Human linkage disequilibrium halves roughly every 40 kb, but this figure varies widely across the genome.
6. Roughly 80% of all recombination occurs in 20% of the genome, due both to recombination hot spots and local variation in the background recombination rate.
7. Linkage disequilibrium decays over a shorter physical span along the chromosome in Africans in a way that is consistent with African populations having a larger long-term effective population size.
8. Human genetic variation shows evidence for a strong haplotype structure, due largely to the small effective size and past demographic history (especially bottlenecks). Most human genetic variation is found within groups, and only about 5% of the variation occurs as differences among groups.
9. Multilocus methods are highly effective at detecting population substructure and individual admixture.
10. The exceptional allelic heterogeneity of many Mendelian disorders suggests that complex traits may also harbor such allelic heterogeneity.
11. Methods for genome-wide association testing demand novel ways to deal with the multiple testing problem, but still they show promise in being able to identify important genes that harbor mutations that inflate risk of common chronic diseases.
12. Many methods exist and more are being developed to make inferences about past roles of natural selection in shaping human genetic variation. These methods identify both positively and negatively selected genes.
13. Population genetics provides tools for analysis of demographic episodes and bouts of natural selection over nearly the past one million years, back to the time of emergence of our species.

PROBLEMS

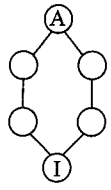
1. Explain why linkage disequilibrium mapping is, in principle, more powerful than conventional mapping using pedigrees.
2. A *founder effect* occurs when a subpopulation is formed from a relatively small number of founding individuals because the sampling variance perturbs the gametic frequencies. What likely effect would such a founder event have on the efficiency of linkage disequilibrium mapping?
3. How does lineage sorting undermine the inference that, when chimpanzees are fixed for a particular nucleotide of a SNP and humans are polymorphic for the SNP, then a mutation must have taken place in the human lineage?
4. In the ascertainment of SNPs, we have seen that the probability of discovering a SNP with two alleles in a sample of two gametes equals $2pq$. What is the chance of discovering a SNP in with two alleles in a sample of three gametes? In a sample of four gametes?
5. In a statistical test for which the null hypothesis is actually true, explain why the expected distribution of P-values is a uniform distribution.
6. What feature of the first phylogenetic tree of human mitochondrial DNA sequences suggested that all human populations derive ultimately from a common ancestral population that lived in Africa?
7. Among human populations sampled from major continental areas on Earth, in which samples is linkage disequilibrium found to decrease the most rapidly as the distance between markers in the genome increases? According to the usual explanation, why does linkage disequilibrium have a shorter range in this continental population than in others?
8. If the mutation rate in human males equals 5 times that in human females (i.e., $\alpha = \mu_m/\mu_f = 5$), then what is the expected ratio of internal branch lengths of gene trees for Y-linked genes versus autosomal genes? What is the expected ratio of branch lengths of gene trees for X-linked genes versus autosomal genes?
9. What value of $4N_e c$ is required for random genetic drift to produce a steady-state value of $E(r^2) = 0.4$, where $E(r^2)$ is the expected value of the squared correlation coefficient of the alleles of two loci present in gametes?
10. If an individual is found to be heterozygous for two linked SNPs, say A/a and B/b, what does it mean to say that the linkage phase of the SNP alleles is unknown?
11. For two genes each with two alleles in a population undergoing random mating, suppose that the gametic frequencies are P_{AB}, P_{aB}, P_{Ab}, and P_{ab}.

What are the expected frequencies of the two linkage phases of the double heterozygous genotypes *A B/a b* and *A b/a B*? Under what condition are the frequencies of the linkage phases equal? What does this condition imply about the value of r^2?

12. Why does a larger inbreeding coefficient increase the statistical power of homozygosity mapping?

13. Transmission disequilibrium tests were carried out to detect genetic risk factors associated with polycystic ovary syndrome (Stewart et al. 2006). Previous studies had suggested an association between the ovarian condition and allele 8 of a genetic marker *D19S884* on chromosome 19. In the transmission disequilibrium tests, among 98 cases with an affected daughter among the offspring, transmission of the *D19S884* allele 8 was observed 59 times and transmission of allele 8 not observed 39 times. Do these results confirm a significant association between allele 8 of *D19S884* and polycystic ovary syndrome?

14. In the pedigree shown here, individual I is the result of mating between half-first cousins related through the common ancestor A. For two autosomal SNPs selected by a recombination distance of $c = 1$ cM (about 1 Mb), what is the probability that individual I has identity by descent for both SNPs and the entire region between them? Assume that A is not inbred.



15. This problem demonstrates an extreme example of the counterintuitive principle that two linked genes can be in linkage disequilibrium with each other even though neither is in linkage disequilibrium with a gene located between them. For this purpose, imagine three linked genes that occur in the order *A B C* along the chromosome. Each gene has two alleles in the population, and if we denote the alleles by uppercase and lowercase letters, then there are eight possible gametic types, namely *A B C*, *A b C*, *a B C*, *a b C* for the *C*-bearing gametes and *A B c*, *A b c*, *a B c*, *a b c* for the *c*-bearing gametes. What frequencies of these gametic types would result in the disequilibrium parameter $D_{AB} = 0$ and $D_{BC} = 0$ but $D_{AC} = D_{max}$. That is, there is complete linkage equilibrium between the genes *A* and *B*, and complete linkage equilibrium between genes *B* and *C*, but there is the maximum possible linkage *disequilibrium* between genes *A* and *C*. (Hint: Four of the gametic types have frequency 0.)

ANSWERS TO CHAPTER-END PROBLEMS

CHAPTER 1

1. The sum of the x values equals 10,134 and that of the squared x values equals 1,053,748. The mean in the sample is 101.34 and the mean square in the sample equals 10,537.48. The mean of the entire population is therefore estimated as 101.34, the variance is estimated as $s^2 = 10{,}537.48 - (101.34)^2 = 267.684$, and the standard deviation is estimated as $s = 16.36$.

2. If the phenotypic values are normally distributed, 68% are expected to be within one standard deviation from the mean. This implies that $(1 - 0.68)/2 = 16\%$ are expected to have a phenotypic value greater than one standard deviation above the mean, and 16% are expected to have a phenotypic value smaller than one standard deviation below the mean. In this case, the mean plus one standard deviation equals 117.70, which rounds to 118. In the sample itself, 16 are therefore expected to have a phenotypic value ≥ 18, and the actual count is 17. Likewise, 16 are expected to have a phenotypic value smaller than the mean minus one standard deviation, which equals $101.34 - 16.36 = 84.98 \approx 85$. Here again the actual count is 17.

3. If the phenotypic values are normally distributed, 95% are expected to be within two standard deviations from the mean. This implies that $(1 - 0.95)/2 = 2.5\%$ are expected to have a phenotypic value greater than two standard deviations above the mean, and 2.5% are expected to have a

phenotypic value smaller than two standard deviations below the mean. In this case, the mean plus two standard deviation equals 134.06, which rounds to 134. In the sample itself, 3 individuals have a phenotypic value exceeding 134, where 2 or 3 would be expected. At the other end of the scale, the mean minus two standard deviations equals 68.61, which rounds to 69, and the actual count of individuals with phenotypic values smaller than 69 is 1.

4. This is an example of the central limit theorem in action. Adding 12 uniform random numbers gives a resulting sum whose distribution is approximately normal.

5. A positive correlation means that the successive random numbers are more similar than expected with independence. This would result in a sum with a smaller variance than expected under the central limit theorem.

6. Most likely the gene is in the X chromosome (X-linked). Females have two X chromosomes and so can be either homozygous or heterozygous, whereas males have only one X chromosome (plus a Y chromosome) and so cannot be heterozygous for a gene in the X chromosome.

7. Heterozygotes would produce both F and S polypeptide chains, which could be assembled into dimers that are FF, FS, and SS. These three molecules would most likely have different electrophoretic mobility, giving three bands in the protein gel.

8. The fast allele is the one whose protein moved fastest through the gel, so the protein band is nearest the bottom. The results indicate 116 *FF*, 98 *FS*, and 26 *SS* genotypes in the sample. The allele frequency p of F is estimated as $(2 \times 116 + 98)/(2 \times 240) = 0.6875$ and the allele frequency q of S as 0.3125. The expected number of each genotype are $p^2 \times 240$, $2pq \times 240$, and $q^2 \times 240$, which work out to 113.44 *FF*, 103.12 *FS*, and 23.44 *SS*. The observed values are 116, 98, and 26, and so the fit seems to be satisfactory.

9. The allele frequencies in the sample are $p = q = 0.5$, and the expected genotype frequencies with Hardy-Weinberg equilibrium are 24.75 for each homozygous genotype and 49.50 for the heterozygous genotype. The sample seems to include too few heterozygous genotypes to be compatible with Hardy-Weinberg proportions. (In Chapter 2 you will learn how to calculate that the probability of obtaining a fit as bad or worse because of chance alone is less than one in 1000.)

10. With Hardy-Weinberg proportions, $P = p^2$, $Q = 2pq$, and $R = q^2$, where p and q are the allele frequencies of A and a. Then $PQ = p^2q^2$ and $Q^2 = 4p^2q^2$, and so $PQ = Q^2/4$.

11. The polymerase chain reaction doubles the amount of a specific fragment of DNA each round (assuming perfect efficiency). This means that 30 rounds of PCR will yield $2^{30} = 1.074 \times 10^9$ copies. In practice, PCR is rarely more than 80% efficient.

12. A sample size of 4 allows $(4 \times 3)/2 = 6$ comparisons. The total number of mismatches is 30, and so the average number of pairwise mismatches in this sample is $\frac{30}{6} = 5$.

13. The sequences are said to be the complete set of polymorphic nucleotides, hence $S = 20$ segregating sites. For these 20 sites, there are 10 sites with the configuration (5, 1), where the first number represents the majority nucleotide and the second the minority nucleotide, 8 sites with the configuration (4, 2), and 2 sites with the configuration (3, 3). The (5, 1) configuration results in 5 pairwise mismatches, the (4, 2) configuration in 8 pairwise mismatches, and the (3, 3) configuration in 9 pairwise mismatches. Altogether there are $(10 \times 5) + (8 \times 8) + (2 \times 9) = 132$ mismatches among 15 pairwise comparisons, for an average of $\Pi = 8.80$. To estimate θ we need the sum $a = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} = 2.28$. The estimate of θ based on S equals $S/a = 8.77$, and that based on Π equals 8.80. These are obviously very close. If these results were found among a larger sample, they would be consistent with the infinite sites model with selectively neutral alleles.

14. If the DNA fingerprints do not match, one can be certain that the sample from the crime scene did not come from the suspect, unless there are errors in handling the evidence or mistakes in the laboratory. This situation is known as exclusion. Such information has resulted in the release of many wrongly accused people. On the other hand, if the DNA fingerprints do match, then either the suspect is the source of the sample from the crime scene, or someone else with the same genotype at the loci examined was the source of the sample. The chance that the suspect is the culprit now depends on how often one is likely to obtain such a match by chance. Thus the interpretation of matching DNA fingerprints, and the strength of the evidence, hinges on issues in population genetics.

CHAPTER 2

1. If the alleles of the other gene, say *A* and *a*, are in random association (linkage equilibrium) with the alleles of the sex-determination gene, then matings of $Mm \times mm$ will occur in random proportions with respect to the *AA*, *Aa*, and *aa* genotypes, and therefore *AA*, *Aa*, and *aa* will be in the Hardy-Weinberg proportions p^2, $2pq$, and q^2. To see this for yourself, let the frequencies of *A* and *a* be p and q, respectively, among both *M*-bearing and *m*-bearing chromosomes. Then *Mm* males will have genotypes *AA*, *Aa*, and

aa at frequencies p^2, $2pq$, and q^2, and *mm* females will have genotypes *AA*, *Aa*, and *aa* at frequencies p^2, $2pq$, and q^2. Random mating between males and females therefore results in random mating between the *AA*, *Aa*, and *aa* genotypes, and maintains the Hardy-Weinberg frequencies.

2. The genotype frequencies are the numbers given in the table multiplied by 10^{-6}, and therefore for trait (a), for example, the frequency of the homozygous recessive is 0.002786. The estimated frequency of the recessive allele q is the square root of the frequency for homozygous recessives, and the expected frequency of heterozygous genotypes is $2q(1-q)$. The values of q and $2q(1-q)$ for the traits are: (a) 0.0528 and 0.1000 (1 in 10 individuals are heterozygous), (b) 0.0256 and 0.0500 (1 in 20), (c) 0.0169 and 0.0333 (1 in 30), (d) 0.0126 and 0.0250 (1 in 40), and (e) 0.0101 and 0.0200 (1 in 50).

3. The chi-square values and their associated P values are: (a) 2.98 (P = 0.084), (b) 7.63 (P = 0.006), (c) 3.63 (P = 0.057), (d) 5.56 (P = 0.018). The P value for population (d) is significant and that for population (b) is highly significant, so in these cases the hypothesis of Hardy-Weinberg proportions should be rejected. The P values for populations (a) and (c) are both greater than 0.05, and so there is no reason to reject the hypothesis in these cases.

4. The normal allele for irregular flowers is dominant because of the approximate 3 normal : 1 peloric ratio in the F_2 generation. The chi-square value equals $(88 - 93.75)^2/93.75 + (37 - 31.25)^2/31.25 = 1.41$ with one degree of freedom (since no parameters were estimated from the data). The associated probability level is 0.235, so the hypothesis of a 3 : 1 ratio cannot be rejected.

5. Expected ratio obtained from the expansion of $[(\frac{3}{4})D + (\frac{1}{4})R]^3$, where D and R represent the dominant and recessive phenotypes, respectively, yielding the ratio 27 : 9 : 9 : 9 : 3 : 3 : 3 : 1 for each 64 progeny. Expectations are 269.6 : 89.9 : 89.9 : 89.9 : 30.0 : 30.0 : 30.0 : 10.0 among 639 progeny, for a chi-square of 2.67 with seven degrees of freedom (because there are eight classes of data and no parameters estimated), for which the associated probability value is about 0.914. This is a very good fit, indeed.

6. The frequency of *AA* equals $p^2 = (1-q)^2 = 1 - 2q + q^2 \approx 1 - 2q$; frequency of *Aa* heterozygotous genotypes equals $2pq = 2(1-q)q = 2q - 2q^2 \approx 2q$; and frequency of *aa* equals $q^2 \approx 0$.

7. Let the recessive allele frequency be q, which equals the frequency of affected males. Then the frequency of heterozygous carrier females

equals $2pq = 2(1 - q)q = 2q - 2q^2 \approx 2q$, because when the recessive allele is rare, then $q^2 \approx 0$. Therefore, the frequency of carrier females is approximately two times the frequency of affected males. For $q = 0.1$, the exact frequency of heterozygous carrier females equals $2(0.1)(0.9) = 0.18$.

8. The expected genotype frequencies with random mating are given by expansion of $(pA + qa)^4$ instead of $(pA + qa)^2$. The result is $(\frac{1}{16})$ *AAAA*, $(\frac{4}{16})$ *Aaaa*, $(\frac{6}{16})$ *AAaa*, $(\frac{4}{16})$ *Aaaa*, and $(\frac{1}{16})$ *aaaa*.

9. (a) $p^2 + pq = p(p + q) = p$; (b) $p - q = (1 - q) - q = 1 - 2q$; (c) $p^2 + q^2 = 1 - 2pq$ follows from the equality $(p + q)^2 = p^2 + 2pq + q^2 = 1$; (d) $(p - q)^2 = p^2 - 2pq + q^2 = (p^2 + q^2) - 2pq = 1 - 2pq - 2pq = 1 - 4pq$.

10. The allele frequency estimates are $p_1 = 0.123$, $p_2 = 0.473$, $p_3 = 0.161$, and $p_4 = 0.243$. Assuming Hardy-Weinberg frequencies, the expected numbers, from left to right across the gel are 15.13, 116.36, 223.73, 39.61, 152.31, 25.92, 59.78, 229.88, 78.25, and 59.05. The chi-square value equals 11.20 and there are 6 degrees of freedom (9 to start with minus 1 for each of three allele frequencies that need to be estimated from the data). The corresponding P value is 0.082, which says that 8.2% of the time chance alone would produce a chi-square value equal to or greater than 11.20. This value is not significant, and there is no reason based on these data to reject the model of Hardy-Weinberg proportions.

11. In a random mating population, the genotypes *AA* and *Aa* occur in the frequencies p^2 and $2pq$, and among these the allele frequency of *a* equals $p^2 \times (0) + 2pq \times (\frac{1}{2}) = pq$. Hence the allele frequency of *a* among individuals with the dominant phenotype equals $pq/(p^2 + 2pq) = q/(1 + q)$. Among matings, there are three types of dominant × dominant matings, namely, *AA* × *AA*, *AA* × *Aa*, and *Aa* × *Aa*, which occur in the proportions p^4, $4p^3q$, and $4p^2q^2$. Only the last can produce homozygous recessive offspring, and the frequency of such offspring is $\frac{1}{4}$. Hence, the proportion of homozygous recessive offspring resulting from dominant × dominant matings equals $p^2q^2/[p^4 + 4p^3q + 4p^2q^2]$, which simplifies to $[q/(1 + q)]^2$. Likewise, there are two types of dominant × recessive matings, namely, *AA* × *aa*, and *Aa* × *aa*, which occur in the proportions $2p^2q^2$ and $4pq^3$. Only the last can produce homozygous recessive offspring, and the frequency of such offspring is $\frac{1}{2}$. Hence, the proportion of homozygous recessive offspring resulting from dominant × recessive matings equals $2pq^3/[2p^2q^2 + 4pq^3]$, which simplifies to $q/(1 + q)$. The ratios $[q/(1 + q)]^2$ and $q/(1 + q)$ are Snyder's ratios.

12. With linkage equilibrium, the expected gametic frequencies are given by the products of the allele frequencies of the two genes, or in this case by

$(x_1 A_1 + x_2 A_2)(y_1 B_1 + y_2 B_2 + y_3 B_3)$. The alleles in gametes and their frequencies for the numerical example given are A_1B_1 (0.3)(0.2) = 0.06; A_1B_2 (0.3)(0.3) = 0.09; A_1B_3 (0.3)(0.5) = 0.15; A_2B_1 (0.7)(0.2) = 0.14; A_2B_2 (0.7)(0.3) = 0.21; and A_2B_2 (0.7)(0.5) = 0.35.

13. (a) Assuming linkage equilibrium, the frequencies of the gametes *A B*, *A b*, *a B*, and *a b* are $P_{AB} = p_A p_B$, $P_{Ab} = p_A q_b$, $P_{aB} = q_a p_B$, and $P_{ab} = q_a q_b$, or (0.7)(0.3) = 0.21, (0.7)(0.7) = 0.49, (0.3)(0.3) = 0.09, (0.3)(0.7) = 0.21, respectively. (b) The theoretical maximum of D equals the smaller of $p_A q_b$ (0.49) and $q_a p_B$ (0.09), hence D_{max} = 0.09. A value of D equal to 50 percent of 0.09 yields D = 0.045. The gametic frequencies required for this value of D are given by $P_{AB} = p_A p_B + D$, $P_{Ab} = p_A q_b - D$, $P_{aB} = q_a p_B - D$ and $P_{ab} = q_a q_b + D$, or P_{AB} = 0.21 + 0.045 = 0.255, P_{Ab} = 0.49 – 0.045 = 0.445, P_{aB} = 0.09 – 0.045 = 0.045, and P_{ab} = 0.21 + 0.045 = 0.255.

14. The calculations are (a) D = –0.0100, D_{min} = –0.1348, D_{max} = 0.0692, D' = 0.0745, r^2 = 0.0028; (b) D = –0.0006, D_{min} = –0.0014, D_{max} = 0.0190, D' = 0.4414, r^2 = 0.0003; (c) D = 0.1726, D_{min} = –0.0630, D_{max} =0.1808, D' = 0.9546, r^2 = 0.8424; (d) D = –0.0088, D_{min} = –0.2451, D_{max} =0.2095, D' = 0.0360, r^2 = 0.0013; (e) D = – 0.1911, D_{min} = –0.2148, D_{max} =0.1549, D' = 0.8897, r^2 = 0.6440. The interpretations are: (a) negligible linkage disequilibrium, (b) substantial linkage disequilibrium in terms of D', but little in terms of r^2, (c) extreme linkage disequilibrium, (d) negligible linkage disequilibrium, and (e) substantial linkage disequilibrium.

15. In every case, $P1_{AA}P1_{ab} - P1_{Ab}P1_{aB}$ and $P2_{AA}P2_{ab} - P2_{Ab}P2_{aB}$ are very close to zero (any slight departure results from round-off error). This calculation verifies that each of the original populations is in linkage equilibrium. Since the admixed population consists of an equal proportion of each of the two original subpopulations, the gametic frequencies in the admixed population are given by the averages, for example, $P_{AB} = (P1_{AB} + P2_{AB})/2$. The results of the analysis are as follows: (a) P_{AB} = 0.2408, P_{Ab} = 0.5120, P_{aB} = 0.0810, P_{ab} = 0.1662, D = –0.0015, D_{min} = –0.1677, D_{max} = 0.0795, D' = 0.0087, r^2 = 0.0001; (b) P_{AB} = 0.3676 P_{Ab} = 0.03515, P_{aB} = 0.21445 P_{ab} = 0.3828, D = 0.1332, D_{min} = –0.2344, D_{max} = 0.1683, D' = 0.7912, r^2 = 0.3031; (c) P_{AB} = 0.0279, P_{Ab} = 0.17935, P_{aB} = 0.16865, P_{ab} = 0.6241, D = –0.0128, D_{min} = –0.0407, D_{max} = 0.1558, D' = 0.3151, r^2 = 0.0063. What these results mean is that the admixed population in (a) shows essentially no linkage disequilibrium, that in (b) shows very substantial linkage disequilibrium, and that in (c) shows considerable linkage disequilibrium as measured by D' but virtually none as measured by r^2.

CHAPTER 3

1. The Wright-Fisher model assumes a population of N diploid, dioecious organisms (so that self-fertilization is theoretically possible), with nonoverlapping generations, in which each organisms has exactly the same chance of reproducing. In the transition matrix, T_{ij} is the probability of the allele frequency changing from $i/(2N)$ to $j/(2N)$ in a single generation of random genetic drift.

2. The Kolmogorov forward equation is based on the magnitude and direction of changes in allele frequency that can occur in the generation immediately preceding the current state. The Kolmogorov backward equation is based on the magnitude and direction of changes in allele frequency that could have occurred in the initial generation of the process.

3. The backward simulation is extremely efficient because each extant allele must trace its ancestry back to the initial population. The forward simulation is extremely inefficient because most lineages going forward in time are destined to become extinct; tracing these lineages requires processing time and memory, although these lineages have no bearing on the current population.

4. For the loss of a single copy in one generation, the probability equals

$$\binom{100}{0}\left(\frac{1}{100}\right)^0\left(\frac{99}{100}\right)^{100} = 0.366$$

and for the loss of two copies in a single generation, the probability equals

$$\binom{100}{0}\left(\frac{2}{100}\right)^0\left(\frac{98}{100}\right)^{100} = 0.133$$

5. The expected heterozygosity in a finite population decreases according to $H_t = [1 - 1/(2N)]H_{t-1}$. In this case, we expect H to go from 0.50 to $0.99 \times 0.50 = 0.495$ in one generation. Since H decreased substantially more than this number, the reduction would appear to require something more than random genetic drift. Given the number of loci, a chi-square test could be done as a formal test of statistical significance.

6. Here Equation 3.14 is the most useful. It states that $H_t = H_0\, e^{-t/(2N)}$, and taking natural logarithm of both sides we obtain $\ln(0.05) = -t/(2N)$. Solv-

ing this for $2N = 10$ yields $t \approx 60$ generations. For a population that is 5 times larger, the time to lose 95% of the original heterozygosity is 5 times longer, or $t \approx 300$ generations.

7. The probability of eventual fixation is $1/(2N)$ where $N = 28$. Hence the probability of eventual fixation is $\frac{1}{56} = 0.018$, and the probability of eventual loss is $1 - (\frac{1}{56}) = \frac{55}{56} = 0.982$. The initial allele frequency is $p = \frac{1}{56}$, and hence the average time to fixation, given eventual fixation, equals $-(4N)[(1-p)/p]\ln(1-p)] = 111$ generations. (Note that this is approximately $4N$ generations.) The average time to loss, given eventual loss, is $-(4N)[p/(1-p)]\ln(p)] = 8.2$ generations. (Note that this is approximately $2\ln(2N)$ generations.]

8. For an X-linked gene, the 28 mice represent 42 copies of the gene, so the probability of ultimate fixation is $\frac{1}{42} = 0.024$ and that of ultimate loss is $1 - 0.024 = 0.976$. The effective size of a population for an X-linked gene equals $\frac{3}{4}$ that of an autosomal gene, and so $N = (\frac{3}{4})28 = 21$. The conditional average times to fixation and loss are therefore 83.0 generations and 7.5 generations. For a Y-linked gene, the population has only 14 copies, and the effective population size for a Y-linked gene equals $(\frac{1}{4})56 = 14$. The probabilities of ultimate fixation or loss equal 0.071 and 0.929, respectively, and the corresponding conditional average times are 54.0 and 11.4 generations, respectively.

9. Use Equation 3.14 to obtain $\ln(0.50) = -t/2N$ with $t = 30$, to obtain $N = 21.6$ individuals.

10. Use Equation 3.13 with $N = 20$ and $t = 8$ to obtain $F_t = 0.18$. (Since self-fertilization cannot occur, use of $N = 20.5$ would be a bit more accurate, but this gives essentially the same answer.) This level of allele identity by descent is just a little smaller than that expected from one generation of brother-sister mating (see Chapter 6).

11. Use Equation 3.14 with $H_t = H_0/x$, which yields $(1/x) = \exp[-t/(2N)]$, or $\ln(1/x) = -t/(2N)$. This result implies that $t = 2N\ln(x)$. For $x = 2$, the value $t = 1.39N$, agreeing with the value stated in the text.

12. After 69 generations of random genetic drift in a diploid subpopulation of effective size 50, the expected value of F_t is given by $1 - [1 - (1/100)]^{69} = 0.50$. The expected genotype frequencies, averaged across populations, are given by Equation 3.15 as 0.22, 0.22, and 0.56 for *AA*, *Aa*, and *aa*, respectively. For a particular subpopulation with an allele frequency of *A* equal to 0.3, the genotype frequencies are given by the Hardy-Weinberg

principle as 0.09, 0.42, and 0.49 for *AA*, *Aa*, and *aa*, respectively, because of random mating within each subpopulation.

13. In the F_1 population, all single-nucleotide polymorphisms are heterozygous so that the initial allele frequencies are $\frac{1}{2}$. Equation 3.14 can be applied to determine the expected fraction of sites that remain heterozygous. With $N = 80$ and $H_0 = 1.00$, for $t = 10$, $H_t = \exp(-\frac{10}{160}) = 0.94 \times 100 = 94$ sites segregating. For $t = 50$, the same approach gives 73 sites segregating.

14. Substitute $H_t = H_0/e$ into Equation 3.14 to obtain $1/e = \exp[-t/(2N)]$. Because $\ln(1/e) = -1$, then $-t/(2N) = -1$, and so $t = 2N$.

15. $N_f = 200$, and the number of breeding males equals $N_f/5 = 40$. In this situation $N_e = 4N_mN_f/(N_m + N_f) = 4(40)(200)/(40 + 200) = 133$.

16. Here $N_e = 4N_mN_f/(N_m + N_f)$. For 10 cows and one bull, $N_e = 4(1)(10)/(1 + 10) = 3.6$; for 40 cows and 1 bull, $N_e = 3.9$; for 10 cows and 2 bulls, $N_e = 6.7$.

17. The appropriate formula is Equation 3.25, which for $N_f = 100$ and $N_m = 10$ yields 37.5 and for $N_f = 10$ and $N_m = 100$ yields 21.4.

18. Using Equation 3.34 with $N = 50$ and $t = 99$ gives $(\frac{1}{100})(\frac{99}{100})^{99} = 0.0037$.

19. Equation 3.36 gives the average time to coalesce from k to $k - 1$ alleles as $4N/[k(k-1)]$, and in this case $N = 30$. For $k = 4$, the average is 10 generations, for $k = 3$ it is 20 generations, and for $k = 2$ it is 60 generations.

20. Use Equation 3.36 with $N = 50$ to obtain the average time to the first coalescence as $200/[k(k-1)] = 10$, or $k = 5$.

21. Use Equation 3.41. With $\theta = 10$, the expected number of sites segregating in samples of size 10, 20, and 50 are 29.3, 36.0, and 45.0, respectively.

CHAPTER 4

1. Various kinds of changes anywhere along the gene can result in loss of gene function, but once a gene has mutated, only very specific kinds of reverse mutations will restore function.

2. The Hill-Robertson effect is a mutual interference in the increase in the frequency of favorable alleles of different genes that exist simultaneously in the same population. It occurs because the favorable mutations almost always take place in different chromosomes, and as they both

increase in frequency, the original linkage disequilibrium between each favorable mutation and its surrounding chromosomal location is amplified. Increased recombination reduces the Hill-Robertson effect, as does an increase in effective population size.

3. Muller's ratchet is the sequential fixation of deleterious alleles that occurs by chance in small populations. Once any deleterious allele is fixed, it is retained in the population except in the unlikely event of a reverse mutation. Repeated population bottlenecks increase the chance fixation of deleterious mutations.

4. The experiment proves the point because the antibiotic-resistant cells in the colonies on the nonselective plate were never themselves exposed to the antibiotic.

5. $P_0 = 11/20 = 0.55$ and $\mu = -[\ln(P_0)]/N = 1.07 \times 10^{-9}$ per generation.

6. Use the Poisson zero term $P_0 = \Pr\{\text{exactly 0 lethals}\} = \exp(-m) = 1 - 0.35 = 0.65$, yielding $m = -\ln(P_0) = 0.43$ as the average number of lethals per chromosome.

7. Dominant lethals $(2 \times 10^{-2}$ to $10 \times 10^{-2})/(5 \times 10^{-4}) = 40$ to 200 rads; recessive visibles $(8 \times 10^{-6})/(7 \times 10^{-8}) = 114$ rads; reciprocal translocations $(2 \times 10^{-4}$ to $5 \times 10^{-4})/(1 \times 10^{-5}$ to $2 \times 10^{-5}) = 10$ to 50 rads. (Incidentally, humans appear to be somewhat less radiation sensitive than mice.)

8. Use $p_t = p_0(1 - \mu)^t$. For 10, 100, 1000, and 10,000 generations, p_t equals 0.99995, 0.9995, 0.995, and 0.95, respectively. Note that the approximation $p_t = 1 - \mu t$ is very accurate in this case.

9. Use $p_t = p_0(1 - \mu)^t$ with $p_0 = 1$, $\mu = 0.01$, and $p_t = 0.90$. Then $t = \ln(0.90)/\ln(0.99) = 10.5$ generations.

10. (a) Use $q_t = q_0 + \mu t$. For $t = 0$ to 12 generations, each four-generation interval increases q by 2×10^{-6}, so $\mu = (2 \times 10^{-6})/4 = 5 \times 10^{-7}$ per generation. (b) For $t = 12$ to 24 generations, each interval increases q by 0.04×10^{-6}, so $\mu = (0.04 \times 10^{-6})/4 = 10^{-8}$ per generation. The novel metabolite decreases the mutation rate; such substances are called *antimutagens*.

11. (a) At equilibrium, $\hat{p} = v/(\mu + v) = \frac{1}{11}$; (b) $\hat{p} = \frac{1}{101}$; (c) $\hat{p} = \frac{1}{2}$; (d) $\hat{p} = \frac{1}{11}$.

12. The equation for reversible mutation implies that $(p_t - \hat{p})/(p_0 - \hat{p}) = (1 - \mu - v)^t$, and halfway to equilibrium means that $p_t - \hat{p} = (\frac{1}{2})(p_0 - \hat{p})$, so that $(1 - \mu - v)^t$ equals $\frac{1}{2}$. Therefore, $t \ln(1 - \mu - v) = \ln(\frac{1}{2})$, or $t = \ln(\frac{1}{2})/\ln(1 - \mu -$

v). Using the approximation, $t \approx -\ln(2)/(\mu + v) \approx 0.69/(\mu + v)$. For the values of μ and v specified, $t \approx 6.3 \times 10^4$ generations.

13. $q_1 = q_0 + \mu_0$; $q_2 = q_1 + \mu_1 = q_0 + \mu_0 + \mu_1$; and so forth, and eventually

$$q_t = q_0 + \sum_{i=0}^{t} \mu_i$$

One can write this equation as $q_t = q_0 + \mu t$ if μ is interpreted as the arithmetic mean of the mutation rates over $t = 0$ to t generations, or

$$\mu = \frac{\sum_{i=0}^{t} \mu_i}{t+1}$$

14. $F = 1/(1 + \theta)$ and therefore $\theta = (1 - F)/F = 7$. With n equally frequent alleles, $p_i = 1/n$, and the homozygosity is $F = \Sigma p_i^2 = n/n^2 = 1/n$, which for $F = 0.125$ yields $n = 8$.

15. Tajima's D compares estimates of $\theta = 4N\mu$ based on the number of segregating sites S in a sample and the average number of pairwise mismatches Π. In the neutral infinite-alleles model at steady state, these estimates of θ should be equal. A negative value of Tajima's D indicates that the allele frequencies in the sample are too skewed, and a positive value of Tajima's D indicates that they are too nearly equal.

16. Let $a = 1 + (\frac{1}{2}) + (\frac{1}{3}) + \cdots + [1/(n-1)] = 3.02$ for $n = 12$. Then $E(S) = \theta a = 50$; hence $\theta = 50/3.02 = 16.56$. Since $E(\Pi) = \theta$, the expected average number of pairwise mismatches equals 16.56. The average number of pairwise mismatches per segregating site equals $16.56/50 = 0.33$. For 12 sequences there are 12 choose 2 = 66 pairwise comparisons, among which a singleton would cause a mismatch in 11. Hence the average number of pairwise mismatches per segregating site resulting from a singleton equals $\frac{11}{66} = 0.17$.

17. The Fu and Li test compares estimates of $\theta = 4N\mu$ based on the number of singleton nucleotides η_e versus the number of nonsingleton nucleotides η_i in the sample. The rationale is that with neutral coalescence, the expected total length of the external branches bears a definite relation to the expected length of the internal branches, and mutations on the external branches give rise to singletons and those on the internal branches to nonsingletons. In the neutral infinite-alleles model at steady state, the estimates of θ based on the singleton and nonsingleton polymorphisms should therefore be equal.

18. Let $a = 1 + (\frac{1}{2}) + (\frac{1}{3}) + \cdots + [1/(n-1)] = 3.25$ for $n = 15$. Then $E(S) = \theta a = 75$; hence $\theta = 75/3.25 = 23.07$. The expected number of singleton polymorphisms in the sample is given by $E(\eta_e) = \theta = 23.07$, and the expected number of nonsingleton polymorphisms is given by $E(\eta_i) = (a-1)\theta = 51.93$.

19. $E(\eta_e) = \theta$ and $E(\eta_i) = (a-1)\theta$; hence $E(\eta_e)/[E(\eta_i) + E(\eta_e)] = 1/a$. For $n = 2$, 5, 10, 20, 50, and 100, the corresponding values of $1/a$ are 1, 0.48, 0.35, 0.28, 0.22, and 0.19.

20. The expected times in generations are (a) $4N/6$, (b) $4N/6$, (c) $4N/2$, and (d) $4N\,[(\frac{1}{6}) + (\frac{1}{2})] = 4N(\frac{4}{6})$. The sum of these lengths equals $4N[(\frac{1}{6}) + (\frac{1}{6}) + (\frac{3}{6}) + (\frac{4}{6})] = 4N(\frac{9}{6}) = 4Na$. The external branches are (a), (b), and (d), and the sum of these branches equals $4N[(\frac{1}{6}) + (\frac{1}{6}) + (\frac{4}{6})] = 4N$. The internal branches consist only of (c), and its length is $4N(\frac{1}{2}) = 4N(a-1)$.

21. $\Pr\{x\} = 0.5255$, $\Pr\{y\} = 0.3285$, $\Pr\{z\} = 0.1460$. The proportion of sample configurations in which the numbers of the alleles are not equal to each other is 85.4%.

22. The possible sample configurations are (2, 0, 0, 1), (1, 1, 1, 0), and (0, 3, 0, 0), and their probabilities are 0.4000, 0.5333, and 0.0667, respectively. Note again that one is very unlikely to obtain a sample with an equal number of copies of each allele.

CHAPTER 5

1. The average fitness in the population tells us only about the average value of the relative fitnesses in the population, with some reference genotype arbitrarily assigned a relative fitness of 1. The relative fitnesses tell us nothing about the absolute fitnesses, and if the absolute fitnesses are less than 1, the population is headed for extinction no matter what may be happening to the average of the relative fitnesses.

2. In runaway sexual selection, the traits elaborated in the male and the basis of choosiness in the female become correlated through hormonal or neuronal pathways, genetic linkage, or other mechanisms, because successful mating pairs produce male progeny that have the genes for male ornamentation and female progeny that have the genes for choosiness based on that ornamentation. The process is runaway because the correlation implies that the male ornamentation and the female choosiness evolve coordinately and reinforce one another. This runaway process can be counteracted by reduced fitness among males with extreme ornamentation or by females that are too fastidious in their choice of mates.

3. In any population of adults, the only matings that produce progeny are those of $Aa \times Aa$, and with Mendelian segregation these matings produce zygotes that are $\frac{1}{4}$ AA, $\frac{1}{2}$ Aa, and $\frac{1}{4}$ aa. The AA genotypes die as embryos, and hence the survivors consist of $\frac{2}{3}$ Aa : $\frac{1}{3}$ aa. This is the equilibrium composition of adults in the population, no matter what system of mating takes place.

4. $p_1 = p_0s_0/(p_0s_0 + q_0)$ and $q_1 = q_0/(p_0s_0 + q_0)$, and hence $(p_1/q_1) = (p_0/q_0)s_0$; similarly, $(p_2/q_2) = (p_1/q_1)s_1 = (p_0/q_0)s_0s_1$. Continuing in this manner yields $p_n/q_n = (p_0/q_0)(s_0s_1 \ldots s_{n-1})$. Setting this equation equal to $p_n/q_n = (p_0/q_0)s^n$ reveals that $s = (s_0s_1 \ldots s_{n-1})^{1/n}$, and so s is the geometric mean of the selection coefficients.

5. Use the general equation for two allele viability selection $p' = (p^2w_{11} + pqw_{12})/\overline{w}$, to get $p_1 = 0.7364$, $p_2 = 0.7682$, and $p_3 = 0.7958$.

6. Defining the fitnesses as $1 - s : 1 : 1 - t$, the equilibrium allele frequency is $\hat{p} = t/(s + t)$, where (a) $s = 0.7$, $t = 0.3$; (b) $s = 0.07$, $t = 0.03$; (c) $s = 0.007$, $t = 0.003$. In all cases $p = 0.3$.

7. $\overline{w} = (0.64)(0.9) + 2(0.8)(0.2)(1) + (0.04)(0.6) = 0.92$. Note that $0.8 = 0.4/(0.4 + 0.1)$, so 0.8 is the equilibrium frequency for this case of overdominance, and in this model the equilibrium occurs at maximum $\overline{w}$.

8. $q' = pq(1 - h)/[p^2 + 2pq(1 - h)] = q(1 - h)/(1 + q - 2qh) \approx q(1 - h)$. Given that $q'/q = 0.99$, this implies that $h = 0.01$.

9. $t = 4[-2.1972 + 11.5129] = 37.26$ generations. Note that the equation can be written as $t = (2/s)[\ln(p_t/q_t) - \ln(p_0/q_0)]$. Under the assumptions given (i.e., significant resistance occurs even when the frequency of the resistance allele is relatively small), $t \approx (2/s)[\ln(p_t) - \ln(p_0)] = (2/s)\ln(p_t/p_0)$. For a large range of values of p_t/p_0 and s, t ranges from 5 to 50. This explains in part why significant pesticide resistance evolves so rapidly in diverse species.

10. For the haploid population, $p' = p/(1 - qs)$. For the diploid population, $p' = [p^2 + pq(1 - s)]/[p^2 + 2pq(1 - s) + q^2(1 - s)^2] = p[p + q(1 - s)]/[p + q(1 - s)]^2 = p/[p + q(1 - s)] = p/(1 - qs)$.

11. Interchange p and q and change the sign of s to obtain $\ln(q_t/p_t) + 1/p_t = [\ln(q_0/p_0) + 1/p_0] - st$, or $\ln(p_t/q_t) - 1/p_t = [\ln(p_0/q_0] - 1/p_0] + st$, where p_t is now the allele frequency of the favored recessive and s is the selection coefficient against individuals with the dominant allele.

12. The derivative $d\Delta p/dp = (\frac{1}{2} - p)(1 - p) - p(\frac{1}{2} - p) - p(1 - p)$. This value is positive for $p = 0$ and $p = 1$ (indicating unstable equilibria) and negative for $p = \frac{1}{2}$ (indicating local stability). The latter point is also globally stable.

13. For a recessive lethal, $w_{11} = w_{12} = 1$, $w_{22} = 0$. Then $q' = pq/(p^2 + pq) = q/(1 + q)$, or $1/q_n = 1 + 1/q_{n-1}$. This equation implies that $1/q_n = n + 1/q_0$, or $q_n = q_0/(1 + nq_0)$. For $q_n = q_0/2$, $n = 1/q_0$ generations.

14. Solve $p' = p(1 - \mu)/[p + q(1 - s)] = p$, which after some cancellation and simplification yields the equilibrium allele frequency $q = \mu/s$.

15. Use the equation $\hat{q} = \mu/hs$ and substitute $\mu = 9 \times 10^{-5}$ and $hs = \frac{1}{2}$, which yields $q = 0.00018$; the expected frequency of affected individuals is $2pq = 0.00036$ or one in 2778 individuals.

16. For the case of a complete recessive, substitute into the equation $\hat{q} = \sqrt{\mu/s}$ to obtain $\hat{q} = \sqrt{(4 \times 10^{-6}/0.2)} = 4.47 \times 10^{-3}$. For the partial dominance case, substitute into the equation $\hat{q} = \mu/hs$ to obtain $\hat{q} = (4 \times 10^{-6})/(0.05 \times 0.2) = 4.00 \times 10^{-4}$. The mere five percent dominance in heterozygous genotypes reduces the equilibrium frequency by more than an order of magnitude.

17. For the first part, use $\hat{q} = \sqrt{(\mu/s)}$ with $\mu = 10^{-6}$ and $s = 0.6$, yielding $\hat{q} = 1.29 \times 10^{-3}$; this would be reduced to $\hat{q} = 1.0 \times 10^{-3}$ if homozygous genotypes did not reproduce at all. That the reduction is very small results from the fact that most of the harmful alleles are present in heterozygous genotypes.

18. The overall frequency of heterozygous genotypes equals $2pq$, and the frequency of heterozygous genotypes resulting from new mutations equals $2p\mu$. (The factor of 2 comes from the fact that each diploid individual has two alleles of the gene, either of which can mutate.) The proportion of heterozygous genotypes resulting from new mutations is therefore $2p\mu/2pq = \mu/q = h$, since $q = \mu/h$.

19. The required conditions are $(1/0.9) + (v_{12}/1) > 2$ and $(1/0.8) + (v_{12}/0.5) > 2$. The first implies that $v_{12} > 0.889$ and the second implies that $v_{12} > 0.375$. Since both conditions must be satisfied, the requirement for a protected polymorphism in this case is $v_{12} > 0.889$.

20. The equation becomes greatly simplified in the case of a recessive lethal, for then we can write $w_{11} = 1$, $w_{12} = 1$, and $w_{22} = 0$ for the selection in the diploid phase and $v_1 = 1 - s$ and $v_2 = 1$ for selection in the haploid phase. Substituting into the equation for the equilibrium condition and solving for p yields the equilibrium allele frequencies $p = 1/(1 + s)$ of A and $q = s/(1 + s)$ of a.

21. Set $\hat{p} = 0.11$ and $k = 0.75$. Then w_{12} satisfies $[2 \times 0.75 \times w_{12} - 1]/(2w_{12} - 1) = 0.11$, or $w_{12} = 0.6953$. (Actually, segregation distorter is active only in males, and a model that takes this into account yields $w_{12} = 0.79$ in males and $w_{12} = 1$ in females).

22. The condition is not sufficient because the homozygous genotypes for some alleles may be superior to the heterozygous genotypes containing other alleles.

23. $p' = (p^2w_{11} + pqw_{12})/\bar{w} = 2c/4c = \frac{1}{2}$. This expression implies that, whatever the allele frequency of p in any generation, it immediately jumps to $\frac{1}{2}$ in the next generation, and so $p = \frac{1}{2}$ is a globally stable equilibrium. The model breaks down when $p = 0$ or $p = 1$ because the equations for w_{11} or w_{22} would require division by 0.

24. The A' and A alleles are overdominant and the equilibrium allele frequency of A' is given by $p = (w_{12} - w_{22})/(2w_{12} - w_{11} - w_{22})$ with $w_{11} = 0.5$, $w_{12} = 1$, and $w_{22} = 0.7$. Hence $p = 0.375$ at equilibrium and $q = 0.625$. The average fitness at equilibrium is given by $\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22} = 0.8125$. The marginal fitness of the new A^* allele is 0.8, because all genotypes bearing this allele have a relative fitness of 0.8. Because the marginal fitness of the A^* allele is less than the average fitness in the population, A^* cannot increase in frequency.

25. If the alleles are additive, then each homozygous genotype has a fitness equal to twice the contribution of the allele it contains, and each heterozygous genotype has the sum of the contributions of the alleles it contains. Hence A_1, A_2, A_3, and A_4 have contributions to fitness of 0.4, 0.3, 0.2, and 0,1, respectively, and the heterozygous genotypes have fitnesses 0.7 for A_1A_2, 0.6 for A_1A_3, 0.5 for A_1A_4, 0.5 for A_2A_3, 0.4 for A_2A_4, and 0.3 for A_3A_4. In a random mating population with equal allele frequencies, each homozygous genotype has a frequency of $\frac{1}{16}$ and each heterozygous genotype has a frequency of $\frac{2}{16}$, hence the average fitness in the population is $(\frac{1}{16})(0.8 + 0.6 + 0.4 + 0.2) + (\frac{2}{16})(0.7 + 0.6 + 0.5 + 0.5 + 0.4 + 0.3) = 0.50$.

CHAPTER 6

1. Two alleles can be identical by descent only if they derive from DNA replication of a single allele in a common ancestor. However, inbreeding means that the parents of an inbred individual share one or more common ancestors, and therefore the extent of shared ancestry is expressed quantitatively in terms of the probability of autozygosity. For a gene in a population that is fixed for some allele, the homozygosity for the fixed

allele equals 1. Inbreeding increases the probability of autozygosity of the gene without increasing its homozygosity.

2. The coefficient of consanguinity is equal to the inbreeding coefficient of a hypothetical offspring, because any offspring is formed from a union of two gametes, each of which carries an allele drawn at random from the alleles in one of the parents. The offspring has to be hypothetical because both individuals could be of the same sex, or one might be deceased.

3. Set $2pq(1-F) = (\frac{1}{2})(1-F) = \frac{1}{3}$, from which it follows that $F = \frac{1}{3}$. As a check, note that $(\frac{1}{4})(1-F) + (\frac{1}{2})F = \frac{1}{3}$ also.

4. $p^2(1-F) + pF = p^2 - p^2F + pF = p^2 + p(1-p)F = p^2 + pqF = p(1-q) + pqF = p - pq + pqF = p - pq(1-F)$.

5. With inbreeding coefficient F, the mean allele frequency among plants equals $[p^2(1-F) + pF](1) + [2pq(1-F)](\frac{1}{2}) = p$. The mean of the squares (MS) of the allele frequency among plants is $MS = [p^2(1-F) + pF](1^2) + [2pq(1-F)](\frac{1}{2})^2 = p - pq(1-F)/2$. The variance in allele frequency among plants $Var = MS - p^2 = (pq/2)(1+F)$. When $F = 0$ (random mating), $Var = pq/2$. When $F = 1$ (complete inbreeding), $Var = pq$.

6. (a) If a male is a common ancestor, then for the loop around that individual to have a probability of $(\frac{1}{2})(1 + F_A)$, one has to define $F_A = 1$ for males. The implication of this definition is that a male must transmit his X chromosome to all of his daughters, which is correct. (b) Because a male contributes his Y chromosome to his sons, paths with two or more consecutive males have probability 0 for the transmission of an X-linked gene.

7. $F_K = 4 \times (\frac{1}{2})^7 = \frac{1}{32}$. Note that A and B are first cousins, as are C and D. The individuals I and J are called double second cousins because they are second cousins along both sides of their ancestry.

8. In any generation, one needs to know only the inbreeding coefficient of the immediate common ancestor of any paths connecting the parents, because in the equation $F_I = \Sigma(\frac{1}{2})^n(1 + F_A)$, any inbreeding in previous generations is taken into account in F_A. In this case, $F_0 = 0$ is given, and $F_1 = 0$ also because the parents share no common ancestors. Continuing, $F_2 = (\frac{1}{2})^3(1 + F_0) = \frac{1}{8}$, $F_3 = (\frac{1}{2})^3(1 + F_1) = \frac{1}{8}$, and $F_4 = (\frac{1}{2})^3(1 + F_2) = \frac{9}{64}$.

9. $F_A = F_B = 0$; $F_C = F_D = 0$; $F_E = F_F = 2(\frac{1}{2})^3 = \frac{1}{4}$; $F_G = F_H = 2(\frac{1}{2})^3 + 4(\frac{1}{2})^5 = \frac{3}{8}$; $F_I = (\frac{1}{2})^3(1 + F_E) + (\frac{1}{2})^3(1 + F_F) + 4(\frac{1}{2})^5 + 8(\frac{1}{2})^7 = \frac{8}{16}$. Note that the pedigree is of three generations of sib mating.

10. $F_t = (\frac{1}{2})^2(1 + F_{t-2}) + (\frac{1}{2})^3(1 + F_{t-3}) + (\frac{1}{2})^4(1 + F_{t-4}) + \cdots = (\frac{1}{2})^2(1 + F_{t-2}) + (\frac{1}{2})F_{t-1} = \frac{1}{4} + (\frac{1}{2})F_{t-1} + (\frac{1}{4})F_{t-2}$. Therefore, $F_0 = 0, F_1 = 0, F_2 = \frac{1}{4}, F_3 = \frac{3}{8}, F_4 = \frac{8}{16}, F_5 = \frac{19}{32}$.

11. $F_0 = 0, F_1 = 0, F_2 = \frac{1}{4}, F_3 = \frac{3}{8}, F_4 = \frac{8}{16}, F_5 = \frac{19}{32} = 0.59$, so the genotype frequencies after five generations are 0.2147, 0.1706, 0.6147; one additional generation of random mating restores the Hardy-Weinberg frequencies of 0.09, 0.42, 0.49.

12. Random-mating frequencies are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$; among offspring of first cousins ($F = \frac{1}{16}$), they are $\frac{17}{64}, \frac{30}{64}, \frac{17}{64}$. The deficiency of heterozygotes equals $(\frac{32}{64} - \frac{30}{64})/(\frac{32}{64}) = \frac{1}{16} = F$.

13. $q = \sqrt{(1/1600)} = \frac{1}{40}$, and expected frequency from first-cousin matings is $(\frac{1}{1600})(\frac{15}{16}) + (\frac{1}{40})(\frac{1}{16}) = \frac{1}{465}$.

14. Proportions with first-cousin parents are 0.016, 0.022, 0.068, 0.120, and 0.391 for the frequencies given. When $q = 1$, the proportion is 0.01, which results from the fact that one percent of the matings are between first cousins.

15. The two-way hybrid has genotype A_1A_2 and gametes $\frac{1}{2}A_1 + \frac{1}{2}A_2$, and so the probability of identity by descent (F) among offspring is $(\frac{1}{2})^2 + (\frac{1}{2})^2 = \frac{1}{2}$; the three-way hybrid has gametes $\frac{1}{4}A_1, \frac{1}{4}A_2, \frac{1}{2}A_3$, and $F = (\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{2})^2 = \frac{3}{8}$; the four-way hybrid has gametes $\frac{1}{4}A_1, \frac{1}{4}A_2, \frac{1}{4}A_3, \frac{1}{4}A_4$, and $F = 4(\frac{1}{4})^2 = \frac{1}{4}$.

16. $F = \frac{1}{2}$ in odd-numbered generations (the progeny of selfing), and $F = 0$ in even-numbered generations (the progeny of random mating).

17. For repeated self-fertilization, the recurrence equation relating F_t to F_{t-1} is given by $F_t = (\frac{1}{2})(1 + F_{t-1})$. Since even a single generation of random mating erases all of the accumulated effects of inbreeding, in the situation described the relation between F_t and F_{t-1} is given by $F_t = (1 - S)(0) + S(\frac{1}{2})(1 + F_{t-1})$. At equilibrium, $F_t = F_{t-1} = F$, and solving the resulting equation for F yields $F = S/(2 - S)$.

18. Average frequency before fusion $= [(q + \varepsilon)^2 + (q - \varepsilon)^2]/2 = q^2 + \varepsilon^2$; average frequency after fusion $= q^2$; the difference equals ε^2, which is the variance in allele frequency among the subpopulations.

19. Multiply $(1 - F_{IS})(1 - F_{ST}) = 1 - F_{IT}$ and simplify. The expression says that the probability of autozygosity within the total population equals the probability of autozygosity because of inbreeding within a subpopula-

tion, plus the probability of autozygosity because of random genetic drift, minus the probability of autozygosity for both reasons.

20. Heterozygosities are 0.54, 0.62, 0.66, respectively, averaging 0.61. The fused population has allele frequencies 0.2, 0.3, 0.5, and heterozygosity 0.62. $F_{ST} = 0.02$. The maximum F_{ST} occurs when each population is fixed for a different allele, and $F_{ST} = 1$ for the set of three as well as for each pairwise comparison.

21. The allele frequency is 0.2 in both subpopulations, hence $H_S = 0.32$. Average $H_I = 0.272$. $F_{IS} = 0.15$. (Note from the genotype frequencies that the populations have inbreeding coefficients 0.1 and 0.2, respectively, which average 0.15). Since the allele frequencies are identical, $H_T = 0.32$ also, and so $F_{ST} = 0$. Since $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$, it follows that $F_{IT} = 0.15$.

22. (a) $F_{ST} = 1 - (1 - \frac{1}{100})^{50} = 0.3950$; $F_{IS} = \frac{1}{4} = 0.25$; $F_{IT} = 1 - (0.75)(0.6050) = 0.5462$. (The exponent in F_{ST} is 50, not 47, because random drift still occurs during the generations of sib mating.) (b) $F_{ST} = 0.3950$; $F_{IS} = 0$; $F_{IT} = 0.3950$.

23. Use Equation 6.19 with $p_t = 0.5$, $p_0 = 0.2$, $p^* = 0.8$, and $m = 0.01$. Then $t = 69$ generations.

24. Use Equation 6.21 with $(1 - m)^{10} = 0.6$ and $\hat{p} = (0.2 + 0.4 + 0.6 + 0.8)/4 = 0.5$, and so the allele frequencies are 0.3204, 0.4401, 0.5599, and 0.6796.

25. Equation 6.21 implies that $p_t - \bar{p} = (p_0 - \bar{p})(1 - m)^t$, and so $\Sigma(p_t - \bar{p})^2 = \Sigma(p_0 - \bar{p})^2(1 - m)^{2t}$, or $\sigma_t^2 = \sigma_0^2(1 - m)^{2t}$.

26. Use Equation 6.23 with $1/(4Nm + 1) < 0.05$, so $m > 4.75/N$.

27. (a) Subtract the equation for y_t from that for x_t to obtain $x_t - y_t = (x_{t-1} - y_{t-1})(1 - u - v)$, from which it follows immediately that $x_t - y_t = (x_0 - y_0)(1 - u - v)^t$. The right-hand side of this equation goes to 0 for all values of $u + v$ such that $0 < u + v < 2$. Hence $x_t - y_t$ goes to 0, and therefore x_t goes to y_t. (b) Multiply the equation for x_t by v and that for y_t by u and add, yielding the result that $vx_t + uy_t = vx_{t-1} + uy_{t-1}$. This equation implies that $vx_t + uy_t$ is a constant for all values of t, and therefore $vx_t + uy_t = vx_0 + uy_0$. At equilibrium, $x_t = y_t = \hat{x} = \hat{y}$, and therefore the equilibrium value of $\hat{x} = \hat{y}$ is given by $u\hat{x} + v\hat{x} = vx_0 + uy_0$, or $\hat{x} = (vx_0 + uy_0)/(u + v)$. (c) In part (a) we showed that $x_t - y_t = (x_0 - y_0)(1 - u - v)^t$. If $u + v < 1$, then $1 - u - v > 0$, and the sign of $(1 - u - v)^t$ is always positive. Hence the difference $x_t - y_t$ becomes smaller in each generation, without changing sign. If $u + v > 1$, then $1 - u - v < 0$, and the sign of $(1 - u - v)^t$ alternates in each generation.

Hence the difference $x_t - y_t$ becomes smaller in each generation, but oscillates in sign.

28. $H_S = \Sigma 2p_i(1 - p_i)/12 = 0.435817$, where p_i is the allele frequency in the ith subpopulation. The average allele frequency across all subpopulations is $p_T = \Sigma p_i/12 = 0.434167$, and $H_T = 2p_T(1 - p_T) = 0.491332$. The average allele frequency in subpopulations above the river is $p_A = 0.288333$ and below the river is $p_B = 0.580000$. Therefore $H_R = [2p_A(1 - p_A) + 2p_B(1 - p_B)]/2 = 0.448797$. For the division based on the river, $F_{SR} = (H_R - H_S)/H_R = 0.028923$, $F_{RT} = (H_T - H_R)/H_T = 0.0865702$, and $F_{ST} = (H_T - H_S)/H_T = 0.112989$. A substantial part of the total variance in allele frequency is therefore due to the presence of the river. For the division based on the knoll, the average allele frequency in subpopulations to the west of the knoll is $p_W = 0.393333$ and to the east of the knoll is $p_E = 0.475000$. The average heterozygosity with this division equals $H_K = [2p_W(1 - p_W) + 2p_E(1 - p_E)]/2 = 0.487997$. For the division based on the knoll, $F_{SK} = (H_K - H_S)/H_K = 0.106928$, $F_{KT} = (H_T - H_K)/H_T = 0.00678711$, and $F_{ST} = (H_T - H_S)/H_T = 0.112989$. Only a very small part of the variance in allele frequency among subpopulations can be accounted for by differentiation of subpopulations on opposite sides of the knoll.

CHAPTER 7

1. The selective constraints are probably weak, as 5×10^{-9} approximates the rate of synonymous substitution.

2. Nondegenerate, 2 (methionine and tryptophan); twofold-degenerate, 24; threefold-degenerate, 3 (isoleucine); and fourfold-degenerate, 32. The average degeneracy at the third position equals $[2(0) + 24(2) + 3(3) + 32(4)]/61 = 3.03$.

3. (a) $k = 1$ means that, in the sequences being compared the average number of nucleotide changes per site equals 1, which also means that some of the nucleotides have changed identity two or more times. (b) Proportion of sites with no substitutions 0.367879; proportion of sites with exactly one substitution, 0.367879; proportion of sites with exactly two substitutions, 0.18394; proportion of sites with more than two substitutions 0.0803014.

4. As d increases toward $\frac{3}{4}$, k increases without limit. A value of $d = \frac{3}{4}$ means that the nucleotide sequences are saturated with mutations; they are no more similar than would be expected by chance alone.

5. Most likely the protein is undergoing very rapid amino acid replacement because of natural selection.

6. Compensatory substitutions might be expected. For example, an A to G substitution at a site in a double-stranded region might be expected to be compensated by a T to C substitution at the site with which it pairs, since the compensatory substitution maintains the base pairing (the original A–T base pair now becomes a G–C base pair). This pattern of substitution is commonly observed.

7. A rate of 0.5% substitution per million years equals $0.005 \times 10^{-6} = 5 \times 10^{-9}$ substitutions per nucleotide site per year. The rate of divergence between two lineages is twice the rate of substitution in either one of them; hence in this case it is 1% per million years.

8. Apply the Jukes-Cantor correction (Equation 7.15) to obtain $k = 0.0517$ when $d = 0.05$, and $k = 2.0310$ when $d = 0.7$. Note that when there are few differences, k and d are nearly equal. When the proportion of sites that differ nears the limiting value of 0.75, there can be more than one expected substitution per site. Inspection of Equation 7.16 shows that the variance in the estimate of k increases as d increases.

9. With only two nucleotides allowed, Equation 7.8 becomes $P_N(t + 1) = (1 - \alpha)P_N(t) + \alpha[1 - P_N(t)]$, and hence $P_N(t + 1) - P_N(t) = \alpha - 2\alpha P_N(t)$. Interpreting this as the differential equation $dP_N(t)/dt = \alpha - 2\alpha P_N(t)$ implies that $P_N(t) = \frac{1}{2} + \frac{1}{2}\exp(-2\alpha t)$. Following the same reasoning described in the text finally results in $k = -(\frac{1}{2})\ln(1 - 2d)$, where d is the proportion of differences between the two sequences.

10. When there are six different nucleotides, each equally likely to mutate to any other, the analog of Equation 7.15 is $k = -(\frac{5}{6})\ln(1 - 6d/5)$. Substitute $d = 0.2$ to obtain $k = 0.2287$. Note that for four-base DNA, application of Equation 7.15 gives $k = 0.2326$, which means that the correction for possible multiple substitutions is greater for four-base than for six-base DNA.

11. If you assume that most differences are synonymous (third position), then the reading frame begins with position 1 at the far left. The sequences code for 20 amino acids among which two are different, and 60 nucleotides among which 13 are different. The sequences include 40 nondegenerate sites, 3 twofold-degenerate sites, 2 threefold-degenerate sites, and 15 fourfold-degenerate sites. The total number of nonsynonymous sites equals $40 + (\frac{2}{3})(3) + (\frac{1}{3})(2) = 42.67$, among which 2 are different. The total number of nonsynonymous sites equals $15 + (\frac{2}{3})(2) + (\frac{1}{3})(3) =$

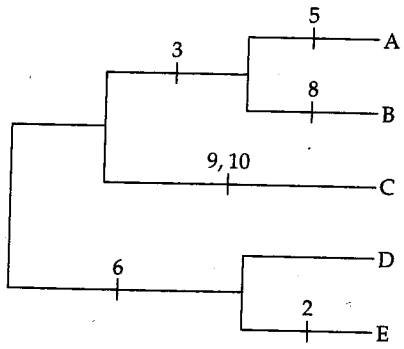
17.33 nonsynonymous sites, of which 11 are different. (a) Use Equation 7.5 with $D = \frac{2}{20}$, $K = 0.1054$. (b) Use Equation 7.15 with $d = \frac{13}{60}$, $k = 0.2557$. (c) Use Equation 7.15 with $d = \frac{2}{40}$, $k = 0.0484$. (d) Use Equation 7.15 with $d = \frac{11}{20}$, $k = 1.4039$. Note that average number of differences per nonsynonymous site is modest but that the average number of differences per synonymous sites exceeds 1. The latter observation suggests that these alleles have been present in the population for a very long time.

12. The rates of loss in the two species are 6.3×10^{-9} per nucleotide per year in *P. pedestris* and 0.49×10^{-9} per nucleotide per year in *D. melanogaster*. For a rate of DNA deletion of d per nucleotide per year, the half-life of a nonessential nucleotide equals t in the equation $(1 - d)^t = 0.5$ and, to an excellent approximation, $e^{-dt} = (1 - d)^t$ and so $t = -\ln(0.5)/d$. For $d = 6.3 \times 10^{-9}$, the value of $t = 110$ million years, and for $d = 0.49 \times 10^{-9}$, the value of $t = 14.3$ million years. This comparison implies that nonessential nucleotides persist almost 8 times longer in *P. pedestris* than in *D. melanogaster*.

13. With $d = \frac{1}{3}$ in Equation 7.15, $k = 0.4408$, and $t = 0.4408/(2 \times 0.01/\text{yr}) = 22.04$ yr, or about 22 years. The date of divergence is therefore approximately 1983 – 22 = 1961.

14. The values of k from Equation 7.15 are HIV1–HIV2 (0.4529), HIV1–VISNA (0.9547), HIV1–MMLV (1.3144), HIV2–VISNA (0.8865), HIV2–MMLV (1.3744), VISNA–MMLV (1.3744). These results indicate that HIV1 and HIV2 are most closely related to each other, more distantly to VISNA, and still more distantly to MMLV. HIV1, HIV2, and VISNA are all about equally distant from MMLV.

15. The number of each variable nucleotide site is shown above the branch in which a mutation most likely occurred to account for the pattern of nucleotide differences among the species.



16. (a) The average allele frequency of a neutral mutation fixed by random genetic drift must be $\frac{1}{2}$, and in a diploid population this frequency implies that the average number of copies of the mutation during the time to fixation must be $2N/2 = N$. Since the average time to fixation equals $4N$ generations, the average number of allele copies that have existed from the origin of the mutation until fixation equals $(4N)(N) = 4N^2$. (b) The expected number of mutations during the process of fixation by random genetic drift equals $(5 \times 10^{-8})(4N^2) = (5 \times 10^{-8})(4)(10^{12}) =$ 200,000 mutations *per nucleotide*!

17. Set $f = c_2$ and solve for λ, yielding $\lambda = (n - 1)/4N$. Then $c_1 = c_2 = 1/(1 + 4N\mu) \approx 1$ because $4N\mu << 1$.

18. When $\lambda = \mu$, then $c_1 = c_2 = 1/n$ and $f \approx 1$.

19. There are 3 types of knockouts, and hence 9 types of double knockouts; each of these can occur in 4 ways (both in the original copy, both in the duplicate copy, the first in the original copy and the second in the duplicate, or the other way around), hence there are $9 \times 4 = 36$ ways in which exactly two knockouts can occur. The probability that exactly two knockouts results in subfunctionalization equals $(2)(2f_Af_B)/[4(f_A + f_B + f_C)^2] = f_Af_B/(f_A + f_B + f_C)^2$.

CHAPTER 8

1. For $x = 78$, the standardized value equals $x' = (78 - 100)/15 = -1.4667$. For the entire set of numbers the standardized values are –1.4667, 1.4000, 2.7333, 1.0667, –0.7333, 1.2000. The two most extreme deviates are –1.4667 and 2.7333, corresponding to the original sample values 78 and 141. Because the normal distribution is symmetrical, half of the normal deviates are expected to be negative and half positive. The expected mean and variance among as sample of normal deviates are 0 and 1, respectively.

2. Let the genotype value of XY males be 1 and that of X′Y males be 0. The mean genotypic value of males is therefore $p(1) + q(0) = p$. The sons of XY males have genotypes XY and X′Y with probabilities p and q, respectively, and so their mean phenotypic value equals $pa + q(-a) = a(p - q)$. Likewise the mean phenotypic value of sons of X′Y fathers equals $a(p - q)$. The mean of the product of the phenotypic values of the sons and the genotypic values of the fathers equals $p(1)a(p - q) + q(0)a(p - q) = ap(p - q)$. The product of the means of the phenotypic values of the sons and the genotypic values of the fathers equals $a(p - q)(p) = ap(p - q)$. Hence the *Covariance(sons, fathers)* = 0, and therefore the regression coefficient equals 0. The result follows from X-linked inheritance, because sons

receive their X chromosome from their mother. With random mating, there is no correlation between the genotype of the father and that of the mother. With assortative mating, however, there would be such a correlation, and in that case the regression coefficient of sons on fathers would not be 0.

3. The midparent (*MP*) mean equals 20.1667 and the offspring (*O*) mean equals 19.8292. The mean of the product of the *MP* and *O* values equals 401.824. The *Covariance*(*O*, *MP*) = 1.93597, *Variance*(*MP*) = 4.22222, and so the regression coefficient of *O* on *MP* = $b_{O,MP}$ = 0.45852. The estimate of the narrow-sense heritability $h^2 = b_{O,MP} = 0.45852$. (We have carried along far more significant digits than justified, to allow you to check your own calculations.)

4. The father (*F*) mean equals 179.000 cm and the son (*S*) mean equals 176.917 cm. The mean of the product of the *F* and *S* values equals 31681.300. The *Covariance*(*F*, *S*) = 13.1667, *Variance*(*F*) = 41.1667, and so the regression coefficient of *S* on *F* = $b_{S,F}$ = 0.3198. The estimate of the narrow-sense heritability $h^2 = 2b_{S,F} = 0.6397$, or about 64%.

5. Let x = first litter, y = second litter. Then $\Sigma x = 104$, $\Sigma x^2 = 1106$, $\Sigma y = 103$, $\Sigma y^2 = 1101$, $\Sigma xy = 1089$, $r = [1089 - (104)(103)/10]/\{[1106 - (104)^2/10][1101 - (103)^2/10]\}^{1/2} = 0.5690$.

6. (a) $a = (23.8 - 19.4)/2 = 2.2$; $d = [25.2 - (23.8 + 19.4)/2] = 3.6$. (b) The mean in the population is $p^2(23.8) + 2pq(25.2) + q^2(19.4)$, where p is the allele frequency of A and q is that of A', with $p + q = 1$. The first derivative of this expression with respect to p equals $2p(23.8) + 2(q - p)(25.2) - 2q(19.4)$, which when set equal to 0 yields $p/q = 4.14286$ or $p = 0.805556$. Alternatively, treat the phenotypic values as if they some sort of measure of fitness, and so the relative fitnesses of AA, AA', and $A'A'$ would be $23.8/25.2 = 0.94444$, $25.2/25.2 = 1.00000$, and $19.4/25.2 = 0.76984$. Writing these in the form $1 - s$, 1, $1 - t$ yields $s = 0.05556$ and $t = 0.23016$, and the equilibrium is $p = t/(s + t) = 0.805556$, which maximizes the "average fitness" and therefore the average phenotypic value.

7. $\sigma_e^2 = 1.5$, $\sigma_g^2 = 6.0 - 1.5 = 4.5$, and the broad-sense heritability $H^2 = 4.5/6.0 = 75\%$.

8. The phenotypic standard deviation = 200 mg, and so the selection differential S = 400 mg. The narrow-sense heritability $h^2 = 10{,}000/40{,}000 = 25\%$, and so expected mean pupal weight among the progeny equals $2000 + (0.25)(400) = 2100$ mg.

9. After one generation, $M' = 20 + 0.30(4) = 21.2$ bristles; after 10 generations (using Equation 8.29), $M' = 20 + 0.25(4)(10) = 32$ bristles.

10. The narrow-sense heritability h^2 is the ratio of the total response to the cumulative selection differential, or $h^2 = 0.15/(0.07 \times 5) = 0.4286$.

11. Using Equation 8.50, $D = 3.73 + 2.01 = 5.74$ and $\sigma_g^2 = (0.87)^2 - (0.60)^2 = 0.3969$, hence $n = (5.74)^2/8(0.3969) = 10.38$. For reasons stated in the text, this is likely to be a substantial underestimate.

12. $R_t = h^2 S_t$ where R_t is the total response and S_t is the cumulative selection differential. If the selection is practiced for t generations, in each of which $S = 10$, then $S_t = 10t$. The desired $R_t = 220 - 180 = 40$ eggs per year, and $h^2 = 0.20$ is given. Therefore, solve $40 = (0.20)(10)t$ for t, which yields $t = 20$.

13. Let M_m be the mean phenotypic value of selected males and M_f that of females. Since both sexes must contribute equally to the progeny, the mean among all selected parents equals $(M_m + M_f)/2$. This average corresponds to M_S in Equation 8.8.

14. Let the phenotypes of AA, AA', $A'A'$ be 1, 0, 0 with frequencies p^2, $2pq$, q^2, so A' is a dominant allele with frequency q. Then $a = \frac{1}{2}$, $d = -\frac{1}{2}$, and $\sigma_a^2 = 2p^3q$. Also the total phenotypic variance $\sigma^2 = p^2 - (p^2)^2 = p^2q(1 + p)$, since in this model we are ignoring the environmental variance. Therefore, $h^2 = 2p^3q/p^2q(1 + p) = 2(1 - q)/(2 - q)$, which is approximately $1 - q$ when q is small.

15. Let fitnesses of AA, AA', $A'A'$ be $1 - s$, 1, $1 - t$, so $a = -(s - t)/2$ and $d = (s + t)/2$. At equilibrium $p = t/(s + t)$ and $q = s/(s + t)$, so $a + (q - p)d = 0$ and therefore $\sigma_a^2 = 0$.

16. Since $i = S/\sigma$ where i is the intensity of selection, S is the selection differential, and σ is the phenotypic standard deviation, Equation 8.10 can be written as $R = i\sigma h^2$. For percent protein, $R = (1.5)(0.45)(0.7) = 0.4725$, so the expected percent protein = 3.3% + 0.47% = 3.77%. For the correlated response use Equation 8.44a, $CR = (1.5)\sqrt{(0.60)}\sqrt{(0.70)}(0.55)(0.65) = 0.3475$, so the expected percent fat = 3.4% + 0.35% = 3.75%. The fat increase corresponds to a direct selection of $i = 0.35/(0.60)(0.65) = 0.897$.

17. For $B = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$ the approximation gives $i = 0.80, 1.25, 1.60, 1.91, 2.21$, respectively.

18. For the A and B loci, $2pq[a + (q - p)d]^2 = 0.18$ and 0.067, respectively, and so $\sigma_a^2 = 0.247$, and $h^2 = \sigma_a^2/\sigma^2 = 0.247$. For the A and B loci, $\sigma_d^2 = (2pqd)^2 =$

0.01 and 0, respectively, and so the broad-sense heritability $H^2 = (\sigma_a^2 + \sigma_d^2)/\sigma^2 = 0.257$.

19. The mean phenotype in the population equals $pa + q(-a)$, and so the phenotypic values of A and A', expressed as deviations from the population mean, are $2qa$ and $-2pa$, respectively. The variance in phenotypic values among the parents is therefore (a) $4pqa^2$. Since A parents yield exclusively A progeny, with a phenotypic value of a, and since A' parents yield exclusively A' progeny, with a phenotypic value of $-a$, the covariance between parents and progeny equals $p(a)(2qa) + q(-a)(-2pa) = 4pqa^2$. The regression coefficient of progeny phenotype on parental phenotype is therefore $4pqa^2/4pqa^2 = 1$. (b) In this case, the offspring mean from A parents equals $(1 - u)a + u(-a)$ and that from A' parents equals $(1 - v)(-a) + v(a)$. The offspring-parent covariance is therefore $p(2qa)[(1 - u)a + u(-a)] + q(-2pa)[(1 - v)(-a) + v(a)] = 4pqa^2(1 - u - v)$. The regression coefficient of progeny phenotype on parental phenotype is therefore $4pqa^2(1 - u - v)/4pqa^2 = 1 - u - v$.

20. The calculations are based on the following table.

Genotype	Frequency	Resistance	SNP genotype	Resistance × SNP genotype
KM	u	a	1	a
Km	v	a	0	0
kM	w	$-a$	1	$-a$
km	x	$-a$	0	0

The mean drug-resistance phenotypic value in the population equals $a(u + v) - a(w + x)$, and the mean SNP genotypic value equals $u + w$. The mean of the products equals $a(u - w)$, and therefore the covariance between drug-resistance phenotype and SNP genotype equals $a(u - w) - (u + w)a(u + v - w - x) = 2aD$ because $D = ux - vw$. The variance in the SNP genotypic values equals $(u + w) - (u + w)^2 = p - p^2 = pq$, and hence the regression coefficient of drug-resistance phenotype on SNP genotype equals $2aD/pq$.

CHAPTER 9

1. The C-value paradox refers to the observation that, among eukaryotes, the size of the genome has no discernable relation to the genetic, metabolic, behavioral, or developmental complexity of the organism. The ratio of largest to smallest is 3325 : 1. The arithmetic, geometric, and harmonic means are $(0.04 + 133)/2 = 66.52$ Gb, $\sqrt{(0.04 \times 133)} = 2.31$ Gb, and $1/[(\frac{1}{2})(\frac{1}{0.04} + \frac{1}{133})] = 0.080$ Gb. The human genome is closest in size to the geometric mean of the other two.

2. The equilibrium frequencies of GC and AT are in inverse proportion to their mutation rates; hence the GC : AT ratio at equilibrium equals $v/\mu = (2 \times 10^{-9})/(1 \times 10^{-9}) = 2$. The GC : AT ratio of 2 corresponds to a percent GC of 66.7%.

3. Isochores are long (> 300 kb) stretches in which the G+C content deviates significantly from that of the genome as a whole. They are prominent features of the genomes of birds and mammals, and were unexpected because in most other organisms the distribution of G+C content across the genome is nearly random.

4. In both cases the expected GC content is a proportion of GC of $p = 0.384$, and the standard deviation across 300 kb equals $\sqrt{[(0.384)(1 - 0.384)/300{,}000]} = 0.00088796$. A 300-kb isochore with a GC proportion of 0.35 deviates by $(0.35 - 0.384)/0.00088796 = -38.3$ standard deviations, and one with a GC proportion of 0.60 deviates by $(0.60 - 0.384)/0.00088796 = +243.3$ standard deviations.

5. Background selection refers to the loss of chromosomes from a population due to the presence of deleterious mutations. In regions of tight linkage, each deleterious mutation that goes to extinction carries along with it whatever neutral alleles may happen to be present in the same region, thereby reducing the level of genetic diversity from that expected with neutrality.

6. The relations among the parameters are easiest to interpret by writing Equation 9.1 as $-\ln(\pi/\pi_0) = U/(2hs + R)$. A value of $-\ln(\pi/\pi_0)$ close to 0 implies that $\pi \approx \pi_0$, which means that the observed level of nucleotide diversity is approximately that expected with neutrality. U is the deleterious mutation rate across the region, and a decrease in U increases π because fewer chromosomes are eliminated by selection. The parameter hs is the reduction in fitness of genotypes that are heterozygous for a deleterious allele; an increase in hs increases π because chromosomes with deleterious mutations are eliminated more rapidly and hence have a smaller effect on distorting the level of genetic diversity. The parameter R is the frequency of recombination across the region, and an increase in R increases π because increasing the rate of recombination decreases the chance that a neutral mutation will be trapped on a chromosome that also contains a deleterious mutation.

7. Haldane's rule is an empirical observation about animal species, which states that when hybridization between species has differential effects on fertility or viability between the sexes, the sex that is most severely affect-

ed is usually the heterogametic sex. Recessive mutations would contribute to Haldane's rule because any such mutations in the X chromosome would be expressed in the heterogametic sex.

8. The critical values of r^2 for $n = 20$, 40, and 80 are 0.192, 0.096, and 0.048, respectively. For an effective size of $N = 10^4$, these correspond to frequencies of recombination of $c = 0.0105\%$, 0.0235%, and 0.0496%, respectively. In the human genome, these values of c correspond roughly to 10.5 kb, 23.5 kb, and 49.6 kb, respectively.

9. $p' = pk/(pk + q/2)$, which implies that $\Delta p = pq(2k - 1)/[1 + p(2k - 1)]$. For $0 < p < 1$, $\Delta p > 0$ for all $k > \frac{1}{2}$, and hence the equilibrium value of p is $p = 1$. This result implies that the sex ratio at equilibrium is $1 - k$ females : k males.

10. In the meiotic drive case, $p' = p^2 + 2pqk$, and so $\Delta p = pq(2k - 1)$. In the selection case, $p' = [p^2(1 + 2s) + pq(1 + s)]/[p^2(1 + 2s) + 2pq(1 + s) + q^2]$, or $\Delta p = pqs/(1 + 2ps)$. Hence when $p \approx 0$, $s \approx 2k - 1$, which for $k = \frac{2}{3}, \frac{3}{4}$, and $\frac{4}{5}$ yields $s = \frac{1}{3}, \frac{1}{2}$, and $\frac{3}{5}$, respectively.

11. $p_0 = \frac{1}{3}$, $p_1 = (\frac{2}{3})(\frac{1}{2}) = \frac{1}{3}$, $p_2 = p_1 \times (\frac{1}{2}) = \frac{1}{6}$, $p_3 = p_2 \times (\frac{1}{2}) = \frac{1}{12}$, $p_4 = p_3 \times (\frac{1}{2}) = \frac{1}{24}$, and $p_5 = p_4 \times (\frac{1}{2}) = \frac{1}{48}$.

12. For IS3 the expected values are 23.7, 18.9, 11.4, 6.8, 4.1, and 6.1. The chi-square value is 13.4, which has an associated P-value of 0.004. For this IS element the model does not provide a satisfactory fit, probably because its mechanisms of regulation differ from that assumed in the model. For IS5, the expected numbers are 47.3, 7.9, 5.3, 3.5, 2.3, and 4.7. The chi-square value equals 4.22, which has an associated probability of 0.24. For this IS element the fit is quite satisfactory.

13. Solve $\alpha = (1 - \alpha)(1 - \phi)$ for ϕ, which yields $\phi = (1 - 2\alpha)/(1 - \alpha)$.

14. $w(49) = \text{Exp}[-0.0002 \times 49] = 0.990128$.

15. One standard deviation equals $\sqrt{(49)} = 7$. Five standard deviations above the mean imply a copy number of $49 + 35 = 84$, and five standard deviations below the mean imply a copy number of $49 - 35 = 14$. The average fitness of the plus-5 genotypes is 0.98334 and that of the minus-5 genotypes is 0.997204. As a percentage of the fitness of diploid genotypes with the average number of copies, the difference between these extremes of copy number is 1.4%.

CHAPTER 10

1. Linkage disequilibrium mapping makes use of information on large numbers of individuals, and the ancestral history of the chromosomes in these individuals includes meiotic divisions and the chance of recombination in every generation back to their coalescence. Conventional pedigree mapping relies on relatively few individuals over a relatively small number of generations.

2. It would likely make linkage disequilibrium mapping more efficient because of increased linkage disequilibrium.

3. The pattern of fixation and polymorphism could also be due to lineage sorting if the population of the common ancestors of chimpanzees and humans had the SNP as a polymorphism but at a later time one of the SNP alleles was lost in the chimpanzee lineage.

4. The chance of discovering a SNP with two alleles is one minus the probability that all of the samples consists entirely of one allele or the other. For a sample of size 2, for example, note that $2pq = 1 - p^2 - q^2$. Hence for a sample of size 3 the probability of SNP discovery is $1 - p^3 - q^3$ and for a sample of size 4 it is $1 - p^4 - q^4$.

5. When the null hypothesis is true, then 1.0 of the tests have a P-value of 1.0 or less, 0.9 have a P-value of 0.9 or less, and so forth down to the conventional statement that 0.05 of the tests have a P value of 0.05 or less. Hence the probability of a P-value equal to any value x equals x, which is the definition of a uniform distribution.

6. The deepest node in the tree had a branch that terminated in mtDNA sequences present exclusively in Africans.

7. Linkage disequilibrium has the shortest range in African populations. This is usually ascribed to African populations being older and having a larger effective population size than populations sampled from different continents.

8. For the Y-linked genes the Y : A ratio is expected to be $\mu_m/[(\frac{1}{2})\mu_m + (\frac{1}{2})\mu_f] = 2\mu_m/(\mu_m + \mu_f) = 2\alpha/(1 + \alpha) = 1.667$. For the X-linked genes the X : A ratio is expected to be $[(\frac{1}{3})\mu_m + (\frac{2}{3})\mu_f]/[(\frac{1}{2})\mu_m + (\frac{1}{2})\mu_f] = (2\mu_m + 4\mu_f)/(3\mu_m + 3\mu_f) = (4 + 2\alpha)/(3 + 3\alpha) = 0.778$.

9. $E(r^2) = 1/(1 + 4N_e c) = 0.4$; hence $4N_e c = 1.5$.

10. The individual's genotype could be $A\ B/a\ b$ or $A\ b/a\ B$, but it is not known which it actually is.

11. The expected frequencies of $A\ B/a\ b$ and $A\ b/a\ B$ are $P_{AB}P_{ab}$ and $P_{aB}P_{Ab}$, respectively. These will be equal when $D = P_{AB}P_{ab} - P_{aB}P_{Ab} = 0$, that is, when there is linkage equilibrium. Although for nonzero values of D the value of r^2 can be anywhere in a range of values, when $D = 0$ then $r^2 = 0$ also.

12. The increase occurs because individuals with an increase in the coefficient of inbreeding have more regions of the genome that are identical by descent and therefore are homozygous.

13. In this study, $a = 59$ and $b = 39$, and the chi-square value is given by $(a - b)^2/(a + b) = 4.08$. The chi-square test has one degree of freedom, and for $\chi^2 = 4.08$, the P value is 0.043. Hence the result does confirm the association.

14. $(\frac{1}{2})^5(1 - c)^6 = (0.99)^6/32 = 0.0294$.

15. Let the gametic types and their frequencies be $A\ B\ C$ $(\frac{1}{4})$, $A\ b\ C$ $(\frac{1}{4})$, $a\ B\ C$ (0), $a\ b\ C$ (0) and $A\ B\ c$ (0), $A\ b\ c$ (0), $a\ B\ c$ $(\frac{1}{4})$, $a\ b\ c$ $(\frac{1}{4})$. Then the mariginal gametic frequencies for A and B are $P_{AB} = \frac{1}{4}$, $P_{Ab} = \frac{1}{4}$, $P_{aB} = \frac{1}{4}$, and $P_{ab} = \frac{1}{4}$ and for B and C they are $P_{BC} = \frac{1}{4}$, $P_{Bc} = \frac{1}{4}$, $P_{bC} = \frac{1}{4}$, and $P_{bc} = \frac{1}{4}$. In these cases $D_{AB} = 0$ and $D_{BC} = 0$. The mariginal gametic frequencies for A and C are $P_{AC} = \frac{1}{2}$, $P_{Ac} = 0$, $P_{aC} = 0$, and $P_{ac} = \frac{1}{2}$. In this case $D_{AC} = (\frac{1}{2})(\frac{1}{2}) - (0)(0) = \frac{1}{4}$, and because the allele frequencies are all $\frac{1}{2}$, $D_{max} = \frac{1}{4}$.

BIBLIOGRAPHY

Aguadé, M., N. Miyashita and C. H. Langley. 1989. Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. Genetics 122:607–615.

Aguadé, M., N. Miyashita and C. H. Langley. 1992. Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. Genetics 132:755–770.

Ajioka, J. W. and D. L. Hartl. 1989. Population dynamics of transposable elements. pp. 939–958. *In* D. E. Berg and M. M. Howe (eds.), *Mobile DNA*. American Society for Microbiology, Washington, DC.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. Genetics 139:1067–1076.

Akashi, H. 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene 205:269–278.

Akashi, H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238:39–51.

Akashi, H., W. Y. Ko, S. Piao, A. John, P. Goel, C. F. Lin and A. P. Vitins. 2006. Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics 172:1711–1126.

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12:1805–1814.

Altenberg, L. and M. W. Feldman. 1987. Selection, generalized transmission and the evolution of modifier genes. I. The reduction principle. Genetics 117:559–572.

Alvarez, L. and K. Jaffe. 2005. Narcissism guides mate selection: Humans mate assortatively, as revealed by facial resemblance, following an algorithm of "self seeking like." Evol. Psychol. 2:177–194.

Aminetzach, Y. T., J. M. Macpherson and D. A. Petrov. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. Science 309:764–767.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. 11:635–641.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149–1152.

Andolfatto, P. and M. Przeworski. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics 158:657–665.

Andolfatto, P., J. D. Wall and M. Kreitman. 1999. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. Genetics 153:1297–1311.

Anholt, R. R. H., R. F. Lyman and T. F. C. Mackay. 1996. Effects of single *P* element insertions on olfactory behavior in *Drosophila melanogaster*. Genetics 143:293–301.

Aoki, K. 1981. Algebra of inclusive fitness. Evolution 35:659–663.

Aquadro, C. F., D. J. Begun and E. C. Kindahl. 1994. Selection, recombination and DNA polymorphism in *Drosophila*. pp. 46–56. *In* B. Golding, (ed.), *Non-Neutral Evolution: Theories and Molecular Data*. Chapman and Hall, New York.

Aquadro, C. F., R. M. Jennings, Jr., M. M. Bland, C. C. Laurie and C. H. Langley. 1992. Patterns of naturally occurring restriction map variation, dopa decarboxylase activity variation and linkage disequilibrium in the Ddc gene region of *Drosophila melanogaster*. Genetics 132:443–452.

Ardlie, K., S. N. Liu-Cordero, M. A. Eberle, M. Daly, J. Barrett, E. Winchester, E. S. Lander and L. Kruglyak. 2001. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am. J. Hum. Genet. 69:582–589.

Arnold, S. J. and M. J. Wade. 1984. On the measurement of natural and sexual selection: Applications. Evolution 38:720–734.

Avise, J. C. 1994. *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.

Avise, J. C., C. Giblin-Davidson, J. Laerm, J. C. Patton and R. A. Lansman. 1979. Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. Proc. Natl. Acad. Sci. USA 76:6694–6698.

Ayala, F. J. and M. L. Tracy. 1974. Genetic differentiation within and between species of the *Drosophila willistoni* group. Proc. Natl. Acad. Sci. USA 71:999–1003.

Ayala, F. J., B. S. W. Chang and D. L. Hartl. 1993. Molecular evolution of the *Rh3* gene in *Drosophila*. Genetica 92:23–32.

Ayala, F. J., D. E. Krane and D. L. Hartl. 1994. Genetic variation in *IncI1-ColIb* plasmids. J. Mol. Evol. 39:129–133.

Bachtrog, D. 2004. Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. Nat. Genet. 36:518–522.

Baird, M., I. Ballazs, A. Giusti, L. Miyazaki, L. Nicholas, K. Wexler, E. Kanter, J. Glasssberg, F. Allen, P. Rubenstein and L. Sussman. 1986. Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its applicability to the determination of paternity. Am. J. Hum. Gen. 39:489–501.

Ballard, J. W. 2000. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. J. Mol. Evol. 51:48–63.

Ballard, J. W. O. and M. Kreitman. 1994. Unraveling selection in the mitochondrial genome of *Drosophila*. Genetics 138: 757–772.

Barrier, M., C. D. Bustamante, J. Y. Yu and M. D. Purugganan. 2003. Selection on rapidly evolving proteins in the Arabidopsis genome. Genetics 163:723–733.

Barton, N. and S. P. Otto. 2005. Evolution of recombination due to random drift. Genetics 169:2353–2370.

Barton, N. H. 1990. Pleiotropic models of quantitative variation. Genetics 124:773–782.

Barton, N. H. and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nat. Rev. Genet. 3:11–21.

Bateson, W. 1909. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, UK.

Bechsgaard, J. S., V. Castric, D. Charlesworth, X. Vekemans and M. H. Schierup. 2006. The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Mol. Biol. Evol. 23:1741–1750.

Beerli, P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22:341–345.

Beerli, P. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Beerli, P. and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. 98:4563–4568.

Begin, M. and D. A. Roff. 2003. The constancy of the G matrix through species divergence

and the effects of quantitative genetic constraints on phenotypic evolution: A case study in crickets. Evolution 57:1107–1120.

Begun, D. J. and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–520.

Begun, D. J. and C. F. Aquadro. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548–550.

Bender, J. 2004. DNA methylation and epigenetics. Annu. Rev. Plant Biol. 55:41–68.

Bensasson, D., D. A. Petrov, D-X. Zhang, D. L. Hartl and G. M. Hewitt. 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. 18:246–253.

Benveniste, R. E. 1985. The contributions of retroviruses to the study of mammalian evolution. pp. 359–417. *In* R. J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum Press, New York.

Berg, D. E. and M. M. Howe (eds.). 1989. *Mobile DNA*. American Society for Microbiology, Washington, DC.

Bergman, A., D. B. Goldstein, K. E. Holsinger and M. W. Feldman. 1995. Population structure, fitness surfaces, and linkage in the shifting balance process. Genet. Res. 66:85–92.

Bergstrom, C. T and J. Pritchard. 1998. Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. Genetics 149:2135–2146.

Bernardi, G. 2004. *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*. Elsevier, Amsterdam.

Bernardi, G. and G. Bernardi. 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24:1–11.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. Science 228:953–958.

Berry, A. and M. Kreitman. 1993. Molecular analysis of an allozyme cline: Alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. Genetics 134:869–893.

Berry, A. J., J. W. Ajioka and M. Kreitman. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics 129:1111–1117.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74:1111–1120.

Betancourt, A. J. and D. C. Presgraves. 2002. Linkage limits the power of natural selection in *Drosophila*. Proc. Natl. Acad. Sci. USA 99:13616–13620.

Betran, E., K. Thornton and M. Long. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12:1854–1859.

Bishop, J. A. and L. M. Cook. 1975. Moths, melanism and clean air. Sci. Am. 232:90–99.

Blumenstiel, J. P. and D. L. Hartl. 2005. Evidence for maternally transmitted siRNA in the repression of transposition in *Drosophila virilis*. Proc. Natl Acad. Sci. USA 102:15965–15970.

Blumenstiel, J. P., D. L. Hartl and E. R. Lozovsky. 2002. Patterns of insertion and deletion in contrasting chromatin domains. Mol. Biol. Evol. 19:2211–2225.

Bonnell, M. L. and R. K. Selander. 1974. Elephant seals: Genetic variation and near extinction. Science 184:908–909.

Bouchard, T. J. and M. McGue. 2003. Genetic and environmental influences on human psychological differences. J. Neurobiol. 54:4–45.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde. E. Minch, J. R. Kidd and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796.

Bromham, L., A. Eyre-Walker, N. H. Smith and J. Maynard Smith. 2003. Mitochondrial Steve: paternal inheritance of mitochondria in humans. Trends Ecol. Evol. 18:2–4.

Brookfield, J. F. Y. and R. M. Badge. 1997. Population genetics models of transposable elements. Genetica 100:281–294.

Brown, W. M. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 77:3605–3609.

Brown, W. M., M. George and A. C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76:1967–1971.

Bruen, T. C., H. Philippe and D. Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665–2681.

Bulmer, M. 1994. *Theoretical Evolutionary Ecology*. Sinauer Associates, Sunderland, MA.

Bulmer, M. G. 1970. *The Biology of Twinning in Man*. Oxford University Press, London.

Bumpus, H. C. 1899. The elimination of the unfit as illustrated by the introduced sparrow *Passer domesticus*. Woods Hole Mar. Biol. Sta. Biol. Lectures 6:209–226.

Buonagurio, D. A., S. Nakada, J. D. Parvin, M. Krystal, P. Palese and W. M. Fitch. 1986. Evolution of human influenza A viruses over 50 years: Rapid uniform rate of change in NS genes. Science 232:980–982.

Bürger, R. 2000 *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester, UK.

Buri, P. 1956. Gene frequency in small populations of mutant *Drosophila*. Evolution 10:367–402.

Burleigh, J. G., A. C. Driskell and M. J. Sanderson. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst. Biol. 55:426–440.

Burzynski, A., M. Zbawicka, D. Skibinski and R. Wenne. 2006. Doubly uniparental inheritance is associated with high polymorphism for rearranged and recombinant control region haplotypes in Baltic *Mytilus trossulus*. Genetics 174:1081–1094.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill and A. G. Clark. 2006. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Bustamante, C., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan and D. L. Hartl. 2002. The cost of inbreeding in *Arabidopsis*. Nature 416:531–534.

Caballero, A. and P. D. Keightley. 1994. A pleiotropic nonadditive model of variation in quantitative traits. Genetics 138:883–900.

Callinan, P. A., J. X. Wang, S. W. Herke, R. K. Garber, P. Liang and M. A. Batzer. 2005. *Alu* retrotransposition-mediated deletion. J. Mol. Biol. 348:791–800.

Cann, R. L., M. Stoneking and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cano, J. M., A. Laurila, J. Palo and J. Merila. 2004. Population differentiation in G matrix structure due to natural selection in *Rana temporaria*. Evolution 58:2013–2020.

Carter, A. B., A. H. Salem, D. J. Hedges, C. N. Keegan, B. Kimball, J. A. Walker, W. S. Watkins, L. B. Jorde and M. A. Batzer. 2004. Genome-wide analysis of the human *Alu Yb*-lineage. Hum. Genomics 1:167–178.

Carvajal-Rodriguez, A., K. A. Crandall and D. Posada. 2006. Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. Mol. Biol. Evol. 23:817–827.

Carvalho, A. B. and A. G. Clark. 1999. Genetic recombination: Intron size and natural selection. Nature 401:344.

Carvalho, A. B. and A. G. Clark. 2005. Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. Science 307:108–110.

Casacuberta, E. and M.-L. Pardue. 2003. Transposon telomeres are widely distributed in the *Drosophila* genus: *TART* elements in the *virilis* group. Proc. Natl. Acad. Sci. USA 100:3363–3368.

Casey, B. 2001. Genetics of human situs abnormalities. Amer. J. Human Genet. 101:356–358.

Castle, W. E. 1921. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science 54:223.

Castro, J. P. and C. M. A. Carareto. 2004. *Drosophila melanogaster P* transposable elements: mechanisms of transposition and regulation. Genetics 121:107–118.

Cavalli-Sforza, L. L. and F. Cavalli-Sforza. 1995. *The Great Human Diasporas: The History of Diversity and Evolution*. Addison-Wesley, New York.

Cavalli-Sforza, L. L. and M. W. Feldman. 1978. Darwinian kin selection and "altruism." Theor. Popul. Biol. 14:268–280

Cavalli-Sforza, L. L. and W. F. Bodmer. 1971. *The Genetics of Human Populations*. W.H. Freeman, San Francisco.

Chakraborty, R. and M. Nei. 1977. Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution 31:347–356.

Chapman, T. 2006. Evolutionary conflicts of interest between males and females. Curr. Biol. 16:R744–754.

Charlesworth, B. 1980. *Evolution in Age-Structured Populations*. Cambridge University Press, Cambridge.

Charlesworth, B. and D. Charlesworth. 1983. The population dynamics of transposable elements. Genet. Res. 42:1–27.

Charlesworth, B. and D. L. Hartl. 1978. Population dynamics of the segregation distorter polymorphism of *Drosophila melanogaster*. Genetics 89:171–192.

Charlesworth, B., C. Bartolome and V. Noel. 2005. The detection of shared and ancestral polymorphisms. Genet. Res. 86:149–157.

Charlesworth, B., D. Charlesworth and N. H. Barton. 2003. The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. System. 34:99–125.

Charlesworth, B., J. A. Coyne and N. Barton. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130:113–1461.

Charlesworth, B., M. T. Morgan and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Charlesworth, B., P. Sniegowski and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–220.

Charlesworth, D., B. Charlesworth and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141:1619–1632.

Charmantier, A. and D. Garant. 2005. Environmental quality and evolutionary potential: lessons from wild populations. Proc. Roy. Soc. B 272:1415–1425.

Chaw, S. M., C. C. Chang, H. L. Chen and W. H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 58:424–441.

Chen, Y. and W. Stephan. 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. Proc. Natl. Acad. Sci. USA. 100:11499–11504.

Cheverud, J. M. and E. J. Routman. 1995. Epistasis and its contribution to genetic variance-components. Genetics 139:1455–1461.

Christiansen, F. B. and O. Frydenberg. 1974. Geographical patterns of four polymorphisms in *Zoarces viviparus* as evidence of selection. Genetics 77:765–770.

Christiansen, F. B., S. P. Otto, A. Bergman and M. W. Feldman. 1998. Waiting with and without recombination: The time to production of a double mutant. Theoret. Pop. Biol. 53:199–215.

Churchill, G. A. and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Civetta, A. and R. S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J. Mol. Evol. 41:1085–1095.

Clark, A. G. 1984. Natural selection with nuclear and cytoplasmic transmission. I. A deterministic model. Genetics 107:679–701.

Clark, A. G. 1988. Deterministic theory of heteroplasmy. Evolution 42:621–626.

Clark, A. G. 1998. Mutation-selection balance with multiple alleles. Genetica 103:41–47.

Clark, A. G. and C. M. S. Lanigan. 1993. Prospects for estimating nucleotide divergence with RAPDs. Mol. Biol. Evol. 10:1096–1111.

Clark, A. G. and M. W. Feldman. 1986. A numerical simulation of the one-locus, multiple-allele fertility model. Genetics 113:161–176.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle and C. F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Amer. J. Hum. Genet. 63:595–612.

Clark, A. G., L. Wang and T. Hulleberg. 1995. *P*-element-induced variation in metabolic regulation in *Drosophila*. Genetics 139:337–348.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams and M. Cargil. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

Clayton, G. A. and A. Robertson. 1955. Mutation and quantitative variation. Am. Nat. 89:151–158.

Clayton, G. A. and A. Robertson. 1957. An experimental check on quantitative genetical theory. II. Long-term effects of selection. J. Genet. 55:152–170.

Clegg, M. T., J. F. Kidwell, M. G. Kidwell and N. J. Daniel. 1976. Dynamics of correlated genetic systems. I. Selection in the region of the *Glued* locus of *Drosophila melanogaster*. Genetics 83:793–810.

Clegg, M. T., R. W. Allard and A. L. Kahler. 1972. Is the gene the unit of selection? Evidence from two experimental plant populations. Proc. Natl. Acad. Sci. USA 69:2474–2478.

Cohen, J. E. 1995. Unexpected dominance of high frequencies in chaotic nonlinear population models. Nature 378:610–612.

Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer and E. C. K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196.

Comeron, J. M. and M. Kreitman. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics 161:389–410.

Conner, J. K. and D. L. Hartl. 2004. *A Primer of Ecological Genetics*. Sinauer Associates, Sunderland, MA.

Cook, L. M. 1965. Inheritance of shell size in the snail *Arianta arbustorum*. Evolution 19:86–94.

Corander J., P. Waldmann and M. J. Sillanpää. 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374.

Corander, J., P. Waldmann, P. Marttinen and M. J. Sillanpää. 2004. BAPS 2: Enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20:2363–2369.

Costantini, M., O. Clay, F. Auletta and G. Bernardi. 2006. An isochore map of human chromosomes. Genome Res. 16:536–541.

Cotterman, C. W. 1940. *A Calculus for Statistico-genetics*. Unpublished PhD thesis, Ohio State University, Columbus, OH.

Coyne, J. A. 1998. Not black and white. Nature 396:35–36.

Coyne, J. A. and H. A. Orr. 2004. *Speciation*. Sinauer Associates, Sunderland, MA.

Craig, N. L., R. Craigie, M. Gellart and A. M. Lambowitz (eds.). 2002. *Mobile DNA II*. American Society for Microbiology, Washington, DC.

Cross, S. R. H. and A. J. Birley. 1986. Restriction endonuclease map variation in the *Adh* region in populations of *Drosophila melanogaster*. Biochem. Genet. 24:415–433.

Crow, J. F. and K. Aoki. 1982. Group selection for a polygenic behavioral trait: A differential proliferation model. Proc. Natl. Acad. Sci. USA 79:2628–2631.

Crow, J. F. and M. J. Simmons. 1983. The mutation load in *Drosophila*. pp. 1–35. *In* M. Ashburner, H. L. Carson and J. N. Thompson (eds.), *The Genetics and Biology of Drosophila*, Vol. 3C. Academic Press, London.

Crow, J. F. and M. Kimura. 1970. *An Introduction to Population Genetic Theory*. Harper & Row, New York.

Curie-Cohen, M. 1982. Estimates of inbreeding in a natural population: A comparison of sampling properties. Genetics 100:339–358.

Curtsinger, J. W. 1984. Evolutionary landscapes for complex selection. Evolution 38:359–367.

Cutler, D. J. 2000. Understanding the overdispersed molecular clock. Genetics. 154:1403–1417.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander. 2001. High-resolution haplotype structure in the human genome. Nat. Genet. 29:229–232.

Darvasi, A. and S. Shifman. 2001. The beauty of admixture. Nat. Genet. 37:118–119.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*. Murray, London.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. Appleton, New York.

Dayhoff, M. O. 1972. *Atlas of Protein Sequence and Structure*, Volume 5. National Biomedical Research Foundation, Silver Spring, MD.

de Bakker, P. I., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly and D. Altshuler. 2005. Efficiency and power in genetic association studies. Nat. Genet. 37:1217–1223.

De Brito, R. A., L. S. Pletscher and J. M. Cheverud. 2005. The evolution of genetic architecture. I. Diversification of genetic backgrounds by genetic drift. Evolution 59:2333–2342.

de Jong, G. 2005. Evolution of phenotypic plasticity: patterns of plasticity and the emer-

gence of ecotypes. New Phytologist 166:101–117.
Deininger, P. L. and M. A. Batzer. 1999. Alu repeats and human disease. Mol. Genet. Metabol. 67:183–193.
Dekkers, J. C. M. and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nature Rev. Genet. 3:22–32.
Depaulis, F., S. Mousset and M. Veuille. 2001. Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol. Biol. Evol. 18:1136–1138.
DePristo, M. A., D. M. Weinreich and D. L. Hartl. 2005. Missense meandering through sequence space. Nature Rev. Genet. 6:678–687.
Dermitzakis, E. T., A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier and S. E. Antonarakis. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302:1033–1035.
Dermitzakis, E. T., A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B. J. Stevenson, V. Flegel, P. Bucher, C. V. Jongeneel and S. E. Antonarakis. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420:578–582.
Devlin, D., S. E. Fienberg, D. P. Resnick and K. Roeder (eds). 1997. *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve.* Springer-Verlag, New York.
Dickerson, R. E. 1971. The structure of cytochrome *c* and the rates of molecular evolution. J. Mol. Evol. 1:26–45.
Dixon, M. T. and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10:256–267.
Dobzhansky, Th. and B. Spassky. 1944. Genetics of natural populations. XI. Manifestation of genetic variants of *Drosophila pseudoobscura* in different environments. Genetics 20:270–290.
Dobzhansky, Th. and B. Spassky. 1963. Genetics of natural populations. XXXIV. Adaptive norm, genetic load and genetic elite in *Drosophila pseudoobscura.* Genetics 48:1467–1485.
Doerge, R. W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. Nature Rev. Genet. 3:43–52.
Doerge, R. W. and G. A. Churchill. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285–294.
DuBose, R. F., D. E. Dykhuizen and D. L. Hartl. 1988. Genetic exchange among natural isolates of bacteria: Recombination within the *phoA* gene of *Escherichia coli.* Proc. Natl. Acad. Sci. USA 85:7036–7040.
Dudley, J. W. and R. J. Lambert. 2004. 100 Generations of selection for oil and protein in corn. Plant Breed. Rev. 24 (Part 1):79–110.
DuMouchel, W. H. and W. W. Anderson. 1968. The analysis of selection in experimental populations. Genetics 58:435–449.

Edwards, A. W. 2002. The fundamental theorem of natural selection. Theor. Popul. Biol. 61:335–337.
Ehrlich, P. and M. Feldman. 2003. Genes and cultures—What creates our behavioral phenome? Curr. Anthropol. 44:87–107.
Eickbush, T. H. 2002. *R2* and related site-specific non-long terminal repeat retrotransposons. pp. 813–835. *In* N. L. Craig, R. Craigie, M. Gellart and A. M. Lambowitz (eds.), *Mobile DNA II.* American Society for Microbiology, Washington, DC.
Ellis, N. A., P. Tippett, A. Petty, M. Reid, P. A. Weller et al. 1994. *PBDX* is the *XG* blood group gene. Nature Genet. 8:285–290.
Emerson, B. C., E. Paradis and C. Thébaud. 2001. Revealing the demographic histories of species using DNA sequences. Trends Ecol. Evol. 16:707–716.
Emigh, T. 1980. A comparison of tests for Hardy–Weinberg equilibrium. Biometrics 36:627–642.
Endler, J. A. 1986. *Natural Selection in the Wild.* Princeton Univ. Press, Princeton, NJ.
Enfield, F. D. 1980. Long term effects of selection: The limits to response. pp. 69–86. *In* A. Robertson (ed.), *Selection Experiments in Laboratory and Domestic Animals.* Commonwealth Agricultural Bureau, Slough, England.
Epling, C. and T. Dobzhansky. 1942. Genetics of natural populations, VI: Microgeographic races in *Linanthus parryae.* Genetics 27:317–332.
Epling, C., H. Lewis and F. M. Ball. 1960. The breeding group and seed storage: a study in population dynamics. Evolution 14:238–255.

Estes, S., B. C. Ajie, M. Lynch and P. C. Phillips. 2005. Spontaneous mutational correlations for life-history, morphological and behavioral characters in *Caenorhabditis elegans*. Genetics 170:645–653.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theoret. Pop. Biol. 3:87–112.

Ewens, W. J. 1979. *Mathematical Population Genetics*. Springer-Verlag, New York.

Ewens, W. J. 1982. On the concept of the effective population size. Theor. Popul. Biol. 21:373–378.

Ewens, W. J. 1989. An interpretation and proof of the fundamental theorem of natural selection. Theor. Popul. Biol. 36:167–180.

Ewens, W. J. 2004. *Mathematical Population Genetics*. Springer: New York.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. Proc. R. Soc. Lond. B 252:237–243.

Eyre-Walker, A. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. Genetics 152:675–683.

Eyre-Walker, A. and L. D. Hurst. 2001. The evolution of isochores. Nature Rev. Genet. 2:549–555.

Falconer, D. 1955. Patterns of response in selection experiments with mice. Cold Spring Harbor Symp. Quant. Biol. 20:178–196.

Falconer, D. 1977. Some results of the Edinburgh selection experiments with mice. pp. 101–115. *In* E. Pollak, O. Kempthorne and T. B. Bailey, Jr. (eds.), *International Conference on Quantitative Genetics*. Iowa State University Press, Ames.

Falconer, D. S. 1985. A note on Fisher's "average effect" and "average excess." Genet. Res. 46:337–347.

Falconer, D. S. and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*, Fourth Edition. Longman, Essex, England.

Fay, J. C. and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Fay, J. C., G. J. Wyckoff and C.-I. Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415:1024–1026.

Fearnhead P., R. M. Harding, J. A. Schneider, S. Myers and P. Donnelly. 2004. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics 167:2067–2081.

Feil, E. J., E. C. Holmes, D. E. Bessen, M.-S. Chan, N. P. J. Day et al. 2001. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA 98:182–187.

Feldman, M. W. and R. C. Lewontin. 1975. The heritability hang-up. Science 190:1163–1168.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Ferreira, M. U. and D. L. Hartl. 2006. *Plasmodium falciparum*: Worldwide sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-2 (MSP-2). Exp. Parasitol. 115:32–40.

Filipski, J., J. P. Thiery and G. Bernardi. 1973. An analysis of the bovine genome by $Cs_2SO_4^-$ Ag^+ density gradient centrifugation. J. Mol. Biol. 80:177–197.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Royal Soc. Edinburgh 52:399–433.

Fisher, R. A. 1922. On the dominance ratio. Proc. Roy. Soc. Edin. 42:321–431.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*, Second Edition. Clarendon, Oxford.

Fitch, W. M. and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. Biochem. Genet. 1:65–71.

Fiumera, A. C., B. L. Dumont and A. G. Clark. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. Genetics 169:243–257.

Frank, S. A. 1991. Haldane's rule: a defense of the meiotic drive theory. Evolution 45:1714–1717.

Fry, J. D., K. A. DeRonde and T. F. C. Mackay. 1995. Polygenic mutation in *Drosophila melanogaster*: Genetic analysis of selection lines. Genetics 139:1293–1307.

Fu, Y. X. and W.-H. Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Fuerst, P. A., R. Chakraborty and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics 86:455–483.

Fullerton, S. M., A. B. Carvalho and A. G. Clark. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. 18:1139–1142.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Galton, F. 1889. *Natural Inheritance*. MacMillan, London.

Game, E. T. and M. J. Caley. 2006. The stability of P in coral reef fishes. Evolution 60:814–823.

Gardner, M. P., K. Fowler, N. H. Barton and L. Partridge. 2005. Genetic variation for total fitness in *Drosophila melanogaster*: Complex yet replicable patterns. Genetics 169:1553–1571.

Garrigan D. and M. F. Hammer. 2006. Reconstructing human origins in the genomic era. Nat. Rev. Genet. 7:669–680.

Gaut, B. S. and A. D. Long. 2003. The lowdown on linkage disequilibrium. Plant Cell 15:1502–1505.

Gaut, B. S. and M. T. Clegg. 1993. Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA 90:5095–5099.

Gerrish, P. J. and R. E. Lenski. 1998. The fate of competing beneficial mutations in an asexual population. Genetica 103:127–144.

Gibbs, K. L. and P. R. Grant. 1987. Oscillating selection on Darwin's finches. Nature 327:511–513.

Gilad, Y., O. Man and G. Glusman G. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res. 15:224–230.

Gilad, Y., V. Wiebe, M. Przeworski, D. Lancet and S. Pääbo. 2004. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. PLoS Biol. 2(1): 0120–0125.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton, FL.

Gillespie, J. H. 1986. Variability of evolutionary rates of DNA. Genetics 113:1077–1091.

Gillespie, J. H. 1989. Lineage effects and the index of dispersion of molecular evolution. Mol. Biol. Evol. 6:636–648.

Gillespie, J. H. 1991. *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Gillespie, J. H. 1999. The role of population size in molecular evolution. Theor. Pop. Biol. 55:145–156.

Gillespie, J. H. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. Genetics 155:909–919.

Gillespie, J. H. 2004. *Population Genetics: A Concise Guide*. Johns Hopkins University Press: Baltimore, MD.

Gillespie, J. H. and C. H. Langley. 1974. A general model to account for enzyme variation in natural populations. Genetics 76:837–848.

Gimenez, O., R. Covas, C. R. Brown, M. D. Anderson, M. B. Brown and T. Lenormand. 2006. Nonparametric estimation of natural selection on a quantitative trait using mark-recapture data. Evolution 60:460–466.

Gojobori, T., E. N. Moriyama and M. Kimura. 1990. Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. USA. 87:10015–10018.

Golding, G. B. 1983. Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol. Biol. Evol. 1:125–142.

Golding, G. B., C. F. Aquadro and C. H. Langley. 1986. Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA 83:427–431.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Goldman, N. and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471.

Goodman, M. 1961. The role of immunologic differences in the phyletic development of human behavior. Hum. Biol. 33: 131–162.

Grant, B. S. 1999. Fine tuning the peppered moth paradigm. Evolution 53:980–984.

Grant, B. S., A. D. Cook, C. A. Clarke and D. F. Owen. 1998. Geographical and temporal variation in the incidence of melanism in peppered moth populations in America and Britain. J. Heredity 89:465–471.

Grant, S. F., G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K. P. Magnusson, G. B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A. Gylfason, J. Saemundsdottir, R. L. Wilensky, M. P. Reilly, D. J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J. R. Gulcher, A. Kong and K. Stefansson. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. 38:320–323.

Graur, D. and W.-H. Li. 2000. *Fundamantals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Green, R. E., J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic and S. Pääbo. 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444:333–336.

Grossman, A. I., L. G. Koreneva and L. E. Ulitskaya. 1970. Variation of the *alcohol dehydrogenase* (*ADH*) locus in natural populations of *Drosophila melanogaster*. Genetika 6:91–96 (in Russian).

Guo, S. W. and E. A. Thompson. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372.

Haddrill, P. R., K. R. Thornton, B. Charlesworth and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15:790–799.

Hahn, M. W. 2006. Detecting natural selection on cis-regulatory DNA. Genetica. http://www.springerlink.com/content/33408721100m3246/.

Haldane, J. B. S. 1922. Sex ratio and unisexual sterility in animal hybrids. J. Genet. 12:101–109.

Haldane, J. B. S. 1930. A mathematical theory of natural and artificial selection (Part VI. Isolation). Proc. Cambr. Philos. Soc. 26:220–230.

Haldane, J. B. S. 1956. The estimation of viabilities. J. Genet. 54:294–296.

Hamilton, W. D. 1964. The genetical evolution of social behaviour I. J. Theor. Biol. 7:1–16.

Hammer, M. F. and L. M. Silver. 1993. Phylogenetic analysis of the α-globin pseudogene-4 (*Hba-ps4*) locus in the house mouse species complex reveals a stepwise evolution of *t* haplotypes. Mol. Biol. Evol. 10:971–1001.

Han, K. D., J. C. Xing, H. Wang, D. J. Hedges, R. K. Garber, R. Cordaux and M. A. Batzer. 2005. Under the genomic radar: The Stealth model of *Alu* amplification. Genome Res. 15:655–664.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox, J. A. Schneider, D. S. Moulin and J. B. Clegg. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. 60:772–789.

Harris, H. 1966. Enzyme polymorphisms in man. Proc. R. Soc. Lond. B 164:298–310.

Harris, H., D. A. Hopkinson and Y. H. Edwards. 1977. Polymorphism and the subunit structure of enzymes. A contribution to the neutralist-selectionist controversy. Proc. Natl. Acad. Sci. USA 74:698–701.

Harris, T., E. F. Cook, R. Garrison et al. 1988. Body mass index and mortality among nonsmoking older persons: The Framingham Heart Study. JAMA 259:1520–1524.

Hart, D. L. and V. Orel. 1992. What did Gregor Mendel think he discovered? Genetics 131:245–253.

Hartl, D. L. 1970. A mathematical model for recessive lethal segregation distorters with differential viabilities in the sexes. Genetics 66:147–163.

Hartl, D. L. 2000a. *A Primer of Population Genetics*. Sinauer Associates, Sunderland, MA.

Hartl, D. L. 2000b. Molecular melodies in high and low C. Nature Rev. Genet. 1:145–149.

Hartl, D. L. 2001. Discovery of the transposable element *mariner*. Genetics 157:471–476.

Hartl, D. L. and R. Campbell. 1982. Allele multiplicity in simple Mendelian disorders. Amer. J. Human Genet. 34:866–873.

Hartl, D. L. and S. A. Sawyer. 1988. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? Genetics 118:537–541.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer. 1994. Selection intensity for codon bias. Genetics 138:227–234.

Hartl, D. L., E. R. Lozovskaya, D. I. Nurminsky and A. R. Lohe. 1997. What restricts the activity of *mariner*-like transposable elements? Trends Genet. 13:197–201.

Hasegawa, M., H. Kishino and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA J. Mol. Evol 22:160–174.

Haymer, D. S. and D. L. Hartl. 1982. The experimental assessment of fitness in *Drosophila*: Comparative measures of competitive reproductive success. Genetics 102:455–466.

Hedrick, P. W. 2005. A standardized genetic differentiation measure. Evolution 59:1633–1638.

Hedrick, P. W. and C. C. Cockerham. 1986. Partial inbreeding: Equilibrium heterozygosity and the heterozygosity paradox. Evolution 40:856–861.

Hegreness, M., N. Shoresh, D. L. Hartl and R. Kishony. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311:1615–1617.

Hey, J. and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760.

Hill, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239.

Hill, W. G. 2005. A century of corn selection. Science 307:683–684.

Hill, W. G. and A. Robertson. 1966. The effect of linkage on the limits to artificial selection. Genet. Res. 8:269–294.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hill, W. G. and B. S. Weir. 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. 54:705–714.

Hoekstra, H. E., K. E. Drumm and M. W. Nachman. 2004. Ecological genetics of adaptive color polymorphism in pocket mice: Geographic variation in selected and neutral genes. Evolution 58:1329–1341.

Holgate, P. 1966. A mathematical study of the founder principle of evolutionary genetics. J. Appl. Prob. 3:115–128.

Houle, D., B. Morikawa and M. Lynch. 1996. Comparing mutational variabilities. Genetics 143:1467–1483.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population without selection. Genet. Res. 50:245–250.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surveys Evol. Biol. 7:1–44.

Hudson, R. R. 1993. The how and why of generating gene genealogies. pp. 23–36. *In* N. Takahata and A. G. Clark (eds.), *Mechanisms of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Hudson, R. R. 2001. Two-locus sampling distributions and their application. Genetics 159:1805–1817.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson, R. R. and N. L. Kaplan. 1995. Deleterious background selection with recombination. Genetics 141:1605–1617.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala. 1994. Evidence for positive selection in the superoxide-dismutase (Sod) region of *Drosophila melanogaster*. Genetics 136:1329–1340.

Hudson, R. R., M. Kreitman and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Huelsenbeck, J. P. and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. J. Mol. Evol. 58:661–672.

Huelsenbeck, J. P., B. Larget, R. E. Miller and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Hughes, A. L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170.

Hurst, L. D. and A. Pomiankowski. 1991. Causes of sex ratio bias may account for unisexual sterility in hybrids: A new explanation of Haldane's rule and related phenomena. Genetics 128:841–858.

Hwang, D. G. and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA. 101:13994–14001.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13–34.

Iltis, H. 1932. *Life of Mendel* (E. and C. Paul, trans.). Norton, New York.

Innan, H. 2006. Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. Genetics 173:1725–1733.

Innan, H. and W. Stephan. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics 159:389–399.

Innan, H. and W. Stephan. 2003. Distinguishing the hitchhiking and background selection models. Genetics 165:2307–2312.

International Human Genome Sequencing and International HapMap Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

International Human Genome Sequencing and International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Ioerger, T. R., A. G. Clark and T.-H. Kao. 1991. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. USA 87:9732–9735.

Iwasa, Y. and A. Pomiankowski. 1995. Continual change in mate preferences. Nature 377:420–422.

Jaenike, J. 1996. Sex-ratio meiotic drive in the *Drosophila quinaria* group. Am. Nat. 148:237–254.

Jansen, R. C. and P. Stam. 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455.

Jeffreys, A. J., A. Ritchie and R. Neumann. 2000. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum. Mol. Genet. 9:725–733.

Jeffreys, A. J., L. Kauppi and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. 29:217–222.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers and P. Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. Nat Genet. 37:601–606.

Jenkins, N. L., G. McColl and G. J. Lithgow. 2004. Fitness cost of extended lifespan in *Caenorhabditis elegans*. Proc Biol. Sci. 271:2523–2526.

Jensen, A. R. 1972. *Genetics and Education.* Methuen, London.

Jensen, M. A., B. Charlesworth and M. Kreitman. 2002. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics 160:493–507.

Jensen, S., M. P. Gassama and T. Heidmann. 1999. Taming of transposable elements by homology-dependent gene silencing. Nature Genet. 21:209–212.

Jessen, T-H., R. E. Weber, G. Fermi, J. Tame and G. Braunitzer. 1991. Adaptation of bird hemoglobins to high altitudes: Demonstration of molecular mechanism by protein engineering. Proc. Natl. Acad. Sci. USA 88:6519–6522.

Jiang, M., J. Ryu, M. Kiraly, K. Duke, V. Reinke and S. K. Kim. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 98:218–223.

Johannsen, W. 1909. *Elemente der exackten Erblichkeitslehre*. Fischer, Jena.

Jones, A. G., S. J. Arnold and R. Borger. 2003. Stability of the G-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. Evolution 57:1747–1760.

Jones, D. T., W. R. Taylor and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comp. Appl. Biosci. 8:275–282.

Judson, O. P. and B. B. Normark. 1996. Ancient asexual scandals. Trends Ecol. Evol. 11:A41–A46.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. pp. 21–132. *In* H. N. Munro (ed.), *Mammalian Protein Metabolism III*. Academic Press, New York.

Jutier, D., N. Derome and C. Montchamp-Moreau. 2004. The sex-ratio trait and its evolution in *Drosophila simulans*: a comparative approach. Genetica 120:87–99.

Kapitonov, V. V. and J. R. Jurka. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biology 3:998–1011.

Kaplan, N. L. and J. F. Y. Brookfield. 1983. Transposable elements in Mendelian populations. III. Statistical results. Genetics 104:485–495.

Karlin S. and J. L. McGregor. 1972. Addendum to a paper of W. Ewens. Theoret. Pop. Biol. 3:113–116.

Karn, M. N. and L. S. Penrose. 1951. Birth weight and gestation time in relation to maternal age, parity and infant survival. Ann. Eugen. 16:147–164.

Kathuria, A., G. N. G. Gordon and A. K. Sheridan. 1996. Commentary on McAllister and Gillis (1996) concerning "Estimation of selection differentials from fish scales: A step towards evaluating genetic alteration of fish size in exploited populations." Can. J. Fish. Aquat. Sci. 53:940–941.

Keightley, P. D. and A. Eyre-Walker. 1999. Terumi Mukai and the riddle of deleterious mutation rates. Genetics 153:515–523.

Keightley, P. D. and S. P. Otto. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. Nature. 443:89–92.

Keightley, P. D. and W. G. Hill. 1988. Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. Genet. Res. 52:33–43.

Keightley, P. D. and W. G. Hill. 1990. Variation maintained in quantitative traits with mutation-selection balance: pleiotropic side-effects on fitness traits. Proc. R. Soc. Lond. Ser. B 242:95–100.

Keightley, P. D., M. J. Evans and W. G. Hill. 1993. Effects of multiple retrovirus insertions on quantitative traits of mice. Genetics 135:1099–1106.

Keith, T. P., L. D. Brooks, R. C. Lewontin, J. C. Martinez-Cruzado and D. L. Rigby. 1985. Nearly identical distributions of *xanthine dehydrogenase* in two populations of *Drosophila pseudoobscura*. Mol. Biol. Evol. 2:206–216.

Kelly, W. G., C. E. Schaner, A. F. Dernburg, M.-H. Lee, S. K. Kim, A. M. Villeneuve and V. Reinke. 2002. X-chromosome silencing in the germline of *C. elegans*. Development 129:479–492.

Kern, A. D., C. D. Jones and D. J. Begun. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 167:725–735.

Kettlewell, H. B. D. 1956. Further selection experiments on industrial melanism in the *Lepidoptera* Heredity 10:287–301.

Kettlewell, H. B. D. 1973. *The Evolution of Melanism: The Study of a Recurring Necessity*. Clarendon, Oxford.

Khil, P. P., B. Oliver and R. D. Camerini-Otero. 2005. X for intersection: retrotransposition both on and off the X chromosome is more frequent. Trends Genet. 21:3–7.

Kibota, T. T. and M. Lynch. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. Nature 381:694–696.

Kim, S. H., N. Elango, C. Warden, E. Vigoda and S. V. Yi . 2006. Heterogeneous genomic molecular clocks in primates. PLoS Genet. 2:1527–1534.

Kim, Y. and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524.

Kim, Y. and W. Stephan. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 155:1415–1427.

Kim, Y. and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777.

Kimura, M. 1955. Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41:144–150.

Kimura, M. 1957. Some problems of stochastic processes in genetics. Ann. Math. Stat. 28:882–901.

Kimura, M. 1964. Diffusion Models in Population Genetics. Methuen, London.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.

Kimura, M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11:247–269.

Kimura, M. 1976. Population genetics and molecular evolution. Johns Hopkins Med. J. 138:253–261.

Kimura, M. 1980a. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kimura, M. 1980b. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudosampling methods. Proc. Natl. Acad. Sci. USA 77:522–526.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M. 1985. The role of compensatory neutral mutations in molecular evolution. J. Genet. 64:7–19.

Kimura, M. 1986. DNA and the neutral theory. Philos. Trans. R. Soc. Lond. B Biol. Sci. 312:343–354.

Kimura, M. 1990. Some models of neutral evolution, compensatory evolution, and the shifting balance process. Theor. Popul. Biol. 37:150–158.

Kimura, M. and J. F. Crow. 1963. The measurement of effective population numbers. Evolution 17:279–288.

Kimura, M. and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.

Kimura, M. and T. Ohta. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61:763–771.

Kimura, M. and T. Ohta. 1971. *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton, NJ.

King, J. L. and T. H. Jukes. 1969. Non-Darwinian evolution: Random fixation of selectively neutral mutations. Science 164:788–798.

Kingman, J. F. C. 1980. *Mathematics of Genetic Diversity*. Society for Industrial and Applied Mathematics, Philadelphia.

Kingman, J. F. C. 1982a. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kingman, J. F. C. 1982b. The coalescent. Stochastic. Proc. Appl. 13:1461–1463.

Kingman, J. F. C. 2000. Origins of the coalescent: 1974–1982. Genetics 156:1461–1463.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Amer. Naturalist 157:245–261.

Kirkpatrick, M. and N. Barton. 1995. Déjà vu all over again. Nature 377:388–389.

Kivisild, T., P. Shen, D. P. Wall, B. Do, R. Sung, K. Davis, G. Passarino, P. A. Underhill, C. Scharfe, A. Torroni, R. Scozzari, D. Modiano, A. Coppa, P. de Knijff, M. Feldman, L. L. Cavalli-Sforza and P. J. Oefner. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373–387.

Kleene, K. C. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mechanisms Develop. 106:3–23.

Kleene, K. C. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Develop. Biol. 277:16–26.

Knott, S. A., R. M. Sibly, R. H. Smith and H. Møller. 1995. Maximum likelihood estimation of genetic parameters in life-history studies using the 'animal model'. Funct. Ecol. 9:122–126.

Ko, W. Y., S. Piao and H. Akashi. 2006. Strong regional heterogeneity in base composition evolution on the *Drosophila* X chromosome. Genetics 174:349–362.

Kolmogorov, A. 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Math. Ann. 104:415–458.

Kondrashov, A. S. 2001. Sex and *U*. Trends Genet. 17:75–77.

Kondrashov, A. S. and M. Turelli. 1992. Deleterious mutations, apparent stabilizing selection and the maintenance of quantitative variation. Genetics 132:603–618.

Kondrashov, A. S., S. Sunyaev and F. A. Kondrashov. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA 99:14878–14883.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulche and K. Stefansson. 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31:241–247.

Kong, X., K. Murphy, T. Raj, C. He, P. S. White and T. C. Matise. 2004. A combined linkage-physical map of the human genome. Am. J. Hum. Genet. 75:1143–1148.

Korpelainen, H. 2004. The evolutionary processes of mitochondrial and chloroplast

genomes differ from those of nuclear genomes. Naturwissenschaften. 91:505–518.
Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412–417.
Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking and S. Pääbo. 1997. Neanderthal DNA sequences and the origin of modern humans. Cell 90:19–30.
Krings, M., H. Geisert, R. W. Schmitz, H. Krainitzki and S. Pääbo. 1999. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc. Natl. Acad. Sci. USA. 96:5581–5585.
Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the 'animal model.' Phil Trans. Roy. Soc. B 359:873–890.
Kulathinal, R. J., B. R. Bettencourt and D. L. Hartl. 2004. Compensated deleterious mutations in insect genomes. Science 306:1553–1554.

Labbe, P., T. Lenormand and M. Raymond. 2005. On the worldwide spread of an insecticide resistance gene: a role for local selection. J. Evol. Biol. 18: 1471–1484.
Lahn, B. T. and D. C. Page. 1999. Four evolutionary strata on the human X chromosome. Science 286:964–967.
Lamason, R. L., M. A. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, M. C. Aros, M. J. Jurynec, X. Mao, V. R. Humphreville, J. E. Humbert, S. Sinha, J. L. Moore, P. Jagadeeswaran, W. Zhao, G. Ning, I. Makalowska, P. M. McKeigue, D. O'donnell, R. Kittles, E. J. Parra, J. N. Mangini, D. J. Grunwald, M. D. Shriver, V. A. Canfield, and K. C. Cheng. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782–1786.
Lande, R. 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet. Res. 26:221–234.
Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.
Lande, R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. Genetics 94:203–215.
Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99:541–553.
Lande, R. and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.
Lande, R. and S. J. Arnold. 1985. Evolution of mating preferences and sexual dimorphism. J. Theor. Biol. 117:651–664.
Lander, E. S. and D. Botstein. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570.
Lander, E. S. and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.
Landry, C. R., J. Oh, D. L. Hartl and D. Cavalieri. 2006. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. Gene 366:343–351.
Langley, C. H. and W. M. Fitch. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.
Langley, C. H., J. F. Y. Brookfield and N. Kaplan. 1983. Transposable elements in Mendelian populations. I. A theory. Genetics 104:457–471.
Laurent, F., H. Lelievre, M. Cornu, F. Vandenesch, G. Carret, J. Etienne and J. P. Flandrois. 2001. Fitness and competitive growth advantage of new gentamicin-susceptible MRSA clones spreading in French hospitals. J. Antimicrob. Chemother. 47:277–283.
Laurie, C. C., S. D. Chasalow, J. R. LeDeaux, R. McCarroll, D. Bush, B. Hauge, C. Lai, D. Clark, T. R. Rocheford and J. W. Dudley. 2004. The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics 168:2141–2155.
Lazzeroni, L. C. 1998. Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am. J. Hum. Genet. 62:159–170.
Le Rouzic, A. and P. Capy. 2005. The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. Genetics 169:1033–1043.
Leamy, L. J. and C. P. Klingenberg. 2005. The genetics and evolution of fluctuating asymmetry. Ann. Rev. Ecol. Syst. 36:1–21.

Leicht, B. G., S. V. Muse, M. Hanczyc and A. G. Clark. 1995. Constraints on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila*. Genetics 139:299–308.

Lemey P., I. Derdelinckx, A. Rambaut, K. Van Laethem, S. Dumont, S. Vermeulen, E. Van Wijngaerden and A. M. Vandamme. 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J. Virol. 79:11981–11989.

Lemos, B., C. D. Meiklejohn, M. Cáceres and D. L. Hartl. 2005. Rates of divergence in gene expression profiles of primates, mice and flies: Stabilizing selection and variability among functional categories. Evolution 59:126–137.

Lerner, I. M. 1958. *The Genetic Basis of Speciation*. John Wiley and Sons, New York.

Levin, D. A. 1978. Genetic variation in annual *Phlox*: Self-compatible versus self-incompatible species. Evolution 32:245–263.

Levings, C. S. III 1983. The plant mitochondrial genome and its mutants. Cell 32:659–661.

Lewontin, R. C. 1974a. The analysis of variance and the analysis of causes. Am. J. Hum. Gen. 26:400–411.

Lewontin, R. C. 1974b. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R. C. 1991. Electrophoresis in the development of evolutionary genetics: milestone or millstone? Genetics 128:657–662.

Lewontin, R. C. and J. Krakauer. 1973. Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics 74:175–195.

Lewontin, R. C. and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609.

Lewontin, R. C. and L. C. Dunn. 1960. The evolutionary dynamics of a polymorphism in the house mouse. Genetics 45:705–722.

Lewontin, R. C., L. R. Ginzburg and S. D. Tuljapurkar. 1978. Heterosis as an explanation for large amounts of genetic polymorphism. Genetics 88:149–170.

Li, J., S. Wang and Z.-B. Zeng. 2006. Multiple-interval mapping for ordinal traits. Genetics 173:1649–1663.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li, W.-H. and C.-I. Wu. 1987. Rates of nucleotide substitution are evidently higher in rodents than in man. Mol. Biol. Evol. 4:74–77.

Li, W.-H. and L. A. Sadler. 1991. Low nucleotide diversity in man. Genetics 129:513–523.

Li, W.-H., C.-C. Luo and C.-I Wu. 1985a. Evolution of DNA sequences. pp. 1–94. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum Press, New York.

Li, W.-H., C.-I Wu and C.-C. Luo. 1985b. A new method for estimating synonymous an nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2:150–174.

Li, W.-H., S. J. Yi and K. Makova. 2002. Male-driven evolution. Curr. Opinion Genet. Develop. 2:650–656.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Liti, G. and E. J. Louis. 2005. Yeast evolution and comparative genomics. Annu. Rev. Microbiol. 59:135–153.

Lofsvold, D. 1986. Quantitative genetics of morphological differentiation in *Peromyscus*. I. Tests of homogeneity of genetic variances and covariances. Evolution 40:559–573.

Lohe, A. R. and D. L. Hartl. 1996. Autoregulation of mariner transposase activity by overproduction inhibition and dominant-negative complementation. Mol. Biol. Evol. 13:549–555.

Lohe, A. R., E. N. Moriyama, D.-A. Lidholm and D. L. Hartl. 1995. Horizontal transmission, vertical degeneration, and stochastic loss of *mariner*-like transposable elements. Mol. Biol. Evol. 12:62–72.

Loisel, D. A., M. V. Rockman, G. A. Wray, J. Altmann and S. C. Alberts. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5′ cis-regulatory region. Proc. Natl. Acad. Sci. USA. 103:16331–16336.

Long, A. D. and C. H. Langley. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research 9:720–731.

Long, A. D., R. F. Lyman, C. H. Langley and T. F. C. Mackay. 1998. Two sites in the *Delta* gene region contribute to naturally occur-

ring variation in bristle number in *Drosophila melanogaster*. Genetics 149:999–1017.

Louis, E. J., E. S. Naumova, A. Lee, G. Naumov and J. E. Haber. 1994. The chromosome end in yeast: Its mosaic nature and influence on recombinational dynamics. Genetics 136:789–802.

Lu, J. and C.-I. Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl. Acad. Sci. USA 102:4063–4067.

Lucotte, G. and G. Mercier. 1998. Distribution of the *CCR5* gene 32-bp deletion in Europe. J. Acquired Immune Deficiency Syndromes and Human Retrovirology 19:174–177.

Lunzer, M., S. P. Miller, R. Felsheim and A. M. Dean. 2005. The biochemical architecture of an ancient adaptive landscape. Science 310:499–501.

Lyman, R. F., F. Lawrence, S. V. Nuzhdin and T. F. C. Mackay. 1996. Effects of single *P* element insertions on bristle number and viability in *Drosophila melanogaster*. Genetics 143:277–292.

Lynch, M. 1994. Neutral models of phenotypic evolution. pp. 86–108. *In* L. Real (ed.), *Ecological Genetics*. Princeton University Press, Princeton, NJ.

Lynch, M. and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Lynch, M. and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Lynch, M. and J. S. Conery. 2003. The origins of genome complexity. Science 21:1401– 1404.

Lynch, M. and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. Evolution 40:915–935.

Lyttle, T. W. 1991. Segregation distorters. Annu. Rev. Genet. 25:511–557.

Mackay, T. F. C. 2001. The genetic architecture of quantitative traits. Ann. Rev. Genet. 35:303–339.

Mackay, T. F. C. 2004. The genetic architecture of quantitative traits: lessons from *Drosophila*. Current Opinion Genet. Develop. 14:253–257.

Mackay, T. F. C., R. F. Lyman and M. S. Jackson. 1992. Effects of *P* element insertions on quantitative traits in *Drosophila melanogaster*. Genetics 180:315–332.

Majerus, M. E. N. 1998. *Melanism: Evolution in Action*. Oxford University Press: Oxford UK.

Makova, K. D. and W.-H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature 416:624–626.

Malécot, G. 1948. *Les mathématiques de l'hérédité*. Masson et Cie, Paris.

Manly, B. F. J. 1985. *The Statistics of Natural Selection on Animal Populations*. Chapman and Hall, London.

Maraganore, D. M., M. de Andrade, T. G. Lesnick, K. J. Strain, M. J. Farrer, W. A. Rocca, P. V. Pant, K. A. Frazer, D. R. Cox and D. G. Ballinger. 2005. High-resolution whole-genome association study of Parkinson disease. Am. J. Hum. Genet. 77:685–593.

Marais, G., T. Domazet-Loso, D. Tautz and B. Charlesworth. 2004. Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. J. Mol. Evol. 59:771–779.

Mark Welch, J. L., D. B. Mark Welch and M. Meselson. 2004. Cytogenetic evidence for asexual evolution of bdelloid rotifers. Proc. Natl. Acad. Sci. USA 101:1618–1621.

Marth G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. J. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok and W. R. Gish. 1999. A general approach to single-nucleotide polymorphism discovery. Nature Genet. 23:452–456.

Martin, E. R., E. H. Lai, J. R. Gilbert, A. R. Rogala, A. J. Afshari, J. Riley, K. L. Finch, J. F. Stevens, K. J. Livak, B. D. Slotterbeck, S. H. Slifer, L. L. Warren, P. M. Conneally, D. E. Schmechel, I. Purvis, M. A. Pericak-Vance, A. D. Roses and J. M. Vance. 2000. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am. J. Hum. Genet. 67:383–394.

Martin, G., S. P. Otto and T. Lenormand. 2006. Selection for recombination in structured populations. Genetics 172:593–609.

Maruyama, K. and D. L. Hartl. 1991. Evolution of the transposable element *mariner* in *Drosophila* species. Genetics 128:319–329.

Matassi G., B. Cherif-Zahar, G. Pesole, V. Raynal and J. P. Cartron. 1999. The members of the RH gene family (RH50 and RH30) followed different evolutionary pathways. J. Mol. Evol. 48:151–159.

May, R. M. (ed.). 1981. *Theoretical Ecology*. Blackwell, Oxford.

May, R. M. 1985. Evolution of pesticide resistance. Nature 315:12–13.

May, R. M. 1995. The cheetah controversy. Nature 374:309–310.

Maynard Smith, J. and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. Genet. Res. 23:23–35.

Maynard Smith, J. and K. C. Sondhi. 1961. The genetics of a pattern. Genetics 45:1039–1050.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly and D. M. Altshuler. International HapMap Consortium. 2006. Common deletion polymorphisms in the human genome. Nat. Genet. 38:86–92.

McDonald, J. H. and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654.

McGuigan, K. 2006. Studying phenotypic evolution using multivariate quantitative genetics. Mol. Ecol. 15:883–896.

McKeigue, P. M., J. R. Carpenter, E. J. Parra and M. D. Shriver. 2000. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann. Hum. Genet. 64:171–186.

McLellan, T. and L. S. Inouye. 1986. The sensitivity of isoelectric focusing and electrophoresis in the detection of sequence differences in proteins. Biochem. Genet. 24:571–577.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

McVean, G., P. Awadalla and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241.

Meiklejohn, C. D., J. Parsch, J. M. Ranz and D. L. Hartl. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. USA 100:9894–9899.

Meiklejohn, C. D., Y. Kim, D. L. Hartl and J. Parsch. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite likelihood estimation. Genetics 168:265–279.

Mendel, G. 1866. Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn 4:3–47.

Merbs, S. L. and J. Nathans. 1992. Absorption-spectra of the hybrid pigments responsible for anomalous color-vision. Science 258: 464–466.

Mercot, H., A. Atlan, M. Jacques and C. Montchamp-Moreau. 1995. Sex-ratio distortion in *Drosophila simulans*: Co-occurrence of a meiotic drive and a suppressor of drive. J. Evol. Biol. 8:283–300.

Meunier, J. and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. 21:984–990.

Meyer, K. and M. Kirkpatrick. 2005. Up hill, down dale: quantitative genetics of curvaceous traits. Phil. Trans. Roy. Soc. B 1459:1443–1455.

Milkman, R. and M. M. Bridges. 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics 126:505–517.

Milkman, R. and M. M. Bridges. 1993. Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. Genetics 133:455–468.

Miller, W. J., J. F. McDonald, D. Nouaud and D. Anxolabéhère. 1999. Molecular domestication—more than a sporadic episode in evolution. Genetica 107:197–207.

Mishmar, D., E. Ruiz-Pesini, P. Golik, V. Macaulay, A. G. Clark, S. Hosseini, M. Brandon, K. Easley, E. Chen, M. D. Brown, R. I. Sukernik, A. Olckers and D. C. Wallace. 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA. 100:171–176.

Mitchell-Olds, T. and R. G. Shaw. 1987. Regression analysis of natural selection: Statistical inference and biological interpretation. Evolution 41:1149–1161.

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu and T. Tasunaga. 1987. Male-driven molecular evolution: A model and nucleotide sequence analysis. Cold Spring Harbor Symp. Quant. Biol. 52:863–867.

Montgomery, E. A. and C. H. Langley. 1983. Transposable elements in Mendelian populations. II. Distribution of three *copia*-like elements in a natural population of *Drosophila melanogaster*. Genetics 104:473–483.

Montooth, K. L., K. T. Siebenthall and A. G. Clark. 2006. Membrane lipid physiology and toxin catabolism underlie ethanol and

acetic acid tolerance in *Drosophila melanogaster*. J. Exp. Biol. 209:3837–3850.
Moore, J. H. 2005. A global view of epistasis. Nature Genet. 37:13–14.
Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA 93:2873–2878.
Moran, N. A. and P. H. Degnan. 2006. Functional genomics of *Buchnera* and the ecology of aphid hosts. Mol. Ecol. 15:1251–1261.
Moran, P. A. P. 1958. Random processes in genetics. Proc. Camb. Phil. Soc. 54:60–71.
Morrell, P. L., K. E. Lundy and M. T. Clegg. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp *spontaneum*) despite migration. Proc. Natl. Acad. Sci. USA 100:10812–10817.
Mosher, M. J., L. J. Martin, L. A. Cupples, Q. Yang, T. D. Dyer, J. T. Williams and K. E. North. 2005. Genotype-by-sex interaction in the regulation of high-density lipoprotein: The Framingham heart study. Human Biol. 77:773–793.
Mourant, A. E., A. C. Kopec and K. Domaniewska-Sobczak. 1976. *The Distribution of Human Blood Groups and other Polymorphisms*, Second Edition. Oxford University Press, New York.
Mukai, T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50:1–19.
Mukai, T., T. K. Watanabe and O. Yamaguchi. 1974. The genetic structure of natural populations of *Drosophila melanogaster*. XII. Linkage disequilibrium in a large local population. Genetics 77:771–793.
Muller, H. J. 1928. The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. Genetics 13:279–357.
Muller, H. J. 1932. Some genetic aspects of sex. Am. Nat. 66:118–138.
Muller, H. J. 1964. The relation of recombination to mutational advance. Mutation Res. 1:2–9.
Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.
Muse, S. V. and B. Gaut. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. 13:105–114.
Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. Mol. Biol. Evol. 11:715–724.
Muse, S. V. and B. S. Weir. 1992. Testing for equality of evolutionary rates. Genetics 132:269–276.
Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Nagy, Z. and M. Chandler. 2004. Regulation of transposition in bacteria. Res. Microbiol. 155:387–398.
Nathans, J., T. P. Piantanida, R. L. Eddy, T. B. Shows and D. S. Hogness. 1986. Molecular genetics of inherited variation in human color vision. Science 232:203–210.
Neel, J. V. and E. A. Thompson. 1978. Founder effect and the number of private polymorphisms observed in Amerindian tribes. Proc. Natl. Acad. Sci. USA 75:1904–1908.
Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70:3321–3323.
Nei, M. 1975. *Molecular Population Genetics and Evolution*. American Elsevier, New York.
Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Nei, M. 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22:2318–2342.
Nei, M. and A. P. Rooney. 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39:121–152.
Nei, M. and D. Graur. 1984. Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. 17:73–118.
Nei, M. and R. K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47:253–259.
Nei, M. and S. Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press: Oxford, UK.
Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.
Nei, M. and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269–5273.

Nevo, E.. 1978. Genetic variation in natural populations: Patterns and theory. Theor. Popul. Biol. 13:121–177.

Ng, P. C. and S. Henikoff. 2005. Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. 7:61–80.

Nickerson, D. A., S. L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle and C. F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. 63:595–612.

Nielsen, D. M., M. G. Ehm and B. S. Wei. 1999. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am. J. Hum. Genet. 63:1531–1540.

Nielsen, R. and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics 158:885–896.

Nielsen, R. and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:920–936.

Nielsen, R. and Z. Yang. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. 20(8):1231–1239.

Nielsen, R., C. D. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams and M. Cargill. 2005b. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biology 3:723–733.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark and C. Bustamante. 2005a. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Nielsen, R., V. L. Bauer Dumont, M. J. Hubisz and C. F. Aquadro. 2006. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol. Biol. Evol. 2006 Oct 13; [Epub ahead of print]

Niimura, Y. and M. Nei. 2006. Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J. Hum. Genet. 51:505–517.

Nilsson, A. I., S. Koskiniemi, S. Eriksson, E. Kugelberg, J. C. D. Hinton and D. I. Anderson. 2005. Bacterial genome size reduction by experimental evolution. Proc. Natl. Acad. Sci. USA 102:12112–12116.

Noonan, J. P., G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, E. M. Rubin. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113–1118.

Nordborg, M. 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63:1237–1240.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genet. 30:190–193.

Nordskog, A. W. and F. G. Giesbrecht. 1964. Regression in egg production in the domestic fowl when selection is relaxed. Genetics 50:407–416.

Novick, A. 1955. Mutagens and antimutagens. Brookhaven Symp. Biol. 8:201–215.

Oakeshott, J. G., J. B. Gibson, P. R. Anderson, W. R. Knibb, D. G. Anderson and R. K. Chambers. 1982. *Alcohol dehydrogenase* and *glycerol-3-phosphate dehydrogenase* clines in *Drosophila melanogaster* on different continents. Evolution 36:86–96.

O'Brien, S. J., D. E. Wildt, D. Goldman, C. R. Merril and M. Bush. 1983. The cheetah is depauperate in genetic variation. Science 221:459–461.

O'Brien, S. J., D. E. Wildt, M. Bush, T. M. Caro, C. FitzGibbon, I. Aggundey and R. E. Leakey. 1987. East African cheetahs: Evidence for two population bottlenecks. Proc. Natl. Acad. Sci. USA 84:508–511.

O'Brien, S. J., M. E. Roelke, L. Marker, A. Newman, C. A. Winkler, D. Meltzer, L. Colly, J. F. Evermann, M. Bush and D. E. Wildt. 1985. Genetic basis for species vulnerability in the cheetah. Science 227:1428–1434.

Ochman, H. 2005. Genomes on the shrink. Proc. Natl. Acad. Sci. USA 102:11959–11960.

Ochman, H., E. Lerat and V. Daubin. 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc. Natl. Acad. Sci. USA 102 Suppl. 1: 6595–6599.

Ogurtsov, A. Y., S. Sunyaev and A. S. Kondrashov. 2004. Indel-based evolutionary distance and mouse-human divergence. Genome Res. 14:1610–1616.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96–98.
Ohta, T. 1982. Allelic and nonallelic homology of a supergene family. Proc. Natl. Acad. Sci. USA 79:3251–3254.
Ohta, T. 1992. The nearly neutral theory of molecular evolution. Ann. Rev. Ecol. System. 23:263–256.
Ohta, T. 1994. Synonymous and non-synonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40:56–63.
Ohta, T. and H. Tachida. 1990. Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics 126:219–229.
Ohta, T. and M. Kimura. 1971a. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68:571–580.
Ohta, T. and M. Kimura. 1971b. On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. 1:18–25.
Oliver, B. 2003. Fast males. Heredity 91:535–536.
Oliver, B. and M. Parisi. 2004. Battle of the Xs. Bioessays 26:543–548.
Orel, V. 1996. *Gregor Mendel: The First Geneticist.* Oxford University Press, Oxford UK.
Orr, H. A. and M. Turelli. 1995. Dominance and Haldane's rule. Genetics 143:613–616.
Orr, H. A. and S. Irving. 2005. Segregation distortion in hybrids between the Bogota and USA subspecies of *Drosophila pseudoobscura*. Genetics 169:671–682.
Ota, T. and M. Nei. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol. Biol. Evol. 11:469–482.
Ott, J. 1985. *Analysis of Human Genetic Linkage.* Johns Hopkins University Press.
Ott, T., J. T. van Dongen, C. Gunther, L. Krusell, G. Desbrosses, H. Vigeolas, V. Bock, T. Czechowski, P. Geigenberger and M. K. Udvardi. 2005. Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. Curr. Biol. 15:531–535.
Otto, S. P. 2003. The advantages of segregation and the evolution of sex. Genetics 164:1099–1118.
Otto, S. P. and N. H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. Genetics 147:879–906.
Pardue, M. L. and P. G. DeBaryshe. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu. Rev. Genet. 37:485–511.
Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley, J. Andrews, S. Eastman and B. Oliver. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science 299:697–700.
Paterson, A. H. 2002. What has QTL mapping taught us about plant domestication? New Phytol. 154:591–608.
Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trudson, K. R. Vyas, K. A. Frazer, S. P. Fodor and D. R. Cox. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 294:1719–1723.
Peetz, E. W., G. Thomson and P. W. Hedrick. 1986. Charge changes in protein evolution. Mol. Biol. Evol. 3:84–94.
Pelz, H. J., S. Rost, M. Hunerberg, A. Fregin, A. C. Heiberg, K. Baert, A. D. MacNicoll, C. V. Prescott, A. S. Walker, J. Oldenburg and C. R. Muller. 2005. The genetic basis of resistance to anticoagulants in rodents. Genetics 170:1839–1847.
Perez-Gonzalez, C. E., W. D. Burke and T. H. Eickbush. 2003. *R1* and *R2* retrotransposition and deletion in the rDNA loci on the X and Y chromosomes of *Drosophila melanogaster*. Genetics 165:675–685.
Petrov, D. A., Y. T. Aminetzach, J. C. Davis, D. Bensasson and A. E. Hirsh. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol. Bio. Evol. 20:880–892.
Phelan, J. P. and S. N. Austad. 1994. Detecting animal models of human aging: Inbred strains often exhibit less biological uniformity than F_1 hybrids. J. Gerontol. 49:B1–B11.
Pietrowski, D., H. Bettendorf, E.-K. Riener, C. Keck, L. A. Hefler et al. 2005. Recurrent pregnancy failure is associated with a polymorphism in the p53 tumour suppressor gene. Human Reproduction 20:848–851.

Piganeau, G., M. Gardner and A. Eyre-Walker. 2004. A broad survey of recombination in animal mitochondria. Mol. Biol. Evol. 21:2319–2325.

Pirchner, F. 1969. *Population Genetics in Animal Breeding*. W. H. Freeman, San Francisco.

Pitnick, S., T. A. Markow and G. S. Spicer. 1995. Delayed male maturity is a cost of producing large sperm in *Drosophila*. Proc. Natl. Acad. Sci. USA 92:10614–10618.

Plass, M. and E. Eyras. 2006. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. BMC Evol. Biol. 6:50.

Pollard, K. S., S. R. Salama, N. Lambert, M. A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, A. D. Kern, C. Dehay, H. Igel, M. Ares, Jr., P. Vanderhaeghen and D. Haussler. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172.

Pond, S. L, S. D. Frost and S. V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Poon, A. and L. Chao. 2004. Drift increases the advantage of sex in RNA bacteriophage Phi 6. Genetics 166:19–24.

Posada, D. 2006. ModelTest Server: A web-based tool for the statistical selection of models of nucleotide substitution online. Nucleic Acids Res. 34:W700–W703.

Powers, L. 1951. Gene analysis by the partitioning method when interactions of genes are involved. Bot. Gaz. 113:1–23.

Prakash, S. 1977. Gene polymorphism in natural populations of *Drosophila persimilis* Genetics 85:513–520.

Presgraves, D. C. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. Current Biol. 15: 1651–1656.

Presgraves, D. C., E. Severance and G. S. Wilkinson. 1997. Sex chromosome meiotic drive in stalk-eyed flies. Genetics 147:1169–1180.

Price, A. H. 2006. Believe it or not, QTLs are accurate! Trends Plant Sci. 11:213–216.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 38:904–909.

Price, T. D. and P. R. Grant. 1985. The evolution of ontogeny in Darwin's finches: A quantitative approach. Amer. Nat. 125:169–188.

Pritchard, J. K., M. Stephens and P. Donnelly. 2000a. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly. 2000b. Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189.

Przeworski, M. and J. D. Wall. 2001. Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. 77:143–151.

Ptak, S. E., A. D. Roeder, M. Stephens, Y. Gilad, S. Pääbo and M. Przeworski. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. PLoS Biol. 2:e155.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer and S. Pääbo. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. 37:429–434.

Pugesek, B. H. and A. Tomer. 1996. The Bumpus house sparrow data: A reanalysis using structural equation models. Evol. Ecol. 10:387–404.

Pybus, O. G., E. C. Holmes and P. H. Harvey. 1999. The mid-depth method and HIV-1: A practical approach for testing hypotheses of viral epidemic history. Mol. Biol. Evol. 16:953–959.

Quesada, U., E. M. Ramírez, J. Rozas and M. Aguadé. 2003. Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics 65:895–900.

Race, R. R. and R. Sanger. 1975. *Blood Groups in Man*, Sixth Edition. J.B. Lippincott, Philadelphia.

Raju, N. B. 1994. Ascomycete spore killers: Chromosomal elements that distort genetic ratios among products of meiosis. Mycologia 86:461–473.

Rand, D. M. and L. M. Kann. 1996. Excess amino acid polymorphism to mitochondrial DNA: Contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. 13:735–748.

Rand, D. M., D. M. Weinreich and B. O. Cezairliyan. 2000. Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrially-encoded proteins. Gene 261:115–125.

Rand, D. M., R. A. Haney and A. J. Fry. 2004. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol. Evol. 19:645–653.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300:1742–1745.

Rawson, P. D. 2005. Nonhomologous recombination between the large unassigned region of the male and female mitochondrial genomes in the mussel, *Mytilus trossulus*. J. Mol. Evol. 61:717–732.

Redfield, R. J. 2001. Do bacteria have sex? Nature Rev. Genet. 2:634–639.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Reed, F. A., R. G. Reeves and C. F. Aquadro. 2005. Evidence of susceptibility and resistance to cryptic X-linked meiotic drive in natural populations of *Drosophila melanogaster*. Evolution 59:1280–1291.

Reed, T. E. and J. V. Neel. 1959. Huntington's chorea in Michigan. Am. J. Hum. Genet. 11:107–136.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley, S. Kresovich, M. M. Goodman and E. S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484.

Rice, W. R. and B. Holland. 2005. Experimentally enforced monogamy: Inadvertent selection, inbreeding, or evidence for sexually antagonistic coevolution? Evolution Int. J. Org. Evolution. 59:682–685.

Rice, W. R., A. D. Stewart, E. H. Morrow, J. E. Linder, N. Orteiza and P. G. Byrne. 2006. Assessing sexual conflict in the *Drosophila melanogaster* laboratory model system. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361:287–299.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen and K. Thornton et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 2005 Jan;15(1):1–18.

Richardson, A. O. and J. D. Palmer. 2006. Horizontal gene transfer in plants. J. Exp. Bot. 2006 Oct 9; [Epub ahead of print]

Rieder, M. J., A. P. Reiner, B. F. Gage, D. A. Nickerson, C. S. Eby, H. L. McLeod, D. K. Blough, K. E. Thummel, D. L. Veenstra and A. E. Rettie. 2005. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. New Eng. J. Med. 352:2285–2293.

Riley, M. A. 1993. Positive selection for colicin diversity in bacteria. Mol. Biol. Evol. 10:1048–1059.

Riley, M. A. and J. E. Wertz. 2002. Bacteriocin diversity: ecological and evolutionary perspectives. Biochimie. 84:357–364.

Risch, N. 1990. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. 46:229–241.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Riska, B., W. R. Atchley and J. J. Rutledge. 1984. A genetic analysis of targeted growth in mice. Genetics 107:79–101.

Robertson, A. 1957. Studies in quantitative inheritance. XI. Genetic and environmental correlation between body size and egg production in *Drosophila melanogaster*. J. Genet. 55:428–443.

Robertson, H. M. 1993. The *mariner* transposable element is widespread in insects. Nature 362:241–245.

Robertson, H. M. and E. G. MacLeod. 1993. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol. Biol. 2:125–139.

Robinson, E. 2006. Insect losses, control costs climb. Western Farm Press (January 21, 2006).

Robinson, H. F., R. E. Comstock and P. H. Harvey. 1949. Estimates of heritability and degree dominance in corn. Agron. J. 41: 353–359.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rosenberg, N. A. and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Rev. Genet. 3:380–390.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky and M. W. Feldman. 2002. Genetic structure of human populations. Science 298:2381–2385.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard and M. W. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1:660–671.

Rost, S., A. Fregin, V. Ivaskevicius, E. Conzelmann, K. Hortnagel, H. J. Pelz, K. Lappegard, E. Seifried, I. Scharrer, E. G. Tuddenham, C. R. Muller, T. M. Strom and J. Oldenburg. 2004. Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427:537–541.

Roughgarden, J. 1979. *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. Macmillan, New York.

Rowan, R. G. and J. A. Hunt. 1991. Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for 5 closely related species of Hawaiian *Drosophila*. Mol. Biol. Evol. 8:49–70.

Roychoudhury, A. K. and M. Nei. 1988. *Human Polymorphic Genes: World Distribution*. Oxford University Press: New York, NY.

Rozas, J., M. Gullaud, G. Blandin and M. Aguadé. 2001. DNA variation at the rp49 gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. Genetics 158:1147–1155.

Roze, D. and N. H. Barton. 2006. The Hill-Robertson effect and the evolution of recombination. Genetics 173:1793–1811.

Rzhetsky, A. and M. Nei. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sabeti, P. C., E. Walsh, S. F. Schaffner, P. Varilly, B. Fry, H. B. Hutcheson, M. Cullen, T. S. Mikkelsen, J. Roy, N. Patterson, R. Cooper, D. Reich, D. Altshuler, S. O'Brien and E. S. Lander. 2005. The case for selection at CCR5-Delta32. PLoS Biol. 3(11):e378. Epub 2005 Nov 1.

Sackton, T. B., R. A. Haney and D. M. Rand. 2003. Cytonuclear coadaptation in *Drosophila*: disruption of cytochrome *c* oxidase activity in backcross genotypes. Evolution 57:2315–2325.

Sagitov, S. and P. Jagers. 2005. The coalescent effective size of age-structured populations. Ann. Appl. Prob. 15:1778–1797.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Salanti, G., G. Amountza, E. E. Ntzani and J. P. A. Loannidis. 2005. Hardy-Weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. Europ. J. Hum. Genet. 13:840–848.

Sanderson, M. J. and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8:374–379.

Satta, Y., C. O'hUigin, N. Takahata and J. Klein. 1993. The synonymous substitution rate of the major histocompatibility complex in primates. Proc. Natl. Acad. Sci. USA 90:7480–7484.

Sawyer, S. A. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

Sawyer, S. A. and D. L. Hartl. 1986. Distribution of transposable elements in prokaryotes. Theor. Popul. Biol. 30:1–17.

Sawyer, S. A. and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–1176.

Sawyer, S. A., D. E. Dykhuizen, R. F. DuBose, L. Green, T. Mutangadura Mhlanga, D. F. Wolczyk and D. L. Hartl. 1987. Distribution and abundance of insertion sequences among

natural isolates of *Escherichia coli*. Genetics 115:51–63.

Sawyer, S. A., R. Kulathinal, C. D. Bustamante and D. L. Hartl. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J. Mol. Evol. 57:S154–S164.

Schaeffer, S. W. and W. W. Anderson. 2005. Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*. Genetics 171:1729–1739.

Schaid, D. J. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. 70:425–434.

Scharloo, W. 1987. Constraints in selection response. pp. 125–150, *In* V. Loeschcke (ed.), *Genetic Constraints on Adaptive Evolution*. Springer-Verlag, Berlin.

Scheiner, S. M. 2002. Selection experiments and the study of phenotypic plasticity. J. Evol. Biol. 15:889–898.

Schemske, D. W. and P. Bierzychudek. 2001. Evolution of flower color in the desert annual *Linanthus parryae:* Wright revisited. Evolution 55:1269–1282.

Schlichting, C. D. and M. Pigliucci. 1994. Gene regulation, quantitative genetics and the evolution of reaction norms. Evol. Ecol. 8:1–15.

Schmalhausen, I. I. 1949. *Factors of Evolution: The Theory of Stabilizing Selection*. Blakiston, Philadelphia.

Scriver, C. R. 2001. Human genetics: Lessons from Quebec populations. Ann. Rev. Genomics and Hum. Genet. 2:69–101.

Scriver, C. R. and P. J. Waters. 1999. Monogenic traits are not simple: lessons from phenylketonuria. Trends Genet. 15:267–272.

Selander, R. K., D. A. Caugant and T. S. Whittam. 1987. Genetic structure and variation in natural populations of *Escherichia coil*. pp. 1625–1648. *In* J. L. Ingraham, K. Brooks Low, B. Magasanik, M. Schaechter and H. E. Umbarger (eds.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. American Society for Microbiology, Washington, DC.

Sen, S., J. M. Satagopan and G. A. Churchill. 2005. Quantitative trait locus study design from an information perspective. Genetics 170:447–464.

Serre, D., A. Langaney, M. Chech, M. Teschler-Nicola, M. Paunovic, P. Mennecier, M. Hofreiter, G. Possnert and S. Paabo. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol. 2(3):E57.

Sharp, A. J., S. Hansen, R. R. Selzer, Z. Cheng, R. Regan, J. A. Hurst, H. Stewart, S. M. Price, E. Blair, R. C. Hennekam, C. A. Fitzpatrick, R. Segraves, T. A. Richmond, C. Guiver, D. G. Albertson, D. Pinkel, P. S. Eis, S. Schwartz, S. J. Knight and E. E. Eichler. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat. Genet. 38:1038–1042.

Sharp, P. M. 1997. In search of molecular darwinism. Nature 385:111–112.

Shaw, C. R. 1965. Electrophoretic variation in enzymes. Science 149:936–943.

Shaw, F. H., R. G. Shaw, G. S. Wilkinson and M. Turelli. 1995. Changes in genetic variances and covariances: G whiz! Evolution 49:1260–1267.

Shaw, R. G. 1987. Maximum likelihood approaches applied to quantitative genetics of natural populations. Evolution 41:812–826.

Sheldahl, L. A., D. M. Weinreich and D. M. Rand. 2003. Recombination, dominance and selection on amino acid polymorphism in the *Drosophila* genome: Contrasting patterns on the X and fourth chromosomes. Genetics 165:1195–1208.

Shen, H., Z. Liu, S. S. Strom, M. R. Spitz, J. E. Lee et al. 2003. p53 codon 72 Arg homozygotes are associated with an increased risk of cutaneous melanoma. J. Invest. Dermatol. 121:1510–1514.

Shields, J. 1962. *Monozygotic Twins Brought Up Apart and Brought Up Together*. Oxford, London.

Shine, R., M. P. LeMaster, I. T. Moore, M. M. Olsson and R. T. Mason. 2001. Bumpus in the snake den: Effects of sex, size, and body condition on mortality of red-sided garter snakes. Evolution 55:598–604.

Shriver, M. D., L. Jin, E. Boerwinkle, R. Deka, R. E. Ferrell and R. Chakraborty. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. 12:914–920.

Siepel, A. and D. Haussler. 2004a. Combining phylogenetic and hidden Markov models in biosequence analysis. J. Comput. Biol. 11:413–428.

Siepel, A. and D. Haussler. 2004b. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. 21:468–488.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Sijen, T. and R. H. A. Plasterk. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. Nature 426:310–314.

Silva, J. C., E. L. Loreto and J. B. Clark. 2004. Factors that affect the horizontal transfer of transposable elements. Curr. Issues Mol. Biol. 6:57–71.

Simmonds, N. W. 1977. Approximations for *i*, intensity of selection. Heredity 34:413–414.

Simmons, M. F. and J. F. Crow. 1977. Mutations affecting fitness in *Drosophila* populations. Annu. Rev. Genet. 11:49–78.

Simmons, M. J., J. D. Raymond, J. B. Niemi, J. R. Stuart and P. J. Merriam. 2004. The P cytotype in *Drosophila melanogaster*: A maternally transmitted regulatory state of the germ line associated with telomeric *P* elements. Genetics 166:243–254.

Simon, M., J. Zieg, M. Silverman, G. Mandel and R. Doolittle. 1980. Phase variation: Evolution of a controlling element. Science 209:1370–1374.

Singh, N. D., J. C. Davis and D. A. Petrov. 2005. X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. Genetics. 171:145–155.

Singh, R. S. and L. R. Rhomberg. 1987. A comprehensive study of genetic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic differentiation. Genetics 117:255–271.

Singh, R. S. and R. J. Kulathinal. 2005. Male sex drive and the masculinization of the genome. BioEssays 27:518–525.

Singh, R. S., R. C. Lewontin and A. A. Felton. 1976. Genetic heterogeneity within electrophoretic "alleles" of *xanthine dehydrogenase* in *Drosophila pseudoobscura*. Genetics 84:609–629.

Sites, J. W. Jr., C. J. Basten and M. A. Asmussen. 1996. Cytonuclear genetic structure of a hybrid zone in lizards of the *Sceloporus grammicus* complex (Sauria, Phrynosomatidae). Mol. Ecol. 5:379–392.

Sjodin, P., I. Kaj, S. M. L. Krone and M. Nordborg. 2005. On the meaning and existence of an effective population size. Genetics 169:1061–1070.

Skibinski, D. O. F., C. Gallagher and C. M. Beynon. 1994. Sex-limited mitochondrial DNA transmission in the marine mussel *Mytilus edulis*. Genetics 138:801–810.

Slate, J. 2005. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. Mol. Ecol. 14:363–379.

Slatkin, M. 1985. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Slatkin, M. and B. Rannala. 2000. Estimating allele age. Annu. Rev. Genomics Hum. Genet. 1:225–249.

Smith, C. 1975. Quantitative inheritance. pp. 382–441. *In* G. Fraser and O. Mayo (eds.), *Textbook of Human Genetics*. Blackwell Scientific, Oxford.

Smith, N. G. C. and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024.

Smithies, O. 1954. Zone electrophoresis in starch gels: Group variation in the serum proteins of normal human adults. Biochem. J. 61:629–641.

Smithies, O. 1995. Early days of gel electrophoresis. Genetics 139:1–3.

Spencer, H. G. and R. W. Marks. 1988. The maintenance of single-locus polymorphism. I. Numerical studies of a viability selection model. Genetics 120:605–613.

Spielman, R. S., R. E. McGinnis and W. J. Ewens. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Amer. J. Hum. Genet. 52:506–516.

Sprague, G. F. 1978. Introductory remarks to the session on the history of hybrid corn. pp. 11–12. *In* D. B. Walden (ed.), *Maize Breeding and Genetics*. John Wiley and Sons, New York.

Stephan, W. 1996. The rate of compensatory evolution. Genetics 144:419–426.

Stephens, J. C. 1985. Statistical methods of DNA sequence analysis: Detection of intragenic

recombination or gene conversion. Mol. Biol. Evol. 2:539–556.

Stephens, J. C., D. E. Reich, D. B. Goldstein, H. D. Shin, M. W. Smith, M. Carrington, C. Winkler, G. A. Huttley, R. Allikmets, L. Schriml, B. Gerrard, M. Malasky, M. D. Ramos, S. Morlot, M. Tzetis, C. Oddoux, F. S. di Giovine, G. Nasioulas, D. Chandler, M. Aseev, M. Hanson, L. Kalaydjieva, D. Glavac, P. Gasparini, E. Kanavakis, M. Claustres, M. Kambouris, H. Ostrer, G. Duff, V. Baranov, H. Sibul, A. Metspalu, D. Goldman, N. Martin, D. Duffy, J. Schmidtke, X. Estivill, S. J. O'Brien and M. Dean. 1998. Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescenceof haplotypes. Am. J. Hum. Genet. 62:1507–1515.

Steppan, S. J., P. C. Phillips and D. Houle. 2002. Comparative quantitative genetics: evolution of the G matrix. Trends Ecol. Evol. 17:320–327.

Stewart, D. R., B. A. Dombroski, M. Urbanek, W. Ankener, K. G. Ewens, J. R. Wood, R. S. Legro, J. F. Strauss, III, A. Dunaif and R. S. Spielman. 2006. Fine mapping of genetic susceptibility to polycystic ovary syndrome on chromosome 19p13.2 and tests for regulatory activity. J. Clin. Endocrinol. Metabol. 91:4112–4117.

Stewart, D. T., C. Saavedra, R. R. Stanwood, A. O. Ball and E. Zouros. 1995. Male and female mitochondrial DNA lineages in the blue mussel (*Mytilus edulis*) species group. Mol. Biol. Evol. 12:735–747.

Stocker, B. A. D. 1949. Measurements of rate of mutation of flagellar antigenic phase in *Salmonella typhimurium*. J. Hyg. 47:398–412.

Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl Acad. Sci. USA 100:9440–9445.

Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Molec. Ecology 14:671–688.

Stumpf, M. P. H. and G. A. T. McVean. 2003. Estimating recombination rates from population-genetic data. Nature Rev. Genet. 4:959–968.

Stylianou, I. M., R. Korstanje, R. Li, S. Sheehan, B. Paigen and G. A. Churchill. 2006. Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. Mamm. Genome. 17:22–36.

Subrahmanyan, L., M. A. Eberle, A. G. Clark, L. Kruglyak and D. A. Nickerson. 2001. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am. J. Hum. Genet. 69:381–395.

Sumiyama, K., T. Kitano, R. Noda, R. E. Ferrell and N. Saitou. 2000. Gene diversity of chimpanzee ABO blood group genes elucidated from exon 7 sequences. Gene 259:75–79.

Sunyaev, S., V. Ramensky, I. Koch, W. Lathe 3rd, A. S. Kondrashov and P. Bork. 2001. Prediction of deleterious human alleles. Hum. Mol. Genet. 10:591–597.

Suoeka, N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85:2653–2657.

Suzuki, Y. and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16:1315–1328.

Sved, J. A. 1975. Fitness of third chromosome homozygotes in *Drosophila melanogaster*. Genet. Res. Camb. 25:197–200.

Sved, J. A. and F. J. Ayala. 1970. A population cage test for heterosis in *Drosophila pseudoobscura*. Genetics 66:97–113.

Swanson, W. J. and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. 3:137–144.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA 98:7375–7379.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607.

Takahata, N. 1983. Population genetics of extranuclear genomes under the neutral mutation hypothesis. Genet. Res. 42:235–256.

Takahata, N. 1984. A model of extranuclear genomes and the substitution rate under within-generation selection. Genet. Res. 44:109–116.

Takahata, N., S. H. Lee and Y. Satta. 2001. Testing multiregionality of modern human origins. Mol. Biol. Evol. 18:172–183.

Tang, H., J. Peng, P. Wang and N. J. Risch. 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28:289–301.

Tang, H., M. Coram, P. Wang, X. Zhu and N. Risch. 2006. Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. 79:1–12.

Tanksley, D. S. 1993. Mapping polygenes. Annu. Rev. Genet. 27:205–233.

Tao, Y. and D. L. Hartl. 2003. Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *Drosophila mauritiana*. III. Heterogeneous accumulation of hybrid incompatibilities, degree of dominance and implications for Haldane's rule. Evolution 57:2580–2598.

Tao, Y., D. L. Hartl and C. C. Laurie. 2001. Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. Proc. Natl. Acad. Sci. USA 98:13183–13188.

Tavaré, S. 1986. Some probalistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

Templeton, A. R. 1982. Adaptation and the integration of evolutionary forces. pp. 15–31. *In* R. Milkman (ed.), *Perspectives on Evolution*. Sinauer Associates, Sunderland, MA.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley and B. S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl Acad. Sci. USA 98:9161–9166.

Teshima, K. M., G. Coop and M. Przeworski. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

Thoday, J. M. 1961. Location of polygenes. Nature 191:368–370.

Thoday, J. M. 1979. Polygene mapping: Uses and limitations. pp. 219–234. *In* J. N. Thompson and J. M. Thoday (eds.), *Quantitative Genetic Variation*. Academic Press, New York.

Thornton, K. and M. Long. 2005. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol. Biol. Evol. 22:273–284.

Thornton, K. and P. Andolfatto. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.

Tiemann-Boege, I., P. Calabrese, D. M. Cochran, R. Sokol and N. Arnheim. 2006. High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. PLoS Genet. 2(5):e70. Epub 2006 May 5.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams and A. G. Clark. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455–462.

Tsaousis, A. D., D. P. Martin, E. D. Ladoukakis, D. Posada and E. Zouros. 2005. Widespread recombination in published animal mtDNA sequences. Mol. Biol. Evol. 22:925–933.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Popul. Biol. 25:138–193.

Turelli, M. 1988. Phenotypic evolution, constant covariances and the maintenance of additive variance. Evolution 42:1342–1347.

Turelli, M., A. A. Hoffman and S. W. McKechnie. 1992. Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. Genetics 132:713–723.

Turelli, M., J. H. Gillespie and R. Lande. 1988. Rate tests for selection on quantitative characters during macroevolution and microevolution. Evolution 42:1085–1089.

Umina, P. A., A. R. Weeks, M. R. Kearney, S. W. McKechnie and A. A. Hoffmann. 2005. A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. Science. 308:691–693.

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino, W. H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J. R. Kidd, S. Q. Mehdi, M. T. Seielstad, R. S.Wells, A. Piazza, R. W. Davis, M. W. Feldman, L. L. Cavalli-Sforza, P. J. Oefner. 2000. Y chromosome sequence variation and the history of human populations. Nature Genet. 26:358–361.

Unseld, M., J. R. Marienfeld, P. Brnadt and A. Brennicke. 1997. The mitochondrial genome

of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat. Genet. 15:57–65.
Uyenoyama, M. and M. W. Feldman. 1980. Theories of kin and group selection: A population genetics perspective. Theor. Popul. Biol. 17:380–414.
Uyenoyama, M. K. 1995. A generalized least-squares estimate of the origin of sporophytic self-incompatibility. Genetics 139:975–992.

Varshney, R. K., A. Graner and M. E. Sorrells. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23:48–55.
Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith et al. 2001.The sequence of the human genome. Science 291:1304–1351.
Verhoeven, K. J. F., K. L. Simonsen and L. M. McIntyre. 2005. Implementing false discovery rate control: Increasing your power. OIKOS 108:643–647.
Verrelli, B. C. and S. A. Tishkoff. 2004. Signatures of selection and gene conversion associated with human color vision variation. Am. J. Hum. Genet. 75:363–375.
Via, S. and R. Lande. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution 39:505–522.
Via, S., R. Gomulkiewicz, G. De Jong, S. M. Scheiner, C. D. Schlichting and P. H. van Tienderen. 1995. Adaptive phenotypic plasticity: Consensus and controversy. TREE 10:212–217.
Vogel, F. and A. G. Motulsky. 1986. *Human Genetics*, Second Edition. Springer-Verlag, New York.
Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4(3):e72.
Volkman, S., D. L. Hartl, D. F. Wirth, K. Nielsen, M. Choi, S. Batalov, Y. Zhou, D. Plouffe, K. Le Roch, R. Abagyan and E. Winzeler. 2002. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. Science 298:216–218.

Wade, M. J. and C. J. Goodnight. 1991. Wright's shifting balance theory: An experimental study. Science 253:1015–1018.
Wahlund, S. 1928. Composition of populations from the perspective of the theory of heredity. Hereditas 11:65–105 (in German).
Wakeley J. and T. Takahashi. 2004. The many-demes limit for selection and drift in a subdivided population. Theor. Popul. Biol. 66:83–91.
Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37:613–623.
Wakeley, J. 1997. Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. 69:45–48.
Wakeley, J. 1999. Nonequilibrium migration in human history. Genetics 153:1863–1871.
Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54:1092–1101.
Wakeley, J. 2003. Polymorphism and divergence for island-model species. Genetics 163:411–420.
Wakeley, J. 2005. The limits of theoretical population genetics. Genetics 169:1–7.
Wall, J. D. and R. R. Hudson. 2001. Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. 18:1134–1135.
Wall, J. D., L. A. Frisse, R. R. Hudson and A. Di Rienzo. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am. J. Hum. Genet. 73:1330–1340.
Walsh, B. 2004. Population- and quantitative-genetic models of selection limits. Plant Breed. Rev. 24 (Part 1):177–225.
Wang Y and B. Rannala. 2005. In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. Am. J. Hum. Genet. 76:1066–1073.
Wang, P. J., J. R. McCarrey, F. Yang and D. C. Page. 2001. An abundance of X-linked genes expressed in spermatogonia. Nature Genet. 27:422–426.
Wang, T. and Z.-B. Zeng. 2006. Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium. BMC Genetics 7:Art. No. 9.
Wang, X. X. and J. Z. Zhang. 2004. Rapid evolution of mammalian X-linked testis-expressed homeobox genes. Genetics 167:879–888.
Watanabe, S., S. Kondo and E. Matsunaga (eds.). 1975. *Human Adaptability, Volume 2: Anthropological and Genetic Studies on the Japanese*. University of Tokyo Press, Tokyo.

Waterston, R. H and Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal et al. 2002. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Watterson, G. A. 1978. The homozygosity test of neutrality. Genetics 88:405–417.

Weber, K. E. 1992. How small are the smallest selectable domains of form? Genetics 130:345–353.

Weber, K. E. 1996. Large genetic change at small fitness cost in large populations of *D. melanogaster* selected for wind tunnel flight: Rethinking fitness surfaces. Genetics 144:205–213.

Weill, M., C. Malcolm, F. Chandre, K. Mogensen, A. Berthomieu, M. Marquine and M. Raymond. 2004. The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. Insect Mol. Biol. 13:1–7.

Weinreich, D. M., N. F. Delaney, M. A. Depristo and D. L. Hartl. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114.

Weir, B. S. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weir, B. S. and C. C. Cockerham. 1984. Estimating *F* statistics for the analysis of population structure. Evolution 38:1358–1370.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen and W. G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15:1468–1476.

Wertz, J. E. and M. A. Riley. 2004. Chimeric nature of two plasmids of *Hafnia alvei* encoding the bacteriocins alveicins A and B. J. Bacteriol. 186:1598–1605.

Whitlock, M. C., P. C. Phillips and K. Fowler. 2002. Persistence of changes in the genetic covariance matrix after a bottleneck. Evolution 56:1968–1975.

Whittam, T. S., H. Ochman and R. K. Selander. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 80:1751–1755.

Wilkinson, G. S., K. Fowler and L. Partridge. 1990. Resistance of genetic correlation structure to directional selection on *Drosophila melanogaster*. Evolution 44:1990–2003.

Williamson, S. H., R. Hernadez, A. Fledel-Alon, L. Zhu, R. Nielsen, C. D. Bustamante. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA. 102:7882–7887.

Woese, C. R. 1981. Archaebacteria. Sci. Am. 244:98–122.

Wolfe, K. H. and W.-H. Li. 2003. Molecular evolution meets the genomics revolution. Nature Genet. Suppl. 33:255–265.

Wolfe, K. H., P. M. Sharp and W.-H. Li. 1989. Mutation rates differ among regions of the mammalian genome. Nature 337:283–285.

Wolfe, K. H., W.-H. Li and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA 84:9054–9058.

Wolfner, M. F. 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity 88:85–93.

Wong, W. S. and R. Nielsen. 2004. Detecting selection in noncoding regions of nucleotide sequences. Genetics 167:949–958.

Wright, S. 1921. Systems of mating. Genetics 6:111–178.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56:330–338.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Wright, S. 1943a. Analysis of local variability of flower color in *Lynanthus parryae*. Genetics 28:139–156.

Wright, S. 1943b. Isolation by distance. Genetics 28:114–138.

Wright, S. 1945. The differential equation of the distribution of allele frequencies. Proc. Natl. Acad. Sci. USA 31:382–389.

Wright, S. 1968. *Evolution and the Genetics of Populations. Vol. 2: Genetic and Biometric Foundations*. University of Chicago Press, Chicago.

Wright, S. 1969. *Evolution and the Genetics of Populations. Vol. 2: The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Wright, S. 1977. *Evolution and the Genetics of Populations. Vol. 3: Experimental Results and Evolutionary Deductions*. University of Chicago Press, Chicago.

Wright, S. 1978. *Evolution and the Genetics of Populations. Vol. 4: Variability within and among Natural Populations*. University of Chicago Press, Chicago.

Wright, S. I. and B. Charlesworth. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics. 168:1071–1076.

Wrobel, B., M. Torres-Puente, N. Jimenez, M. A. Bracho, I. Garcia-Robles, A. Moya and F. Gonzalez-Candelas. 2006. Analysis of the overdispersed clock in the short-term evolution of hepatitis C virus: Using the *E1/E2* gene sequences to infer infection dates in a single source outbreak. Mol. Biol. Evol. 23:1242–1253.

Wu, C.-I. and A. W. Davis. 1993. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. Am. Nat. 142:187–212.

Wu, C.-I. and E. Y. Xu. 2003. Sexual antagonism and X inactivation: The SAXI hypothesis. Trends Genet. 19:243–247.

Wu, C.-I. and W.-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci USA 82:1741–1745.

Wyckoff, G. J., W. Wang and C.-I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309.

Xu, J., A. Turner, J. Little, E. R. Bleecker and D. A. Meyers. 2002. Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: Hint for genotyping error? Hum. Genet. 111:573–574.

Yang, Y. Y., F. J. Lin and H. Y. Chang. 2004. Sex ratio distortion in hybrids of *Drosophila albomicans* and *D. nasuta*. Zool. Studies 43:622–628.

Yang, Z. 1996a. Maximum likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Yang, Z. 1996b. Among site rate variation and its impact on phylogenetic analysis. TREE 11:367–371.

Yang, Z. and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409–418.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17:32–43.

Yang, Z. and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

Yang, Z., W. S. Wong and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.

Yi, N., B. S. Yandell, G. A. Churchill, D. B. Allison, E. J. Eisen and D. Pomp. 2005. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170:1333–1344.

Yip, S. P. 2002. Sequence variation at the human *ABO* locus. Ann. Hum. Genet. 66:1–27.

Yokoyama, R. and S. Yokoyama. 1990. Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. Proc. Natl. Acad. Sci. USA 87:9315–9318.

Yokoyama, S. and F. B. Radlwimmer. 2001. The molecular genetics and evolution of red and green color vision in vertebrates. Genetics 158:1697–1710.

Yokoyama, S., L. Chung and T. Gojobori. 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. 5:237–251.

Yokoyama, S., W. T. Starmer, Y. Takahashi and T. Tada. 2006. Tertiary structure and spectral tuning of UV and violet pigments in vertebrates. Gene 365:95–103.

Youssoufian, H., H. H. Kazazian, Jr., D. G. Phillips, S. Aronis, G. Tsiftis, V. A. Brown and S. E. Antonarakis. 1986. Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. Nature. 324:380–382.

Zeng, Z.-B. 1992. Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: A further improved method. Genetics 131:987–1001.

Zeng, Z.-B. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457–1468.

Zeng, Z.-B., C. H. Kao and C. J. Basten. 1999. Estimating the genetic architecture of quantitative traits. Genet. Res. 74:279–289.

Zhang, J., H. F. Rosenberg and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA 95:3708–3713.

Zhang, J., R. Nielsen and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang, M., K. L. Montooth, M. T. Wells, A. G. Clark and D. Zhang. 2005. Mapping multiple quantitative trait loci by Bayesian classification. Genetics 169:2305–2318.

Zhang, X.-S. and W. G. Hill. 2002. Joint effect of pleiotropic selection and real stabilizing selection on the maintenance of quantitative genetic variation at mutation-selection balance. Genetics 162:459–471.

Zhang, X.-S. and W. G. Hill. 2003. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. Evolution 57:1761–1775.

Zhang, X.-S., J. Wang and W. G. Hill. 2002. Pleiotropic model of maintenance of quantitative genetic variation at mutation-selection balance. Genetics 161:419–433.

Zhang, X.-S., W. Jinliang and W. G. Hill. 2004. Influence of dominance, leptokurtosis and pleiotropy of deleterious mutations on quantitative genetic variation at mutation-selection balance. Genetics 166:597–610.

Zhang, Z., T. M. Hambuch and J. Parsch. 2004. Molecular evolution of sex-biased genes in *Drosophila*. Mol. Biol. Evol. 21:2130–2139.

Zhu, L. and C. D. Bustamante. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics 170:1411–1421.

Zollner, S. and J. K. Pritchard. 2005. Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092.

Zuckerkandl, E. and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. pp. 189–225. *In* M. Kasha and B. Pullman (eds.), *Horizons in Biochemistry*. Academic Press, New York.

Zurovcova, M. and W. F. Eanes. 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene *Dhc-Yh3* of *Drosophila melanogaster* and *D. simulans*. Genetics 153:1709–1715.

AUTHOR INDEX

Aguadé, M., 251, 353
Ajioka, J. W., 501
Akashi, H., 350
Akey, J. M., 557
Altenberg, L., 182
Alvarez, L., 294
Aminetzach, Y. T., 501
Anderson, W. W., 86, 220
Andolfatto, P., 136, 354, 480, 556
Anholt, R.R.H., 367
Aoki, K., 245, 247
Aquandro, C. F., 251, 353, 355, 476, 534
Ardlie, K., 536
Arnold, S. J., 242, 413, 414
Austad, N., 392
Avise, J. C., 188, 190
Ayala, F. J., 30, 166, 211, 212, 235

Bachtrog, D., 475, 480
Badge, M., 501
Ballard, J. W. D., 363, 365
Barrier, M., 486
Barton, N. H., 183, 185, 242, 366, 418, 420
Bateson, W., 395
Batzer, A., 505, 506
Bechsgaard, J. S., 235
Beerli, P., 304, 305, 306, 309
Begin, M., 415
Begun, D. J., 251, 353, 476
Bender, J., 500
Bensasson, D., 381
Benveniste, R. E., 335
Berg, D. E., 498, 501
Bergman, A., 246
Bergstrom, C. T., 365
Bernardi, G., 373, 471, 472
Berry, A., 34, 233, 234, 475
Bersaglieri, T., 548
Betancourt, A. J., 481
Bierzychudek, P., 285, 286
Birley, A. J., 37, 82
Bishop, J. A., 64, 301
Blumenstiel, J. P., 500, 509, 510
Bodmer, W. F., 68, 218, 224
Bonferroni, C. E., 62
Bonnell, M. L., 122
Botstein, D., 428, 545
Bouchard, T. J., 401
Braverman, J. M., 479, 480
Bridges, M. M., 188
Bromham, L., 190
Brookfield, J. F. Y., 501
Brown, W. M., 333, 363
Bruen, T. C., 190
Bulmer, M. G., 231, 401
Bumpus, H. C., 413, 414
Buonagurio, D. A., 335
Bürger, R., 420
Buri, P., 119
Burleigh, J. G., 370
Burzynski, A., 365
Bustamante, C. D., 143, 556

Caballero, A., 420
Caley, J., 415
Callinan, P. A., 505
Cann, R. L., 559, 560
Cano, J. M., 415
Cantor, C. K., 326
Capy, P., 513
Carareto, A., 508
Carter, A. B., 507
Carvajal-Rodriguez, A., 143
Carvalho, A. B., 185, 498
Casacuberta, E., 500
Casey, B., 373
Castle, W. E., 423

Castro, J. P., 508
Cavalli-Sforza, F., 539
Cavalli-Sforza, L. L., 68, 218, 224, 244, 539
Chakraborty, R., 123
Chandler, M., 502
Chao, L., 187
Chapman, T., 242
Charlesworth, B., 232, 236, 251, 353, 371, 478, 482, 496, 501, 506
Charlesworth, D., 478, 482, 506
Charmantier, A., 403
Chaw, S. M., 333
Chen, Y., 358
Chesser, R. K., 285
Cheverud, J. M., 396
Christiansen, F. B., 185, 233
Churchill, G. A., 429
Civetta, A., 493
Clark, A. G., 170, 185, 228, 232, 329, 347, 349, 365, 367, 487, 498, 520, 535, 548, 555
Clayton, G. A., 416
Clegg, M. T., 210, 211, 274, 353
Cockerham, C. C., 284
Cohen, J. E., 231
Collard, B. C. Y., 429
Comeron, J. M., 185
Conery, S., 470
Conner, J. K., 402
Cook, L. M., 64, 301, 364
Corander, J., 288, 541
Costantini, M., 472
Cotterman, C. W., 258
Coyne, J. A., 64, 495, 496
Craig, N., 498, 501
Cross, S. R. H., 37, 82
Crow, J. F., 22, 105, 121, 126, 184, 228, 232, 247, 265, 266, 419, 548
Curie-Cohen, M., 285
Curtsinger, J. W., 218
Cutler, D. J., 338, 339

Daly, M. J., 534, 535
Darvasi, A., 545
Darwin, C., 199, 200, 242
Davis, A. W., 496
De Bakker, P. I., 542
De Brito, R. A., 415, 491
DeBaryshe, G., 500
Degnan, P. H., 187, 376
Deininger, P. L., 505, 506
Dekkers, J. C. M., 429
Depaulis, F., 138
DePristo, M. A., 490
Dermitzakis, E. T., 334
Dickerson, R. E., 332
Dixon, M. T., 490
Dobzhansky, T., 211, 278, 285
Doerge, R. W., 426, 428, 429
Driskell, A. C., 370
DuBose, R. F., 188
Dudley, J. W., 365, 366
DuMouchel, W. H., 220
Dunn, L. C., 236
Duret, L., 373
Dyer, K. A., 344, 487

Eanes, W. F., 475, 480
Edwards, Y. H., 212
Ehrlich, J., 402
Eickbush, T. H., 500
Ellis, N. A., 72
Emerson, B. C., 133, 135, 136
Emigh, T., 59
Endler, J. A., 420
Enfield, F. D., 359, 372
Epling, C. H., 278, 285
Estes, S., 420
Ewens, W. J., 104, 105, 121, 167, 168, 212, 218, 232, 240
Eyras, E., 334
Eyre-Walker, A., 373, 419, 471, 472

Falconer, D. S., 8, 34, 370, 406, 410, 411, 428
Fay, J. C., 488
Feil, E. J., 188
Feldman, M. W., 182, 232, 244, 400, 402
Felsenstein, J., 131, 134, 136, 304, 305, 306, 309, 329, 367, 369, 370
Ferreira, M. U., 349
Filipski, J., 471
Fisher, R. A., 12, 102, 106, 183, 212, 242, 248, 359, 425, 428
Fitch, W. M., 334, 338, 367
Fiumera, A. C., 493
Force, A., 376
Frank, S. A., 494, 496
Fry, J. D., 418
Frydenberg, O., 233
Fu, Y., 178, 179, 362
Fuerst, P. A., 170
Fullerton, S. M., 373

Gabriel, S. B., 534
Galton, F., 4, 5, 6, 7, 8–9, 10, 14, 358, 359, 360
Game, E. T., 415
Garant, D., 403
Garrigan, D., 561
Gaut, B. S., 344, 353, 483, 486
Gerrish, P. J., 183
Gibbs, K. L., 414
Giesbrecht, F. G., 410
Gilad, Y., 378
Gilks, W. R., 305
Gillespie, J. H., 24, 162, 251, 338
Gimenez, O., 413, 414
Gingerich, P. D., 338
Gojobori, T., 335, 342–343, 486, 487
Golding, G. B., 334, 505
Goldman, N., 329, 344, 486
Goldstein, D. B., 329
Goodnight, C. J., 246
Grant, B. S., 64, 65, 551
Grant, P. R., 414, 415
Graur, D., 162, 321
Green, P., 334
Grossman, A. I., 284
Guo, S. W., 59, 61

Haddrill, P. R., 354, 482, 488
Hahn, M. W., 353
Haigh, J., 251, 556
Haldane, J. B. S., 228, 244, 266, 308, 495
Hamilton, W. D., 244
Hammer, M. F., 236, 561
Han, K. D., 507, 508
Harding, R. M., 520
Hardy, G. H., 48
Harris, H., 21, 22, 413
Hartl, D. L., 5, 155, 162, 218, 236, 237, 349, 350, 402, 470, 488, 496, 500, 501, 502, 512, 513, 556
Hasegawa, M., 328
Haussler, D., 334, 359, 555
Haymer, D. S., 218
Hedrick, P. W., 283

Hegreness, M., 183
Henikoff, S., 354
Hey, J., 306, 307
Hill, W. G., 83, 241, 366, 416, 417, 420, 483, 552
Hillis, D. M., 490
Hoekstra, H. E., 308, 309
Holgate, P., 123
Holland, B., 243
Hospital, F., 429
Houle, D., 418
Howe, M. M., 498, 501
Hubby, J. L., 16
Hudson, R. R., 128, 137, 138, 142, 143, 175, 177, 303, 352, 361, 362, 480, 538, 548, 556
Huelsenbeck, J. P., 344, 371, 487
Hughes, A. L., 349
Hunt, J. A., 510
Hurst, L. D., 373, 471, 472, 494, 496
Hwang, D. G., 334

Ikemura, T., 350
Iltis, H., 89
Innan, H., 480, 490
Inouye, L. S., 23
International HapMap Consortium 2005, 87, 144, 523, 534, 538
International Human Genome Sequencing Consortium, 508, 509
Ioerger, T. R., 235, 371
Irving, S., 497
Iwasa, Y., 242

Jaenike, J., 497
Jaffe, K., 294
Jagers, P., 133
Jansen, R. C., 428
Jeffreys, A. J., 537, 539
Jenkins, N. L., 208
Jensen, M. A., 477
Jensen, S., 508
Jessen, T.-H., 491, 492
Jiang, M., 497
Johanssen, W., 366
Jones, D. T., 415
Judson, O. P., 182
Jukes, T. H., 161, 326, 335
Jurka, J. R., 501
Jutier, D., 494, 497

Kann, L. M., 365
Kapitonov, V. V., 501
Kaplan, N. L., 480, 501
Karlin, S., 168
Karn, M. N., 413
Kathuria, A., 411
Keightley, P. D., 241, 366, 367, 418, 420
Keith, T. P., 170, 171
Kelly, W. G., 497
Kern, A. D., 493
Kettlewell, H. B. D., 65
Khil, P. P., 495
Kibota, T. T., 419
Kim, Y., 143, 336, 478, 480, 556
Kimura, M., 22, 35, 105, 106, 108, 109, 112, 121, 126, 137, 161, 172, 184, 232, 249, 250, 265, 318, 319, 328, 330, 331, 338, 345, 358, 483, 490, 532, 548
King, J. L., 161, 335
Kingman, J. F. C., 128, 133
Kingsolver, J. G., 420
Kirkpatrick, M., 242, 406
Kishino, H., 328
Kivisild, T., 363, 559
Kleene, K. C., 493
Klingenberg, C. P., 392
Knott, S. A., 403
Kolmogorov, A., 106
Kondrashov, A. S., 186, 419, 420, 490
Kong, A., 537
Korpelainen, H., 365
Krakauer, J., 557
Kreitman, M., 34, 185, 233, 234, 347, 348, 353, 358, 359, 365, 488, 555
Krings, M., 138, 561
Kruuk, L. E. B., 403
Kulathinal, R., 490, 493
Kumar, S., 162

Labbe, P., 358, 359
Lahn, B. T., 498
Lambert, J. R., 366
Lande, R., 242, 406, 413, 414, 415, 416, 420, 423, 424
Lander, E. S., 428, 545
Landry, C. R., 403
Langley, C. H., 24, 338, 483, 501
Lanigan, C. M. S., 329
Laurent, F., 203
Laurie, C. C., 366, 424, 425
Lazzeroni, L. C., 553
Le Rouzic, A., 513
Leamy, L. J., 392
Leicht, B. G., 359
Lemey, P., 335
Lemos, B., 418
Lenski, R. E., 183
Lerner, J. M., 410
Levin, D. A., 265
Levings, C. S., 365
Lewontin, R. C., 16, 23, 224, 236, 400, 402, 557
Li, W.-H., 162, 176, 178, 179, 325, 337, 347, 362, 428, 493, 495, 508, 520
Lindblad-Toh, K., 87
Liti, G., 474
Lofsvold, D., 415
Lohe, A. R., 500, 505, 512
Long, Y. A. D., 87, 375, 483
Louis, E. J., 474
Lu, J., 495
Lunzer, M., 340
Lyman, R. F., 367, 421, 422
Lynch, M., 366, 376, 402, 416, 418, 419, 470
Lyttle, T. W., 494, 497

Mackay, T. F. C., 358, 366, 370, 389, 406, 410, 411, 418, 421, 422, 426, 428
MacLeod, E. G., 512
Majerus, M. E. N., 64, 65
Makova, K. D., 495
Malécot, G., 258
Malthus, T., 206
Manly, B. F. J., 170
Maraganore, D. M., 554
Marais, G., 350
Margoliash, E., 334, 366
Marks, R. W., 224
Marth, G. T., 529
Martin, G., 185, 553
Maruyama, K., 513
May, R. M., 220, 231, 392
Maynard Smith, J., 251, 373, 556
McCarroll, S. A., 523
McDonald, J. H., 353, 488, 555
McGregor, J. L., 168
McGue, M., 401

McGuigan, K., 415
McKeigue, P. M., 544
McLellan, T., 23
McVean, G. T., 140, 141, 143, 538
Meiklejohn, C. D., 483, 494, 497
Mendel, G., 4, 5
Merbs, S. L., 378
Mercot, H., 496
Merikangas, K., 521, 552
Meunier, J., 373
Meyer, K., 406
Milkman, R., 188
Miller, R. H., 500
Mishmar, D., 365, 559
Mitchell-Olds, T., 414
Miyata, T., 531
Montgomery, E. A., 501
Montooth, K. L., 233
Moore, J. H., 395
Moran, N. A., 104, 105, 187, 376
Morrell, P. L., 275
Mosher, M. J., 389
Motulsky, A. G., 69
Mourant, A. E., 68
Mukai, T., 52, 81, 267, 419
Muller, H. J., 183, 185, 228, 419
Murphy, W. J., 337
Muse, S. V., 337, 344, 486
Myers, S., 538

Nagy, Z., 502
Nathans, J., 378
Neel, J. V., 123, 227
Nei, M., 35, 63, 123, 162, 176, 283, 285, 321, 342, 349, 368, 369, 376, 377, 378, 486
Nevo, E., 21
Ng, P., 354
Nielsen, R., 58, 306, 307, 329, 344, 347, 349, 351, 359, 486, 487, 495, 526, 529, 555, 556
Niimura, Y., 378
Nilsson, A. I., 187
Nilsson-Ehle, H., 375
Nordborg, M., 87, 128, 133, 136, 138, 140, 562
Nordskog, A. W., 410
Normark, B. B., 182
Novick, A., 155

O'Brien, S. J., 23, 122, 392
Oakeshott, J. G., 233, 234
Ochman, H., 187, 188
Ogurtsov, A. Y., 329
Ohta, T., 22, 105, 112, 249, 337, 338, 339, 358, 374, 375, 483, 532
Oliver, B., 493, 497
Orel, V., 5
Orr, H. A., 495, 496, 497
Ota, T., 376, 377
Ott, J., 537
Ott, T., 330
Otto, S. P., 183, 185, 241

Page, D. C., 498
Palmer, J. D., 371
Pardue, M.-L., 500
Parisi, M., 494, 497
Paterson, A. H., 426
Pauling, L., 329
Peetz, E. W., 346
Pelz, H. J., 220
Penrose, L. S., 413
Perez-Gonzalez, C. E., 500
Petrov, D. A., 505, 511
Phelan, J. P., 392
Pietrowski, D., 54
Piganeau, G., 190
Pitnick, S., 493
Plass, M., 334
Plasterk, R. H. A., 500
Pollard, K. S., 358, 555
Pomiankowski, A., 242, 494, 496
Pond, S. L., 329
Poon, A., 187
Posada, D., 329
Powers, L., 381
Prakash, S., 70
Presgraves, D. C., 355, 481, 497
Price, A. H., 426, 429
Price, A. L., 542
Price, T. D., 415
Pritchard, J. K., 287, 365, 541, 553
Przeworski, M., 481, 536, 556
Ptak, S. E., 539
Pugesek, B. H., 414
Punnett, R. C., 49
Pybus, O. G., 135, 136

Quesada, U., 483

Race, R. R., 72
Radlwimmer, F. B., 492
Raju, N. B., 236
Rand, D. M., 365
Rannala, B., 548, 553
Ranz, J. M., 494, 497
Rawson, P. D., 365
Redfield, R. J., 183, 187
Redon, R., 523
Reed, T. E., 227, 497
Reich, D., 543
Remington, D. L., 483
Rhomberg, L. R., 303
Rice, W. R., 243
Richards, S. Y., 347
Richardson, A. O., 371
Rieder, M. J., 220
Riley, M. A., 235, 349
Risch, N., 521, 551, 552
Riska, B., 406
Robertson, H. M., 241, 391, 416, 483, 511, 512
Robinson, A. R., 215
Roff, D. A., 415
Ronquist, F., 371
Rooney, A. P., 376
Rosenberg, N. A., 128, 133, 136, 140, 287, 288, 541, 542
Rost, S., 220
Rotimi, C., 561
Roughgarden, J., 231
Routman, E. J., 396
Rowan, R. G., 510
Roychoudhury, A. K., 63
Rozas, J., 481, 482, 483
Roze, D., 183
Rzhetsky, A., 369

Sabeti, P. C., 550, 557
Sackton, T. B., 365
Sadler, L. A., 520
Sagitov, S., 133
Saitou, N., 368
Salanti, G., 58
Sanderson, M. J., 370
Sanger, R., 72
Satta, Y., 235
Sawyer, S. A., 188, 488, 489, 501, 502, 503, 504, 517, 556
Schaeffer, S. W., 86
Schaid, D. J., 83, 532, 552
Scharloo, W., 373, 375

Scheiner, S. M., 406
Schemske, D. W., 285, 286
Schmalhausen, I. I., 403
Scriver, C. R., 123, 547
Selander, R. K., 23, 122
Sen, S., 428
Sharp, P. M., 523
Shaw, C. R., 16
Shaw, F. H., 415
Shaw, R. G., 403, 414, 415
Shen, S., 54
Shields, J., 401
Shifman, S., 545
Shine, R., 413
Shriver, M. D., 329
Siepel, A., 334, 359, 555
Sijen, T., 500
Silva, J. C., 513
Silver, L. M., 236
Simmons, M. F., 228, 266, 419, 500
Simon, M., 157
Singh, R. S., 24, 303, 350, 493
Sites, J. W., Jr., 365
Sjodin, P., 133, 134
Skibinski, D. G., 365
Slate, J., 426
Slatkin, M., 303, 548
Smith, C., 402
Smith, N. H., 488
Smithies, O., 16
Sondhi, K. C., 373
Spassky, B., 211
Spencer, H. G., 224
Spielman, R. S., 551
Sprague, G. F., 275
Stam, L. F., 428
Stephan, W., 143, 358, 478, 480, 490, 556
Stephens, J. C., 188, 548
Steppan, S. J., 415
Stewart, D. T., 170, 365, 565
Stocker, B. A. D., 157
Storey, J. D., 62, 554
Storz, J. F., 480
Stumpf, M. P. H., 140
Subrahmanyan, L., 536
Sumiyama, K., 371
Sunyaev, S., 354
Suoeka, N., 471
Suzuki, Y., 487
Sved, J. A., 211, 212, 266
Swanson, W. J., 486, 493
Szent-Gyorgyi, A., 16

Tachida, H., 249
Tajima, F., 35, 133, 175, 176, 177, 178, 362
Takahata, N., 365, 561
Tang, H., 541, 544
Tanksley, D. S., 426
Tao, Y., 494, 496, 497
Tavaré, S., 328
Templeton, A. R., 218
Tenaillon, M. I., 87
Teshima, K. M., 557
Thoday, J. M., 425
Thompson, E. A., 59, 61, 123
Thornton, J. M., 136, 375
Tibshirani, R., 62, 554
Tiemann-Boege, I., 537
Tishkoff, S. A., 548
Tomer, A., 414
Tracy, M. L., 166
Tsaousis, A. D., 190
Turelli, M., 415, 416, 417, 420, 496

Umina, P. A., 233
Underhill, P. A., 559, 561
Unseld, M., 364
Uyenoyama, M. K., 235, 244

Vacquier, V. D., 493
Varshney, R. K., 429
Venter, J. C., 520
Verhoeven, K.J.F., 62
Verrelli, B. C., 378
Via, S., 406
Vogel, F., 68
Voight, B. F., 557
Volkman, S., 475

Wade, M. J., 246, 413
Wakeley, J., 127, 142, 175, 306, 334, 488
Walhlund, S. G. W., 288
Wall, J. D., 138, 536
Walsh, J. B., 366, 402, 418
Wang, L., 428, 495, 498, 553
Watanabe, S., 38
Waters, P. J., 547
Waterston, G. A., 530
Watterson, G. A., 35, 137, 169, 172, 175, 354
Wayne, R. K., 392
Weber, K. E., 371, 412
Weill, M., 215
Weinberg, W., 48
Weinreich, D. M., 340
Weir, B. S., 59, 82, 83, 284, 337, 532, 540, 552
Welch, M., 182
Wertz, J. E., 235, 349
Whitlock, M. C., 415
Whittam, T. S., 171, 188
Wilkinson, G. S., 415
Williamson, S., 528, 529, 557, 558
Woese, C. R., 366
Wolfe, K. H., 333, 472, 493
Wolfner, M. F., 493
Wong, C., 359
Wright, S., 102, 105, 106, 108, 116, 121, 226, 245, 248, 258, 263, 266, 279, 280, 281, 283, 285, 308, 379, 423
Wright, S. I., 353
Wrobel, B., 339
Wu, C.-I., 337, 495, 496, 497
Wyckoff, G. J., 493

Xu, J., 58, 497

Yang, Z., 329, 334, 344, 486, 487, 497
Yano, T., 328
Yi, N., 428
Yokoyama, R., 492
Yokoyama, S., 335, 378, 492
Youssoufian, H., 334

Zeng, Z. B., 423, 428
Zhang, D., 428
Zhang, J., 344
Zhang, X.-S., 420, 421, 422, 494, 495
Zhu, L., 143
Zollner, S., 553
Zuckerkandl, E., 329
Zurovcova, M., 475, 480

SUBJECT INDEX

Italic page numbers refer to illustrations

ABO blood groups, 68–69, 233, 371
Acinonyx jubatus (African cheetah), 23, 421–422
adaptation
  evolutionary, 200
  human-specific, 554–558
  neutral polymorphisms and, 161
  norm of reaction and, 435–436
  rapid evolution, 486
  species-specific change, 358
  structure-function analysis, 491–492
adaptive mutations, 359
adaptive topography, 225–226, 246
additive genetic covariance, 441
additive genetic variance, 415
  age and, 436–437
  heterozygosity and, 446
  mutations and, 446, 448
  natural selection and, 429
  total genetic variance and, 453
Adh genes
  geneaology, 358, *359*
  nucleotide diversity, 347–348, *348*
  Phlox cuspidata, 70
  polymorphisms, 233, *234*
admixture mapping, 544, *545*
African American subpopulations, 289
  linkage disequilibrium, 544
  in protein-coding genes, *521*
  SNP distribution, *540*
African cheetah (*Acinonyx jubatus*), 23, 421–422
African subpopulations, *543*, 561
age/aging
  body weight and, 443, 444
  phenotypes and, 436–437
  population structure and, 232
AIDS, 57, 549–550
Ainu people, Northern Japan, 38
air pollution, 64, 301–302
albinism, 289
alcohol dehydrogenase
  Adh alleles, 70, 233, *234*, 352–353, 358, *359*
  gene coding for, 33, *33*
  Phlox cuspidata alleles, 70
  polymorphisms, 233, *234*
alkaline phosphatase (*phoA*) gene, 188
allele frequencies
  adaptive topography, 225–226
  allele age estimated via, 548–549
  artificial selection and, 412–413
  doubletons, 527
  electrophoretic analysis, 19
  equilibrium, 226–227

equilibrium values, 156–157, *157*, 216, *217*
estimates by species, *302*
Ewens-Watterson test, 168–172
fitness and, *213*
in gametes, 230
in haploid organisms, 205–206
inbreeding and, 261
mutation pressure and, 152–154, *153*, 156–157
mutation-selection balance and, *228*
neutral alleles, 318–321, *319*
one-way migration and, 296–297, *296*
overdominance and, *221*
random genetic drift and, *99*, 100, 158–161, 286
relative fitness and, 207–212
selection and, *360*
singletons, 527, 528
site frequency spectra, *249*, 525
in subpopulations, *299*
truncation points and, *413*
Wright-Fisher model and, *103*
allele frequency spectra, 166–167, *477*, *478*
alleles
absorption time, 112–113
additive, 405
ancestral lineage of, 159–160, *159*, *160*
ancestry of, *129*
coalescence, 128–138
estimated age of, 548–549
evolutionary history, 172
fixation index, 116
fixation of, 96, 97, 99
identical by descent, 115
identical by kind, 118
identical by state, 118
infinite-alleles model, 162–172
lost, 97, 99
Mendelian variation and, 12
mutant, 145
neutral, *319*
persistence of, *113*
polymorphism, 18
random association, *74*
random sampling of, 158, *159*
rare, 19
recessive, 288–294
segregating, 97
selectively neutral, 112
time to fixation, 112–113
allelic configurations, 167
allozygosity, definition, 259
allozygous genotypes, 163
allozymes, 18, 20–21, 23–24
α-globin genes, 329, *330*, *331*, *377*
altruism, 243, 244, *244*
Alu elements, 507, *509*
*Alu*I restriction enzyme, 26
AluYb subfamilies, 507, *508*
amino acid polymorphisms, 165
amino acid replacements
compensatory, 489–491
constraints on, 346, 349
in *Drosophila*, 480–481
index of dispersion, 493
male-biased genes, 494
pathogenic, 490
patterns of, 340–351
positive selection and, 486, 487–489
purifying selection and, 484–485
rates, 322–326
amoeba (*Amoeba dubia*), 470
amplified fragment length polymorphism (AFLP), 30–31, *31*
ancestral lineages, 529, *529*
ancestral recombination graphs, 142
ancestral sequences, *323*
Andean goose (*Chloephaga melanoptera*), 491
angiosperms, 364, *364*
anonymous polymorphism, 519
Anopheles gambiae (mosquito), 490
Anser anser (graylag goose), 491, 492
Anser indicus (bar-headed goose), 491
Antirrhinum majus (snapdragon), 89
α_1-antitruypsin deficiency, 289
aphids (*Buchnera aphidicola*), 187
apolipoprotein E (*ApoE*) genotyping, *553*
apoptosis, 486
approximate Bayesian coalescence methods, 143
Arabidopsis lyrata, 486
Arabidopsis thaliana, 486
inbreeding in, 87
mtDNA genome, 364
outcrossing in, 87
SI locus, 235
Archea, reclassification, 366
Arianta arbustorum (snail), 394
artificial selection
allele frequencies and, 412–413
correlated responses and, 440
G changes and, 445
mechanisms, 182
pleiotropic effects and, 439–440
realized heritability and, 394–403
ascertainment bias, 524–527, *525*

asexual reproduction, 182, 185
Ashkenazi Jew subpopulations, 289
Asian subpopulations, *540*
assemblages, human, 305
association, genome-wide, 553–554
association mapping, *553*
association tests, 62, 390, *392*
assortative mating, 294–295
Astyanax fasciatus (blind cave fish), 492
Australian lungfish (*Neoceratodus forsteri*), 470
autozygosity, 164, 258–259
autozygous genotypes, 163
average distance method, 366–367

background selection, 356, 478, *479*, 480
bacteria, piecewise recombination, 187–189
bacterial chemostat, 154, *154*, *155*
bacteriophage ϕ6, 187
balance hypothesis, 15
balanced polymorphisms, 216
balancer chromosomes, 211, 448
balancing selection, 359
bar-headed goose (*Anser indicus*), 491
barley (*Hordeum vulgare*), 274
Bateson, William, 425
Bayesian coalescence methods, 143
Bayesian methods, 370
bean (*Phaseolus vulgaris*), 398, *398*
beta globin gene, 331, *342*, 487, 520
β-gene deletion, 377
biased gene conversion, 472
binomial distributions, 96
binomial sampling, 95–101
biometricians, 12
birth-and-death process, 376–378, *377*
Biston betularia (peppered moth), 64, 65, 300, *301*, 302
blind cave fish (*Astyanax fasciatus*), 492
blood types, 68–69, 233, 371
body weight
  age and, *436*
  mortality rates and, 443, 444
Bonferroni, Carlo Emilio, 62
Bonferroni correction, 62
bootstrapping analysis, 370
bottlenecks, population size and, 122–123
Brassica campestris (turnip), 364
breast cancer (*BRCA1*) gene, 547
Brevundimonas diminuta, 470
Brownian motion, 105
Buchnera aphidicola (aphids), 187
budding yeast. *see Saccharomyces cerevisiae*
bugs-in-a-box analogy, 134
Bumpus, Hermon C., 443, 444

C peptide, 347
C-value paradox, 470–471, *471*
Caenorhabditis elegans
  codon bias, 350
  daf-2 gene, 208
  mariner-like elements, *511*, 512
  mutation accumulation experiments, 449
  X chromosome, 497
camouflage, 64–65, *65*
cancer, 54–55, 547
candidate genes
  identification of, 386
  for quantitative traits, 452–461
carbamate insecticides, 215
case-control designs, 550
cat flea (*Ctenocephalides felis*), 512
categorical traits, 387, 433–442
Caucasian subpopulations, 431. *see also* European subpopulations
CCR5 gene, 57, 60–61, 549–550
central limit theorem, 9
Centre d'Étude du Polymorphism Humaine (CEPH), 522
Chaetodipus intermedius (rock-pocket mice), 308
cheetah (*Acinonynx jubatus; A.j. raineyi*), 23, 421–422
chemokine receptor CCR5, 549–550
chemostats, bacterial, 154, *154*, *155*
chi-square tests, 55–56, *57*
chimpanzees
  ABO blood groups, 371
  recombination hot spots, 539
  red opsin gene (*OPN1LW*), 378
Chinese subpopulations, 522
Chloephaga melanoptera (Andean goose), 491
chloroplast DNA (cpDNA), 362–366, 363–364
chloroplasts, transmission, 363–364
choline kinase gene, 501
chromosome 5, human, *535*
clade of haplotypes, 549
classical hypothesis, 15
clines, definition, 233
clonal frame, 189
clonal interference, 183, 185
coalescence
  of ancestral lineages, 258
  applications in methods in, 137–138
  estimations of migration, 303–307
  inbreeding and, 259
  models with mutation, 136–137
  population growth and, 134–136

random genetic drift and, 128–138
in single nucleotide polymorphisms, *141*
time to, 130
coalescent effective size, 133
coalescent trees. *see also* gene trees
branch length, 139
interval length in, 139
sample size of 5, *179*
segregating sites, 139
codon substitution models, 343–344, 555
codons
evolution of, 487
usage biases, 349–351, *351*
colicins, 349
color, industrial melanism, 64
color vision, 377–378, 492
competition, male-male, 242, 243
complementary base pairing, 356–357
complex traits, 385–386
composite interval mapping, 458
composite likelihood analyses, 143
composite linkage disequilibrium estimator, 532
composite transposons, 502–505, 504
concerted evolution, 372, *373*, 374–375, *377*
concordance, twin data and, 439, *439*
congenital total color blindness, 123
conjunction, recombination and, 187
continuous traits, 387, 407–408
continuous variation, 5
corn (*Zea mays*), *395*, 396
correlation coefficients, 391, *392*
courtship, 242–243
covariance
additive genetic, 441
concept of, 390
environmental, 441
CRH1 gene, 433–434, *434*
Ctenocephalides felis (cat flea), 512
Culex pipiens, 356, *357*
cumulative selection differential, 415–416, *416*
cystic fibrosis, 66, 289
cystic fibrosis gene (*CFTR*), 547, 548

daf-2 gene, 208
Darwin, Charles, 14, 199
Darwinian fitness, 206, 207
demasculinization, 497–498
demes, 46, 127
density-dependent selection, 230
diabetes mellitus, 437
diastrophic dystrophy, 123
dichotomous traits, 387
differential selection, 229–230
diffusion approximation, 105–113
diffusion equation, *109*
diffusion theory, 548–549
diploid organisms
allele frequency in, 208–212
selection in, 206–215
sporophytes, 236
directional selection, 394, 449
discrete variation, 5, 10–12
disease carriers, 4
dispersion index, 338–339
dissortative mating, 294
distributions
binomial, 96
of heights, *6*, *7*
leptokurtic, 450–451
of means, 7
normal, 5–6, *6*, *9*, *10*, *407*, *408*
parameters, 389–390
phenotypic, 4, *421*
population parameters, 389
posterior, 143
variance and, 7
divergence
definition, 321
interspecific, 554–555
mtDNA restriction sites, 333, *333*
in noncoding sequences, 354
polymorphism and, 487–489
divergence rates, 354
divergent evolution, *377*
diversifying selection, 234–235
dizygotic twins, 431
DNA. *see also* mitochondrial DNA
chloroplast, 362–366
continuous variation and, 5
electrophoresis of, 16, *17*
genomic, 471–473
noncoding regions, 34
organelle, *364*
polymerase chain reactions, 28–31
randomly amplified polymorphic, 329
selfish, 501
Southern blot procedure, 27, *27*
transposable elements, 151
transposons, 499–500
DNA barcoding project, 39
DNA fingerprinting, 38, 39
DNA polymerases, 362
DNA repair, 182–183
DNA sequences
aligned, 34
evolutionary history, 172

genetic variation and, 34
polymorphisms in, 25–38
population genetics and, 4
sample, *173*
DNAsp software, 329
domesticated animals, 4, 394–403. *see also* artificial selection
dominance
complications of, 62–63
fitness and, *309*
heritability and, 405–406
selection coefficients and, *309*
dominance theory, 496
dominance variance, 423, *424*
dosage compensation, 497
Drosophila erecta, 512
Drosophila mauritana, 155
Drosophila melanogaster
Adh alleles, 233, 284, 347–348, *348*, 352–353
Adh gene, *359*
Adh region, 33–34, *33*, *37*
allele frequencies, 303, *303*
amino acid replacements, 488, *489*
autozygosity in, 265
bristles, 387, *388*, *451*, 452, 455
codon bias, 350
codon-usage bias, 351, *351*
Curly offspring, 266
D. persimilis divergence time, 307
$E6^F$ allele, 52
$E6^S$ allele, 52
esterases, 81
ESTs, 486
genome size, 381
GI (*Glued* eyes) allele, 210, *210*
jockey transposable element, 510, *510*
local recombination rates, 355
male accessory gland proteins, 493, *494*
mutation accumulation experiments, 448
noncoding sequences, 354
polymorphisms, 476–477
random genetic drift in, 100, *101*
recombination rates, 476
response to selection, *392*
segregation distortion in, 236
shared polymorphisms, 371
somatic cell chromosomes, *475*
substitution rates, *343*
superoxide dismutase in, 137
transposable elements, 505
wing veins, 403, *404*
X chromosome, 497
Y chromosome, 480
Drosophila persimilis, 70, 307
Drosophila pseudobscura, 490
balancer chromosomes, 211
homozygosity of samples, 169–170
inversions, 86
xanthine dehydrogenase gene in, 24
Xdh gene, 172
Drosophila sechellia, 352–353
Drosophila simulans
Adh alleles, 353
amino acid replacements, 488, *489*
codon bias, 350
ESTs, 486
haplotypes, *482*
local recombination rates, 355
Rh3 gene, 30
RpL32 gene, 481
substitution rates, *343*
Wolbachia infection, 366
Drosophila spp.
chromosomal variation, 15
haplotype blocks in genome, 87
linkage disequilibrium and, 483–484
P-element insertions, 396–397
proportion of polymorphic genes, 22
recombination rates, 185
superoxide dismutase, 556
transposons, 500, 501
wing veins, 442
Y chromosome, 498
Drosophila subobscura, 86
Drosophila yakuba, 350, 353, 355, 488
dual sampling process models, 365
Duchene muscular dystrophy (*DMD*) gene, 547

*Eco*RI restriction enzyme, 26, 30–31
eelpout fish (*Zoarces viviparus*), 233
effective population size. *see also* population size
homozygosity and, 164
mutation-selection balance and, 309
random genetic drift and, *118*, 121–128
simulations and, 142
in subdivided populations, 127–128
unequal sex ratios and, 123–125, *124*
Egypt, migrations, 305–307, *306*
eigenvalue effective size, 121
electrophoresis, 16–18, *17*
elephant seals, 123
EM algorithm, 68–69
ENCODE project, 523
ENCODE region (ENm001), *336*
endangered species, 4

environment
covariance, 393
genotype/phenotype gap and, 16
heritabilities and, *430*, 434–435
heterogenous, 232–234
phenotypic variance and, 416–422
substitution rates and, 338
twin studies and, 431
environmental covariance, 441
environmental variance, 420
enzymes, 18, 24. *see also* restriction enzymes; *specific* enzymes
epistasis, 239–240, 425–426
epistatic variance, 424
equilibrium frequencies, 226–227
equilibrium values
allele frequency and, 156–157, *157*, *217*
local stability, 216, 220–222
with selection, 215–226
stable, 223–225
Escherichia coli
6-phosphogluconate dehydrogenase coding, 381
chromosome, 189
codon bias, 350
gene identity, *171*
gene trees, 189
insertion sequences, 502–504
IS*30* elements, *503*
mutation accumulation experiments, 449
mutation rates, *155*
phoA gene in, *188*
Ester-2 allele, 356, *357*
Esterase-B gene, 274
Esterase-D gene, 274
estimates, of parameters, 390
ethanol toxicity, 233, *234*
euchromatin, 510, 511
eukaryotes
mutation rates, 185–187
recombination in, 182
transposable elements in, 505
European subpopulations, 289, 522, *540*. *see also* Caucasian subpopulations
evolution
α-globin genes, *330*, *331*
accelerated, 487
chloroplast DNA, 362–366
concerted, 372, *373*, 374–375, *377*
context and, 251
divergent, *377*
evidence for selection, 285–286
intercorrelated traits, 444–445
mariner-like elements, *513*
migration and, 296
mitochondrial DNA, 362–366
molecular, 318–321
nucleotide substitution rates, 322
patterns of polymorphisms in, 39
of pseudogenes, 346–347
quantitative genetics, 385–466
random genetic drift and, 446–448
of recombination rates, 241
of resistance to insecticides, 215
shifting balance theory, 246
evolutionary adaptation, 200
evolutionary quantitative genetics, 443–452
Ewens sampling formula, 166–168
Ewens-Watterson sampling theory, *168*
Ewens-Watterson test, 168–172
exact tests, 82
expressed sequence tags (ESTs), 486
extended haplotype homozygosity (EHH) test, 557
external branches, 178–180
extinction, risk of, 159–160

false discovery rates, 554
false positives, 62
familial hyperchylomicronemia, 123
faster-male theory, 496
fecundity selection, 206, 231–232
female reproduction, 496
fingerprints, 430
Fisher, Ronald Aylmer, 12–14
Fisher-Muller model, 183, 185
fitness
adaptive topologies and, 225–226
altruism and, 243
Darwinian, 206
differential selection, 229–230
dominance and, *309*
gametic selection and, 236
heritability in the narrow sense, *429*
heterogametic sex, 496
heterogenous environments and, 233–234
infinite-alleles model, *360*
intercorrelated traits and, 444–445
interdeme selection and, *247*
kin selection and, 243–244
L. parryae subpopulations, 285–286
Malthusian, 206
marginal, 212–214
mutation load and, 228–229
natural selection and, 386
neutral theory and, 15

population growth and, 201–206
random mating and, 217
relative, 207
reproductive, 440
survival and, 200
two-locus, 239–240, *239*
fixation
background selection and, 480
of mutations, 477
fixation index, 116
adaptive mutations, 359
definition, 281–285
heterozygosities and, *287*
island model of migration and, *300*
recursion for *F*, 116–117, *117*
Wahlund's principle and, 291
flour beetles (*Tribolium castaneum*), *389*, 402
fluctuating asymmetry, 422
founder effects, 123, 482
fourfold degenerate sites, 340
fraternal twins, 431
frequency-dependent selection, 230. *see also* stabilizing selection
frequency of recombination, 75–82
F_{ST} tests, 557
Fu and Li test of fit, 178–182, 362
full likelihood analyses, 142–143
fundamental theorem of natural selection, 212

G matrix, 445
Galton, F., 4, 5–6, 9, 14, 389, 390
gametes
allele frequencies in, 230
random union of, 50–52
gametic phase disequilibrium, 78
gametic selection, 206, 236
gametogenesis, 497
gametophytes, 236
G+C content, 470
gene conversion
biased, 472
definition, 86, 372, 374
genetic variation and, 474, *474*
linkage disequilibrium decay and, 536
unbiased, 473
gene duplication, 372, *373*
gene flow
among human assemblages, 305
among subpopulations, 113–114
migration and, 295
gene identity, *171*
gene mapping resolution, 521
gene pools, 138
gene transfer, horizontal, 371
gene trees. *see also* coalescent trees
coalescence in, *132*, 134–136, *135*
construction of, 358
geneologies, 358–362
generation-time effect, 337–338
genes
elimination of, 376–378
frequencies, *96*
geneologies, 358–362
heterozygosity, 162–163
linkage, *76*
mosaic ancestry, 188–189
nucleotide sites, *30*
paralogous copies, 376
polymorphism of, 20–21
genetic architecture, 550
genetic coadaptation, 86
genetic codes, *341*
genetic counseling, 4
genetic differentiation, 275
genetic divergence, 288, 299–303
genetic "draft," 251
genetic drift, 48, 285–286, 359
genetic hitchhiking, 211, 251
genetic mapping, 4
genetic variance-covariance matrix, 445
genetic variation
additive, *424*
DNA sequence and, 34
maintenance of, 14–15
Mendelian inheritance and, 53–54
mutations and, 151–158
in natural populations, 432–433
organization of, 45–92
polymorphism and, 321
population genetics and, 4
random genetic drift and, 446–448
superoxide dismutase gene, 361
total, *424*
genomes
composition of, 470–473
definition, 5
sizes of, 470–473
genomic DNA, 471–473
genomics, 452–453, 469
genotype-by-environment associations, 420
genotype-by-environment interactions, 418–419
genotype-by-sex interactions
QTLs and, 419–420
genotype frequencies
Hardy-Weinberg equilibrium and, 54–55
inbreeding and, 259–265, *262*, *263*

in populations, 46
in subpopulations, 277, 291–293
with three or more alleles, 67–70
genotypes
allozygous, 163
autozygous, 163
definition, 3
heterozygosity, 115–116
phenotypic variance and, 416–422
random mating of, 50–52
selection and, 206–208
two-locus, *406*
genotypic variance, 420
genotypic variation, 422–426
genotyping, 460, 523
Geomys pinetis (pocket gophers), 189–190, *190*
germ plasm preservation, 4
GI (*Glued* eyes) allele, 210, *210*
Gillespie, John, 251
global warming, 233
globally stable equilibrium, 216
glucose-6-phosphate dehydrogenase (G6PD), 218, *219*, 549
glycophorin A, 79–80
glycophorin B, 79–80
Gonodontis bidentata (scalloped hazel moth), 64, 300, *301*, 302
graft acceptance, 23
graylag goose (*Anser anser*), 491, 492
Grundlach-Bundschu Winery, 169–170, *171*
Guinea pigs, 409–411, *409, 410*

Haldane, J.B.S., 228, 244
Haldane-Muller principle, 228–229, 449
Haldane's rule, 495–497
haploid organisms
allele frequencies in, 205–206
gametophytes, 236
selection in, 200–206, *202*
haplotype blocks, 87
haplotypes
definition, 34, 80
description, 481
geneaology of, *534*
nonrandomness, 481
SNP inferences, 524
Haploview, 534
HapMap Database, 523
HapMap Project, 521–524, 524, *527, 528*, 532
HAR1 gene, 358, 555
Hardy, G. H., 48
Hardy-Weinberg equilibrium (HWE)
allele frequencies and, 49
exact test for, 58–60
frequency of heterozygotes and, 66, *67*
permutation test for, 60–61
sample size and, 58
subdivided populations and, 292
testing for, 54–67
in three or more alleles, 67–70, *68*
in X-linked genes, 70–73
Hardy-Weinberg principle
demonstration, *51*
departure from, 527
genotype frequencies and, 18, 19
implications of, 52–54
overview, 48–50
random mating of genotypes in, 50–52
random union of gametes in, 50–52
harmonic mean, 122
heights, distribution of, *6, 7*
hemoglobins, 345, 491–492. *see also* sickle-cell anemia
Hemophilus influenzae, 189
heritability
age and, 437
in the broad sense, 423, *430*
cumulative selection differential and, *416*
dominance and, 405–406
family resemblance and, 388–394
linearity of change, 415–416
in the narrow sense, 423, *425, 429*, 441, *460*
quantitative genetics and, 393
realized, 394–403
twin studies, 430–432
variance ratios and, 423
HeT-A, 500
heterochromatin, 510, 511
heterogametic sex, 496
heterogeneity, 486
heterogenous environments, 233–234
heteroplasmy, 365
heterosis, 275
heterozygosities
additive genetic variance and, 446
allele numbers and, 162
deficiencies of, 115
fixation indexes and, 287–287, *287*
frequency of, 66, *67*
inbreeding and, 293–294
polymorphisms and, *21*, 22–23
population size and, *321*
random mating and, 281–285
selection and, *250*
theoretical curves, *119*
virtual, 118

heterozygote inferiority, 222–223, *223*
heterozygote superiority, 216–220
heterozygous genotypes
allozyme studies of, *21*
equilibrium, 166
meiotic drive and, 236–238
Mendelian variation and, 12
hidden Markov-chain models, 555
hierarchical *F*-statistics, 282–283
hierarchical population structures, 276
Hill-Robertson effect, 183, 185, 241, 477
histone H4, 333, 345
hitchhiking effects, 211, 251, 366, 477, 480
HKA test, 351–354, *352*, 362
"HKY" model, 328
HLA regions, *533*
Homo erectus, 559
homozygosity mapping, 544–547, *546*
homozygous genotypes, 12, 166, *260*, 264
Hopi subpopulations, 289
Hordeum vulgare (barley), 274
horizontal transmission, 371, 511–513. *see also* lateral gene transfer
house sparrows (*Passer domesticus*), 443–444
Human Gene Mutation Database, 547
human genome
DNA transposons in, 500
linkage disequilibrium and, 483–484, 532–539, 542–543
mitochondrial, 363
recombination hot spots, *538*, 539
single nucleotide polymorphisms in, 61–62
whole-genome shotgun sequence, 520–521
Human Genome Project, 519
human immunodeficiency virus (HIV), 135–136, 335, 381–382, 549–550
human populations
ABO blood groups, 371
aboriginal populations, 233
assortative mating in, 294
heritability in the broad sense, *430*
inbreeding in, 266–268
mitochondrial pathologies, 365
Neanderthals and, 138
olfactory receptors, 378
origins of, 558–562
polymorphisms, 520–531
population genetics, 519–565
red opsin gene (*OPN1LW*), 378
subpopulations, 305
twin studies, 430–432
Huntington disease, 227–228
hybrid dysgenesis, 508
hybrid sterilization, 496
hybrid vigor, 275
hydatidiform moles, 524
HyPHY software, 329
hypothesis testing, 358–362, 370. *see also specific tests*

Icelandic populations, 537
identical by descent
altruism and, *244*
definition, 115, 258
impact of migration, 299
random genetic drift and, 257
identical by kind, 118
identical by states, 260–261
identical twins, 243, 431
identity by descent, 162–163, *374*
IM model, 306
immune systems, 486
inbred populations, 544–547
inbreeding, 257–275
definition, 87
genetic effects of, 265–268, 421
genotype frequencies and, 259–265, *262*, *263*
in human populations, 266–268
Neanderthals and humans, 138
rare recessive alleles and, *268*
reduced recombination and, 87
regular systems of mating and, 272–275
self-fertilization and, 87
inbreeding coefficient
altruism and, 244, *244*
calculation of, 268–272, *269*, *270*, *271*, *272*, *273*
definition, 258–259
fixation index and, 293–294
inbreeding depression, 266
inbreeding effective size, 121
incomplete penetrance, 437
indels (insertion/deletion) polymorphisms, 523
index of dispersion, 493
individual selection, 397–400
individuals, selection and, 200
industrial melanism, 64, 301–302
Inferring Phylogenies (Felsenstein), 131
infinite-alleles model, 162–172, *168*, *169*, 176–178, 320, *360*
infinite-series model, 172–182, *174*, 175
infinite-sites model, 35, 136
infinitesimal model, 455
influenza virus, *335*
insecticides, 215. *see also* pesticides
insects

inbreeding, 257
mariner-like elements, 512
insertion/deletion (indel) polymorphisms, 523
insertion sequences, 502–505
insertions, 396
instantaneous rate matrix, 344
insulin, 335, 347
intensity of selection, 441
interaction variance, 424
interdeme selection, 245–247, *247*
γ-interferon, 333, 345
interleukin 3, 486
internal branches, 178–180
International HapMap Consortium 2005, 61, 62, 521
International HapMap Project. *see* HapMap Project
interspecific divergence, 554–555
interval mapping, 458
intestinal lactase gene, 549
introgression, 306, 496
introns
folding, 356–357
polymorphisms in, 32
size, 185
inversions, 15
IQ scores, 4
island model of migration, 297–299, *297*, *300*
isochores, composition, *471*
isolate breaking. *see* Wahlund principle
isolation, migration and, 306
isolation by distance, 281, 285–286

Japanese populations, 431, 522
jockey transposable element, 509, *510*
Jukes-Cantor model, 326–328

Kimura, Motoo, 318
kin selection, 243–245, 247
Kingman coalescent, 131
Kolmogorov backward equation, 110–111, *111*
Kolmogorov forward equation, 107–108, *107*

large-X effect, 496
lateral gene transfer, 512–513. *see also* horizontal transmission
least squares method, 368
leptokurtic distributions, 450–451
leptokurtosis, 451–452, *451*
liability, towards traits, 387, 437, *438*
life histories, *47*
likehood ratios, 459
likelihood analyses, 142–143, 432–433
Linanthus parryae, 278–281, *279*, 285–286, *286*
lineage sorting, 529, *529*
LINEs (long interspersed elements), 498, 508, *509*
linkage
epistasis and, 239–240
linkage disequilibrium and, 73–84
mapping methods and, 550–552
population substructures and, *85*
quantitative trait loci, 386
linkage disequilibrium, 74–75
in admixed populations, 544
causes of, 84–87
correlated responses and, 440
decay of, *543*
disappearance of, *79*
genetic distance and, 483–484, *484*
human genome, 532–539
human populations, 542–543
local, 521
local distortions in, 556–557
mapping, 143–145, 552–553
measurement of, 80–81
measures of, 83–84
negative, 183, 185
observation of, 79
parameter, 78
population admixtures and, 85–86
population growth and, 482
population shrinkage and, 482
random genetic drift and, 143–144, *144*
reduced recombination and, 86–87
selection and, 481
self-fertilization and, 274
sexual selection and, 242
linkage equilibrium, 74–75, 77, 82–83, 274
linkage phase, 524
lipoprotein lipase gene, *535*
local populations, 46
local recombination rates, 354–356, *355*
local stability, 216, 220–222
locally stable equilibrium, 216
loss-of-function mutations, 376
LTRs (long terminal repeats), 498

macromolecules
electrophoresis of, 16
folding, 356–357
maize genome. *see also* corn (*Zea mays*)
haplotype blocks in, 87
linkage disequilibrium and, 483–484
mtDNA, 364
transposable elements, 501

major histocompatibility loci, 23, 371
malaria. *see also Plasmodium falciparum*
β-thalassemia and, 377
G6PD deficiency and, 549
in sickle-cell anemia, 218
sickle-cell anemia and, *219*, 224–225
male reproduction, 493–495, 496
Malthus, Thomas, 206
malthusian fitness, 206
marbled lungfish (*Protopterus aethiopicus*), 515
marginal fitness, 212–214
marginal overdominance, 232–233
mariner-like elements (*MLEs*), 511, *513*
mariner transposable element, 511–513, *511*
marker-assisted selection, 459–461, *460*
Markov Chain Monte Carlo (MCMC) algorithm, 143, 288, 538, 541
Markov chains, description, 102
mate preference, 242
mating systems, 272–275
maximum likelihood (ML) method, 432–433
maximum parsimony method, 369, *560*
McDonald-Kreitman tests, 351–354, 555–556
mean population fitness, 212
means, of populations, 6–8
means (arithmetic), 6, 7
MEGA software, 329
meiosis
cross-over events, 552
male, 537
recombination and, 183
meiotic drive, 206, 236–238, *238*, 496–497
meiotic pairing, 372
melanocortin-1 receptor, 308
Mendel, Gregor, 4, 5
Mendelian inheritance, 54
Mendelian populations, 46
Mendelian variation, discrete, 10–12
merosoite surface protein-2 gene, 349
methacillin-resistant *Staphylococcus aureus* (MRSA), 203–204
metric traits, 449. *see also* continuous traits
Metropolis–Hastings Markov chain Monte Carlo method, 305, 307
midparent, 394, 428
migrant organisms per generation (*Nm*), *302*
MIGRATE, 305
migration, 295–309
definition, 295
environmental gradients and, 233
genetic divergence and, 299–303
human, 519, 559
interdeme selection and, 246
island model of, 297–299, *297*
isolation and, 306
mutation and, 299–300
one-way, 295–297
selection balance and, 307–309
migration rates, 295, 303–307
migration-selection balance, 307–309
minimum evolution trees, 369
mismatch repair, 473
mitochondrial DNA
animal, 189–191
effective population size for, 125
evolution, 362–366
hitchhiking by, 366
human, 559
maximum parsimony tree, *560*
Neanderthal, 138
restriction sites, *333*
selection in, 365–366
transmission, 305
variation in, 364–365
MMLV virus, 382
modern synthesis, 54
molecular clocks, 329–340
dispersion index, 338–339
generation times and, 337–338
rates of, 332–337, *332*
molecular domestication, 500–501
molecular evolution
faster-male, 493–495
neutral theory and, 318–321
neutral theory of, 161–162
objectives, 469
X chromosome genes, 495
molecular phylogenetics, 366–372
molecular population genetics, 15–24, 317–383
molecular sequence divergence, 321–329
molecular weights, 17–18, *17*
Mongoloid groups, 38
monkeys, olfactory receptors, 378
monomorphism, 18, *18*
monozygotic twins, 431
Moran model, 104–105
morphs, 64
mosquitoes, 215
most recent common ancestors, 133, *367*
mountain grasshopper (*Podisma pedestris*), 381
Muller, H. J., 228
Muller's ratchet, 185–187, *186*
multi-factor inheritance, 12–14
multigene families, 372–378
birth-and-death process, 376–378, *377*
identity by descent in, *374*

subfunctionalization, 375–376, *376*
multiple drug resistance, 505
mutation-accumulation experiments, 448–449
mutation bias, 471, 472
mutation-drift equilibrium, 447–448, 448
mutation load, 228–229
mutation pressure, 152–154, *153*, 156–157
mutation rates
in a bacterial chemostat, *155*
coalescence and, 139
deleterious, 186
female, 531
generational change and, 152
Haldane-Muller principle, 228–229
infinite-series model and, 173
male, 531
mutation accumulation experiments and, 450
neutral, 320
nucleotide sequence length and, 172
population size and, 22, 185
recombination rates and, 355
mutation-selection balance, 226–229, 365, 449–452, 548
mutational diversity, 544–547
mutational variance, 448–449
mutations
additive genetic variance, 448
bidirectional, 248
coalescent models with, 136–137
definition, 151
elimination of, 477
fixation of, 160–161, 477
fixation rates, 359
genetic variation and, 15
infinite-alleles model, *360*
inhomogeneities in process of, 530
irreversible, 153–155
Jukes-Cantor model, 326–328
loss-of-function, 376
median persistence time, 448
migration and, 299–300
mtDNA, 365–366
in neutral amino acids, 165
neutral theory and, 318–321
neutrality hypothesis, 161
pleiotropic, 450
pleiotropic effects, 229
population growth and, 528
random genetic drift and, 158–161, *447*
recombination and, 182–191
recurrent, 20, 229
reversible, 156–158
selection and, 396–397
myoglobins, 23
Mytilus edulis (mussel), 365
mytochondrial myopathies, 365

National Institute of Environmental Health and Sciences (NIEHS), *528*
National Institutes of Health (NIH), 523
natural selection. *see also* selection
additive genetic variance and, 429
allele presentation and, 224
Darwinian concept of, 199
fundamental theorem of, 212
genetic variation and, 14–15
in human lineage, 554
interference with, 349
mechanisms of, 182
noncoding DNA and, 354
nonsynonymous sites, 341–342
norm of reaction and, 435–436
phenotypic variation and, 386
substitution rates and, 338
uniparental transmission and, 364
4*Nc* values, 538, 539
Neanderthals, 138, 561–562, 562
nearly neutral theory, 248–251, 339–340
negative selection, 339
neighbor-joining method, *368*
Neisseria meningitis, 189
Neoceratodus forsteri (Australian lungfish), 470
net selection gradient, 445
neutral amino acids, 165
neutral mutation rates, 320, 351
neutral theory
description, 15
molecular clocks and, 338–339
of molecular evolution, 161–162
molecular evolution and, 318–321
principles of, 318–321
neutrality hypothesis, 161
neutrally stable equilibrium, 216
Nm estimates, 301–303
non-Mendelian segregation, 236–238
noncoding sequences, 334, 354
nonoverlapping generations model, 47–48, *47*
nonrandom mating, 217–218
nonrecombinant gametes, 75–82
nonsynonymous divergence, 484–485
nonsynonymous nucleotide substitutions, 340–347, *345*
nonsynonymous polymorphisms, 32–34
normal distributions
central limit theorem and, 9, *10*
continuous variation and, 5–6, *6*

normal probability density function, 6
normalizing transformation, 411
norms of reaction, 433–437
Norway rats, 216
Nubia, migrations, 305–307, *306*
nuclear genomes, 365
nucleic acids
electrophoresis of, 16, *17*
probes, *26*
nucleotide diversity
definition, 347, 480
nucleotide polymorphism and, 175–176
in protein-coding genes, *521*
nucleotide mismatches
average number of, 36
calculation of, 34–35
infinite-series model and, 173, *174*
simulations and, 142
nucleotide polymorphism, 175–176
nucleotide sequence divergence, 351–354
nucleotide substitution
in the mammalian genome, 530
models, 329
nonsynonymous, 340–347
overdispersed clocks and, 338
patterns of, 340–351
rates, 326–328, *327*
silent, 340–347
synonymous, 340–347
null hypotheses, rejection of, 62

offspring
random genetic drift and, 126–127
variance in number of, 126–127
Ohta, Tomoko, 249, 339
olfactory receptors (ORs), 378
one-egg twins, 431
one-way migration, 295, *295*
opsin proteins, 492
organelle genes, 123–125
organophosphates, 215
The Origin of Species (Darwin), 199
out-of-Africa hypothesis, 561
outcrossing
artificial selection and, 395
definition, 87, 265
linkage disequilibrium and, 87
overdominance, 216–220, 232–233, *424*. *see also* stabilizing selection

P-elements, 396–397, *451*
p53 protein, 54–55
PAML software, 329
panmictic index, 273
paralogous copies, 376
parameters
definition, 6
of distributions, 389–390
true values of, 8
parasites, gene elimination, 376
parent-offspring covariance, *427*
parent-offspring regression, 393
parsimony analyses, 369, 529, *560*
Passer domesticus (house sparrows), 443
pathogenic deviations, 489–491
PBDX gene, 72
pedigree studies, 537
penetrance, incomplete, 437
peppered moth (*Biston betularia*), 64, 65, 300, *301*, 302
Perlegen Biosciences, *528*
pesticides, resistance to, 215, 356, *357*, 501
PfEMP1 protein, 474
Phaseolus vulgaris (bean), 398, *398*
phenotypes
age and, 436–437
artificial selection and, 413–415
definition, 3
diabetes, 437
diabetes as continuous trait, 437
environment and, 16, *419*
evolutionary biology and, 385
natural selection and, 182
population averages, 432
segregation of alleles and, *13*, *14*
selection and, 200
variation in natural populations, 5–12
phenotypic covariance, 426–433
phenotypic plasticity, 433–437
phenotypic variance, 416–426, *417*
phenylalanine hydroxylase (*PAH*) gene, 547, *548*
phenylketonuria, 66, 547, *548*
Phlox cuspidata, 70
phoA gene, 188, *188*
6-phosphogluconate dehydrogenase coding, 381
phylogenetic trees
Bayesian methods, 370
bootstrapping, 370
construction, 368–369
neighbor-joining method, 368, *368*
statistical confidence, 370
phylogenetics, molecular, 366–372
phyloHMM tool, 555
plants, organelle DNA, *364*

plasmids, assembly of, 505
Plasmodium falciparum. see also malaria
  A-T base pairs, 380
  polymorphisms in, 349, 474
  sickle-cell anemia-prone populations, 218, *219*
pleiotropic effects, 439–442
pocket gophers (*Geomys pinetis*), 189–190, *190*
Podisma pedestris (mountain grasshopper), 381
Poisson random field tests, 555–556
polymerase chain reactions (PCRs), 28–31, *29*
polymorphic genes, 20
polymorphisms
  allozymes, 20–21
  anonymous, 519
  balanced, 216
  definition, 321
  divergence and, 487–489
  in DNA sequences, 25–38
  electrophoretic separations, 18, *18*
  estimation of, 23–24
  genes, 20–21
  genome-wide patterns, 473–484
  heterozygosity and, *21*, 22–23
  human, 287–288, 520–531
  insertion/deletion, 523
  intraspecific, 351
  L. parryae subpopulations, 285–286
  in noncoding sequences, 354
  nonsynonymous, 32–34
  nucleotide, 175–176
  nucleotide divergence and, 351–354
  in p53, 54
  prediction of damage, 354
  random genetic drift and, 318
  recombination rates and, 475–477, *476*
  self-fertilization and, 274
  sickle-cell anemia, 218
  within species, 347–349
  in subtelomeric regions, 473–475
  synonymous, 32–34
  trans-species, 371–372, *372*
  uniparental transmission and, 364
  utility of, 38–39
polypeptide chains, 18
population admixtures, 482
population genetics
  focus of, 3
  human, 519–565
  Mendelian disease and, 547–550
  molecular, 15–24, 317–383
  relevance of, 4–5
population genomics, 468–517, 470
population screening, 4
population size. *see also* effective population size
  bottlenecks and, 122–123
  density-dependent selection and, 231
  fluctuation in, 121–123
  haplotype structure and, 482
  heterozygosity and, *321*
  linkage disequilibrium and, *144*
  mutation rates and, 22, 185
  neutral mutation rate and, 339–340
  random genetic drift and, 246
  recombination and, *184*, 241
  selection and, 248–251
  variation and, 251
populations
  admixed, 544, *545*
  age-structured, 232
  definition of, 45–46
  distribution parameters, 389
  growth, 134–136, *201*, 231, 527–529
  hierarchical structures, 276
  impact of migration on, 295–309
  inbred, 544–547
  mean fitness of, 209
  mean of, 6–8
  sampling, 6
  stratification, 540–542
  structures, 275–288, 278–280, *280*, 526
  subdivisions, 113–121, 245–247, 275–288
  with submaximal fitness peak, 225–226
  substructures of, 85–86, *85*
  variance within, 6–8
positive selection, 486, 495
posterior distributions, 143
predation, 64–65
prediction equations, 399
preproinsulin, 347
primers, adapters, 31
Princess bean (*Psophocarpus tetragonolobus*), 396
Principal Components Analysis (PCA), 542
private alleles, 302–303
PRO119Ala replacement, 491
probes, nucleotides in, 26
progeny. *see* offspring
prokaryotes, recombination in, 187–189
proteins
  amino acid replacements, 322–326
  electrophoresis of, 16, *17*
  folding, 490–491
  mammalian genome coding for, 162
  molecular clocks, *332*
  monomorphism, 18, *18*
  nucleotide substitution rates, 347

polymorphism, 18, *18*
sequencing, 4
stability of, 490–491
Protopterus aethiopicus (marbled lungfish), 515
pseudogenes, 346–347
Psophocarpus tetragonolobus (Princess bean), 396
Punnet square, *49*, 50
purifying selection, 354, *360*, 484–485, 530, 557
pyloric stenosis, 438

quantitative genetics
evolutionary, 385–466, 443–452
social applicability of, 430
quantitative trait loci (QTLs), 386
description, 453
genotype-by-sex interactions and, 419–420
mapping, 455–461, *456*
mutation rates and, 396
significance testing of, 459
quantitative trait nucleotides (QTNs), 453, 551
quantitative traits
candidate genes, 452–461
description, 385
genetic models, 403–416
mutation-selection balance for, 449–452
normal distribution of, *407*, *408*
phenotypic distribution, *421*
types of, 387–388

RAG1 gene, 501
RAG2 gene, 501
random cohort designs, 550
random genetic drift
adaptive topography, 225–226
allele frequency and, 48, 286
among subpopulations, *114*
binomial sampling and, 95–101
coalescence and, 128–145
diffusion approximation, 105–113
effective population size and, *118*, 121–128
gene trees, 128–138
mutations and, 158–161, *447*
in organelle genomes, 365
phenotypic evolution and, 446–448
polymorphisms and, 318
population size and, 246
in subdivided populations, 113–121, 245–246
theoretical results, *108*
within-deme phase, 127–128
Wright-Fisher model of, 102–105
random mating
average fitness and, 217
definition, 46
genetic variation and, 46–48
heterozygosity and, 281–285
in isolated subpopulations, 277
parent-offspring covariance and, 427
X-linked recessive alleles, *73*
randomly amplified polymorphic DNA (RAPD), 329
realized heritability, 399
recessive alleles, 288–294, 289
reciprocals, sums of, *36*
recombinant gametes, 75–82
recombination
benefits of, 241
coalescent models with, 140–143
conjunction, 187
DNA repair and, 182–183
euchromatin and, 511
evolutionary benefit of, *184*
gene conversion and, 536–537
hot spots of, *538*
mitochondrial DNA and, 189–191
mutation and, 182–191
phoA gene in, *188*
piecewise, 187–189
selection and, 239
in single nucleotide polymorphisms, *141*
transduction, 187
transformation in, 187
recombination rates
evolution of, 241
linkage disequilibrium and, 532
local, 354–356, *355*, 537–539
polymorphism and, 475–477, *476*
variation and, 356
red-green colorblindness, 377–378
red opsin gene (*OPN1LW*), 378, 492
regression
concept of, 390
graphical interpretation, 391, *392*
toward the mean, 183
regression coefficients, 391
regression lines, 390–391
regular systems of mating, 272–275
relative fitness, 207, 226
relative rate tests, 336–337
relatives
covariance among, 426–433
phenotypic covariance, *428*
relaxed selection, 400
reproduction systems, 182, 185, 272–275, 493–496
reproductive fitness, 440

resistance
to drugs, 505
to insecticides, 215
to pesticides, 215, 356, *357*, 501
to warfarin, 220
resistance transfer factors, 505
response to selection, 399
restricted maximum likelihood (REML), 432–433
restriction enzymes. *see also* enzymes
definition, 25
function, 25–28, *25*
mtDNA, 363
restriction fragment length polymorphisms (RFLPs), 28, *28*, 36–37
restriction fragment patterns, 329
restriction sites, 25
retrotransposons, 498, 505
reverse selection, 400
reverse transcriptases, 498
RhCE genes, 63–64
RhD genes, 63–64
Rhesus (Rh) blood groups, 63
rifin gene, 474
RNA
compensatory changes, 490
mitochondrial, 363
5S, 366
28S, 500
secondary structures, 350
substitution models for, 356–358
transfer, 349–350
RNA replicase, 335
rock-pocket mice (*Chaetodipus intermedius*), 308
RpL32 gene, 481, *482*, 483
runaway sexual selection, 242

Saccharomyces cerevisiae, 433–434, 474
Salmonella enterica, 187
sample sizes, 58
samples, 6
scaling effects, 411–412
scalloped hazel moth (*Gonodontis bidentata*), 64, 300, *301*, 302
scattergrams, 389
Schmalhausen, Ivan, 433
Seattle SNP project, *521*
segregating sites, 139, 173
segregation
of allels, *13*, *14*
definition, 5
non-Mendelian, 236–238
phenotypic variation and, 5
segregation distortion, 206, 236
selection
amino acid substitutions and, 346
background, 356
Darwinian, 199–255
density-dependent, 230
differential, 229–230
in diploid organisms, 206–215
directional, 449
diversifying, 234–235
equilibrium values with, 215–226
evidence for, 285–286
fecundity, 206, 231–232
in finite populations, 248–251
frequency-dependent, 230, *360*
gametic, 206, 236
genome scans for, 557–558
heterozygote inferiority and, 222–223, *223*
individual, 397–400
intensity of, 441
interdeme, 245–247
kin, 243–245
limits to, 400–403
marginal fitness and, 212–214
marker-assisted, 459–461
migration-selection balance, 307–309
in mtDNA, 365–366
mutation-selection balance, 226–229
mutations and, 396–397
overdominance and, 216–220, *217*
phenotypic, 443–444
phenotypic divergence and, 446
positive, 486
purifying, 354, *360*
relaxed, 400
response to, 399, *402*
reverse, 400
sexual, 206, 241–243
sexually antagonistic, 497
stabilizing, *360*, 449
strength of, 556
truncation, 397–398, *397*
for viability, *209*
selection coefficients, 226, 309, *309*
selection differential, 399
selection gradients, 445
selection indexes, 440
selection-mutation balance model, 451
selection plateaus, 400–403
selective constraints, 346
selective sweeps, 356, 477, 556
self-fertilization, 87, *259*, 273–274
self-incompatibility, 235, 371
self-pollinization, 235

self-sterility, 235
selfish DNA, 501
sensory perception, 486
sequential importance sampling, 143
sex chromosomes, 123–125, 492–493
sex-ratio segregation, 496
sex ratios, 123–125, *124*
sexes, differential selection, 229–230
sexual conflict, 242–243
sexual reproduction, 182, 185
sexual selection, 206, 241–243, 492–493
sexually antagonistic selection, 497
shared polymorphisms, 371–372
shifting balance theory, 245–246
sickle-cell anemia, 218, *219*, 224–225, 289
silent nucleotide substitutions, 340–347
simple Mendelian variation, 10–12, *11*
simulations, ancestral histories and, 142
SINEs (short interspersed elements), 498, 508, *509*
single nucleotide polymorphisms (SNPs)
  anonymous, 532
  ascertainment bias of genotypes, 524–527, *525*
  association mapping, *553*
  coalescence, *141*
  density of, *522*
  description, 32
  disease risk and, 541
  flanking, 542
  genetic diversity and, 266
  human, 521
  in the human genome, 61–62
  linkage disequilibrium and, 79, 145
  mapping, 453–454
  proximity, 532, *533*
  public resources, 521–524
  recombination in, *141*
  saturation, 552
site frequency spectra, 525, 527–529, *528*, 557–558, *558*
snapdragon (*Antirrhinum majus*), 11, *11*, 89
software, 329
Southern blot procedure, *27*
speciation, 372, 495–497, 513
"sperm killers," 236
spermatogenesis, 486
spore killers, 236
sporophytes, 236
stabilizing selection, *360*, 449, 450, 451, *451*
standard deviation, 7, 8, *10*
standard error, 8
Staphylococcus aureus, 203, *204*
star phylogenies, 134
statistical significance, 62, 82–83
steady state, 166
stevor gene, 474
Streptococcus aureus, 189
Streptococcus pneumoniae, 189
Streptococcus pyogenes, 189
stress conditions, 434
STRUCTURE program, 541
structured coalescent, 549
subfunctionalization, 375–376
subpopulations
  F-statistic in, 283
  heterozygosity in, *287*
  human, 286–288, 289, 305
  interdeme selection, 245–247
  isolation, 276–278, *277*
substitution models, 555
substitution rates, *250*, 333, 362
subtelomeric regions, 473–475
Sudan, migrations, 305–307, *306*
sulfur dioxide, atmospheric, 65
superoxide dismutase, 137, 361, 556
supertrees, 370
supreme law of unreason, 14
survival, natural selection and, 199–200
"survival of the fittest" concept, 200
symbionts, gene elimination, 376
synonymous divergence, 484–485
synonymous nucleotide substitutions, 340–347, *345*
synonymous polymorphisms, 32–34
Szent-Gyorgyi, Albert, 16

T-cell receptor beta protein, *546*
Tajima's *D* statistic, 177–178, 362
TART array, 500
Tay-Sachs disease, 123, 289
Taylor series expansion, 221
telomeres, 473
β-thalassemia, 377
threshold traits. *see also* dichotomous traits
  concordance in, 439, *439*
  disease and, 437–439
  liability towards, *438*
traits
  addititive genetic variance, 415
  age and, 436–437
  categorical, 387
  complex, 385–386
  continuous, 8–10, 387
  continuously varying, 385
  genetic models, 403–416
  intercorrelated, 444–445

liability towards, 387
Mendelian, 5
metric, 387
multifactorial inheritance of, 12–14, *13*, 385
quantitative, 385
response to selection, 402–403, *402*
trans-species polymorphisms, 371–372, *372*
transduction, 187
transformation, 187
Transib elements, 501
transition probability, 102
transitions, 328
transmission disequilibrium test (TDT), 551
transposable elements
in bacteria, 502–505
classes of, *499*
definition, 151, 396
in eukaryotes, 505
horizontal transmission of, 511–513
jockey, 509, 510, *510*
mariner, 511–513, *511*
population dynamics of, 501, 505–506
types of, 498–501
transposase proteins, 499–500, 500
transposition, 497, 502–505
transposition rates, 506–511
transposons. *see* composite transposons; DNA transposons; retrotransposons; transposable elements
transversions, 328
Tribolium castaneum (flour beetles), *389*, 402
Trichoplax adhaerence (placozoan), 515
Triticum vulgare (wheat), 405
truncation points, 398
truncation selection, 397–398, *397*, 408–409, *413*
tumor suppression, 486
turkeys, 440
turnip (*Brassica campestris*), 364
twin studies, 430–432, 439, *439*
two-egg twins, 431
twofold degenerate sites, 340

unbiased estimators, 390
unequal crossing-over, 372, 473–474, *474*
United States subpopulations, 522
uniting gametes, 263–264, *264*
unstable equilibrium, 216
UPGMA (unweighted pair-group method with arithmetic mean) method, 366–367

var genes, 474
variance
age and, 436–437
definition (mathematical), 6
distribution and, 7
dominance, *424*
environmental, 420
epistatic, 424
genetic, *424*
genotypic, 420
interaction, 424
phenotypic, 416–426
of populations, 6–8, 389
variance effective size, 121
variation
continuous, 5
discrete, 5
genotypic, 422–426
phenotypic, 5–12
viability
expression of, 207
selection for, 206, *209*, *238*
VISNA virus, 382
VNTR loci, 329

Wahlund principle, 288–294, *289*
Walhlund, Sten Gösta William, 288
warfarin resistance, 220
weak selection, 248–251
Weinberg, Wilhelm, 48
wheat (*Triticum vulgare*), 405
white Leghorn chickens, 425, 440
wildlife refuges, 4
Wolbachia infection, 366
Wright-Fisher model, *98*, 102–105, *103*
Wright's F statistic, 281–285

X chromosomes, 495, 497–498, 531
X linkage, 70–73, *71*
X-linked genes, 230, 350, 495
xanthine dehydrogenase (*Xdh*) gene, 24, 70, 170, 172

Y chromosomes
diversity in, 480
mutation rates, 531
polymorphisms, 475, *475*
recombination rates in, 185
Yanomama Indians, 287
Yoruban populations, 522, 542

Zea mays (corn), *395*, 396
Zoarces viviparus (eelpout fish), 233
zoo mating programs, 4