

hyperref hyperref

GRAHAM COOP

POPULATION AND
QUANTITATIVE
GENETICS

Author: Graham Coop

Author address: Department of Evolution and Ecology & Center for Population Biology,
University of California, Davis.

To whom correspondence should be addressed: gmccoop@ucdavis.edu

This work is licensed under a Creative Commons Attribution 3.0 Unported License.

<http://creativecommons.org/licenses/by/3.0/>

i.e. you are free to reuse and remix this work, but please include an attribution to the original.

Typeset using L^AT_EX and the LetterSpace=10TUFTE-LATEX book style.

The L^AT_EX code and R code for this book are kept here <https://github.com/cooplab/popgen-notes/> and again are under a Creative Commons Attribution 3.0 Unported License.

Updated on August 2019

This book was developed from my set of notes for the Population Biology graduate group core class (PBG200A) and Undergraduate Population and Quantitative Genetics class (EVE102) at UC Davis. Thanks to the many students who've read these notes and suggested improvements. Thanks to Simon Aeschbacher, Vince Buffalo, and Erin Calfee who read and extensively edited earlier drafts of these notes. To illustrate these notes I've used old scientific and natural history illustrations, in part because they are out of copyright but mainly because they bring me joy. Many of the old images come from Biodiversity Heritage Library a consortium of natural history institutions that are digitizing their collections and make them freely available online. If you enjoy the images consider donating to the BHL. Many of the data and simulation graphics in the book were prepared in LetterSpace=10R (2018), the code for each is linked to from the caption of each figure. In many cases data were extracted from old figures using the WebPlotDigitizer tool, as such I advise re-extracting the data if you wish to use it for research purposes.

Contents

1	<i>Introduction</i>	9
2	<i>Allele and Genotype Frequencies</i>	13
3	<i>Genetic Drift and Neutral Diversity</i>	49
4	<i>Phenotypic Variation and the Resemblance Between Relatives</i>	95
5	<i>The Response to Phenotypic Selection</i>	115
6	<i>One-Locus Models of Selection</i>	137
7	<i>The interaction of Selection, Mutation, and Migration.</i>	173
8	<i>The Impact of Genetic Drift on Selected Alleles</i>	189
9	<i>The Effects of Linked Selection.</i>	201
10	<i>Interaction of Multiple Selected Loci</i>	219
11	<i>Bibliography</i>	233

1

2 Introduction

LetterSpace=10BIOLOGICAL EVOLUTION IS THE CHANGE OVER TIME IN

4 THE GENETIC COMPOSITION OF A POPULATION.¹ Our population is
made up of a set of interbreeding individuals, the genetic composition
6 of which is made up of the genomes that each individual carries.
The genetic composition of the population alters due to the death
8 of individuals or the migration of individuals in or out of the population.
If our individuals vary in the number of children they have,
10 this also alters the genetic composition of the population in the next
generation. Every new individual born into the population subtly
12 changes the genetic composition of the population. Their genome is
a unique combination of their parents' genomes, having been shuf-
14 fled by segregation and recombination during meioses, and possibly
changed by mutation. These individual events seem minor at the
16 level of the population, but it is the accumulation of small changes
in aggregate across individuals and generations that is the stuff of
18 evolution. It is the compounding of these small changes over tens,
hundreds, and millions of generations that drives the amazing diver-
20 sity of life that has emerged on this earth.

Population genetics is the study of the genetic composition of
22 natural populations and its evolutionary causes and consequences.
Quantitative genetics is the study of the genetic basis of phenotypic
24 variation and how phenotypic changes evolve over time. Both fields
are closely conceptually aligned as we'll see throughout these notes.
26 They seek to describe how the genetic and phenotypic composition
of populations can be changed over time by the forces of mutation,
28 recombination, selection, migration, and genetic drift. To understand
how these forces interact, it is helpful to develop simple theoretical
30 models to help our intuition. In these notes we will work through
these models and summarize the major areas of population- and
32 quantitative-genetic theory.

While the models we will develop will seem naïve, and indeed

¹ LetterSpace=10DOBZHANSKY, T., 1951
Genetics and the Origin of Species (3rd Ed.
ed.), pp. 16

"All models are wrong but some are useful" - LetterSpace=10Box (1979).

³⁴ they are, they are nonetheless incredibly useful and powerful. Through-
 out the course we will see that these simple models often yield accu-
³⁶ rate predictions, such that much of our understanding of the process
 of evolution is built on these models. We will also see how these
³⁸ models are incredibly useful for understanding real patterns we see
 in the evolution of phenotypes and genomes, such that much of our
⁴⁰ analysis of evolution, in a range of areas from human medical genetics
 to conservation, is based on these models. Therefore, population
⁴² and quantitative genetics are key to understanding various applied
 questions, from how medical genetics identifies the genes involved in
⁴⁴ disease to how we preserve species from extinction.

Population genetics emerged from early efforts to reconcile Mendelian genetics with Darwinian thought. Part of the power of population genetics comes from the fact that the basic rules of transmission genetics are simple and nearly universal. One of the truly remarkable things about population genetics is that many of the important ideas and mathematical models emerged before the 1940s, long before the mechanistic-basis of inheritance (DNA) was discovered, and yet the usefulness of these models has not diminished. This is a testament to the fact that the models are established on a very solid foundation, building from the basic rules of genetic transmission combined with simple mathematical and statistical models.

Much of this early work traces to the ideas of R.A. Fisher, Sewall Wright, and J.B.S. Haldane, who, along with many others, described the early principals and mathematical models underlying our understanding of the evolution of populations. Building on this conceptual fusion of genetics and evolution, there followed a flourishing of evolutionary thought, the modern evolutionary synthesis, combining these ideas with those from the study of speciation, biodiversity, and paleontology. In total this work showed that both short-term evolutionary change and the long-term evolution of biodiversity could be well understood through the gradual accumulation of evolutionary change within and among populations. This evolutionary synthesis continues to this day, combining new insights from genomics, phylogenetics, ecology, and developmental biology.

Population and quantitative genetics are a necessary but not sufficient description of evolution; it is only by combining the insights of many fields that a rich and comprehensive picture of evolution emerges. We certainly do not need to know the genes underlying the displays of the birds of paradise to study how the divergence of these displays, due to sexual selection, may drive speciation. Indeed, as we'll see in our discussion of quantitative genetics, we can predict how populations respond to selection, including sexual selection and assortative mating, without any knowledge of the loci involved. Nor

See LetterSpace=10PROVINE (2001) for a history of early population genetics.

LetterSpace=10PROVINE, W. B., 2001
The origins of theoretical population genetics: with a new afterword. University of Chicago Press

"LetterSpace=10DOBZHANSKY (1951) once defined evolution as 'a change in the genetic composition of the populations' an epigram that should not be mistaken for the claim that everything worth saying about evolution is contained in statements about genes"

- LetterSpace=10LEWONTIN

- ⁷⁸ do we need to know the precise selection pressures and the ordering of genetic changes to study the emergence of the tetrapod body plan.
- ⁸⁰ We do not necessarily need to know all the genetic details to appreciate the beauty of these, and many other, evolutionary case studies.
- ⁸² However, every student of biology gains from understanding the basics of population and quantitative genetics, allowing them to base
- ⁸⁴ their studies on a solid bedrock of understanding of the processes that underpin all evolutionary change.

Allele and Genotype Frequencies

88 In this chapter we will work through how the basics of Mendelian
 89 genetics play out at the population level in sexually reproducing
 90 organisms.

91 Loci and alleles are the basic currency of population genetics—and
 92 indeed of genetics. If all individuals in the population carry the
 93 same allele, we say that the locus is *monomorphic*; at this locus there is
 94 no genetic variability in the population. If there are multiple alleles in
 95 the population at a locus, we say that this locus is *polymorphic* (this is
 96 sometimes referred to as a segregating site).

97 Table 2.1 shows a small stretch of orthologous sequence for the
 98 ADH locus from samples from *Drosophila melanogaster*, *D. simulans*,
 99 and *D. yakuba*. *D. melanogaster* and *D. simulans* are sister species and
 100 *D. yakuba* is a close outgroup to the two. Each column represents a
 101 single haplotype from an individual (the individuals are diploid but
 102 were inbred so they're homozygous for their haplotype). Only sites
 103 that differ among individuals of the three species are shown. Site 834
 104 is an example of a polymorphism; some *D. simulans* individuals carry
 105 a C allele while others have a T. *Fixed differences* are sites that differ
 106 between the species but are monomorphic within the species. Site 781
 107 is an example of a fixed difference between *D. melanogaster* and the
 108 other two species.

109 We can also annotate the alleles and loci in various ways. For
 110 example, position 781 is a non-synonymous fixed difference. We
 111 call the less common allele at a polymorphism the *minor allele* and
 112 the common allele the *major allele*, e.g. at site 1068 the T allele is
 113 the minor allele in *D. melanogaster*. We call the more evolutionarily
 114 recent of the two alleles the *derived allele* and the older of the two the
 115 *ancestral allele*. We infer that the T allele at site 1068 is the derived
 116 allele because the C is found in both other species, suggesting that
 117 the T allele arose via a C → T mutation.

118 **Question 1.** A) How many segregating sites does the sample from

A *locus* (plural: *loci*) is a specific spot in the genome. A locus may be an entire gene, or a single nucleotide base pair such as A-T. At each locus, there may be multiple genetic variants segregating in the population—these different genetic variants are known as *alleles*.

pos.	con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	g	h	i	j	k	l	NS/S
781	G	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	NS	
789	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	S	
808	A	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	NS	
816	G	T	T	T	T	-	-	-	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	S	
834	T	-	-	-	-	-	-	-	-	C	C	-	-	C	-	-	-	-	-	-	-	-	-	-	S	
859	C	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	NS	
867	C	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	A	G	G	G	G	G	G	S	
870	C	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S	
950	G	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	S	
974	G	-	-	-	-	-	-	-	-	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	S	
983	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	S	
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1034	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	-	C	-	C	C	S		
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1089	C	-	-	-	-	-	-	-	-	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	NS	
1101	G	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	NS	
1127	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	S	
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1160	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	S	

D. simulans have in the ADH gene?

B) How many fixed differences are there between *D. melanogaster* and *D. yakuba*?

2.1 Allele frequencies

Allele frequencies are a central unit of population genetics analysis, but from diploid individuals we only get to observe genotype counts. Our first task then is to calculate allele frequencies from genotype counts. Consider a diploid autosomal locus segregating for two alleles (A_1 and A_2). We'll use these arbitrary labels for our alleles, merely to keep this general. Let N_{11} and N_{12} be the number of A_1A_1 homozygotes and A_1A_2 heterozygotes, respectively. Moreover, let N be the total number of diploid individuals in the population. We can then define the relative frequencies of A_1A_1 and A_1A_2 genotypes as $f_{11} = N_{11}/N$ and $f_{12} = N_{12}/N$, respectively. The frequency of allele A_1 in the population is then given by

$$p = \frac{2N_{11} + N_{12}}{2N} = f_{11} + \frac{1}{2}f_{12}. \quad (2.1)$$

Note that this follows directly from how we count alleles given individuals' genotypes, and holds independently of Hardy–Weinberg proportions and equilibrium (discussed below). The frequency of the alternate allele (A_2) is then just $q = 1 - p$.

2.1.1 Measures of genetic variability

Nucleotide diversity (π) One common measure of genetic diversity is the average number of single nucleotide differences between haplotypes chosen at random from a sample. This is called *nucleotide diversity* and is often denoted by π . For example, we can calculate π for our ADH locus from Table 2.1 above: we have 6 sequences from

Table 2.1: Variable sites in exons 2 and 3 of the ADH gene in *Drosophila* LetterSpace=10McDONALD and LetterSpace=10KREITMAN (1991). The first column (pos.) gives the position in the gene; exon 2 begins at position 778 and we've truncated the dataset at site 1175. The second column gives the consensus nucleotide (con.), i.e. the most common base at that position; individuals with nucleotides that match the consensus are marked with a dash. The first columns of sequence (a-l) are from *D. melanogaster*; the next columns (a-f) give sequences from *D. simulans*, and the final set of columns (a-l') from *D. yakuba*. The last column shows whether the difference is a non-synonymous (N) or synonymous (S) change.

¹⁴⁴ *D. simulans* (a-f), there's a total of 15 ways of pairing these sequences, and

$$\pi = \frac{1}{15} ((2+1+1+1+0) + (3+3+3+2) + (0+0+1) + (0+1) + (1)) = 1.2\bar{6} \quad (2.2)$$

¹⁴⁶ where the first bracketed term gives the pairwise differences between a and b-f, the second bracketed term the differences between b and c-f and so on.

¹⁴⁸ Our π measure will depend on the length of sequence it is calculated for. Therefore, π is usually normalized by the length of sequence, to be a per site (or per base) measure. For example, our ADH sequence covers 397bp of DNA and so $\pi = 1.2\bar{6}/397 = 0.0032$ per site in *D. simulans* for this region. Note that we could also calculate π per synonymous site (or non-synonymous). For synonymous site π , we would count up number of synonymous differences between our pairs of sequences, and then divide by the total number of sites where a synonymous change could have occurred.¹

¹⁵⁸ *Number of segregating sites.* Another measure of genetic variability is the total number of sites that are polymorphic (segregating) in our sample. One issue is that the number of segregating sites will grow as we sequence more individuals (unlike π). Later in the course, we'll talk about how to standardize the number of segregating sites for the number of individuals sequenced (see eqn (3.38)).

¹⁶⁴ *The frequency spectrum.* We also often want to compile information about the frequency of alleles across sites. We call alleles that are found once in a sample *singletons*, alleles that are found twice in a sample *doubletons*, and so on. We count up the number of loci where an allele is found i times out of n , e.g. how many singletons are there in the sample, and this is called the *frequency spectrum*. We'll want to do this in some consistent manner, such as calculating the frequency spectrum of the minor allele or the derived allele.

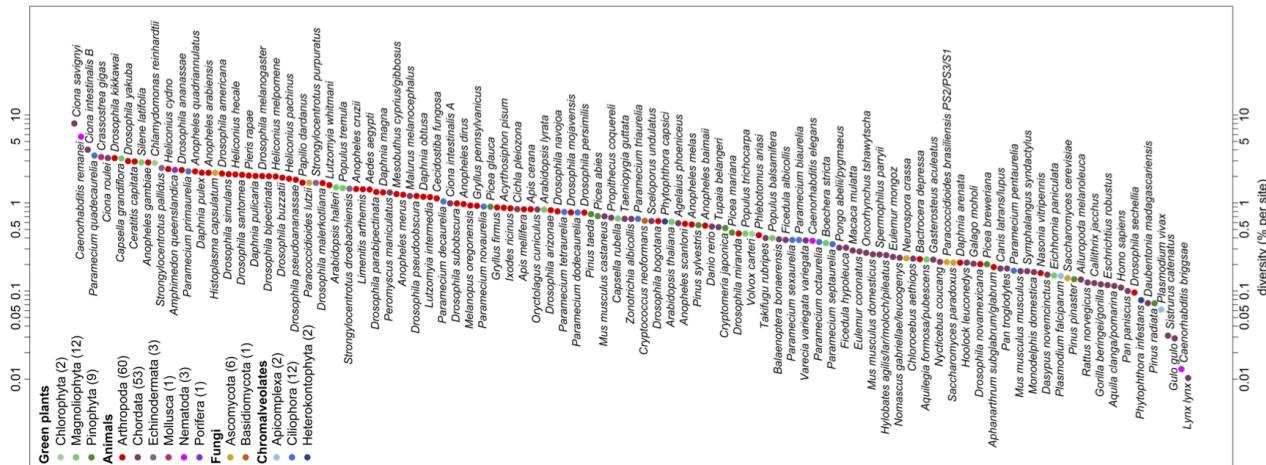
¹⁷² **Question 2.** How many minor-allele singletons are there in *D. simulans* in the ADH region?

¹⁷⁴ *Levels of genetic variability across species.* Two observations have puzzled population geneticists since the inception of molecular population genetics. The first is the relatively high level of genetic variation observed in most obligately sexual species. This first observation, in part, drove the development of the Neutral theory of molecular evolution, the idea that much of this molecular polymorphism may simply reflect a balance between genetic drift and mutation. The second observation is the relatively narrow range of polymorphism

¹ Technically we would need to divide by the total number of possible point mutations that would result in a synonymous change; this is because some mutational changes at a particular nucleotide will result in a non-synonymous or synonymous change depending on the base-pair change.

across species with vastly different census sizes. This observation represented a puzzle as the Neutral theory predicts that levels of genetic diversity should scale with population size. Much effort in theoretical and empirical population genetics has been devoted to trying to reconcile models with these various observations. We'll return to discuss these ideas throughout our course.

The first observations of molecular genetic diversity within natural populations were made from surveys of allozyme data, but we can revisit these general patterns with modern data.



For example, LetterSpace=10 LEFFLER *et al.* (2012) compiled data on levels of within-population, autosomal nucleotide diversity (π) for 167 species across 14 phyla from non-coding and synonymous sites (Figure 2.2). The species with the lowest levels of π in their survey was Lynx, with $\pi = 0.01\%$, i.e. only 1/10000 bases differed between two sequences. In contrast, some of the highest levels of diversity were found in *Ciona savignyi*, Sea Squirts, where a remarkable 1/12 bases differ between pairs of sequences. This 800-fold range of diversity seems impressive, but census population sizes have a much larger range.

2.1.2 Hardy–Weinberg proportions

Imagine a population mating at random with respect to genotypes, i.e. no inbreeding, no assortative mating, no population structure, and no sex differences in allele frequencies. The frequency of allele A_1 in the population at the time of reproduction is p . An A_1A_1 genotype is made by reaching out into our population and independently drawing two A_1 allele gametes to form a zygote. Therefore, the probability that an individual is an A_1A_1 homozygote is p^2 . This proba-

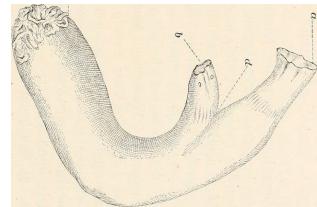


Figure 2.1: Sea Squirt (*Ciona intestinalis*). Einleitung in die vergleichende gehirnphysiologie und Vergleichende psychologie. Loeb, J. 1899. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. No known copyright restrictions.

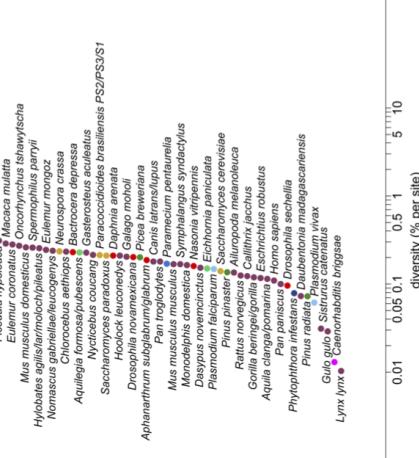


Figure 2.2: Levels of autosomal nucleotide diversity for 167 species across 14 phyla. Figure 1 from LetterSpace=10 LEFFLER *et al.* (2012), licensed under CC BY 4.0. Points are ranked by their π , and coloured by their phylum. Note the log-scale.

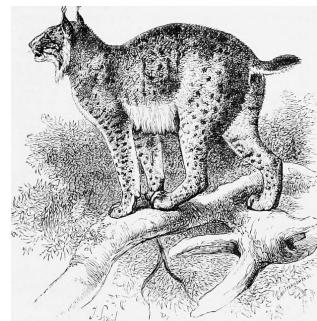


Figure 2.3: Eurasian Lynx (*Lynx lynx*). An introduction to the study of mammals living and extinct. Flower, W.H. and Lydekker, R. 1891. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. No known copyright restrictions.

bility is also the expected frequencies of the A_1A_1 homozygote in the population. The expected frequency of the three possible genotypes are

$$\frac{f_{11}}{p^2} \quad \frac{f_{12}}{2pq} \quad \frac{f_{22}}{q^2}$$

Note that we only need to assume random mating with respect to our focal allele in order for these expected frequencies to hold in the zygotes forming the next generation. Evolutionary forces, such as selection, change allele frequencies within generations, but do not change this expectation for new zygotes, as long as p is the frequency of the A_1 allele in the population at the time when gametes fuse.

Question 3. On the coastal islands of British Columbia there is a subspecies of black bear (*Ursus americanus kermodei*, Kermode's bear). Many members of this black bear subspecies are white; they're sometimes called spirit bears. These bears aren't hybrids with polar bears, nor are they albinos. They are homozygotes for a recessive change at the MC1R gene. Individuals who are *GG* at this SNP are white while *AA* and *AG* individuals are black.

Below are the genotype counts for the MC1R polymorphism in a sample of bears from British Columbia's island populations from LetterSpace=10RITLAND *et al.* (2001).

<i>AA</i>	<i>AG</i>	<i>GG</i>
42	24	21

What are the expected frequencies of the three genotypes under HWE?

See Figure 2.5 for a nice empirical demonstration of Hardy-Weinberg proportions. The mean frequency of each genotype closely matches its HW expectations, and much of the scatter of the dots around the expected line is due to our small sample size (~ 60 individuals). While HW often seems like a silly model, it often holds remarkably well within populations. This is because individuals don't mate at random, but they do mate at random with respect to their genotype at most of the loci in the genome.

Question 4. You are investigating a locus with three alleles, A, B, and C, with allele frequencies p_A , p_B , and p_C . What fraction of the population is expected to be homozygotes under Hardy-Weinberg?

Microsatellites are regions of the genome where individuals vary for the number of copies of some short DNA repeat that they carry. These regions are often highly variable across individuals, making them a suitable way to identify individuals from a DNA sample.

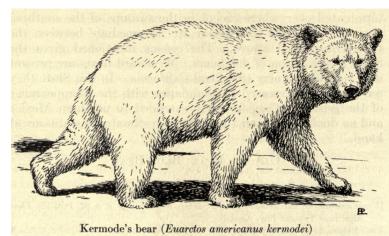


Figure 2.4: Kermode's bear.
Extinct and vanishing mammals of the western hemisphere.
1942. Glover A. Image from the Biodiversity Heritage Library.
Contributed by Prelinger Library. Not in copyright.

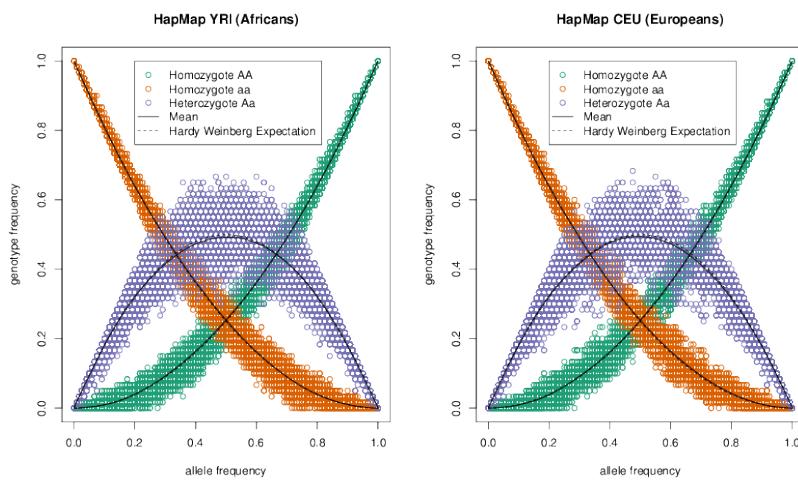


Figure 2.5: Demonstrating Hardy-Weinberg proportions using 10,000 SNPs from the HapMap European (CEU) and African (YRI) populations. Within each of these populations the allele frequency against the frequency of the 3 genotypes; each SNP is represented by 3 different coloured points. The solid lines show the mean genotype frequency. The dashed lines show the predicted genotype frequency from Hardy-Weinberg equilibrium. Code here. Blog post on figure here.

This so-called DNA fingerprinting has a range of applications from establishing paternity and identifying human remains to matching individuals to DNA samples from a crime scene. The FBI make use of the CODIS database². The CODIS database contains the genotypes of over 13 million people, most of whom have been convicted of a crime. Most of the profiles record genotypes at 13 microsatellite loci that are tetranucleotide repeats (since 2017, 20 sites have been genotyped).

The allele counts for two loci (D16S539 and TH01) are shown in table 2.2 and 2.3 for a sample of 155 people of European ancestry. You can assume these two loci are on different chromosomes.

allele name	80	90	100	110	120	121	130	140	150
allele count	3	34	13	102	97	1	44	13	3

allele name	60	70	80	90	93	100	110
allele counts	84	42	37	67	77	1	2

258 **Question 5.** You extract a DNA sample from a crime scene. The genotype is 100/80 at the D16S539 locus and 70/93 at TH01.

260 A) You have a suspect in custody. Assuming this suspect is innocent and of European ancestry, what is the probability that their genotype would match this profile by chance (a false-match probability)?

264 B) The FBI uses ≥ 13 markers. Why is this higher number necessary to make the match statement convincing evidence in court?

² CODIS: Combined DNA Index System

Table 2.2: Data for 155 Europeans at the D16S539 microsatellite from CODIS from LetterSpace=10ALGEE-HEWITT *et al.* (2016). The top row gives the number of tetranucleotide repeats for each allele, the bottom row gives the sample counts.

Table 2.3: Same as 2.2 but for the TH01 microsatellite.

266 C) An early case that triggered debate among forensic geneticists
was a crime among the Abenaki, a Native American community in
268 Vermont (see LetterSpace=10LEWONTIN, 1994, for discussion). There
was a DNA sample from the crime scene, and the perpetrator was
270 thought likely to be a member of the Abenaki community. Given
that allele frequencies vary among populations, why would people
272 be concerned about using data from a non-Abenaki population to
compute a false match probability?

274 **2.2 Allele sharing among related individuals and Identity by Descent**

276 All of the individuals in a population are related to each other by a
giant pedigree (family tree). For most pairs of individuals in a pop-
278 ulation these relationships are very distant (e.g. distant cousins),
while some individuals will be more closely related (e.g. sibling/first
280 cousins). All individuals are related to one another by varying lev-
els of relatedness, or *kinship*. Related individuals can share alleles
282 that have both descended from the shared common ancestor. To be
shared, these alleles must be inherited through all meioses connect-
284 ing the two individuals (e.g. surviving the $1/2$ probability of segrega-
tion each meiosis). As closer relatives are separated by fewer meioses,
286 closer relatives share more alleles. In Figure 2.6 we show the sharing
of chromosomal regions between two cousins. As we'll see, many
288 population and quantitative genetic concepts rely on how closely
related individuals are, and thus we need some way to quantify the
290 degree of kinship among individuals.

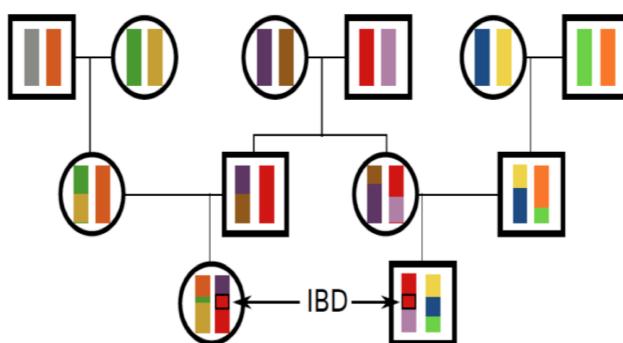


Figure 2.6: First cousins sharing a stretch of chromosome identical by descent. The different grandparental diploid chromosomes are coloured so we can track them and recombinations between them across the generations. Notice that the identity by descent between the cousins persists for a long stretch of chromosome due to the limited number of generations for recombination.

292 We will define two alleles to be identical by descent (IBD) if they
are identical due to transmission from a common ancestor in the past
few generations³. For the moment, we ignore mutation, and we will
294 be more precise about what we mean by 'past few generations' later

³ LetterSpace=10COTTERMAN, C. W., 1940 A calculus for statistico-genetics. Ph. D. thesis, The Ohio State University; and LetterSpace=10MALÉCOT, G., 1948 Les mathématiques de l'hérédité

on. For example, parent and child share exactly one allele identical by descent at a locus, assuming that the two parents of the child are randomly mated individuals from the population. In Figure 2.12, I show a pedigree demonstrating some configurations of IBD.

One summary of how related two individuals are is the probability that our pair of individuals share 0, 1, or 2 alleles identical by descent (see Figure 2.7). We denote these probabilities by r_0 , r_1 , and r_2 respectively. See Table 2.4 for some examples. We can also interpret these probabilities as genome-wide averages. For example, on average, at a quarter of all their autosomal loci full-sibs share zero alleles identical by descent.

One summary of relatedness that will be important is the probability that two alleles picked at random, one from each of the two different individuals i and j , are identical by descent. We call this quantity the *coefficient of kinship* of individuals i and j , and denote it by F_{ij} . It is calculated as

$$F_{ij} = 0 \times r_0 + \frac{1}{4}r_1 + \frac{1}{2}r_2. \quad (2.3)$$

The coefficient of kinship will appear multiple times, in both our discussion of inbreeding and in the context of phenotypic resemblance between relatives.

Relationship (i,j)*	r_0	r_1	r_2	F_{ij}
parent-child	0	1	0	$\frac{1}{4}$
full siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Monozygotic twins	0	0	1	$\frac{1}{2}$
1 st cousins	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{16}$

Question 6. What are r_0 , r_1 , and r_2 for 1/2 sibs? (1/2 sibs share one parent but not the other).

Our r coefficients are going to have various uses. For example, they allow us to calculate the probability of the genotypes of a pair of relatives. Consider a biallelic locus where allele A_1 is at frequency p , and two individuals who have IBD allele sharing probabilities r_0 , r_1 , r_2 . What is the overall probability that these two individuals are both homozygous for allele 1? Well that's

$$\begin{aligned} P(A_1A_1) = & P(A_1A_1|0 \text{ alleles IBD})P(0 \text{ alleles IBD}) \\ & + P(A_1A_1|1 \text{ allele IBD})P(1 \text{ allele IBD}) \\ & + P(A_1A_1|2 \text{ alleles IBD})P(2 \text{ alleles IBD}) \end{aligned} \quad (2.4)$$

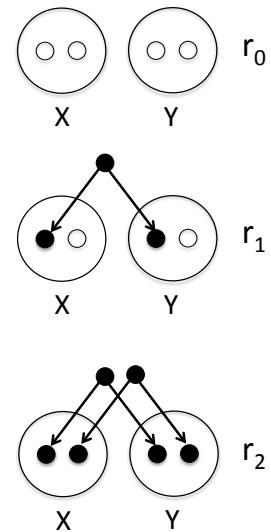


Figure 2.7: A pair of diploid individuals (X and Y) sharing 0, 1, or 2 alleles IBD where lines show the sharing of alleles by descent (e.g. from a shared ancestor).

Table 2.4: Probability that two individuals of a given relationship share 0, 1, or 2 alleles identical by descent on the autosomes. *Assuming this is the only close relationship the pair shares.

Or, in our r_0 , r_1 , r_2 notation:

$$\begin{aligned} P(A_1A_1) = & P(A_1A_1|0 \text{ alleles IBD})r_0 \\ & + P(A_1A_1|1 \text{ allele IBD})r_1 \\ & + P(A_1A_1|2 \text{ alleles IBD})r_2 \end{aligned} \quad (2.5)$$

- If our pair of relatives share 0 alleles IBD, then the probability that they are both homozygous is $P(A_1A_1|0 \text{ alleles IBD}) = p^2 \times p^2$, as all four alleles represent independent draws from the population. If they share 1 allele IBD, then the shared allele is of type A_1 with probability p , and then the other non-IBD allele, in both relatives, also needs to be A_1 which happens with probability p^2 , so $P(A_1A_1|1 \text{ allele IBD}) = p \times p^2$. Finally, our pair of relatives can share two alleles IBD, in which case $P(A_1A_1|2 \text{ alleles IBD}) = p^2$, because if one of our individuals is homozygous for the A_1 allele, both individuals will be. Putting this all together our equation (2.5) becomes

$$P(A_1A_1) = p^4r_0 + p^3r_1 + p^2r_2 \quad (2.6)$$

Note that for specific cases we could also calculate this by summing over all the possible genotypes their shared ancestor(s) had; however, that would be much more involved and not as general as the form we have derived here.

We can write out terms like eq (2.6) for all of the possible configurations of genotype sharing/non-sharing between a pair of individuals. Based on this we can write down the expected number of polymorphic sites where our individuals are observed to share 0, 1, or 2 alleles.

Question 7. The genotype of our suspect in Question 5 turns out to be 100/80 for D16S539 and 70/80 at TH01. The suspect is not a match to the DNA from the crime scene; however, they could be a sibling.

Calculate the joint probability of observing the genotype from the crime and our suspect:

- A) Assuming that they share no close relationship.
- B) Assuming that they are full sibs.
- C) Briefly explain your findings.

There's a variety of ways to estimate the relationships among individuals using genetic data. An example of using allele sharing to identify relatives is offered by the work of Nancy Chen (in collaboration with Stepfanie Aguillon, see LetterSpace=10CHEN *et al.*, 2016; LetterSpace=10AGUILLO *et al.*, 2017). LetterSpace=10CHEN *et al.* has collected genotyping data from thousands of Florida Scrub Jays at over ten thousand loci. These Jays live at the Archbold field



Figure 2.8: Florida Scrub-Jays (*Aphelocoma coerulescens*).

The birds of America : from drawings made in the United States and their territories. 1880. Audubon J.J. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

site, and have been carefully monitored for many decades allowing the pedigree of many of the birds to be known. Using these data she estimates allele frequencies at each locus. Then by equating the observed number of times that a pair of individuals share 0, 1, or 2 alleles to the theoretical expectation, she estimates the probability of r_0 , r_1 , and r_2 for each pair of birds. A plot of these are shown in Figure 2.9, showing how well the estimates match those known from the pedigree.

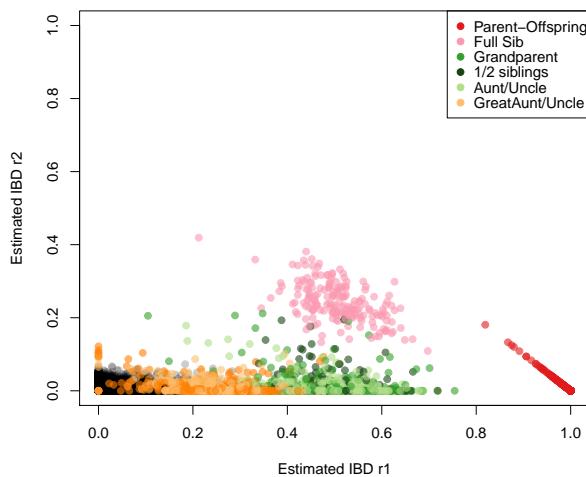


Figure 2.9: Estimated coefficient of kinship from Florida Scrub Jays. Each point is a pair of individuals, plotted by their estimated IBD (r_1 and r_2) from their genetic data. The points are coloured by their known pedigree relationships. Note that most pairs have low kinship, and no recent genealogical relationship, and so appear as black points in the lower left corner. Thanks to Nancy Chen for supplying the data. [Code here.](#)

Sharing of genomic blocks among relatives. We can more directly see the sharing of the genome among close relatives using high-density SNP genotyping arrays. Below we show a simulation of you and your first cousin's genomic material that you both inherited from your shared grandmother. Colored purple are regions where you and your cousin will have matching genomic material, due to having inherited it IBD from your shared grandmother.

You and your first cousin will share at least one allele of your genotype at all of the polymorphic loci in these purple regions. There's a range of methods to detect such sharing. One way is to look for unusually long stretches of the genome where two individuals are never homozygous for different alleles. By identifying pairs of individuals who share an unusually large number of such putative IBD blocks, we can hope to identify unknown relatives in genotyping datasets. In fact, companies like 23&me and Ancestry.com use signals of IBD to help identify family ties.

As another example, consider the case of third cousins. You share

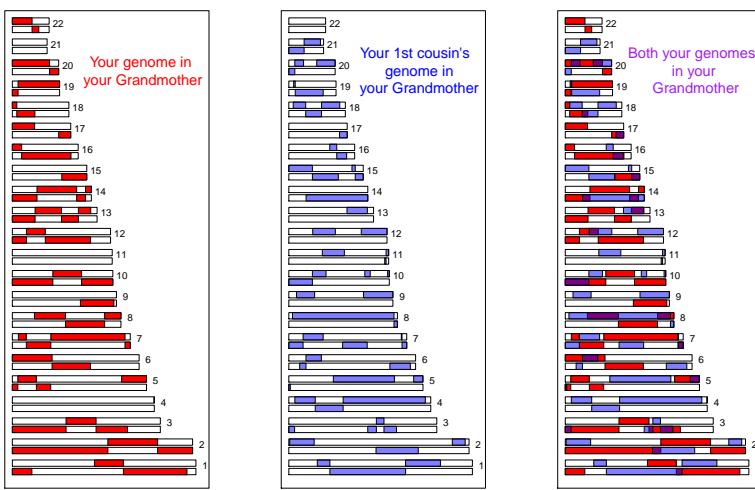


Figure 2.10: A simulation of sharing between first cousins. The regions of your grandmother's 22 autosomes that you inherited are coloured red, those that your cousins inherited are coloured blue. In the third panel we show the overlapping genomic regions in purple, these regions will be IBD in you and your cousin. If you are full first cousins, you will also have shared genomic regions from your shared grandfather, not shown here. Details about how we made these simulations here.

one of eight sets of great-great grandparents with each of your (likely many) third cousins. On average, you and each of your third cousins each inherit one-sixteenth of your genome from each of those two great-great grandparents. This turns out to imply that on average, a little less than one percent of your and your third cousin's genomes ($2 \times (1/16)^2 = 0.78\%$) will be identical by virtue of descent from those shared ancestors. A simulated example where third cousins share blocks of their genome (on chromosome 16 and 2) due to their great, great grandmother is shown in Figure 2.11.

Note how if you compare Figure 2.11 and Figure 2.10, individuals inherit less IBD from a shared great, great grandmother than from a shared grandmother, as they inherit from more total ancestors further back. Also notice how the sharing occurs in shorter genomic blocks, as it has passed through more generations of recombination during meiosis. These blocks are still detectable, and so third cousins can be detected using high-density genotyping chips, allowing more distant relatives to be identified than single marker methods alone.⁴ More distant relations than third cousins, e.g. fourth cousins, start to have a significant probability of sharing none of their genome IBD. But you have many fourth cousins, so you will share some of your genome IBD with some of them; however, it gets increasingly hard to identify the degree of relatedness from genetic data the deeper in the family tree this sharing goes.

⁴ Indeed the suspect in case of the Golden State Killer was identified through identifying third cousins that genetically matched a DNA sample from an old crime scene (see a here for more details).

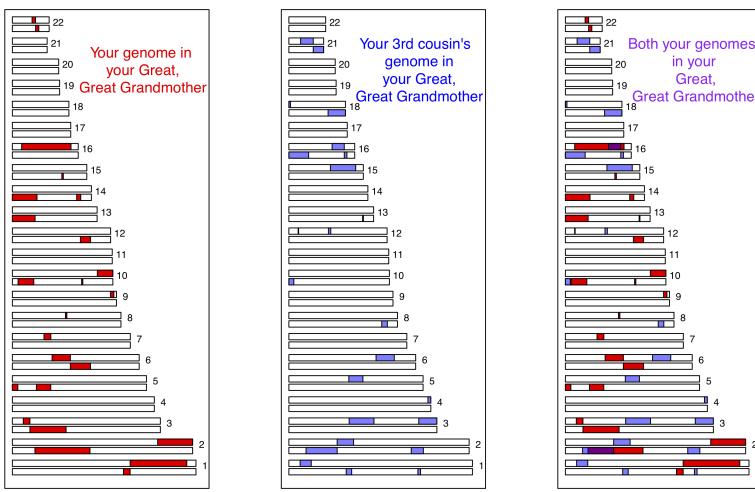


Figure 2.11: A simulation of sharing between third cousins, the details are the same as in Figure 2.10.

400 2.2.1 Inbreeding

We can define an inbred individual as an individual whose parents 402 are more closely related to each other than two random individuals drawn from some reference population.

When two related individuals produce an offspring, that individual 404 can receive two alleles that are identical by descent, i.e. they can 406 be homozygous by descent (sometimes termed autozygous), due to 408 the fact that they have two copies of an allele through different paths 410 through the pedigree. This increased likelihood of being homozygous 412 relative to an outbred individual is the most obvious effect of 414 inbreeding. It is also the one that will be of most interest to us, as it underlies a lot of our ideas about inbreeding depression and population structure. For example, in Figure 2.12 our offspring of first 416 cousins is homozygous by descent having received the same IBD 418 allele via two different routes around an inbreeding loop.

As the offspring receives a random allele from each parent (i and 416 j), the probability that those two alleles are identical by descent is 418 equal to the kinship coefficient F_{ij} of the two parents (Eqn. 2.3). This follows from the fact that the genotype of the offspring is made by sampling an allele at random from each of our parents.

$$\begin{array}{ccc} f_{11} & f_{12} & f_{22} \\ \hline (1-F)p^2 + Fp & (1-F)2pq & (1-F)q^2 + Fq \end{array}$$

420 The only way the offspring can be heterozygous (A_1A_2) is if their

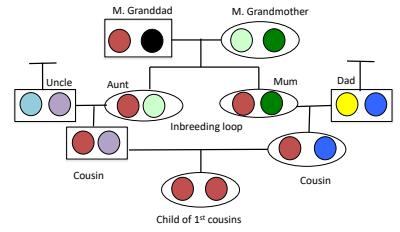


Figure 2.12: Alleles being transmitted through an inbred pedigree. The two sisters (mum and aunt) share two alleles identical by descent (IBD). The cousins share one allele IBD. The offspring of first cousins is homozygous by descent at this locus.

Table 2.5: Generalized Hardy-Weinberg

⁴²² two alleles at a locus are not IBD (otherwise they would necessarily be homozygous). Therefore, the probability that they are heterozygous is

$$(1 - F)2pq, \quad (2.7)$$

⁴²⁴ where we have dropped the indices i and j for simplicity. The offspring can be homozygous for the A_1 allele in two different ways.
⁴²⁶ They can have two non-IBD alleles that are not IBD but happen to be of the allelic type A_1 , or their two alleles can be IBD, such that they ⁴²⁸ inherited allele A_1 by two different routes from the same ancestor. Thus, the probability that an offspring is homozygous for A_1 is

$$(1 - F)p^2 + Fp. \quad (2.8)$$

⁴³⁰ Therefore, the frequencies of the three possible genotypes can be written as given in Table 2.5, which provides a generalization of the ⁴³² Hardy–Weinberg proportions.

⁴³⁴ Note that the generalized Hardy–Weinberg proportions completely specify the genotype probabilities, as there are two parameters (p and F) and two degrees of freedom (as p and q have to sum to one). ⁴³⁶ Therefore, any combination of genotype frequencies at a biallelic site can be specified by a combination of p and F .

⁴³⁸ **Question 8.** The frequency of the A_1 allele is p at a biallelic locus. Assume that our population is randomly mating and that the genotype frequencies in the population follow from HW. We select two ⁴⁴⁰ individuals at random to mate from this population. We then mate the children from this cross. What is the probability that the child ⁴⁴² from this full sib-mating is homozygous?

⁴⁴⁴ *Multiple inbreeding loops in a pedigree.* Up to this point we have assumed that there is at most one inbreeding loop in the recent family ⁴⁴⁶ history of our individuals, i.e. the parents of our inbred individual have at most one recent genealogical connection. However, an individual who has multiple inbreeding loops in their pedigree can be ⁴⁴⁸ homozygous by descent thanks to receiving IBD alleles via multiple ⁴⁵⁰ different different loops. To calculate inbreeding in pedigrees of arbitrary complexity, we can extend beyond our original relatedness ⁴⁵² coefficients r_0 , r_1 , and r_2 to account for higher order sharing of alleles ⁴⁵⁴ IBD among relatives. For example, we can ask, what is the probability that *both* of the alleles in the first individual are shared IBD with one allele in the second individual? There are nine possible relatedness ⁴⁵⁶ coefficients in total to completely describe kinship between two diploid individuals, and we won't go in to them here as it's a lot to

⁴⁵⁸ keep track of. However, we will show how we can calculate the inbreeding coefficient of an individual with multiple inbreeding loops
⁴⁶⁰ more directly.

⁴⁶² Let's say the parents of our inbred individual (B and C) have K
⁴⁶⁴ shared ancestors, i.e. individuals who appear in both B and C's re-
⁴⁶⁶ cent family trees. We denote these shared ancestors by A_1, \dots, A_K ,
⁴⁶⁸ and we denote by n the total number of individuals in the chain from
⁴⁷⁰ B to C via ancestor A_i , including B, C, and A_i . For example, if B is
C's aunt, then B and C share two ancestors, which are B's parents
and, equivalently, C's grandparents. In this case, there are $n=4$ indi-
viduals from B to C through each of these two shared ancestor. In the
general case, the kinship coefficient of B and C, i.e. the inbreeding
coefficient of their child, is

$$F = \sum_{i=1}^K \frac{1}{2^{n_i}} (1 + f_{A_i}) \quad (2.9)$$

⁴⁷² where f_{A_i} is the inbreeding coefficient of the ancestor A_i . What's
⁴⁷⁴ happening here is that we sum over all the mutually-exclusive paths
in the pedigree through which B and C can share an allele IBD. With
⁴⁷⁶ probability $1/2^{n_i}$, a pair of alleles picked at random from B and C is
descended from the same ancestral allele in individual A_i , in which
⁴⁷⁸ case the alleles are IBD.⁵ However, even if B inherits the maternal
allele and C inherits the paternal allele of shared ancestor A_i , if A_i
was themselves inbred, with probability f_{A_i} those two alleles are
themselves IBD. Thus a shared *inbred* ancestor further increases the
kinship of B and C.

⁵ For example, in the case of our aunt-nephew case, assuming that the aunt's two parents are their only recent shared ancestors, then $F = 1/2^4 + 1/2^4 = 1/8$, in agreement with the answer we would obtain from eqn (2.3).

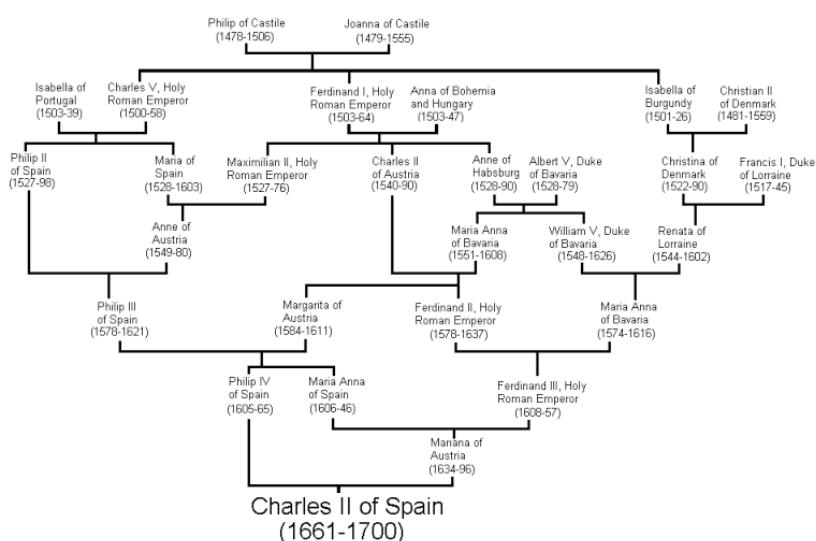


Figure 2.13: The pedigree of King Charles II of Spain. Pedigree from wikimedia drawn by Lec CRP1, public domain.



Figure 2.14: Charles II of Spain (by Juan Carreño de Miranda, 1685). Public Domain.

Multiple inbreeding loops increase the probability that a child is homozygous by descent at a locus, which can be calculated simply by plugging in F , the child's inbreeding coefficient, into our generalized HW equation.

As one extreme example of the impact of multiple inbreeding loops in an individual's pedigree, let's consider king Charles II of Spain, the last of the Spanish Habsburgs. Charles was the son of Philip IV of Spain and Mariana of Austria, who were uncle and niece. If this were the only inbreeding loop, then Charles would have had an inbreeding coefficient of $1/8$. Unfortunately for Charles, the Spanish Habsburgs had long kept wealth and power within their family by arranging marriages between close kin. The pedigree of Charles II is shown in Figure 2.13, and multiple inbreeding loops are apparent. For example, Phillip III, Charles II's grandfather and great-grandfather, was himself a child of an uncle-niece marriage.

LetterSpace=10 ALVAREZ *et al.* (2009) calculated that Charles II had an inbreeding coefficient of 0.254, equivalent to a full-sib mating, thanks to all of the inbreeding loops in his pedigree. Therefore, he is expected to have been homozygous by descent for a full quarter of his genome. As we'll talk about later in these notes, this means that Charles may have been homozygous for a number of recessive disease alleles, and indeed he was a very sickly man who left no descendants due to his infertility.⁶ Thus plausibly the end of one of

⁶ Pedro Gargantilla, who performed Charles's autopsy, stated that his body "did not contain a single drop of blood; his heart was the size of a peppercorn; his lungs corroded; his intestines rotten and gangrenous; he had a single testicle, black as coal, and his head was full of water." While some of this description may refer to actual medical conditions, some of these details seem a little unlikely. See here.

504 the great European dynasties came about through inbreeding.

2.2.2 Calculating inbreeding coefficients from genetic data

506 If the observed heterozygosity in a population is H_O , and we assume
that the generalized Hardy–Weinberg proportions hold, we can set
508 H_O equal to f_{12} , and solve Eq. (2.7) for F to obtain an estimate of the
inbreeding coefficient as

$$\hat{F} = 1 - \frac{f_{12}}{2pq} = \frac{2pq - f_{12}}{2pq}. \quad (2.10)$$

510 As before, p is the frequency of allele A_1 in the population. This
can be rewritten in terms of the observed heterozygosity (H_O) and
512 the heterozygosity expected in the absence of inbreeding, $H_E = 2pq$,
as

$$\hat{F} = \frac{H_E - H_O}{H_E} = 1 - \frac{H_O}{H_E}. \quad (2.11)$$

514 Hence, \hat{F} quantifies the deviation due to inbreeding of the observed
heterozygosity from the one expected under random mating, relative
516 to the latter.

Question 9. Suppose the following genotype frequencies were ob-
518 served for an esterase locus in a population of *Drosophila* (A denotes
the “fast” allele and B denotes the “slow” allele):

	AA	AB	BB
	0.6	0.2	0.2

520 What is the estimate of the inbreeding coefficient at the esterase
locus?

If we have multiple loci, we can replace H_O and H_E by their means
524 over loci, \bar{H}_O and \bar{H}_E , respectively. Note that, in principle, we could
also calculate F for each individual locus first, and then take the aver-
526 age across loci. However, this procedure is more prone to introducing
a bias if sample sizes vary across loci, which is not unlikely when we
528 are dealing with real data.

Genetic markers are commonly used to estimate inbreeding for
530 wild and/or captive populations of conservation concern. As an
example of this, consider the case of the Mexican wolf (*Canis lupus*
532 *baileyi*), also known as the lobo, a sub-species of gray wolf.

They were extirpated in the wild during the mid-1900s due to
534 hunting, and the remaining five lobos in the wild were captured
to start a breeding program. LetterSpace=10VONHOLDT *et al.* (2011)
536 estimated the current-day, average expected heterozygosity to be 0.18,
based on allele frequencies at over forty thousand SNPs. However,
538 the average lobos individual was only observed to be heterozygous



Figure 2.15: Grey wolf (*Canis lupus*). Dogs, jackals, wolves, and foxes: a monograph of the Canidae. 1890. y J.G. Keulemans. Image from the Biodiversity Heritage Library. Contributed by University of Toronto - Gerstein Science Information Centre. Not in copyright.

at 12% of these SNPs. Therefore, the average inbreeding coefficient for the lobo is $F = 1 - 0.12/0.18$, i.e. $\sim 33\%$ of a lobo's genome is homozygous due to recent inbreeding in their pedigree.

Genomic blocks of homozygosity due to inbreeding. As we saw above, close relatives are expected to share alleles IBD in large genomic blocks. Thus, when related individuals mate and transmit alleles to an inbred offspring, they transmit these alleles in big blocks through meiosis. An example, lets return to the case of our hypothetical first cousins from Figure 2.6. If this pair of individuals had a child, one possible pattern of genetic transmission is shown in Figure 2.16. The child has inherited the red stretch of chromosome via two different routes through their pedigree from the grandparents. This is an example of an autozygous segment, where the child is homozygous by descent at all of the loci in this red region. The inbreeding

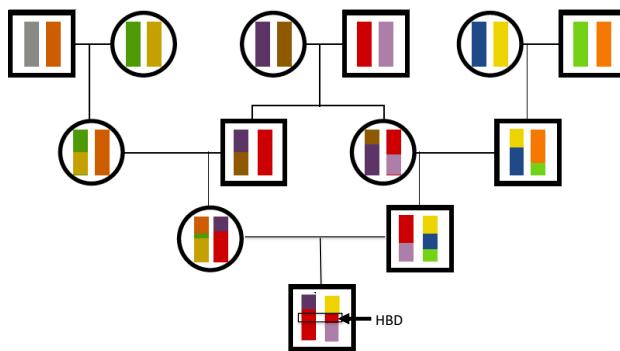


Figure 2.16: .

coefficient of the child sets the proportion of their genome that will be in these autozygous segments. For example, a child of first full cousins is expected to have $1/16$ of their genome in these segments. The more distant the loop in the pedigree, the more meioses that chromosomes have been through and the shorter individual blocks will be. A child of first cousins will have longer blocks than a child of second cousins, for example.

Individuals with multiple inbreeding loops in their family tree can have a high inbreeding coefficient due to the combined effect of many small blocks of autozygosity. For example, Charles II had an inbreeding coefficient that is equivalent to that of the child of full-sibs, with a quarter of his genome expected to homozygous by descent, but this would be made up of many shorter blocks.

We can hope to detect these blocks by looking for unusually long genomic runs of homozygosity (ROH) sites in an individual's genome. One way to estimate an individual's inbreeding coefficient is then to total up the proportion of an individual's genome that falls in

570 such ROH regions. This estimate is called F_{ROH} .

An example of using F_{ROH} to study inbreeding comes from the
572 work of LetterSpace=10SAMS and LetterSpace=10BOYKO (2018b), who
identified runs of homozygosity in 2,500 dogs, ranging from 500kb
up to many megabases. Figure 2.18 shows the distribution of

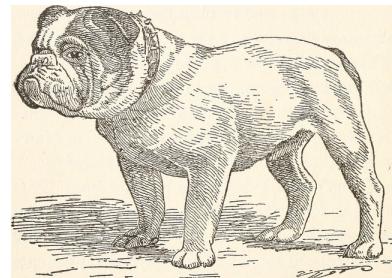
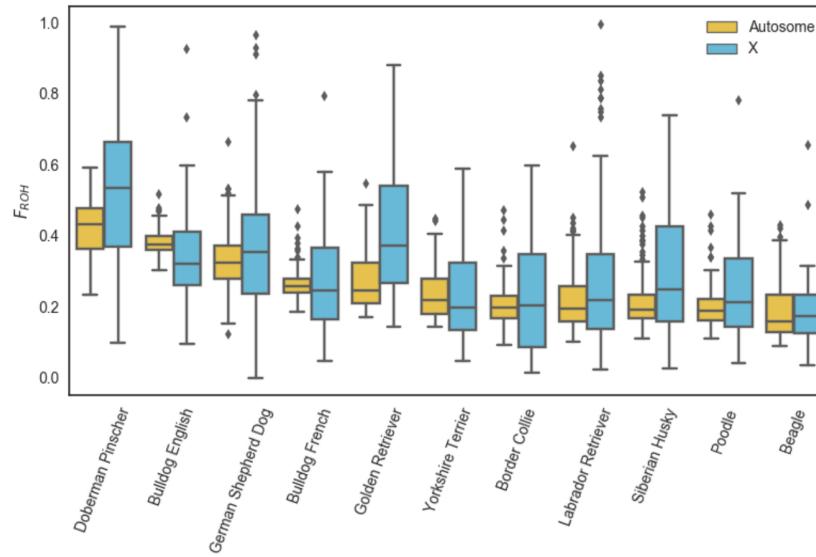


Figure 2.18: The distribution of F_{ROH} for individuals from 10 different dog breeds. The breeds are Doberman Pinscher, English Bulldog, German Shepherd Dog, Bulldog French, Golden Retriever, Yorkshire Terrier, Border Collie, Labrador Retriever, Siberian Husky, Poodle, and Beagle. The figure is licensed under CC BY 4.0.

574 F_{ROH} of individuals in each dog breed for the X and autosome. In

576 Figure 2.19 this is broken down by the length of ROH segments.

Dog breeds have been subject to intense breeding that has resulted
578 in high levels of inbreeding. Of the population samples examined,
Doberman Pinschers have the highest levels of their genome in runs
580 of homozygosity (F_{ROH}), somewhat higher than English bulldogs.
In 2.19 we can see that English bulldogs have more short ROH than
582 Doberman Pinschers, but that Doberman Pinschers have more of
their genome in very large ROH (> 16Mb). This suggests that En-
584 glish bulldogs have had long history of inbreeding but that Dober-
man Pinschers have a lot of recent inbreeding in their history.

586 2.3 Summarizing population structure

LetterSpace=10INDIVIDUALS RARELY MATE COMPLETELY AT RANDOM;
588 your parents weren't two Bilateria plucked at random from the tree
of life. Even within species, there's often geographically-restricted
590 mating among individuals. Individuals tend to mate with individuals
from the same, or closely related sets of populations. This form of
592 non-random mating is called population structure and can have

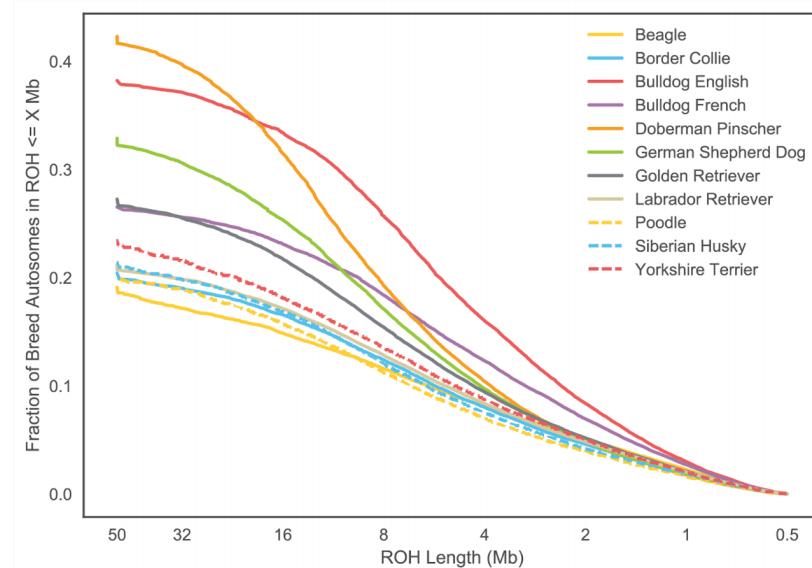


Figure 2.19: Cumulative density of ROH length, measured in megabases (Mb) from LetterSpace=10SAMS and LetterSpace=10BOYKO (2018a) for various dog breeds (licensed under CC BY 4.0). Note that longer lengths of ROH are on the left of the plot.

profound effects on the distribution of genetic variation within and among natural populations.
594

2.3.1 Inbreeding as a summary of population structure.

596 It turns out that statements about inbreeding represent one natural way to summarize population structure. We defined inbreeding
598 as having parents that are more closely related to each other than two individuals drawn at random from some reference population. The
600 question that naturally arises is: Which reference population should we use? While I might not look inbred in comparison to allele frequencies in the United Kingdom (UK), where I am from, my parents certainly are not two individuals drawn at random from the world-wide population. If we estimated my inbreeding coefficient F using allele frequencies within the UK, it would be close to zero, but would
602 likely be larger if we used world-wide frequencies. This is because there is a somewhat lower level of expected heterozygosity within the
604 UK than in the human population across the world as a whole.
606

608 LetterSpace=10WRIGHT⁷ developed a set of 'F-statistics' (also
610 called 'fixation indices') that formalize the idea of inbreeding with respect to different levels of population structure. See Figure 2.20 for a schematic diagram. Wright defined F_{XY} as the correlation between random gametes, drawn from the same level X , relative to level Y .
612 We will return to why F-statistics are statements about correlations between alleles in just a moment. One commonly used F-statistic is
614 F_{IS} , which is the inbreeding coefficient between an individual (I) and
616

⁷ LetterSpace=10WRIGHT, S., 1943 Isolation by Distance. Genetics 28(2): 114–138; and LetterSpace=10WRIGHT, S., 1949 The Genetical Structure of Populations. Annals of Eugenics 15(1): 323–354

the subpopulation (S). Consider a single locus, where in a subpopulation (S) a fraction $H_I = f_{12}$ of individuals are heterozygous. In this subpopulation, let the frequency of allele A_1 be p_S , such that the expected heterozygosity under random mating is $H_S = 2p_S(1 - p_S)$. We will write F_{IS} as

$$F_{IS} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_S q_S}, \quad (2.12)$$

a direct analog of eqn. 2.10. Hence, F_{IS} is the relative difference between observed and expected heterozygosity due to a deviation from random mating within the subpopulation. We could also compare the observed heterozygosity in individuals (H_I) to that expected in the total population, H_T . If the frequency of allele A_1 in the total population is p_T , then we can write F_{IT} as

$$F_{IT} = 1 - \frac{H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T}, \quad (2.13)$$

which compares heterozygosity in individuals to that expected in the total population. As a simple extension of this, we could imagine comparing the expected heterozygosity in the subpopulation (H_S) to that expected in the total population H_T , via F_{ST} :

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T}. \quad (2.14)$$

We can relate the three F -statistics to each other as

$$(1 - F_{IT}) = \frac{H_I}{H_S} \frac{H_S}{H_T} = (1 - F_{IS})(1 - F_{ST}). \quad (2.15)$$

Hence, the reduction in heterozygosity within individuals compared to that expected in the total population can be decomposed to the reduction in heterozygosity of individuals compared to the subpopulation, and the reduction in heterozygosity from the total population to that in the subpopulation.

If we want a summary of population structure across multiple subpopulations, we can average H_I and/or H_S across populations, and use a p_T calculated by averaging p_S across subpopulations (or our samples from sub-populations). For example, the average F_{ST} across K subpopulations (sampled with equal effort) is

$$F_{ST} = 1 - \frac{\bar{H}_S}{H_T}, \quad (2.16)$$

where $\bar{H}_S = 1/K \sum_{i=1}^K H_S^{(i)}$, and $H_S^{(i)} = 2p_i q_i$ is the expected heterozygosity in subpopulation i . It follows that the average heterozygosity of the sub-populations $\bar{H}_S \leq H_T$,⁸ and so $F_{ST} \geq 0$ and $F_{IS} \leq F_{IT}$.

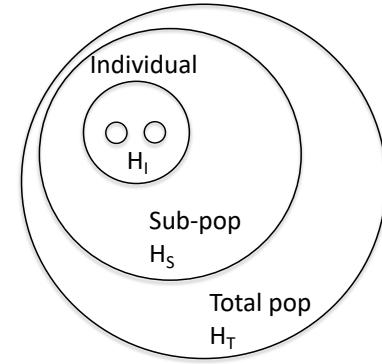


Figure 2.20: The hierarchical nature of F -statistics. The two dots within an individual represent the two alleles at a locus for an individual I . We can compare the heterozygosity in individuals (H_I), to that found by randomly drawing alleles from the sub-population (S), to that found in the total population (T).

⁸ This observation that the average heterozygosity of the sub-populations must be less than or equal to that of the total population is called the Wahlund effect.

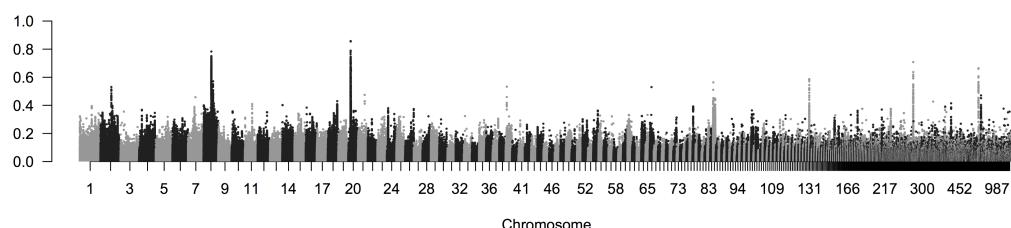
646 Furthermore, if we have multiple sites, we can replace H_I , H_S , and H_T with their averages across loci (as above). ⁹

648 As an example of comparing a genome-wide estimate of F_{ST} to that at individual loci we can look at some data from blue- and
650 golden-winged warblers (*Vermivora cyanoptera* and *V. chrysoptera* 1-
2 & 5-6 in Figure 2.21).

652 These two species are spread across eastern Northern America, with the golden-winged warbler having a smaller, more northerly range. They're quite different in terms of plumage, but have long been known to have similar songs and ecologies. The two species 654 hybridize readily in the wild; in fact two other previously-recognized species, Brewster's and Lawrence's warbler (4 & 3 in 2.21), are actually found to just be hybrids between these two species. The golden-winged warbler is listed as 'threatened' under the Canadian endangered species act. The golden-winged warbler's habitat is under pressure from human activity and increased hybridization with the blue-winged warbler, which is moving north into its range, also poses a significant issue. LetterSpace=10Toews *et al.* (2016) investigated 656 the population genomics of these warblers, sequencing ten golden- and ten blue-winged warblers. They found very low divergence 658 among these species, with a genome-wide $F_{ST} = 0.0045$. In Figure 2.22, per SNP F_{ST} is averaged in 2000bp windows moving along the genome. The average is very low, but some regions of very high



Figure 2.21: Blue-, golden-winged, and Lawrence's warblers (*Vermivora*).
The warblers of North America. Chapman, F.M. 1907.
Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.



668 F_{ST} stand out. Nearly all of these regions correspond to large allele
670 frequency differences at loci in, or close, to genes known to be involved in plumage colouration differences in other birds. To illustrate
672 these frequency differences LetterSpace=10Toews *et al.* genotyped a SNP in each of these high- F_{ST} regions. Here's their genotyping counts from the SNP, segregating for an allele 1 and 2, in the *Wnt* region, a key regulatory gene involved in feather development:

Species	11	12	22
Blue-winged	2	21	31
Golden-winged	48	12	1

Question 10. With reference to the table of *Wnt*-allele counts:

Figure 2.22: F_{ST} between blue- and golden-winged warbler population samples at SNPs across the genome. Each dot is a SNP, and SNPs are coloured alternating by scaffold. Thanks to David Toews for the figure.

- 678 **A)** Calculate F_{IS} in blue-winged warblers.
 680 **B)** Calculate F_{ST} for the sub-population of blue-winged warblers
 compared to the combined sample.
 681 **C)** Calculate mean F_{ST} across both sub-populations.

682 *Interpretations of F-statistics* Let us now return to Wright's definition
 684 of the F -statistics as correlations between random gametes, drawn
 from the same level X , relative to level Y . Without loss of generality,
 we may think about X as individuals and S as the subpopulation.
 686 Rewriting F_{IS} in terms of the observed homozygote frequencies (f_{11} ,
 f_{22}) and expected homozygosities (p_S^2, q_S^2) we find

$$F_{IS} = \frac{2p_S q_S - f_{12}}{2p_S q_S} = \frac{f_{11} + f_{22} - p_S^2 - q_S^2}{2p_S q_S}, \quad (2.17)$$

688 using the fact that $p^2 + 2pq + q^2 = 1$, and $f_{12} = 1 - f_{11} - f_{22}$. The
 form of eqn. (2.17) reveals that F_{IS} is the covariance between pairs
 690 of alleles found in an individual, divided by the expected variance
 under binomial sampling. Thus, F -statistics can be understood as the
 692 correlation between alleles drawn from a population (or an individ-
 ual) above that expected by chance (i.e. drawing alleles sampled at
 694 random from some broader population).

We can also interpret F -statistics as proportions of variance ex-
 696 plained by different levels of population structure. To see this, let
 us think about F_{ST} averaged over K subpopulations, whose fre-
 698 quencies are p_1, \dots, p_K . The frequency in the total population is
 $p_T = \bar{p} = 1/K \sum_{i=1}^K p_i$. Then, we can write

$$F_{ST} = \frac{2\bar{p}\bar{q} - \frac{1}{K} \sum_{i=1}^K 2p_i q_i}{2\bar{p}\bar{q}} = \frac{\left(\frac{1}{K} \sum_{i=1}^K p_i^2 + \frac{1}{K} \sum_{i=1}^K q_i^2 \right) - \bar{p}^2 - \bar{q}^2}{2\bar{p}\bar{q}} = \frac{\text{Var}(p_1, \dots, p_K)}{\text{Var}(\bar{p})}, \quad (2.18)$$

700 which shows that F_{ST} is the proportion of the variance explained by
 the subpopulation labels.

702 2.3.2 Other approaches to population structure

There is a broad spectrum of methods to describe patterns of popula-
 704 tion structure in population genetic datasets. We'll briefly discuss two
 broad-classes of methods that appear often in the literature: assign-
 706 ment methods and principal components analysis.

2.3.3 Assignment Methods

708 Here we'll describe a simple probabilistic assignment to find the
 probability that an individual of unknown population comes from

710 one of K predefined populations. For example, there are three broad
 712 populations of common chimpanzee (*Pan troglodytes*) in Africa: western, central, and eastern. Imagine that we have a chimpanzee whose
 714 population of origin is unknown (e.g. it's from an illegal private collection). If we have genotyped a set of unlinked markers from a panel
 716 of individuals representative of these populations, we can calculate
 the probability that our chimp comes from each of these populations.

718 We'll then briefly explain how to extend this idea to cluster a set of
 720 individuals into K initially unknown populations. This method is a
 simplified version of what population genetics clustering algorithms
 such as STRUCTURE and ADMIXTURE do.¹⁰

A simple assignment method We have genotype data from unlinked S
 722 biallelic loci for K populations. The allele frequency of allele A_1 at locus l in population k is denoted by $p_{k,l}$, so that the allele frequencies
 724 in population 1 are $p_{1,1}, \dots, p_{1,L}$ and population 2 are $p_{2,1}, \dots, p_{2,L}$ and
 so on.

726 You genotype a new individual from an unknown population
 at these L loci. This individual's genotype at locus l is g_l , where g_l
 728 denotes the number of copies of allele A_1 this individual carries at
 this locus ($g_l = 0, 1, 2$).

730 The probability of this individual's genotype at locus l conditional
 on coming from population k , i.e. their alleles being a random HW
 732 draw from population k , is

$$P(g_l | \text{pop } k) = \begin{cases} (1 - p_{k,l})^2 & g_l = 0 \\ 2p_{k,l}(1 - p_{k,l}) & g_l = 1 \\ p_{k,l}^2 & g_l = 2 \end{cases} \quad (2.19)$$

Assuming that the loci are independent, the probability of the
 734 individual's genotype across all S loci, conditional on the individual
 coming from population k , is

$$P(\text{ind.} | \text{pop } k) = \prod_{l=1}^S P(g_l | \text{pop } k) \quad (2.20)$$

736 We wish to know the probability that this new individual comes
 from population k , i.e. $P(\text{pop } k | \text{ind.})$. We can obtain this through
 738 Bayes' rule

$$P(\text{pop } k | \text{ind.}) = \frac{P(\text{ind.} | \text{pop } k)P(\text{pop } k)}{P(\text{ind.})} \quad (2.21)$$

where

$$P(\text{ind.}) = \sum_{k=1}^K P(\text{ind.} | \text{pop } k)P(\text{pop } k) \quad (2.22)$$

¹⁰ LetterSpace=10PRITCHARD, J. K., LetterSpace=10M. STEPHENS, and LetterSpace=10P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945–959; and LetterSpace=10ALEXANDER, D. H., LetterSpace=10J. NOVEMBRE, and LetterSpace=10K. LANGE, 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19(9): 1655–1664

740 is the normalizing constant. We interpret $P(\text{pop } k)$ as the prior probability
 741 of the individual coming from population k , and unless we
 742 have some other prior knowledge we will assume that the new individual has an equal probability of coming from each population

744 $P(\text{pop } k) = 1/K$.

We interpret

$$P(\text{pop } k|\text{ind.}) \quad (2.23)$$

746 as the posterior probability that our new individual comes from each of our $1, \dots, K$ populations.

748 More sophisticated versions of this are now used to allow for hybrids, e.g., we can have a proportion q_k of our individual's genome 750 come from population k and estimate the set of q_k 's.

Question 11.

752 Returning to our chimp example, imagine that we have genotyped a set of individuals from the Western and Eastern populations at two 754 SNPs (we'll ignore the central population to keep things simpler). The frequency of the capital allele at two SNPs (A/a and B/b) is given by

Population	locus A	locus B
Western	0.1	0.85
Eastern	0.95	0.2

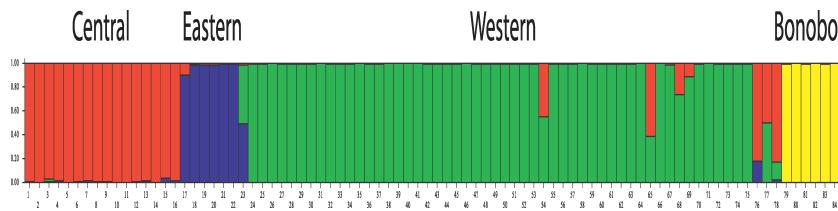
758 **A)** Our individual, whose origin is unknown, has the genotype AA at the first locus and bb at the second. What is the posterior probability
 760 that our individual comes from the Western population versus Eastern chimp population?

762 **B)** Let's assume that our individual is a hybrid. At each locus, with probability q_W our individual draws an allele from the Western population
 764 and with probability $q_C = 1 - q_W$ they draw an allele from the Eastern population. What is the probability of our individual's 766 genotype given q_C ?

Optional You could plot this probability as a function of q_W . How
 768 does your plot change if our individual is heterozygous at both loci?

770 *Clustering based on assignment methods* While it is great to be able to assign our individuals to a particular population, these ideas can be pushed to learn about how best to describe our genotype data in terms of discrete populations without assigning any of our individuals to populations *a priori*. We wish to cluster our individuals into 772 K unknown populations. We begin by assigning our individuals at random to these K populations.

- 776 1. Given these assignments we estimate the allele frequencies at all
of our loci in each population.
- 778 2. Given these allele frequencies we chose to reassign each individual
to a population k with a probability given by eqn. (2.20).
- 780 We iterate steps 1 and 2 for many iterations (technically, this ap-
proach is known as *Gibbs Sampling*). If the data is sufficiently infor-
782 mative, the assignments and allele frequencies will quickly converge
on a set of likely population assignments and allele frequencies for
these populations.



784 To do this in a full Bayesian scheme we need to place priors on
786 the allele frequencies (for example, one could use a beta distribu-
tion prior). Technically we are using the joint posterior of our al-
788 lele frequencies and assignments. Programs like STRUCTURE, use
this type of algorithm to cluster the individuals in an “unsuper-
790 vised” manner (i.e. they work out how to assign individuals to an
unknown set of populations). See Figure 2.23 for an example of
792 LetterSpace=10BECQUET *et al.* using STRUCTURE to determine the
population structure of chimpanzees.

794 STRUCTURE-like methods have proven incredible popular and
useful in examining population structure within species. However,
796 the results of these methods are open to misinterpretation; see Let-
terSpace=10LAWSON *et al.* (2018) for a recent discussion. Two common
798 mistakes are 1) taking the results of STRUCTURE-like approaches
for some particular value of K and taking this to represent the best
800 way to describe population-genetic variation. 2) Thinking that these
clusters represent ‘pure’ ancestral populations.

802 There is no right choice of K , the number of clusters to partition
into. There are methods of judging the ‘best’ K by some statistical
804 measure given some particular dataset, but that is not the same as
saying this is the most meaningful level on which to summarize
806 population structure in data. For example, running STRUCTURE
on world-wide human populations for low value of K will result
808 in population clusters that roughly align with continental popula-
tions (LetterSpace=10ROSENBERG *et al.*, 2002). However, that does
810 not tell us that assigning ancestry at the level of continents is a par-

Figure 2.23: LetterSpace=10BECQUET *et al.* (2007) genotyped 78 common chimpanzee and 6 bonobo at over 300 polymorphic markers (in this case microsatellites). They ran STRUCTURE to cluster the individuals using these data into $K = 4$ populations. In LetterSpace=10BECQUET *et al.* (2007) above figure they show each individual as a vertical bar divided into four colours depicting the estimate of the fraction of ancestry that each individual draws from each of the four estimated populations (licensed under CC BY 4.0). We can see that these four colours/populations correspond to: Red, central; blue, eastern; green, western; yellow, bonobo.

ticularly meaningful way of partitioning individuals. Running the same data for higher value of K, or within continental regions, will result in much finer-scale partitioning of continental groups (LetterSpace=10ROSENBERG *et al.*, 2002; LetterSpace=10LI *et al.*, 2008). No one of these layers of population structure identified is privileged as being more meaningful than another.

It is tempting to think of these clusters as representing ancestral populations, which themselves are not the result of admixture. However, that is not the case, for example, running STRUCTURE on world-wide human data identifies a cluster that contains many European individuals, however, on the basis of ancient DNA we know that modern Europeans are a mixture of distinct ancestral groups.

2.3.4 Principal components analysis

Principal component analysis (PCA) is a common statistical approach to visualize high dimensional data, and used by many fields. The idea of PCA is to give a location to each individual data-point on each of a small number principal component axes. These PC axes are chosen to reflect major axes of variation in the data, with the first PC being that which explains largest variance, the second the second most, and so on. The use of PCA in population genetics was pioneered by Cavalli-Sforza and colleagues and now with large genotyping datasets, PCA has made a comeback.¹¹

Consider a dataset consisting of N individuals at S biallelic SNPs. The i^{th} individual's genotype data at locus ℓ takes a value $g_{i,\ell} = 0, 1$, or 2 (corresponding to the number of copies of allele A_1 an individual carries at this SNP). We can think of this as a $N \times S$ matrix (where usually $N \ll S$).

Denoting the sample mean allele frequency at SNP ℓ by p_ℓ , it's common to standardize the genotype in the following way

$$\frac{g_{i,\ell} - 2p_\ell}{\sqrt{2p_\ell(1 - p_\ell)}} \quad (2.24)$$

i.e. at each SNP we center the genotypes by subtracting the mean genotype ($2p_\ell$) and divide through by the square root of the expected variance assuming that alleles are sampled binomially from the mean frequency ($\sqrt{2p_\ell(1 - p_\ell)}$). Doing this to all of our genotypes, we form a data matrix (of dimension $N \times S$). We can then perform principal component analysis of this data matrix to uncover the major axes of genotype variance in our sample. Figure 2.24 shows a PCA from LetterSpace=10BECQUET *et al.* (2007) using the same chimpanzee data as in Figure 2.23.

It is worth taking a moment to delve further into what we are doing here. There's a number of equivalent ways of thinking about

¹¹ LetterSpace=10MENOZZI, P., LetterSpace=10A. PIAZZA, and LetterSpace=10L. CAVALLI-SFORZA, 1978 Synthetic maps of human gene frequencies in Europeans. *Science* 201(4358): 786–792; and LetterSpace=10PATTERSON, N., LetterSpace=10A. L. PRICE, and LetterSpace=10D. REICH, 2006 Population structure and eigenanalysis. *PLoS genetics* 2(12): e190

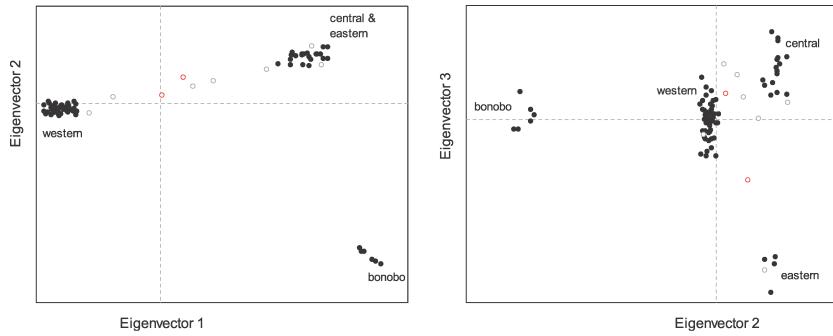


Figure 2.24: Principal Component Analysis by LetterSpace=10BECQUET *et al.* (2007) using the same chimpanzee data as in Figure 2.23. Here LetterSpace=10BECQUET *et al.* (2007) plot the location of each individual on the first two principal components (called eigenvectors) in the left panel, and on the second and third principal components (eigenvectors) in the right panel (licensed under CC BY 4.0). In the PCA, individuals identified as all of one ancestry by STRUCTURE cluster together by population (solid circles). While the nine individuals identified by STRUCTURE as hybrids (open circles) for the most part fall at intermediate locations in the PCA. There are two individuals (red open circles) reported as being of a particular population but that but appear to be hybrids.

what PCA is doing. One of these ways is to think that when we do PCA we are building the individual by individual covariance matrix and performing an eigenvalue decomposition of this matrix (with the eigenvectors being the PCs). This individual by individual covariance matrix has entries the $[i, j]$ given by

$$\frac{1}{S-1} \sum_{\ell=1}^S \frac{(g_{i,\ell} - 2p_\ell)(g_{j,\ell} - 2p_\ell)}{2p_\ell(1-p_\ell)} \quad (2.25)$$

Note that this is the covariance, and is very similar to those we encountered in discussing F -statistics as correlations (equation (2.17)), except now we are asking about the covariance between two individuals above that expected if they were both drawn from the total sample at random (rather than the covariance of alleles within a single individual). So by performing PCA on the data we are learning about the major (orthogonal) axes of the kinship matrix.

As an example of the application of PCA, let's consider the case of the putative ring species in the Greenish warbler (*Phylloscopus trochiloides*) species complex. This set of subspecies exists in a ring around the edge of the Himalayan plateau. LetterSpace=10ALCAIDE *et al.* (2014) collected 95 Greenish warbler samples from 22 sites around the ring, and the sampling locations are shown in Figure 2.25.

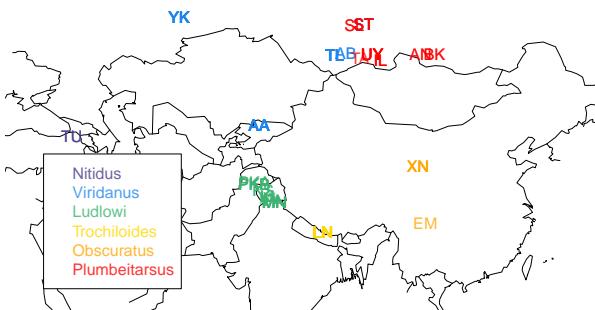


Figure 2.25: The sampling locations of 22 populations of Greenish warblers from LetterSpace=10ALCAIDE *et al.* (2014). The samples are coloured by the subspecies. Code here.

It is thought that these warblers spread from the south, northward in two different directions around the inhospitable Himalayan plateau, establishing populations along the western edge (green and blue populations) and the eastern edge (yellow and red populations). When they came into secondary contact in Siberia, they were reproductively isolated from one another, having evolved different songs and accumulated other reproductive barriers from each other as they spread independently north around the plateau, such that *P. t. viridanus* (blue) and *P. t. plumbeitarsus* (red) populations presently form a stable hybrid zone.

LetterSpace=10ALCAIDE *et al.* (2014) obtained sequence data for their samples at 2,334 snps. In Figure 2.27 you can see the matrix of kinship coefficients, using (2.25), between all pairs of samples. You can already see a lot about population structure in this matrix. Note how the red and yellow samples, thought to be derived from the Eastern route around the Himalayas, have higher kinship with each other, and blue and the (majority) of the green samples, from the Western route, form a similarly close group in terms of their higher kinship.

We can then perform PCA on this kinship matrix to identify the major axes of variation in the dataset. Figure 2.28 shows the samples plotted on the first two PCs. The two major routes of expansion clearly occupy different parts of PC space. The first principal component distinguishes populations running North to South along the western route of expansion, while the second principal component distinguishes among populations running North to South along the Eastern route of expansion. Thus genetic data supports the hypothesis that the Greenish warblers speciated as they moved around the Himalayan plateau. However, as noted by LetterSpace=10ALCAIDE



Figure 2.26: Greenish warbler, subsp. *viridanus* (*Phylloscopus trochiloides viridanus*).

Coloured figures of the birds of the British Islands. 1885.
Lilford T. L. P. Image from the Biodiversity Heritage Library.
Contributed by American Museum of Natural History Library.
Not in copyright. (Greenish warblers are rare visitors to the UK.)

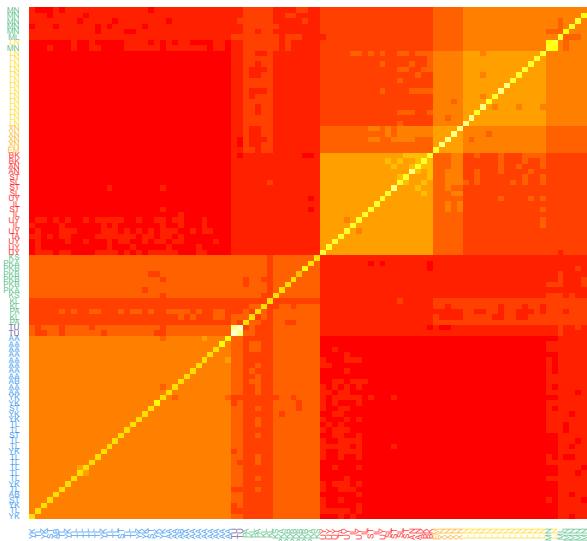


Figure 2.27: The matrix of kinship coefficients calculated for the 95 samples of Greenish warblers. Each cell in the matrix gives the pairwise kinship coefficient calculated for a particular pair. Hotter colours indicating higher kinship. The x and y labels of individuals are the population labels from Figure 2.25, and coloured by subspecies label as in that figure. The rows and columns have been organized to cluster individuals with high kinship. [Code here.](#)

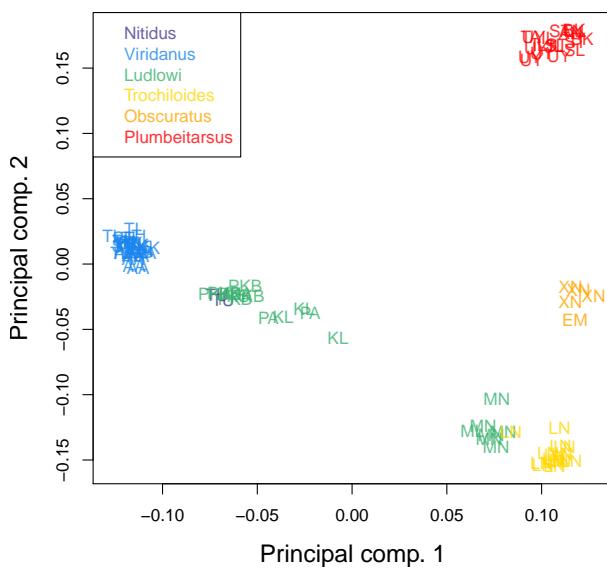


Figure 2.28: The 95 greenish warbler samples plotted on their locations on the first two principal components. The labels of individuals are the population labels from Figure 2.25, and coloured by subspecies label as in that figure. [Code here.](#)

et al. (2014), it also suggests additional complications to the traditional view of these warblers as an unbroken ring species, a case of speciation by continuous geographic isolation. The *Ludlowi* sub-species shows a significant genetic break, with the southern most MN samples clustering with the *Trochiloides* subspecies, in both the PCA and kinship matrix (Figures 2.28 and 2.27), despite being much more geographically close to the other *Ludlowi* samples. This suggests that genetic isolation is not just a result of geographic distance, and other biogeographic barriers must be considered in the case of this broken ring species.

Finally, while PCA is a wonderful tool for visualizing genetic data, care must be taken in its interpretation. The U-like shape in the case of the Greenish warbler PC might be consistent with some low level of gene flow between the red and the blue populations, pulling them genetically closer together and helping to form a genetic ring as well as a geographic ring. However, U-like shapes are expected to appear in PCAs even if our populations are just arrayed along a line, and more complex geometric arrangements of populations in PC space can result under simple geographic models (LetterSpace=10NOVEMBRE and LetterSpace=10STEPHENS, 2008). Inferring the geographical and population-genetic history of species requires the application of a range of tools; see LetterSpace=10ALCAIDE *et al.* (2014) and LetterSpace=10BRADBURD *et al.* (2016) for more discussion of the Greenish warblers.

2.3.5 Correlations between loci, linkage disequilibrium, and recombination

924

Up to now we have been interested in correlations between alleles
 926 at the same locus, e.g. correlations within individuals (inbreeding)
 or between individuals (relatedness). We have seen how relatedness
 928 between parents affects the extent to which their offspring is inbred.
 We now turn to correlations between alleles at different loci.

930 *Recombination* To understand correlations between loci we need
 to understand recombination a bit more carefully. Let us consider
 932 a heterozygous individual, containing AB and ab haplotypes. If no
 recombination occurs between our two loci in this individual, then
 934 these two haplotypes will be transmitted intact to the next genera-
 tion. While if a recombination (i.e. an odd number of crossing over
 936 events) occurs between the two parental haplotypes, then $1/2$ the time
 the child receives an Ab haplotype and $1/2$ the time the child receives
 938 an aB haplotype. Effectively, recombination breaks up the association
 between loci. We'll define the recombination fraction (r) to be the
 940 probability of an odd number of crossing over events between our
 loci in a single meiosis. In practice we'll often be interested in rela-
 942 tively short regions such that recombination is relatively rare, and so
 we might think that $r = r_{BP}L \ll \frac{1}{2}$, where r_{BP} is the average recombi-
 944 nation rate (in Morgans) per base pair (typically $\sim 10^{-8}$) and L is the
 number of base pairs separating our two loci.

946 *Linkage disequilibrium* The (horrible) phrase linkage disequilibrium
 (LD) refers to the statistical non-independence (i.e. a correlation)
 948 of alleles in a population at different loci. It's an awful name for
 a fantastically useful concept; LD is key to our understanding of
 950 diverse topics, from sexual selection and speciation to the limits of
 genome-wide association studies.

952 Our two biallelic loci, which segregate alleles A/a and B/b , have
 allele frequencies of p_A and p_B respectively. The frequency of the
 954 two locus haplotype AB is p_{AB} , and likewise for our other three
 combinations. If our loci were statistically independent then $p_{AB} =$
 956 $p_A p_B$, otherwise $p_{AB} \neq p_A p_B$. We can define a covariance between the
 A and B alleles at our two loci as

$$D_{AB} = p_{AB} - p_A p_B \quad (2.26)$$

958 and likewise for our other combinations at our two loci (D_{Ab} , D_{aB} , D_{ab}).
 Gametes with two similar case alleles (e.g. A and B, or a and b) are
 960 known as *coupling* gametes, and those with different case alleles are
 known as *repulsion* gametes (e.g. a and B, or A and b). Then, we can

962 think of D as measuring the *excess* of coupling to repulsion gametes.
 These D statistics are all closely related to each other as $D_{AB} = -D_{Ab}$
 964 and so on. Thus we only need to specify one D_{AB} to know them all,
 so we'll drop the subscript and just refer to D . Also a handy result is
 966 that we can rewrite our haplotype frequency p_{AB} as

$$p_{AB} = p_A p_B + D. \quad (2.27)$$

If $D = 0$ we'll say the two loci are in linkage equilibrium, while if
 968 $D > 0$ or $D < 0$ we'll say that the loci are in linkage disequilibrium
 (we'll perhaps want to test whether D is statistically different from
 970 0 before making this choice). You should be careful to keep the con-
 cepts of linkage and linkage disequilibrium separate in your mind.
 972 Genetic linkage refers to the linkage of multiple loci due to the fact
 that they are transmitted through meiosis together (most often be-
 974 cause the loci are on the same chromosome). Linkage disequilibrium
 merely refers to the covariance between the alleles at different loci;
 976 this may in part be due to the genetic linkage of these loci but does
 not necessarily imply this (e.g. genetically unlinked loci can be in LD
 978 due to population structure).

i

980 **Question 12.** You genotype 2 bi-allelic loci (A & B) segregat-
 ing in two mouse subspecies (1 & 2) which mate randomly among
 982 themselves, but have not historically interbreed since they speci-
 ated. On the basis of previous work you estimate that the two loci
 984 are separated by a recombination fraction of 0.1. The frequencies of
 haplotypes in each population are:

Pop	p_{AB}	p_{Ab}	p_{aB}	p_{ab}
1	.02	.18	.08	.72
2	.72	.18	.08	.02

- A) How much LD is there within species? (i.e. estimate D)
 988 B) If we mixed individuals from the two species together in equal
 proportions, we could form a new population with p_{AB} equal to the
 990 average frequency of p_{AB} across species 1 and 2. What value would
 D take in this new population before any mating has had the chance
 992 to occur?

Our linkage disequilibrium statistic D depends strongly on the
 994 allele frequencies of the two loci involved. One common way to par-
 tially remove this dependence, and make it more comparable across
 996 loci, is to divide D through by its maximum possible value given the
 frequency of the loci. This normalized statistic is called D' and varies
 998 between +1 and -1. In Figure 2.29 there's an example of LD across
 the TAP2 region in human and chimp. Notice how physically close

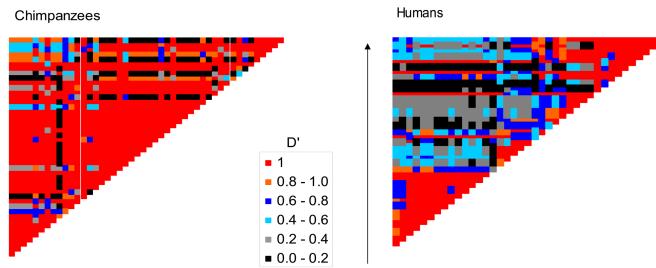


Figure 2.29: LD across the TAP2 gene region in a sample of Humans and Chimps, from LetterSpace=10PTAK *et al.* (2004), licensed under CC BY 4.0. The rows and columns are consecutive SNPs, with each cell giving the absolute D' value between a pair of SNPs. Note that these are different sets of SNPs in the two species, as shared polymorphisms are very rare.

1000 SNPs, i.e. those close to the diagonal, have higher absolute values of
 1002 D' as closely linked alleles are separated by recombination less often
 1004 allowing high levels of LD to accumulate. Over large physical dis-
 1006 tances, away from the diagonal, there is lower D' . This is especially
 1008 notable in humans as there is an intense, human-specific recombi-
 1010 nation hotspot in this region, which is breaking down LD between
 1012 opposite sides of this region.

Another common statistic for summarizing LD is r^2 which we

1008 write as

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \quad (2.28)$$

As D is a covariance, and $p_A(1-p_A)$ is the variance of an allele
 1010 drawn at random from locus A , r^2 is the squared correlation coeffi-
 1012 cient. Note that this r in r^2 is NOT the recombination fraction.

1014 Figure 2.31 shows r^2 for pairs of SNPs at various physical dis-
 1016 tances in two population samples of *Mus musculus domesticus*. Again
 1018 LD is highest between physically close markers as LD is being gener-
 1020 ated faster than it can decay via recombination; more distant markers
 have much lower LD as here recombination is winning out. Note the
 decay of LD is much slower in the advanced-generation cross popu-
 lation than in the natural wild-caught population. This persistence of
 LD across megabases is due to the limited number of generations for
 recombination since the cross was created.

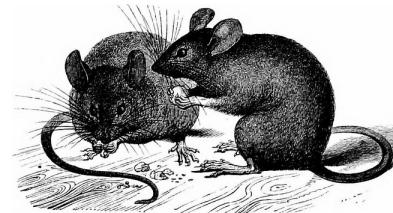
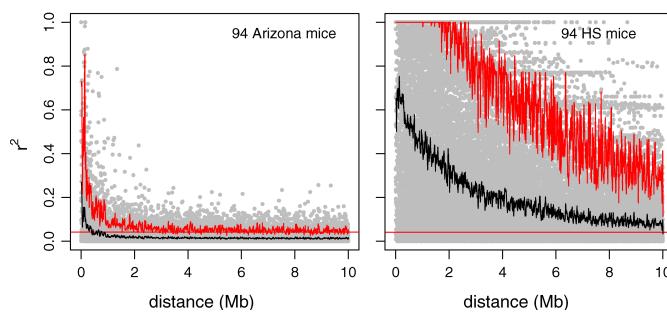


Figure 2.30: *Mus musculus*.
 A history of British quadrupeds, including the Cetacea. 1874.
 Bell T., Tomes, R. F.m Alston E. R. Image from the Biodiversity
 Heritage Library. Contributed by Cornell University Library.
 No known copyright restrictions.

Figure 2.31: The decay of LD for auto-
 somal SNPs in *Mus musculus domesticus*,
 as measured by r^2 , in a wild-caught
 mouse population from Arizona and
 a set of advanced-generation crosses
 between inbred lines of lab mice. Each
 dot gives the r^2 for a pair of SNPs a
 given physical distance apart, for a total
 of ~ 3000 SNPs. The solid black line
 gives the mean, the jagged red line the
 95th percentile, and the flat red line a
 cutoff for significant LD. From Letter-
 Space=10LAURIE *et al.* (2007), licensed
 under CC BY 4.0.

The generation of LD. Various population genetic forces can generate LD. Selection can generate LD by favouring particular combinations of alleles. Genetic drift will also generate LD, not because particular combinations of alleles are favoured, but simply because at random particular haplotypes can by chance drift up in frequency. Mixing between divergent populations can also generate LD, as we saw in the mouse question above.

The decay of LD due to recombination We will now examine what happens to LD over the generations if, in a very large population (i.e. no genetic drift and frequencies of our loci thus follow their expectations), we only allow recombination to occur. To do so, consider the frequency of our AB haplotype in the next generation, p'_{AB} . We lose a fraction r of our AB haplotypes to recombination ripping our alleles apart but gain a fraction $rp_A p_B$ per generation from other haplotypes recombining together to form AB haplotypes. Thus in the next generation

$$p'_{AB} = (1 - r)p_{AB} + rp_A p_B \quad (2.29)$$

The last term above, in eqn 2.29, is $r(p_{AB} + p_{Ab})(p_{AB} + p_{aB})$ simplified, which is the probability of recombination in the different diploid genotypes that could generate a p_{AB} haplotype.

We can then write the change in the frequency of the p_{AB} haplotype as

$$\Delta p_{AB} = p'_{AB} - p_{AB} = -rp_{AB} + rp_A p_B = -rD \quad (2.30)$$

So recombination will cause a decrease in the frequency of p_{AB} if there is an excess of AB haplotypes within the population ($D > 0$), and an increase if there is a deficit of AB haplotypes within the population ($D < 0$). Our LD in the next generation is

$$\begin{aligned} D' &= p'_{AB} - p'_A p'_B \\ &= (p_{AB} + \Delta p_{AB}) - (p_A + \Delta p_A)(p_B + \Delta p_B) \\ &= p_{AB} + \Delta p_{AB} - p_A p_B \\ &= (1 - r)D \end{aligned} \quad (2.31)$$

where we can cancel out Δp_A and Δp_B above because recombination only changes haplotype, not allele, frequencies. So if the level of LD in generation 0 is D_0 , the level t generations later (D_t) is

$$D_t = (1 - r)^t D_0 \quad (2.32)$$

Recombination is acting to decrease LD, and it does so geometrically at a rate given by $(1 - r)$. If $r \ll 1$ then we can approximate this by an exponential and say that

$$D_t \approx D_0 e^{-rt} \quad (2.33)$$

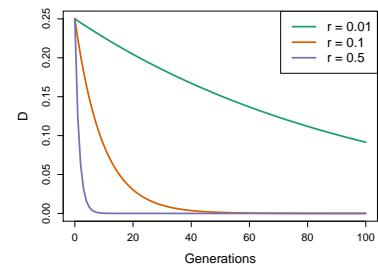


Figure 2.32: The decay of LD from an initial value of $D_0 = 0.25$ over time (Generations) for a pair of loci a recombination fraction r apart. Code here.

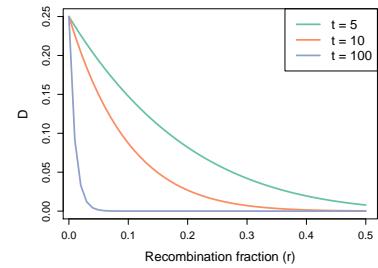


Figure 2.33: The decay of LD from an initial value of $D_0 = 0.25$ due to recombination over t generations, plotted across possible recombination fractions (r) between our pair of loci. Code here.

1048 **Question 13.** You find a hybrid population between the two
 mouse subspecies described in the question above, which appears
 1050 to be comprised of equal proportions of ancestry from the two sub-
 species. You estimate LD between the two markers to be 0.0723.
 1052 Assuming that this hybrid population is large and was formed by a
 single mixture event, can you estimate how long ago this population
 1054 formed?

1056 A particularly striking example of the decay of LD generated by
 the mixing of populations is offered by the LD created by the in-
 terbreeding between humans and Neanderthals. Neanderthals and
 1058 modern Humans diverged from each other likely over half a million
 years ago, allowing time for allele frequency differences to accumu-
 1060 late between the Neanderthal and modern human populations. The
 two populations spread back into secondary contact when humans
 moved out of Africa over the past hundred thousand years or so. One
 1062 of the most exciting findings from the sequencing of the Neanderthal
 genome was that modern-day people with Eurasian ancestry carry a
 few percent of their genome derived from the Neanderthal genome,
 via interbreeding during this secondary contact. To date the timing of
 1064 this interbreeding, LetterSpace=10SANKARARAMAN *et al.* (2012) looked
 at the LD in modern humans between pairs of alleles found to be de-
 rived from the Neanderthal genome (and nearly absent from African
 1066 populations). In Figure 2.35 we show the average LD between these
 1068 populations. In Figure 2.35 we show the average LD between these
 1070 loci as a function of the genetic distance (r) between them, from the
 work of LetterSpace=10SANKARARAMAN *et al.*.

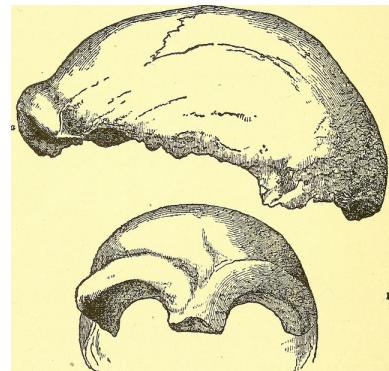
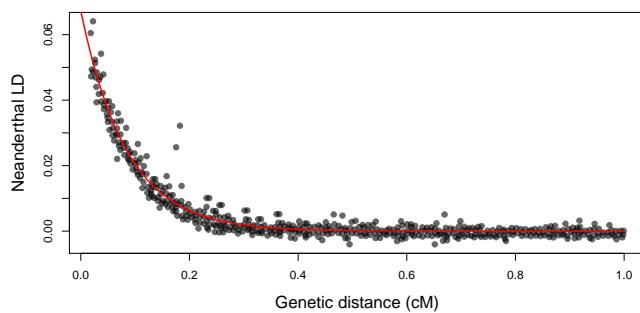


Figure 2.34: The earliest discovered fossil of a Neanderthal, fragments of a skull found in a cave in the Neander Valley in Germany.

Man's place in nature. 1890. Huxley, T. H. Image from the Internet Archive. Contributed by The Library of Congress. No known copyright restrictions.



1072 Assuming a recombination rate r , we can fit the exponential decay
 1074 of LD predicted by eqn. (2.33) to the data points in this figure; the fit
 is shown as a red line. Doing this we estimate $t = 1200$ generations,
 1076 or about 35 thousand years (using a human generation time of 29
 years). Thus the LD in modern Eurasians, between alleles derived
 1078 from the interbreeding with Neanderthals, represents over thirty

Figure 2.35: The LD between putative-Neanderthal alleles in a modern European population (the CEU sample from the 1000 Genomes Project). Each point represents the average D statistic between a pair of alleles at loci at a given genetic distance apart (as given on the x-axis and measured in centiMorgans (cM)). The putative Neanderthal alleles are alleles where the Neanderthal genome has a derived allele that is at very low frequency in a modern-human West African population sample (thought to have little admixture from Neanderthals). The red line is the fit of an exponential decay of LD, using non-linear least squared (nls in R).

thousand years of recombination slowly breaking down these old
1080 associations.¹²

¹² The calculation done by LetterSpace=10SANKARARAMAN *et al.* (2012) is actually a bit more involved as they account for inhomogeneity in recombination rates and arrive at a date of 47,334 – 63,146 years.

3

1082 Genetic Drift and Neutral Diversity

LetterSpace=10RANDOMNESS IS INHERENT TO EVOLUTION, from the
1084 lucky birds blown of course to colonize some new oceanic island, to
which mutations arise first in the HIV strain infecting an individual
1086 taking anti-retroviral drugs. One major source of stochasticity in
evolutionary biology is genetic drift. Genetic drift occurs because
1088 more or less copies of an allele by chance can be transmitted to the
next generation. This can occur because, by chance, the individuals
1090 carrying a particular allele can leave more or less offspring in the
next generation. In a sexual population, genetic drift also occurs
1092 because Mendelian transmission means that only one of the two
alleles in an individual, chosen at random at a locus, is transmitted to
1094 the offspring.

Genetic drift can play a role in the dynamics of all alleles in all
1096 populations, but it will play the biggest role for neutral alleles. A
neutral polymorphism occurs when the segregating alleles at a poly-
1098 morphic site have no discernible differences in their effect on fitness.
We'll make clear what we mean by "discernible" later, but for the
1100 moment think of this as "no effect" on fitness.

The neutral theory of molecular evolution. The role of genetic drift
1102 in molecular evolution has been hotly debated since the 60s when
the Neutral theory of molecular evolution was proposed (see LetterSpace=10OHTA and LetterSpace=10GILLESPIE, 1996, for a history)¹. The central premise of Neutral theory theory is that patterns
1104 of molecular polymorphism within species and substitution between
species can be well understood by supposing that the vast majority of
1106 these molecular polymorphisms and substitutions were neutral alleles,
whose dynamics were just subject to the vagaries of genetic drift
1108 and mutation. Early proponents of this view suggested that the vast
majority of new mutations are either neutral or highly deleterious
1110 (e.g. mutations that disrupt important protein functions). This latter
1112

¹ LetterSpace=10KIMURA, M., 1968 Evolutionary rate at the molecular level. *Nature* 217(5129): 624–626; LetterSpace=10KING, J. L. and LetterSpace=10T. H. JUKES, 1969 Non-darwinian evolution. *Science* 164(3881): 788–798; and LetterSpace=10KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press

class of mutations are too deleterious to contribute much to common
 1114 polymorphisms or substitutions between species, because they are
 quickly weeded out of the population by selection.

1116 Neutral theory can sound strange given that much of the time our
 first brush with evolution often focuses on adaptation and pheno-
 1118 typic evolution. However, proponents of this world-view didn't deny
 the existence of advantageous mutations, they simply thought that
 1120 beneficial mutations are rare enough that their contribution to the
 bulk of polymorphism or divergence can be largely ignored. They
 1122 also often thought that much of phenotypic evolution may well be
 adaptive, but again the loci responsible for these phenotypes are a
 1124 small fraction of all the molecular change that occur. The neutral
 theory of molecular evolution was originally proposed to explain
 1126 protein polymorphism. However, we can apply it more broadly to
 think about neutral evolution genome-wide. With that in mind, what
 1128 types of molecular changes could be neutral? Perhaps:

1. Changes in non-coding DNA that don't disrupt regulatory se-
 1130 quences. For example, in the human genome only about 2% of
 the genome codes for proteins. The rest is mostly made up of old
 1132 transposable element and retrovirus insertions, repeats, pseudo-
 genes, and general genomic clutter. Current estimates suggest
 1134 that, even counting conserved, functional, non-coding regions,
 less than 10% of our genome is subject to evolutionary constraint
 1136 (LetterSpace=10RANDS *et al.*, 2014).

2. Synonymous changes in coding regions, i.e. those that don't
 1138 change the amino-acid encoded by a codon.
 3. Non-synonymous changes that don't have a strong effect on the
 1140 functional properties of the amino acid encoded, e.g. changes that
 don't change the size, charge, or hydrophobic properties of the
 1142 amino acid too much.
 4. An amino-acid change with phenotypic consequences, but little
 1144 relevance to fitness, e.g. a mutation that causes your ears to be a
 slightly different shape, or that prevents an organism from living
 1146 past 50 in a species where most individuals reproduce and die by
 their 20s.

1148 There are counter examples to all of these ideas, e.g. synonymous
 changes can affect the translation speed and accuracy of proteins and
 1150 so are subject to selection. However, the list above hopefully con-
 vinces you that the general thinking that some portion of molecular
 1152 change may not be subject to selection isn't as daft as it may have
 initially sounded.

1154 Various features of molecular polymorphism and divergence have
 1155 been viewed as consistent with the neutral theory of molecular evo-
 1156 lution. The two we'll focus on in this chapter are the high level of
 1158 molecular polymorphism in many species (see for example Figure
 1160 2.2) and the molecular clock. We'll see that various aspects of the
 1162 original neutral theory have merit in describing some features and
 1164 types of molecular change, but we'll also see that it is demonstrably
 1166 wrong in some cases. We'll also see the primary utility of the neu-
 1168 tral theory isn't whether it is right or wrong, but that it serves as a
 1170 simple null model that can be tested and in some cases rejected, and
 1172 subsequently built on. The broader debate currently in the field of
 1174 molecular evolution is the balance of neutral, adaptive, and deleteri-
 1176 ous changes that drive different types of evolutionary change.

3.1 Loss of heterozygosity due to drift.

1168 Genetic drift will, in the absence of new mutations, slowly purge our
 1169 population of neutral genetic diversity, as alleles slowly drift to high
 1170 or low frequencies and are lost or fixed over time.

1172 Imagine a randomly mating population of a constant size N
 1173 diploid individuals, and that we are examining a locus segregating
 1174 for two alleles that are neutral with respect to each other. This pop-
 1176 ulation is randomly mating with respect to the alleles at this locus.
 See Figures 3.1 and 3.2 to see how genetic drift proceeds, by tracking
 1178 alleles within a small population.

1178 In generation t our current level of heterozygosity is H_t , i.e. the
 1180 probability that two randomly sampled alleles in generation t are
 non-identical is H_t . Assuming that the mutation rate is zero (or van-
 1182 ishingely small), what is our level of heterozygosity in generation
 $t + 1$?

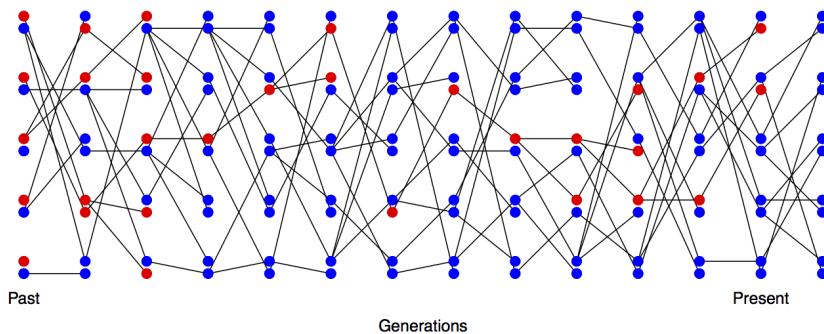


Figure 3.1: Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals over the generations, with lines showing transmission. In the first generation every individual is a heterozygote. Code here.

1182 In the next generation ($t + 1$) we are looking at the alleles in the
 1183 offspring of generation t . If we randomly sample two alleles in gener-

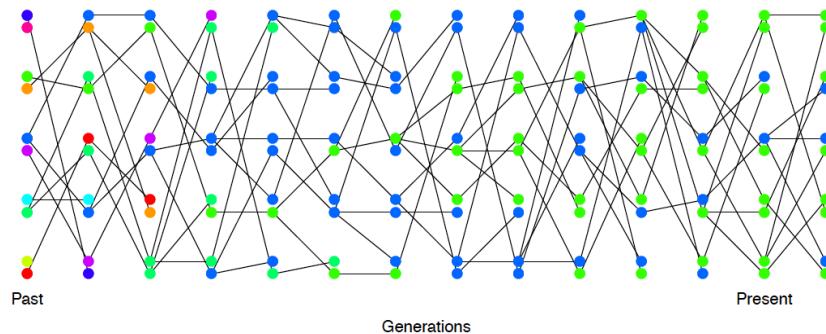


Figure 3.2: Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals. In the first generation I colour every allele a different colour so we can track their descendants. Code here.

¹¹⁸⁴ ation $t + 1$ which had different parental alleles in generation t , that is just like drawing two random alleles from generation t . So the probability that these two alleles in generation $t + 1$, that have different parental alleles in generation t , are non-identical is H_t .

¹¹⁸⁸ Conversely, if the two alleles in our pair had the same parental allele in the proceeding generation (i.e. the alleles are identical by ¹¹⁹⁰ descent one generation back) then these two alleles must be identical (as we are not allowing for any mutation).

¹¹⁹² In a diploid population of size N individuals there are $2N$ alleles. The probability that our two alleles have the same parental allele in ¹¹⁹⁴ the proceeding generation is $1/(2N)$ and the probability that they have different parental alleles is $1 - 1/(2N)$. So by the above argument, ¹¹⁹⁶ the expected heterozygosity in generation $t + 1$ is

$$H_{t+1} = \frac{1}{2N} \times 0 + \left(1 - \frac{1}{2N}\right) H_t \quad (3.1)$$

¹¹⁹⁸ Thus, if the heterozygosity in generation 0 is H_0 , our expected heterozygosity in generation t is

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \quad (3.2)$$

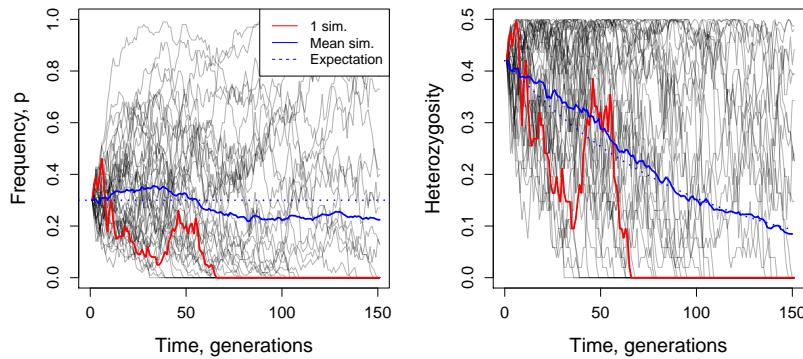
¹²⁰⁰ i.e. the expected heterozygosity within our population is decaying geometrically with each passing generation. If we assume that $1/(2N) \ll 1$ then we can approximate this geometric decay by an ¹²⁰² exponential decay (see Question 2 below), such that

$$H_t = H_0 e^{-t/(2N)} \quad (3.3)$$

i.e. heterozygosity decays exponentially at a rate $1/(2N)$.

¹²⁰⁴ In Figure 3.3 we show trajectories through time for 40 independently simulated loci drifting in a population of 50 individuals. Each ¹²⁰⁶ population was started from a frequency of 30%. Some drift up and some drift down, eventually being lost or fixed from the population, but, on average across simulations, the allele frequency doesn't

change. We also track heterozygosity, you can see that heterozygosity sometimes goes up, and sometimes goes down, but on average we are losing heterozygosity, and this rate of loss is well predicted by eqn. (3.2).



1212

Question 1. You are in charge of maintaining a population of delta smelt in the Sacramento river delta. Using a large set of microsatellites you estimate that the mean level of heterozygosity in this population is 0.005. You set yourself a goal of maintaining a level of heterozygosity of at least 0.0049 for the next two hundred years. Assuming that the smelt have a generation time of 3 years, and that only genetic drift affects these loci, what is the smallest fully outbreeding population that you would need to maintain to meet this goal?

1222 Note how this picture of decreasing heterozygosity stands in contrast to the consistency of Hardy-Weinberg equilibrium from the previous chapter. However, our Hardy-Weinberg *proportions* still hold in forming each new generation. As the offspring genotypes in the next generation ($t + 1$) represent a random draw from the previous generation (t), if the parental frequency is p_t , we *expect* a proportion $2p_t(1 - p_t)$ of our offspring to be heterozygotes (and HW proportions for our homozygotes). However, because population size is finite, the observed genotype frequencies in the offspring will (likely) not match exactly with our expectations. As our genotype frequencies likely change slightly due to sampling, biologically this reflects random variation in family size and Mendelian segregation, the allele frequency will change. Therefore, while each generation represents a sample from Hardy-Weinberg proportions based on the generation before, our genotype proportions are not at an equilibrium (an unchanging state) as the underlying allele frequency changes over the

Figure 3.3: Change in allele frequency and loss of heterozygosity over time for 40 replicates. Simulations of genetic drift in a diploid population of 50 individuals, in the absence of new mutations. We start 40 independent, biallelic loci each with an initial allele at 30% frequency. The left panel shows the allele frequency over time and the right panel shows the heterozygosity over time, with the mean decay matching eqn. (3.2). Code here.

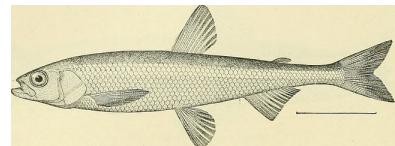


Figure 3.4: Pond smelt (*Hypomesus olidus*), a close relative of delta smelt. Bulletin of the United States Fish Commission, 1906. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

generations. We'll develop some mathematical models for these allele frequency changes later on. For now, we'll simply note that under our simple model of drift (formally the Wright-Fisher model), our allele count in the $t + 1^{th}$ generation represents a binomial sample (of size $2N$) from the population frequency p_t in the previous generation. If you've read to here, please email Prof Coop a picture of JBS Haldane in a striped suit with the title "I'm reading the chapter 3 notes". (It's well worth googling JBS Haldane and to read more about his life; he's a true character and one of the last great polymaths.)

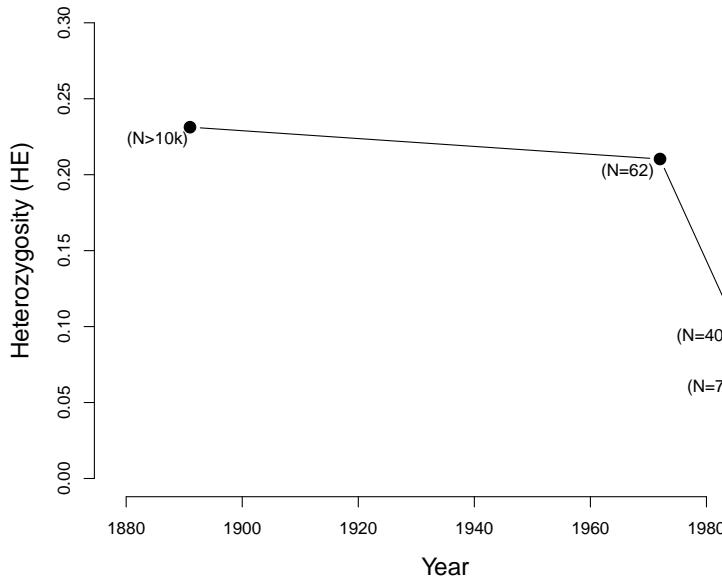


Figure 3.6: Loss of heterozygosity in the Black-footed Ferrets in their declining population. Numbers in brackets give estimated number of individuals alive at that time. Data from LetterSpace=10 WISELY *et al.* (2002). Code here.

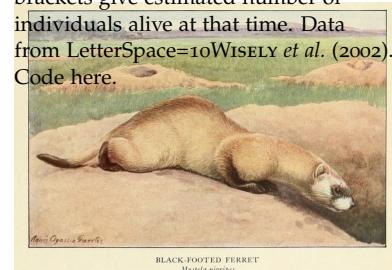


Figure 3.5: The black-footed ferret (*M. nigripes*). Wild animals of North America, The National geographical society, 1918. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

To see how a decline in population size can affect levels of heterozygosity, let's consider the case of black-footed ferrets (*Mustela nigripes*). The black-footed ferret population has declined dramatically through the twentieth century due to destruction of their habitat. In 1979, when the last known black-footed ferret died in captivity, they were thought to be extinct. In 1981, a very small wild population was rediscovered (40 individuals), but in 1985 this population suffered a number of disease outbreaks. All of the 18 remaining wild individuals were brought into captivity, 7 of which reproduced. Thanks to intense captive breeding efforts and conservation work, a wild population of over 300 individuals has been established since. However, because all of these individuals are descended from those 7 individuals who survived the bottleneck, diversity levels remain low.

LetterSpace=10 WISELY *et al.* measured heterozygosity at a number of

microsatellites in individuals from museum collections, showing the sharp drop in diversity as population sizes crashed (see Figure 3.6).

Question 2. In mathematical population genetics, a commonly used approximation is $(1 - x) \approx e^{-x}$ for $x \ll 1$ (formally, this follows from the Taylor series expansion of $\exp(-x)$, ignoring second order and higher terms of x). This approximation is especially useful for approximating a geometric decay process by an exponential decay process, e.g. $(1 - x)^t \approx e^{-xt}$. Using your calculator, or R, check how good of an approximation this is compared to the exact expression for two values of x , $x = 0.1$, and 0.01 , across two different values of t , $t = 5$ and $t = 50$. Briefly comment on your results.

3.1.1 Levels of diversity maintained by a balance between mutation and drift

Next we're going to consider the amount of neutral polymorphism that can be maintained in a population as a balance between genetic drift removing variation and mutation introducing new neutral variation, see Figure 3.7 for an example. Note in our example, how no single allele is maintained at a stable equilibrium, rather an equilibrium level of polymorphism is maintained by a constantly shifting set of alleles.

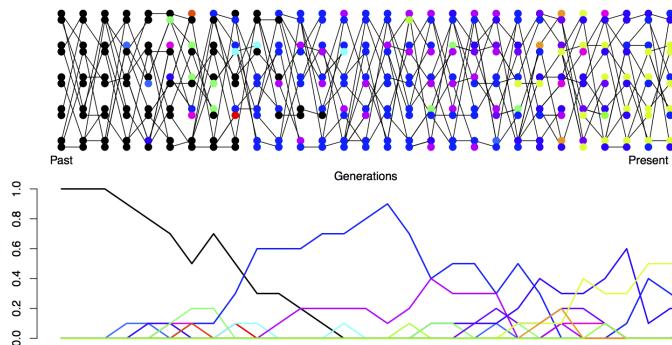


Figure 3.7: Mutation-drift balance. A diploid population of 5 individuals. In the first generation everyone has the same allele (black). Each generation the transmitted allele can mutate and we generate a new colour. In the bottom plot, I trace the frequency of alleles in our population over time. The mutation rate we use is very high, simply to maintain diversity in this small population. Code here.

The neutral mutation rate. We'll first want to consider the rate at which neutral mutations arise in the population. Thinking back to our discussion of the neutral theory of molecular evolution, let's suppose that there are only two classes of mutation that can arise in our genomic region of interest: neutral mutations and highly deleterious mutations. The total mutation rate at our locus is μ per generation, i.e. per transmission from parent to child. A fraction C of our mutations are new alleles that are highly deleterious and so quickly

removed from the population. We'll call this C parameter the constraint, and it will differ according to the genomic region we consider.
 1290 The remaining fraction $(1 - C)$ are our neutral mutations, such that
 1292 our neutral mutation rate is $(1 - C)\mu$. This is the per generation rate.

Question 3. It's worth taking a minute to get familiar with both
 1294 how rare, and how common, mutation is. The per base pair mutation
 1296 rate in humans is around 1.5×10^{-8} per generation. That means, on
 1298 average, we have to monitor a site for ~ 66.6 million transmissions
 from parent to child to see a mutation. Yet populations and genomes
 are big places, so mutations are common at these levels.

A) Your autosomal genome is ~ 3 billion base pairs long (3×10^9).
 1300 You have two copies, the one you received from your mum and one
 from your dad. What is the average (i.e. the expected) number of
 1302 mutations that occurred in the transmission from your mum and
 your dad to you?

B) The current human population size is ~ 7 billion individuals.
 1304 How many times, at the level of the entire human population, is a
 1306 single base-pair mutated in the transmission from one generation to
 the next?

1308 *Levels of heterozygosity maintained as a balance between mutation and selection.* Looking backwards in time from one generation to the
 1310 previous generation, we are going to say that two alleles which have
 1312 the same parental allele (i.e. find their common ancestor) in the preceding generation have *coalesced*, and refer to this event as a *coalescent event*.

1314 The probability that our pair of randomly sampled alleles have
 coalesced in the preceding generation is $1/(2N)$, and the probability
 1316 that our pair of alleles fail to coalesce is $1 - 1/(2N)$.

The probability that a mutation changes the identity of the transmitted allele is μ per generation. So the probability of no mutation occurring is $(1 - \mu)$. We'll assume that when a mutation occurs it
 1318 creates some new allelic type which is not present in the population. This assumption (commonly called the infinitely-many-alleles model) makes the math slightly cleaner, and also is not too bad an assumption biologically. See Figure 3.7 for a depiction of mutation-drift
 1320 balance in this model over the generations.

This model lets us calculate when our two alleles last shared a common ancestor and whether these alleles are identical as a result
 1326 of failing to mutate since this shared ancestor. For example, we can
 1328 work out the probability that our two randomly sampled alleles coalesce 2 generations in the past (i.e. they fail to coalesce in generation

¹³³⁰ 1 and then coalesce in generation 2), and that they are identical as

$$\left(1 - \frac{1}{2N}\right) \frac{1}{2N} (1 - \mu)^4 \quad (3.4)$$

Note the power of 4 is because our two alleles have to have failed to ¹³³² mutate through 2 meioses each.

More generally, the probability that our alleles coalesce in generation $t + 1$ (counting backwards in time) and are identical due to no mutation to either allele in the subsequent generations is ¹³³⁴

$$P(\text{coal. in } t+1 \text{ & no mutations}) = \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t (1 - \mu)^{2(t+1)} \quad (3.5)$$

¹³³⁶ To make this slightly easier on ourselves let's further assume that $t \approx t + 1$ and so rewrite this as:

$$P(\text{coal. in } t+1 \text{ & no mutations}) \approx \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t (1 - \mu)^{2t} \quad (3.6)$$

¹³³⁸ This gives us the approximate probability that two alleles will coalesce in the $(t + 1)^{\text{th}}$ generation. In general, we may not know ¹³⁴⁰ when two alleles may coalesce: they could coalesce in generation $t = 1, t = 2, \dots$, and so on. Thus, to calculate the probability that two ¹³⁴² alleles coalesce in *any* generation before mutating, we can write:

$$\begin{aligned} P(\text{coal. in any generation & no mutations}) &\approx P(\text{coal. in } t = 1 \text{ & no mutations}) + \\ &\quad P(\text{coal. in } t = 2 \text{ & no mutations}) + \dots \\ &= \sum_{t=1}^{\infty} P(\text{coal. in } t \text{ generations & no mutation}) \end{aligned}$$

¹³⁴⁴ which follows from basic probability and the fact that coalescing in a particular generation is mutually exclusive with coalescing in a different generation.

While we could calculate a value for this sum given N and μ , it's difficult to get a sense of what's going on with such a complicated expression. Here, we turn to a common approximation in population genetics (and all applied mathematics), where we assume that $1/(2N) \ll 1$ and $\mu \ll 1$. This allows us to approximate the geometric decay as an exponential decay. Then, the probability two alleles coalesce in generation $t + 1$ and don't mutate can be written as:

$$P(\text{coal. in } t+1 \text{ & no mutations}) \approx \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t (1 - \mu)^{2t} \quad (3.7)$$

$$\approx \frac{1}{2N} e^{-t/(2N)} e^{-2\mu t} \quad (3.8)$$

$$= \frac{1}{2N} e^{-t(2\mu + 1/(2N))} \quad (3.9)$$

¹³⁴⁶ Then we can approximate the summation by an integral, giving us:

$$\frac{1}{2N} \int_0^\infty e^{-t(2\mu+1/(2N))} dt = \frac{1/(2N)}{1/(2N) + 2\mu} = \frac{1}{1 + 4N\mu} \quad (3.10)$$

The equation above gives us the probability that our two alleles
¹³⁴⁸ coalesce at some point in time, and do not mutate before reaching
 their common ancestor. Equivalently, this can be thought of as the
¹³⁵⁰ probability our two alleles coalesce *before* mutating, i.e. that they are
 homozygous.

¹³⁵² Then, the complementary probability that our pair of alleles are
 non-identical (or heterozygous) is simply one minus this. The follow-
¹³⁵⁴ ing equation gives the equilibrium heterozygosity in a population at
 equilibrium between mutation and drift:

$$H = \frac{4N\mu}{1 + 4N\mu} \quad (3.11)$$

¹³⁵⁶ The compound parameter $4N\mu$, the population-scaled mutation rate,
 will come up a number of times so we'll give it its own name:

$$\theta = 4N\mu \quad (3.12)$$

¹³⁵⁸ So all else being equal, species with larger population sizes should
 have proportionally higher levels of neutral polymorphism.

¹³⁶⁰ **Question 4.** The sequence-level heterozygosity in *Capsella grandiflora* (grand shepherd's purse) is $\sim 2\%$ per base. Assuming a mu-
¹³⁶² tation rate of $10^{-9} bp^{-1}$ per generation, what is your estimate of the
 population size of *C. grandiflora*?

¹³⁶⁴ 3.1.2 The effective population size

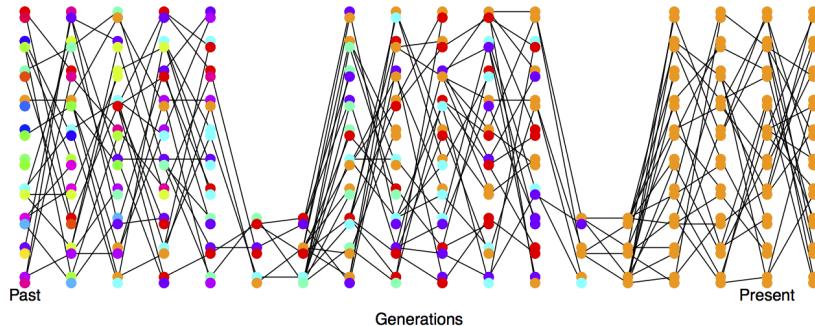
In practice, populations rarely conform to our assumptions of being
¹³⁶⁶ constant in size with low variance in reproductive success. Real pop-
 ulations experience dramatic fluctuations in size, and there is often
¹³⁶⁸ high variance in reproductive success. Thus rates of drift in natural
 populations are often a lot higher than the census population size
¹³⁷⁰ would imply. See Figure 3.8 for a depiction of a repeatedly bottle-
 necked population losing diversity at a fast rate.

To cope with this discrepancy, population geneticists often invoke
 the concept of an *effective population size* (N_e). In many situations (but
¹³⁷⁴ not all), departures from model assumptions can be captured by
 substituting N_e for N .

If population sizes vary rapidly in size, we can (if certain condi-
 tions are met) replace our population size by the harmonic mean
¹³⁷⁸ population size. Consider a diploid population of variable size,

This result was derived by LetterSpace=10KIMURA and LetterSpace=10CROW (1964) and LetterSpace=10MALÉCOT (1948) (see LetterSpace=10MALÉCOT, 1969, for an English translation, the lack of earlier translation meant this result was missed). Technically we're assuming that every new mutation creates a new allele, the so-called "infinitely many alleles" model, otherwise our pair of sequences could be identical due to repeat or back mutation. See this GENETICS blog post and LetterSpace=10EWENS (2016) for a nice discussion of the history.

the effective population size (N_e) is the population size that would result in the same rate of drift in an idealized population of constant size (following our modeling assumptions) as that observed in our true population .



whose size is N_t t generations into the past. The probability our pairs of alleles have not coalesced by generation t is given by

$$\prod_{i=1}^t \left(1 - \frac{1}{2N_i}\right) \quad (3.13)$$

Note that this simply collapses to our original expression $\left(1 - \frac{1}{2N}\right)^t$ if N_i is constant. Under this model, the rate of loss of heterozygosity in this population is equivalent to a population of effective size

$$N_e = \frac{1}{\frac{1}{t} \sum_{i=1}^t \frac{1}{N_i}}. \quad (3.14)$$

This is the harmonic mean of the varying population size.²

Thus our effective population size, the size of an idealized constant population which matches the rate of genetic drift, is the harmonic mean true population size over time. The harmonic mean is very strongly affected by small values, such that if our population size is one million 99% of the time but drops to 1000 every hundred or so generations, N_e will be much closer to 1000 than a million.

Figure 3.8: Loss of heterozygosity over time in a bottlenecking population. A diploid population of 10 individuals, that bottlenecks down to three individuals repeatedly. In the first generation, I colour every allele a different colour so we can track their descendants. There are no new mutations. Code here.

² To see this, note that if $1/(N_i)$ is small, then we can approximate (3.13) using the exponential approximation:

$$\prod_{i=1}^t \exp\left(-\frac{1}{2N_i}\right) = \exp\left(-\sum_{i=1}^t \frac{1}{2N_i}\right). \quad (3.15)$$

When we put the product inside the exponent, it becomes a sum. We can also write the probability of not coalescing by generation t in a population of constant size (N_e) as an exponential, so that it takes the same form as the expression above on the right. Comparing the exponent in the two cases, we see

$$\frac{t}{2N_e} = \sum_{i=1}^t 1/(2N_i) \quad (3.16)$$

So that if we want a constant effective population size (N_e) that has the same rate of loss of heterozygosity as our variable population, we need to rearrange and solve this equation to give (3.14).

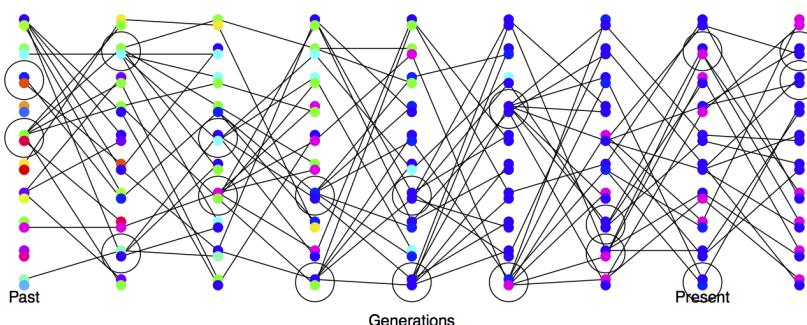


Figure 3.9: High variance on reproductive success increases the rate of genetic drift. A diploid population of 10 individuals, where the circled individuals have much higher reproductive success. In the first generation I colour every allele a different colour so we can track their descendants, there are no new mutations. Code here.

Variance in reproductive success will also affect our effective population size. Even if our population has a large constant size N individuals, if only small proportion of them get to reproduce, then

¹³⁹⁴ the rate of drift will reflect this much smaller number of reproducing
¹³⁹⁵ individuals. See Figure 3.9 for a depiction of the higher rate of drift
¹³⁹⁶ in a population where there is high variance in reproductive success.

To see one example of this, consider the case where N_F of females
¹³⁹⁸ get to reproduce and N_M males get reproduce. While every individual
¹⁴⁰⁰ has a mother and a father, not every individual gets to be a parent.
¹⁴⁰² In practice, in many animal species far more females get to repro-
¹⁴⁰⁴ duce than males, i.e. $N_M < N_F$, as a few males get many mating
¹⁴⁰⁶ opportunities and many males get no/few mating opportunities (see
¹⁴⁰⁸ LetterSpace=10JANICKE *et al.*, 2016, for a broad analysis, and note that
¹⁴¹⁰ there are certainly many exceptions to this general pattern). When our
¹⁴¹² two alleles pick an ancestor, 25% of the time our alleles were both
¹⁴¹⁴ in a female ancestor, in which case they are IBD with probability
¹⁴¹⁶ $1/(2N_F)$, and 25% of the time they are both in a male ancestor, in
¹⁴¹⁸ which case they coalesce with probability $1/(2N_M)$. The remaining
¹⁴²⁰ 50% of the time, our alleles trace back to two individuals of different
¹⁴²² sexes in the prior generation and so cannot coalesce. Therefore, our
¹⁴²⁴ probability of coalescence in the preceding generation is

$$\frac{1}{4} \left(\frac{1}{2N_M} \right) + \frac{1}{4} \left(\frac{1}{2N_F} \right) \quad (3.17)$$

¹⁴²⁶ i.e. the rate of coalescence is the harmonic mean of the two sexes'
¹⁴²⁸ population sizes, equating this to $\frac{1}{2N_e}$ we find

$$N_e = \frac{4N_F N_M}{N_F + N_M} \quad (3.18)$$

¹⁴³⁰ Thus if reproductive success is very skewed in one sex (e.g. $N_M \ll$
¹⁴³² $N_F/2$), our effective population size will be much reduced as a re-
¹⁴³⁴ sult. For more on how different evolutionary forces affect the rate of
¹⁴³⁶ genetic drift, and their impact on the effective population size, see
¹⁴³⁸ LetterSpace=10CHARLESWORTH (2009).

Question 5. You are studying a population of 500 male and 500
¹⁴⁴⁰ female Hamadryas baboons. Assume that all of the females but only
¹⁴⁴² 1/10 of the males get to mate: **A)** What is the effective population
¹⁴⁴⁴ size for the autosome?
¹⁴⁴⁶ **B)** Do you expect the *ratio* of X-chromosome to autosomal diver-
¹⁴⁴⁸ sity to be higher or lower in this species compared to a species where
¹⁴⁵⁰ the sexes have more similar variance in reproductive success? Explain
¹⁴⁵² the intuition behind your answer.

3.2 The Coalescent and patterns of neutral diversity

¹⁴⁵⁴ "Life can only be understood backwards; but it must be lived for-
¹⁴⁵⁶ wards" – Kierkegaard



Figure 3.10: Male Hamadryas baboons.
¹⁴⁵⁸ Up to ten females live in a harem with a single male.
¹⁴⁶⁰ Brehm's Tierleben (Brehm's animal life). Brehm, A.E. 1893.
¹⁴⁶² Image from the Biodiversity Heritage Library. Contributed by University of Illinois Urbana-Champaign. Not in copyright.

1430 *Pairwise Coalescent time distribution and the number of pairwise differences.* Thinking back to our calculations we made about the loss of
 1432 neutral heterozygosity and equilibrium levels of diversity (in Sections
 3.1 and 3.1.1), you'll note that we could first specify which generation
 1434 a pair of sequences coalesce in, and then calculate some properties
 of heterozygosity based on that. That's because neutral mutations do
 1436 not affect the probability that an individual transmits an allele, and
 so don't affect the way in which we can trace ancestral lineages back
 1438 through the generations.

As such, it will often be helpful to consider the time to the common ancestor of a pair of sequences, and then think of the impact of that time to coalescence on patterns of diversity. See Figure 3.11 for an example of this.

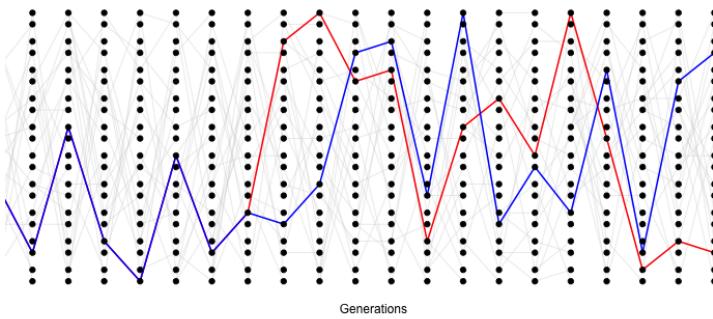


Figure 3.11: A simple demonstration of the coalescent process. The simulation consists of a diploid population of 10 individuals (20 alleles). In each generation, each individual is equally likely to be the parent of an offspring (and the allele transmitted is indicated by a light grey line). We track a pair of alleles, chosen in the present day, back 14 generations until they find a common ancestor. Code here.

The probability that a pair of alleles have failed to coalesce in t
 1444 generations and then coalesce in the $t + 1$ generation back is

$$P(T_2 = t + 1) = \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t \quad (3.19)$$

Thus the coalescent time of our pair of alleles is a Geometrically distributed random variable, where the probability of success is $1/(2N)$; we denote this by $T_2 \sim \text{Geo}(1/(2N))$. The expected (i.e. the mean
 1446 over many replicates) coalescent time of a pair of alleles is then
 1448

$$\mathbb{E}(T_2) = 2N \quad (3.20)$$

generations.

Conditional on a pair of alleles coalescing t generations ago, there are $2t$ generations in which a mutation could occur. If the per generation mutation rate is μ , then the expected number of mutations between a pair of alleles coalescing t generations ago is $2t\mu$ (the alleles

Using our exponential approximation, we can see that 3.19 is

$$\approx \frac{1}{2N} e^{-t/(2N)} \quad (3.21)$$

and so think of a continuous random variable, i.e. we could say that the coalescent time of a pair of sequences (T_2) is approximately exponentially distributed with a rate $1/(2N)$, i.e. $T_2 \sim \text{Exp}(1/(2N))$. Formally we can do this by taking the limit of the discrete process more carefully.

have gone through a total of $2t$ meioses since they last shared a common ancestor). So we can write the expected number of mutations (S_2) separating two alleles drawn at random from the population as

$$\begin{aligned}\mathbb{E}(S_2) &= \sum_{t=0}^{\infty} \mathbb{E}(S_2|T_2 = t)P(T_2 = t) \\ &= \sum_{t=0}^{\infty} 2\mu t P(T_2 = t) \\ &= 2\mu \mathbb{E}(T_2) \\ &= 4\mu N\end{aligned}\tag{3.22}$$

We'll assume that mutation is rare enough that it never happens at the same basepair twice, i.e. no multiple hits, such that we get to see all of the mutation events that separate our pair of sequences.³ Thus the number of mutations between a pair of sites is the observed number of differences between a pair of sequences. In the previous chapter we denote the observed number of pairwise differences at putatively neutral sites separating a pair of sequences as π (we usually average this over a number of pairs of sequences for a region). Therefore, under our simple, neutral, constant population-size model we expect

$$\mathbb{E}(\pi) = 4N\mu = \theta\tag{3.23}$$

So we can get an empirical estimate of θ from π , let's call this $\hat{\theta}_\pi$, by setting $\hat{\theta}_\pi = \pi$, i.e. our observed level of pairwise genetic diversity. If we have an independent estimate of μ , then from setting $\pi = \hat{\theta}_\pi = 4N\mu$ we can furthermore obtain an estimate of the population size N that is consistent with our levels of neutral polymorphism. If we estimate the population size this way, we should call it the effective coalescent population size (N_e). It's best to think about N_e estimated from neutral diversity as a long-term effective population size for the species, but there are many caveats that come along with that assumption. For example, past bottlenecks and population expansions are all subsumed into a single number and so this estimated N_e may not be very representative of the population size at any time. That said, it's not a bad place to start when thinking about the rate of genetic drift for neutral diversity in our population over long time-periods.⁴

Lets take a moment to distinguish our expected heterozygosity (eqn. 3.11) from our expected number of pairwise differences (π). Our expected heterozygosity is the probability that two alleles at a locus, sampled from a population at random, are different from each other. If one or more mutations have occurred since a pair of alleles last shared a common ancestor, then our sequences will be different from each other. On the other hand, our π measure keeps track of the

³ This is called the infinitely-many-sites assumption, which should be fine if $N\mu_{BP} \ll 1$, where μ_{BP} is the mutation rate per base pair).

⁴ Up to this point we've been describing only neutral processes, however, selection can also alter levels of polymorphism. For example, if some synonymous sites directly experience selection, then even if we use π calculated for synonymous changes we may underestimate the coalescent effective population size. As we'll see later in the notes, selection at linked sites can also impact neutral diversity. As such, if we can, we may want to use genomic sites subject to the weakest selective constraints, and also far from gene-dense or otherwise very constrained regions of the genome, to estimate N_e from π . But even then caution is warranted.

average total number of differences between our loci. As such, π is often a more useful measure, as it records the number of differences between the sequences, not just whether they are different from each other (however, for certain types of loci, e.g. microsatellites, heterozygosity is often used as we cannot usually count up the minimum number of mutations in a sensible way). In the case where our locus is a single basepair, the two measures will usually be close to one another, as $H \approx \theta$ for small values of θ . For example, comparing two sequences at random in humans, $\pi \approx 1/1000$ per basepair, and the probability that a specific base pair differs between two sequences is $\approx 1/1000$. However, these two quantities start to differ from each other when we consider regions with higher mutation rates. For example, if we consider a 10kb region, our mutation rate will 10,000 times larger than a single base pair. For this length of sequence the probability that two randomly chosen haplotypes differ is quite different from the number of mutational differences between them. (Try a mutation rate of 10^{-8} per base and a population size of 10,000 in our calculations of $E[\pi]$ and H to see this.)

Question 6. LetterSpace=10ROBINSON *et al.* (2016) found that the endangered Californian Channel Island fox on San Nicolas had very low levels of diversity ($\pi = 0.000014\text{bp}^{-1}$) compared to its close relative the California mainland gray fox (0.0012bp^{-1}).

A) Assuming a mutation rate of 2×10^{-8} per bp, what effective population sizes do you estimate for these two populations?

B) Why is the effective population size of the Channel Island fox so low? [Hint: quickly google Channel island foxes to read up on their history, also to see how ridiculously cute they are.]

Question 7. In your own words describe why the coalescent time of a pair of lineages scales linearly with the (effective) population size.

More details on the pairwise coalescent and the randomness of mutation.

We've derived the expected number of differences between a pair of sequences and talked about how variable the coalescent time is for a pair of sequences. The mutation process is also very variable; even if two sequences coalesce in the very distant past by chance, they may still be identical in the present if there was no mutation during that time.

Conditional on the coalescent time t , the probability that our pair of alleles are separated by S_2 mutations since they last shared a common ancestor is

$$P(S_2 | T_2 = t) = \binom{2t}{j} \mu^j (1 - \mu)^{2t-j} \quad (3.24)$$



Figure 3.12: Gray Fox, *Urocyon cinereoargenteus*.

Diseases and enemies of poultry. Pearson and Warren. (1897)
Image from the Biodiversity Heritage Library. Contributed by University of California Libraries. Not in copyright.

i.e. mutations happen in j generations and do not happen in $2t - j$ generations (with $\binom{2t}{j}$ ways this combination of events can possibly happen). Assuming that $\mu \ll 1$ and that $2t - j \approx 2t$, then we can approximate the probability that we have S_2 mutations as a Poisson distribution:

$$P(S_2|T_2 = t) = \frac{(2\mu t)^j e^{-2\mu t}}{j!} \quad (3.25)$$

i.e. a Poisson with mean $2\mu t$. We'll not make much use of this result, but it is very useful in thinking about how to simulate the process of mutation.

3.3 The coalescent process of a sample of alleles.

Usually we are not just interested in pairs of alleles, or the average pairwise diversity. Generally we are interested in the properties of diversity in samples of a number of alleles drawn from the population. Instead of just following a pair of lineages back until they coalesce, we can follow the history of a sample of alleles back through the population.

Consider first sampling three alleles at random from the population. The probability that all three alleles choose exactly the same ancestral allele one generation back is $1/(2N)^2$. If N is reasonably large, then this is a very small probability. As such, it is very unlikely that our three alleles coalesce all at once, and in a moment we'll see that it is safe to ignore such unlikely events.

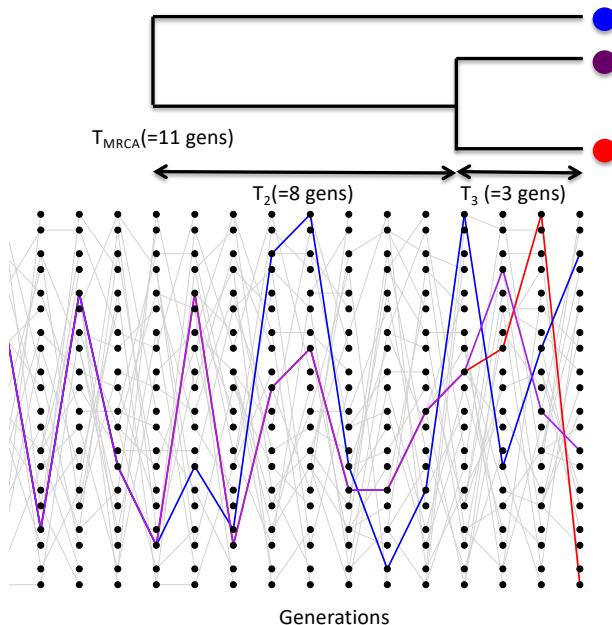


Figure 3.13: A simple simulation of the coalescent process for three lineages. We track the ancestry of three modern-day alleles, the first pair (blue and purple) coalesce four generations back, after which there are only two independent lineages we are tracking. This pair then coalesces twelve generations in the past. Note that different random realizations of this process will differ from each other a lot. The T_{MRCA} is $T_3 + T_2$. The total time in the tree is $T_{tot} = 3T_3 + 2T_2 = 25$ generations. Code here.

The probability that a specific pair of alleles find a common ancestor in the preceding generation is still $1/(2N)$. There are three possible pairs of alleles, so the probability that no pair finds a common ancestor in the preceding generation is

$$\left(1 - \frac{1}{2N}\right)^3 \approx \left(1 - \frac{3}{2N}\right) \quad (3.26)$$

In making this approximation we are multiplying out the right hand side and ignoring terms of $1/N^2$ and higher. See Figure 3.13 for a random realization of this process.

More generally, when we sample i alleles there are $\binom{i}{2}$ pairs,⁵ i.e. $i(i-1)/2$ pairs. Thus, the probability that no pair of alleles in a sample of size i coalesces in the preceding generation is

$$\left(1 - \frac{1}{(2N)}\right)^{\binom{i}{2}} \approx \left(1 - \frac{\binom{i}{2}}{2N}\right) \quad (3.27)$$

while the probability any pair coalesces is $\approx 2N/\binom{i}{2}$.

We can ignore the possibility that more than pairs of alleles (e.g. tripletons) simultaneously coalesce at once as terms of $1/N^2$ and higher can be ignored as they are vanishingly rare. Obviously in reasonable sample sizes there are many more triples ($\binom{i}{3}$) and higher order combinations than there are pairs ($\binom{i}{2}$), but if $i \ll N$ then we are safe to ignore these terms.

When there are i alleles, the probability that we wait until the $t + 1$ generation before any pair of alleles coalesces is

$$P(T_i = t + 1) = \frac{\binom{i}{2}}{2N} \left(1 - \frac{\binom{i}{2}}{2N}\right)^t \quad (3.28)$$

Thus the waiting time to the first coalescent event while there are i lineages is a geometrically distributed random variable with probability of success $\binom{i}{2}/2N$, which we denote by

$$T_i \sim \text{Geo}\left(\frac{\binom{i}{2}}{2N}\right). \quad (3.29)$$

The mean waiting time till any of pair within our sample coalesces is

$$\mathbb{E}(T_i) = \frac{2N}{\binom{i}{2}} \quad (3.30)$$

After a pair of alleles first finds a common ancestral allele some number of generations back in the past, we only have to keep track of that common ancestral allele for the pair when looking further into the past. Thus when a pair of alleles in our sample of i alleles coalesces, we then switch to having to follow $i - 1$ alleles back in time. Then when a pair of these $i - 1$ alleles coalesce, we then only

⁵ said as “ i choose 2”

To see the continuous time version of this, note that (3.28) is

$$\approx \frac{\binom{i}{2}}{2N} \exp\left(-\frac{\binom{i}{2}}{2N}t\right) \quad (3.31)$$

The waiting time T_i to the first coalescent event in a sample of i alleles is thus exponentially distributed with rate $\binom{i}{2}/2N$, i.e. $T_i \sim \text{Exp}\left(\frac{\binom{i}{2}}{2N}\right)$.

¹⁵⁷² have to follow $i - 2$ alleles back. This process continues until we
¹⁵⁷⁴ coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).

Simulating a coalescent genealogy To simulate a coalescent genealogy at a locus for a sample of n alleles we therefore simply follow the following algorithm:

- ¹⁵⁷⁸ 1. Set $i = n$.
- ¹⁵⁸⁰ 2. Simulate a random variable to be the time T_i to the next coalescent event from $T_i \sim \text{Exp}\left(\frac{\binom{i}{2}}{2N}\right)$
- ¹⁵⁸² 3. Choose a pair of alleles to coalesce at random from all possible pairs.
- ¹⁵⁸⁴ 4. Set $i = i - 1$
5. Continue looping steps 2-4 until $i = 1$, i.e. the most recent common ancestor of the sample is found.

¹⁵⁸⁶ By following this algorithm we are generating realizations of the genealogy of our sample.

¹⁵⁸⁸ 3.3.1 Expected properties of coalescent genealogies and mutations.

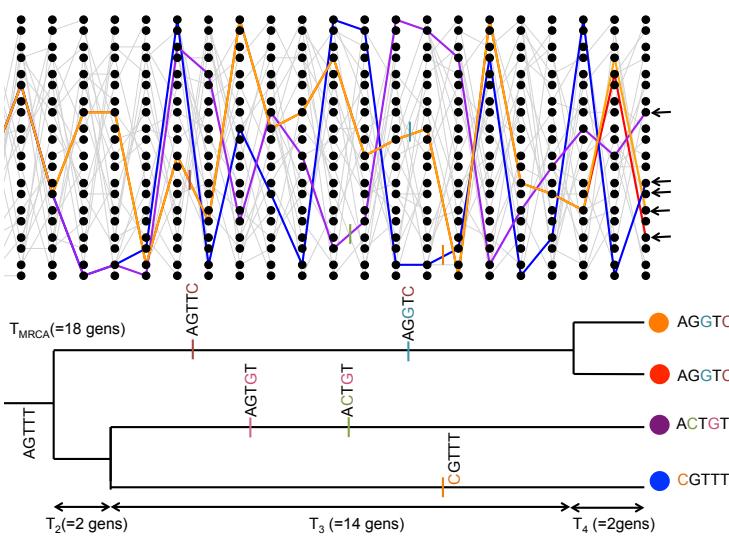


Figure 3.14: A simple coalescent tree from a single coalescent simulation, tracing the genealogy of 4 alleles with mutational changes marked with dashes showing transitions away from the MRCA sequence (AGTTT). The T_{MRCA} is $T_4 + T_3 + T_2$. The total time in the tree is $T_{tot} = 4T_4 + 3T_3 + 2T_2 = 54$ generations. Code here.

The expected time to the most recent common ancestor. We will first consider the time to the most recent common ancestor of the entire sample (T_{MRCA}). This is

$$T_{MRCA} = \sum_{i=n}^2 T_i \quad (3.32)$$

generations back, where we are summing from $i = n$ alleles counting backwards to $i = 2$ alleles (see Figure 3.14 for example). As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^2 \mathbb{E}(T_i) = \sum_{i=n}^2 2N / \binom{i}{2} \quad (3.33)$$

Using the fact that $\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$ and a bit of rearrangement, we can rewrite this as

$$\mathbb{E}(T_{MRCA}) = 4N \left(1 - \frac{1}{n} \right) \quad (3.34)$$

So the average T_{MRCA} scales linearly with population size N . Interestingly, as we move to larger and larger samples (i.e. $n \gg 1$), the average time to the most recent common ancestor converges on $4N$. What's happening here is that in large samples our lineages typically coalesce rapidly at the start and very soon coalesce down to a much smaller number of lineages.

Question 8. Assume an autosomal effective population of 10,000 individuals (roughly the long-term human estimate) and a generation time of 30 years. What is the expected time to the most recent common ancestor of a sample of 20 people? What is this time for a sample of 500 people?

The expected total time in a genealogy and the number of segregating sites. Mutations fall on specific lineages of the coalescent genealogy and are transmitted to all descendants of their lineage. Furthermore, under the infinitely-many-sites assumption, each mutation creates a new segregating site. The mutation process is a *Poisson process*, and the longer a particular lineage, i.e. the more generations of meioses it represents, the more mutations that can accumulate on it. The total number of segregating sites in a sample is thus a function of the *total* amount of time in the genealogy of the sample, or the sum of all the branch lengths on the genealogical tree, T_{tot} . Our total amount of time in the genealogy is

$$T_{tot} = \sum_{i=n}^2 iT_i \quad (3.35)$$

¹⁶²⁰ as when there are i lineages, each contributes a time T_i to the total
¹⁶²¹ time (see Figure 3.14 for an example). Taking the expectation of the
¹⁶²² total time in the genealogy,

$$\mathbb{E}(T_{tot}) = \sum_{i=n}^2 i \frac{2N}{\binom{i}{2}} = \sum_{i=n}^2 \frac{4N}{i-1} = \sum_{i=n-1}^1 \frac{4N}{i} \quad (3.36)$$

¹⁶²³ we see that our expected total amount of time in the genealogy scales
¹⁶²⁴ linearly with our population size N . Our expected total amount
¹⁶²⁵ of time is also increasing with sample size n , but is doing so very
¹⁶²⁶ slowly. This again follows from the fact that in large samples, the
¹⁶²⁷ initial coalescence usually happens very rapidly, so that extra sam-
¹⁶²⁸ ples add little to the total amount of time in the genealogical tree.

We saw above that the number of mutational differences between
¹⁶³⁰ a pair of alleles that coalescence T_2 generations ago was Poisson with
¹⁶³¹ a mean of $2\mu T_2$, where $2T_2$ is the total branch length in this simple
¹⁶³² 2-sample genealogical tree. A mutation that occurs on any branch of
¹⁶³³ our genealogy will cause a segregating polymorphism in the sample
¹⁶³⁴ (meeting our infinitely-many-sites assumption). Thus, if the total
¹⁶³⁵ time in the genealogy is T_{tot} , there are T_{tot} generations for mutations.
¹⁶³⁶ So the total number of mutations segregating in our sample (S) is
¹⁶³⁷ Poisson with mean μT_{tot} . Thus the expected number of segregating
¹⁶³⁸ sites in a sample of size n is

$$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^1 \frac{4N\mu}{i} = \theta \sum_{i=n-1}^1 \frac{1}{i} \quad (3.37)$$

Note that this is growing with the sample size n , albeit very slowly
¹⁶⁴⁰ (roughly at the rate of the log of the sample size). We can use this
¹⁶⁴¹ formula to derive another estimate of the population scaled muta-
¹⁶⁴² tion rate θ , by setting our observed number of segregating sites in a
¹⁶⁴³ sample (S) equal to this expectation. We'll call this estimator $\hat{\theta}_W$:

$$\hat{\theta}_W = \frac{S}{\sum_{i=n-1}^1 1/i} \quad (3.38)$$

¹⁶⁴⁴ This estimator of θ was devised by LetterSpace=10WATTERSON (1975),
¹⁶⁴⁵ hence the W .

¹⁶⁴⁶ *The neutral site-frequency spectrum.* We can use our coalescent process
¹⁶⁴⁷ to find the expected number of derived alleles present i times out of
¹⁶⁴⁸ a sample size n , e.g. how many singletons ($i = 1$) do we expect to
¹⁶⁴⁹ find in our sample? For example, in Figure 3.14 in our sample of four
¹⁶⁵⁰ sequences, there are 3 singletons and 2 doubletons. The number of
¹⁶⁵¹ sites with these different allele frequencies depends on the lengths
¹⁶⁵² of specific genealogical branches. A mutation that falls on a branch
¹⁶⁵³ with i descendants will create a derived allele with frequency i . For

To get a better sense of how T_{tot} grows with the sample size, we can approximate the sum 3.36 by an integral, which will work for large n . The result is
 $\int_1^{n-1} \frac{4N}{i} di = 4N \log(n-1).$

1654 example, in our example tree in Figure 3.14, the total number of
 1655 generations where a mutation could arise and be a doubleton is
 1656 $T_3 + 2T_2$, the total length of the branch ancestral to just the orange
 1657 and red allele ($T_3 + T_2$) plus the branch ancestral to just the blue and
 1658 purple allele (T_2).

To see how we could go about working this out, lets start by con-
 1660 sidering the simple coalescent tree, shown in Figure 3.15, for sam-
 1662 ple of 3 alleles drawn from a population. Mutations that fall on the
 1664 branches coloured in black will be derived singletons, while mu-
 1666 tations that fall along the orange branch will be doubletons in the
 sample. The total number of generations where a singleton mutation
 could arise is $3T_3 + T_2$. Note that we only count the time where there
 are two lineages (T_2) once. So our expected number of singletons,
 using eqn (3.30), is

$$\mathbb{E}(S_i) = \mu (3\mathbb{E}(T_3) + \mathbb{E}(T_2)) = \mu \left(3 \frac{2N}{3} + 2N \right) = \theta \quad (3.39)$$

1668 By similar logic, the time where doubletons could arise is T_2 and our
 1669 expected number of doubletons is $\mathbb{E}(S_i) = \theta/2$. Thus, there are on
 1670 average half as many doubletons as singletons.

Extending this logic to larger samples might be doable, but is
 1672 tedious (I mean really tedious: for 10 alleles there are thousands
 1674 of possible tree shapes and the task quickly gets impossible even
 computationally). A nice, relatively simple proof of the neutral site
 frequency spectrum is given by (LetterSpace=10HUDSON, 2015), but
 1676 we won't give this here. The general form is:

$$\mathbb{E}(S_i) = \frac{\theta}{i} \quad (3.40)$$

i.e. there are twice as many singletons as doubletons, three times
 1678 as many singletons as tripletons, and so on. The other thing that
 1679 will be helpful for us to know is that neutral alleles at intermediate
 1680 frequency tend to be old, and those that are rare in the sample are
 1682 young. We expect to see a lot more rare alleles in our sample than
 common alleles.

Question 9. There are two possible tree shapes that could relate
 1684 four samples. Draw both of them and separately colour (or otherwise
 1685 mark) the branches by where singletons, doubletons, and triplet
 1686 derived alleles could arise.

We can also ask the probability of observing a derived allele seg-
 1688 regating at frequency i/n given that the site is polymorphic in our
 1689 sample of size n (i.e. given that $0 < i < n$). We can obtain this prob-
 1690 ability by dividing the expected number of sites segregating for an

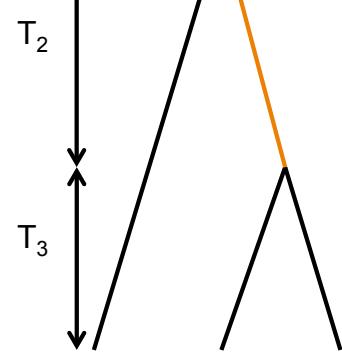


Figure 3.15: A tree for three samples; note that this is the only possible tree shape (treating the tips as unlabeled, i.e. I don't care which pair of sequences carry a doubleton, just that any two sequences carry a derived allele).

allele at frequency i by the expected number segregating at all of the
 1692 possible allele frequencies for polymorphisms in our sample

$$P(i|0 < i < n) = \frac{\mathbb{E}(S_i)}{\sum_{j=1}^{n-1} \mathbb{E}(S_j)} = \frac{1/i}{\sum_{j=1}^{n-1} 1/j}. \quad (3.41)$$

We can interpret this probability as the fraction of polymorphic sites
 1694 we expect to find at a frequency i/n .

Tests based on the site frequency spectrum Population geneticists
 1696 have proposed a variety of ways to test whether an observed site
 frequency spectrum conforms to its neutral, constant-size expecta-
 1698 tions. These tests are useful for detecting population size changes
 using data across many loci, or for detecting the signal of selec-
 1700 tion at individual loci. One of the first tests was proposed by (Let-
 terSpace=10TAJIMA, 1989), and is called Tajima's D . Tajima's D is
 1702

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{C} \quad (3.42)$$

where the numerator is the difference between the estimate of θ
 1704 based on pairwise differences and that based on segregating sites.
 As these two estimators both have expectation θ under the neutral,
 1706 constant-size model, the expectation of D is zero. The denominator C
 is a positive constant; it's the square-root of an estimator of the vari-
 1708 ance of this difference under the constant population size, neutral
 model. This constant was chosen for D to have mean zero and vari-
 1710 ance 1 under the null model, so we can test for departures from this
 simple null model.

An excess of rare alleles compared to the constant-size, neutral
 1712 model will result in a negative Tajima's D , because each additional
 1714 rare allele increases the number of segregating sites by 1, but only
 has a small effect on the number of pairwise differences between
 1716 samples. In contrast, a positive Tajima's D reflects an excess of inter-
 1718 mediate frequency alleles relative to the constant-size, neutral expec-
 tation. Alleles at intermediate-frequency increase pairwise diversity
 more per segregating site than typical, thus increasing θ_π more than
 1720 θ_W .

3.3.2 Demography and the coalescent

We've already seen how changes in population size can change the
 1722 rate at which heterozygosity is lost from the population (see the dis-
 1724 cussion around eqn. (3.13)). If the population size in generation i is
 1726 N_i , the probability that a pair of lineages coalesce is $1/(2N_i)$; this con-
 forms to our intuition that if the population size is small, the rate at
 which pairs of lineages find their common ancestor is faster. We can

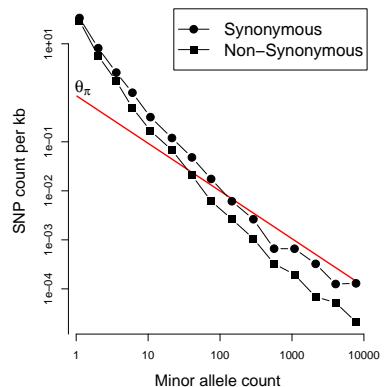
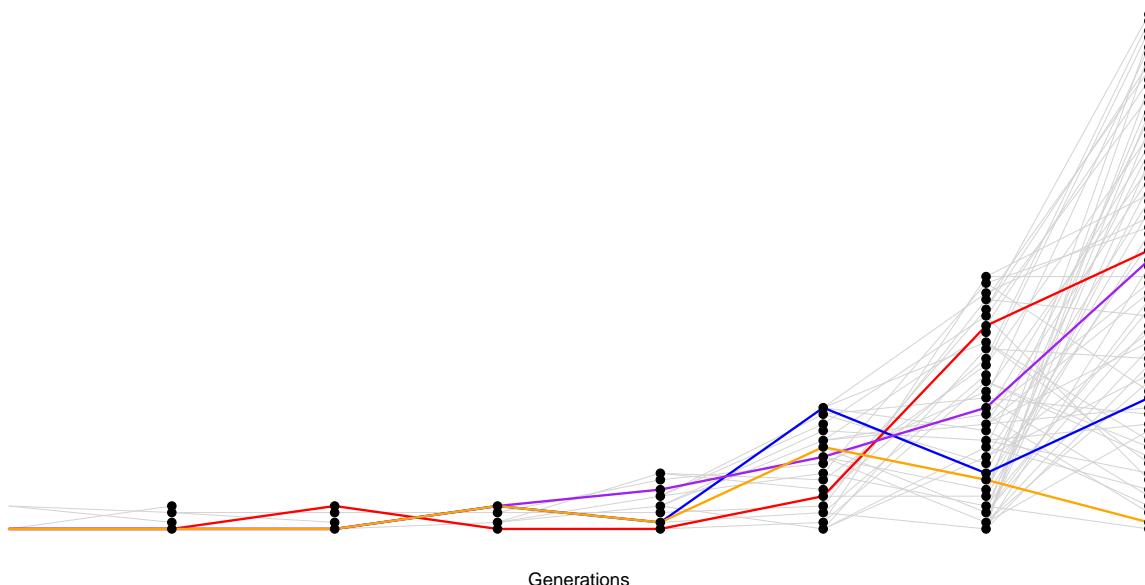


Figure 3.16: Data from 202 genes from 14002 people of European ancestry (28004 alleles). Note the double log-scale. The red line gives the neutral, constant population size estimate of the site frequency spectrum, our equation (3.40), using a θ estimated from π . Note how the non-synonymous changes are even more skewed towards rare alleles, likely due to selection against non-synonymous alleles preventing them from reaching high frequency. Data from LetterSpace=10NELSON *et al.* (2012). Code here.

1728 potentially accommodate rapid random fluctuations in population size by simply using the effective population size N_e in place of N .
 1730 However, longer term more systematic changes in population size will distort the coalescent genealogies, and hence patterns of diversity, in more systematic ways.
 1732

We can see how demography potentially distorts the observed frequency spectrum away from the neutral expectation in a very large sample of humans shown in Figure 3.16. For comparison, the neutral frequency spectrum, eqn (3.40), is shown as a red line. There are vastly more rare alleles than expected under our neutral, constant-size-size model, but the neutral prediction and reality agree somewhat more for alleles that are more common.



1740 Why is this? Well, these patterns are likely the result of the very recent explosive growth in human populations. If the population
 1742 has grown rapidly, then the pairwise-coalescent rate in the past may be much higher than the coalescent rate closer to the present. (see
 1744 Figure 3.17).

One consequence of a recent population expansion is that there
 1746 is much less genetic diversity in the population than you'd predict using the census population size. Humans are one example of this effect; there are 7 billion of us alive today, but this is due to very rapid population growth over the past thousand to tens of thousands of years. Our level of genetic diversity is very much lower than you'd predict given our census size, reflecting our much smaller ancestral
 1750 population. A second consequence of recent population expansion is

Figure 3.17: A realization of the coalescent process in a growing population. The population underwent a period of doubling every generation. The initial population size of just two individuals, maintained for a number of generations, is obviously highly unrealistic but serves our purpose. Code here.

that the deeper coalescent branches are much more squished together
 1754 in time compared to those in a constant-sized population. Mutations
 on deeper branches are the source of alleles at more intermediate
 1756 frequencies, and so there are even fewer intermediate-frequency
 alleles in growing populations. That's why there are so many rare
 1758 alleles, especially singletons, in this large sample of Europeans.

Another common demographic scenario is a population bottleneck.
 1760 In a bottleneck, the population size crashes dramatically, and
 subsequently recovers. For example, our population may have had
 1762 size N_{Big} and crashed down to N_{Small} . One example of a bottleneck
 is shown in Figure 3.18. Looking at a sample of lineages drawn

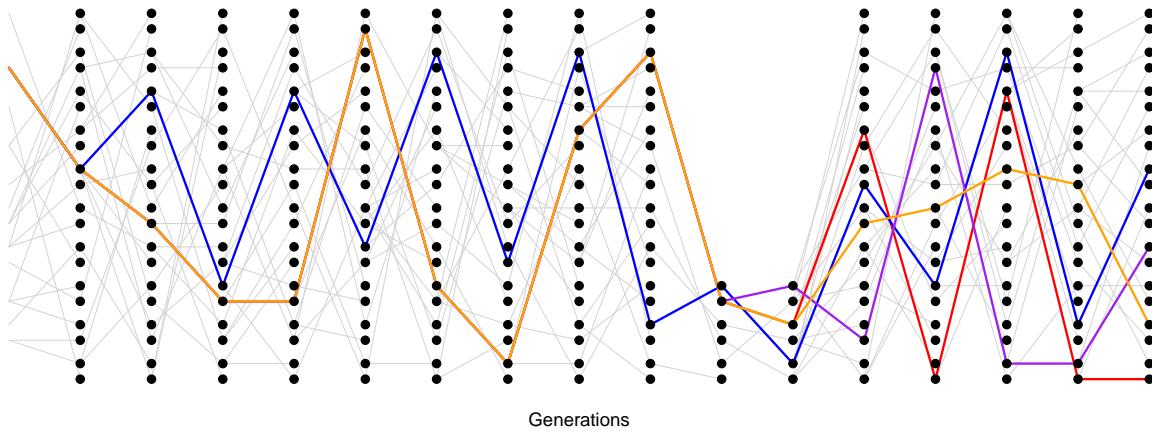
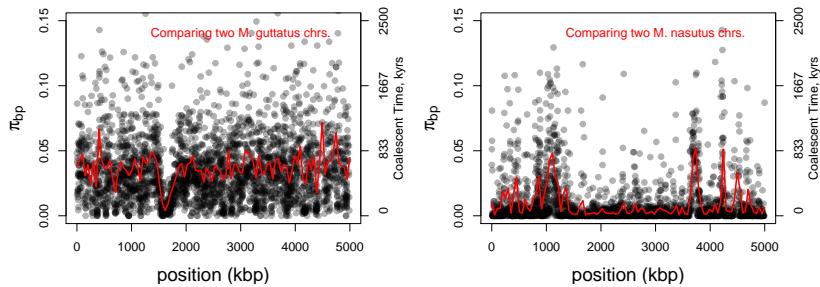


Figure 3.18: A realization of the coalescent process in a bottlenecked population. Our population underwent a bottleneck eight generations in the past. Code here.

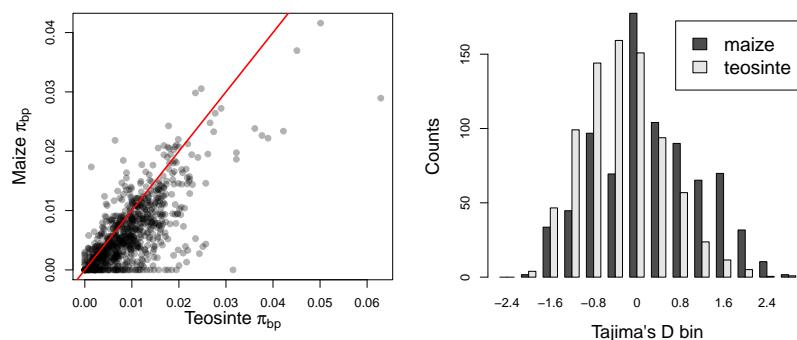
1764 from the population today, if the bottleneck was somewhat recent
 ($\ll N_{\text{Big}}$ generations in the past) many of our lineages will not have
 1766 coalesced before reaching the bottleneck, moving backward in time.
 But during the bottleneck our lineages coalesce at a much higher rate,
 1768 such that many of our lineages will coalesce if the bottleneck lasts
 long enough ($\sim N_{\text{Small}}$ generations). If the bottleneck is very strong,
 1770 then all of our lineages will coalesce during the bottleneck, and the
 resulting site frequency spectrum may look very much like our pop-
 1772 ulation growth model (i.e. an excess of rare alleles). However, if some
 pairs of lineages escape coalescing during the bottleneck, they will
 1774 coalesce much more deeply in time (e.g. the blue and orange ances-
 tral lineages in 3.18).

Figure 3.19: Diversity along a region of **LetterSpace=10** in *Mimulus nasutus*. **GENETICS** 73 sampled from two individuals, the red line is a moving average (data from LetterSpace=10 BRANDVAIN *et al.*). Pairwise coalescent times (t) estimated assuming $t = \pi/2\mu$ using $\mu_{BP} = 10^{-9}$. Code here.



1776 An example of this is shown Figure 3.19, data from LetterSpace=10 BRANDVAIN
 et al.. *Mimulus nasutus* is a selfing species that arose recently from an
 1778 out-crossing progenitor *M. guttatus*, and experienced a strong bottle-
 neck. *M. guttatus* has very high levels of genetic diversity ($\pi = 4\%$
 1780 at synonymous sites), but *M. nasutus* has lost much of this diversity
 1782 ($\pi = 1\%$). Looking along the genome, between a pair of *M. guttatus*
 chromosomes, levels of diversity are fairly uniformly high.

But in comparing two *M. nasutus* chromosomes, diversity is low
 1784 because the pair of lineages generally coalesce recently. Yet in a few
 places we see levels of diversity comparable to *M. guttatus*; these
 1786 regions correspond to genomic sites where our pair of lineages fail to
 coalesce during the bottleneck and subsequently coalesce much more
 deeply in the ancestral *M. guttatus* population.



1788 Mutations that arise on deeper lineages will be at intermediate
 1790 frequency in our sample, and so mild bottlenecks can lead to an
 excess of intermediate frequency alleles compared to the standard
 1792 constant-size model. This can skew Tajima's D (see eqn 3.42) towards
 positive values and away from its expectation of zero. One example
 1794 of this skew is shown in Figure 3.21. Maize (*Zea mays* subsp. *mays*)

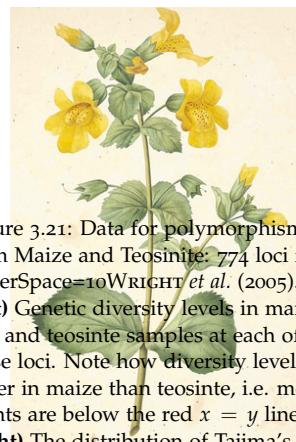
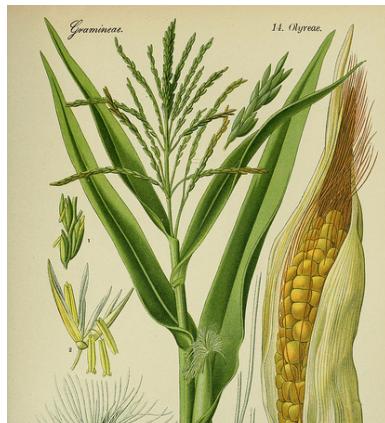


Figure 3.21: Data for polymorphism from Maize and Teosinte: 774 loci from LetterSpace=10 WRIGHT *et al.* (2005). **Left)** Genetic diversity levels in maize and teosinte samples at each of these loci. Note how diversity levels are lower in maize than teosinte, i.e. most points are below the red $x = y$ line. **Right)** The distribution of Tajima's D in maize and teosinte, see how the maize distribution is shifted towards positive values. **Yellow Monkey Flower** *Mimulus luteus*. Code here.

Choix des plus belles fleurs et des plus beaux fruits. Pierre-Joseph Redouté. (1833). Contributed to Flickr by Swallowtail Garden Seeds. Public Domain.



was domesticated from its wild progenitor teosinte (*Zea mays* subsp. *parviglumis*) roughly ten thousand years ago. We can see how the bottleneck associated with domestication has resulted in a loss of genetic diversity in maize compared to teosinte, and the polymorphism that remains is somewhat skewed towards intermediate frequencies resulting in more positive values of Tajima's D.

Question 10. LetterSpace=10VOIGHT *et al.* (2005) sequenced 40 autosomal regions from 15 diploid samples of Hausa people from Yaounde, Cameroon. The average length of locus they sequenced for each region was 2365bp. They found that the average number of segregating sites per locus was $S = 11.1$ and the average $\pi = 0.0011$ per base over the loci. Is Tajima's D positive or negative? Is a demographic model with a bottleneck or growth more consistent with this result?

3.4 Molecular Evolution and the fixation of neutral alleles

"history is just one damn thing after another" -Arnold Toynbee

It is very unlikely that a rare neutral allele accidentally drifts up to fixation; more likely, such an allele will be eventually lost from the population. However, populations experience a large and constant influx of rare alleles due to mutation, so even if it is very unlikely that an individual allele fixes within the population, some neutral alleles will fix by chance.

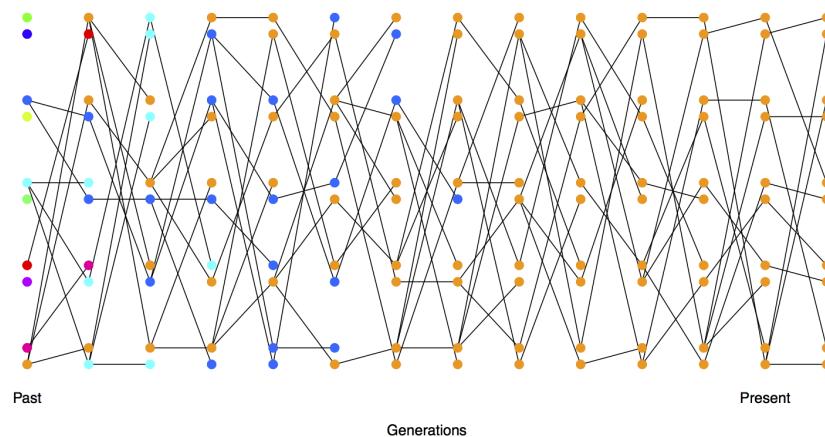


Figure 3.23: Each allele initially present in a small diploid population is given a different colour so we can track their descendants over time. By the 9th generation, all of the alleles present in the population can trace their ancestry back to the orange allele. Code here.

Probability of the eventual fixation of a neutral allele An allele which reaches fixation within a population is an ancestor to the entire population. In a particular generation there can only be a single allele

that all other alleles at the locus in a later generation can claim as an ancestor (See Figure 3.23). At a neutral locus, the actual allele does not affect the number of descendants that the allele has (this follows from the definition of neutrality: neutral alleles don't leave more or less descendants on average than other neutral alleles). An equivalent way to state this is that the allele labels don't affect anything; thus the alleles are *exchangeable*. As a consequence of being exchangeable, any allele is equally likely to be the ancestor of the entire population. In a diploid population of size N , there are $2N$ alleles, all of which are equally likely to be the ancestor of the entire population at some later time point. So if our allele is present in a single copy, the chance that it is the ancestor to the entire population in some future generation is $1/(2N)$, i.e. the chance our neutral allele is eventually fixed is $1/(2N)$. In Figure 3.23, our orange allele in the first generation is one of 10 differently coloured alleles, and so has a $1/10$ chance of being the ancestor of the entire population at some later time point (and in this simulation it does become the common ancestor, by the 9th generation).

More generally, if our neutral allele is present in i copies in the population, of $2N$ alleles, the probability that this allele becomes fixed is $i/(2N)$, i.e. the probability that a neutral allele is eventually fixed is simply given by its frequency (p) in the population. (We can also derive this result by letting $N_s \rightarrow 0$ in eqn. (8.11), a result we'll encounter later.)

How long does it take on average for such an allele to fix within our population? Well, in developing equation (3.34) we've seen that it takes $4N$ generations for a large sample of alleles to all trace their ancestry back to a single most recent common ancestral allele. Any single-base pair change which arose as a single mutation at a locus, and fixed in the population, must have been present in the sequence transmitted by the most recent common ancestor of the population at that locus. Thus it must take roughly $4N$ generations for a neutral allele present in a single copy within the population to fix. This argument can be made more precise, but in general we would still find that it takes $\approx 4N$ generations for a neutral allele to go from its introduction to fixation with the population.

Rate of substitution of neutral alleles A substitution between populations that do not exchange gene flow is simply a fixation event within one population. The rate of substitution is therefore the rate at which new alleles fix in the population, so that the long-term substitution rate is the rate at which mutations arise that will eventually become fixed within our population.

Lets assume, based on our discussion of the neutral theory of

molecular evolution, that there are only two classes of mutational changes that can occur with a region, highly deleterious mutations and neutral mutations. A fraction C of all mutational changes are highly deleterious, and cannot possibly contribute to substitution nor polymorphism. The other $1 - C$ fraction of mutations are neutral. If our total mutation rate is μ per transmitted allele per generation, then a total of $2N\mu(1 - C)$ neutral mutations enter our population each generation.

Each of these neutral mutations has a $1/(2N)$ probability chance of eventually becoming fixed in the population. Therefore, the rate at which neutral mutations arise that eventually become fixed within our population is

$$2N\mu(1 - C) \frac{1}{2N} = \mu(1 - C) \quad (3.43)$$

Thus the rate of substitution, under a model where newly arising alleles are either highly deleterious or neutral, is simply given by the mutation rate of neutral alleles, i.e. $\mu(1 - C)$.

Consider a pair of species that have diverged for T generations, i.e. orthologous sequences shared between the species last shared a common ancestor T generations ago. If these species have maintained a constant μ over that time, they will have accumulated an average of

$$2\mu(1 - C)T \quad (3.44)$$

neutral substitutions. This assumes that T is a lot longer than the time it takes to fix a neutral allele, such that the total number of alleles introduced into the population that will eventually fix is the total number of substitutions.

This is a really pretty result as the population size has completely canceled out of the neutral substitution rate. However, there is another way to see this in a more straight forward way. If I look at a sequence in me compared to, say, a particular chimp, I'm looking at the mutations that have occurred in both of our germlines since they parted ways T generations ago. Since neutral alleles do not alter the probability of their transmission to the next generation, we are simply looking at the mutations that have occurred in $2T$ generations worth of transmissions. Thus the average number of neutral mutational differences separating our pair of species is simply $2\mu(1 - C)T$.

A number of observations follow under this model, from equation (3.44). The first is that a primary determinant of patterns of molecular evolution in a genomic region is the level of constraint (C). This pattern generally seems to hold empirically: non-coding regions often evolve more rapidly than coding regions, synonymous substitutions accumulate faster than nonsynonymous, and nonsynonymous

"functionally less important molecules or parts of a molecule evolve faster than more important ones."

- LetterSpace=10KIMURA and LetterSpace=10OHTA (1974)

substitutions accumulate faster in less vital proteins than ones that are absolutely necessary for early development. Note that this is not a unique prediction of the neutral model, e.g. less constrained regions may also be better able to evolve adaptively. However, it is a fantastically useful general insight, e.g. it allows us to spot putatively functional non-coding regions by looking for genomic regions that have very low levels of divergence among distantly related species.

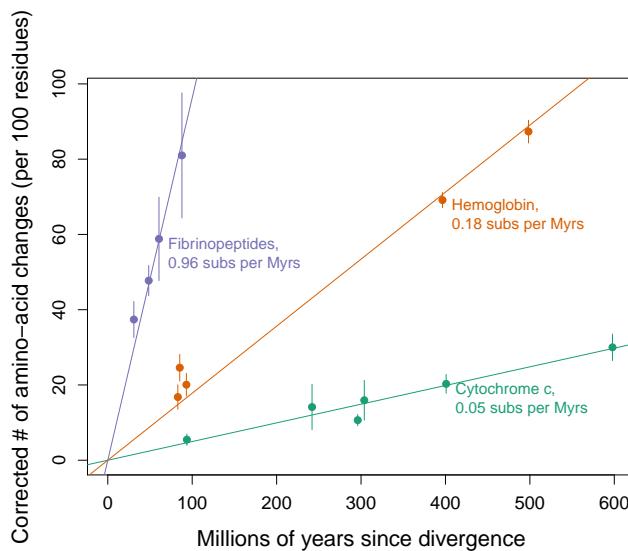


Figure 3.24: The numbers of substitutions in three proteins, corrected for multiple hits, between various pairs of groups plotted against the time these groups shared a common ancestor in the fossil record. Data from LetterSpace=10DICKERSON (1971). The lines give the linear regression through the origin for each protein. The slope of the regression is given next to the protein name. Code here. See (LetterSpace=10ROBINSON *et al.*, 2016) who revisited this classic study and confirmed the conclusions.



Figure 3.25: Eastern diamondback rattlesnake (*Crotalus adamanteus*). North American herpetology. Holbrook, J. E. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

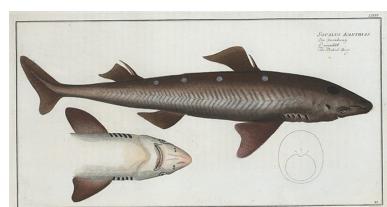


Figure 3.26: Spiny dogfish (*Squalus acanthias*). Rare Book Division, The New York Public Library. "Squalus Acanthias, The Picked-Dog" The New York Public Library Digital Collections. 1785 - 1797.

The second important insight, and critical for the development of the neutral theory, is that equation (3.44) is seemingly consistent with LetterSpace=10ZUCKERKANDL and LetterSpace=10PAULING (1965)'s hypothesis of a surprisingly constant, protein molecular clock. The protein molecular clock is the observation that for some proteins there's a linear relationship between the number of non-synonymous (NS) substitutions and the time species last shared a common ancestor in the fossil record. LetterSpace=10DICKERSON (1971) provided an example of this observation (Figure 3.24), by comparing various organisms whose molecular sequences were available to him. For example, he found that humans and rattlesnakes, who last share a common ancestor in the fossil record around 300 million years, are separated by roughly 15 NS substitutions per 100 sites in the Cytochrome c protein. While, humans and dogfish, which diverged around 400 million years, are separated by 19 NS substitutions per 100 sites in this gene.

In equation (3.44) we double the amount of time separating a pair of species T , we double the number of substitutions predicted.

Note that for this to be true T must be measured in generations.

To explain a protein molecular clock between species that clearly differed dramatically in generation time it was hypothesized that the mutation rate actually scaled with generation time, i.e. short-lived organisms introduced less mutations per generation, e.g. as they had fewer rounds of mitosis. This generation-time assumption meant that the mutation rate per year could be constant, such that μT would be a constant for pairs of species that had diverged for similar geological times, which are measured in years, even if the organisms differed in generation time. This assumption would allow neutral theory to be consistent with a protein molecular clock measured in years. We now know that this critical generation time assumption is false, organisms with shorter generation times have somewhat higher mutation rates per year, and so a strict neutral model is inconsistent with the protein molecular clock. We'll return to these ideas when we discuss the fate of very weakly selected mutations in Chapter 8 and LetterSpace=10OHTA (1973)'s Nearly Neutral theory. If you are still reading this send Graham a picture of Tomoko Ohta receiving the Crafoord Prize, an analog of the Nobel prize for biology, for her contributions to molecular evolution.

The contribution of ancestral polymorphism to divergence. If we are considering T to represent the divergence between long-separated species, then we can think of T as the time that the species split. However, for more recently diverged populations and species, we need to include the fact that the sorting of ancestral polymorphism contributes to divergence among species. In Figure 3.27, we see our two populations split T_s generations ago. However, the coalescence of our A and B lineage is necessarily deeper in time than T_s . The top mutation was polymorphic in the ancestral population but now contributes to the divergence between A and B. Assuming that our ancestral population had effective size N_A individuals, and that our populations split cleanly with no subsequent gene flow, then

$$T = T_s + 2N_A. \quad (3.45)$$

If our species split time is very large compared to $2N$ then we can think of T as the split time.

Question 11. For this, and the next question, assume that humans and chimp diverged around 5.5×10^6 years ago, have a generation time 20 years, that the speciation occurred instantaneously in allopatry with no subsequent gene flow, and the ancestral effective population size of the human and chimp common ancestor population was 10,000 individuals.

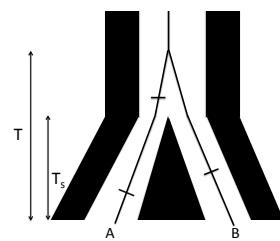


Figure 3.27: The genealogy of two alleles one sampled from population A and B. Mutations on the lineages are shown as dashes. The pair of alleles coalesce in the ancestral population of A and B. The two populations split T_s generations ago, with no subsequent gene flow, but the two lineages must coalesce deeper in time.

Nachman and Crowell sequenced 12 pseudogenes in human and

1970 chimp and found substitutions at 1.3% of sites.

A) What is the mutation rate per site per generation at these

1972 genes?

B) All of the pseudogenes they sequenced are on the autosomes.

1974 What would your prediction be for pseudogenes on the X and Y chromosomes, given that there are fewer rounds of replication in the
 1976 female germline than in the male germline.

3.5 Tests of molecular evolution.

3.5.1 Comparing the rates of non-synonymous to synonymous substitutions d_N/d_S

1980 One common tool in molecular evolution is to compare the estimated
 1982 number (or rates) of substitutions in different classes of genomic
 1984 sites, for example the ratio of the number of non-synonymous to syn-
 1986 onymous substitutions in a given gene. The simplest way to calculate
 d_N is to count up the non-synonymous changes and divide by the
 1990 total number of positions in the gene where a non-synonymous point
 1992 mutation could occur. We can do likewise for synonymous changes
 d_S , and then take the ratio d_N/d_S . This is a helpful conceptual way
 1994 to think about what d_N/d_S represents, however, this ignores the fact
 1996 that some changes are more likely to occur by mutation than others
 1998 and also does not account for multiple hits (multiple mutations at
 2000 the same bp position). Therefore, in practice the ratio d_N/d_S is more
 2002 typically calculated by model-based likelihood and bayesian methods
 that can account for these features.

2004 For the vast majority of protein-coding genes in the genome we
 see that $d_N/d_S < 1$. This observation is consistent with the view that
 2006 non-synonymous sites are much more constrained than synonymous
 sites, i.e. that most non-synonymous mutations are deleterious and
 quickly removed from the population. If we are willing to make the
 assumption that all synonymous changes are neutral, $d_S = 2T\mu$,
 then we can estimate the degree of constraint on non-synonymous
 sites. (Note that synonymous changes can sometimes be subject to
 both positive and negative selection, but this neutral assumption is a
 useful starting place.)

2004 Assume that a fraction C of non-synonymous changes are too
 deleterious to contribute to polymorphism. Then, after T generations
 2006 of divergence have elapsed between two populations, we'd expect d_N
 neutral non-synonymous substitutions, where

$$d_N = 2T(1 - C)\mu \quad (3.46)$$

2008 Dividing by d_S , we find

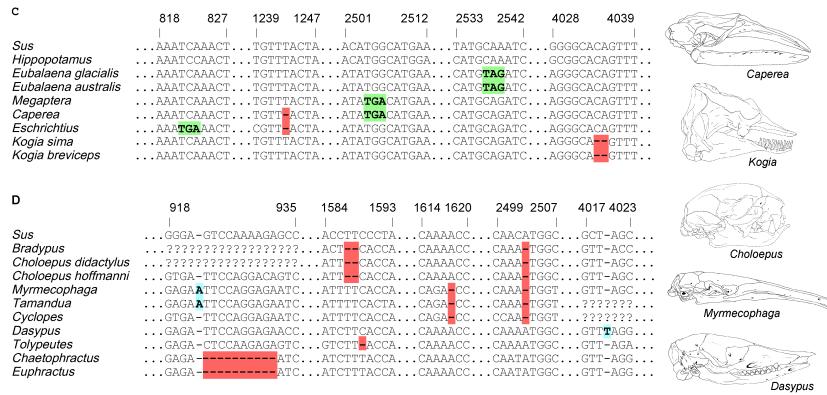
$$d_N/d_S = (1 - C) \quad (3.47)$$

Therefore, if we assume that non-synonymous mutations can only be
 2010 strongly deleterious or neutral, we estimate the fraction of mutational
 changes that are constrained by negative selection as $C = 1 - d_N/d_S$.
 2012 C has the interpretations of being the fraction of non-synonymous
 mutations that are quickly weeded out of the population by selection,
 2014 and so do not contribute to divergence among species.

We can test whether our gene is evolving in a constrained way at
 2016 the protein level by estimating d_N/d_S and testing whether this is sig-
 nificantly less than 1. A d_N/d_S test can provide evolutionary evidence
 2018 that a stretch of DNA proposed to be protein-coding is subject to se-
 lective constraint, and so likely does encode for a functional protein.
 2020 We can also perform a d_N/d_S test on specific branches of a phylogeny
 for a gene, to test on which branches the gene is subject to constraint,
 2022 or to test for changes in the level of constraint across the phylogeny.

Loss of constraint at pseudogenes. While most protein genes evolve
 2024 under constraint, we can find examples of genes that are evolving
 in a less constrained manner. The simplest example of this is where
 2026 the gene has lost function. Genes can lose function because of inacti-
 vating mutations that stop them being transcribed or translated into
 2028 functional proteins. Such genes are called 'pseudogenes'. When a
 gene completely loses function there is no longer selection against
 2030 non-synonymous changes and so such mutations are just as free to
 accumulate as synonymous changes, and so $d_N/d_S = 1$. Pseudo-
 2032 genes are a wonderful example of the extension of Darwin's ideas
 about vestigial traits ('Rudimentary organs') to the DNA level; we
 2034 can still recognize a once useful word (gene) whose spelling is slowly
 degrading. Our genomes are filled with old pseudogenes whose
 2036 original meanings (functional protein coding sequences) are slowly
 being eroded through the accumulation of neutral substitutions. One
 2038 nice example of a gene that has repeatedly lost function, i.e. become
 repeatedly pseudogenized, is the Enamlin gene from the study of
 2040 LetterSpace=10MEREDITH *et al.* (2009).

"Rudimentary organs may be com-
 pared with the letters in a word, still
 retained in the spelling, but become
 useless in the pronunciation, but which
 serve as a clue .. for its derivation." –
 LetterSpace=10DARWIN (1859) pg. 455



The protein Enamlin is a key structural protein involved in the outer cap of enamel on teeth. Various mammals have secondarily evolved diets that do not require hard teeth, and so greatly reduced the selection pressure for hard enamel, or even teeth at all. For example, two-toed sloths (*Choloepus*), Pygmy sperm whales (*Kogia*), and aardvark (*Orycteropus*) all lack enamel on teeth. Other mammals have lost their teeth entirely, e.g. giant anteaters (*Myrmecophaga*) and Baleen whales. Due to this relaxation of constraint on the phenotype, the Enamlin gene has accumulated pseudogenizing substitutions such as premature stop codons and frameshift mutations (see Figure 3.28 for examples). LetterSpace=10MEREDITH *et al.* sequenced Enamlin across a range of species and found that none of the species with enamel have frameshift mutations in Enamlin, while 17/20 of species that lack enamel or teeth have frameshifts in Enamlin, and all of them carry premature stop codons (Figure 3.30).

The branches of the Enamlin phylogeny with a functional Enamlin gene (black) had an estimated $d_N/d_S = 1.051$, consistent with the protein evolving in a constrained manner. In contrast, the branches with a pseudogenized Enamlin (red) had $d_N/d_S = 1.02$, consistent with the gene evolving an unconstrained way. The branches where the gene was likely transitioning from a functional to non-function state, i.e. pre-mutation (blue) and mixed (purple), had intermediate values of $d_N/d_S = 0.83 - 0.98$, consistent with a transition from a constrained to unconstrained mode of protein evolution somewhere along these branches of the phylogeny.

Adaptive evolution and d_N/d_S . Clearly genes are not only subject to neutral and deleterious mutations; beneficial mutations must also arise and fix from time to time. Let's assume that a fraction B of non-synonymous mutations that arise are beneficial such that $2N\mu B$

Figure 3.28: Examples of frameshift mutations (insertions blue, deletions red) and premature stop codons in Enamlin in Cetacea and Xenarthra. Figure from LetterSpace=10MEREDITH *et al.* (2009), licensed under CC BY 4.0.



Figure 3.29: Two-toed sloth (*Choloepus hoffmanni*). An introduction to the study of mammals, living and extinct, 1891. Flower W. H. and Lydekker R. Image from the Biodiversity Heritage Library. Contributed by University of Toronto. Not in copyright.

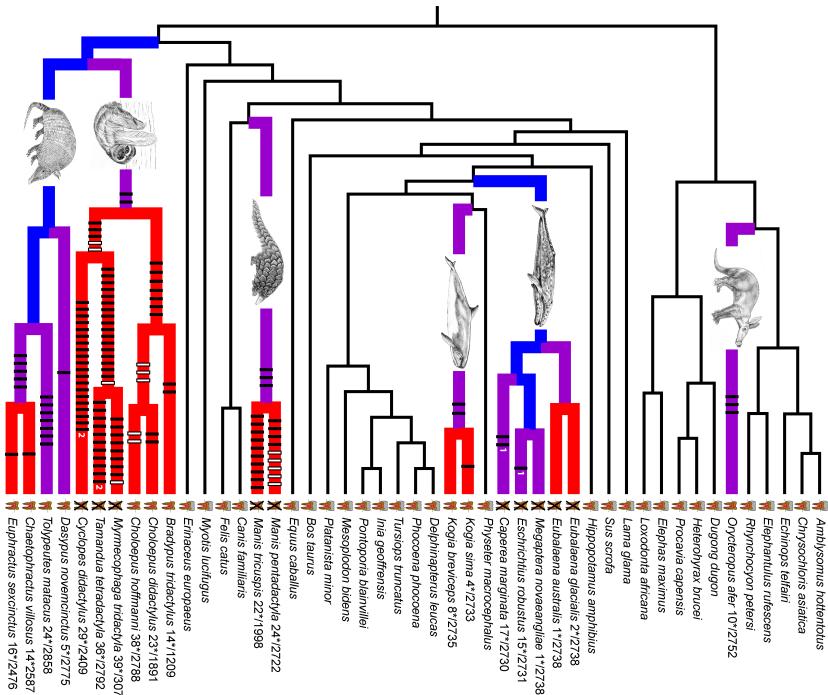


Figure 3.30: The tooth symbol next to each taxon shows whether they have teeth with enamel, lack enamel, or lack teeth. Branches of the phylogeny are coloured by whether their Enamlin is functional (black), pre-mutation (blue), mixed (purple), or pseudogenic (red). The black and white vertical bars on branches show frameshift mutations. The numbers after taxon names indicate minimum number of stop codons in the sequence divided by the length of the sequence. Figure from LetterSpace=10MEREDITH *et al.* (2009), licensed under CC BY 4.0.

2070 beneficial mutations arise per generation. Newly arisen beneficial alleles are not destined to fix in the population, as they may be lost to 2072 genetic drift when they are rare in the population (we'll discuss how to calculate the fixation probability for beneficial alleles in Chapter 2074 8). A newly arisen beneficial allele reaches fixation in the population with probability f_B from its initial frequency of $1/2N$. This fixation probability may be much higher than that of neutral mutations, but still much less than 1. If $2T$ generations of divergence have elapsed 2076 between the two populations then a total of 2078

$$dN = 2T(1 - C - B)\mu + 2T \times (2N\mu B) \times f_B \quad (3.48)$$

non-synonymous substitutions will have accumulated. Then

$$d_N/d_S = (1 - C - B) + 2NBf_B \quad (3.49)$$

2080 assuming again that all synonymous mutations are neutral. Note that this means that our estimates of C using $1 - d_N/d_S$ will be a lower 2082 bound on the true constraint if even a small fraction of mutations are beneficial. Those cases where the gene is evolving more rapidly 2084 at the protein level than at synonymous sites, i.e. $d_N/d_S > 1$, are potentially strong candidates for positive selection rapidly driving 2086 change at the protein level. We can identify genes that have d_N/d_S significantly greater than one, either on the complete gene phylogeny, or

2088 on particular branches. Note that is a very conservative test that few
genes in the genome meet, as many genes that are fixing adaptive
2090 non-synonymous substitutions will have $d_N/d_S < 1$; even if adap-
tive mutations are common, genes may still evolve in a constrained
2092 way (i.e. $d_N/d_S < 1$) if the rapid fixation of beneficial mutations due
to positive selection is outweighed by the loss of non-synonymous
2094 mutations to negative selection.

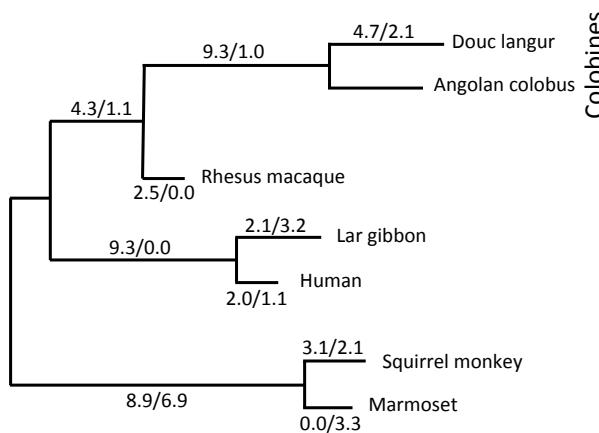


Figure 3.31: A phylogram for the primate lysozyme gene, data from LetterSpace=10YANG (1998). For each branch, the numbers give the estimated average number of non-synonymous to synonymous changes in the lysozyme protein.

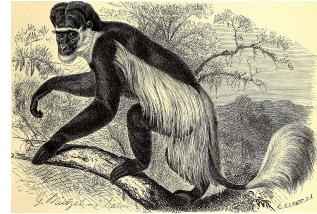


Figure 3.32: Abyssinian black-and-white colobus (*Colobus guereza*). A member of the leaf-eating Colobines.
Brehm's Tierleben, Brehm, A.E. 1893. Image from the Biodiversity Heritage Library. Contributed by University of Illinois Urbana-Champaign. Not in copyright.



Figure 3.33: (hoatzin (*Opisthocomus hoazin*)). A leaf-eating bird.
A history of birds (1910) Pyrcraft, W.P. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

2096 A classic example for looking at adaptive evolution using dN/dS
is the evolution of the lysozyme protein in primates (LetterSpace=10MESSIER
and LetterSpace=10STEWART, 1997; LetterSpace=10YANG, 1998), see
2098 the phylogeny in Figure 3.31. The lysozyme protein is a key com-
ponent for the breakdown of bacterial walls. It shows very fast pro-
2100 tein evolution, notably on the lineages leading to apes (e.g. gibbons
and humans) and Colobines (e.g. colobus and langur monkeys).
2102 Colobines have leaf-based diets. They digest these leaves by bact-
rial fermentation in their foregut, and then use lysozymes to break
2104 down the bacteria to extract energy from the leaves. In Colobines,
the lysozyme protein has evolved to work well in the high-PH en-
2106 vironment of the stomach. Remarkably, the Colobine lysozyme
has convergently evolved this activity via very similar amino-acid
2108 changes at 5 key residuals in cows and Hoatzins (a leaf eating bird,
LetterSpace=10KORNEGAY *et al.*, 1994)

2110 *The McDonald-Kreitman test* As noted above, a big issue with using
 d_N/d_S to detect adaptation is that it is very conservative. For a more
2112 powerful test of rapid divergence, what we need to do is adjust for
the level of constraint a gene experiences at non-synonymous sites.
2114 One way to do this is to use polymorphism data as an internal con-
trol. If we see little non-synonymous polymorphism at a gene, but a

2116 lot of synonymous polymorphism, we now know that there is likely
 2117 strong constraint on the gene (i.e. high C), thus we expect d_N/d_S to
 2118 be low. LetterSpace=10McDONALD and LetterSpace=10KREITMAN
 2119 (1991) devised a simple test of the neutral theory of molecular evo-
 2120 lution at a gene based on this intuition (building on the concep-
 2121 tually similar HKA test LetterSpace=10HUDSON *et al.*, 1987). Let-
 2122 terSpace=10McDONALD and LetterSpace=10KREITMAN took the case
 2123 where we have polymorphism data at a gene for one species and
 2124 divergence to a closely related species. They partitioned polymor-
 2125 phism and fixed differences in their sample into non-synonymous
 2126 and synonymous changes:

	Poly.	Fixed
Non-Syn.	P_N	D_N
Syn.	P_S	D_S
Ratio	P_N/P_S	D_N/D_S

2128 Under neutral theory, we expect a smaller number of non-synonymous
 2129 to synonymous fixed differences ($D_N/D_S < 1$) and exactly the same
 2130 expectation holds for polymorphism (P_N/P_S). Let's consider a gene
 2131 with L_S and L_N sites where synonymous and non-synonymous mu-
 2132 tations could arise respectively. We can think of the underlying gene
 2133 genealogy at our gene, see Figure 3.34, with the total time on the co-
 2134 alescent genealogy within the species as T_{tot} and the total time for
 2135 fixed differences between our species as T'_{div} . Then under neutrality
 2136 we expect $\mu L_N(1 - C)T_{tot}$ non-synonymous polymorphisms (i.e. our
 2137 number of segregating sites), and $\mu L_N(1 - C)T'_{div}$ non-synonymous
 2138 fixed differences. We can then fill out the rest of our table as follows:

	Poly.	Fixed
Non-Syn.	$\mu L_N(1 - C)T_{tot}$	$\mu L_N(1 - C)T'_{div}$
Syn.	$\mu L_N T_{tot}$	$\mu L_S T'_{div}$
Ratio	$L_N(1 - C)/(L_S)$	$L_N(1 - C)/(L_S)$

2140 Therefore, we expect the ratio of non-synonymous to synonymous
 2141 changes to be the same for polymorphism and divergence under a
 2142 strict neutral model. We can test this expectation of equal ratios via
 2143 the standard tests of a 2×2 table. If the ratio of N/S is significantly
 2144 higher for divergence than polymorphism we have evidence that
 2145 non-synonymous substitutions are accumulating more rapidly than
 2146 we would predict given levels of constraint alone.

As example of a Mcdonald-Kreitman table consider the work of
 2148 LetterSpace=10FRENTIU *et al.* (2007) on the molecular evolution of L
 2149 Photopigment opsin in Admiral (*Limenitis*) butterflies, responsible for
 2150 colour vision in the long-wavelength part of the visual spectrum. Let-
 2151 terSpace=10FRENTIU *et al.* found that the sensitivity of this opsin had

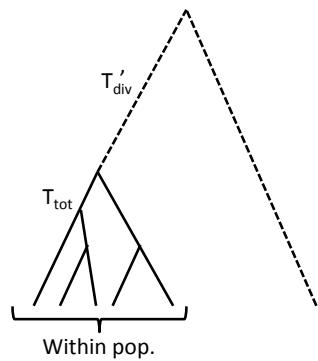


Figure 3.34: An example ogene genealogy for a set of alleles sampled within a population and a single allele sampled from a distantly-related species.

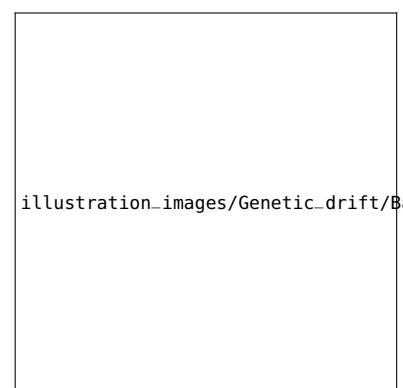


Figure 3.35: White admiral (*Limenitis arthemis*) and Viceroy (*Limenitis archippus*). *Basilarchia* is the old genus that these two species were originally placed in. Viceroy and Monarch butterflies are Müllerian mimics.

Field book of insects (1918). Lutz, F.E. . illustrations by Edna L. Beutenmüller. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

shifted towards blue-shifted in its sensitivity in *L. archippus archippus* (viceroy) compared to *L. arthemis astyanax*. To test whether this molecular evolution reflected positive selection they sequenced 24 *L. arthemis astyanax* individuals and one *L. archippus archippus* sequence. They identified 11 polymorphic sites in *L. arthemis astyanax* and 16 fixed differences, which break down as follows:

	Poly.	Fixed
Non-Syn.	2	12
Syn.	9	4
Ratio	2/9	3/1

Note the strong excess of non-synonymous to synonymous divergence compared to polymorphism (p-value of 0.006, Fisher's exact test), which is consistent with the gene evolving in an adaptive manner among the two species. We would expect roughly only 3 non-synonymous substitutions out of 16 substitutions if the gene was evolving neutrally ($16 \times 2/11$).

3.6 Neutral diversity and population structure

We've considered alleles drawn from a randomly-mating population, and divergence among alleles drawn from two distantly-related populations. We'll now turn to consider divergence among more closely related populations. In thinking about the coalescent within populations we made the assumption that any pair of lineages is equally likely to coalesce with each other. However, when there is population structure this assumption is violated.

We have previously written the measure of population structure F_{ST} as

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad (3.50)$$

where H_S is the probability that two alleles sampled at random from a subpopulation differ, and H_T is the probability that two alleles sampled at random from the total population differ.

A simple population split model Imagine a population of constant size of N_e diploid individuals that T generations in the past split into two daughter populations (sub-populations) each of size N_e individuals, which do not subsequently exchange migrants. In the current day we sample an equal number of alleles from both subpopulations.

Consider a pair of alleles sampled within one of our sub-populations and think about their per site heterozygosity. These alleles have experienced a population of size N_e and so the probability that they differ is $H_S \approx 4N_e\mu$ (assuming that $N_e\mu \ll 1$, using our equation 3.11 for heterozygosity within a population).

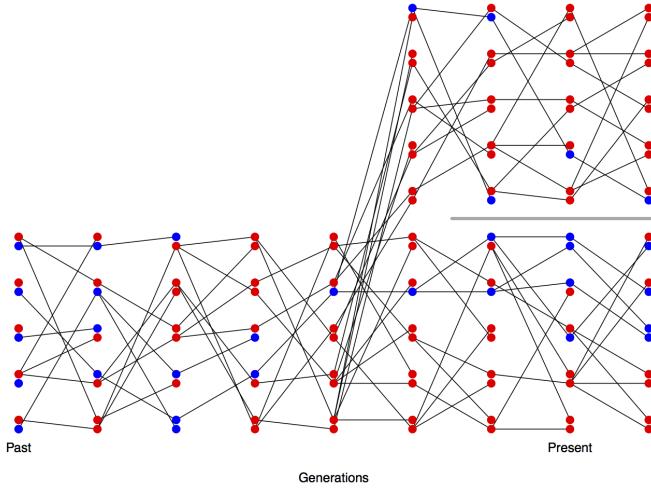


Figure 3.36: Change in allele frequencies following a population split. Code here.

2188 The heterozygosity in our total population is a little more tricky to
 2189 calculate. Assuming that we equally sample both sub-populations,
 2190 when we draw two alleles from our total sample, 50% of the time
 2191 they are drawn from the same subpopulation and 50% of the time
 2192 they are drawn from different subpopulations. Therefore, our total
 heterozygosity is given by

$$H_T = \frac{1}{2}H_S + \frac{1}{2}H_B \quad (3.51)$$

2194 where H_B is the probability that a pair of alleles drawn from our two
 2195 different sub-populations differ from each other. A pair of alleles
 2196 from different sub-populations cannot find a common ancestor with
 2197 each other for at least T generations into the past as they are in dis-
 2198 tinct populations (not connected by migration). Once our alleles find
 2199 themselves back in the combined ancestral population it takes them
 2200 on average $2N$ generations to coalesce. So the total opportunity for
 2201 mutation between our pair of alleles sampled from different popu-
 2202 lations is $2(T + 2N)$ generations of meioses, such that the probability
 that our pairs of alleles is different is

$$H_B \approx 2\mu(T + 2N) \quad (3.52)$$

2204 We can plug this into our expression for H_T , and then that in turn
 into F_{ST} . Doing so we find that

$$F_{ST} \approx \frac{\mu T}{\mu T + 4N_e \mu} = \frac{T}{T + 4N_e} \quad (3.53)$$

2206 Note that μ cancels out of this equation. In this simple toy model,
 2207 F_{ST} is increasing because the amount of between-population diversity
 2208 increases with the divergence time of the two populations (initially

linearly with T). F_{ST} grows at a rate give by $T/(4N_e)$ so that differentiation will be higher between populations separated by long divergence times or with small effective population sizes.

Question 12. The genome-wide F_{ST} between Bornean and Sumatran orang-utan species samples (*Pongo pygmaeus* and *Pongo abelii*) is ≈ 0.37 (LetterSpace=10LOCKE *et al.*, 2011), representing a deep population split between the species (potentially with little subsequent gene flow). Within the populations the genome-wide average Watterson's θ is $\theta_W = 1.4\text{kb}^{-1}$, estimated from the number of segregating sites. Assume a generation time of 20 years, and a mutation rate of 2×10^{-8} per base per generation. How far in the past did the two populations diverge?

A simple model of migration between an island and the mainland. We can also use the coalescent to think about patterns of differentiation under a simple model of migration-drift equilibrium. Let's consider a small island population that is relatively isolated from a large mainland population, where both of these populations are constant in size. We'll assume that the expected heterozygosity for a pair of alleles sampled on the mainland is H_M .

Our island has a population size N_I that is very small compared to our mainland population. Each generation some low fraction m of our individuals on the island have migrant parents from the mainland the generation before. Our island may also send migrants back to the mainland, but these are a drop in the ocean compared to the large population size on the mainland and their effect can be ignored.

If we sample an allele on the island and trace its ancestral lineage backward in time, each generation our ancestral allele has a low probability m of being descended from the mainland in the preceding generation (if we go back far enough the allele eventually has to be descended from an allele on the mainland). The probability that a pair of alleles sampled on the island are descended from a shared recent common ancestral allele on the island is the probability that our pair of alleles coalesces before either lineage migrates. For example, the probability that our pair of alleles coalesces $t + 1$ generations back on the island is

$$\frac{1}{2N_I} (1-m)^{2(t+1)} \left(1 - \frac{1}{2N_I}\right)^t \approx \frac{1}{2N_I} \exp\left(-t\left(\frac{1}{2N_I} + 2m\right)\right), \quad (3.54)$$

with the approximation following from assuming that $m \ll 1$ & $\frac{1}{(2N_I)} \ll 1$ (note that this is very similar to our derivation of heterozygosity above). The probability that our alleles coalesce before either

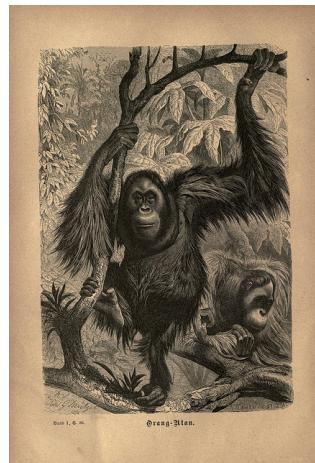


Figure 3.37: Orangutan (*Pongo*).
Brehms thierleben, allgemeine kunde des thierreichs. Brehm, A. E. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

2248 one of them migrates off the island, irrespective of the time, is

$$\int_0^\infty \frac{1}{2N_I} \exp\left(-t\left(\frac{1}{2N_I} + 2m\right)\right) dt = \frac{1/(2N_I)}{1/(2N_I) + 2m}. \quad (3.55)$$

2250 Let's assume that the mutation rate is very low such that it is very
2252 unlikely that the pair of alleles mutate before they coalesce on the
2254 island. Therefore, the only way that the alleles can be different from
2256 each other is if one or other of them migrates to the mainland, which
2258 happens with probability

$$1 - \frac{1/(2N_I)}{1/(2N_I) + 2m} \quad (3.56)$$

2254 Conditional on one or other of our alleles migrating to the mainland,
2256 both of our alleles represent independent draws from the mainland
2258 and so differ from each other with probability H_M . Therefore, the
2260 level of heterozygosity on the island is given by

$$H_I = \left(1 - \frac{1/(2N_I)}{1/(2N_I) + 2m}\right) H_M \quad (3.57)$$

2258 So the reduction of heterozygosity on the island compared to the
2260 mainland is

$$F_{IM} = 1 - \frac{H_I}{H_M} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{1 + 4N_I m}. \quad (3.58)$$

2260 The level of inbreeding on the island compared to the mainland will
2262 be high if the migration rate is low and the effective population size
2264 of the island is low, as allele frequencies on the island are drifting
2266 and diversity on the island is not being replenished by migration.
2268 The key parameter here is the number individuals on the island
2270 replaced by immigrants from the mainland each generation ($N_I m$).

2266 We have framed this problem as being about the reduction in
2268 genetic diversity on the island compared to the mainland. However,
2270 if we consider collecting individuals on the island and mainland in
2272 proportion to their population sizes, the total level of heterozygosity
2274 would be $H_T = H_M$, as samples from our mainland would greatly
2276 outnumber those from our island. Therefore, considering the island
2278 as our sub-population, we have derived another simple model of F_{ST} .

Question 13. You are investigating a small river population of
2274 sticklebacks, which receives infrequent migrants from a very large
2276 marine population. At a set of putatively neutral biallelic markers the
2278 freshwater population has frequencies:

0.2, 0.7, 0.8

2278 at the same markers the marine population has frequencies:
2280 0.4, 0.5 and 0.7.

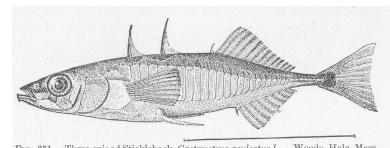


Fig. 3.38.—Three-spined Stickleback, *Gasterosteus aculeatus* L., Woods Hole, Mass.

Figure 3.38: Three-spined stickleback (*Gasterosteus aculeatus*). Illustration by Elizabeth Gould. Jordan, David Starr (1907) Fishes, New York City, NY: Henry Holt and Company. Image from Wikimedia Commons Public domain.

2280 From studying patterns of heterozygosity at a large collection
of markers, you have estimated the long term effective size of your
2282 freshwater population is 2000 individuals.

2284 What is your estimate of the migration rate from the marine popu-
lations into the river?

2286 *Incomplete lineage sorting* Because it can take a long time for an poly-
morphism to drift up or down in frequency, multiple population
splits may occur during the time an allele is still segregating. This
2288 can lead to incongruence between the overall population tree and the
information about relationships present at individual loci. In Figure
2290 3.38 and 3.39 we show a simulations of three populations where the
bottom population splits off from the other two first, followed by the
2292 subsequent splitting of the the top and the middle populations. We
start both simulations with a newly introduced red allele being poly-
2294 morphic in the combined ancestral population. The most likely fate
of this allele is that it is quickly lost from the population, but some-
2296 times the allele can drift up in frequency and be polymorphic when
the populations split, as the alleles in our two figures have done. If
2298 the allele is lost/fixed in the descendant populations before the next
population split, our allele configuration will agree with the popula-
2300 tion tree, as it does in Figure 3.38, and so too the gene tree will agree
with population tree (as shown in the left side of Figure 3.40). How-
2302 ever, if the allele persists as a polymorphism in the ancestral popula-
tion till the top and the middle populations split, then the allele can
2304 fix in one of these populations and not the other. Such an event can
lead to a substitution pattern that disagrees with the population tree,
2306 as in Figure 3.39. If we were to construct a phylogeny using the varia-
tion at this site we would see a disagreement between the gene tree
2308 and population tree. In Figure 3.39 an allele drawn from the top and
the bottom populations are necessarily more closely related to each
2310 other than either is to an allele drawn from population 2; tracing our
allelic lineages from the top and bottom populations back through
2312 time, they must coalesce with each other before we reach the point
where the red mutation arose; in contrast, a lineage from the middle
2314 population cannot have coalesced with either other lineage until past
the time the red mutation arose. An example of this ‘incomplete lin-
2316 eage sorting’ in terms of the underlying tree is shown on the right
side of Figure 3.40 .

2318 A natural pedigree analogy to incomplete lineage sorting is the
fact that while two biological siblings are more closely related to
2320 each other genealogically than either is to their cousin, at any given
locus one of the siblings can share an allele IBD with their cousin
2322 that they do not share with their own sibling, due to the randomness

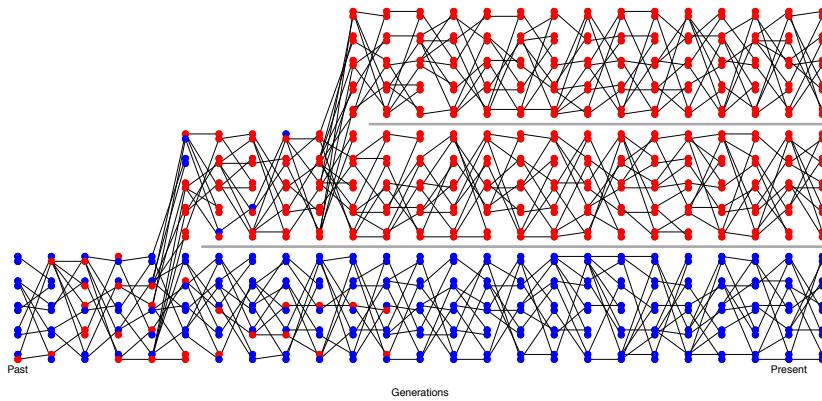


Figure 3.39: An example of alleles assorting among three populations such that there is no incomplete lineage sorting. [Code here.](#)

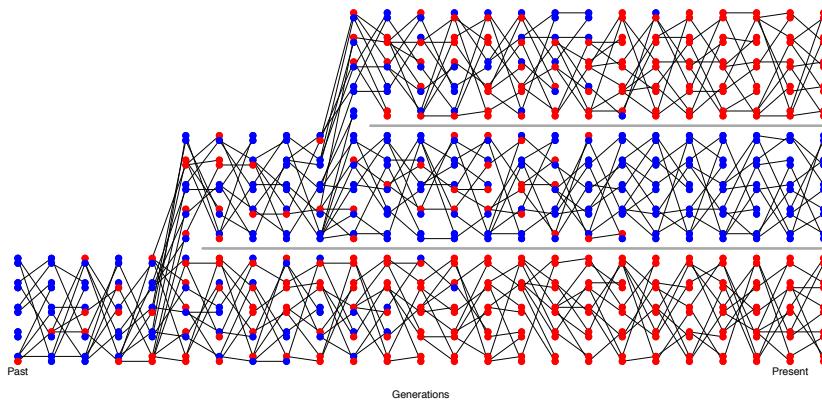


Figure 3.40: An example of alleles assorting among three populations leading to incomplete lineage sorting. [Code here.](#)



Figure 3.41: The population tree of three populations ((A, B), C) is shown blocked out with black shapes. Two different coalescent trees are relating a single allele drawn from A, B, and C are shown with thinner lines.

of Mendelian segregation down their pedigree. In these cases, the
2324 average relatedness of the individuals/populations disagrees with
the patterns of relatedness at a particular locus.

2326 As an empirical example of incomplete lineage sorting, let's consider
the work of LetterSpace=10JENNINGS and LetterSpace=10EDWARDS
2328 who sequenced a single allele from three different species of Aus-
tralian grass finches (*Poephila*): two sister species of long-tailed
2330 finches (*Poephila acuticauda* and *P. hecki*) and the black-throated finch
(*Poephila cincta*, see Figure 3.41). They collected sequence data for 30
2332 genes, and constructed phylogenetic gene trees at each of these loci,
resulting in 28 well-resolved gene trees. 16 of the gene trees showed
2334 *P. acuticauda* and *P. hecki* as sisters with *P. cincta* (the tree ((A,H),C)),
while for twelve genes the gene tree was discordant with the pop-
2336 ulation tree: for seven of their genes *P. hecki* fell as an outgroup to
the other two and at five *P. acuticauda* fell as an outgroup (the trees
2338 ((A,C),H) and ((H,C),A) respectively).

Let's use the coalescent to understand this discordance between
2340 gene trees and species trees. Let's assume that two sister populations
(A & B) split t_1 generations in the past, with a deeper split from a
2342 third outgroup population (C) t_2 generations in the past. We'll as-
sume that there's no gene flow among our populations after each
2344 split. We can trace back the ancestral lineages of our three alleles.
The first opportunity for the A & B lineages to coalesce is t_1 gener-
2346 ations ago. If they coalesce with each other in their shared ancestral
population before t_2 in the past (left side of Figure 3.40) their gene
2348 tree will definitely agree with the population tree. So the only way
for the gene tree to disagree with the population tree is for the A &
2350 B lineages to fail to coalesce in their shared ancestral population be-
tween t_1 and t_2 ; this happens with probability $(1 - 1/2N)^{t_2 - t_1}$. We'll
2352 get a discordant gene tree if A & B make it back to the shared ances-
tral population with C without coalescing, and then one or the other
2354 of them coalesces with the C lineage before they coalesce with each
other. This happens with probability 2/3, as at the first pairwise-
2356 coalescent event there are are three possible pairs of lineages that
could coalesce, two of which (A & C and B & C) result in a dis-
cordant tree. So the probability that we get a coalescent tree that is
2358 discordant with the population tree is

$$\frac{2}{3} (1 - 1/2N)^{t_2 - t_1}. \quad (3.59)$$

2360 Thus we should expect gene-tree population-tree discordance when
populations split in rapid succession and/or population sizes are
2362 large.

Question 14. Let's return to LetterSpace=10JENNINGS and Let-



Figure 3.42: Banded Grass Finch (*P. cincta*). Illustration by Elizabeth Gould. Birds of Australia Gould J. 1840. CC BY 4.0 uploaded to Flickr by rawpixel.com.

²³⁶⁴ terSpace=10EDWARDS's Australian grass finches example. They estimated that the ancestral population size of our two long-tailed
²³⁶⁶ finches was four hundred thousand. What is your best estimate of the inter-speciation time, i.e. $t_2 - t_1$?

²³⁶⁸ *Testing for gene flow.* We often want to test whether gene flow has occurred between populations. For example, we might want to establish a case that interbreeding between humans and Neanderthals occurred or demonstrate that gene flow occurred after two populations began to speciate. A broad range of methods have been designed to test for gene flow and to estimate gene flow rates, based on neutral expectations. Here we'll briefly just discuss one method based on some simple coalescent ideas. Above we assumed that gene-tree population-tree discordance was due to incomplete lineage sorting due to populations rapidly splitting. However, gene flow among populations can also lead to gene-tree discordance. While both ILS and gene flow can lead to discordance, under simplifying assumptions, ILS implies more symmetry in how these discordances manifest themselves.

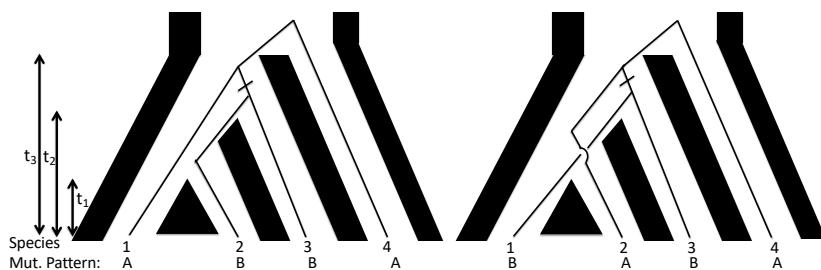


Figure 3.43: In both the left and right trees ILS has occurred between our single lineages sampled from populations A, B, and C. Imagine that population D is a somewhat distant outgroup such that the lineages from A through C (nearly) always coalesce with each other before any coalesce with D. The small dash on the branch indicates the mutation A→B occurring, giving rise to the ABBA or BABA mutational pattern shown at the bottom.

²³⁸² Take a look at Figure 3.42. In both cases the lineages from A and B fail to coalesce in their initial shared ancestral population, and one or the other of them coalesces with the lineage from C before they coalesce with each other. Each option is equally likely; therefore the mutational patterns ABBA and BABA are equally likely to occur under ILS.⁶

²³⁸⁸ However, if gene flow occurs from population C into population B, in addition to ILS the lineage from B can more recently coalesce with the lineage from C, and so we should see more ABBAAs than BABAs. To test for this effect of gene flow, we can sample a sequence from each of our 4 populations and count up the number of sites that show the two mutational patterns consistent with the gene-tree discordance n_{ABBA} and n_{BABA} and calculate

$$\frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}} \quad (3.60)$$

⁶ here we have to assume no structure in the ancestral population.

This statistic will have expectation zero if the gene-tree discordance is due to ILS and will be skewed negative if gene flow occurred from C into B (and skewed positive if gene flow occurred from C into A).
2396

Phenotypic Variation and the Resemblance Between Relatives

LetterSpace=10THE DISTINCTION BETWEEN GENOTYPE AND PHENOTYPE is one of the most useful ideas in Biology.¹ The genotype of an individual (the genome), for most purposes, is decided when the sperm fertilizes egg. The phenotype of an individual represents any measurable aspect of an organism.

Your height, to the amount of RNA transcribed from a given gene, to what you ate last Tuesday: all of these are phenotypes. Nearly any phenotype we can choose to measure about an organism represents the outcome of the information encoded by their genome played out through an incredibly complicated developmental, physiological and/or behavioural processes that in turn interact with a myriad of environmental and stochastic factors. Honestly it boggles the mind how organisms work as well as they do, let alone that I managed to eat lunch last Tuesday.

There are many different ways to think about studying the path from genotype through to phenotype. The one we will take here is to think about how phenotypic variation among individuals in a population arises as a result of genetic variation in the population. One simple way to measure this genotype-phenotype relationship is to calculate the phenotypic mean for each genotype at a locus. For example, LetterSpace=10WANG *et al.* (2018) explored the genetic basis of budset time in European aspen (*Populus tremula*); the effect of one specific SNP on that phenotype is shown in Figure 4.2. Budset timing is a key trait underlying local adaptation to varying growing season length. The associated SNP falls in a gene (*PtFT2*) that is known to play a strong role in flowering time regulation in other plants.

One way for us to assess the relationship between genotype and phenotype is to find the best fitting linear line through the data, i.e.

¹ LetterSpace=10JOHANNSEN, W., 1911
The Genotype Conception of Heredity.
The American Naturalist 45(531): 129–159

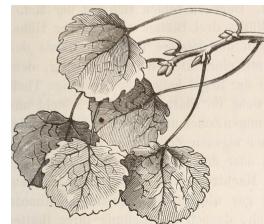


Figure 4.1: European aspen *P. tremula*.
Der baum. H. Schacht. 1860. BHL Image from the Biodiversity Heritage Library. Contributed by The Library of Congress. Not in copyright.

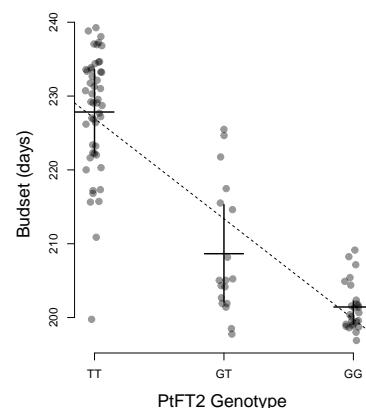


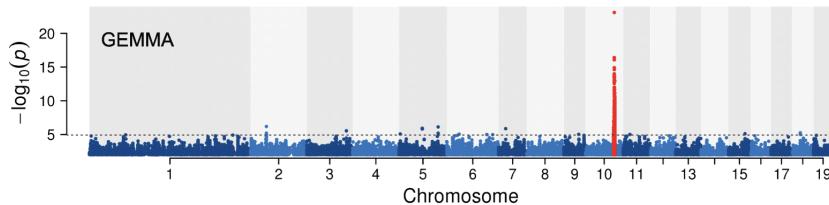
Figure 4.2: The effect of a flowering time gene (*PtFT2*) SNP on budset time in European aspen. Each dot gives the genotype-phenotype combination for an individual. The horizontal lines give the budset mean for each genotype and the vertical lines show the inter-quartile range. The dotted line gives the linear regression of phenotype on genotype. Thanks to Pär Ingvarsson for sharing these data from LetterSpace=10WANG *et al.* (2018).

2430 fit a linear regression of phenotypes for our individuals on their
genotypes at a particular SNP (l):

$$X \sim \mu + a_l G_l \quad (4.1)$$

2432 In the equation above, X is a vector of the phenotypes of a set of in-
dividuals and G_l is our vector of genotypes at locus l , with $G_{i,l}$ taking
2434 the value 0, 1, or 2 depending on whether our individual i is ho-
mozygote, heterozygote, or the alternate homozygote at our locus of
2436 interest. Here μ is our phenotypic mean. The slope of this regression
line (a_l) has the interpretation of being the average effect of substitut-
2438 ing a copy of allele 2 for a copy of allele 1. In our Aspen example the
slope is -13.6 , i.e. swapping a single T for a G allele moves the bud-
2440 set forward by 13.6 days, such that the GG homozygote is predicted
to set buds 27.2 days earlier than the TT homozygote.

2442 As a measure of the significance of this genotype-phenotype rela-
tionship, we can calculate the p-value of our regression. To try and
2444 identify loci that are associated with our trait genome-wide, we can
conduct this regression at each SNP we genotype in the genome.
2446 One common way to display the results of such an analysis (called
a genome-wide association study or GWAS for short) is to plot the
2448 logarithm of the p-value for each SNP along genome (a so-called
Manhattan plot). Here's one from LetterSpace=10WANG *et al.* (2018)
2450 for their Aspen budset phenotype



The SNP with the most significant p-value is the PtFT2 SNP. Note
2452 that other SNPs in the surrounding region also light up as showing
a significant association with budset timing. This is because loci that
2454 are in LD with a functional locus may in turn show an association,
not because they directly affect the phenotype, but simply because
2456 the genotypes at the two loci are themselves non-randomly associ-
ated. Below is a zoomed in version (Figure 2 in LetterSpace=10WANG
2458 *et al.* (2018)) with SNPs coloured by the strength of their LD with the
putatively functional SNP. Note how SNPs in strong LD with the
2460 functional allele (redder points) have more significant p-values.

Variation in some traits seems to have a relatively simple genetic
2462 basis. In our Aspen example there is one clear large-effect locus,
which explains 62% of the variation in budset. Note that even in this

Figure 4.3: Manhattan plot of the p-
value of the linear association between
genotype and budset in Aspen. Each
dot represents the test at a single SNP,
plotted at its physical coordinate in the
genome. Different chromosomes are
plotted in alternating colours. The SNPs
surrounding the PtFT2 gene are shown
in red. From LetterSpace=10WANG *et al.*
(2018), licensed under CC BY 4.0.

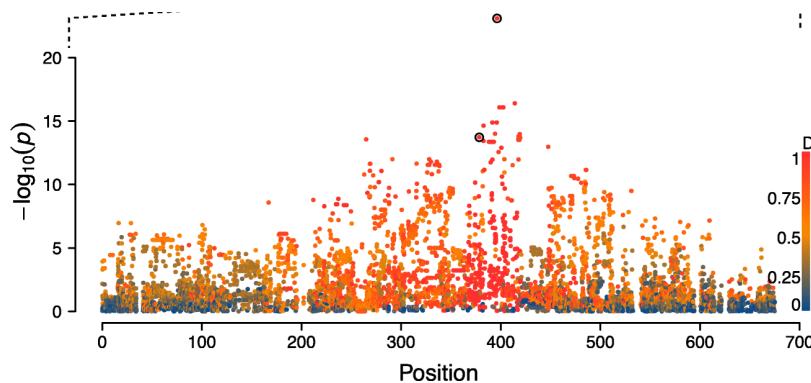


Figure 4.4: The Manhattan plot zoomed in on the top-hit (red SNPs from Figure 4.3). SNPs are now coloured by their D' value with the most significant SNP. D' is the LD covariance between a pair of loci (D) normalized by the largest value D can take given the allele frequencies. Figure from LetterSpace=10WANG *et al.* (2018), licensed under CC BY 4.0.

case, where we have an allele with a very strong effect on a phenotype, this is not an allele *for* budget, nor is PtFT2 a gene *for* budget. It is an allele that is associated with budget in the sampled environments and populations. In a different set of environments, this allele's effects may be far smaller, and a different set of alleles may contribute to phenotype variation. PtFT2, the gene our focal SNP falls close to, is just one of many genes and molecular pathways involved in budget. A mutant screen for budget may uncover many genes with larger effects; this gene is just a locus that happens to be polymorphic in this particular set of genotyped individuals.

While phenotypic variation for some phenotypes has a relatively simple genetic basis, many phenotypes are likely much more genetically complex, involving the functional effect of many alleles at hundreds or thousands of polymorphic loci. For example hundreds of small effect loci affecting human height have been mapped in European populations to date. Such genetically complex traits are called polygenic traits.

In this chapter, we will use our understanding of the sharing of alleles between relatives to understand the phenotypic resemblance between relatives in quantitative phenotypes. This will allow us to understand the contribution of genetic variation to phenotypic variation. In the next chapter, we will then use these results to understand the evolutionary change in quantitative phenotypes in response to selection.

4.0.1 A simple additive model of a trait

Let's imagine that the genetic component of the variation in our trait is controlled by L autosomal loci that act in an additive manner. The frequency of allele 1 at locus l is p_l , with each copy of allele 1 at this locus increasing your trait value by a_l above the population mean. The phenotype of an individual, let's call her i , is X_i . Her genotype

"All that we mean when we speak of a gene [allele] for pink eyes is, a gene which differentiates a pink eyed fly from a normal one —not a gene [allele] which produces pink eyes per se, for the character pink eyes is dependent on the action of many other genes." - LetterSpace=10STURTEVANT (1915)

2494 at SNP l is $G_{i,l}$. Here $G_{i,l} = 0, 1$, or 2 , representing the number of
 2495 copies of allele 1 she has at this SNP. Her expected phenotype, given
 2496 her genotype at all L SNPs, is then

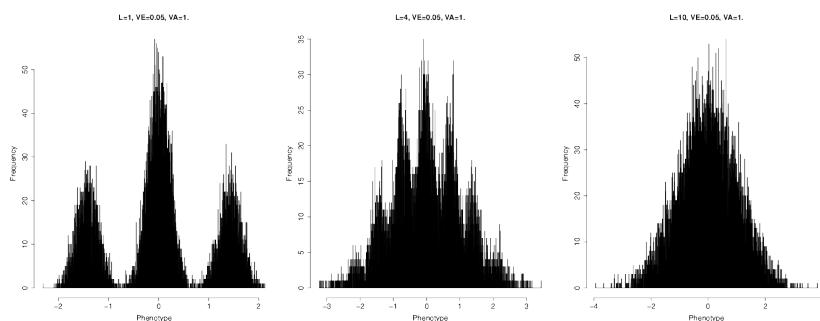
$$\mathbb{E}(X_i|G_{i,1}, \dots, G_{i,L}) = \mu + X_{A,i} = \mu + \sum_{l=1}^L G_{i,l}a_l \quad (4.2)$$

2498 where μ is the mean phenotype in our population, and $X_{A,i}$ is the de-
 2500 deviation away from the mean phenotype due to her genotype. Now in
 reality the phenotype is a function of the expression of those alleles
 2502 in a particular environment. Therefore, we can think of this expected
 phenotype as being an average across a set of environments that
 occur in the population.

When we measure our individual's observed phenotype we see

$$X_i = \mu + X_{A,i} + X_{E,i} \quad (4.3)$$

2504 where X_E is the deviation from the mean phenotype due to the envi-
 2506 ronment. This X_E includes the systematic effects of the environment
 our individual finds herself in and all of the noise during develop-
 2508 ment, growth, and the various random insults that life throws at our
 2510 individual. If a reasonable number of loci contribute to variation in
 our trait then we can approximate the distribution of $X_{A,i}$ by a nor-
 2512 mal distribution due to the central limit theorem (see Figure 4.5).
 Thus if we can approximate the distribution of the effect of envi-
 2514 ronmental variation on our trait ($X_{E,i}$) also by a normal distribution,
 which is reasonable as there are many small environmental effects,
 then the distribution of phenotypes within the population (X_i) will be
 normally distributed (see Figure 4.5).



2516 Note that as this is an additive model; we can decompose eqn. 4.3
 into the effects of the two alleles at each locus and rewrite it as

$$X_i = \mu + X_{iM} + X_{iP} + X_{E,i} \quad (4.4)$$

2518 where X_{iM} and X_{iP} are the contribution to the phenotype of the al-
 leles that our individual received from her mother (maternal alleles)

Figure 4.5: The convergence of the phenotypic distribution to a normal distribution. Each of the three histograms shows the distribution of the phenotype in a large sample, for increasingly large numbers of loci ($L = 1, 4$, and 10 , with the proportion of variance explained held at $V_A = 1$). I have simulated each individual's phenotype following equations 4.2 and 4.3. Specifically, we've simulated each individual's biallelic genotype at L loci, assuming Hardy-Weinberg proportions and that the allele is at 50% frequency. We assume that all of the alleles have equal effects and combine them additively together. We then add an environmental contribution, which is normally distributed with variance 0.05. Note that in the left two pictures you can see peaks corresponding to different genotypes due to our low environmental noise (in practice we can rarely see such peaks for real quantitative phenotypes). Code here.

2520 and father (paternal alleles) respectively. This will come in handy in
 just a moment when we start thinking about the phenotypic covariance
 2522 of relatives.

2524 Now obviously this model seems silly at first sight as alleles don't
 only act in an additive manner, as they interact with alleles at the
 same loci (dominance) and at different loci (epistasis). Later we'll
 2526 relax this assumption, however, we'll find that if we are interested
 in evolutionary change over short time-scales it is actually only the
 2528 "additive component" of genetic variation that will (usually) concern
 us. We will define this more formally later on, but for the moment
 2530 we can offer the intuition that parents only get to pass on a single
 allele at each locus on to the next generation. As such, it is the effect
 2532 of these transmitted alleles, averaged over possible matings, that is
 an individual's average contribution to the next generation (i.e. the
 2534 additive effect of the alleles that their genotype consists of).

4.0.2 Additive genetic variance and heritability

2536 As we are talking about an additive genetic model, we'll talk about
 the additive genetic variance (V_A), the phenotypic variance due to the
 2538 additive effects of segregating genetic variation. This is a subset of
 the total genetic variance if we allow for non-additive effects.

2540 The variance of our phenotype across individuals (V) we can write
 as

$$V = \text{Var}(X_A) + \text{Var}(X_E) = V_A + V_E \quad (4.5)$$

2542 In writing the phenotypic variance as a sum of the additive and environmental contributions, we are assuming that there is no covariance
 2544 between $X_{G,i}$ and $X_{E,i}$, i.e. there is no covariance between genotype
 and environment.

2546 Our additive genetic variance can be written as

$$V_A = \sum_{l=1}^L \text{Var}(G_{i,l}a_l) \quad (4.6)$$

2548 where $\text{Var}(G_{i,l}a_l)$ is the contribution of locus l to the additive variance among individuals. Assuming random mating, and that our loci are in linkage equilibrium, we can write our additive genetic variance
 2550 as

$$V_A = \sum_{l=1}^L a_l^2 2p_l(1 - p_l) \quad (4.7)$$

2552 where the $2p_l(1 - p_l)$ term follows from the binomial sampling of two alleles per individual at each locus.

Question 1. You have two biallelic SNPs contributing to variance
 2554 in human height. At the first SNP you have an allele with an additive

effect of 5cm which is found at a frequency of 1/10,000. At the second SNP you have an allele with an additive effect of -0.5cm segregating at 50% frequency. Which SNP contributes more to the additive genetic variance? Explain the intuition of your answer.

An example of calculating polygenic scores. Now we don't usually get to see the individual loci contributing to highly polygenic traits. Instead, we only get to see the distribution of the trait in the population. However, with the advent of GWAS in human genetics we can see some of the underlying genetics using the many trait-associated loci identified to date. Using the estimated effect sizes at each locus, each one of which is tiny, we can calculate the weighted sum over an individual's genotype as in equation 4.2. This weighted sum is called the individual's polygenic score. To illustrate how polygenic scores work, we can take a set of 1700 SNPs, each chosen as the SNP with the strongest signal of association with height in 1700 roughly independent bins spaced across the genome. The effects of these SNPs are tiny; the medium, absolute additive effect size is 0.07cm. Figure 4.6 shows the distribution of a thousand individuals' polygenic scores calculated using these 1700 SNPs (simulated genotypes using the UKBB frequencies). The standard deviation of these polygenic scores $\sim 2\text{cm}$. The individuals with higher polygenic scores for height are predicted to be taller than the individuals with lower polygenic scores.

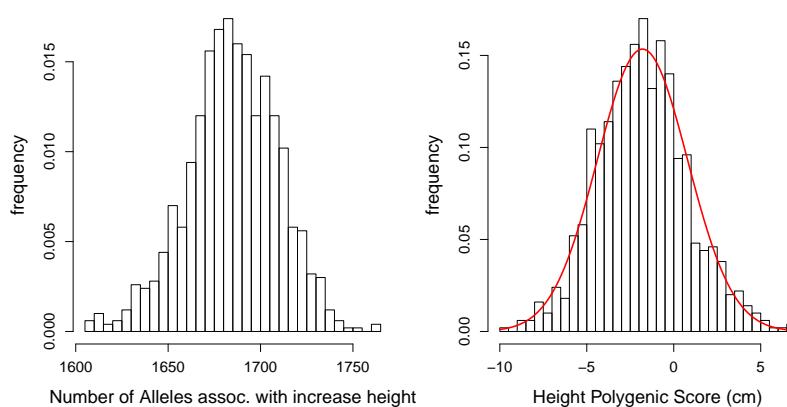


Figure 4.6: **Left)** The distribution of the number of height-increasing alleles that individuals carry at 1700 SNPs associated with height in the UK Biobank, for a sample of 1000 individuals. **right)** The distribution of the polygenic scores for these 1000 individuals. Plotted on top is a normal distribution with the same mean and variance. The empirical variance of these polygenic scores is 0.13, the additive genetic variance calculated by equation (4.7) is 0.135, so the two are in good agreement. Code here.

2578 The narrow sense heritability We would like a way to think about what proportion of the variation in our phenotype across individuals **2580** is due to genetic differences as opposed to environmental differences.

Such a quantity will be key in helping us think about the evolution
 2582 of phenotypes. For example, if variation in our phenotype had no
 genetic basis, then no matter how much selection changes the mean
 2584 phenotype within a generation the trait will not change over genera-
 tions.

2586 We'll call the proportion of the variance that is genetic the *heritabil-
 ity*, and denote it by h^2 . We can then write heritability as

$$h^2 = \frac{Var(X_A)}{V} = \frac{V_A}{V} \quad (4.8)$$

2588 Remember that we are thinking about a trait where all of the alleles
 act in a perfectly additive manner. In this case our heritability h^2
 2590 is referred to as the *narrow sense heritability*, the proportion of the
 variance explained by the additive effect of our loci. When we allow
 2592 dominance and epistasis into our model, we'll also have to define the
broad sense heritability (the total proportion of the phenotypic variance
 2594 attributable to genetic variation).

The narrow sense heritability of a trait is a useful quantity; indeed
 2596 we'll see shortly that it is exactly what we need to understand the
 evolutionary response to selection on a quantitative phenotype. We
 2598 can calculate the narrow sense heritability by using the resemblance
 between relatives. For example, if the phenotypic differences between
 2600 individuals in our population were solely determined by environ-
 mental differences experienced by these different individuals, we
 2602 should not expect relatives to resemble each other any more than ran-
 dom individuals drawn from the population. Now the obvious caveat
 2604 here is that relatives also share an environment, so may resemble
 each other due to shared environmental effects.

2606 Note that the heritability is a property of a sample from the pop-
 ulation in a particular set of environments at a particular time.

2608 Changes in the environment may change the phenotypic variance.
 Changes in the environment may also change how our genetic al-
 2610 leles are expressed through development and so change V_A . Thus
 estimates of heritability are not transferable across environments or
 2612 populations.

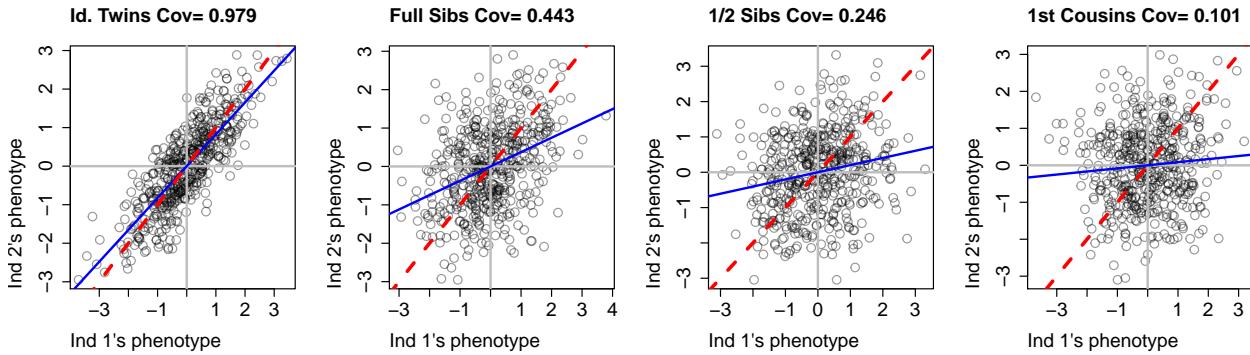
4.0.3 The covariance between relatives

2614 So we'll go ahead and calculate the covariance in phenotype between
 two individuals (1 and 2) who have phenotypes X_1 and X_2 respec-
 2616 tively. To think about imagine plotting the phenotypes of, say, sisters
 against each other. The x and y coordinates of each point will be the,
 2618 say, heights of the pair of siblings. Do tall women tend to have tall
 sisters, do short women tend to have short sisters? How much do
 2620 their phenotypes covary. If some of the variation in our phenotype is

genetic we expect identical twins to resemble each other more than
 2622 full siblings, who in turn will resemble each other more than half-sibs
 and so on out (see Figure 4.7). Under our simple additive model of
 2624 phenotypes we can write the covariance as

$$\text{Cov}(X_1, X_2) = \text{Cov}((X_{1M} + X_{1P} + X_{1E}), ((X_{2M} + X_{2P} + X_{2E})) \quad (4.9)$$

We can expand this out in terms of the covariance between the vari-
 2626 ous components in these sums.



To make our task easier, we will make two commonly made as-
 2628 sumptions:

1. We can ignore the covariance of the environments between indi-
 2630 viduals (i.e. $\text{Cov}(X_{1E}, X_{2E}) = 0$)
2. We can ignore the covariance between the environment of one
 2632 individual and the genetic variation in another individual (i.e.
 $\text{Cov}(X_{1E}, (X_{2M} + X_{2P})) = 0$). (We can actually incorporate these
 2634 effects in later if we choose too.)

The failure of these assumptions to hold can undermine our esti-
 2636 mates of heritability, but we'll return to that later. Moving forward
 with these assumptions, we can simplify our original expression
 2638 above and write our phenotypic covariance between our pair of indi-
 viduals as

$$\text{Cov}(X_1, X_2) = \text{Cov}((X_{1M}, X_{2M}) + \text{Cov}(X_{1M}, X_{2P}) + \text{Cov}(X_{1P}, X_{2M}) + \text{Cov}(X_{1P}, X_{2P}) \quad (4.10)$$

This equation is saying that, under our simple additive model, we
 2640 can see the covariance in phenotypes between individuals as the
 covariance between the maternal and paternal allelic effects in our in-
 2642 dividuals. We can use our results about the sharing of alleles between
 2644 relatives to obtain these covariance terms. But before we write down
 the general case, let's quickly work through some examples.

Figure 4.7: Covariance of phenotypes between pairs of individuals of a given relatedness. Each point gives the phenotypes of a different pair of individuals. The additive genetic variance is held constant at $V_A = 1$, such that the expected covariances ($2F_{1,2}V_A$) should be 1, 0.5, 0.25, and 0.125 respectively din good agreement with the empirical covariances reported in the title of each graph. The data were simulated as described in the caption of Figure 4.5. The blue line shows $x = y$ and the red line shows the best fitting linear regression line. Code here.

2646 *The covariance between identical twins* Let's first consider the case
 of a pair of identical twins from two unrelated parents. Our pair of
 2648 twins share their maternal and paternal allele identical by descent
 $(X_{1M} = X_{2M}$ and $X_{1P} = X_{2P}$). As their maternal and paternal
 2650 alleles are not correlated draws from the population, i.e. have no
 probability of being *IBD* as we've said the parents are unrelated,
 2652 the covariance between their effects on the phenotype is zero (i.e.
 $\text{Cov}(X_{1P}, X_{2M}) = \text{Cov}(X_{1M}, X_{2P}) = 0$). In that case, eqn. 4.10 is

$$\text{Cov}(X_1, X_2) = \text{Cov}((X_{1M}, X_{2M}) + \text{Cov}(X_{1P}, X_{2P}) = 2\text{Var}(X_{1M}) = V_A \quad (4.11)$$

2654 Now in general identical twins are not going to be super helpful
 for us in estimating h^2 , because under models with non-additive
 2656 effects, identical twins will have higher covariance than we'd expect
 just based on the alleles they share. This is because identical twins
 2658 don't just share alleles, they share their entire genotypes, and thus
 resemble each other in phenotype also because of shared dominance
 2660 effects.

2662 *The covariance in phenotype between mother and child* If a mother and
 father are unrelated individuals (i.e. are two random draws from
 the population) then this mother and her child share one allele IBD
 2664 at each locus (i.e. $r_1 = 1$ and $r_0 = r_2 = 0$). Half the time our
 mother (ind 1) transmits her paternal allele to the child (ind 2), in
 2666 which case $X_{P1} = X_{M2}$, and so $\text{Cov}(X_{P1}, X_{M2}) = \text{Var}(X_{P1})$, and
 all the other covariances in eqn. 4.10 are zero. The other half of the
 2668 time she transmits her maternal allele to the child, in which case
 $\text{Cov}(X_{M1}, X_{M2}) = \text{Var}(X_{M1})$ and all the other terms are zero. By this
 2670 argument, $\text{Cov}(X_1, X_2) = \frac{1}{2}\text{Var}(X_{M1}) + \frac{1}{2}\text{Var}(X_{P1}) = \frac{1}{2}V_A$.

2672 *The covariance between general pairs of relatives under an additive model*
 The two examples above make clear that to understand the covariance
 between phenotypes of relatives, we simply need to think about
 2674 the alleles they share IBD. Consider a pair of relatives (1 and 2) with
 a probability r_0 , r_1 , and r_2 of sharing zero, one, or two alleles IBD
 2676 respectively. When they share zero alleles $\text{Cov}((X_{1M} + X_{1P}), (X_{2M} +$
 $X_{2P})) = 0$, when they share one allele $\text{Cov}((X_{1M} + X_{1P}), (X_{2M} +$
 2678 $X_{2P})) = \text{Var}(X_{1M}) = \frac{1}{2}V_A$, and when they share two alleles
 $\text{Cov}((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = V_A$. Therefore, the general covariance
 2680 between two relatives is

$$\text{Cov}(X_1, X_2) = r_0 \times 0 + r_1 \frac{1}{2}V_A + r_2 V_A = 2F_{1,2}V_A \quad (4.12)$$

So under a simple additive model of the genetic basis of a phenotype, to measure the narrow sense heritability we need to measure

the covariance between pairs of relatives (assuming that we can remove the effect of shared environmental noise). From the covariance between relatives we can calculate V_A , and we can then divide this by the total phenotypic variance to get h^2 .

Question 2. A) In polygynous red-winged blackbird populations (i.e. males mate with several females), paternal half-sibs can be identified. Suppose that the covariance of tarsus lengths among half-sibs is 0.25 cm^2 and that the total phenotypic variance is 4 cm^2 . Use these data to estimate h^2 for tarsus length in this population.

B) Why might paternal half-sibs be preferable for measuring heritability than maternal half-sibs?

Parent-midpoint offspring regression Another way that we can estimate the narrow sense heritability is through the regression of child's phenotype on the parental mid-point phenotype. The parental mid-point phenotype is simply the average of the mum and dad's phenotype. We denote the child's phenotype by X_{kid} and mid-point phenotype by X_{mid} , so that if we take the regression $X_{kid} \sim X_{mid}$ this regression has slope $\beta = \text{Cov}(X_{kid}, X_{mid}) / \text{Var}(X_{mid})$. The covariance of $\text{Cov}(X_{kid}, X_{mid}) = \frac{1}{2}V_A$, and $\text{Var}(X_{mid}) = \frac{1}{2}V$, as by taking the average of the parents we have halved the variance, such that the slope of the regression is

$$\beta_{mid,kid} = \frac{\text{Cov}(X_{kid}, X_{mid})}{\text{Var}(X_{mid})} = \frac{V_A}{V} = h^2 \quad (4.13)$$

i.e. the regression of the child's phenotype on the parental midpoint phenotype is an estimate of the narrow sense heritability. This way of estimating heritability has the problem of not controlling for environmental correlations between relatives. But it's a useful way to think about heritability and will be directly relevant to our discussion of the response to selection in the next chapter.

Our regression allows us to attempt to predict the phenotype of the child given the phenotypes of the parents; how well we can do this depends on the slope. If the slope is close to zero then the parental phenotypes hold no information about the phenotype of the child, while if the slope is close to one then the parental mid-point is a good guess at the child's phenotype.

More formally, the expected phenotype of the child given the parental phenotypes is

$$\mathbb{E}(X_{kid}|X_{mum}, X_{dad}) = \mu + \beta_{mid,kid}(X_{mid} - \mu) = \mu + h^2(X_{mid} - \mu) \quad (4.14)$$

which follows from the definition of linear regression. So to find the child's predicted phenotype, we simply take the mean phenotype



Figure 4.8: Red-winged blackbird and Tricoloured Red-winged blackbirds (*Agelaius phoeniceus* and *Agelaius tricolor*).

Bird-lore (1899). National Association of Audubon Societies for the Protection of Wild Birds and Animals. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

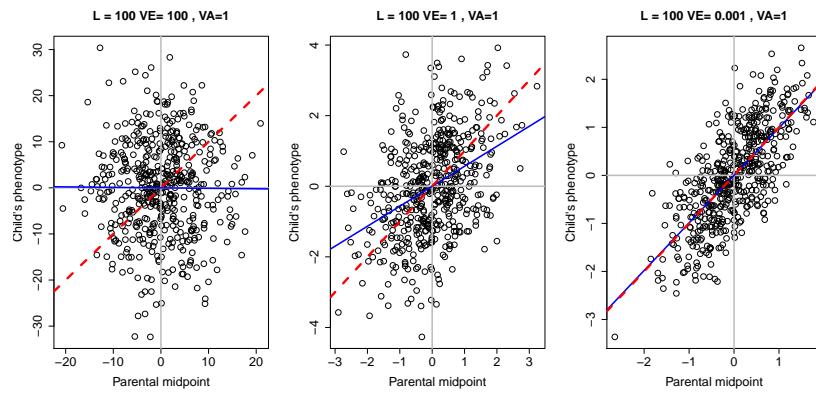


Figure 4.9: Regression of child's phenotype on the parental mid-point phenotype. The three panels show decreasing levels of environmental variance (V_E) holding the additive genetic variance constant ($V_A = 1$). In these figures, we simulate 100 loci, as described in the caption of Figure 4.5. We simulate the genotypes and phenotypes of the two parents, and then simulate the child's genotype following mendelian transmission. The blue line shows $x = y$ and the red line shows the best fitting linear regression line. Code here.

2720 and add on the difference between our parental mid-point and the population mean, multiplied by our narrow sense heritability.

2722 **Question 3.** Briefly explain what Galton meant by 'regression towards mediocrity', and why he observed this pattern in light of
 2724 Mendelian inheritance.

2726 *Estimating additive genetic variance across a variety of different relationships.* In many natural populations we may have access to individuals with a range of different relationships to each other (e.g. through
 2728 monitoring of the paternity of individuals), but relatively few pairs of individuals for a specific relationship (e.g. sibs). We can try and use
 2730 this information on various relatives as fully as possible in a mixed model framework. Building from equation 4.3, we can write an individual's phenotype X_i as
 2732

$$X_i = \mu + X_{A,i} + X_{E,i} \quad (4.15)$$

2734 where $X_{E,i} \sim N(0, V_E)$ and $X_{A,i}$ is normally distributed across individuals with covariance matrix $V_A A$, where the entries for a pair of individuals i and j are $A_{ij} = 2F_{i,j}$ and $A_{ii} = 1$. Given the matrix
 2736 A we can estimate V_A . We can also add fixed effects into this model
 2738 to account for generation effects, additional mixed effects could also be included to account for shared environments between particular individuals (e.g. a shared nest). This approach is sometimes called
 2740 the "animal model".

4.1 Multiple traits

2742 Traits often covary with each other, both due to environmentally induced effects (e.g. due to the effects of diet on multiple traits) and
 2744 due to the expression of underlying genetic covariance between traits.

Genetic covariance, in turn, can reflect pleiotropy, a mechanistic effect
 2746 of an allele on multiple traits (e.g. variants that affect skin pigmentation often affect hair color), the genetic linkage of loci independently
 2748 affecting multiple traits, or the effects of assortative mating.

Consider two traits $X_{1,i}$ and $X_{2,i}$ in an individual i . These traits
 2750 could be, say, the individual's leg length and nose length. As before,
 we can write these as

$$\begin{aligned} X_{1,i} &= \mu_1 + X_{1,A,i} + X_{1,E,i} \\ X_{2,i} &= \mu_2 + X_{2,A,i} + X_{2,E,i} \end{aligned} \quad (4.16)$$

2752 As before we can talk about the total phenotypic variance (V_1, V_2),
 environmental variance ($V_{1,E}$ and $V_{2,E}$), and the additive genetic
 2754 variance for trait one and two ($V_{1,A}, V_{2,A}$). But now we also have
 to consider the total covariance between trait one and trait two,
 2756 $V_{1,2} = \text{Cov}(X_1, X_2)$, as well as the environmentally induced co-
 variance ($V_{E,1,2} = \text{Cov}(X_{1,E}, X_{2,E})$) and the additive genetic covariance
 2758 ($V_{A,1,2} = \text{Cov}(X_{1,A}, X_{2,A})$). To better understand the covariance arising
 due to pleiotropy, let's think about a set of L SNPs contributing
 2760 to our two traits. If the additive effect of an allele at the i^{th} SNP is $\alpha_{i,1}$
 and $\alpha_{i,2}$ on traits 1 and 2, then the additive covariance between our
 2762 traits is

$$V_{A,1,2} = \sum_{i=1}^L 2\alpha_{i,1}\alpha_{i,2}p_i(1-p_i) \quad (4.17)$$

2764 assuming our loci are in linkage disequilibrium. Thus a genetic correlation arises due to pleiotropy, because loci that tend to affect trait 1
 also systematically affect trait 2. For example, alleles associated with
 2766 later Age at Menarche (AAM) in European females also tend to be
 positively associated with height (see Figure 4.10), thereby creating a
 2768 genetic correlation between AAM and height.

We can store our variance and covariance values in matrices, a
 2770 way of gathering these terms that will be useful when we discuss
 selection:

$$\mathbf{V} = \begin{pmatrix} V_1 & V_{1,2} \\ V_{1,2} & V_2 \end{pmatrix} \quad (4.18)$$

2772 and

$$\mathbf{G} = \begin{pmatrix} V_{1,A} & V_{A,1,2} \\ V_{A,1,2} & V_{2,A} \end{pmatrix} \quad (4.19)$$

Here we've shown the matrices for two traits, but we can generalize
 2774 this to an arbitrary number of traits.

We can estimate these quantities, in a similar way as before, by
 2776 studying the covariance in different traits between relatives:

$$\text{Cov}(X_{1,i}, X_{2,j}) = 2F_{i,j}V_{A,1,2} \quad (4.20)$$

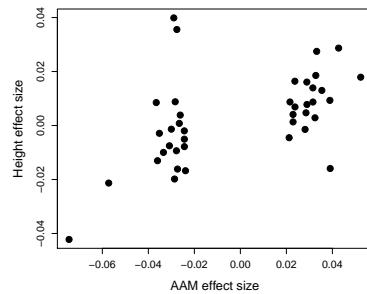


Figure 4.10: The additive effect sizes of loci associated with female Age at Menarche (AAM) and their effect size on Height in a European population. Data from LetterSpace=10PICKRELL *et al.* (2016). Code here.

We can also talk about the genetic correlation between two phenotypes

$$r_g = \frac{V_{A,1,2}}{\sqrt{V_{A,1}V_{A,2}}} \quad (4.21)$$

where $V_{A,1}$ and $V_{A,2}$ are the additive genetic variance for trait 1 and

2 respectively. Here, r_g tells us to what extent the additive genetic variance in two traits is correlated.

One type of genetic covariance we often think about is the covariance of male and female phenotypes. For example, below is the relationship between the forehead patch size for Pied fly-catcher fathers and their sons and daughters. The phenotype has been standardized to have mean 0 and variance 1 in each group. The phenotypic covariance of the sample of fathers and sons is 0.35, while the phenotypic covariance of fathers and daughter is 0.23.

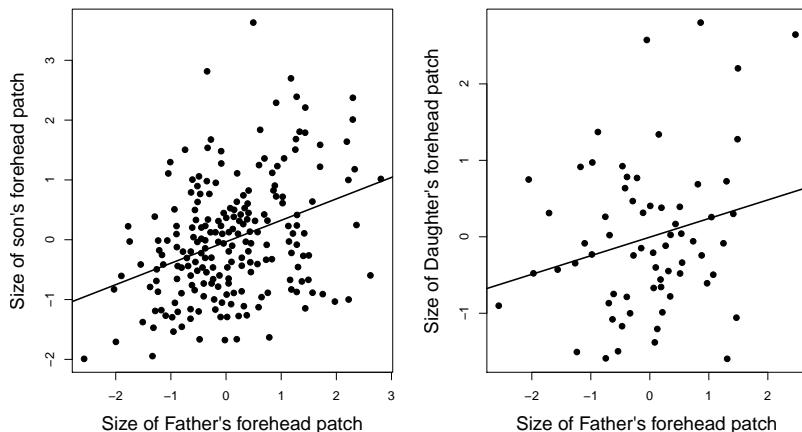


Figure 4.11: Relationship of standardized forehead patch size between fathers and sons and daughters in Pied fly-catchers. Data from LetterSpace=10POTTI and LetterSpace=10CANAL. Code here.



Figure 4.12: *Ficedula hypoleuca*, Pied fly-catcher.

Coloured illustrations of British birds, and their eggs (1842-1850). London :G.W. Nicklinson. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

Question 4. Assume we can ignore the effect of the shared environment in our Pied fly-catcher example.

A) What is the additive genetic covariance between male and female patch size?

B) What is the additive genetic correlation of male and female patch size? You can assume that the additive genetic variance is the same in males and females.

2796 4.1.1 Non-additive variation.

Up to now we've assumed that our alleles contribute to our phenotype in an additive fashion. However, that does not have to be the case as there may be non-additivity among the alleles present at a

2800 locus (*dominance*) or among alleles at different loci (*epistasis*). We can
 2801 accommodate these complications into our models. We do this by
 2802 partitioning our total genetic variance into independent variance
 2803 components.

2804 *Dominance.* To understand the effect of dominance, let's consider
 2805 how the allele that a parent transmits influences their offspring's
 2806 phenotype. A parent transmits one of their two alleles at a locus to
 2807 their offspring. Assuming that individuals mate at random, this allele
 2808 is paired with another allele drawn at random from the population.
 2809 For example, assume your mother transmitted an allele 1 to you:
 2810 with probability p it would be paired with another allele 1, and you
 2811 would be a homozygote; and with probability q it's paired with a 2
 2812 allele and you're a heterozygote.

Now consider an autosomal biallelic locus ℓ , with frequency p for
 2814 allele 1, and genotypes 0, 1, and 2 corresponding to how many copies
 2815 of allele 1 individuals carry. We'll denote the mean phenotype of an
 2816 individual with genotype 0, 1, and 2 as $\bar{X}_{\ell,0}$, $\bar{X}_{\ell,1}$, $\bar{X}_{\ell,2}$ respectively.
 This mean is taking an average phenotype over all the environments
 2818 and genetic backgrounds the alleles are present on. We'll mean center
 2819 (MC) these phenotypic values, setting $\bar{X}'_{\ell,0} = \bar{X}_{\ell,0} - \mu$, and likewise
 2820 for the other genotypes.

We can think about the average (marginal) MC phenotype of an
 2822 individual who received an allele 1 from their parent as the average
 2823 of the MC phenotype for heterozygotes and 11 homozygotes, weighted
 2824 by the probability that the individual has these genotypes, i.e. the
 2825 probability they receive an additional allele 1 or an allele 2 from their
 2826 other parent:

$$a_{\ell,1} = p\bar{X}'_{\ell,2} + q\bar{X}'_{\ell,1}, \quad (4.22)$$

Similarly, if your parent transmitted an 2 allele to you, your average
 2828 MC phenotype would be

$$a_{\ell,2} = p\bar{X}'_{\ell,1} + q\bar{X}'_{\ell,0} \quad (4.23)$$

Let's now consider the average phenotype of an offspring of each
 2830 of our three genotypes

genotype:	0,	1,	2.
additive genetic value:	$a_{\ell,2} + a_{\ell,0}$,	$a_{\ell,1} + a_{\ell,0}$,	$a_{\ell,1} + a_{\ell,2}$

2832 i.e. the mean phenotype of each genotypes' offspring averaged over
 2833 all possible matings to other individuals in the population (assuming
 2834 individuals mate at random). These are the additive MC genetic val-
 2835 ues (breeding values) of our genotypes. Here we are simply adding
 2836 up the additive contributions of the alleles present in each genotype
 and ignoring any non-additive effects of genotype.

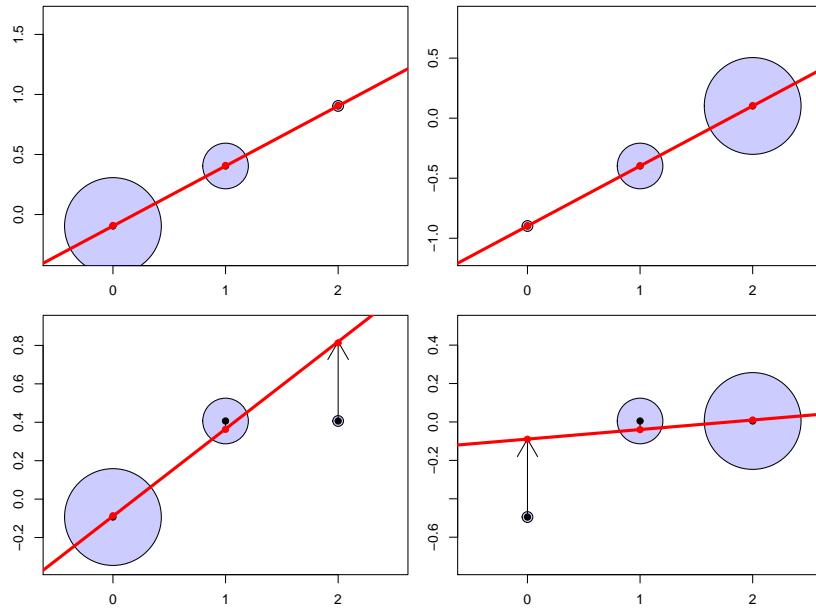


Figure 4.13: The average mean-centered (MC) phenotypes of each genotype. **Top Row:** Additive relationship between genotype and phenotype. **Bottom Row:** Allele 1 is dominant over allele 2, such that the heterozygote has the same phenotype as the 22 genotype (2). The area of each circle is proportion to the fraction of the population in each genotypic class (p^2 , $2pq$, and q^2). One the left column $p = 0.1$ and the right column is $p = 0.9$. The additive genetic values of the genotypes are shown as red dots. The regression between phenotype and additive genotype is shown as a red line. The black vertical arrows show the difference between the average MC phenotype and additive genetic value for each genotype. Code here.

2838 To illustrate this, in Figure 4.13 we plot two different cases of dominance relationships; in the top row an additive polymorphism and in
 2840 the second row a fully dominant allele. The additive genetic values
 2842 of the genotypes are shown as red dots. Note that the additive values
 2844 of the genotypes line up with the observed MC phenotypic means in
 2846 the top row, when our alleles interact in a completely additive manner.
 2848 Our additive genetic values always fall along a linear line (the red line in our figure). The additive values are falling along the best
 2850 fitting line of linear regression for our population, when phenotype
 2852 is regressed against the additive genotype (0, 1, 2 copies of allele
 2854 1) across all individuals in our population. Note in the dominant
 2856 case the additive genetic values differ from the observed phenotypic
 2858 means, and are closer to the observed values for the genotypes that
 2860 are most common in the population.

2852 The difference in the additive effect of the two alleles $a_{\ell,2} - a_{\ell,1}$
 2854 can be interpreted as an average effect of swapping an allele 1 for an
 2856 allele 2; we'll call this difference $\alpha_\ell = a_{\ell,2} - a_{\ell,1}$. Our α_ℓ is also the
 2858 slope of the regression of phenotype against genotype (the red line
 2860 in Figure 4.13). Note that the slope of our regression of phenotype
 2862 on genotype (α_ℓ) does not depend on the population allele frequency
 2864 for our completely additive locus (top row of 4.13). In contrast, when
 2866 there is dominance, the slope between genotype and phenotype (α_ℓ)
 2868 is a function of allele frequency (bottom row of 4.13). When a domi-
 2870 nant allele (1) is rare there is a strong slope of phenotype on geno-
 2872 type, bottom left Figure 4.13. This strong slope is because replacing

a single copy of the 2 allele with a 1 allele in an individual has a big effect on average phenotype, as it will most likely move an individual from being a 22 homozygote to being a 12 heterozygote. In contrast, when the dominant allele (1) is common in the population, replacing a 2 allele by a 1 allele in an individual on average has little phenotypic effect, leading to a weak slope bottom right Figure 4.13. This small effect is because as we are mainly turning heterozygotes into homozygotes (11), who have the same mean phenotype as each other.

As an example of how dominance and population allele frequencies can change the additive effect of an allele, let's consider the genetics of the age of sexual maturity in Atlantic Salmon. A single allele of large effect segregates in Atlantic Salmon that influences the sexual maturation rate in salmon (LetterSpace=10AYLLON *et al.*, 2015; LetterSpace=10BARSON *et al.*, 2015), and hence the timing of their return from the sea to spawn (sea age). The allele falls close to the autosomal gene VGLL3 (LetterSpace=10COUSMINER *et al.*, 2013, variation at this gene in humans also influences the timing of puberty). The left side of Figure 4.15 shows the age at sexual maturity in males. The allele (E) associated with slower sexual maturity is recessive in males. While the LL homozygotes mature on average a whole year later, the additive effect of the allele is weak while the L allele is rare in the population. The right panel shows the effect of the L allele in females. Note how the allele is much more dominant in females, and has a much more pronounced additive effect. The dominance of an allele is not a fixed property of the allele but rather a statement of the relationship of genotype to phenotype, such that the dominance relationship between alleles may vary across phenotypes and contexts (e.g. sexes).

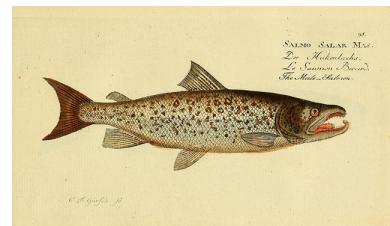


Figure 4.14: Atlantic Salmon (*Salmo salar*).

Histoire naturelle des poissons. 1796. Bloch, M. E. Image from the Biodiversity Heritage Library. Contributed by Ernst Mayr Library, Museum of Comparative Zoology. Not in copyright.

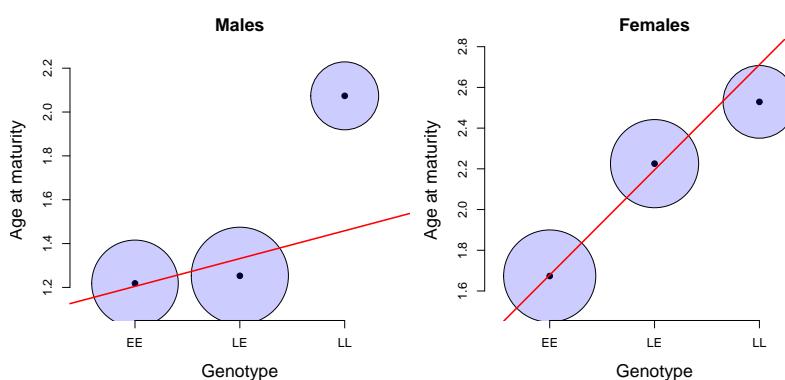


Figure 4.15: The average age at sexual maturity for each genotype, broken down by sex. The area of each circle is proportional to the fraction of the population in each genotypic class. The regression between phenotype and additive genotype is shown as a red line. Data from LetterSpace=10BARSON *et al.* (2015). Code here.

The variance in the population phenotype due to these additive

breeding values at locus ℓ , assuming HW proportions, is

$$\begin{aligned} V_{A,\ell} &= p^2(2a_{\ell,2})^2 + 2pq(a_{\ell,1} + a_{\ell,2})^2 + q^2(2a_{\ell,0})^2 \\ &= 2(pa_{\ell,1}^2 + qa_{\ell,2}^2) \\ &= 2pq\alpha_\ell^2 \end{aligned} \quad (4.24)$$

The total additive variance for the whole genotype can be found by
2892 summing the individual additive genetic variances over loci

$$V_A = \sum_{\ell=1}^L V_{A,\ell} = \sum_{\ell=1}^L 2p_\ell q_\ell \alpha_\ell^2. \quad (4.25)$$

Having assigned the additive genetic variance to be the variance
2894 explained by the additive contribution of the alleles at a locus, we
define the dominance variance as the population variance among
2896 genotypes at a locus due to their deviation from additivity. We can
calculate how much each genotypic mean deviates away from its
2898 additive prediction at locus ℓ (the length of the arrows in Figure
4.13). For example, the heterozygote deviates

$$d_{\ell,1} = \bar{X}'_{\ell,1} - (a_{\ell,1} + a_{\ell,2}) \quad (4.26)$$

2900 away from its additive genetic value, with similar expressions for
each of the homozygotes ($d_{\ell,0}$ and $d_{\ell,2}$). We can then write the domi-
2902 nance variance at our locus as the genotype-frequency weighted sum
of our squared dominance deviations

$$V_{D,\ell} = p^2 d_{\ell,0}^2 + 2pq d_{\ell,1}^2 + q^2 d_{\ell,2}^2. \quad (4.27)$$

2904 Writing our total dominance variance as the sum across loci

$$V_D = \sum_{\ell=1}^L V_{D,\ell}. \quad (4.28)$$

Having now partitioned all of the genetic variance into additive and
2906 dominant terms, we can write our total genetic variance as

$$V_G = V_A + V_D. \quad (4.29)$$

We can do this because by construction the covariance between our
2908 additive and dominant deviations for the genotypes is zero. We
can define the narrow sense heritability as before $h^2 = V_A/V_P =$
2910 $V_A/(V_G + V_E)$, which is the proportion of phenotypic variance due to
additive genetic variance. We can also define the total proportion of
2912 the phenotypic variance due to genetic differences among individu-
als, as the broad-sense heritability $H^2 = V_G/(V_G + V_E)$.

2914 When dominance is present in the loci influencing our trait ($V_D >$
0), we need to modify our phenotype covariance among relatives

Relationship (i,j)*	$Cov(X_i, X_j)$
parent-child	$1/2V_A$
full siblings	$1/2V_A + 1/4V_D$
identical (monozygotic) twins	$V_A + V_D$
1 st cousins	$1/8V_A$

Table 4.1: Phenotypic covariance between some pairs of relatives, include the dominance variation. *Assuming this is the only relationship the pair of individuals share (above that expected from randomly sampling individuals from the population).

2916 to account for this non-additivity. Specifically, our equation for the
 2917 covariance among a general pair of relatives (eqn. 4.12 for additive
 2918 variation) becomes

$$Cov(X_1, X_2) = 2F_{1,2}V_A + r_2V_D \quad (4.30)$$

2919 where r_2 is the probability that the pair of individuals share 2 alleles
 2920 identical by descent, making the same assumptions (other than additivity)
 2921 that we made in deriving eqn. 4.12. In table 4.1 we show
 2922 the phenotypic covariance for some common pairs of relatives. The
 2923 regression of offspring phenotype on parental midpoint still has a
 2924 slope V_A/V_P .

2925 Full sibs and parent-offspring have the same covariance if there
 2926 is no dominance variance (as they have the same kinship coefficient
 $F_{1,2}$). However, when dominance is present ($V_D > 0$), full-sibs re-
 2927 semble each other more than parent-offspring pairs. That's because
 2928 parents and offspring share precisely one allele, while full-sibs can
 2929 share both alleles (i.e. the full genotype at a locus) identical by de-
 2930 scent. We can attempt to estimate V_D by comparing different sets
 2931 of relationships. For example, non-identical twins (full sibs born at
 2932 same time) should have 1/2 the phenotypic covariance of identical
 2933 twins if $V_D = 0$. Therefore, we can attempt to estimate V_D by look-
 2934 ing at whether identical twins have more than twice the phenotypic
 2935 covariance than non-identical twins.

2936 The most important aspect of this discussion for thinking about
 2937 evolutionary genetics is that the parent-offspring covariance is still
 2938 only a function of V_A . This is because our parent (e.g. the mother)
 2939 transmits only a single allele, at each locus, to its offspring. The other
 2940 allele the offspring receives is random (assuming random mating), as
 2941 it comes from the other unrelated parent (the father). Therefore, the
 2942 average effect on the child's phenotype of an allele the child receives
 2943 from their mother is averaged over all possible random alleles the
 2944 child could receive from their father (weighted by their frequency
 2945 in the population). Thus we only care about the additive effect of
 2946 the allele, as parents transmit only alleles (not genotypes) to their
 2947 offspring. This means that the short-term response to selection, as
 2948 described by the breeder's equation, depends only on V_A and the ad-
 2949 ditive effect of alleles. Therefore, if we can estimate the narrow-sense

heritability we can predict the short-term response. However, if alleles display dominance, our value of V_A will change as alleles at our loci change in frequency, e.g. as dominant alleles become common in the population their contribution to V_A decreases. Therefore, if there is dominance our value of V_A will not be constant across generations.

Up to this point we have only considered dominance and not epistasis. However, we can include epistasis in a similar manner (for example among pairs of loci). This gets a little tricky to think about, so we will only briefly explain it. We can first estimate the additive effect of the alleles by considering the effect of the alleles averaging over their possible genetic backgrounds (including the other interacting alleles they are possibly paired with), just as before. We can then calculate the additive genetic variance from this. We can estimate the dominance variance, by calculating the residual variance among genotypes at a locus unexplained by the additive effect of the loci. We can then estimate the epistatic variance by estimating the residual variance left unexplained among the two locus genotypes after accounting for the additive and dominant deviations calculated from each locus separately. In practice these high variance components are hard to estimate, and usually small as much of our variance is assigned to the additive effect. Again we would find that we mostly care about V_A for predicting short-term evolution, but that the contribution of loci to the additive genetic variance will depend on the epistatic relationships among loci.

Question 5. How could you use 1/2 sibs vs. full-sibs to estimate V_D ? Why might this be difficult in practice? Why are identical vs. non-identical twins better suited for this?

Question 6. Can you construct a case where $V_A = 0$ and $V_D > 0$? You need just describe it qualitatively; you don't need to work out the math. (tricker question).

5

The Response to Phenotypic Selection

Evolution by natural selection requires:

1. Variation in a phenotype
2. That survival is non-random with respect to this phenotypic variation.
3. That this variation is heritable.

Points 1 and 2 encapsulate our idea of Natural Selection, but evolution by natural selection will only occur if the 3rd condition is also met.¹ It is the heritable nature of variation that couples change within a generation due to natural selection to change across generations (evolutionary change).

Let's start by thinking about the change within a generation due to directional selection, where selection acts to change the mean phenotype within a generation. For example, a decrease in mean height within a generation, due to taller organisms having a lower chance of surviving to reproduction than shorter organisms. Specifically, we'll denote our mean phenotype at reproduction by μ_S , i.e. after selection has acted, and our mean phenotype before selection acts by μ_{BS} . This second quantity may be hard to measure, as obviously selection acts throughout the life-cycle, so it might be easier to think of this as the mean phenotype if selection hadn't acted. So the change in mean phenotype within a generation is $\mu_S - \mu_{BS} = S$.

We are interested in predicting the distribution of phenotypes in the next generation. In particular, we are interested in the mean phenotype in the next generation to understand how directional selection has contributed to evolutionary change. We'll denote the mean phenotype in offspring, i.e. the mean phenotype in the next generation before selection acts, as μ_{NG} . The change across generations we'll call the response to selection R and put this equal to $\mu_{NG} - \mu_{BS}$.

The mean phenotype in the next generation is

$$\mu_{NG} = \mathbb{E}(\mathbb{E}(X_{kid}|X_{mum}, X_{dad})) \quad (5.1)$$

See LetterSpace=10LEWONTIN (1970). Note that these requirements are not specific to DNA, i.e. the concept of evolution by natural selection is substrate independent.

¹ Some people consider natural selection to only operate on heritable phenotype variation and so require all three conditions to say that natural selection occurs. This is mostly a semantic point, however, it is useful to be able to distinguish the action of selection from a possible response.

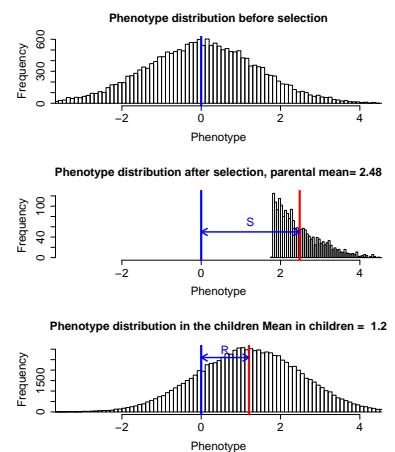


Figure 5.1: **Top.** Distribution of a phenotype in the parental population prior to selection, $V_A = V_E = 1$. **Middle.** Only individuals in the top 10% of the phenotypic distribution are selected to reproduce; the resulting shift in the phenotypic mean is S . **Bottom.** Phenotypic distribution of children of the selected parents; the shift in the mean phenotype is R . Code here.

3012 where the outer expectation is over possible pairs of randomly mating
 3013 individuals who survive to reproduce. We can use eqn. 4.14 to
 obtain an expression for this expectation:

$$\mu_{NG} = \mu_{BS} + \beta_{mid,kid}(\mathbb{E}(X_{mid}) - \mu_{BS}) \quad (5.2)$$

So to obtain μ_{NG} we need to compute $\mathbb{E}(X_{mid})$, the expected mid-point phenotype of pairs of individuals who survive to reproduce.

3016 Well this is just the expected phenotype in the individuals who survived to reproduce (μ_S), so

$$\mu_{NG} = \mu_{BS} + h^2(\mu_S - \mu_{BS}) \quad (5.3)$$

So we can write our response to selection as

$$R = \mu_{NG} - \mu_{BS} = h^2(\mu_S - \mu_{BS}) = h^2S \quad (5.4)$$

3020 So our response to selection is proportional to our selection differential, and the constant of proportionality is the narrow sense heritability. This equation is sometimes termed the Breeder's equation.
 3022 It is a statement that the evolutionary change across generations (R) is proportional to the change caused by directional selection within a generation (S), and that the strength of this relationship is determined by the narrow sense heritability (h^2).
 3024

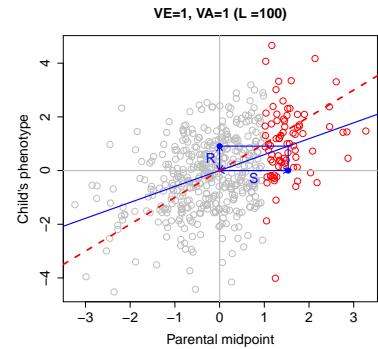


Figure 5.2: A visual representation of the Breeder's equation. Regression of child's phenotype on parental mid-point phenotype ($V_A = V_E = 1$). Under truncation selection, only individuals with phenotypes > 1 (red) are bred. Code here.

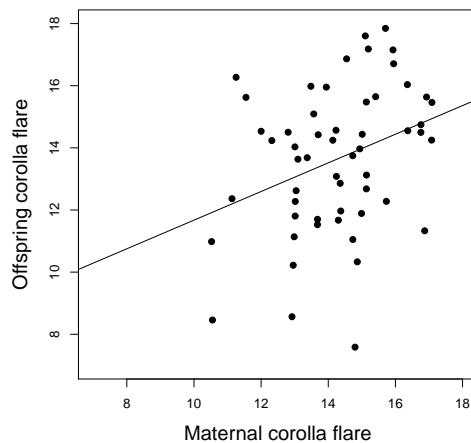


Figure 5.3: The relationship between maternal and offspring corolla flare (flower width) in *P. viscosum*. From LetterSpace=10GALEN's data the covariance of mother and child is 1.3, while the variance of the mother is 2.8. Data from LetterSpace=10GALEN (1996). Code here.



Figure 5.4: Sticky jacob's ladder (*Polemonium viscosum*). Flowers of Mountain and Plain (1920). Clements, E. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden, Mertz Library. Not in copyright. Cropped from original.

Question 1. LetterSpace=10GALEN (1996) explored selection on flower shape in *P. viscosum*. She found that plants with larger corolla flare had more bumblebee visits, which resulted in higher seed set and a 17% increase in corolla flare in the plants contributing to the next generation. Based on the data in the caption of Figure 5.3 what is the expected response in the next generation?

To understand the genetic basis of the response to selection take
3034 a look at Figure 5.5. The setup is the same as in our previous sim-
ulation figures. The individuals who are selected to form our

figures/QT3_w_genosums.pdf

Figure 5.5: **Top.** Distribution of the number of up alleles in the parental population prior to selection (red), for the selected individuals the top 10% of the population (blue) **Bottom.** The same distribution for the offspring of the selected parents in the next generation (green). Code here.

3036 next generation carry more alleles that increase the phenotype, in the
current range of environments currently experienced by the popula-
3038 tion. The average individual before selection carried 100 of these 'up'
alleles, the average individual surviving selection 108 'up' alleles.
3040 As individuals faithfully transmit their alleles to the next generation
the average child of the selected parents carries 108 up alleles. Note
3042 that the variance has changed little, the children have plenty of varia-
tion in their genotype, such that selection can readily drive evolution
3044 in future generations. The average frequency of an 'up' allele has
changed from 50% to 54%. Our gains due to selection will be stably
3046 inherited to future generations.

The long-term response to selection If our selection pressure is sus-
3048 tained over many generations, we can use our breeder's equation to
predict the response. If we are willing to assume that our heritability

3050 does not change and we maintain a constant selection gradient, then
3051 after n generations our phenotype mean will have shifted

$$nh^2S \quad (5.5)$$

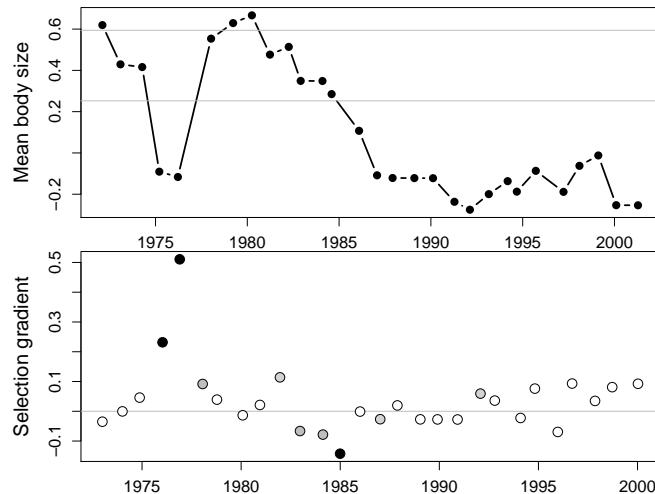
3052 i.e. our population will keep up a linear response to selection.

Question 2. A population of red deer were trapped on Jersey (an
3054 island off of England) during the last inter-glacial period. From the
fossil record ² we can see that the population rapidly adapted to their
3056 new conditions. Within 6,000 years they evolved from an estimated
mean weight of the population of 200kg to an estimated mean weight
3058 of 36kg (a 6 fold reduction)! You estimate that the generation time
of red deer is 5 years and, from a current day population, that the
3060 narrow sense heritability of the phenotype is 0.5.

A) Estimate the mean change per generation in the mean body
3062 weight.

B) Estimate the change in mean body weight caused by selection
3064 within a generation. State your assumptions.

C) Assuming we only have fossils from the founding popula-
3066 tion and the population after 6000 years, should we assume that the
calculations accurately reflect what actually occurred within our pop-
3068 ulation?



² LetterSpace=10LISTER, A., 1989. Rapid dwarfing of red deer on Jersey in the last interglacial. *Nature* 342(6249): 539

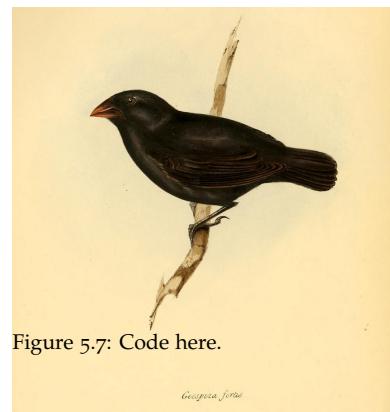


Figure 5.7: Code here.

Figure 5.6: Medium ground-finch (*Geospiza fortis*).
The zoology of the voyage of H.M.S. Beagle. Birds Part 3. (1841)
Gould G. Edited by Darwin, C. Illustration by Elizabeth Gould.
Image from the Biodiversity Heritage Library. Contributed by
Natural History Museum Library, London . Not in copyright.

3070 Alternative formulations of the Breeder's equation. A change in mean
phenotype within a generation occurs because of the differential fit-
3071 ness of our organisms. To think more carefully about this change
within a generation, let's think about a simple fitness model where

our phenotype affects the viability of our organisms (i.e. the probability they survive to reproduce). The probability that an individual has a phenotype X before selection is $p(X)$, so that the mean phenotype before selection is

$$\mu_{BS} = \mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx \quad (5.6)$$

The probability that an organism with a phenotype X survives to reproduce is $w(X)$, and we'll think about this as the fitness of our organism. The probability distribution of phenotypes in those who do survive to reproduce is

$$\mathbb{P}(X|\text{survive}) = \frac{p(x)w(x)}{\int_{-\infty}^{\infty} p(x)w(x)dx}. \quad (5.7)$$

where the denominator is a normalization constant which ensures that our phenotypic distribution integrates to one. The denominator also has the interpretation of being the mean fitness of the population, which we'll call \bar{w} , i.e.

$$\bar{w} = \int_{-\infty}^{\infty} p(x)w(x)dx. \quad (5.8)$$

Therefore, we can write the mean phenotype in those who survive to reproduce as

$$\mu_S = \frac{1}{\bar{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx \quad (5.9)$$

If we mean center our population, i.e. set the phenotype before selection to zero, then

$$S = \frac{1}{\bar{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx = \mathbb{E}(Xw(X)) \quad (5.10)$$

where the final part follows from the fact that the integral is taking the mean of $Xw(X)$ over the population.

As our phenotype is mean centered ($\mathbb{E}(X) = 0$), we can see that S has the form of a covariance between our phenotype X and our relative fitness $w(X)$

$$S = \mathbb{E}(Xw(X)) - \mathbb{E}(X)\mathbb{E}(w(X)) = \text{Cov}(X, w(X)/\bar{w}) \quad (5.11)$$

Thus our change in mean phenotype is directly a measure of the covariance of our phenotype and our fitness. Rewriting our breeder's equation using this observation we see

$$R = \frac{V_A}{V} \text{Cov}(X, w(X)/\bar{w}) \quad (5.12)$$

we see that the response to selection is due to the fact that our fitness (viability) of our organisms/parents covaries with our phenotype, and that our child's phenotype is correlated with our parent's phenotype.



illustration_images/Quant_gen/red_deer/Red_deer.pn

Figure 5.8: Red deer (*Cervus elaphus*). British mammals. Thorburn, A. (1920) Image from the Biodiversity Heritage Library. Contributed by Field Museum of Natural History Library. Licensed under CC BY-2.0.

Fisher's fundamental theorem of natural selection If we choose fitness to be our phenotype ($X = w(X)/\bar{w}$), then the response in fitness is

$$\begin{aligned} R &= \frac{V_A}{V} \text{Cov}(w(X)/\bar{w}, w(X)/\bar{w}) = \frac{V_A}{V} V \\ &= V_A \end{aligned} \quad (5.13)$$

i.e. the response to selection is equal to the additive genetic variance for fitness. Or as Fisher put it

"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time." -LetterSpace=10FISHER (1930) (pg 37)

Fisher called this 'the fundamental theorem of natural selection'. Our proof here is just a sketch, and more formal approaches are needed to show it in generality. There has been much nashing of teeth over exactly how broadly this result holds, and exactly what Fisher meant (see LetterSpace=10EWENS, 2010, for a recent overview).

Fitness Gradients and linear regressions To understand this in more detail let imagine that we calculate the linear regression of an individual i 's mean-centered phenotype (X_i) on fitness (W_i), i.e.

$$W_i \sim \beta X_i + \bar{w} \quad (5.14)$$

The best fitting slope of this regression (β), lets call it the 'fitness gradient', is given by

$$\beta = \text{Cov}(X, w(X)/\bar{w}) / V \quad (5.15)$$

i.e. the fitness gradient is the covariance of phenotype-fitness covariance divided by the phenotypic variance. Using this result we can rewrite the breeder's equation as

$$R = V_A \beta \quad (5.16)$$

i.e. we'll see a directional response to selection if there is a linear relationship of phenotype on fitness, and if there is additive genetic variance for the phenotype. As one example of a fitness gradient, in Figure 5.9 the lifetime reproductive success (LRS) of male Red Deer is plotted against the weight of their antlers. The red line gives the linear regression of fitness (LRS) on antler mass and the slope of this line is the fitness gradient (β).

Fitness landscapes When we talk about evolution we often talk of a population exploring an adaptive landscape with natural selection pushing a population towards higher fitness states corresponding to

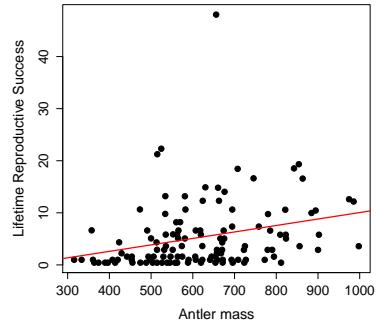


Figure 5.9: Lifetime reproductive success (LRS) of male Red Deer as a function of their antler mass. Data from LetterSpace=10KRUUK *et al.* (2002), see the paper for discussion of the complexities of equating this selection gradient with the evolutionary response. Code here..

figures/Response_to_sel/fitness_landscape_1D_w_wbar.pdf

Figure 5.10: A population evolving on a (guassian) fitness surface. The bottom panel shows the expected individual fitness ($w()$) and mean fitness as a function of phenotype. The red line shows the best fitting linear approximation to the relationship between phenotype and fitness, eqn (??), whose slope is β . The top panel shows the distribution of the phenotype before and after selection. Code here..

peaks in this landscape (see e.g. Figure ??). LetterSpace=10LANDE (1976) found an evocative formulation of the Breeder's equation which aids our intuition of phenotypic fitness landscapes. LetterSpace=10LANDE showed that, if the phenotype is normally distributed, the response to selection (R) could be written in terms of the gradient (derivative) of the mean fitness (\bar{w}) of the population as a function of the mean phenotype:

$$R = \frac{V_A}{\bar{w}} \frac{\partial \bar{w}}{\partial \bar{x}} \quad (5.18)$$

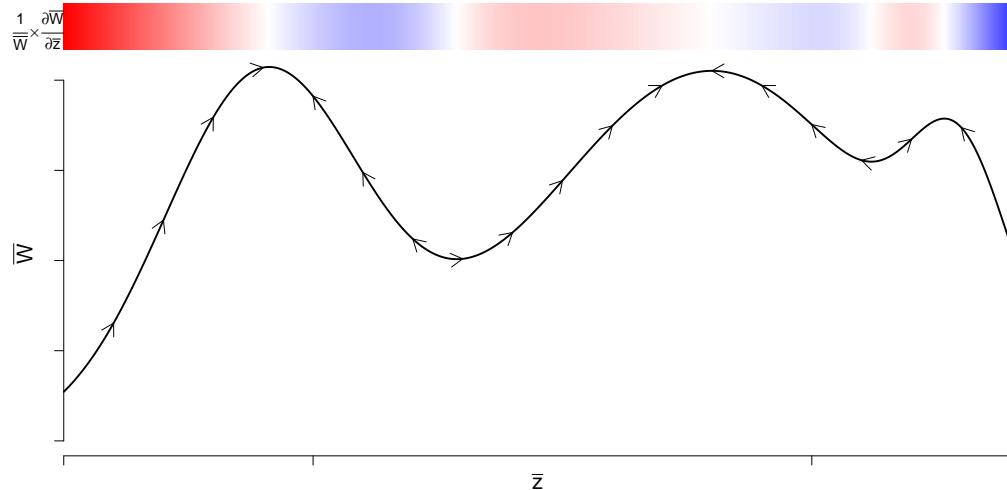
What does this mean? Well V_A/\bar{w} is always positive, so the direction our population responds to selection is predicted by the sign of the derivative. If increasing the mean phenotype of the population slightly would increase mean fitness ($\partial \bar{w} / \partial \bar{x} > 0$) our population will respond that generation by evolving toward higher values of the trait ($R > 0$), left panel of Figure 5.10. Conversely if decreasing the population mean phenotype slightly would increase the mean fitness ($\partial \bar{w} / \partial \bar{x} < 0$) the population will that generation evolve towards lower values of the phenotype, middle panel of Figure 5.10. Thus we can think of the population as evolving on an adaptive landscape where the elevation is given by the population mean fitness. Natural selection operates on the basis of individual-level fitness, but as a result of this our population is increasing in its average fitness, it is becoming more adapted. What happens when it reaches the top?

This follows from the fact that we can then move the derivative inside the integral of \bar{w}

$$\begin{aligned} \frac{1}{\bar{w}} \frac{\partial \bar{w}}{\partial \bar{x}} &= \frac{1}{\bar{w}} \int_{-\infty}^{\infty} w(x) \frac{\partial p(x)}{\partial \bar{x}} dx \\ &= \int_{-\infty}^{\infty} \frac{w(x)}{\bar{w}} \frac{(x - \bar{x})}{V} dx \\ &= \frac{\text{cov}(w(x), x)}{\text{var}(x)} \end{aligned} \quad (5.17)$$

which is β . The middle line holds when $p(x)$ is the normal distribution.

CAVEATS



Well at the top of a peak $\partial \bar{w} / \partial \bar{x} = 0$, as it is a local maximum, and so $R = 0$. Assuming that the relationship between fitness and phenotype stays constant, our population will stay at the top of the fitness peak. This view of natural selection does not imply that the population is evolving to the best possible state. Our population is just marching

up the hill of mean fitness end panel Figure 5.10. However, this peak
 3156 isn't necessarily the highest fitness peak it's just which ever peak was
 closest and so our population can become trapped on a local, but not
 3158 global peak of fitness (see, for example Figure ??).

One nice example documenting adaptive evolution to a new fit-
 3160 ness optimum is offered by a remarkable time-series of stickleback
 evolution from a fossil lake-bed in Nevada (LetterSpace=10BELL *et al.*,
 3162 2006). In this lake the layers of sediment are laid down each year
 allowing a very detailed time series with over five thousand fossils
 3164 measured. The time-series documents the evolution towards a new
 set of optimum phenotypes in the fifteen thousand years after the ini-
 3166 tial invasion of the lake by a heavily armoured stickleback species. In
 Figure ?? the population mean number of touching pterygiophores,
 3168 the bones supporting the dorsal spines, through the fossil record.
 Note how quickly the species evolves toward its new value, presum-
 3170 ably a fitness optimum in their new environment, and the long time
 subsequent time interval over which the population mean phenotype
 3172 fluctuates about its new value.

LetterSpace=10HUNT *et al.* (2008) fitted a model of a population
 3174 adapting to a fitness landscape, with a single peak, to these time-
 series data. Their fitted fitness surface is shown in the lower panel of
 3176 Figure 5.12 . The arrows show the moves that the population mean
 phenotype is making on this inferred fitness surface. The population
 3178 initially takes large steps up toward the peak of this surface and
 subsequently fluctuates around the peak. Under the interpretation
 that there is a single stationary peak these fluctuations represent
 3180 genetic drift randomly knocking the population off its optimum,
 with selection acting to restore the population towards this local
 3182 optimum.

3184 peaks in the fitness landscape

Stabilizing and Disruptive selection Up to now we have just looked
 3186 at directional selection, where selection acts to change the mean
 phenotype. However, we can also use quantitative genetic models
 3188 to describe other modes of selection, extending from effects on the
 population mean the next natural step is to think about selection
 3190 which acts on the population variance. Selection might act against
 more strongly against individuals in the tails of the distribution, with
 3192 those closer to the mean phenotype having higher fitness, which
 lowers the variance. Selection could also disfavour individuals close
 3194 to the population mean, with individuals with extreme phenotypes
 having higher fitness, which acts to increase the fitness.

3196 Directional selection occurs because of the covariance between our
 phenotype and fitness, eqn (5.11). Just as we expressing directional

Figure 5.11: Fossil stickleback. Photo by Peter J. Park from LetterSpace=10Losos *et al.* (2013), licensed under CC BY 4.0.

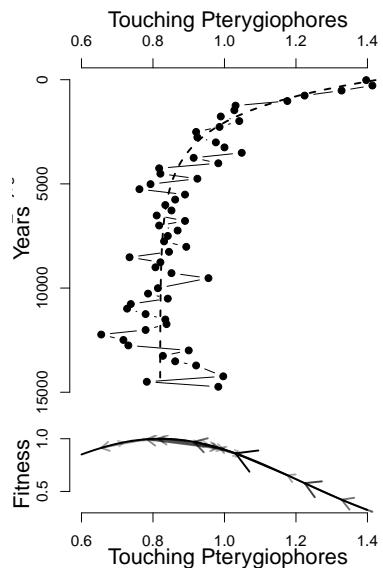


Figure 5.12: Top) A time series of stickleback phenotypic evolution from the fossil record. After a heavily armoured stickleback invades the lake it quickly evolves towards touching pterygiophores (the bones supporting the dorsal spines). Fossil measurements means are calculated in 250 year bins. Bottom) How our population moves on the inferred fitness landscape. The arrows show each move made by the population in the 250 intervals. Data from LetterSpace=10BELL *et al.* (2006) and LetterSpace=10HUNT *et al.* (2008) Code here.

selection as a covariance allowed us to characterize directional selection as the linear relationship between fitness and phenotype, β , we can summarize the variance reducing selection by including a quadratic term in the regression of fitness on phenotype

$$w_i \sim \beta x_i + 1/2\gamma x_i^2 + \bar{w} \quad (5.19)$$

This γ , the coefficient of the quadratic term in our model, is the quadratic selection gradient: the covariance of fitness and the squared deviation from the phenotypic mean (μ_{BS}), i.e.

$$\gamma = \frac{\text{Cov}(w(X), (X - \mu_{BS})^2)}{V^2} \quad (5.20)$$

Our γ describes the curvature of the fitness surface around the mean.

Values of $\gamma < 0$ are consistent with stabilizing selection, reducing the variance. While values of $\gamma > 0$ are consistent with disruptive selection, increasing the variance.

Under stabilizing selection the individuals with extreme phenotypes in either tail have lower fitness, the result of which is to reduce the phenotypic variance within a generation. A classic case of stabilizing selection is birth weight in humans (LetterSpace=10KARN and LetterSpace=10PENROSE, 1951). Mary Karn collected data for nearly fourteen thousand pregnancies from 1935-46 for birth weight and mortality. These data are replotted in Figure 5.13. The variance of all births is 1.575lb^2 , while in live births this was reduced to 1.26lb^2 , a 20% reduction in variance due to stabilizing selection. It is worth noting, that this selection pressure has been greatly reduced over the decades in societies with access to good prenatal care (LetterSpace=10ULIZZI and LetterSpace=10TERRENATO, 1992).

In Central Africa, Black-bellied seedcrackers (*P. ostrinus*) show remarkable size polymorphism in their beaks (Figure 5.15). The small-beaked individuals feed on soft seeds from one species of marsh sedge while the big-beaked individuals feed on hard seeds from another sedge, which requires ten times the force to crack. LetterSpace=10SMITH (1993) recorded the fates of hundreds of juveniles, and found that individuals with intermediate beak sizes survived at much lower rates, Figure 5.15, because they were not well adapted to either seed resource. Break length is subject to disruptive selection, as can also be seen by the significant negative quadratic term in the regression of survival probability on break length. The variance of mandible in the total sample of individuals was 0.5mm^2 in the survivors this variance increased by a factor of almost $\times 2.5$ to 1.3mm^2 .

To illustrate how directional selection and quadratic terms play off during adaptation, lets consider the goldenrod gall fly (*Eurosta solidaginis*), aka the goldenrod ball gallmaker. See Figure 5.17. As it's

Just like how β could be interpreted as the mean gradient of the fitness surface, our γ is the mean curvature of the fitness surface

$$\gamma = \mathbb{E} [\partial^2 w(x)/\partial x^2] = \int \partial^2 w(x)/\partial x^2 p(x) dx \quad (5.21)$$

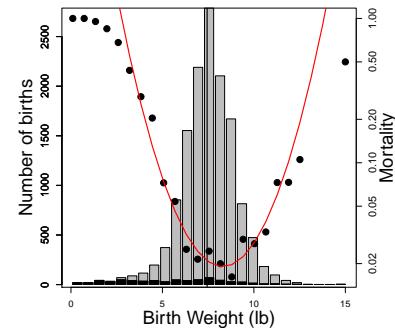


Figure 5.13: Bars show the total number of births with different birth weights (left axis). Dots show the mortality probability for different birth-weight bins (right axis). Data from LetterSpace=10KARN and LetterSpace=10PENROSE (1951) Table 2, collapsing male and female births, Code here.

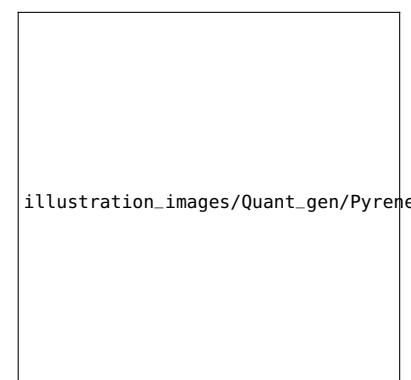


Figure 5.14: Lesser seedcracker *Pyrenestes minor* a close relative of the Black-bellied seedcracker, whose beak is about the same size as the smallest Black-bellied individuals.

The birds of Africa, comprising all the species which occur in the Ethiopian region. (c)86 Salter, W. L. Prints by H. Grönwald

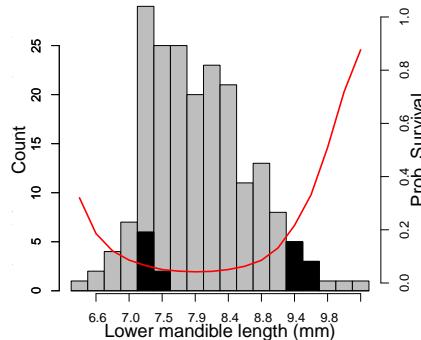
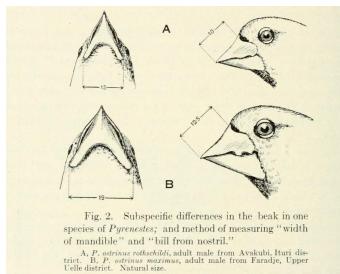


Figure 5.15: Left An illustration of the remarkable variation in beak size within Black-bellied seedcrackers (*P. ostrinus*). Right A histogram of a beak size measurement in Black-bellied seedcrackers, all juveniles are shown in white the black bars show the survivors. The red curve shows the best fitting linear and quadratic model to the probability of survival, fitted using a binomial generalized linear models with a logit link function.

Left illustration from: Size variation in *Pyrenestes* by Chapin J.P. in the Bulletin of the American Museum of Natural History (Vol. XLIX 1923) Image from the Biodiversity Heritage Library. Contributed by Toronto Library. Not in copyright.

wonderful name implies this insect lays its eggs in Goldenrod plants, and the larvae release chemicals forcing the plant to form a gall that forms a home for the larvae as they develop. While this seems like a pretty sweet deal for the larvae, it is not without its perils.

When the small, ball galls fall risk of parasitism from parasitoid wasps. This selection drives strong positive directional selection on gall size, with little stabilizing selection, notice that the good agreement between the linear selection gradient and the fit including a linear and quadratic term. However, bigger galls fall under the pall of predation from downy woodpeckers and black-capped chickadees, who seek out the tasty larvae. Thus intermediate size galls are favoured, a fitness peak that the population quickly reaches this fitness peak. Once on this peak, there is no directional selection, i.e. no linear slope, but there is strong stabilizing selection, i.e. a quadratic term.

3252 5.0.1 The response of multiple traits to selection, the multivariate breeder's equation.

3254 We can generalize these results for multiple traits, to ask how selection on multiple phenotypes plays out over short time intervals.³
3256 Considering two traits we can write our responses in both traits as

$$\begin{aligned} R_1 &= V_{A,1}\beta_1 + V_{A,1,2}\beta_2 \\ R_2 &= V_{A,2}\beta_2 + V_{A,1,2}\beta_1 \end{aligned} \quad (5.22)$$

where the 1 and 2 index our two different traits. Here $V_{A,1,2}$ is our additive covariance between our traits. Our selection gradient for trait 1, β_1 , represents the change in fitness changing trait 1 alone holding everything else constant. This is a statement that our response in any one phenotype is modified by selection on other traits that covary with that trait. This offers a good way to think about how

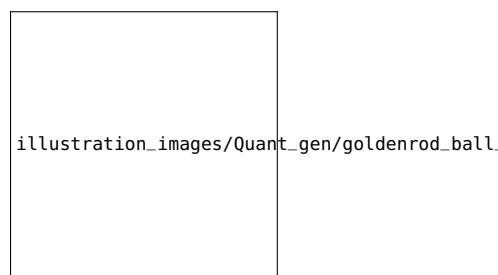


Figure 5.16: The gall formed by the goldenrod ball gallmaker (*Eurosta solidaginis*) in a goldenrod plant. The one on the right is cut to show a partial cross-section.

Annual report of the New York State Museum (1917) Image from the Biodiversity Heritage Library. Contributed by The LuEsther T Mertz Library, the New York Botanical Garden. Not in copyright.

³ LetterSpace=10LANDE, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution 33(1Part2): 402–416



Journal_figs/Quant_gen/Weis_Gorman_gall_size_stabilizing_sel/gall_size.pdf

Figure 5.17: Fitness surface for gall diameter in goldenrod ball gallmakers. The dots are the measured survival probabilities of different sized galls. The solid line is a fitted individual fitness surface ($w()$). Dotted line is \bar{w} plotted as a function of the population mean assuming a normal distribution with a standard deviation of 2mm. Data from LetterSpace=10WEIS and LetterSpace=10GORMAN (1990), Code here.

genetic trade offs play out over short-term evolution.

We can also write this in matrix form. We can write our change in the mean of our multiple phenotypes within a generation as the vector \mathbf{S} and our response across multiple generations as the vector \mathbf{R} . These two quantities are related by

$$\mathbf{R} = \mathbf{G}\mathbf{V}^{-1}\mathbf{S} = \mathbf{G}\beta \quad (5.23)$$

where \mathbf{V} and \mathbf{G} are our matrices of the variance-covariance of phenotypes and additive genetic values (eqn. (4.19) (4.18)) and β is a vector of selection gradients (i.e. the change within a generation as a fraction of the total phenotypic variance).

Question 3. You collect observations of red deer within a generation, recording an individual's number of offspring and phenotypes for a number of traits which are known to have additive genetic variation. Using your data, you construct the plots shown in Figure 5.18 (standardizing the phenotypes). Answer the following questions by choosing one of the bold options. Briefly justify each of your answers with reference to the breeder's equation and multi-trait breeder's equation.

A) Looking just at figure 5.18 A, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

B) Looking just at figures 5.18 B and C, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

C) Looking at figures 5.18 A, B, and C, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

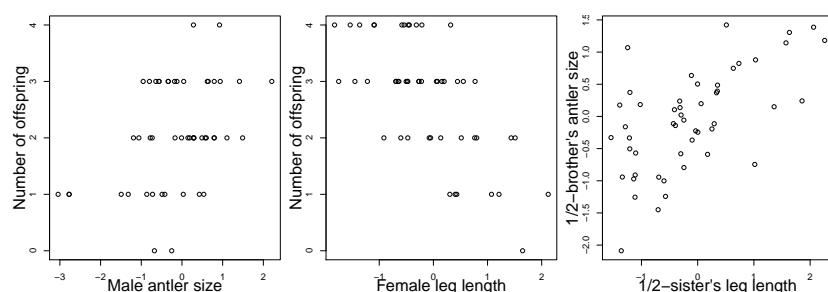
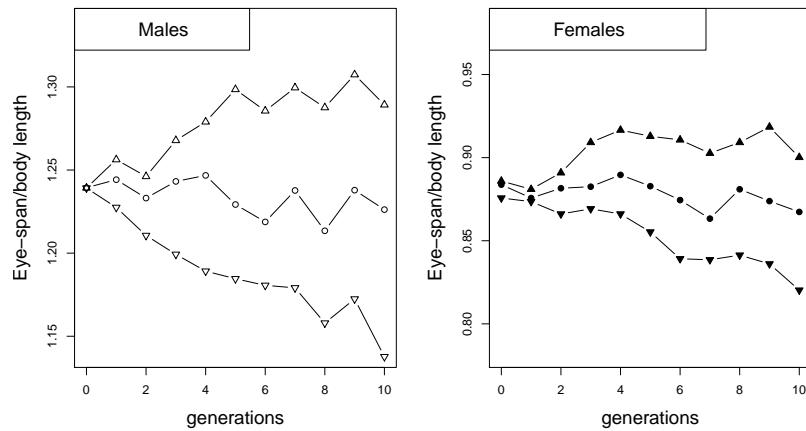


Figure 5.18: Observations of red deer within a generation; recording an individual's number of offspring and phenotypes (simulated data), which are known to have additive genetic variation. The figures left to right are A-C. (Data are simulated. Code here.)

As an example of correlated responses to selection, consider the

LetterSpace=10WILKINSON (1993) selection experiment on Stalk-eyed flies (*Cyrtodiopsis dalmani*). Stalk-eyed flies have evolved amazingly long eye-stalks. In the lab, LetterSpace=10WILKINSON established six populations of wild-caught flies and selected up and down on males

3294 eye-stalk to body size ratio for 10 generations (left plot in Figure
 5.19). Despite the fact that he did not select on females, he saw a
 3296 correlated response in the females from each of the lines (right plot),
 because of the genetic correlation between male and female body
 3298 proportions.



Question 4.

3300 At the end of ten generations in LetterSpace=10WILKINSON's ex-
 3302 periment (Figure 5.19), the males from the up- and down-selected
 3304 lines had mean eye-stalk to body ratios of 1.29 and 1.14 respectively,
 while the females from the up- and down-selected lines had means of
 0.9 and 0.82.

A) LetterSpace=10WILKINSON estimated that by selecting the
 3306 top/bottom 10 males, he had on average shifted the mean body
 3308 ratio by 0.024 within each generation. What is the male heritability
 of eye-stalk to body-length ratio?

B) Assume that the additive genetic variance of male and female
 3310 phenotypes are equal and that there is no direct selection on female
 3312 body-proportion in this experiment, i.e. that all of the response in
 females is due to correlated selection. Can you estimate the male-
 female genetic correlation of the eye-stalk ratio?

3314 *Estimating multivariate selection gradients* We can estimate multivari-
 ate directional (β) and quadratic selection gradients (γ) just as we did
 3316 for a single traits (x_1 and x_2), using linear models. For example, for
 two traits we can write

$$w_i \sim \beta_1 x_{1,i} + 1/2\gamma_1 x_{1,i}^2 + \beta_2 x_{2,i} + 1/2\gamma_2 x_{2,i}^2 + \gamma_{1,2} x_{1,i} x_{2,i} + \bar{w} \quad (5.24)$$

Figure 5.19: LetterSpace=10WILKINSON selected two of populations for flies for increased and eye-stalk to body length ratio in males (mean shown as up triangles), and two for a decreased ratio (down triangles), by taking the top 10 males with the highest (lowest) ratio out of 50 measures. He also established two control populations (circles). He constructed each generation of females by sampling 10 at random from each population. Data from LetterSpace=10WILKINSON (1993). Code here.

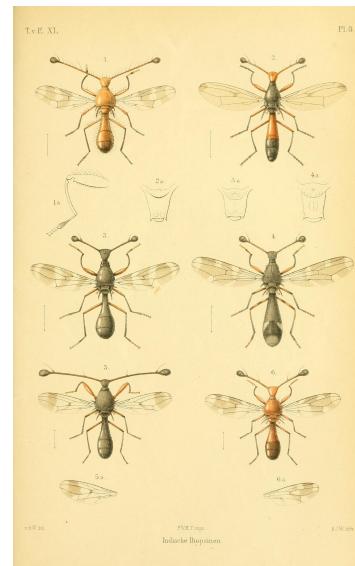


Figure 5.20: Stalk-eyed Flies (*Diopsidae*).
 Diptera. van der Wulp. 1898. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

3318 where β_1 and γ_1 are the directional and quadratic selection gradients
for trait one, and similarly for trait two. The covariance selection
3320 gradient between between traits is given by $\gamma_{1,2}$.

3322 LetterSpace=10BRODIE III (1992)'s work provides a nice example
of selection on multiple predation-avoidance traits in northwestern
garter snakes (*Thamnophis ordinoides*). LetterSpace=10BRODIE III re-
3324 leased hundreds on snakes born in the lab into the wild, and then
performed mark-recapture observations to monitor their fate.

3326 Before releasing them he measured how stripey they were, and
their behavioural tendency to reversals of direction during simulated
3328 flight from a predator flight. His quadratic fitness surface is shown
in Figure 5.22, based on fitting the regression given by eqn (5.24) to
3330 juvenile survival. He found that neither single trait directional or
quadratic gradients were significant, ie there was no apparent selec-
3332 tion on one trait ignoring the other. However, there was a significant,
negative covariance. The individuals with the highest chance of sur-
3334vival are either highly striped and perform few reversals (top left
3336 corner), or have little striping but reverse course frequently (bottom
right corner).

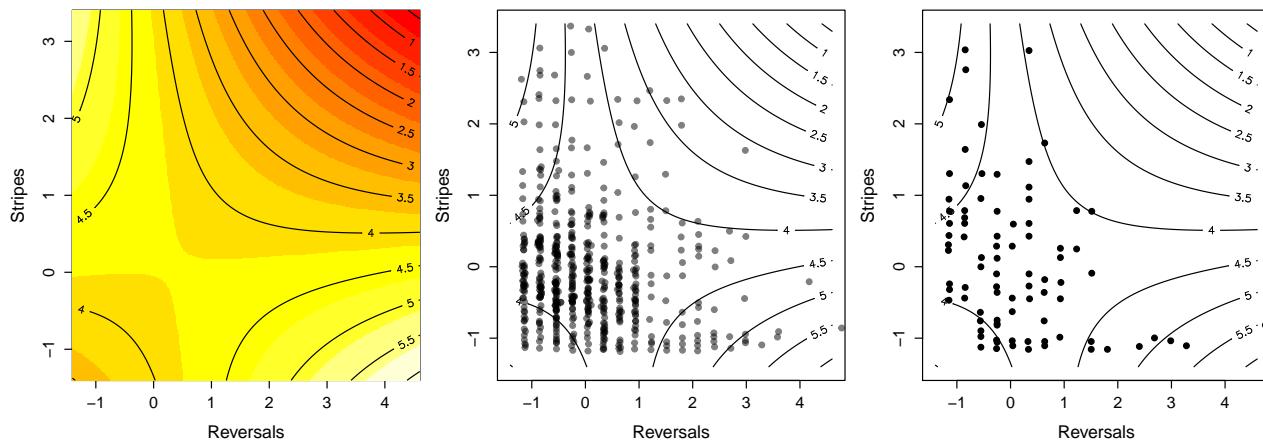


Figure 5.21: Northwestern garter snake (*Eutaenia cooperi*, now *Thamnophis ordinoides*)

The natural history of Washington territory, with much relating to Minnesota, Nebraska, Kansas, Oregon, and California (1859). Cooper J.G. and Stuckley G. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

5.1 Some applications of the multivariate trait breeder's equation

3338 The multivariate breeders equation has a lot of different uses in un-
derstanding the response of multiple traits to selection. It also offers
3340 some insights into kin selection and sexual selection. We'll discuss
these next.

3342 *Hamilton's Rule and the evolution of altruistic and selfish behaviours* In-
dividuals frequently behave in ways that sacrifice their own fitness

LetterSpace=10MAYNARD SMITH (1964) coined the name kin selection to de-
scribe Hamilton's approach to this problem. It's also sometimes called the
inclusive fitness approach, as we need to include not just one individual's
fitness but the weighted sum of all the fitness of all their relatives.

for the benefit of others. That selection favours such apparent acts of altruism is puzzling at first sight. LetterSpace=10HAMILTON (1964a,b) supplied the first general evolutionary explanation of such altruism. His intuition was that while an individual is losing out of some reproductive output, the alleles underlying an altruistic behaviour can still spread in the population if this cost is outweighed by benefits gained through the transmission of these alleles through a related individual. Note that this means that the allele is not acting in an self-sacrificing manner, even though individuals may as a result.

Altruism reflects social interactions. So as a simple model let's imagine that individuals interact in pairs, with our focal individual i being paired with an individual j . This could be pairs of siblings interacting. Imagine that individuals have two possible phenotypes $X = 1$ or 0 , corresponding to providing or withholding some small act of 'altruism' (we could just as easily flip these labels and call them an unselfish act and a selfish act respectively). Our pairs of individuals interacting could, for example, be siblings sharing a nest. The altruistic trait could be as simple as growing at a slightly slower rate so as to reducing sibling-competition for food from parents, or more complicated acts of altruism such as children foregoing their own reproduction so as to help their parents raise their siblings.

Providing the altruistic act has a cost C to the fitness of our individual and failing to provide this act has no cost. Receiving this altruistic act confers a fitness benefit B over individuals who did not receive this act. LetterSpace=10HAMILTON's rule states that such a trait will spread through the population if

$$2FB > C \quad (5.25)$$

where F is the average kinship coefficient between the interacting individuals (i and j). In the usual formulation of Hamilton's Rule our $2F$ is replaced by the 'Coefficient of relationship', which is the proportion of alleles shared between the individuals. Here we use two times the kinship coefficient to keep things inline with our notation for these chapters. Note that if our individuals are themselves inbred we need to do a little more careful to reconcile these two measures. So the altruistic behaviour will spread even if it is costly to the individual if its cost is paid off by the benefit to sufficiently related individuals.

As one example of kin-selection consider LetterSpace=10KRAKAUER (2005)'s work on co-operative courtship in wild turkeys (*Meleagris gallopavo*). Male turkeys often form display partnerships, with a subordinate male helping a dominant male with displaying to females and defending the females from other groups of males.

These pairs are often full brothers ($F = 0.25$), with the subordinate

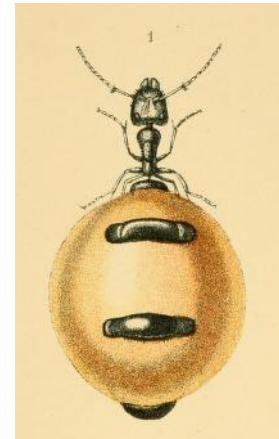


Figure 5.23: Australian Honey-pot Ant (*Camponotus inflatus*). Honey ants are gorged with honeydew collected by their nest mates, till they swell to the size of grapes, and used as a food storage device.

Ants, bees, and wasps; a record of observations on the habits of the social Hymenoptera (1897) Lubbock, J. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.



Figure 5.24: Turkey (*Meleagris gallopavo*). Bilder-atlas zur Wissenschaftlich-populären Naturgeschichte der Vögel in ihren sAdmmtlichen Hauptformen (1864). Wien, K. Hof Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

male often being the younger of the two. The subordinate male often loses out on mating opportunities over their entire lifetime by acting as a wingman to their older brothers. LetterSpace=10KRAKAUER (2005) estimated that dominant males gained an extra 6.1 offspring when they display with a partner than males who display alone. While the subordinate males lose out on fathering 0.9 offspring compared to solitary males. Thus the costs of helping by subordinate males is more than compensated by the fitness gains of their brothers ($(2 \times 0.25) \times 6.1 > 0.9$), and so the evolution of this altruistic helping in co-operative courtship is potentially well explained by kin-selection (see LetterSpace=10AKÇAY and LetterSpace=10VAN CLEVE, 2016, for more analysis).

Question 5. How would this answer be changed if the male Turkey partnerships were only 1/2 sibs, or first cousins?

Where does this result come from? Well, we can use our quantitative genetics framework to gain some intuition by deriving a simple version of Hamilton's Rule by thinking about the phenotypes of an individual's kin as genetically correlated phenotypes. To sketch a proof of this result, let's assume that our focal i individual's fitness can be written as

$$W(i, j) = W_0 + W_i + W_j \quad (5.26)$$

where W_i is the contribution of the fitness of the individual i due to their own phenotype, and W_j is the contribution to our individual i 's fitness due to the interacting individual j 's behaviour (i.e. j 's phenotype). With the benefit B and cost C , our $W(i, j)$ are depicted in Figure 5.25.

Following our multivariate breeder's equation, we can write the expected change of our behavioural phenotype as

$$R = \beta_i V_A + \beta_j V_{A,i,j}, \quad (5.27)$$

Our altruistic phenotype is increasing in the population if $R > 0$, i.e. if

$$\beta_i V_A + \beta_j V_{A,i,j} > 0 \quad (5.28)$$

The slope β_i of the regression of our focal individual's behavioural phenotype on fitness is proportional to $-C$. The slope β_j of the regression of our interacting partner's phenotype on our focal individual's fitness is proportional to B (with the same constant of proportionality). Therefore, our altruistic phenotype is increasing in the population if

$$\begin{aligned} \beta_i V_A + \beta_j V_{A,i,j} &> 0 \\ B \frac{V_{A,i,j}}{V_A} &> C \end{aligned} \quad (5.29)$$

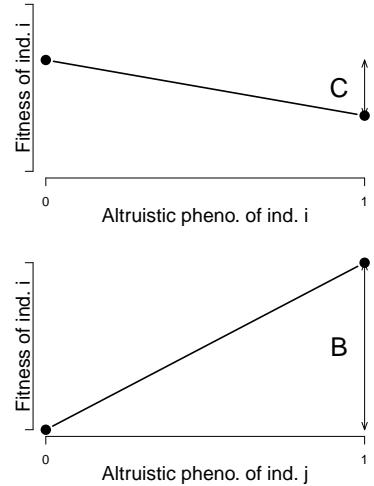


Figure 5.25: **Top)** The fitness of individual i as a function of their behavioural phenotype, where altruistic/non-altruistic behavioural phenotypes are encoded as 1 and 0 respectively. The direct fitness cost of behaving altruistically is C . **Bottom)** The fitness of our focal individual i as a function of the behavioural phenotype of their interacting partner (j). Our focal individual gets an increase B in fitness if their partner behaves altruistically. Code here.

Here we're following a simplified version of LetterSpace=10QUELLER (1992)'s treatment, to re-derive Hamilton's rule in a quantitative genetics framework (Hamilton's original papers did this in a population genetics framework).

So what's the average genetic covariance between individual i and j 's altruistic phenotype? Well it's the same behavioural phenotype in both individuals, so the phenotypes are genetically correlated if our individuals are related to each other. The covariance of the same phenotype between two individuals is just $2F_{i,j}V_A$ (see (4.12)). So our altruistic phenotype is increasing in the population if

$$B \frac{2F_{i,j}V_A}{V_A} > C \\ 2F_{i,j}B > C \quad (5.30)$$

Seen from this perspective, LetterSpace=10HAMILTON's rule is simply a statement that altruistic behaviours can spread via kin-selection, if the average cost to an individual of carrying altruistic alleles is paid back through the average benefit of interacting with altruistic relatives (kin)

3432 Sexual selection and the evolution of mate preference by indirect benefits.
 Organisms often put an enormous effort into finding and attracting mates, sometimes at a considerable cost to their chances of survival.
 Why are individuals so choosy about who they mate with, particularly when their choice seems to be based on elaborate characters and arbitrary displays that surely lower the viability of their mates?

One major reason why individuals evolve to be choosy about who they mate with is that it can directly impact their fitness. By choosing a mate with particular characteristics, individuals can gain more parental care for their offspring, avoid parasites, or be choosing a mate with higher fertility. For example, female glow-worms flash at night to attract males flying by. Females with larger, brighter lanterns have higher fecundity, so males with a preference for brighter flashes will gain a direct benefit to their own fitness. (Note that males will benefit even if these differences in female fecundity are entirely driven by differences in environment, and so non-heritable.) Indeed male glow worms have evolved to be attracted to brighter flashing lures.

However, even in the absence of direct benefits of choice, selection can still indirectly favour the evolution of choosiness. These indirect benefits occur because individuals can have higher fitness offspring by choosing a mate whose phenotype indicates high viability (the so-called good genes hypothesis), or by choosing a mate whose phenotype is simply attractive, and likely to produce similarly attractive offspring (the 'runaway' or sexy sons hypothesis).

We'll denote a display trait, e.g. tail length, in males by σ and a preference trait in females by φ . Our display trait is under direct selection in males, such that its response to selection can be written

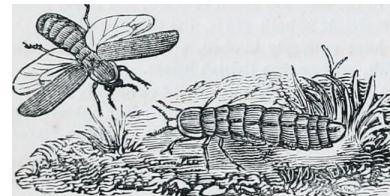


Figure 5.26: Male (left) and female (right) common glow worm (*Lampyris noctiluca*).

The animal kingdom : arranged after its organization; forming a natural history of animals, and an introduction to comparative anatomy. (1863) Cuvier, G. Image from the Biodiversity Heritage Library. Contributed by University of Toronto - Gerstein Science Information Centre. Not in copyright.

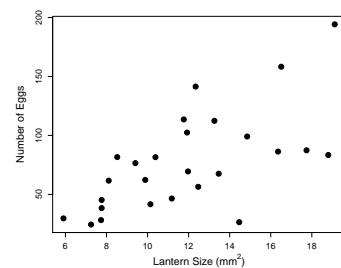


Figure 5.27: Female Glow worms who have the largest, and therefore brightest, lanterns have the highest fecundity. Data from LetterSpace=10HOPKINS *et al.* (2015). Code here.

3460 as

$$R_{\sigma} = \beta_{\sigma} V_{A,\sigma} \quad (5.31)$$

Let's assume that the female preference trait, the degree to which
 3462 females are attracted to long tails, is not under direct selection $\beta_{\varphi} =$
 0. Then the response to selection of the preference trait can be written
 3464 as

$$R_{\varphi} = \beta_{\varphi} V_{A,\varphi} + \beta_{\sigma} V_{A,\varphi\sigma} = \beta_{\sigma} V_{A,\varphi\sigma} \quad (5.32)$$

So the female preference will respond to selection if it is genetically
 3466 correlated with the male trait, i.e. if $V_{A,\varphi\sigma}$ is not zero. There's a
 3468 number of different ways this genetic correlation could arise; the simplest is that the loci underlying the male trait may have a pleiotropic effect on female preference. However, female preference may often have quite a distinct genetic basis from male display traits.

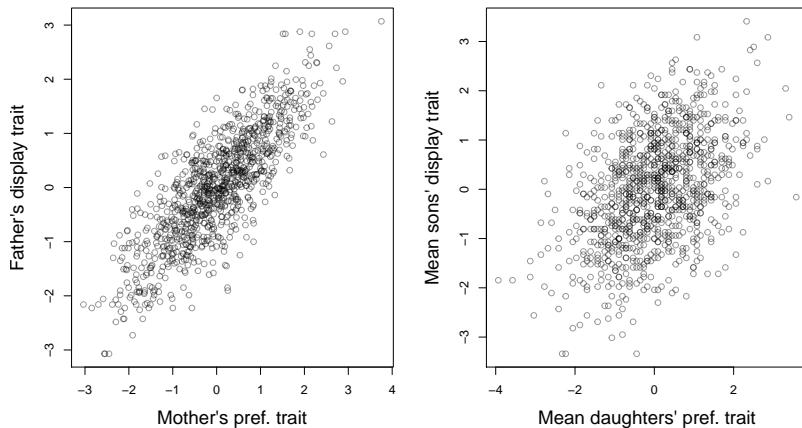


Figure 5.28: **Left**) Assortative mating between males and females. Males vary in a display trait (e.g. tail length), females vary in their preference for this trait. We see evidence of assortative mating as females with a preference for a particular value of the male trait tend to mate with those males. **Right**) As both male trait and female preference are genetic this establishes a genetic correlation in the next generation. This is simulated data. Code here.

3470 A more general way in which trait-preference genetic correlations
 3472 may arise is through assortative mating. As females vary in their
 3474 tail-length preference, the ones with a preference for longer tails
 3476 will mate with long-tailed males and the opposite for females with a
 3478 preference for shorter-tails. Therefore, a genetic correlation between
 Figure 5.28).

3478 The males with the longer tails will also carry the alleles associated
 3480 with the preference for longer tails, as their long-tailed dads
 3482 tended to mate with females with a genetic preference for long tails.
 Similarly, the males with shorter tails will carry alleles associated
 3484 with the preference for shorter tails. Thus if there is direct selection
 for males with longer tails, then the female preference for longer tails
 will increase too, as it is genetically correlated via assortative mating.

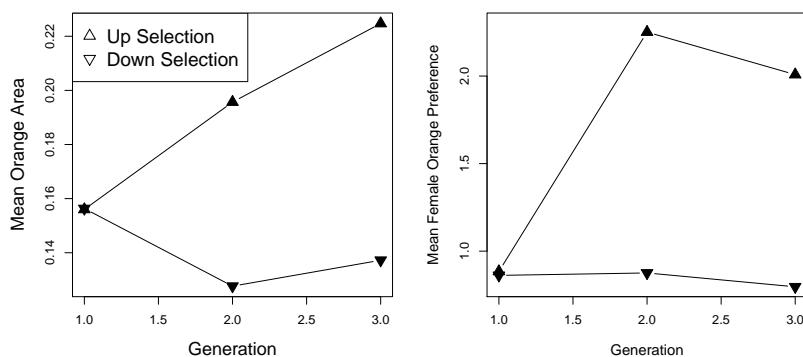


Figure 5.29: Mean phenotypes for the two up- and two down-selected populations of Guppies. Left panel: A response to selection was seen due to the direct selection on male colouration. Right panel: An indirect, correlated response was also seen in female preference. Data from LetterSpace=10HOUDE (1994). Code here.

As an example of how direct selection on display traits can drive the evolution of preference traits, let's consider some data from guppies. Guppies (*Poecilia reticulata*) are a classic system for studying the interplay of natural and sexual selection. In some populations of guppies, females show a preference for males with more orange colouration.

LetterSpace=10HOUDE established four replicate population pairs of guppies and selected one of each pair for an increased or decreased orange coloration in males, selecting the top/bottom 20 out of 50 males. She randomly chose females from each population to form the next generation, and so did not exert direct selection on females. She measured the response to selection on male colouration and on female preference for orange (left and right panels of Figure 5.1 respectively). In the lines that were selected for more orange males females showed an increased preference for orange. While in those lines that she selected males for less orange in their display females showed a decreased preference for orange. This is consistent with indirect selection on female orange preference as a response to selection on male colouration, due to a genetic correlation between female preference and male trait. It is *a priori* unlikely that pleiotropy is the source of the genetic correlation between these traits, rather it is likely caused by females assortative mating with males that match their colour preference.

Returning to our bird tail example, what could drive the direct selection on male tail length? The selection for longer tails in males could come about because longer tails are genetic correlated with higher male viability, for example perhaps only males who gather an excess of food have the resources to invest in growing long tail, i.e. a long tail is an honest signal. This would be a good genes explanation of female mate choice evolution.



Figure 5.30: Guppy (*Poecilia reticulata*). From a set of 1962 stamps of Hungary. Contributed to wikimedia by Darjac, not covered by copyright

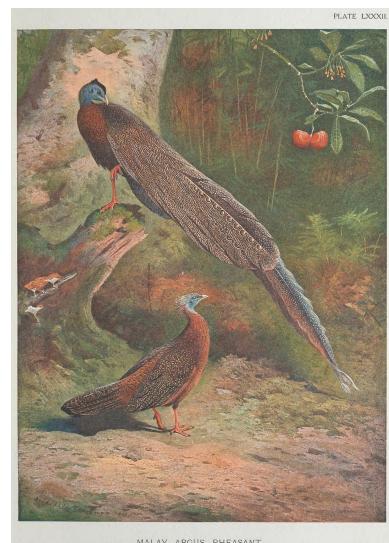


Figure 5.31: Argus Pheasant. A photograph of the phasiana (cav8). Photo: W. Images from

There's another subtler way that selection could favour our male trait. Imagine that the variation in female preference trait is because some females have no strong preference for the male-tail length, but some females have a strong preference for males with longer tails. Males with longer tails would then have higher fecundity than the short-tailed males as there's a subset of females who are strongly attracted to long tails, and these males also get to mate with the other females. Thus selection favours long-tailed males, and so indirectly favours female preference for longer tails; females with a preference for longer-tails have sons who in turn who are more attractive. This model is sometimes called the sexy-son model. It is also called the Fisherian runaway model (LetterSpace=10FISHER, 1915), as female preference and male trait can coevolve in an escalating fashion driving more and more extreme preferences for arbitrary traits. Thus many extravagant display traits in males and females may exist purely because individuals find them beautiful and are attracted to them.

One-Locus Models of Selection

3534 "Socrates consisted of the genes his parents gave him, the experiences
they and his environment later provided, and a growth and develop-
3536 ment mediated by numerous meals. For all I know, he may have been
very successful in the evolutionary sense of leaving numerous off-
3538 spring. His phenotype, nevertheless, was utterly destroyed by the hem-
lock and has never since been duplicated. The same argument holds
3540 also for genotypes. With Socrates' death, not only did his phenotype
disappear, but also his genotype.[...] The loss of Socrates' genotype
3542 is not assuaged by any consideration of how prolifically he may have
reproduced. Socrates' genes may be with us yet, but not his genotype,
3544 because meiosis and recombination destroy genotypes as surely as
death." -LetterSpace=10WILLIAMS (1966)

3546 Individuals are temporary, their phenotypes are temporary, and
their genotypes are temporary. However, the alleles that individuals
3548 transmit across generations have permanence. Sustained pheno-
typic evolutionary change due to natural selection occurs because of
3550 changes in the allelic composition of the population. To understand
these changes, we need to understand how the frequency of alleles
3552 (genes) changes over time due to natural selection.

As we have seen, natural selection occurs when there are differ-
3554 ences between individuals in fitness. We may define fitness in various
ways. Most commonly, it is defined with respect to the contribution
3556 of a phenotype or genotype to the next generation. Differences in
fitness can arise at any point during the life cycle. For instance, differ-
3558 ent genotypes or phenotypes may have different survival probabili-
ties from one stage in their life to the stage of reproduction (viability),
3560 or they may differ in the number of offspring produced (fertility),
or both. Here, we define the absolute fitness of a genotype as the
3562 expected number of offspring of an individual of that genotype. Dif-
ferences in fitness among genotypes drive allele frequency change.
3564 In this chapter we'll study the dynamics of alleles at a single locus.
In this chapter we'll ignore the effects of genetic drift, and just study
3566 the deterministic dynamics of selection. We'll return to discuss the

interaction of selection and drift in the next chapter.

3568 **6.0.1 Haploid selection model**

We start out by modeling selection in a haploid model, as this is mathematically relatively simple. Let the number of individuals carrying alleles A_1 and A_2 in generation t be P_t and Q_t . Then, the relative frequencies at time t of alleles A_1 and A_2 are $p_t = P_t / (P_t + Q_t)$ and $q_t = Q_t / (P_t + Q_t) = 1 - p_t$. Further, assume that individuals of type A_1 and A_2 on average produce W_1 and W_2 offspring individuals, respectively. We call W_i the absolute fitness.

3576 Therefore, in the next generation, the absolute number of carriers of A_1 and A_2 are $P_{t+1} = W_1 P_t$ and $Q_{t+1} = W_2 Q_t$, respectively. The 3578 mean absolute fitness of the population at time t is

$$\bar{W}_t = W_1 \frac{P_t}{P_t + Q_t} + W_2 \frac{Q_t}{P_t + Q_t} = W_1 p_t + W_2 q_t, \quad (6.1)$$

3580 i.e. the sum of the fitness of the two types weighted by their relative frequencies. Note that the mean fitness depends on time, as it is a function of the allele frequencies, which are themselves time dependent.

3584 As an example of a rapid response to selection on an allele in a haploid population, we can consider some data on the evolution of drug resistant viruses. LetterSpace=10FEDER *et al.* (2017) studied viral dynamics in a macaque infected with a strain of simian immunodeficiency virus (SHIV) that carries the HIV-1 reverse transcriptase coding region. The viral load of the macaque's blood plasma is shown as a black line in Figure 6.1. Twelve weeks after infection, the 3590 macaque was treated with an anti-retroviral drug that targeted the the virus' reverse transcriptase protein. Note how the viral load initially starts to drop once the drug is administered, suggesting that the absolute fitness of the original strain is less than one ($W_2 < 1$) in the presence of the drug (as their numbers are decreasing). However, the viral population rebounds as a mutation that confers drug 3596 resistance to the anti-retroviral drug arises in the SHIV and starts to spread. Viruses carrying this mutation (let's call them allele 1) likely have absolute fitness $W_1 > 1$. The frequency of the drug-resistant allele is shown in red; it quickly spreads from being undetectable in week 13, to being fixed in the SHIV population in week 20.

3602 The rapid spread of this drug-resistant allele through the population is driven by the much greater relative fitness of the drug-resistant allele over the original strain in the presence of the anti-retroviral drug.

The main focus of LetterSpace=10FEDER *et al.*'s work was modeling the complicated spatial dynamics of drug-resistant SHIV adaptation in different organ systems.

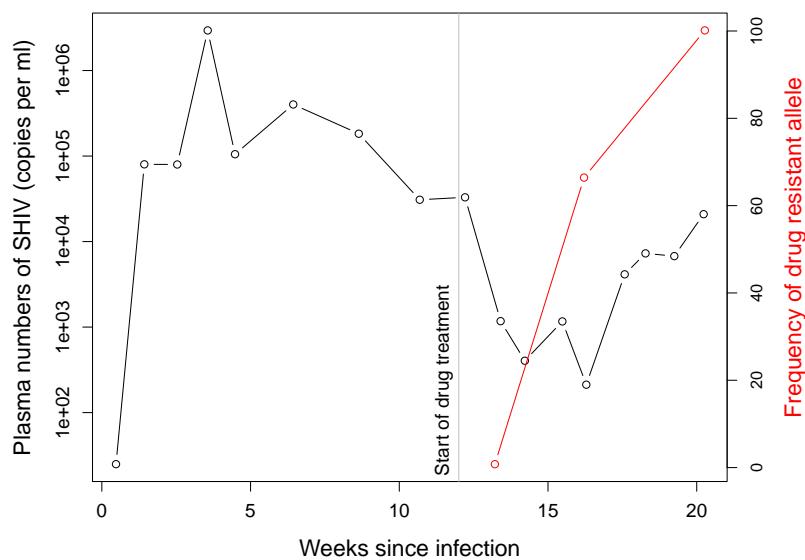


Figure 6.1: The rapid evolution of drug-resistant SHIV. The viral load of SHIV in the blood of a macaque (black line), the frequency of a drug resistance mutation (red line). Data from LetterSpace=10FEDER et al. (2017). Code here.

The frequency of allele A_1 in the next generation is given by

$$p_{t+1} = \frac{P_{t+1}}{P_{t+1} + Q_{t+1}} = \frac{W_1 P_t}{W_1 P_t + W_2 Q_t} = \frac{W_1 p_t}{W_1 p_t + W_2 q_t} = \frac{W_1}{\bar{W}_t} p_t. \quad (6.2)$$

Importantly, eqn. (6.2) tells us that the change in p only depends on a ratio of fitnesses. Therefore, we need to specify fitness only up to an arbitrary constant. As long as we multiply all fitnesses by the same value, that constant will cancel out and eqn. (6.2) will hold.

Based on this argument, it is very common to scale absolute fitnesses by the absolute fitness of one of the genotypes, e.g. the most or the least fit genotype, to obtain relative fitnesses. Here, we will use w_i for the relative fitness of genotype i . If we choose to scale by the absolute fitness of genotype A_1 , we obtain the relative fitnesses $w_1 = W_1/W_1 = 1$ and $w_2 = W_2/W_1$.

Without loss of generality, we can therefore rewrite eqn. (6.2) as

$$p_{t+1} = \frac{w_1}{\bar{w}} p_t, \quad (6.3)$$

dropping the subscript t for the dependence of the mean fitness on time in our notation, but remembering it. The change in frequency from one generation to the next is then given by

$$\Delta p_t = p_{t+1} - p_t = \frac{w_1 p_t}{\bar{w}} - p_t = \frac{w_1 p_t - \bar{w} p_t}{\bar{w}} = \frac{w_1 p_t - (w_1 p_t + w_2 q_t) p_t}{\bar{w}} = \frac{w_1 - w_2}{\bar{w}} p_t q_t, \quad (6.4)$$

recalling that $q_t = 1 - p_t$.

Assuming that the fitnesses of the two alleles are constant over time, the number of the two allelic types τ generations after time t

are $P_{t+\tau} = (W_1)^\tau P_t$ and $Q_{t+\tau} = (W_2)^\tau Q_t$, respectively. Therefore, the
 3624 relative frequency of allele A_1 after τ generations past t is

$$p_{t+\tau} = \frac{(W_1)^\tau P_t}{(W_1)^\tau P_t + (W_2)^\tau Q_t} = \frac{(w_1)^\tau P_t}{(w_1)^\tau P_t + (w_2)^\tau Q_t} = \frac{p_t}{p_t + (w_2/w_1)^\tau q_t}, \quad (6.5)$$

where the last step includes dividing the whole term by $(w_1)^\tau$ and
 3626 switching from absolute to relative allele frequencies.

Rearranging eqn. (6.5) and setting $t = 0$, we can work out the time
 3628 τ for the frequency of A_1 to change from p_0 to p_τ . First, we write

$$p_\tau = \frac{p_0}{p_0 + (w_2/w_1)^\tau q_0} \quad (6.6)$$

and rearrange this to obtain

$$\frac{p_\tau}{q_\tau} = \frac{p_0}{q_0} \left(\frac{w_1}{w_2} \right)^\tau. \quad (6.7)$$

3630 Solving this for τ yields

$$\tau = \log \left(\frac{p_\tau q_0}{q_\tau p_0} \right) / \log \left(\frac{w_1}{w_2} \right). \quad (6.8)$$

In practice, it is often helpful to parametrize the relative fitnesses
 3632 w_i in a specific way. For example, we may set $w_1 = 1$ and $w_2 = 1 - s$,
 where s is called the selection coefficient. Using this parametrization,
 3634 s is simply the difference in relative fitnesses between the two alleles.
 Equation (6.5) becomes

$$p_{t+\tau} = \frac{p_t}{p_t + q_t(1-s)^\tau}, \quad (6.9)$$

3636 as $w_2/w_1 = 1 - s$. Then, if $s \ll 1$, we can approximate $(1 - s)^\tau$ in the
 denominator by $\exp(-s\tau)$ to obtain

$$p_{t+\tau} \approx \frac{p_t}{p_t + q_t e^{-s\tau}}. \quad (6.10)$$

3638 This equation takes the form of a logistic function. That is because we
 are looking at the relative frequencies of two ‘populations’ (of alleles
 3640 A_1 and A_2) that are growing (or declining) exponentially, under the
 constraint that p and q always sum to 1.

3642 Moreover, eqn. (6.7) for the number of generations τ it takes for a
 certain change in frequency to occur becomes

$$\tau = -\log \left(\frac{p_\tau q_0}{q_\tau p_0} \right) / \log(1 - s). \quad (6.11)$$

3644 Assuming again that $s \ll 1$, this simplifies to

$$\tau \approx \frac{1}{s} \log \left(\frac{p_\tau q_0}{q_\tau p_0} \right). \quad (6.12)$$

One particular case of interest is the time it takes to go from an
 3646 absolute frequency of 1 to near fixation in a population of size N . In
 this case, we have $p_0 = 1/N$, and we may set $p_\tau = 1 - 1/N$, which is
 3648 very close to fixation. Then, plugging these values into eqn. (6.12), we
 obtain

$$\begin{aligned}\tau &= \frac{1}{s} \log \left(\frac{1 - 2/N + 1/N^2}{1/N^2} \right) \\ &\approx \frac{1}{s} (\log(N) + \log(N - 2)) \\ &\approx \frac{2}{s} \log(N)\end{aligned}\tag{6.13}$$

3650 where we make the approximations $N^2 - 2N + 1 \approx N^2 - 2N$ and
 later $N - 2 \approx N$.

3652 **Question 1.** In our example of the evolution of drug resistance,
 the drug-resistant SHIV virus spread from undetectable frequencies
 3654 to $\sim 65\%$ frequency by 16 weeks post infection. An estimated ef-
 fective population size of SHIV is 1.5×10^5 , and its generation time
 3656 is ~ 1 day. Assuming that the mutation arose as a single copy al-
 lele very shortly the start of drug treatment at 12 weeks, what is the
 3658 selection coefficient favouring the drug resistance allele?

Haploid model with fluctuating selection Selection pressures may
 3660 change while a polymorphism persists in the population due to en-
 vironmental changes. We can use our haploid model to consider this
 3662 case where the fitnesses depend on time (LetterSpace=10DEMPSTER,
 1955), and say that $w_{1,t}$ and $w_{2,t}$ are the fitnesses of the two types in
 3664 generation t . The frequency of allele A_1 in generation $t + 1$ is

$$p_{t+1} = \frac{w_{1,t}}{\bar{w}_t} p_t,\tag{6.14}$$

which simply follows from eqn. (6.3). The ratio of the frequency of
 3666 allele A_1 to that of allele A_2 in generation $t + 1$ is

$$\frac{p_{t+1}}{q_{t+1}} = \frac{w_{1,t}}{w_{2,t}} \frac{p_t}{q_t}.\tag{6.15}$$

Therefore, if we think of the two alleles starting in generation t at
 3668 frequencies p_t and q_t , then τ generations later,

$$\frac{p_{t+\tau}}{q_{t+\tau}} = \left(\prod_{i=t}^{\tau-1} \frac{w_{1,i}}{w_{2,i}} \right) \frac{p_t}{q_t}.\tag{6.16}$$

The question of which allele is increasing or decreasing in fre-
 3670 quency comes down to whether $\left(\prod_{i=t}^{\tau-1} w_{1,i}/w_{2,i} \right)$ is > 1 or < 1 . As it is

a little hard to think about this ratio, we can instead take the τ^{th} root of it and consider

$$\sqrt[\tau]{\left(\prod_{i=t}^{\tau-1} \frac{w_{1,i}}{w_{2,i}}\right)} = \frac{\sqrt[\tau]{\prod_{i=t}^{\tau-1} w_{1,i}}}{\sqrt[\tau]{\prod_{i=t}^{\tau-1} w_{2,i}}} \quad (6.17)$$

The term

$$\sqrt[\tau]{\prod_{i=t}^{\tau-1} w_{1,i}} \quad (6.18)$$

is the geometric mean fitness of allele A_1 over the τ generations past generation t . Therefore, allele A_1 will only increase in frequency if it has a higher geometric mean fitness than allele A_2 (at least in our simple deterministic model). This implies that an allele with higher geometric mean fitness can even invade and spread to fixation if its (arithmetic) mean fitness is lower than the dominant type. To see this consider two alleles that experience the fitnesses given in Table 6.1. The allele A_1 does much better in dry years, but suffers in wet years; while the A_2 is generalist and is not affected by the variable environment. If there is an equal chance of a year being wet or dry, the A_1 allele has higher (arithmetic) mean fitness, but it will be replaced by the A_2 allele as the A_2 allele has higher geometric mean fitness (See Figure 6.2).

	A_1	A_2
Dry	2	1.57
Wet	1.16	1.57

Table 6.1: Fitnesses of two alleles in wet and dry years. Means calculated assuming equal chances of wet and dry years. The geometric mean is calculated as $\sqrt{w_{\text{wet}}w_{\text{dry}}}$. Example numbers taken from LetterSpace=10SEGER and LetterSpace=10BROCKMANN (1987).

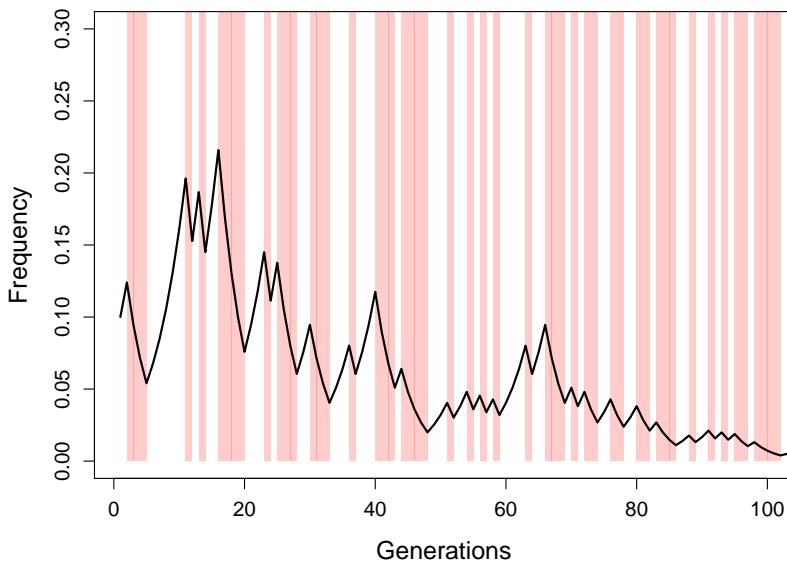


Figure 6.2: An example frequency trajectory of the A_2 allele under variable environments (using the fitnesses from Table 6.1). Dry years (generations) are shown in red, wet years in white. Note how the A_2 allele decreases in frequency in the dry years as A_1 has higher fitness, and yet the A_2 allele still wins out. Code here.

Evolution of bet hedging Don't put your eggs in one basket, it makes
 3688 a lot of sense to spread your bets. Financial advisors often advise you
 to diversify your portfolio, rather than placing all your investments
 3690 in one stock. Even if that stock looks very strong, you can come a
 cropper that 1/20 times some particular part of the market crashes.
 3692 Likewise, evolution can result in risk averse strategies. Some species
 of bird lay multiple nests of eggs; some plants don't put all of their
 3694 energy into seeds that will germinate next year. It can even make
 sense to hedge your bets even if that comes at an average cost (LetterSpace=10SEGER and LetterSpace=10BROCKMANN, 1987).
 3696

To see this lets think more about geometric fitness. We can write
 3698 the fitness in a given generation i as $w_i = 1 + s_i$, such that we can
 write your geometric fitness as

$$\bar{g} = \sqrt[\tau]{\prod_{i=t}^{\tau-1} 1 + s_i} \quad (6.19)$$

when we think about products it's often natural to take the log to
 turn it into a sum

$$\begin{aligned} \log(\bar{g}) &= \frac{1}{\tau} \sum_{i=t}^{\tau-1} \log(1 + s_i) \\ &= \mathbb{E} [\log(1 + s_i)] \end{aligned} \quad (6.20)$$

equating the mean and the expectation. Assuming that s_i is small
 $\log(1 + s_i) \approx s_i - s_i^2/2$, ignoring terms s_i^3 and higher then this is

$$\begin{aligned} \log(\bar{g}) &\approx \mathbb{E} [s_i - s_i^2/2] \\ &= \mathbb{E} [s_i] - \text{var}(s_i)/2 \end{aligned} \quad (6.21)$$

3700 So genotypes with high arithmetic mean fitness can be selected
 against, i.e. have low geometric mean fitness against, if their fitness
 3702 has too high a variance across generations (LetterSpace=10GILLESPIE,
 1973, 1977). See our example above, Table 6.1 and Figure 6.2).

3704 A classic example of bet-hedging is in delayed seed germination
 in plants (LetterSpace=10COHEN, 1966). In variable environments,
 3706 such as deserts, it may make sense to spread your bets over years by
 having only a proportion of your seeds germinate in the first year.
 3708 However, delaying germination can come at a cost due to seed mor-
 tality. LetterSpace=10GREMER and LetterSpace=10VENABLE (2014),
 3710 using data from a long-term study various species of Sonoran Desert
 winter showed that annual plants were indeed pursuing adaptive

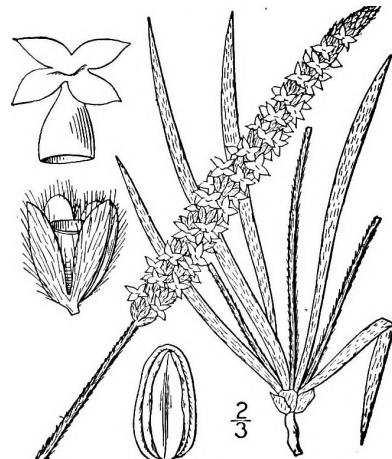


Figure 6.3: Woolly plantain (*Plantago patagonica*). One of the desert annuals shown to have a bet-hedging germination strategy by LetterSpace=10GREMER and LetterSpace=10VENABLE (2014).

An illustrated flora of the northern United States, Canada and the British possessions, from Newfoundland to the parallel of the southern boundary of Virginia, and from the Atlantic Ocean westward to the 102d meridian (1913) Britton, N.L. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. Not in copyright.

³⁷¹² bet-hedging strategies. The plant species with the highest variation in among-year yield had the lowest germination fraction per year.
³⁷¹⁴ Further, LetterSpace=10GREMER and LetterSpace=10VENABLE showed through modeling life that by having per-year germination proportions < 1 all of the species were achieving higher geometric fitness at the expense of arithmetic fitness in the variable desert environment.
³⁷¹⁶ See Figure 6.4 for an example of bet hedging in Woolly plantain.

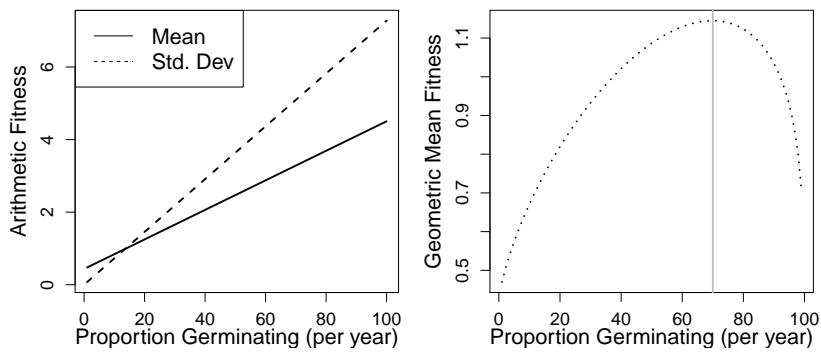


Figure 6.4: *Plantago patagonica*'s arithmetic fitness is an increasing function of the proportion of seeds germinating, due to seeds not surviving a germination delay. However, the standard deviation of fitness also increases with this proportion as they are more likely to have all of their seeds germinate in a bad year. Thus *Plantago patagonica* can achieve higher geometric fitness by only having a proportion of their seeds germinate. Thanks to Jenny Gremer for sharing these data from LetterSpace=10GREMER and LetterSpace=10VENABLE (2014)

³⁷²⁰ Delayed reproduction is also a common example of bet-hedging in micro-organisms. For example, the Chicken Pox virus, varicella zoster virus, has a very long latent phase. After it causes chicken pox
³⁷²² it enters a latent phase, residing, inactive, in neurons in the spinal cord, only to emerge 5-40 years later to cause the disease shingles. It
³⁷²⁴ is hypothesized that the virus actively suppresses itself as a strategy to allow it to emerge at a later time point as insurance against there
³⁷²⁶ being no further susceptible hosts at the time of its first infection (LetterSpace=10STUMPF *et al.*, 2002).

³⁷²⁸ 6.0.2 Diploid model

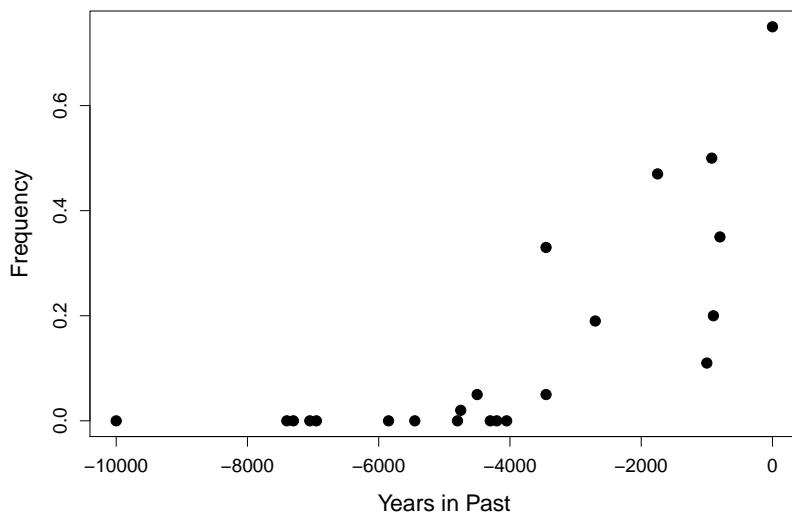


Figure 6.5: Frequency of the Lactase persistence allele in ancient and modern samples from Central Europe. Data compiled by LetterSpace=10MARCINIAK and LetterSpace=10PERRY (2017) from various sources. Thanks to Stephanie Marciniak for sharing these data. Code here.

We will now move on to a diploid model of a single locus with two segregating alleles. As an example of the change in the frequency of an allele driven by selection, let's consider the evolution of Lactase persistence. A number of different human populations that historically have raised cattle have convergently evolved to maintain the expression of the protein Lactase into adulthood (in most mammals the protein is switched off after childhood), with different lactase-persistence mutations having arisen and spread in different pastoral human populations. This continued expression of Lactase allows adults to break down Lactose, the main carbohydrate in milk, and so benefit nutritionally from milk-drinking. This seems to have offered a strong fitness benefit to individuals in pastoral populations.

With the advent of techniques to sequence ancient human DNA, researchers can now potentially track the frequency of selected mutations over thousands of years. The frequency of a Lactase persistence allele in ancient Central European populations is shown in Figure 6.5. The allele is absent more than 5,000 years ago, but now found at frequency of upward of 70% in many European populations.

We will assume that the difference in fitness between the three genotypes comes from differences in viability, i.e. differential survival of individuals from the formation of zygotes to reproduction. We denote the absolute fitnesses of genotypes A_1A_1 , A_1A_2 , and A_2A_2 by W_{11} , W_{12} , and W_{22} . Specifically, W_{ij} is the probability that a zygote of genotype A_iA_j survives to reproduction. Assuming that individuals



Figure 6.6: Auroch (*Bos primigenius*). Aurochs are an extinct species of large wild cattle that cows were domesticated from.

Dictionnaire des sciences naturelles. 1816 Cuvier, F.G. Image from the Internet Archive. Contributed by NCSU Libraries. No known copyright restrictions.

mate at random, the number of zygotes that are of the three genotypes and form generation t are

$$Np_t^2, \quad N2p_tq_t, \quad Nq_t^2. \quad (6.22)$$

The mean fitness of the population of zygotes is then

$$\bar{W}_t = W_{11}p_t^2 + W_{12}2p_tq_t + W_{22}q_t^2. \quad (6.23)$$

Again, this is simply the weighted mean of the genotypic fitnesses.

How many zygotes of each of the three genotypes survive to re-

produce? An individual of genotype A_1A_1 has a probability of W_{11}

of surviving to reproduce, and similarly for other genotypes. There-

fore, the expected number of A_1A_1 , A_1A_2 , and A_2A_2 individuals who survive to reproduce is

$$NW_{11}p_t^2, \quad NW_{12}2p_tq_t, \quad NW_{22}q_t^2. \quad (6.24)$$

It then follows that the total number of individuals who survive to

reproduce is

$$N \left(W_{11}p_t^2 + W_{12}2p_tq_t + W_{22}q_t^2 \right). \quad (6.25)$$

This is simply the mean fitness of the population multiplied by the population size (i.e. $N\bar{w}$).

The relative frequency of A_1A_1 individuals at reproduction is

simply the number of A_1A_1 genotype individuals at reproduction

$(NW_{11}p_t^2)$ divided by the total number of individuals who survive to reproduce ($N\bar{W}$), and likewise for the other two genotypes. There-

fore, the relative frequency of individuals with the three different genotypes at reproduction is

$$\frac{NW_{11}p_t^2}{N\bar{W}}, \quad \frac{NW_{12}2p_tq_t}{N\bar{W}}, \quad \frac{NW_{22}q_t^2}{N\bar{W}} \quad (6.26)$$

(see Table 6.2).

	A_1A_1	A_1A_2	A_2A_2
Absolute no. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	W_{11}	W_{12}	W_{22}
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_tq_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{W_{11}}{\bar{W}} p_t^2$	$\frac{W_{12}}{\bar{W}} 2p_tq_t$	$\frac{W_{22}}{\bar{W}} q_t^2$

Table 6.2: Relative genotype frequencies after one episode of viability selection.

As there is no difference in the fecundity of the three genotypes,

the allele frequencies in the zygotes forming the next generation are simply the allele frequency among the reproducing individuals of the

previous generation. Hence, the frequency of A_1 in generation $t + 1$ is

$$p_{t+1} = \frac{W_{11}p_t^2 + W_{12}p_tq_t}{\bar{W}}. \quad (6.27)$$

³⁷⁷⁸ Note that, again, the absolute value of the fitnesses is irrelevant to the frequency of the allele. Therefore, we can just as easily replace the ³⁷⁸⁰ absolute fitnesses with the relative fitnesses. That is, we may replace W_{ij} by $w_{ij} = W_{ij}/\bar{W}$, for instance.

³⁷⁸² Each of our genotype frequencies is responding to selection in a manner that depends just on its fitness compared to the mean fitness ³⁷⁸⁴ of the population. For example, the frequency of the 11 homozygotes increases from birth to adulthood in proportion to W_{11}/\bar{W} . In fact, ³⁷⁸⁶ we can estimate this fitness ratio for each genotype by comparing the frequency at birth compared to adults. As an example of this ³⁷⁸⁸ calculation, we'll look at some data from sticklebacks.

³⁷⁹⁰ Marine threespine stickleback (*Gasterosteus aculeatus*) independently colonized and adapted to many freshwater lakes as glaciers receded following the last ice age, making sticklebacks a wonderful ³⁷⁹² system for studying the genetics of adaptation. In marine habitats, most of the stickleback have armour plates to protect them from ³⁷⁹⁴ predation, but freshwater populations repeatedly evolve the loss of armour plates due to selection on an allele at the *Ectodysplasin* gene (EDA). This allele is found as a standing variant at very low ³⁷⁹⁶ frequency marine populations; LetterSpace=10BARRETT *et al.* took advantage of this fact and collected and bred a population of marine ³⁷⁹⁸ individuals carrying both the low- (L) and completely-plated (C) alleles. They introduced the offspring of this cross into four freshwater ponds and monitored genotype frequencies ¹ over their life courses:

	CC	LC	LL
Juveniles	0.55	0.23	0.22
Adults	0.21	0.53	0.26
Adults/Juv. (W_{\bullet}/\bar{W})	0.4	2.3	1.2
rel. fitness (W_{\bullet}/W_{12})	0.17	1.0	0.54

³⁸⁰² The heterozygotes have increased in frequency dramatically in the population as their fitness is more than double the mean fitness of the population. We can also calculate the relative fitness of each ³⁸⁰⁴ genotype by dividing through by the fitness of the fittest genotype, the heterozygote in this case (doing this cancels through \bar{W}). The ³⁸⁰⁶ relative fitness of the CC is $\sim 1/5$ of the heterozygote. Note that this calculation does not rely on the genotype frequencies being at their ³⁸⁰⁸ HWE in the juveniles.

³⁸¹² **Question 2. A** What is the frequency of the low-plated EDA allele (L) at the start of the stickleback experiment?

B What is the frequency in the adults?

³⁸¹⁴ **Question 3.** For many generations you have been studying an annual wildflower that has two color morphs, orange and white. You

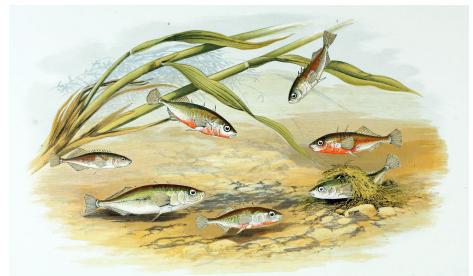


Figure 6.7: Freshwater threespine Stickleback (*G. aculeatus*).
British fresh-water fishes. Houghton W 1879. Image from the Biodiversity Heritage Library. Contributed by Ernst Mayr Library, Harvard.. Not in copyright.

¹ The actual dynamics observed by LetterSpace=10BARRETT *et al.* are more complicated as in the very young fish selection reverses direction.

3816 have discovered that a single bi-allelic locus controls flower color,
 with the white allele being recessive. The pollinator of these plants
 3818 is an almost blind bat, so individuals are pollinated at random with
 respect to flower color. Your population census of 200 individuals
 3820 showed that the population consisted of 168 orange-flowered individ-
 uals, and 32 white-flowered individuals.

3822 Heavy February rainfall creates optimal growing conditions for
 an exotic herbivorous beetle with a preference for orange-flowered
 3824 individuals. This year it arrives at your study site with a ravenous
 appetite. Only 50% of orange-flowered individuals survive its wrath,
 3826 while 90% of white-flowered individuals survive until the end of the
 growing season.

3828 **A** What is the initial frequency of the white allele, and what do
 you have to assume to obtain this?

3830 **B** What is the frequency of the white allele in the seeds forming
 the next generation?

3832 The change in frequency from generation t to $t + 1$ is

$$\Delta p_t = p_{t+1} - p_t = \frac{w_{11}p_t^2 + w_{12}p_tq_t}{\bar{w}} - p_t. \quad (6.28)$$

3834 To simplify this equation, we will first define two variables \bar{w}_1 and \bar{w}_2
 as

$$\bar{w}_1 = w_{11}p_t + w_{12}q_t, \quad (6.29)$$

$$\bar{w}_2 = w_{12}p_t + w_{22}q_t. \quad (6.30)$$

These are called the marginal fitnesses of allele A_1 and A_2 , respec-
 3836 tively. They are so called as \bar{w}_1 is the average fitness of an allele A_1 ,
 i.e. the fitness of A_1 in a homozygote weighted by the probability
 3838 it is in a homozygote (p_t) plus the fitness of A_1 in a heterozygote
 weighted by the probability it is in a heterozygote (q_t). We further
 3840 note that the mean relative fitness can be expressed in terms of the
 marginal fitnesses as

$$\bar{w} = \bar{w}_1p_t + \bar{w}_2q_t, \quad (6.31)$$

3842 where, for notational simplicity, we have omitted subscript t for the
 dependence of mean and marginal fitnesses on time.

3844 We can then rewrite eqn. (6.28) using \bar{w}_1 and \bar{w}_2 as

$$\Delta p_t = \frac{(\bar{w}_1 - \bar{w}_2)}{\bar{w}} p_t q_t. \quad (6.32)$$

3846 The sign of Δp_t , i.e. whether allele A_1 increases or decreases in fre-
 quency, depends only on the sign of $(\bar{w}_1 - \bar{w}_2)$. The frequency of
 3848 A_1 will keep increasing over the generations so long as its marginal
 fitness is higher than that of A_2 , i.e. $\bar{w}_1 > \bar{w}_2$, while if $\bar{w}_1 < \bar{w}_2$, the

frequency of A_1 will decrease. Note the similarity between eqn. (6.32) and the respective expression for the haploid model in eqn. (6.4). (We will return to the special case where $\bar{w}_1 = \bar{w}_2$ shortly).

We can also rewrite (6.28) as

$$\Delta p_t = \frac{1}{2} \frac{p_t q_t}{\bar{w}} \frac{d\bar{w}}{dp}, \quad (6.33)$$

the demonstration of which we leave to the reader. This form shows that the frequency of A_1 will increase ($\Delta p_t > 0$) if the mean fitness is an increasing function of the frequency of A_1 (i.e. if $\frac{d\bar{w}}{dp} > 0$). On the other hand, the frequency of A_1 will decrease ($\Delta p_t < 0$) if the mean fitness is a decreasing function of the frequency of A_1 (i.e. if $\frac{d\bar{w}}{dp} < 0$). Thus, although selection acts on individuals, under this simple model, selection is acting to increase the mean fitness of the population. The rate of this increase is proportional to the variance in allele frequencies within the population ($p_t q_t$).

Question 4. Show that eqns. (6.33) and (6.32) are equivalent.

(Trickier question.)

So far, our treatment of the diploid model of selection has been in terms of generic fitnesses w_{ij} . In the following, we will use particular parametrizations to gain insight about two specific modes of selection: directional selection and heterozygote advantage.

6.0.3 Diploid directional selection

Directional selection means that one of the two alleles always has higher marginal fitness than the other one. Let us assume that A_1 is the fitter allele, so that $w_{11} \geq w_{12} \geq w_{22}$, and hence $\bar{w}_1 > \bar{w}_2$. As we are interested in changes in allele frequencies, we may use relative fitnesses. We parameterize the reduction in relative fitness in terms of a selection coefficient, similar to the one we met in the haploid selection section, as follows:

genotype	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$
absolute fitness	W_{11}	$\geq W_{12} \geq$	W_{22}
relative fitness (generic)	$w_{11} = W_{11}/W_{11}$	$w_{12} = W_{12}/W_{11}$	$w_{22} = W_{22}/W_{11}$
relative fitness (specific)	1	$1 - sh$	$1 - s$.

Here, the selection coefficient s is the difference in relative fitness between the two homozygotes, and h is the dominance coefficient. For selection to be directional, we require that $0 \leq h \leq 1$ holds. The dominance coefficient allows us to move between two extremes. One is when $h = 0$, such that allele A_1 is fully dominant and A_2 fully recessive. In this case, the heterozygote $A_1 A_2$ is as fit as the $A_1 A_1$

homozygote genotype. The inverse holds when $h = 1$, such that allele A_1 is fully recessive and A_2 fully dominant.

We can then rewrite eqn. (6.32) as

$$\Delta p_t = \frac{p_t h s + q_t s(1-h)}{\bar{w}} p_t q_t, \quad (6.34)$$

where

$$\bar{w} = 1 - 2p_t q_t s h - q_t^2 s. \quad (6.35)$$

Question 5. Throughout the Californian foothills are old copper and gold-mines, which have dumped out soils that are polluted with heavy metals. While these toxic mine tailing are often depauperate of plants, *Mimulus guttatus* and a number of other plant species have managed to adapt to these harsh soils. LetterSpace=10WRIGHT *et al.* (2015) have mapped one of the major loci contributing to the adaptation to soils at two mines near Copperopolis, CA. LetterSpace=10WRIGHT *et al.* planted homozygote seedlings out in the mine tailings and found that only 10% of the homozygotes for the non-copper-tolerant allele survived to flower, while 40% of the copper-tolerant seedlings survived to flower.

A) What is the selection coefficient acting against the non-copper-tolerant allele on the mine tailing?

B) The copper-tolerant allele is fairly dominant in its action on fitness. If we assume that $h = 0.1$, what percentage of heterozygotes should survive to flower?

Question 6. Comparing the red ($h = 0$) and black ($h = 0.5$) trajectories in Figure 6.8, provide an explanation for why A_1 increases faster initially if $h = 0$, but then approaches fixation more slowly compared to the case of $h = 0.5$.

To see how dominance affects the trajectory of a real polymorphism, we'll consider an example from a colour polymorphism in red foxes (*Vulpes vulpes*). There are three colour morphs of red foxes: silver, cross, and red (see Figure 6.11), with this difference primarily controlled by a single polymorphism with genotypes RR, Rr, and rr respectively. The fur pelts of the silver morph fetched three times the price for hunters compared to cross (a smoky red) and red pelts, the latter two being seen as roughly equivalent in worth. Thus the desirability of the pelts acts as a recessive trait, with much stronger selection against the silver homozygotes. As a result of this price difference, silver foxes were hunted more intensely and declined as a proportion of the population in Eastern Canada, see Figure 6.10, as documented by LetterSpace=10ELTON, from 16% to 5% from 1834 to 1937. LetterSpace=10HALDANE reanalyzed these data and showed

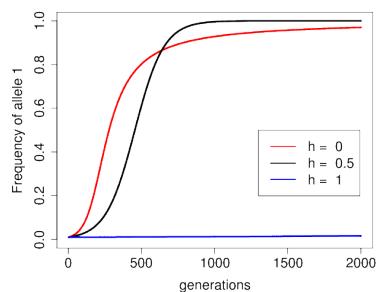


Figure 6.8: The trajectory of the frequency of allele A_1 , starting from $p_0 = 0.01$, for a selection coefficient $s = 0.01$ and three different dominance coefficients. The recessive beneficial allele ($h = 1$) will eventually fix in the population, but it takes a long time. Code here.

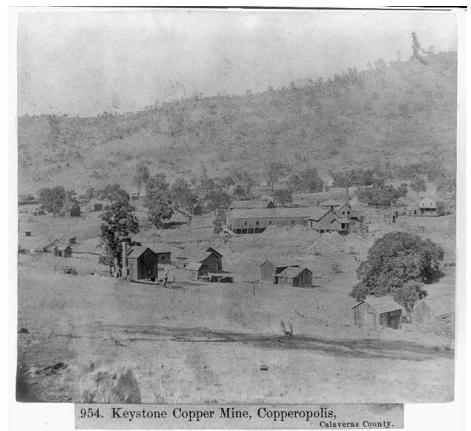


Figure 6.9: Keystone Copper Mine 1866, Copperopolis, Calaveras County. Image from picryl. Source Library of Congress, Public Domain.

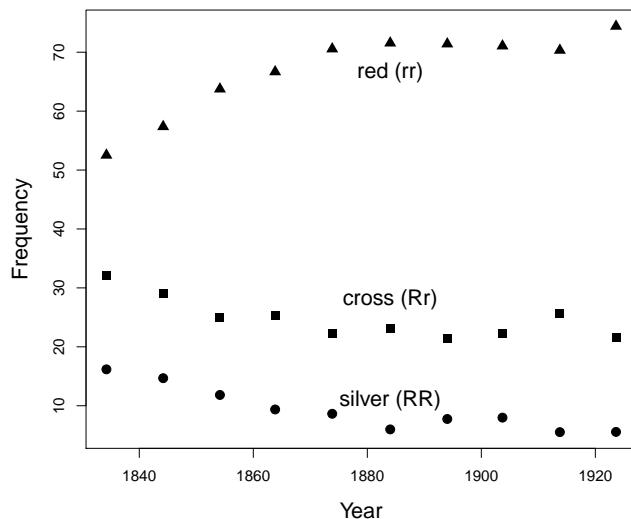


Figure 6.10: The frequency of red, cross, and silver fox morphs over the decades in Eastern Canada. These data are well described by recessive selection acting against the silver fox morph. Data from LetterSpace=10 ELTON (1942), compiled by LetterSpace=10 ALLENDORF and LetterSpace=10 HARD (2009). Code here.

3922 that they were consistent with recessive selection acting against the
silver morph alone.

3924 Note how the heterozygotes (cross) decline somewhat as a result
of selection on the silver homozygotes, but overall the R allele is
3926 slow to respond to selection as it is ‘hidden’ from selection in the
heterozygote state.

3928 *Directional selection on an additive allele.* A special case is when $h = 0.5$. This case is the case of no dominance, as the interaction among
3930 alleles with respect to fitness is strictly additive. Then, eqn. (6.34)
simplifies to

$$\Delta p_t = \frac{1}{2} \frac{s}{\bar{w}} p_t q_t. \quad (6.36)$$

3932 If selection is very weak, i.e. $s \ll 1$, the denominator (\bar{w}) is close to
1 and we have

$$\Delta p_t = \frac{1}{2} s p_t q_t. \quad (6.37)$$

3934 It is instructive to compare eqn. (6.37) to the respective expression
under the haploid model. To this purpose, start from the generic
3936 term for Δp_t under the haploid model in eqn. (6.4) and set $w_1 = 1$
and $w_2 = 1 - s$. Again, assume that s is small, so that eqn. (6.4)
3938 becomes $\Delta p_t = s p_t q_t$. Hence, if s is small, the diploid model of
directional selection without dominance is identical to the haploid
3940 model, up to a factor of 1/2. That factor is due to the choice of the
parametrisation; we could have set $w_{11} = 1$, $w_{12} = 1 - s$, and $w_{22} =$

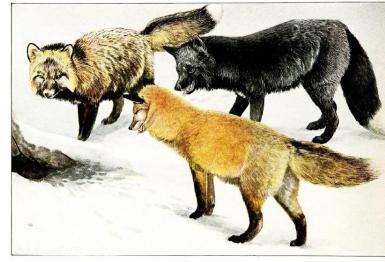


Figure 6.11: Three colour morphs in red fox *V. vulpes*, cross, red, and silver foxes from left to right.
The larger North American mammals" Nelson, E.W., Fuertes,
L.A. 1916. Image from the Biodiversity Heritage Library.
Contributed by Cornell University Library. No known copyright
restrictions.

³⁹⁴² 1 – 2s in our diploid model instead, in which case the agreement with the haploid model would be perfect.

³⁹⁴⁴ From this analogy, we can borrow some insight we gained from the haploid model. Specifically, the trajectory of the frequency of allele A_1 in the diploid model without dominance follows a logistic growth curve similar to (6.10). From this similarity, we can extrapolate from Equation (6.12) to find the time it takes for our diploid, beneficial, additive allele (A_1) to move from frequency p_0 to p_τ :

$$\tau \approx \frac{2}{s} \log \left(\frac{p_\tau q_0}{q_\tau p_0} \right) \quad (6.38)$$

³⁹⁵⁰ generations; this just differs by a factor of 2 from our haploid model. Using this result we can find the time it takes for our favourable, ³⁹⁵² additive allele (A_1) to transit from its entry into the population ($p_0 = 1/(2N)$) to close to fixation ($p_\tau = 1 - 1/(2N)$):

$$\tau \approx \frac{4}{s} \log(2N) \quad (6.39)$$

³⁹⁵⁴ generations. Note the similarity to eqn. 6.13 for the haploid model, with a difference by a factor of 2 due to the choice of parametrization ³⁹⁵⁶ (and that the number of alleles is $2N$ in the diploid model, rather than N). Doubling our selection coefficient halves the time it takes for ³⁹⁵⁸ our allele to move through the population.

Question 7. Gulf killifish (*Fundulus grandis*) have rapidly adapted ³⁹⁶⁰ to the very high pollution levels in the Houston shipping canal since the 1950s. One of the ways that they've adapted is through the deletion ³⁹⁶² of their aryl hydrocarbon receptor (AHR) gene. Oziolor et al. estimated that individuals who were homozygote for the intact AHR ³⁹⁶⁴ gene had a relative fitness of 20% of that of homozygotes for the deletion. Assuming an effective population size of 200 thousand individuals, how long would it take for the deletion to reach fixation, ³⁹⁶⁶ starting as a single copy in this population?

³⁹⁶⁸ Directional selection on genotypes is expected to remove variation from populations, yet we see plentiful phenotypic and genetic ³⁹⁷⁰ variation in every natural population. Why is this? Three broad explanations for the maintenance of polymorphisms are

- ³⁹⁷² 1. Variation is maintained by a balance of genetic drift and mutation (we discussed this explanation in Chapter 3).
- ³⁹⁷⁴ 2. Selection can sometimes act to maintain variation in populations (balancing selection).
- ³⁹⁷⁶ 3. Deleterious variation can be maintained in the population as a balance between selection removing variation and mutation constantly introducing new variation into the population.

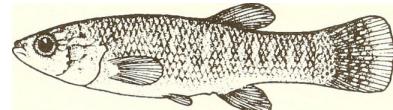


Figure 6.12: Gulf killifish (*Fundulus grandis*).

Distribution and abundance of fishes and invertebrates in Gulf of Mexico estuaries. Nelson D M and Pattiello M E Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. No known copyright restrictions.

We'll turn to these latter two explanations through the rest of the chapter. Note that these explanations are not mutually exclusive, and each of them will explain some proportion of the variation.

3982 6.0.4 Heterozygote advantage

One form of balancing selection occurs when the heterozygotes are fitter than either of the homozygotes. In this case, it is useful to parameterize the relative fitnesses as follows:

genotype	A_1A_1	A_1A_2	A_2A_2
absolute fitness	w_{11}	$w_{12} < w_{22}$	w_{22}
relative fitness (generic)	$w_{11} = W_{11}/W_{12}$	$w_{12} = W_{12}/W_{12}$	$w_{22} = W_{22}/W_{12}$
relative fitness (specific)	$1 - s_1$	1	$1 - s_2$

3988 Here, s_1 and s_2 are the differences between the relative fitnesses of the two homozygotes and the heterozygote. Note that to obtain 3990 relative fitnesses we have divided absolute fitness by the heterozygote fitness. We could use the same parameterization as in the model of 3992 directional selection, but the reparameterization we have chosen here makes the math easier.

3994 In this case, when allele A_1 is rare, it is often found in a heterozygous state, while the A_2 allele is usually in the homozygous state, 3996 and so A_1 is more fit and increases in frequency. However, when the allele A_1 is common, it is often found in a less fit homozygous state, 3998 while the allele A_2 is often found in a heterozygous state; thus it is now allele A_2 that increases in frequency at the expense of allele A_1 . 4000 Thus, at least in the deterministic model, neither allele can reach fixation and both alleles will be maintained at an equilibrium frequency 4002 as a balanced polymorphism in the population.

4004 We can solve for this equilibrium frequency by setting $\Delta p_t = 0$ in eqn. (6.32), i.e. $p_t q_t (\bar{w}_1 - \bar{w}_2) = 0$. Doing so, we find that there are three equilibria, all of which are stable. Two of them are not very interesting ($p = 0$ or $q = 0$), but the third one is the polymorphic equilibrium, where $\bar{w}_1 - \bar{w}_2 = 0$ holds. Using our s_1 and s_2 4008 parametrization above, we see that the marginal fitnesses of the two alleles are equal when

$$p_e = \frac{s_2}{s_1 + s_2} \quad (6.40)$$

4010 for the equilibrium frequency of interest. This is also the frequency 4012 of A_1 at which the mean fitness of the population is maximized. The highest possible fitness of the population would be achieved if every 4014 individual was a heterozygote. However, Mendelian segregation of alleles in the gametes of heterozygotes means that a sexual population can never achieve a completely heterozygote population. This

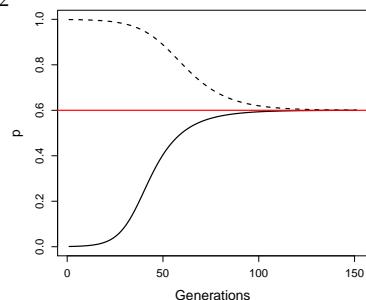


Figure 6.13: Two allele frequency trajectories of the A_1 allele subject to heterozygote advantage ($w_{11} = 0.9$, $w_{12} = 1$, and $w_{22} = 0.85$). In one simulation the allele is started from being rare in the population ($p = 1/1000$, solid line) and increases in frequency/ In the other simulation the allele is almost fixed ($p = 999/1000$, dashed line). In both cases the frequency moves toward the equilibrium frequency. The red line shows the equilibrium frequency (p_e). Code here.

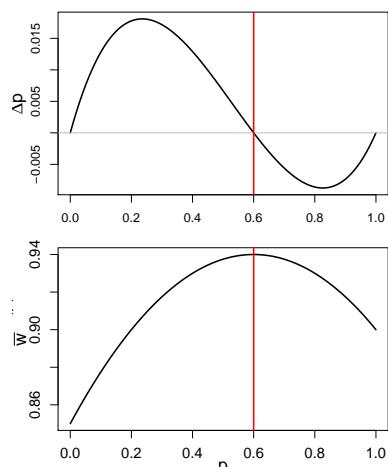
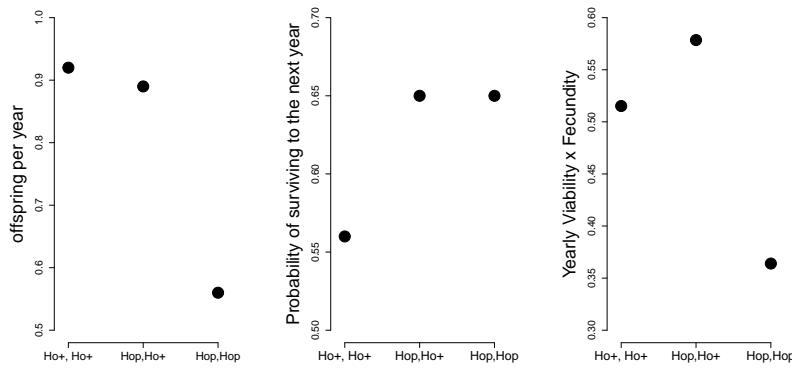


Figure 6.14: Top) The change in frequency of an allele with heterozygote advantage within a generation (Δp) as a function of the allele frequency. Fitnesses as in Figure 6.13. Note how the frequency change is positive below the equilibrium frequency (p_e) and negative

4016 equilibrium frequency represents an evolutionary compromise between the advantages of the heterozygote and the comparative costs
 4018 of the two homozygotes.



One example of a polymorphism maintained by heterozygote advantage is a horn-size polymorphism found in Soay sheep, a population of feral sheep on the island of Soay (about 40 miles off the coast of Scotland). The horns of the soay sheep resemble those of the wild Mouflon sheep, and the male Soay sheep use their horns to defend females during the rut. LetterSpace=10JOHNSTON *et al.* (2013) found a large-effect locus, at the RXFP2 gene, that controls much of the genetic variation for horn size. Two alleles Ho^p and Ho^+ segregate at this locus. The Ho^+ allele is associated with growing larger horns, while the Ho^p allele is associated with smaller horns, with a reasonable proportion of Ho^p homozygotes developing no horns at all. LetterSpace=10JOHNSTON *et al.* (2013) found that the Ho locus had substantial effects on male, but not female, fitness (see Figure 6.16).

The Ho^p allele has a mostly recessive effect on male fecundity, with the Ho^p homozygotes having lower yearly reproductive success presumably due to the fact that they perform poorly in male-male competition (left plot Figure 6.16). Conversely, the Ho^+ has a mostly recessive effect on viability, with Ho^+ homozygotes having lower yearly survival (middle plot Figure 6.16), likely because they spend little time feeding during the rut and so lose substantial body weight. Thus both of the homozygotes suffer from trade-offs between viability and fecundity. As a result, the Ho^pHo^+ heterozygotes have the highest fitness (right plot Figure 6.16). The allele is thus balanced at intermediate frequency (50%) in the population due to this trade off between fitness at different life history stages.

Question 8. Assume that the frequency of the Ho^P allele is 10%, that there are 1000 males at birth, and that individual adults mate at

Figure 6.15: For the three Soay sheep genotypes: the offspring per year (left), the probability of surviving a year (middle), and the product of the two (right). Thanks to Susan Johnston for supplying these simplified numbers from LetterSpace=10JOHNSTON *et al.* (2013). Code here.

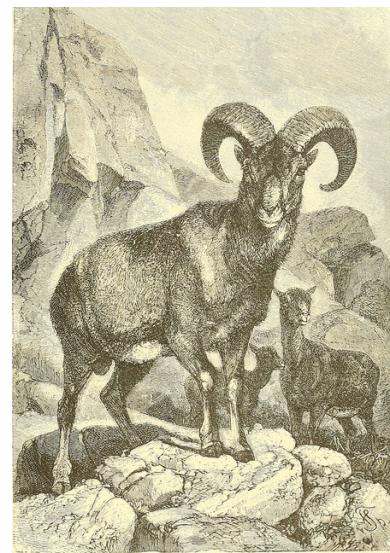


Figure 6.16: Mouflon (*Ovis orientalis orientalis*).
 Animate creation. (1898). Wood, J. G. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

The fitnesses here are chosen to roughly match those of the real Soay sheep example, as a full model would require us to more carefully model the life-histories of the sheep.

random.

A) What is the expected number of males with each of the three genotypes in the population at birth?

B) Assume that a typical male individual of each genotypes has the following probability of surviving to adulthood:

$$\begin{array}{cccccc} \text{Ho}^+ & \text{Ho}^+ & \text{Ho}^+ & \text{Ho}^p & \text{Ho}^p & \text{Ho}^p \\ 0.5 & & & 0.8 & & 0.8 \end{array}$$

Making the assumptions from above, how many males of each genotype survive to reproduce? C) Of the males who survive to reproduce, let's say that males with the Ho^+Ho^+ and Ho^+Ho^p genotype have on average 2.5 offspring, while Ho^pHo^p males have on average 1 offspring. Taking into account both survival and reproduction, how many offspring do you expect each of the three genotypes to contribute to the total population in the next generation?

D) What is the frequency of the Ho^+ allele in the sperm that will form this next generation?

E) How would your answers to B-D change if the Ho^p allele was at 90% frequency?

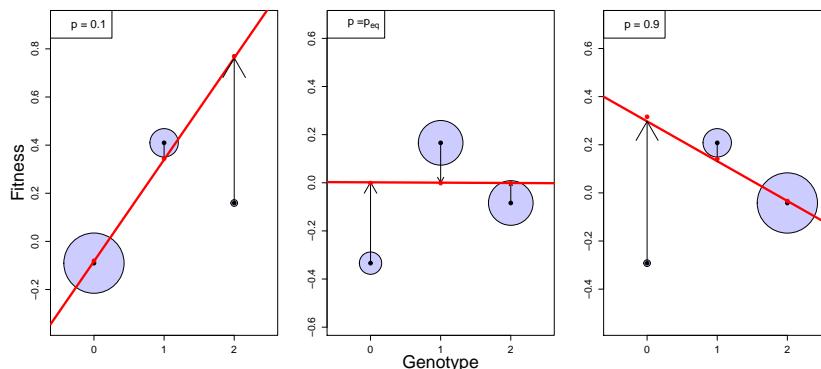


Figure 6.17: The deviation of the fitness of each genotype away from the mean population fitness (\bar{w}) is shown as black dots. The area of each circle is proportion to the fraction of the population in each genotypic class (p^2 , $2pq$, and q^2). The additive genetic fitness of each genotype is shown as a red dot. The linear regression between fitness and additive genotype is shown as a red line. The black vertical arrows show the difference between the average mean-centered phenotype and additive genetic value for each genotype. The left panel shows $p = 0.1$ and the right panel shows $p = 0.9$; in the middle panel the frequency is set to the equilibrium frequency. Code here.

To push our understanding of heterozygote advantage a little further, note that the marginal fitnesses of our alleles are equivalent to the additive effects of our alleles on fitness. Recall from our discussion of non-additive variation (Section 4.1.1) that the difference in the additive effects of the two alleles gives the slope of the regression of additive genotypes on fitness, and that there is additive variance in fitness when this slope is non-zero. So what's happening here in our heterozygote advantage model is that the marginal fitness of the A_1 allele, the additive effect of allele A_1 on fitness, is greater than the marginal fitness of the A_2 allele ($\bar{w}_1 > \bar{w}_2$) when A_1 is at low frequency in the population. In this case, the regression of fitness on the number of A_1 alleles in a genotype has a positive slope. This is true when the frequency of the A_1 allele is below the equilibrium

frequency. If the frequency of A_1 is above the equilibrium frequency, then the marginal fitness of allele A_2 is higher than the marginal fitness of allele A_1 ($\bar{w}_1 < \bar{w}_2$) and the regression of fitness on the number of copies of allele A_1 that individuals carry is negative. In both cases there is additive genetic variance for fitness ($V_A > 0$) and the population has a directional response. Only when the population is at its equilibrium frequency, i.e. when $\bar{w}_1 = \bar{w}_2$, is there no additive genetic variance ($V_A = 0$), as the linear regression of fitness on genotype is zero.

Underdominance. Another case that is of potential interest is the case of fitness underdominance, where the heterozygote is less fit than either of the two homozygotes. Underdominance can be parametrized as follows:

genotype	A_1A_1	A_1A_2	A_2A_2
absolute fitness	w_{11}	$> w_{12} <$	w_{22}
relative fitness (generic)	$w_{11} = W_{11}/W_{12}$	$w_{12} = W_{12}/W_{12}$	$w_{22} = W_{22}/W_{12}$
relative fitness (specific)	$1 + s_1$	1	$1 + s_2$

Underdominance also permits three equilibria: $p = 0$, $p = 1$, and a polymorphic equilibrium $p = p_U$. However, now only the first two equilibria are stable, while the polymorphic equilibrium is unstable. If $p < p_U$, then Δp_t is negative and allele A_1 will be lost, while if $p > p_U$, allele A_1 will become fixed.

While strongly-selected, underdominant alleles might not spread within populations (if $p_U \gg 0$), they are of special interest in the study of speciation and hybrid zones. That is because alleles A_1 and A_2 may have arisen in a stepwise fashion, i.e. not by a single mutation, but in separate subpopulations. In this case, heterozygote disadvantage will play a potential role in species maintenance.

Question 9. You are studying the polymorphism that affects flight speed in butterflies. The polymorphism does not appear to affect fecundity. Homozygotes for the B allele are slow in flight and so only 40% of them survive to have offspring. Heterozygotes for the polymorphism (Bb) fly quickly and have a 70% probability of surviving to reproduce. The homozygotes for the alternative allele (bb) fly very quickly indeed, but often die of exhaustion, with only 10% of them making it to reproduction.

- 4110 A) What is the equilibrium frequency of the B allele?
- B) Calculate the marginal absolute fitnesses of the B and the b allele at the equilibrium frequency.



Figure 6.18: In *Pseudacraea eurytus* there are two homozygotes morphs that mimic a different blue and orange butterfly; the heterozygote fails to mimic either successfully and so suffers a high rate of predation (LetterSpace=10OWEN and LetterSpace=10CHANTER, 1972). Illustrations of new species of exotic butterflies (1868) Hewitson. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

figures/het_disadvant_dp_wbar.pdf

Figure 6.19: **Left)** Two allele frequency trajectories of an A_1 allele subject to heterozygote disadvantage ($w_{11} = 1.1$, $w_{12} = 1$, and $w_{22} = 1.2$). The allele is started from just above and below the equilibrium frequency, in both cases the frequency move away the equilibrium frequency. The red line shows the unstable equilibrium frequency (p_e). **Middle)** The change in frequency of an allele with heterozygote disadvantage within a generation (Δp) as a function of the allele frequency. Fitnesses as in Figure 6.13. Note how the frequency change is negative below the equilibrium frequency (p_e) and positive above. **Right)** Mean fitness (\bar{w}) as a function of the allele frequency. Code here.

Diploid fluctuating fitness Selection pressures fluctuate over time and can potentially maintain polymorphisms in the population. Two examples of polymorphisms fluctuating in frequency in response to temporally-varying selection are shown in Figure 6.20; thanks to the short lifespan of *Drosophila* we can see seasonally-varying selection. The first example is an inversion allele in *Drosophila pseudoobscura* populations. Throughout western North America, two orientations of the chromosome, two 'inversion alleles', exist: the Chiricahua and Standard alleles. LetterSpace=10DOBZHANSKY (1943) and LetterSpace=10WRIGHT and LetterSpace=10DOBZHANSKY (1946) investigated the frequency of these inversion alleles over four years at a number of locations and found that their frequency fluctuated systematically over the seasons in response to selection (left side of 6.20). If you're still reading these notes send Prof. Coop a picture of Dobzhansky; Dobzhansky was one of the most important evolutionary geneticists of the past century and spent a bunch of time at UC Davis in his later years. Our second example is an insertion-deletion polymorphism in the Insulin-like Receptor gene in *Drosophila melanogaster*. LetterSpace=10PAABY *et al.* (2014) tracked the frequency of this allele over time and found it oscillated with the seasons (right side of 6.20). She and her coauthors also determined that these alleles had large effects on traits such as developmental time and fecundity, which could mediate the maintenance of this polymorphism through life-history trade-offs.

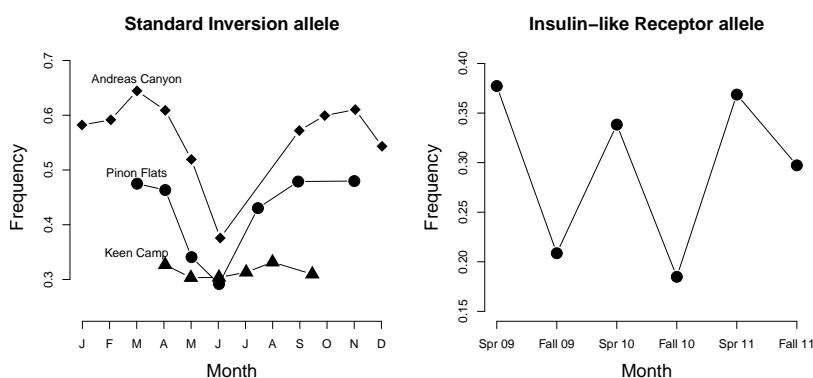


Figure 6.20: **Left)** Seasonal variation in the frequency of the 'Standard' inversion allele in *Drosophila pseudoobscura* for three populations from Mount San Jacinto, CA. These frequencies are an average over four years. Data from LetterSpace=10WRIGHT and LetterSpace=10DOBZHANSKY (1946). **Right)** The frequency of an allele at the Insulin-like Receptor gene over three years in *Drosophila melanogaster* samples from an Orchard in Pennsylvania. Data from LetterSpace=10PAABY *et al.* (2014). Code here.

To explore temporal fluctuations in fitness, we'll need to think about the diploid absolute fitnesses being time-dependent, where the three genotypes have fitnesses $w_{11,t}$, $w_{12,t}$, and $w_{22,t}$ in generation t . Modeling the diploid case with time-dependent fitness is much less tractable than the haploid case, as segregation makes it tricky to keep track of the genotype frequencies. However, we can make

some progress and gain some intuition by thinking about how the frequency of allele A_1 changes when it is rare (following the work of LetterSpace=10 HALDANE and LetterSpace=10 JAYAKAR, 1963).

When A_1 is rare, i.e. $p_t \ll 1$, the frequency of A_1 in the next generation (6.27) can be approximated as

$$p_{t+1} \approx \frac{w_{12}}{\bar{w}} p_t. \quad (6.41)$$

To obtain this equation, we have ignored the p_t^2 term (because it is very small when p_t is small) and we have assumed that $q_t \approx 1$ in the numerator. Following a similar argument to approximate q_{t+1} , we can write

$$\frac{p_{t+1}}{q_{t+1}} = \frac{w_{12,t} p_t}{w_{22,t} q_t}. \quad (6.42)$$

Starting from out from p_0 and q_0 in generation 0, then $t + 1$ generations later we have

$$\frac{p_{t+1}}{q_{t+1}} = \left(\prod_{i=0}^t \frac{w_{12,i}}{w_{22,i}} \right) \frac{p_0}{q_0}. \quad (6.43)$$

From this we can see, following our haploid argument from above, that the frequency of allele A_1 will increase when rare only if

$$\frac{\sqrt[t]{\prod_{i=0}^t w_{12,i}}}{\sqrt[t]{\prod_{i=0}^t w_{22,i}}} > 1, \quad (6.44)$$

i.e. if the heterozygote has higher geometric mean fitness than the A_2A_2 homozygote.

The question now is whether allele A_1 will approach fixation in the population, or whether there are cases in which we can obtain a balanced polymorphism. To investigate that, we can simply repeat our analysis for $q \ll 1$, and see that in that case

$$\frac{p_{t+1}}{q_{t+1}} = \left(\prod_{i=0}^t \frac{w_{11,i}}{w_{12,i}} \right) \frac{p_0}{q_0}. \quad (6.45)$$

Now, for allele A_1 to carry on increasing in frequency and to approach fixation, the A_1A_1 genotype has to be out-competing the heterozygotes. For allele A_1 to approach fixation, we need the geometric mean of $w_{11,i}$ to be greater than the geometric mean fitness of heterozygotes ($w_{12,i}$). At the same time, if heterozygotes have higher geometric mean fitness than the A_1A_1 homozygotes, then the A_2 allele will increase in frequency when it is rare.

Intriguingly, we can thus have a balanced polymorphism even if the heterozygote is never the fittest genotype in any generation, as long as the heterozygote has a higher geometric mean fitness than either of the homozygotes. In this case, the heterozygote comes out

ahead when we think about long-term fitness across heterogeneous environmental conditions, despite never being the fittest genotype in any particular environment.

As a toy example of this type of balanced polymorphism, consider a plant population found in one of two different environments each generation. These occur randomly; $1/2$ of time the population experiences the dry environment and with probability $1/2$ it experiences the wet environment. The absolute fitnesses of the genotypes in the different environments are as follows:

Environment	AA	Aa	aa
Wet	6.25	5.0	3.75
Dry	3.85	5.0	6.15
arithmetic mean	5.05	5.0	4.95

Let's write $w_{AA,dry}$ and $w_{AA,wet}$ for the fitnesses of the AA homozygote in the two environments. Then, if the two environments are equally common, $\prod_{i=0}^t w_{AA,i} \approx w_{AA,dry}^{t/2} w_{AA,wet}^{t/2}$ for large values of t . To obtain an estimate of this product normalized over the t generations, we can take the t^{th} root to obtain the geometric mean fitness. Taking the t^{th} root, we find the geometric mean fitness of the AA allele is $w_{AA,dry}^{1/2} w_{AA,wet}^{1/2}$. Doing this for each of our genotypes, we find the geometric mean fitnesses of our alleles to be:

	AA	Aa	aa
Geometric mean	4.91	5.0	4.80

i.e. the heterozygote has higher geometric mean fitnesses than either of the homozygotes, despite not being the fittest genotype in either environment (nor having the highest arithmetic mean fitness). So the A_1 allele can invade the population when it is rare as it spread thanks to the higher fitness of the heterozygotes. Similarly the A_2 allele can invade the population when it is rare. Thus both alleles will persist in the population due to the environmental fluctuations, and the higher geometric mean fitness of the heterozygotes.

Negative frequency-dependent selection. In the models and examples above, heterozygote advantage maintains multiple alleles in the population because the common allele has a disadvantage compared to the other rarer allele. In the case of heterozygote advantage, the relative fitnesses of our three genotypes are not a function of the other genotypes present in the population. However, there's a broader set of models where the relative fitness of a genotype depends on the genotypic composition of the population; this broad family of models is called frequency-dependent selection. Negative frequency-dependent selection, where the fitness of an allele (or phenotype)

This example is loosely based on the work of LetterSpace=10SCHEMSKE and LetterSpace=10BIERZYCHUDEK (2001) on *Linanthus parryae*, a desert annual, endemic to California. There are blue- and a white-flowered colour morphs polymorphic many populations, with this polymorphism being controlled by a single dominant allele. The blue-flowered plants produce more seeds in dry years, i.e. they have higher fitness in these years, while the white-flowered plants have higher seed production in wet years. Thus both morphs can potentially be maintained in the population. See LetterSpace=10TURELLI *et al.* (2001) for a more detailed analysis.

decreases as it becomes more common in the population, can act to maintain genetic and phenotypic diversity within populations. While cases of long-term heterozygote advantage may be somewhat rare in nature, negative frequency-dependent selection is likely a common form of balancing selection.

One common mechanism that may create negative frequency-dependent selection is the interaction between individuals within or among species. For example, negative frequency-dependent dynamics can arise in predator-prey or pathogen-host dynamics, where alleles conferring common phenotypes are at a disadvantage because predators or pathogens learn or evolve to counter the phenotypic effects of common alleles.

As one example of negative frequency-dependent selection, consider the two flower colour morphs in the deceptive Elderflower orchid (*Dactylorhiza sambucina*). Throughout Europe, there are populations of these orchids polymorphic for yellow- and purple-flowered individuals, with the yellow flower corresponding to a recessive allele. Neither of these morphs provide any nectar or pollen reward to their bumblebee pollinators. Thus these plants are typically pollinated by newly emerged bumblebees who are learning about which plants offer food rewards, with the bees alternating to try a different coloured flower if they find no food associated with a particular flower-colour morph (LetterSpace=10SMITHSON and LetterSpace=10MACNAIR, 1997). LetterSpace=10GIGORD *et al.* (2001) explored whether this behaviour by bees could result in negative frequency-dependent selection; out in the field, the researchers set up experimental orchid plots in which they varied the frequency of the two colour morphs. Figure 6.22 shows their measurements of the relative male and female reproductive success of the yellow morph across these experimental plots. When the yellow morph is rare, it has higher reproductive success than the purple morph, as it receives a disproportionate number of visits from bumblebees that are dissatisfied with the purple flowers. This situation is reversed when the yellow morph becomes common in the population; now the purple morph outperforms the yellow morph. Therefore, both colour morphs are maintained in this population, and presumably Europe-wide, due to this negative frequency-dependent selection.

Negative frequency-dependent selection can also maintain different breeding strategies due to interactions amongst individuals within a population. One dramatic example of this occurs in ruffs (*Philomachus pugnax*), a marsh-wading sandpiper that summers in Northern Eurasia. The males of this species lek, with the males gathering on open ground to display and attract females. There are three different male morphs differing in their breeding strategy. The



Figure 6.21: Elderflower orchid (*Dactylorhiza sambucina*).

Abbildungen der in Deutschland und den angrenzenden gebieten vorkommenden grundformen der orchideenarten (1904). Mäjiller, W. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden. Not in copyright.

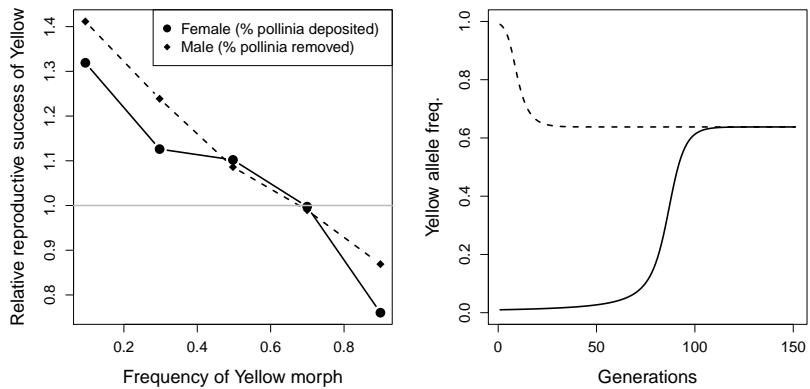


Figure 6.22: **Left)** Measures of the relative male- and female- reproductive success of the yellow Elderflower orchid morph as a function of the yellow morph in experimental plots. **Right)** Two allele frequency trajectories of the Yellow allele subject to negative frequency scheme given in the left plot (for an initial frequency of 0.01 and 0.99, solid and dotted line respectively). Note that the yellow Male reproductive success is measured in terms of the % of pollinia removed front a plant and female reproductive success is measured in terms of the % of stigmas receiving pollinia on a plant. These measures are made relative by dividing the reproductive success of the yellow morph by the mean of the yellow and purple morphs. Pollinia are the pollen masses of orchids, and other plants, where individual pollinium are transferred as a single unit by pollinators. Data from LetterSpace=10GIGORD *et al.* (2001). Code here.

4254 large majority of males are 'Independent', with black or chestnut ruff
 4255 plumage, and try to defend and display on small territories. 'Satel-
 4256 lite' males, with white ruff plumage, make up $\sim 16\%$ of males and
 4257 do not defend territories, but rather join in displays with Indepen-
 4258 dent males and opportunistically mate with females visiting the lek.
 Finally, the rare 'Faeder' morph was only discovered in 2006 (LetterSpace=10JUKEMA and LetterSpace=10PIERSMA, 2006) and makes
 4260 up less than 1% of males. These Faeder males are female mimics who
 4262 hang around the territories of Independents and try to 'sneak' in
 4263 matings with females. Faeder males have plumage closely resembling
 4264 that of females and a smaller body size than other males, but with
 4266 larger testicles (presumably to take advantage of rare mating oppor-
 4267 tunities). All three of these morphs, with their complex behavioural
 4268 and morphological differences, are controlled by three alleles at a sin-
 4269 gle autosomal locus, with the Satellite and Faeder alleles being genet-
 4270 ically dominant over the high frequency Independent allele. The
 4271 genetic variation for these three morphs is potential maintained by
 4272 negative frequency-dependent selection, as all three male strategies
 4273 are likely at an advantage when they are rare in the population. For
 example, while the Satellites mostly lose out on mating opportunities
 4274 to Independents, they may have longer life-spans and so may have
 equal life-time reproductive success (LetterSpace=10WIDEMO, 1998).
 However, Satellite and Faeder males are totally reliant on the lekking
 4276 Independent males, and so both of these alternative strategies can-
 4278 not become overly common in the population. The locus controlling
 4280 these differences has been mapped, and the underlying alleles have
 persisted for roughly four million years (LetterSpace=10KÜPPER *et al.*,
 2016; LetterSpace=10LAMICHHANEY *et al.*, 2016). While this mating
 4282 system is bizarre, the frequency dependent dynamics mean that it



Figure 6.23: Lekking Ruffs (*Philomachus pugnax*). Three Independent males, one Satellite male, and one female (or Faeder male?).

Painting by Johann Friedrich Naumann (1780–1857). Public Domain, wikimedia.

has been around longer than we've been using stone tools.

4284 While these examples may seem somewhat involved, they must
be simple compared to the complex dynamics that maintain the hun-
4286 dreds of alleles present at the genes in the Major histocompatibility
complex (MHC). MHC genes are key to the coordination of the verte-
4288 brate immune system in response to pathogens, and are likely caught
in an endless arms race with pathogens adapting to common MHC
4290 alleles, allowing rare MHC alleles to be favoured. Balancing selection
at the MHC locus has maintained some polymorphisms for tens of
4292 millions of years, such that some of your MHC alleles may be geneti-
cally more closely related to MHC alleles in other primates than they
4294 are to alleles in your close human friends.

6.1

4296 We have seen that when selection acts in a simple manner it can act
to increase the mean fitness of the population. However, when the
4298 absolute fitnesses of individuals are frequency dependent, e.g. de-
pend on the strategies deployed by others in the population, natural
4300 selection is not guaranteed to increase mean fitness. One place where
this is particularly apparent is in the evolution of a 50/50 sex ratio. In
4302 fact as we'll see that selection can drive the evolution of traits that are
actively harmful to the fitness of an individual, when selection acts
4304 below the level of an individual.

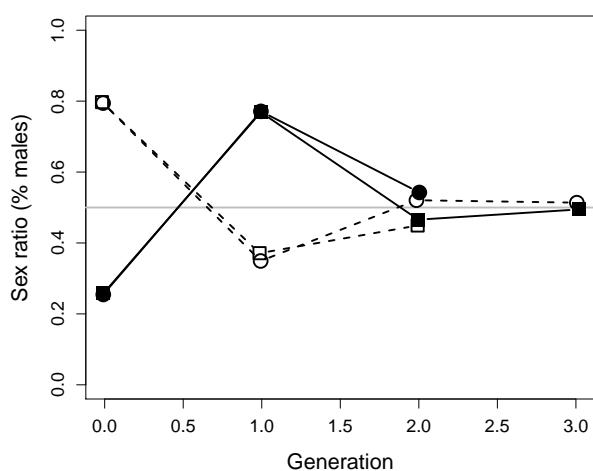


Figure 6.24: LetterSpace=10BASOLO (1994) explored sex ratio dynamics in platyfish (*Xiphophorus maculatus*), which has manipulable sex ratio due to its three factor sex determination. She started two replicates with a strong female bias (black) and two replicates with strong male bias (white). In all four cases the sex ratio quickly oscillated to a 50/50 sex ratio. Data from LetterSpace=10BASOLO (1994), Code here.

In many species, regardless of the mechanism of sex determination, the sex ratio is close to 50/50. Yet this is far from the optimum sex ratio from the perspective of the population viability. In many species females are the limiting sex, investing more in gametes and (sometimes) more in parental care, thus a population having many females and few males would offer the fastest rate of population growth (i.e. the highest mean fit). Imagine if the population sex ratio was strongly skewed towards females. A rare autosomal allele that caused a mother to produce sons would have high fitness, as the mother's sons would have high reproductive success in this population of most females. Thus our initially rare allele would initially increase in frequency. Conversely if the sex ratio was strongly skewed towards males, a rare autosomal allele that causes a mother to produce daughters would spread. So selection on autosomal alleles favours the production of the rare sex, a form of negative frequency dependence, this pushes the sex ratio away from being too skewed. Only the 50/50 sex ratio is evolutionarily stable as there is no rarer sex, and so no (autosomal) sex-ratio-altering mutation can invade a population with a 50/50. The 50/50 sex ratio is an example of an evolutionary stable strategy (ESS), described in more detail in Section 6.1.2. Our population is held well away from its female-bias optimum for population growth as individual-level selection favours the production of the rarer sex, which results in a 50/50 sex ratio.

Adaptive adjustments to sex ratio in response to local mate competition. There are, however, situations where we see strong deviations away

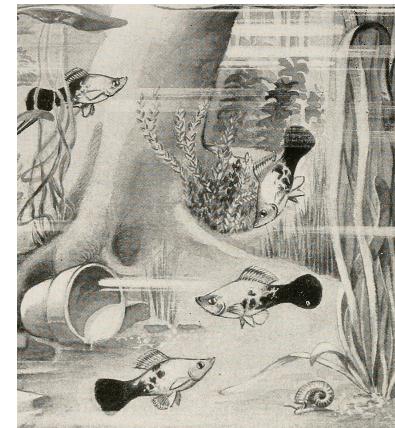


Figure 6.25: Poecilid Hybrid, *Xiphophorus helleri* × *Platypoecilus maculatus*. Aquatic life, chapter by Curtis E.S. (1915) Image from the Biodiversity Heritage Library. Contributed by Harvard University, Museum of Comparative Zoology, Ernst Mayr Library. Not in copyright.

"An ESS is a strategy such that, if all the members of a population adopt it, then no mutant strategy could invade the population under the influence of natural selection" LetterSpace=10SMITH (1982), pg 10.

A version of this sex ratio argument was first put forward by Düring in 1884 and popularized by LetterSpace=10FISHER (1930), see LetterSpace=10EDWARDS (1998).

4330 from a 50/50 sex ratio. This can represent an adaptive strategy to situations where individuals compete against relatives for access to resources or mating opportunities. to see this consider fig wasps. There are many species of fig wasp, which form a tight pollination symbiosis with many species of Fig. Wasp females enter the inverted fig flower structure, top right Figure 6.26, pollinating the flowers.

4336 They lay their eggs in some of the flowers, which form galls in response. The young, wingless, male wasps emerge from their galls first, Figure 6.27f, but they never leave the fig. Their only role in this is to fertilize the female wasps, Figure 6.27d, in the fig and then die. The female offspring, Figure 6.27a & e, emerge in the fig just as the male fig flowers are emerging, the female wasps burrow out and take the fig pollen with them as they fly off.

4342 Female wasps have control over the sex of their offspring, what's a wasp mamma to do, i.e. what is their optimal strategy? They have this degree of control as sex determination in wasps is haplo-diploid, with fertilized eggs developing as diploid females and unfertilized males; thus by choosing to lay fertilized eggs they can control their number of daughters. If a female wasp lays her eggs into a fig with no other eggs, her sons will mate with her daughters and then die. Thus a lone female can maximize her contribution to the next generation by having many daughters, and just enough sons to fertilize them. And that's exactly what female wasps do, in many species of fig wasp 95% of individuals born are female.

4354 6.1.1 Selfish genetic elements and selection below the level of the individual.

4356 *Selfish sex chromosomes and sex ratio distortion* Now from the perspective of the autosomes a 50/50 sex ratio normally represents a stable strategy, but all is not harmonious in the genome. In systems with XY sex determination, male fertilization by Y-bearing sperm leads to sons, while male fertilization by X-bearing sperm leads to daughters. From the viewpoint of the X chromosome the Y-bearing sperm, and a male's sons, are an evolutionary deadend. We can imagine a mutation arising on the X chromosome that causes a poison to be released during gametogenesis that kills Y-bearing sperm. This would cause much of the ejaculate of the males carrying this mutation to be X-bearing sperm, and so these males would have mostly daughters. Such an allele would potentially spread in the population as it is over transmitted through males, even if it somewhat reduces the fitness of the individuals who carry it. The spread of this allele would strongly bias the population sex ratio towards females. Such 'selfish' X alleles turn out to be relatively common. They do not spread because they

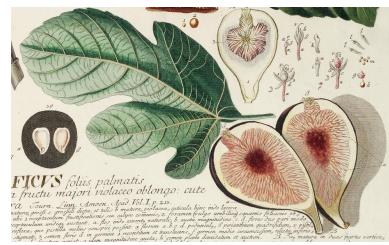


Figure 6.26: Common Fig (*Ficus carica*). Despite urban legends the crunch in figs isn't dead wasps, edible figs are dioecious and female wasps can't lay in the female flowers that form the fruit we eat.

Planta selectae quarum imagines ad exemplaria naturalia Londini, in hortis curiosorum nutrita (1750) Trew, C.J. Image from the Biodiversity Heritage Library. Contributed by Missouri Botanical Garden. Not in copyright.

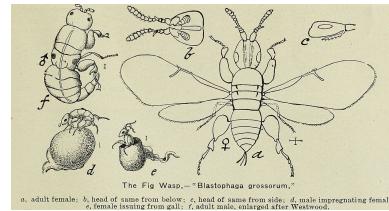


Figure 6.27: Life stages of Fig wasp (*Blastophaga psenes*, synonym *Blastophaga grossorum*); the primary pollinator of the common fig *Ficus carica*.

A descriptive catalogue of fruit and forest trees, vines and shrubs, choice palms and roses (1903) by Fancher Creek Nurseries. Image from the Biodiversity Heritage Library. Contributed by National Agricultural Library, USDA. Not in copyright.

figures/sex_ratio_distortor.pdf

Figure 6.28: The increase in frequency of a sex-ratio distorting X allele in the population of X chromosomes (solid line) and the frequency of males in the population. Males carrying the selfish X allele have 99% daughters, and the selfish X allele reduces the viability of the carries by 20% in a dominant manner. The model set up as in LetterSpace=10EDWARDS (1961), Code here.

are good for the individual, they can often substantially low the fitness of the bearer, but rather they spread because they are favoured due to selection below the level of the individual.

One example of a selfish X chromosome allele is the *Winters sex-ratio* system found in *Drosophila simulans*, so named as it was found in flies collected around Winters, California (just a few miles down the road from Davis). In a cross the selfish X chromosome carrying males have > 80% daughters. The gene responsible, Dox (*Distorter on the X*), appears to be a transposition from a parental gene, and produces a transcript which targets a region on the Y chromosome preventing the Y-bearing sperm from developing LetterSpace=10TAO *et al.* (see Figure 6.29 from 2007).

Journal_figs/single_locus_selection/Winters_sex_ratio_drive/Fer

The spread of such selfish sex chromosomes, distorting the sex ratio strongly away from 50/50, and can have profound effects for population growth rates.² However, the other sex chromosome and autosomes are not helpless against the spread of selfish sex chromosome elements. In the case of a selfish X chromosome that has achieved appreciable frequency in the population, there will be a strong excess of females in the population, such that suppressors of

Figure 6.29: Top) Normally developing spermatids in *D. simulans*. **Bottom)** Abnormally developing spermatids in Figure 6.30: Mechanistic and Evolutionary Model for sex-ratio Distortion Left) The X-linked *dox* gene evolved to target the Y chromosome, blocking Y-spermatids from becoming sperm. Right) Subsequently *Dox* was retrotransposed to an autosome forming the *Nmy* gene. *Nmy* was subsequently rearranged by a small duplication, and now blocks the action of *dox* by the formation of a hairpin small interfering RNA. Figure from LetterSpace=10FERREE and LetterSpace=10BARBASH (2007), licensed under CC BY 4.0. See LetterSpace=10BARBASH dox cartoon.png update on the fascinating biology and further loci uncovered in this system.

² Indeed people have long discussed using selfish Y chromosomes, driving an over production of sons, for population control of malaria-spreading mosquitos. Natural selfish systems on the Y appear rare, likely because of its low gene content.

drive can arise on the autosomes and spread due to the fact that they allele causes the male bearer to produce sons and so spread due to Fisherian sex-ratio advantage. This has happened in the case of the Winters sex chromosome system. An autosomal allele has spread through the population that suppresses the selfish X chromosome, restoring the 50/50 sex ratio. Now the sex ratio distorter can only be found by crosses to naive populations, where the suppressor has not spread yet. The autosomal suppressor gene turns out to be a duplicate of the selfish dox gene, *NMY* (Not Much Yang), that moved to the autosome through retrotransposition and now blocks the action of dox through RNA-interference degradation of the dox transcript (see LetterSpace=10TAO *et al.*, 2007, , see Figure 6.30).

Conflict due to maternally transmitted elements. Chromosomes transmitted maternally, i.e. only through mothers, also have divergent interests from the individual. Many plants are hermaphrodite producing both pollen and seeds. But from the perspective on the Mitochondria in an individual, pollen is a waste of energy as the Mitochondria won't be transmitted through it. Thus a mutation that arises on the Mitochondria abolishing male sexual function, pollen, and shunting energy into other processes, can spread as this selfish outcome is favourable from the perspective of the mitochondrial allele.

The self spread of a CMS allele creates a population of females and hermaphrodite plants (a gynodioecious population). This strong excess of female plants in turn can select for the spread of autosomal suppressors of CMS that are favoured by producing the rarer gamete (pollen), and so restore the population to hermaphroditism. The spread of such Cytoplasmic Male Sterility (CMS) alleles, and subsequent autosomal suppression, is thought to be common in hermaphrodite species and often uncovered in crosses between diverged hermaphrodite populations. The discovery or deliberate creation of CMS alleles in agricultural plants is prized because it gives breeders more control over hybridization as they can more carefully control the pollen donor to the plants.

The maternal transmission of mtDNA also causes genetic conflicts in organisms with separate sexes. Males are an evolutionary dead end as far as mitochondria are concerned, and so mitochondrial mutations that lower a male's fitness are not removed from the population of mitochondria. Thus the Mitochondria genome may be a hotspot of alleles that are deleterious in males (an effect termed the "Mother's curse"). One example of a male-deleterious mitochondrial mutations underlying LeberâŽs 'hereditary optic neuropathy' (LHON) in humans. LHON causes degeneration of the



Figure 6.31: Bladder Campion (*Silene vulgaris*), on left, has both hermaphrodite and female plants due to a CMS and nuclear restorer polymorphisms (LetterSpace=10CHARLESWORTH and LetterSpace=10LAPORTE, 1998). (*S. nutans* on right)

Billeder af nordens flora (1917). Mentz, A Image from the Biodiversity Heritage Library. Contributed by The LuEsther T Mertz Library, the New York Botanical Garden. Not in copyright.



Figure 6.32: Arrival of the fille du roi, the 'king's daughters' to Quebec city in 1667. Painting by Eleanor Fortescue-Brickdale. The fille du roi were some 800 women whose emigration to New France (Quebec) was paid for by a program established by King Louis XIV of France to address the strong gender imbalance of the new colony. You can read more in this Atlantic article by Sarah Zhang.

Painting from the Library and Archives Canada collection, Wikimedia, Public Domain.

4434 optic nerve and loss of vision in teenage males (with much lower penetrance in women). One such LHON mutation is present at low frequency in the Quebec population. The Québécois population grew rapidly from a relatively small number of founders, leading to the prevalence of some disease mutations due to the founder effect. Thanks to the detailed genealogical records kept by French Canadian since the founding of Quebec, we know that nearly all the Québécois LHON alleles are descended from the mitochondria of 4442 a single woman, one of the fille du roi, who arrived in Quebec City in 1669 (LetterSpace=10LABERGE et al., 2005). Using the genealogy, 4444 LetterSpace=10MILLOT et al. (2017) tracked all of her mitochondrial descendants, individuals who's mother was in her matrilineal line, and 4446 so identified all the individuals in the Québécois who carried this allele. There was no significant difference in the fitness of females who 4448 carried or didn't carry the mutation. In contrast, the fitness of male carriers of the mutation was only 65.3% that of male non-carriers. 4450 This mitochondria mutation has increased in frequency slightly over the past 290 years, despite its strong effects in males, due to the fact 4452 that its effects have no consequence for female fitness.

Question 10. The frequency of the LHON allele was roughly 4454 1/2000 in 1669. If females suffered the same ill consequences as males what would be the frequency today? [assume there are ~29 years a 4456 generation]

Question 11. Kin selection has been proposed as a way that the 4458 male deleterious effects of mitochondrial mutation could be removed from the population. Can you explain this idea?



Figure 6.33: male Eggspot butterfly (*Hypolimnas bolina*).
P. Cramer's Uitlandsche kapellen (1780) Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

4460 It's not just chromosomes that get in on the act of the battle of
 4462 the sexes. Numerous arthropods, including a high proportion of in-
 4464 sects, are infected with the intracellular bacteria *Wolbachia*, which
 4466 are passed to offspring through the maternal cytoplasm. As they are
 4468 only transmitted by females, *Wolbachia* increase their transmission in
 4470 a variety of selfish ways including feminization of males and killing
 4472 male embryos. In one dramatic case, a male-killing *Wolbachia* strain
 4474 forced a sex ratio of 100 females to every 1 male in *Hypolimnas bolina*
 (eggspot butterflies) throughout Southeast Asia. This extreme sex
 ratio persisted for many decades, according to the analysis of mu-
 seum collections from the late 19C, before the sex ratio was rapidly
 restored to 50/50 by the spread of an autosomal suppressing allele.
 The autosomal suppressor allele spread very rapidly within popula-
 tions taking just 5 years to spread through the population from 2001
 to 2006.

4476 *Autosomal selfish systems* Self genetic systems can also arise and
 cause genetic conflicts on the autosomes. The interests of autosomal
 4478 alleles are usually relatively well aligned with promoting the fitness.
 However, these interests can diverge during meiosis and gametogen-
 4480 esis. After all, There are two alleles at each autosomal locus but only
 one of them will get passed to a child, therefore there can be com-
 petition between alleles in an individual to be transmitted gamete to the
 4482 next generation (Figure 6.34).

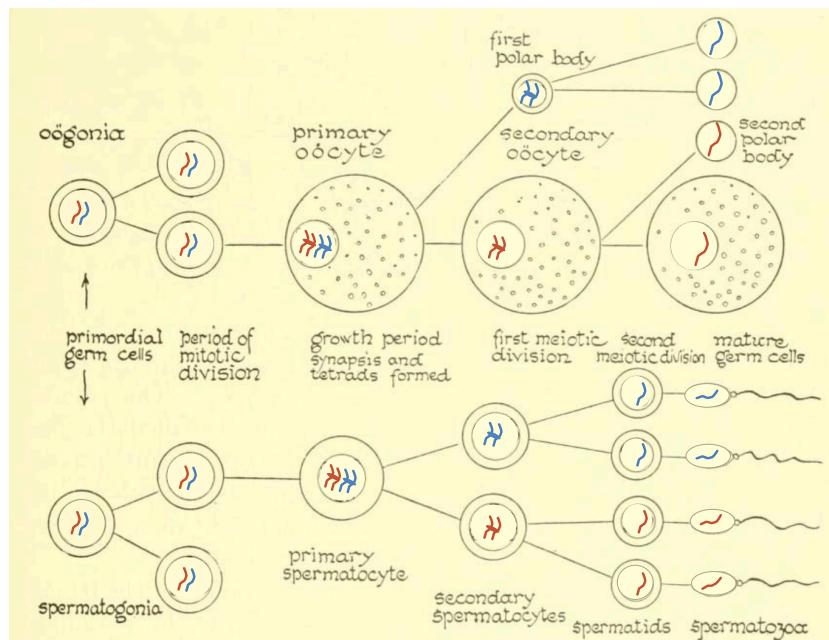


Figure 6.34: The two copies of a chromosome are shown in red and blue through the process of female and male meiosis and gametogenesis. Crossovers are omitted to keep things simpler, they should occur between at least one red and one blue chromosome. Modified from original to include chromosomes transmitted.

Biology; the story of living things (1937). Hunter, G.W., Walter H.E. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. No known copyright restrictions.

Male and female gametogenesis offer different opportunities for selfish systems. Just as how selfish X chromosome systems can spread by targeting sperm that carry the Y chromosome, selfish autosomal alleles can spread by targeting sperm carrying the other chromosome in heterozygotes. Both the Segregation Distortion allele and T-allele s in Drosophila and mouse respectively are selfish autosomal systems that game transmission in heterozygotes by killing off sperm that don't carry the allele in heterozygotes.

In females meiosis there is a unique opportunity for cheating. In male meiosis all four products of meiosis become gametes. However, only 1 of the four products of female meiosis becomes the egg, the other 3 products are fated to become the polar bodies. Thus alleles can cheat in female meiosis by preferentially getting transmitted into the egg rather than the polar body. If an allele on red chromosome can manipulate any asymmetry of meioses so that it can be present in the egg > 50% of the time it will be a transmission advantage in female heterozygotes.

To see how such drivers can spread through the population lets consider the case of a population where an allele drives in both male and female gametogenesis. (Most selfish alleles will be sex-specific, but that makes the math a little more tricky.) Imagine a randomly-mating population of hermaphrodites. In this population, a derived allele (D) segregates that distorts transmission in its favour over the ancestral allele (d) in the production of all the gametes of heterozygotes. The drive leads to a fraction α of the gametes of heterozygotes (D/d) to carry the D allele ($\alpha \geq 0.5$). The D allele causes viability problems such that the relative fitnesses are $w_{dd} = 1, 1 > w_{Dd} \geq w_{DD}$. If the D allele is currently at frequency p in the population at birth, its frequency at birth in the next generation will be

$$p' = \frac{w_{DD}p^2 + w_{Dd}\alpha 2pq}{\bar{w}} \quad (6.46)$$

when $\alpha = 1/2$, i.e. fair Mendelian transmission this is exactly the same as our directional selection, which results in our D allele being selected out of the population (blue line, Figure 6.35). However, if $\alpha > 1/2$, i.e. our deleterious allele cheats, it can potentially increase in the population when it is rare (red and black lines, Figure 6.35)). However, the allele can become trapped in the population at a polymorphic equilibrium if its cost is sufficient in homozygotes. This is akin to the case of heterozygote advantage, but now our allele offers no advantage to heterozygote but has a self advantage in heterozygotes.

Question 12. (Tricker question) Thinking of our autosomal driver from equation 6.46. **B)** Imagine the cost of the driver were additive,

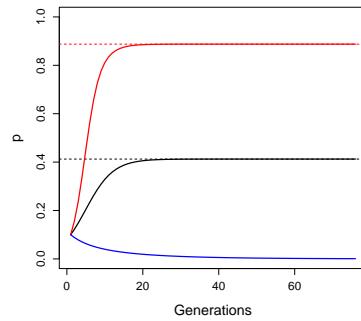


Figure 6.35:

4524 i.e. $w_{dd} = 1$, $w_{Dd} = 1 - e$, $w_{DD} = 1 - 2e$. Under what conditions can
 4526 the driver invade the population? Can a polymorphic equilibrium be
 maintained?

4528 **A)** Imagine the allele is completely recessive, i.e. $w_{dd} = w_{Dd} = 1$.
 4530 What conditions do you need for a polymorphic equilibrium to be
 maintained? What is the equilibrium frequency of this balanced
 polymorphism?

4532 Many of the known autosomal drive systems are polymorphic in
 4534 populations, unable to reach fixation in the population due to their
 costs in homozygotes. It seems like that this represents an ascertainment bias, and that many over selfish systems, which had lower
 selective costs, have swept to fixation.

4536 **6.1.2 Appendix: ESS for the sex ratio**

4538 Let R be the sources and C_σ and C_φ be the cost of producing a son
 4540 and daughter respectively. If our focal mother directs s of her effort
 towards sons and $(1 - s)$ of her effort towards daughters, she'll pro-
 duces $\frac{Rs}{C_\sigma}$ sons and $\frac{R(1-s)}{C_\varphi}$ daughters. We will assume that the mean
 reproductive value of daughters is 1. Given this the reproductive
 4542 value of sons is the average number of matings that a male will have,
 i.e. the ratio # females/# males. So if the population has a sex ratio s_p , the
 4544 fitness of our focal female is

$$W(s, s_p) = \left(\frac{R(1-s)}{C_\varphi} \times 1 \right) + \left(\frac{Rs}{C_\sigma} \times \frac{R(1-s_p)/C_\varphi}{Rs_p/C_\sigma} \right) \quad (6.47)$$

4546 expressing fitness in terms the number of grandkids our focal female
 is expected to have.

4548 To find the ESS we want a sex ratio s^* for the population that no
 mutant has higher fitness, i.e. $W(s^*, s^*) > W(s, s^*)$ for $s \neq s^*$. We can
 find this by

$$\frac{\partial W(s, s_p)}{\partial s} \Big|_{s^*=s=s_p} = 0 \quad (6.48)$$

4550 taking the derivative of Eqn 6.47 we obtain

$$\frac{\partial W(s, s_p)}{\partial s} = -\frac{R}{C_\varphi} + \frac{R}{C_\sigma} \left(\frac{R(1-s_p)/C_\varphi}{Rs_p/C_\sigma} \right) \quad (6.49)$$

setting $s^* = s = s_p$ and rearranging

$$\frac{R}{C_\varphi} = \frac{R}{C_\sigma} \left(\frac{R(1-s^*)/C_\varphi}{Rs^*/C_\sigma} \right) \quad (6.50)$$

4552 which is satisfied when $s^* = 1/2$, i.e. devoting equal resources to male
 and female offspring is the ESS, which corresponds to a 50/50 sex
 4554 ratio if male and female offspring are equally costly.

7

4556 *The interaction of Selection, Mutation, and Migration.*

7.0.1 Mutation-selection balance

4558 Mutation is constantly introducing new alleles into the population.
Therefore, variation can be maintained within a population not only
4560 if selection is balancing (e.g. through heterozygote advantage or fluctuating selection over time, as we have seen in the previous section),
4562 but also due to a balance between mutation introducing deleterious alleles and selection acting to purge these alleles from the population
4564 (LetterSpace=10HALDANE, 1937). To study mutation-selection balance,
4566 we return to the model of directional selection, where allele A_1 is advantageous, i.e.

genotype	A_1A_1	A_1A_2	A_2A_2
absolute fitness	W_{11}	$\geq W_{12} \geq$	W_{22}
relative fitness	$w_{11} = 1$	$w_{12} = 1 - sh$	$w_{22} = 1 - s.$

4568 We'll begin by considering the case where allele A_2 is not completely recessive ($h > 0$), so that the heterozygotes suffer at least some disadvantage. We denote by $\mu = \mu_{1 \rightarrow 2}$ the mutation rate per generation from A_1 to the deleterious allele A_2 , and assume that there is no reverse mutation ($\mu_{2 \rightarrow 1} = 0$). Let us assume that selection against A_2 is relatively strong compared to the mutation rate, so that it is justified
4572 to assume that A_2 is always rare, i.e. $q_t = 1 - p_t \ll 1$. Compared to previous sections, for mathematical clarity, we also switch from
4574 following the frequency p_t of A_1 to following the frequency q_t of A_2 . Of course, this is without loss of generality. The change in frequency
4576 of A_2 due to selection can be written as
4578

$$\Delta_S q_t = \frac{\bar{w}_2 - \bar{w}_1}{\bar{w}} p_t q_t \approx -hsq_t. \quad (7.1)$$

This approximation can be found by assuming that $q^2 \approx 0$, $p \approx 1$,
4580 and that $\bar{w} \approx w_1$. All of these assumptions make sense if $q \ll 1$. From eqn. (7.1) we see that selection acts to reduce the frequency of
4582 A_2 (as both h and s are positive), and it does so geometrically across

the generations. That is, if the initial frequency of A_2 is q_0 , then its frequency at time t is approximately

$$q_t = q_0(1 - hs)^t. \quad (7.2)$$

We will now consider the change in frequency induced by mutation. Recalling that μ is the mutation rate from A_1 to A_2 per generation, the frequency of A_2 after mutation is

$$q' = \mu p_t + q_t = \mu(1 - q_t) + q_t. \quad (7.3)$$

Assuming that $\mu \ll 1$ and that $q \ll 1$, the change in the frequency of allele A_2 due to mutation ($\Delta_M q_t$) can be approximated by

$$\Delta_M q_t = q' - q_t = \mu. \quad (7.4)$$

Hence, when A_2 is rare and the mutation rate is low, mutation acts to linearly increase the frequency of the deleterious allele A_2 .

If selection is to balance deleterious mutation, their combined effect over one generation has to be zero. Therefore, to find the mutation-selection equilibrium, we set

$$\Delta_M q_t + \Delta_S q_t = 0, \quad (7.5)$$

insert eqns. (7.1) and (7.4), and solve for q to obtain

$$q_e = q_t = \frac{\mu}{hs}. \quad (7.6)$$

We see that the frequency of the deleterious allele A_2 is balanced at a frequency equal to the mutation rate (μ) divided by the reduction in relative fitness in the heterozygote (hs).

It is worth pointing out that the fitness of the $A_2 A_2$ homozygote has not entered this calculation, as A_2 is so rare that it is hardly ever found in the homozygous state. Therefore, if A_2 has any deleterious effect in a heterozygous state (i.e. if $h > 0$), it is this effect that determines the frequency at which A_2 is maintained in the population. Also, note that by writing the total change in allele frequency as $\Delta_M q_t + \Delta_S q_t$ we have implicitly assumed that we can ignore terms of order $\mu \times s$. That is, we have assumed that mutation and selection are both relatively weak. This assumption is valid under our prior assumption that both μ and s are small.

If an allele is truly recessive (although few likely are), we have $h = 0$, and so eqn. (7.6) is not valid. However, we can make an argument similar to the one above to show that, for truly recessive alleles,

$$q_e = \sqrt{\frac{\mu}{s}}. \quad (7.7)$$



Figure 7.1: Oblong-winged katydid.
Field book of insects (1918). Lutz, F.E. Illustrations by Edna L. Beutemiller. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

4612 **Question 1.** Oblong-winged katydids (*Amblycorypha oblongifolia*)
 4614 are usually green. However, some are bright pink, thanks to an ery-
 4616 thrism mutation (a nice example of early Mendelian reasoning in a
 4618 wonderfully titled paper¹). This pink condition is thought to be due
 to a dominant mutation (Crew, 2013). Assume that roughly one in
 ten thousand katydids is bright pink and that the mutation rate at the
 gene underlying this condition is 10^{-5} . What is the relative fitness of
 heterozygotes for the pink mutation?

¹ LetterSpace=10WHEELER, W. M., 1907
 Pink Insect Mutants. The American
 Naturalist 41(492): 773–780

4620 *The genetic load of deleterious alleles* What effect do such deleterious
 mutations at mutation–selection balance have on the population?
 4622 It is common to quantify the effect of deleterious alleles in terms
 of a reduction of the mean relative fitness of the population. For a
 4624 single site at which a deleterious mutation is segregating at frequency
 $q_e = \mu/(hs)$, the population mean relative fitness is reduced to

$$\bar{w} = 1 - 2p_e q_e hs - q_e^2 s \approx 1 - 2\mu. \quad (7.8)$$

4626 Somewhat remarkably, the drop in mean fitness due to a site segre-
 gating at mutation–selection balance is independent of the selection
 4628 coefficient against the heterozygote; it depends only on the mutation
 rate. Intuitively this is because, given a fixed mutation rate, less dele-
 4630 terious alleles can rise to a higher equilibrium frequency, and thus
 contribute the same total load as more deleterious (rarer) alleles, but
 4632 this load is spread across more individuals in the population. Note
 that this result applies only if the mutation is not totally recessive, i.e.
 4634 if $h > 0$.

4636 A fitness reduction of 2μ is very small, given that the mutation
 rate of a gene is likely $< 10^{-5}$. However, if there are many loci seg-
 4638 regating at mutation–selection balance, small fitness reductions can
 accumulate to a substantial so-called genetic load, a major cause of
 variation in fitness-related traits among individuals. For example,
 4640 the human genome contains over twenty thousand genes, and many
 other functional regions, the vast majority of which will be subject
 4642 to purifying selection against mutations that disrupt their function.
 In humans, most loss of function (LOF) variants, which severely dis-
 4644 rupt a protein-coding gene, are found at low frequencies. However,
 each human genome typically carries over a hundred LOF variants
 4646 (LetterSpace=10MACARTHUR *et al.*, 2012; LetterSpace=10LEK *et al.*,
 2016). Not every LOF allele will be deleterious; some could even be
 4648 advantageous. However, the combined load of these LOF alleles must
 on average lower our fitness, otherwise selection wouldn't be remov-
 4650 ing them from the population. Each one of us carries a unique set of
 these LOF alleles, usually in a heterozygous state. We differ slightly

4652 in how many of these alleles we carry. For example, the left side of
 4653 Figure 7.2 shows the distribution of the number of LOF alleles car-
 4654 ried by 769 individuals of Dutch ancestry. The individuals who carry
 4655 fewer of these LOF alleles will on average have higher fitness than
 4656 those individuals with more.

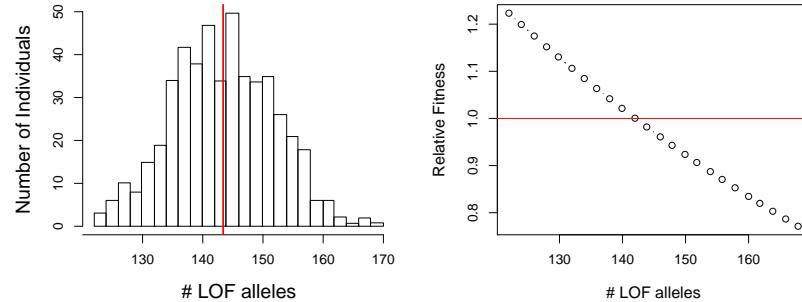


Figure 7.2: **Left)** The distribution of LOF alleles in 769 individuals from the Genome of the Netherlands project. Data from LetterSpace=10FRANCIOLI *et al.*. The average individual (red line) carries 144 LOF alleles. **Right).** The relative fitness of individuals carrying these varying numbers of LOF alleles, assuming multiplicative selection and a selection coefficient of $sh = 10^{-2}$ acting against these alleles (LetterSpace=10CASSA *et al.*, 2017). Code here.

4658 How do these differences across individuals in total LOF muta-
 4659 tions mount up? Well, if we are willing to assume that the fitness
 4660 costs of deleterious alleles interact multiplicatively, we can make
 4661 some progress. If an individual who carries one LOF mutation has
 4662 a fitness $1 - hs$, then an individual who's heterozygote for two LOF
 4663 mutations would have fitness $(1 - hs)^2$, and an individual who is
 4664 heterozygote for L LOF alleles would have fitness $(1 - hs)^L$. The
 4665 right-hand side of Figure 7.2 shows the predicted fitness of individ-
 4666 uals carrying varying number of LOF alleles, relative to the mean
 4667 fitness of the sample, using this multiplicative model. We don't yet
 4668 know how much lower the fitness of these individuals really is, nor
 4669 do we know how most of these LOF alleles manifest their fitness con-
 4670 sequences through disease and other mechanisms. However, it's a
 4671 reasonable guess that this variation in LOF alleles, presumably main-
 4672 tained by mutation-selection balance, is a major source of variation in
 4673 fitness.

7.0.2 Inbreeding depression

4674 All else being equal, eqn. (7.6) suggests that mutations that have a
 4675 smaller effect in the heterozygote can segregate at higher frequency
 4676 under mutation-selection balance. As a consequence, alleles that
 4677 have strongly deleterious effects in the homozygous state can still
 4678 segregate at low frequencies in the population, as long as they do not
 4679 have too strong a deleterious effect in heterozygotes. Thus, outbred
 4680 populations may have many alleles with recessive deleterious effects

segregating within them.

4682 **Question 2.** Assume that a deleterious allele has a relative fitness .99 in heterozygotes and a relative fitness 0.2 when present in the homozygote state. Assume that the deleterious allele is at a frequency 10^{-3} at birth and the genotype frequencies follow from HWE. Only 4684 considering the fitness effects of this locus, and measuring fitness 4686 relative to the most fit genotype, answer the following questions:

- 4688 A) What is the average fitness of an individual in the population?
B) What is the average fitness of the child of a full-sib mating?

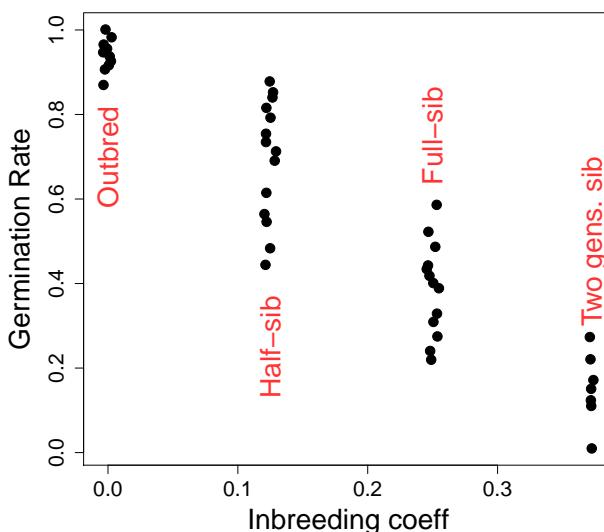


Figure 7.3: Data showing inbreeding depression over different degrees of inbreeding in *S. latifolia*. Each point is the mean seed germination rates for different family crosses. Data from LetterSpace=10RICHARDS. Code here.

4690 One consequence of segregating for low-frequency recessive deleterious alleles is that inbreeding can reduce fitness. In typically outbred populations, the mean fitness of individuals decreases with the inbreeding coefficient, i.e. so-called ‘inbreeding depression’ is a common observation. This wide-spread observation dates back to systematic surveys of inbreeding depression by LetterSpace=10DARWIN 4694 (1876). Inbreeding depression is likely primarily a consequence of being homozygous at many loci for alleles with recessive deleterious effects.

4700 One example of inbreeding depression is shown in Figure 7.3. White campion (*Silene latifolia*) is a dioecious flowering plant; dioecious means that the males and females are separate individuals. 4702 LetterSpace=10RICHARDS performed crosses to create offspring who were outbred, the offspring of half-sibs, full-sibs, and of two generations of full-sib mating. He measured their germination success, which is plotted in Figure 7.3. Note how the fitness of individuals declines with increased inbreeding.

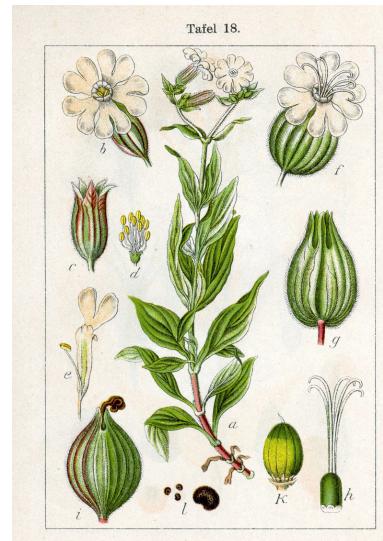


Figure 7.4: White campion (*S. latifolia*). Deutslands Flora in Abbildungen (1796). Johann Georg Sturm (Painter: Jacob Sturm). Public Domain, wikipedia.

Purging the inbreeding load. Populations that regularly inbreed over sustained periods of time are expected to partially purge this load of deleterious alleles. This is because such populations have exposed many of these alleles in a homozygous state, and so selection can more readily remove these alleles from the population.

If the population has sustained inbreeding, such that individuals in the population have an inbreeding coefficient F , deleterious alleles at each locus will find a new equilibrium frequency. Assuming the mutation-selection model, now with inbreeding, the equilibrium frequency is

$$q_e = \frac{\mu}{(h(1 - F) + F)s} \quad (7.9)$$

The frequency of the deleterious allele is decreased due to the allele now being expressed in homozygotes, and therefore exposed to selection, more often due to inbreeding. Thus, all else being equal, populations with a high degree of inbreeding will purge their load.

7.0.3 Migration-selection balance

Another reason for the persistence of deleterious alleles in a population is that there is a constant influx of maladaptive alleles from other populations where these alleles are locally adaptive. Migration-selection balance seems unlikely to be as broad an explanation for the persistence of deleterious alleles genome-wide as mutation-selection balance. However, a brief discussion of such alleles is worthwhile, as it helps to inform our ideas about local adaptation.

Local adaptation can occur over a range of geographic scales. Local adaptation is relatively unimpeded by migration at broad geographically scales, where selection pressures change more slowly than distances over which individuals typically migrate over a number of generations. Adaptation can, however, potentially occur on much finer geographic scales, from kilometers down to meters in some species. On such small scales, dispersal is surely rapidly moving alleles between environments, but local adaptation is maintained by the continued action of selection. An example of adaptation at fine-scales is shown in Figure 7.6 . LetterSpace=10JAIN and LetterSpace=10BRADSHAW (1966) studied the patterns of heavy-metal resistance in plants on mine tailings and in nearby meadows, a set of classic studies of population differences maintained by local adaptation to different soils. Even at these very short geographically scales, over which seed and pollen will definitely move, we see strong local adaptation. Zinc-intolerant alleles are nearly absent from the mine tailings because they prevent plants from growing on these zinc-heavy soils; conversely, zinc-tolerant alleles do not spread into



Figure 7.5: Sweet vernal grass (*Anthoxanthum odoratum*).

Billeder af nordens flora (1917). Mertz, A & Ostenfeld, C H. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden. Not in copyright.

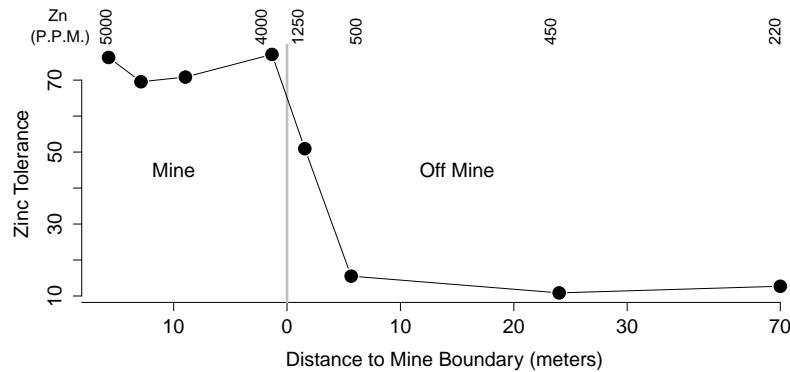


Figure 7.6: Data showing the Zinc tolerance of *Anthoxanthum odoratum* on and off of the Trelogan Mine, Flintshire, North Wales. The numbers along the top give the soil contamination of Zinc in parts per million. Data from LetterSpace=10JAIN and LetterSpace=10BRADSHAW (1966). Code here.

the meadow populations, likely due to some trade-off or fitness cost of zinc-tolerance.

As a first pass at developing a model of local adaptation, let's consider a haploid two-allele model with two different populations, see Figure 7.7, where the relative fitnesses of our alleles are as follows

allele	1	2
population 1	1	$1-s$
population 2	$1-s$	1

As a simple model of migration, let's suppose within a population a fraction of m individuals are migrants from the other population, and $1-m$ individuals are from the same population.

To quickly sketch an equilibrium solution to this scenario, we'll take an approach analogous to our mutation-selection balance model. To do this, let's assume that selection is strong compared to migration ($s \gg m$), such that allele 1 will be almost fixed in population 1 and allele 2 will be almost fixed in population 2. If that is the case, migration changes the frequency of allele 2 in population 1 (q_1) by

$$\Delta_{Mig.} q_1 \approx m \quad (7.10)$$

while as noted above $\Delta_S q_1 = -sq_1$, so that migration and selection are at an equilibrium when $0 = \Delta_S q_1 + \Delta_{Mig.} q_1$, i.e. an equilibrium frequency of allele 2 in population 1 of

$$q_{e,1} = \frac{m}{s} \quad (7.11)$$

Here, migration is playing the role of mutation and so migration-selection balance (at least under strong selection) is analogous to mutation-selection balance.

We can use this same model by analogy for the case of migration-selection balance in a diploid model. For the diploid case, we replace

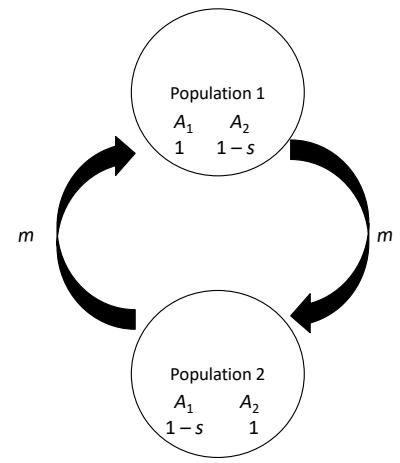


Figure 7.7: Setup of a two-population haploid model of local adaptation.

4770 our haploid s by the cost to heterozygotes hs from our directional
 selection model, resulting in a diploid migration-selection balance
 4772 equilibrium frequency of

$$q_{e,1} = \frac{m}{hs} \quad (7.12)$$

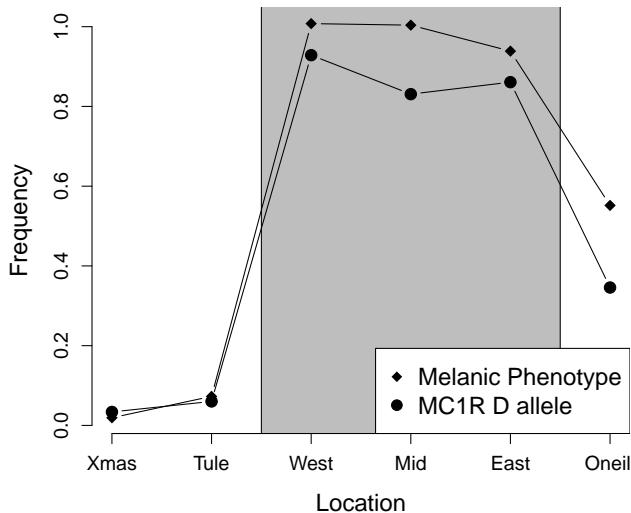


Figure 7.8: Frequency of melanic mice on the lava flow, and at nearby locations (diamonds). Frequency of MC1R melanic allele at same locations. Data from LetterSpace=10HOEKSTRA *et al.* (2004). Code here.

4774 As an example of fine-scale local adaptation due to a single locus,
 consider the case of the rock pocket mice adapting to lava flows.
 Throughout the deserts of the American Southwest there are old lava
 4776 flows, where the rocks and soils are much dark than the surrounding
 desert.

4778 Many populations of small animals that live on these flows have
 evolved darker pigmentation to be cryptic against this dark substrate
 4780 and better avoid visual predators. One example of such a locally
 adapted population are the rock pocket mice (*Chaetodipus intermedius*)
 4782 who live on the Pinacate lava flow on the Arizona-Mexico border,
 studied by LetterSpace=10HOEKSTRA *et al.* (2004). These mice have
 4784 much darker, more melanic pelts than the mice who live on nearby
 rocky outcrops (see Figure 7.8). LetterSpace=10NACHMAN *et al.* (2003)
 4786 determined that a dominant allele (D) at MC1R is the primary de-
 terminant of this melanic phenotype. The frequency of this allele
 4788 across study sites is shown in Figure 7.8. LetterSpace=10HOEKSTRA
et al. (2004) found that other, unlinked markers showed little differen-
 4790 tiation over these populations, suggesting that the migration rate is
 high.

4792 **Question 3.** LetterSpace=10HOEKSTRA *et al.* (2004) found that



Figure 7.9: Two species from the genus *Chaetodipus*, pocket mice, formally known as *Perognathus*.

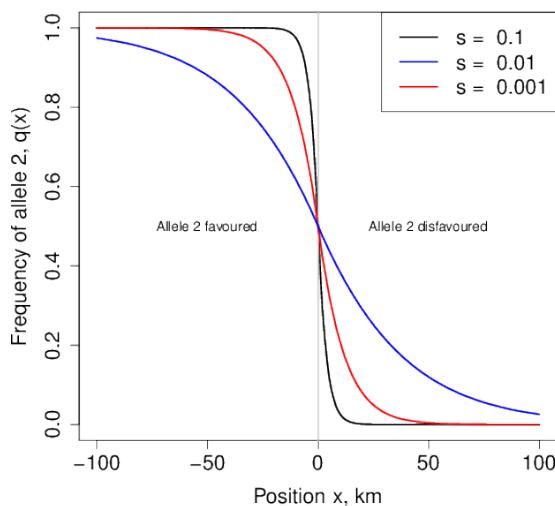
Wild animals of North America, intimate studies of big and little creatures of the mammal kingdom (1918), Nelson, E. W. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

the dark D allele was at 3% frequency at the Tule Mountains study site. Using F_{ST} -based approaches, for unlinked markers, they estimated that the per individual migration rate was $m = 7.0 \times 10^{-4}$ per generation between this site and the Pinacate lava flow. What is the selection coefficient acting against the dark D allele at the Tule Mountains site?

The width of a genetic cline. We can also extend these ideas beyond our discrete model to a model of a population spread out on a landscape where individuals migrate in a more continuous fashion. For simplicity, let's assume a one dimensional habitat, where the habitat makes a sharp transition in the middle of our region. You could imagine this to be a set of populations sampled along a transect through some environmental transition. Our individuals disperse to live on average σ miles away from where they were born (we can think of this as our individuals migrating a random distance drawn from a normal distribution, with mean zero, and σ being the standard deviation of this distribution). . We'll think of a bi-allelic model where the homozygotes for allele 1 have an additive selective advantage s over allele 2 homozygotes to the east of our habitat transition (left of zero in Figure 7.10). This flips to allele 2 having the same advantage s west of the transition (right of zero). If you've read this send Prof Coop a picture of the East and West Beast.

"Upon an island hard to reach, the East Beast sits upon his beach. Upon the west beach sits the West Beast. Each beach beast thinks he's the best beast."
 – Theodor Seuss Geisel

Figure 7.10: An equilibrium cline in allele frequency (the frequency of allele 2, $q(x)$) is shown. Our individuals disperse an average distance of $\sigma = 1$ miles per generation, and our allele 2 has a relative fitness of $1 + s$ and $1 - s$ on either side of the environmental change at $x = 0$. Code here.



With this setup, we get an equilibrium distribution of our two alleles, where to the left of zero our allele 2 is at higher frequency, while

to the right of zero allele 1 predominates. As we cross from the left to the right side of our range, the frequency of our allele 2 decreases in a smooth cline. The frequency of allele 2, $q(x)$, is shown as a function of location along the cline for a variety of selection coefficients (s) in Figure 7.10. The width of this cline, i.e. the geographic distance over which the allele frequency changes, depends on the relative strengths of dispersal and selection. If selection is strong compared to dispersal, then selection acts to remove maladaptive alleles much faster than migration acts to move alleles across the environmental transition. Thus the allele frequency transition would be very rapid, and the cline narrow, as we move across the environmental transition. In contrast, if individuals disperse long distances and selection is weak, many alleles are being moved back and forth over the environmental transition much faster than selection can act against these alleles and so the cline would be very wide.

The width of our cline, i.e. the distance over which we make this shift from allele 2 to allele 1 predominating, can be defined in a number of different ways. One way to define the cline width, which is simple to define but perhaps hard to measure accurately, is via the slope (i.e. the tangent) of $q(x)$ at $x = 0$. See Figure 7.11. Under this definition, the cline width is approximately

$$0.6\sigma/\sqrt{s} \text{ miles}, \quad (7.13)$$

note that the units are miles here just because we defined the average dispersal distance (σ) in miles above. Thus the cline will be wider if individuals disperse further, higher σ , and if selection is weaker, smaller s . The appendix below talks through the math underlying these ideas in more detail.

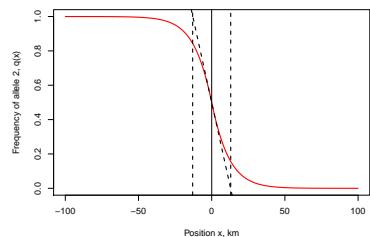


Figure 7.11: An equilibrium cline in allele frequency from Figure 7.10, $s = 0.01$. Vertical lines show the cline width. The diagonal line shows the tangent to the cline at its midpoint. Code here.

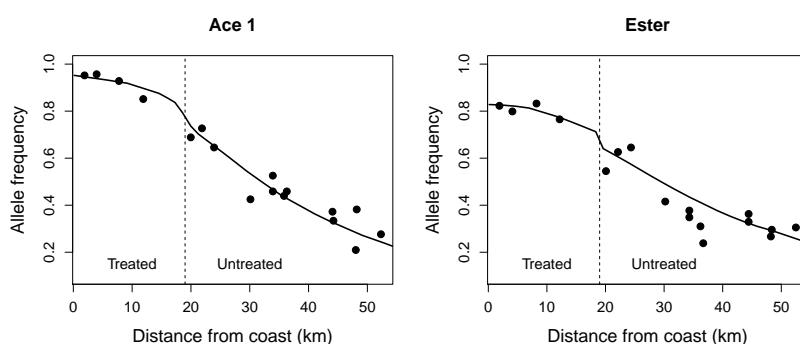


Figure 7.12: Allele frequency clines of two pesticide resistance alleles, at the Ace 1 and Ester genes, in the mosquito *Culex pipiens*. The dotted line shows where we move from pesticide-treated to untreated areas as we move away from the French coast. The dots show observed allele frequencies, the solid lines clines fit under a migration-selection balance model of a cline. These allele frequencies represent collections over two summers, the frequencies of the alleles are substantially reduced in the winter due to the reduced use of pesticides. Data from LetterSpace=10LENORMAND *et al.* (1999). Code here.

LetterSpace=10LENORMAND *et al.* (1999) collected mosquitoes (*Culex pipiens*) in a north-south transect moving away from the Southern French coast. Areas near the coast were treated with

pesticides, and the mosquitos have evolved resistance, but areas just a few tens of kilometers from the coast were untreated. LetterSpace=10 LENORMAND *et al.* estimated the frequency of two un-linked, pesticide-resistance alleles, and found them at high frequency near the coast but found that their frequencies declined rapidly moving inland. LetterSpace=10 LENORMAND *et al.* fit migration-selection cline models to their data, similar to those in Figure 7.10, with the pesticide-resistance alleles having an selection advantage (s) in treated areas an a cost (c) in untreated areas (they didn't enforce the selective advantage and cost being symmetric).

They estimated that a higher selective advantage for the Ace 1 allele than Ester allele ($s = 0.33$ and $s = 0.19$ respectively) and a higher cost to the Ace 1 allele than Ester allele in untreated areas ($c = 0.11$ and $c = 0.07$ respectively) potentially explaining the less extreme cline for Ester allele than the Ace 1 allele. Despite these strong selection pressures we still see a cline over tens of kilometers because dispersal is relatively high ($\sigma = 6.6\text{km per generation}$).

Hybrid zones Local adaptation isn't the only way that selection can generate strong spatial patterns. We can also see strong selection-driven clines when partially-reproductively isolated species spread back in to secondary contact they can hybridize bringing alleles together that may not work well with each other. One simple model of is to think about an under-dominant polymorphism, i.e. where the heterozygote has lower fitness. The two ancestral populations are alternatively fixed for the two fitter homozygote states, e.g. ancestral population 1 fixed A_1A_1 and ancestral population two the A_2A_2 . The hybrid population forming at the mating edge between the two ancestral populations has a high frequency of the less fit heterozygotes. Thus hybrids are at a disadvantage, potentially acting to keep the two populations from collapsing into each other.

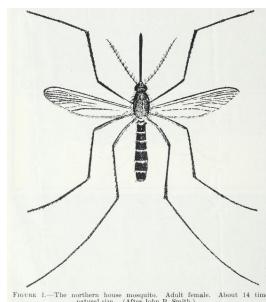


FIGURE 1.—The northern house mosquito. Adult female. About 14 times natural size. (After John B. Smith.)

Figure 7.13: mosquito (*Culex pipiens*). Domestic mosquitoes (1939). Bishopp, F. C. Image from the Biodiversity Heritage Library. Contributed by U.S. Department of Agriculture, National Agricultural Library. Not in copyright.

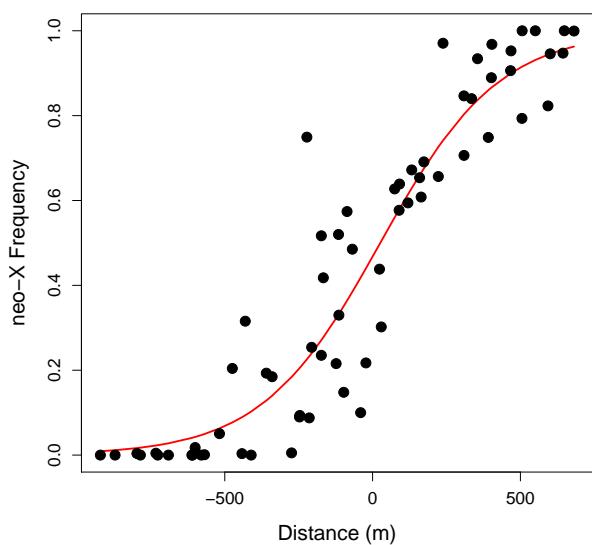


Figure 7.14: The frequency of the southern neo-X chromosome moving along a valley transect (more southern locations to the right of the graph). This represents data from four different valleys in the French Alps over less than a kilometer, each point represents a sample of 20 males. The red curve is the fitted cline under a model of heterozygote disadvantage (LetterSpace=10BAZYKIN, 1969). Data from LetterSpace=10BARTON and LetterSpace=10HEWITT (1981), Code here.

4876 Two previously isolated populations of the short-horned grasshopper *Podisma pedestris* have spread into secondary contact in the French
 4878 Alps, probably after the last ice age. The population that has spread into the Alps from the south has a large section of novel X chromo-
 4880 some, due to a chromosomal fusion. This 'neo-X' is absent in the populations that spread from the North into the Alps. The two pop-
 4882ulations meet in many valleys running through the Alps, and repeated form a narrow hybrid zone, with the frequency of the neo-X chromo-
 4884 some forming a very steep cline transitioning in frequency over a few hundred meters (LetterSpace=10BARTON and LetterSpace=10HEWITT,
 4886 1981). One potential reason for this steep cline is that females who are heterozygous for the neo-X (neo-X/old-X) may have reduced fit-
 4888 ness, consistent with an underdominant polymorphism. The neo-X allele cannot spread into the northern population as it cannot in-
 4890 crease in frequency when rare. Conversely the northern population cannot displace the neo-X, as the old-X is at a disadvantage. This
 4892 spatial distribution at this locus is a tension zone between the two populations, where neither population can make ground on the other due to the low fitness of the hybrid.

4894 We can use our same continuous model of migration and selection to study this setup. Assuming that the homozygotes are equally fit, and that the heterozygotes relative fitness is reduced by a selection

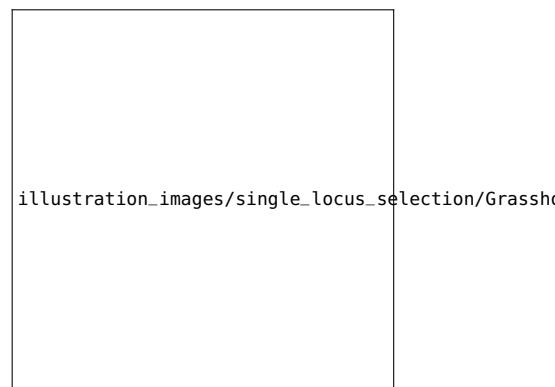


Figure 7.15: 7. *Podisma pedestris*, a species of short-horned grasshoppers; from a page illustrating *Orthoptera*. Illustration from Brockhaus and Efron Encyclopedic Dictionary (1890) Image wikipedia, public domain.

4898 coefficient s_h , the width of the cline is

$$\frac{\sigma}{\sqrt{s_h}} \quad (7.14)$$

The stronger the selection the more abrupt the transition between the 4900 populations. These wingless grasshoppers move $\sigma \sim 20$ meters a generation. Thus a reduction in the relative fitness of the hybrid would 4902 be needed to explain this hybrid zone with a width of ~ 800 m.

More generally we can see tension zones arise when hybrids have 4904 reduced fitness compared to either species. For example, this can occur due to be due to bad epistatic interactions between alleles from 4906 each species. If selection is strong enough on hybrids, often because many loci are involved in incompatibilities between the species, the 4908 entire genome can be tied up in a tension zone between the two species.

4910 *Appendix: Some theory of the spatial distribution of allele frequencies under deterministic models of selection*

4912 Imagine a continuous haploid population spread out along a line. Each individual disperses a random distance Δx from its birthplace 4914 to the location where it reproduces, where Δx is drawn from the probability density $g(\cdot)$. To make life simple, we will assume that 4916 $g(\Delta x)$ is normally distributed with mean zero and standard deviation σ , i.e. migration is unbiased and individuals migrate an average 4918 distance of σ .

The frequency of allele 2 at time t in the population at spatial location x is $q(x, t)$. Assuming that only dispersal occurs, how does 4920 our allele frequency change in the next generation? Our allele frequency in the next generation at location x reflects the migration 4922 from different locations in the proceeding generation. Our population at location x receives a contribution $g(\Delta x)q(x + \Delta x, t)$ of allele 4924 2 from the population at location $x + \Delta x$, such that the frequency of 4926 our allele at x in the next generation is

$$q(x, t + 1) = \int_{-\infty}^{\infty} g(\Delta x)q(x + \Delta x, t)d\Delta x. \quad (7.15)$$

To obtain $q(x + \Delta x, t)$, let's take a Taylor series expansion of $q(x, t)$:

$$q(x + \Delta x, t) = q(x, t) + \Delta x \frac{dq(x, t)}{dx} + \frac{1}{2}(\Delta x)^2 \frac{d^2q(x, t)}{dx^2} + \dots \quad (7.16)$$

4928 then

$$q(x, t + 1) = q(x, t) + \left(\int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x \right) \frac{dq(x, t)}{dx} + \frac{1}{2} \left(\int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x \right) \frac{d^2q(x, t)}{dx^2} + \dots \quad (7.17)$$

Because $g(\)$ has a mean of zero, $\int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x = 0$, and has because $g(\)$ has variance σ^2 , $\int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x = \sigma^2$. All higher order terms in our Taylor series expansion cancel out (as all high moments of the normal distribution are zero). Looking at the change in allele frequency, $\Delta q(x, t) = q(x, t + 1) - q(x, t)$, so

$$\Delta q(x, t) = \frac{\sigma^2}{2} \frac{d^2 q(x, t)}{dx^2} \quad (7.18)$$

This is a diffusion equation, so that migration is acting to smooth out allele frequency differences with a diffusion constant of $\frac{\sigma^2}{2}$. This is exactly analogous to the equation describing how a gas diffuses out to equal density, as both particles in a gas and our individuals of type 2 are performing Brownian motion (blurring our eyes and seeing time as continuous).

We will now introduce fitness differences into our model and set the relative fitnesses of allele 1 and 2 at location x to be 1 and $1 + s\gamma(x)$. To make progress in this model, we'll have to assume that selection isn't too strong, i.e. $s\gamma(x) \ll 1$ for all x . The change in frequency of allele 2 obtained within a generation due to selection is

$$q'(x, t) - q(x, t) \approx s\gamma(x)q(x, t)(1 - q(x, t)) \quad (7.19)$$

i.e. logistic growth of our favoured allele at location x . Putting our selection and migration terms together, we find the total change in allele frequency at location x in one generation is

$$q(x, t + 1) - q(x, t) = s\gamma(x)q(x, t)(1 - q(x, t)) + \frac{\sigma^2}{2} \frac{d^2 q(x, t)}{dx^2} \quad (7.20)$$

In deriving this result, we have essentially assumed that migration acted upon our original allele frequencies before selection, and in doing so have ignored terms of the order of σs .

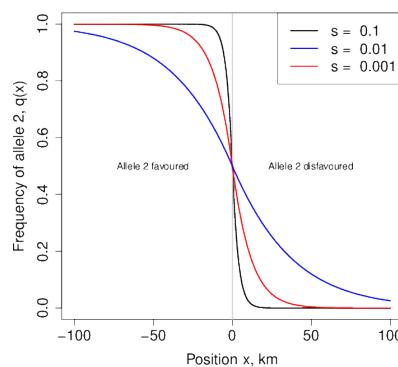


Figure 7.16: An equilibrium cline in allele frequency. Our individuals disperse an average distance of $\sigma = 1$ km per generation, and our allele 2 has a relative fitness of $1 + s$ and $1 - s$ on either side of the environmental change at $x = 0$.

The cline in allele frequency associated with a sharp environmental transition. To make progress, let's consider a simple model of local adaptation where the environment abruptly changes. Specifically, we assume that $\gamma(x) = 1$ for $x < 0$ and $\gamma(x) = -1$ for $x \geq 0$, i.e. our allele 2 has a selective advantage at locations to the left of zero, while this allele is at a disadvantage to the right of zero. In this case we can get an equilibrium distribution of our two alleles, where to the left of zero our allele 2 is at higher frequency, while to the right of zero allele 1 predominates. As we cross from the left to the right side of our range, the frequency of our allele 2 decreases in a smooth cline.

Our equilibrium spatial distribution of allele frequencies can be found by setting the left-hand side of eqn. (7.20) to zero to arrive at

$$s\gamma(x)q(x)(1-q(x)) = -\frac{\sigma^2}{2} \frac{d^2q(x)}{dx^2} \quad (7.21)$$

We then could solve this differential equation with appropriate boundary conditions ($q(-\infty) = 1$ and $q(\infty) = 0$) to arrive at the appropriate functional form for our cline. While we won't go into the solution of this equation here, we can note that by dividing our distance x by $\ell = \sigma/\sqrt{s}$, we can remove the effect of our parameters from the above equation. This compound parameter ℓ is the characteristic length of our cline, and it is this parameter which determines over what geographic scale we change from allele 2 predominating to allele 1 predominating as we move across our environmental shift.

Cline arising from an underdominant polymorphism

8

The Impact of Genetic Drift on Selected Alleles

4974 "Natural selection is a mechanism for generating an exceedingly high
4976 degree of improbability." –R.A. Fisher

4978 In the previous chapter we assumed that the selection acting on
our alleles was strong enough that we could ignore the action of
4980 genetic drift in shaping allele frequencies. However, genetic drift
affects all alleles, and so in this chapter we explore the interaction
4982 of selection and drift. Strongly selected alleles can be lost from the
population via drift when they are rare in the population, while both
weakly beneficial and weakly deleterious alleles are subject to the
4984 random whims of genetic drift throughout their entire time in the
population. Understanding the interaction of selection and genetic
4986 drift is key to understanding the extent to which small populations
may be mutation-limited in their rates of adaptation, and how rates
4988 of molecular and genome evolution may differ across taxa.

8.1 Stochastic loss of strongly selected alleles

4990 Even strongly beneficial alleles can be lost from the population when
they are sufficiently rare. This is because the number of offspring left
4992 by individuals to the next generation is fundamentally stochastic. A
selection coefficient of $s=1\%$ is a strong selection coefficient, which
4994 can drive an allele through the population in a few hundred genera-
tions once the allele is established. However, if individuals have
4996 on average a small number of offspring per generation, the first in-
dividual to carry our beneficial allele, who has on average 1% more
4998 children than their peers, could easily have zero offspring, leading to
the loss of our allele before it ever gets a chance to spread.

5000 To take a first stab at this problem, let's think of a very large hap-
loid population in which a single individual starts with the selected
5002 allele, and ask about the probability of eventual loss of our selected
allele starting from this single copy. To derive this probability of loss

5004 (p_L), we'll make use of a simple argument (derived from branching
5005 processes LetterSpace=10FISHER, 1923; LetterSpace=10HALDANE,
5006 1927). Our selected allele will be eventually lost from the population
if every individual with the allele fails to leave descendants. Well

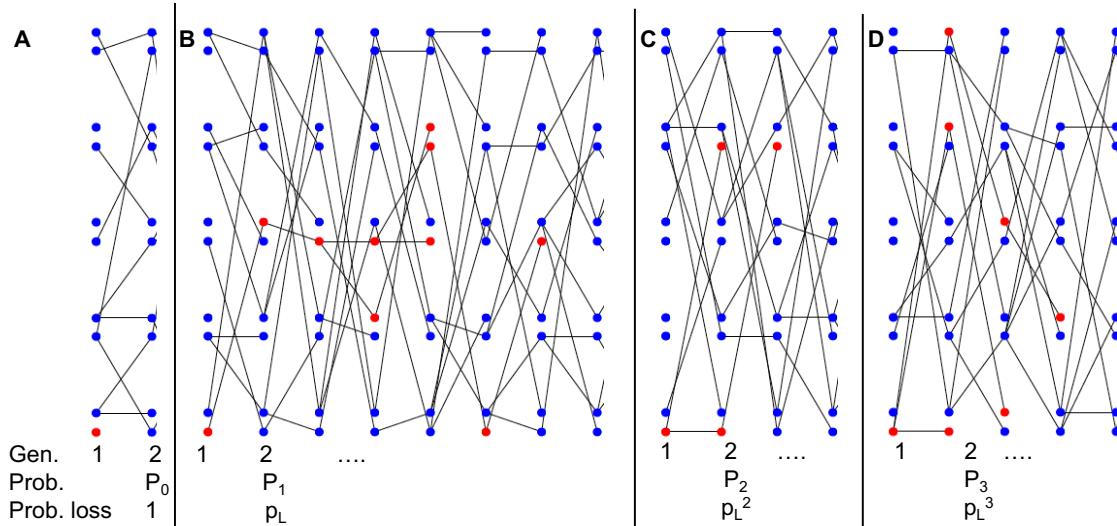


Figure 8.1: Four different outcomes of a selected allele present as a single copy in the population, leaving zero, one, two, three offspring in the next generation.

5008 we can think about different cases:

- 5009 1. In our first generation, with probability P_0 our individual allele
5010 leaves no copies of itself to the next generation, in which case our
allele is lost (Figure 8.1A).
- 5012 2. Alternatively, our allele could leave one copy of itself to the next
generation (with probability P_1), in which case with probability p_L
5014 this copy eventually goes extinct (Figure 8.1B).
- 5016 3. Our allele could leave two copies of itself to the next generation
5017 (with probability P_2), in which case with probability p_L^2 both of
these copies eventually go extinct (Figure 8.1C).
- 5018 4. More generally, our allele could leave could leave k copies ($k > 0$)
5019 of itself to the next generation (with probability P_k), in which case
5020 with probability p_L^k all of these copies eventually go extinct (e.g.
Figure 8.1D).

5022 Summing over these probabilities, we see that

$$p_L = \sum_{k=0}^{\infty} P_k p_L^k \quad (8.1)$$

We'll now need to specify P_k , the probability that an individual carrying our selected allele has k offspring. In order for this population

to stay constant in size, we'll assume that individuals without the selected mutation have on average one offspring per generation, while individuals with our selected allele have on average $1 + s$ offspring per generation. We'll assume that the number of offspring an individual has is Poisson distributed with mean given by 1 or $1 + s$, i.e. the probability that an individual with the selected allele has i children is

$$P_i = \frac{(1+s)^i e^{-(1+s)}}{i!} \quad (8.2)$$

Substituting P_k into the equation above, we see

$$\begin{aligned} p_L &= \sum_{k=0}^{\infty} \frac{(1+s)^k e^{-(1+s)}}{k!} p_L^k \\ &= e^{-(1+s)} \left(\sum_{k=0}^{\infty} \frac{(p_L(1+s))^k}{k!} \right) \end{aligned} \quad (8.3)$$

The term in the brackets is itself an exponential expansion, so we can rewrite this equation as

$$p_L = e^{(1+s)(p_L - 1)} \quad (8.4)$$

Solving for p_L would give us our probability of loss for any selection coefficient. Let's rewrite our result in terms of the probability of escaping loss, $p_F = 1 - p_L$. We can rewrite eqn. (8.4) as

$$1 - p_F = e^{-p_F(1+s)} \quad (8.5)$$

To gain an approximate solution for this result, let's consider a small selection coefficient $s \ll 1$ such that $p_F \ll 1$ and then use a Taylor series to expand out the exponential on the right hand side (ignoring terms of higher order than s^2 and p_F^2):

$$1 - p_F \approx 1 - p_F(1+s) + p_F^2(1+s)^2/2 \quad (8.6)$$

Solving this we find that

$$p_F = 2s. \quad (8.7)$$

Thus even an allele with a 1% selection coefficient has a 98% probability of being lost when it is first introduced into the population by mutation.

If the mutation rate towards our advantageous allele is μ , and there are N individuals in our haploid population, then $N\mu$ advantageous mutations arise per generation. Each of these new beneficial mutations has a probability p_F of fixing. Thus the number of advantageous mutations arising per generation that will eventually fix in the population is $N\mu p_F$, and the waiting time for a mutation that will fix to arise is the reciprocal of this: $1/N\mu p_F$. Thus, in adapting to a

novel selection pressure via new mutations, the population size, the mutational target size, and the selective advantage of new mutations all matter. One reason why combinations of drugs are used against viruses like HIV and malaria is that, even if the viruses adapt to one of the drugs, the viral load (N) of the patient is greatly reduced, making it very unlikely that the population will manage to fix a second drug-resistant allele.

Diploid model of stochastic loss of strongly selected alleles. We can also adapt this result to a diploid setting. Assuming that heterozygotes for the 1 allele have on average $1 + hs$ children, the probability allele 1 is not lost, starting from a single copy in the population, is

$$p_F = 2hs \quad (8.8)$$

for $h > 0$. Note this is a slightly different parameterization from our diploid model in the previous chapter; here h is the dominance of our positively selected allele, with $h = 1$ corresponding to the full selective advantage expressed in an individual with only a single copy. Thus the probability that a beneficial allele is not lost depends just on the relative fitness advantage of the heterozygote; this is because when the allele is rare it is usually present in heterozygotes and so its probability of escaping loss just depends on the fitness of these individuals compared to homozygotes for the ancestral allele (assuming an outbred population).

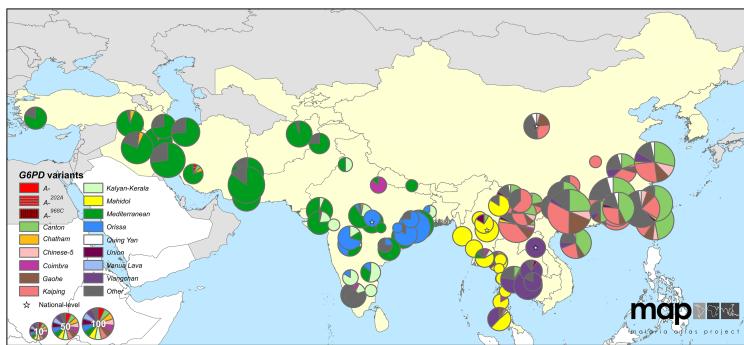


Figure 8.2: **Map of G6PD-deficiency allele frequencies across Asia.** The pie chart shows the frequency of G6PD-deficiency alleles. The size of the pie chart indicates the number of G6PD-deficient individuals sampled. Countries with endemic malaria are colored yellow. Figure from LetterSpace=10HOWES *et al.* (2013), licensed under CC BY 4.0.

Over roughly the past ten thousand years, adaptive alleles conferring resistance to malaria have arisen in a number of genes and spread through human populations in areas where malaria is endemic (LetterSpace=10KWIATKOWSKI, 2005). One particularly impressive case of convergent evolution in response to selection pressures imposed by malaria are the numerous changes throughout the G6PD gene, which include at least 15 common variants in Central and Eastern Asia alone that lower the activity of the enzyme (Let-

terSpace=10HOWES *et al.*, 2013). These alleles are now found at a combined frequency of around 8% frequency in malaria endemic areas, rarely exceeding 20% (LetterSpace=10HOWES *et al.*, 2012). Whether these variants *all* confer resistance to malaria is unknown, but a number of these alleles have demonstrated effects against malaria and are thought to have a selective advantage to heterozygotes $sh > 5\%$ where malaria is endemic (LetterSpace=10RUWENDE *et al.*, 1995; LetterSpace=10TISHKOFF *et al.*, 2001; LetterSpace=10LOUICHAROEN *et al.*, 2009).

With a 5% advantage in heterozygotes, a G6PD allele present as a single copy would only have a 10% probability of fixing in the population. If that's so, how come malaria adaptation has repeatedly occurred via changes at G6PD? Well, maybe adaptation didn't start from a single copy of the selected allele. How many copies of the G6PD-deficiency alleles do we expect were segregating in the population before selection pressures changed?

In the absence of malaria, these G6PD alleles are deleterious with carriers suffering from G6PD deficiency, leading to hemolytic anemia when individuals are exposed to a variety of different compounds, notably those present in fava beans. There's upward of one hundred bases where G6PD-deficiency alleles can arise, so assuming a mutation rate of $\approx 10^{-8}$ per base pair per generation, we can roughly estimate the rate of mutations arising that affect the G6PD gene as $\mu \approx 10^{-6}$ per generation. In the absence of malaria, the selective cost of being a heterozygote carrier of a G6PD-deficient allele must have been on the order of 5% or more, and thus the frequency of the allele under mutation-selection balance would have been $\approx 10^{-6}/0.05 = 2 \times 10^{-5}$. Assuming an effective population size of 2 – 20 million individuals, roughly five to ten thousand years ago that means that there would have been forty to four hundred copies of the G6PD-deficiency allele present in the population when selection pressures shifted at the introduction of malaria. The chance that one of these newly adaptive alleles is lost is 90% but the chance that they're all lost is $< (0.9)^{40} \approx 0.02$, i.e. there would have been a greater than 98% chance that adaptation would occur via one or more alleles at G6PD. How many alleles would escape drift? Well with 40 – 400 copies of the allele pre-malaria, and each of them having a 10% probability of escaping drift, we expect between 4 and 40 G6PD alleles to escape drift and contribute to adaptation. We see 15 common G6PD alleles in Eurasia, so our simple model of adaptation from mutation-selection balance seems reasonable.

Question 1. 'Haldane's sieve' is the name for the idea that the mutations that contribute to adaptation are likely to be dominant or at least co-dominant.

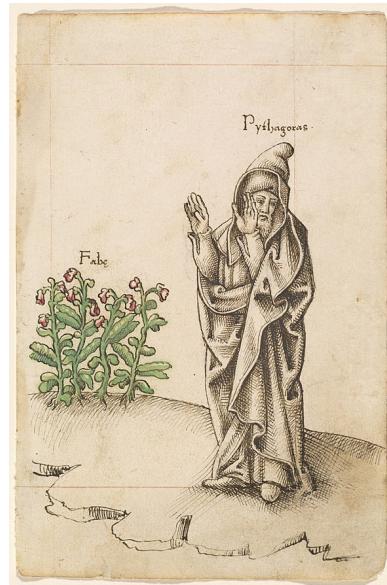


Figure 8.3: Pythagoras's "just say no to fava beans" campaign. Pythagoras prohibited the consumption of fava beans by his followers; perhaps because favaism, the anemia induced in G6PD-deficient individuals by fava beans, is relatively common in the Mediterranean due to adaptation to endemic malaria. French early 16th Century. Woodner Collection, National Gallery of Art. Public Domain, wikipedia.
A full analysis of this case requires modeling of G6PD's X chromosome inheritance, and the randomness in the number of copies of the allele present at mutation-selection balance (LetterSpace=10RALPH and LetterSpace=10COOP, 2015).



Figure 8.4: Haldane's sieve. To our knowledge Haldane never wore a sieve, but we assume he owned one. Sieve, Flickr licensed under CC BY 2.0. Haldane, Public Domain wikipedia.

A) Briefly explain this argument with a verbal model relating to the results we've developed in the last two chapters.

B) Haldane's sieve is thought to be less important for adaptation from previously deleterious standing variation, than adaptation from new mutation. Can you explain the intuition behind of this idea?

C) Haldane's sieve is likely to be less important in inbred, e.g. selfing, populations. Why is this?

Question 2. Melanic squirrels suffer a higher rate of predation (due to hawks) than normally pigmented squirrels. Melanism is due to a dominant, autosomal mutation. The frequency of melanic squirrels at birth is 4×10^{-5} .

A) If the mutation rate to new melanic alleles is 10^{-6} , assuming the melanic allele is at mutation-selection equilibrium, what is the reduction in fitness of the heterozygote?

Suddenly levels of pollution increase dramatically in our population, and predation by hawks now offers an equal (and opposite) advantage to the dark individuals as it once offered to the normally pigmented individuals.

B) What is the probability that a single copy of this allele (present just once in the population) is lost?

C) If the population size of our squirrels is a million individuals, and is at mutation-selection balance, what is the probability that the population adapts from one or more allele(s) from the standing pool of melanic alleles?

8.2 The interaction between genetic drift and weak selection.

For strongly selected alleles, once the allele has escaped initial loss at low frequencies, its path will be determined deterministically by its selection coefficients. However, if selection is weak compared to genetic drift, the stochasticity of reproduction can play a role in the trajectory an allele takes even when it is common in the population. If selection is sufficiently weak compared to genetic drift, then genetic drift will dominate the dynamics of alleles and they will behave like they're effectively neutral. Thus, the extent to which selection can shape patterns of molecular evolution will depend on the relative strengths of selection and genetic drift. But how weak must selection on an allele be for drift to overpower selection? And do these interactions between selection and drift have longterm consequences for genome-wide patterns evolution?

To model selection and drift each generation, we can first calculate the deterministic change in our allele frequency due to selection using our deterministic formula. Then, using our newly calculated



Figure 8.5: cress bug (*Asellus aquaticus*) in the isopod family *Asellidae*.
Brehms Tierleben. Allgemeine Kunde des Tierreichs (1911).
Brehm A.E. Image from the Biodiversity Heritage Library.
Contributed by Smithsonian Libraries. Not in copyright.

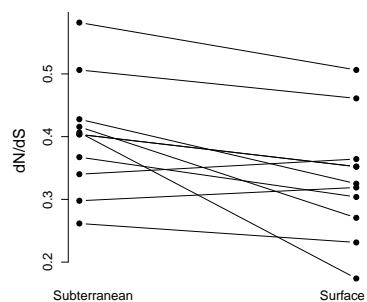


Figure 8.6: Asellid isopods have repeatedly invaded subterranean, groundwater habitats from surface-water habitats, and leading to a genome-wide increase in d_N/d_S and larger genomes (Data from Lefébure et al., 2017, comparing independent invasions across pairs). One possible

5166 expected allele frequency, we can binomially sample two alleles for
each of our offspring to construct the next generation. This approach
5168 to jointly modeling genetic drift and selection is called the Wright-
Fisher model.

5170 Under the Wright-Fisher model, we will calculate the expected
change in allele frequency due to selection and the variance around
5172 this expectation due to drift. To make our calculations simpler, let's
assume an additive model, i.e. $h = 1/2$, and that $s \ll 1$ so that $\bar{w} \approx 1$.
5174 Using our directional selection deterministic model, from Chapter 6,
and these approximations gives us our deterministic change due to
5176 selection

$$\Delta_S p = \mathbb{E}(\Delta p) = \frac{s}{2} p(1-p) \quad (8.9)$$

To obtain our new frequency in the next generation, p_1 , we binomially sample from our new deterministic frequency $p' = p + \Delta_S p$, so
5178 the variance in our allele frequency change from one generation to
the next is given by

$$Var(\Delta p) = Var(p_1 - p) = Var(p_1) = \frac{p'(1-p')}{2N} \approx \frac{p(1-p)}{2N}. \quad (8.10)$$

where the previous allele frequency p drops out because it is a constant
5182 and the variance in our new allele frequency follows from the fact that we are binomially sampling $2N$ new alleles from a frequency p' to form the next generation.

To get our first look at the relative effects of selection vs. drift
5186 we can simply look at when our change in allele frequency caused by selection within a generate is reasonably faithfully passed down through the generations. In particular, if our expected change in allele frequency is much greater than the variance around this change, genetic drift will play little role in the fate of our selected allele (once the allele is not at low copy number within the population).
5190 When does selection dominant genetic drift? This will happen if $\mathbb{E}(\Delta p) \gg Var(\Delta p)$, i.e. when $|Ns| \gg 1$. Conversely, any hope of our selected allele following its deterministic path will be quickly undone if our change in allele frequencies due to selection is much less than the variance induced by drift. So if the absolute value of our population-size-scaled selection coefficient $|Ns| \ll 1$, then drift will dominate the fate of our allele.

To make further progress on understanding the fate of alleles with
5200 selection coefficients of the order $1/N$ requires more careful modeling. However, under our diploid model, with an additive selection coefficient s , we can obtain the probability that allele 1 fixes within the population, starting from a frequency p :

$$p_F(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \quad (8.11)$$

To see this denote our new count of allele 1 by i , then

$$\begin{aligned} Var(p_1 - p) &= Var\left(\frac{i}{2N} - p\right) = Var\left(\frac{i}{2N}\right) \\ &= \frac{Var(i)}{(2N)^2} \end{aligned}$$

and from binomial sampling $Var(i) = 2Np'(1-p')$ and so we arrive at our answer. Assuming that $s \ll 1$, $p' \approx p$, then in practice we can use

$$Var(\Delta p) = Var(p' - p) \approx p(1-p)/2N.$$

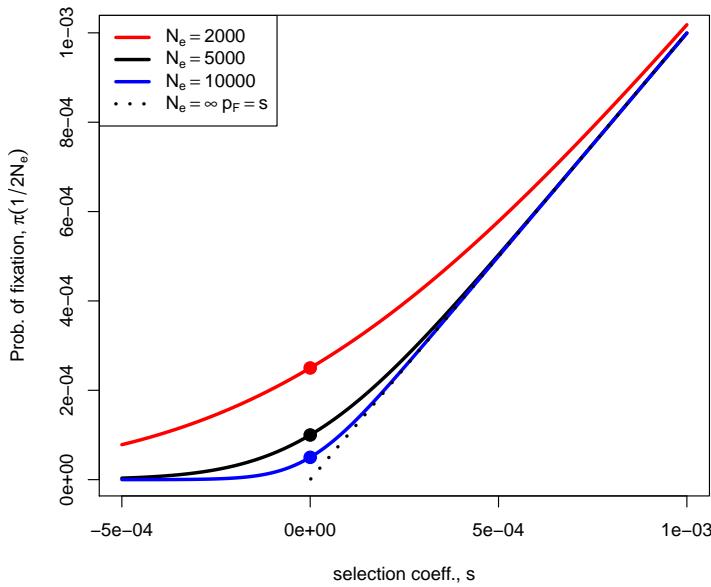


Figure 8.7: The probability of the fixation of a new mutation with selection coefficient s ($h = 1/2$) in a diploid population of effective size N_e . The dashed line gives the infinite population solution. The dots give the solution for $s \rightarrow 0$, i.e. the neutral case, where the probability of fixation is $1/(2N_e)$. Code here.

5204 The proof of this result is sketched out below (see Section 8.2.1). A
new allele that arrives in the population at frequency $p = 1/(2N)$ has
5206 a probability of reaching fixation of

$$p_F \left(\frac{1}{2N} \right) = \frac{1 - e^{-s}}{1 - e^{-2Ns}} \quad (8.12)$$

If $s \ll 1$ but $Ns \gg 1$ then $p_F(\frac{1}{2N}) \approx s$, which nicely gives us back
5208 the result that we obtained above for an allele under strong selection
(eqn. (8.8)). Our probability of fixation (eqn. (8.12)) is plotted as a
5210 function of s and N in Figure 8.8. To recover our neutral result, we
can take the limit $s \rightarrow 0$ to obtain our neutral fixation probability,
5212 $1/(2N)$.

In the case where Ns is close to 1, then

$$p_F \left(\frac{1}{2N} \right) \approx \frac{s}{1 - e^{-2Ns}} \quad (8.13)$$

5214 This is greater than our earlier result $p_F = s$ from the branching
process argument (using our additive model of $h = 1/2$), increasingly
5216 so for smaller N . Why is this? The reason why is that p_F is really the
probability of "never being lost" in an infinitely large population. So
5218 to persist indefinitely, the allele has to escape loss permanently, by
never being absorbed by the zero state. When the population size
5220 is finite, to fix we only need to reach a size $2N$ individuals. Weakly

beneficial mutations ($Ns > 1$) are slightly more likely to fix than the neutral probability, as they only have to reach $2N$ to never be lost.

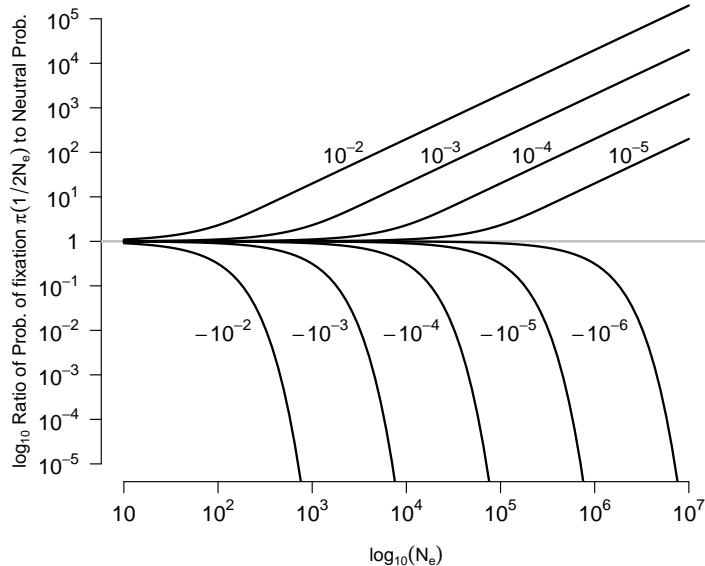


Figure 8.8: The probability of the fixation of a new mutation with selection coefficient s relative to the neutral fixation probability ($1/2N_e$) as a function of the effective size N_e . The selection coefficient is shown next to the line. Note how quickly the probabilities move away from the neutral expectation as $N_e s$ moves passed 1. [Code here.](#)

If, for selection to operate on an allele, we need the selection coefficient to satisfy $|Ns| \gg 1$, then that holds if $|s| \gg 1/N$. Well, effective population sizes are often reasonably large, on the order of hundreds of thousands or millions of individuals, thus selection coefficients on the order of 10^{-5} to 10^{-6} can be effectively selected upon, i.e. selection equivalent to individuals have incredibly slight advantages in terms of the number of offspring they leave to the next generation. While we are incapable of detecting measuring all but the large fitness effect sizes, except in some elegant experiments (e.g. in microbes), such small effects are visible to selection in large populations. Thus, if consistent selection pressures are exerted over long time periods, natural selection can potentially finely tune various aspects of an organism.

As one example of this fine-tuning, consider how carefully crafted and optimized the sequence of codons is for translation. Due to the degeneracy of the protein code, multiple codons code for the same amino-acid. For example, there are six different codons that can code leucine. While these synonymous codons are equivalent at the protein level, cells do differ in the number of tRNA molecules that bind these codons and so the efficacy and accuracy with which proteins can be formed through translation and folding. These slight differ-

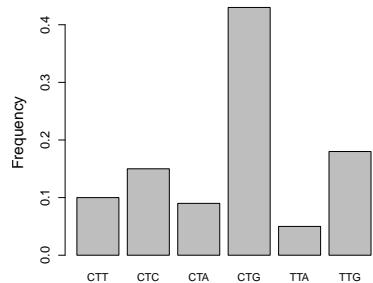


Figure 8.9: Data from *Drosophila melanogaster* on the frequency of different codons for Leucine. Data from Genscript. [Code here.](#)

5244 ences in translation rates likely often correspond to tiny differences in
 fitness, but do they matter?

5246 In many organisms there is a strong bias in the codons to encode
 particular amino-acids, see Figure 8.9, with the most abundant codon
 5248 matching the most abundant tRNA in cells. This 'codon bias' likely
 reflects the combined action of weak selection and mutational pres-
 5250 sure, pushing the codon composition of the genome and tRNA abun-
 dances towards an adaptive compromise. These selection pressures
 5252 have acted over long time periods, as codon usage patterns are often
 very similar for species that diverged over many tens of millions of
 5254 years ago. Compared to other genes, highly expressed genes show
 a strong bias towards using codons matching abundant tRNAs, con-
 5256 sistent with the idea that the synonymous codon content of highly
 5258 expressed genes is evolving to optimize their translation (see Figure
 8.10 for an early example). These patterns likely represent the action
 of selection pressures that are incredibly weak on average, but that
 5260 have played out over vast time-periods.

5262 *The fixation of slightly deleterious alleles.* From Figure 8.8 we can see
 that weakly deleterious alleles can also fix, especially in small pop-
 5264ulations. To understand how likely it is that deleterious alleles by
 chance reach fixation by genetic drift, let's assume a diploid model
 with additive selection (with a selection coefficient of $-s$ against our
 5266 allele 2).

5268 If $Ns \gg 1$ then our deleterious allele (allele 2) cannot possibly
 reach fixation. However, if Ns is not large, then the probability of
 fixation

$$p_F \left(\frac{1}{2N} \right) \approx \frac{s}{e^{2Ns} - 1} \quad (8.14)$$

5270 for our single-copy deleterious allele. So deleterious alleles can fix
 within populations (albeit at a low rate) if Ns is not too large. As
 5272 above, this is because while deleterious mutations will never escape
 loss in infinite populations, they can become fixed in finite popula-
 5274 tion by reaching $2N$ copies.

5276 **Question 3.** An additive mutation arises that lowers the relative
 fitness of heterozygotes by 10^{-5} . What is the probability that this
 mutation fixes in a diploid population with effective size of 10^4 ?
 5278 What is the probability it fixes in a population of effective size 10^6 ?
 By comparing both to their neutral probability describe the intuition
 5280 behind this result.

5282 LetterSpace=10OHTA proposed the 'nearly-neutral' theory of
 molecular evolution in a series of papers¹. She suggested that a rea-
 sonable fraction of newly arising functional mutations may have

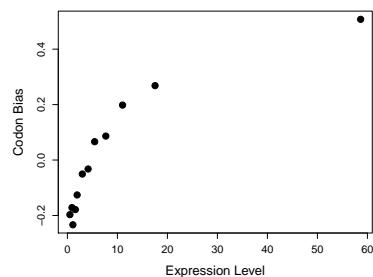


Figure 8.10: A measure of unequal codon frequencies (F) plotted in bins of gene expression (E) for genes across the *Drosophila melanogaster* genome. Data from LetterSpace=10HEY and LetterSpace=10KIMAN (2002). Code here.

¹ LetterSpace=10OHTA, T., 1972 Population size and rate of evolution. Journal of Molecular Evolution 1(4): 305–314; LetterSpace=10OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246(5428): 96; and LetterSpace=10OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. Journal of Molecular Evolution 26(1-2): 1–6

5284 very weak selection coefficients, such that species with smaller effective
5285 population sizes may have higher rates of fixation of these very
5286 weakly deleterious alleles. In effect, her suggestion is that the constraint
5287 parameter C of a functional region is not a fixed property, but
5288 rather depends on the ability of the population to resist the influx of
very weakly deleterious mutations.

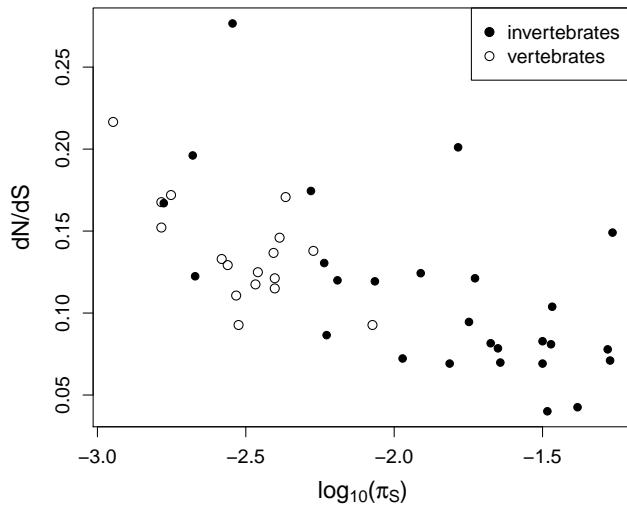


Figure 8.11: Data from 44 metazoan species from Cuttlefish to Sifakas. Each dot represents the average of over many genes plotting d_N/d_S against synonymous diversity (π_S). Data from LetterSpace=10GALTIER (2016). Code here.

5290 Across species, genome-wide averages of d_N/d_S do seem to be cor-
5291 related with measures of the effective population size (such as syn-
5292 onymous diversity), see Figure 8.11. This evidence supports the idea
5293 that in species with smaller effective population sizes (lower π_S), pro-
5294 teins may be subject to lower degrees of constraint, as very weakly
5295 deleterious mutations are able to fix. Thus, some reasonable propor-
5296 tion of functional substitutions in populations with small effective
5297 population sizes, such as humans, may be mildly deleterious.

5298 8.2.1 Appendix: The fixation probability of weakly selected alleles

What is the probability a weakly beneficial or deleterious additive
5300 allele fixes in our population? We'll let $P(\Delta p)$ be the probability that
our allele frequency shifts by Δp in the next generation. Using this,
5302 we can write our probability $p_F(p)$ in terms of the probability of
achieving fixation averaged over the frequency in the next generation

$$p_F(p) = \int p_F(p + \Delta p)P(\Delta p)d(\Delta p) \quad (8.15)$$



Figure 8.12: Common Cuttlefish (*Sepia officinalis*).

Cefalopodi viventi nel Golfo di Napoli (1896). Jatta G. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

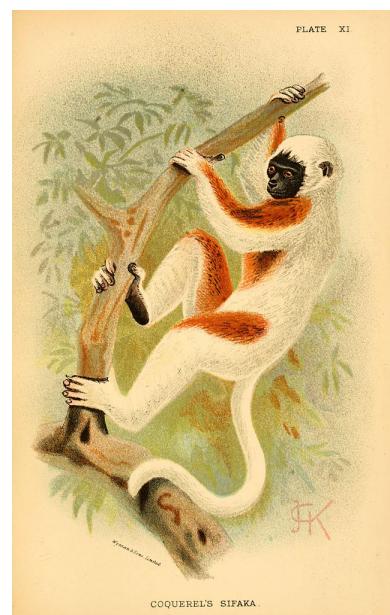


Figure 8.13: Coquerel's Sifaka (*Propithecus coquereli*).

A hand-book to the primates (1894). Forbes, H. O. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

5304 This is very similar to the technique that we used when deriving our probability of escaping loss in a very large population above.

5306 So we need an expression for $p_F(p + \Delta p)$. To obtain this, we'll do a Taylor series expansion of $p_F(p)$, assuming that Δp is small:

$$p_F(p + \Delta p) \approx p_F(p) + \Delta p \frac{dp_F(p)}{dp} + (\Delta p)^2 \frac{d^2 p_F(p)}{dp^2}(p) \quad (8.16)$$

5308 ignoring higher order terms.

Taking the expectation over Δp on both sides, as in eqn. 8.15, we
5310 obtain

$$p_F(p) = p_F(p) + \mathbb{E}(\Delta p) \frac{dp_F(p)}{dp} + \mathbb{E}((\Delta p)^2) \frac{d^2 p_F(p)}{dp^2} \quad (8.17)$$

Well, $\mathbb{E}(\Delta p) = \frac{s}{2}p(1-p)$ and $\text{Var}(\Delta p) = \mathbb{E}((\Delta p)^2) - \mathbb{E}^2(\Delta p)$, so if
5312 $s \ll 1$ then $\mathbb{E}^2(\Delta p) \approx 0$, and $\mathbb{E}((\Delta p)^2) = \frac{p(1-p)}{2N}$. Substituting in these values and subtracting p from both sides of our equation, this leaves
5314 us with

$$0 = \frac{s}{2}p(1-p) \frac{dp_F(p)}{dp} + \frac{p(1-p)}{2N} \frac{d^2 p_F(p)}{dp^2} \quad (8.18)$$

and we can specify the boundary conditions to be $p_F(1) = 1$ and

5316 $p_F(0) = 0$. Solving this differential equation is a somewhat involved process, but in doing so we find that

$$p_F(p) = \frac{1 - e^{-2Ns p}}{1 - e^{-2Ns}} \quad (8.19)$$

5318 This proof can be extended to alleles with arbitrary dominance, however, this does not lead to a analytically tractable expression so we do
5320 not pursue this here.

9

5322 The Effects of Linked Selection.

LetterSpace=10GENETIC DRIFT IS NOT THE ONLY SOURCE OF RANDOMNESS in the dynamics of alleles. Alleles also experience random fluctuations in frequency due to the fact that they present on a set of random genetic backgrounds with different fitnesses. For example, when a beneficial allele arises via a single mutation, it arises on a particular genetic background, i.e. a particular haplotype (Figure 9.1A). Imagine this mutation arising in a region with no recombination, or in an organism where genetic exchange is rare. If our beneficial allele becomes established in the population, i.e. escapes loss by genetic drift in those first few generations, it will start to increase in frequency rapidly. As it rises in frequency, so will the alleles that happened to be present on the haplotype that the mutation arose on (if those other alleles are neutral or at least not too deleterious). These other alleles are getting to 'hitchhiking' along. The alleles that are not on that particular background are swept out of the population, so the net effect of this selective sweep is to remove genetic diversity from the population. Diversity will eventually recover, as new mutations arise and some slowly drift up in frequency. But in the short-term, selective sweeps remove genetic variation from populations.

LetterSpace=10WILLIAMS and LetterSpace=10PENNINGS (2019) have visualized selective sweeps in HIV. In Figure 9.1B) we see a set of HIV haplotypes sampled from a patient before and after of a selective sweep of a drug-resistant mutation. The patient is taking a retrotransposase inhibitor (Efavirenz), but sadly within 161 days a drug-resistant mutation that changes the HIV retrotransposase protein has arisen and spread. Note how a particular haplotype is now fixed in the sample, and little genetic diversity remains, due to the hitchhiking effect of the strong selective sweep of this allele.

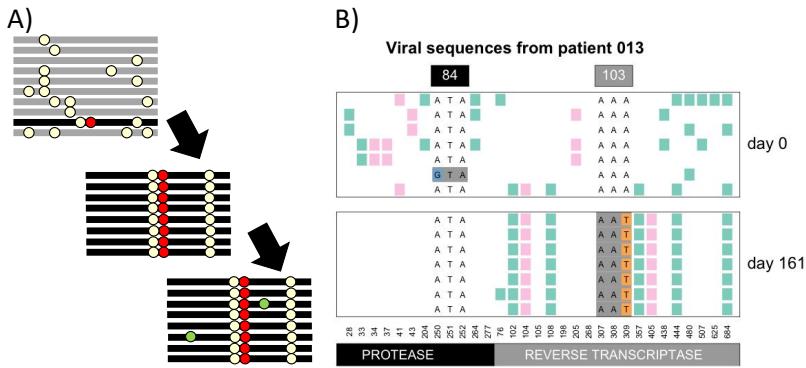


Figure 9.1: **A)** In the top panel, a selected mutation (red dot) arises on a particular haplotype in the population. It sweeps to fixation, carrying with it the haplotype on which it arose, middle panel, erasing the standing genetic diversity in the region. The bottom panel is some time after the selective sweep when some new neutral alleles (green dots) have started to drift up in frequency. **B)** Top panel: HIV sequences from a patient at the start of drug treatment in the protease and retrotransposase coding regions. Bottom panel: A sample 161 days later, after a drug resistant mutation has spread, the $A \rightarrow T$ in the 103rd codon of retrotransposase. Each row is a haplotype, with the alleles present shown as coloured blocks. Figure B from LetterSpace=10WILLIAMS and LetterSpace=10PENNINGS (2019), licensed under CC BY 4.0.

To better understand hitchhiking, first let's imagine examining variation at a locus fully linked to our selected locus, just after our sweep reached fixation. Neutral alleles sampled at this locus must trace their ancestral lineages back to the neutral allele on whose background the selected allele initially arose (Figure 9.6). This is because that background neutral allele, which existed τ generations ago, is the ancestor of the entire population at this fully linked locus. Our individuals who carry the beneficial allele are, from the perspective of these alleles, experiencing a rapidly expanding population. Therefore, a pair of neutral alleles sampled at our linked neutral locus will be forced to coalesce $\approx \tau$ generations ago. A newly derived allele with an additive selection coefficient s will take a time $\tau = 4 \log(2N)/s$ generations to reach fixation within our population (see eqn. (6.39)). This is a very short-time scale compared to the average neutral coalescent time of $2N$ generations for a pair of alleles. Thus we expect little variation, as few mutations will have arisen on these very short branches, and those that have done will likely be singletons in our sample.

Now let's think about a sweep in a recombining region. Again the selected mutation arises on a particular haplotype, and it and its haplotype starts to increase in frequency in the population. However, now recombination events can occur between haplotypes carrying and not carrying the selected allele, in individuals who are heterozygote for the selected allele. These recombination events allow alleles that were not present on the original selected haplotype to avoid being swept out of the population, and also decouple the selected allele somewhat from hitchhiking alleles, preventing many of them from hitchhiking all the way to fixation. Far out from the selected site, the recombination rate is high enough that alleles that were present on the original background barely get to hitchhike along at all, as re-

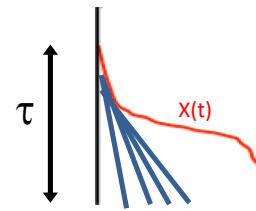


Figure 9.2: The coalescent of 4 lineages, marked in blue, at a locus completed linked to our selected allele. The frequency trajectory of the selected allele $X(t)$ is shown in red.

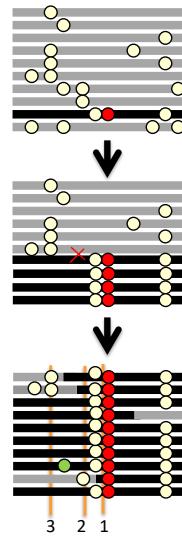


Figure 9.3: A cartoon depiction of a sweep of a red beneficial allele over three time points. The haplotype that the beneficial arose on by mutation is shown in black. The three vertical orange lines mark the loci shown in Figure 9.4. Neutral alleles segregating prior to the sweep appear as white circles, new mutations after the sweep as green circles.

combination breaks up their association with the selected allele very rapidly.

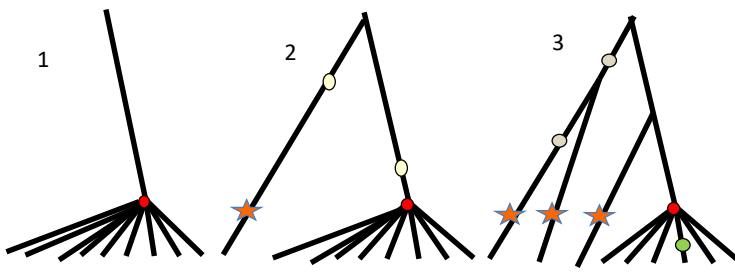


Figure 9.4: Coalescent genealogies at three loci different distances along the genome from a selective sweep. The locations of these three loci along the genome are marked in Figure 9.3. The selected mutation is shown in red. Lineages descended from recombination events during the sweep are marked in stars. Neutral mutations close to each of the loci are shown on the genealogy.

What do the coalescent genealogies look like at loci various dis-

tances away from the selected site? Well, close to the selected site all our alleles in the present day trace back to a most recent common ancestor allele present on that selected haplotype, and so are all forced to coalesce around τ generations ago (locus 1). Slightly further out from the selected site (locus 2), we have lineages that don't trace their ancestry back to the original selected haplotype, but instead are descended from recombinant haplotypes that recombined onto the sweep (the haplotype 2 from the bottom). These lineages can coalesce neutrally with the other ancestral lineages over far deeper time scales and mutations on these deeper lineages correspond to the standing diversity present in our population prior to the sweep. As we move even further out from the selected site (locus 3), we encounter more and more lineages descended from recombinant haplotypes that coalesce neutrally much deeper in time than τ , allowing diversity to recover to background levels as we move away from the selected site.

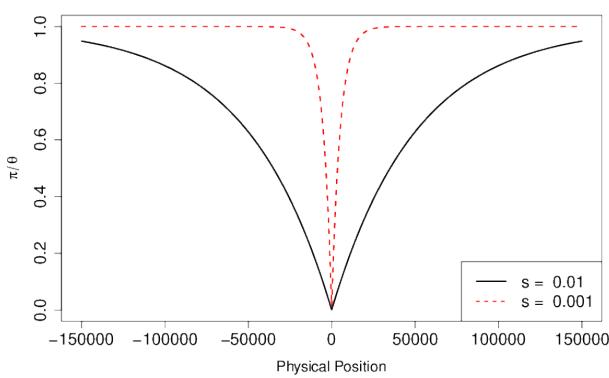


Figure 9.5: The expected reduction in diversity compared to its neutral expectation as a function of the distance away from a site where a selected allele has just gone to fixation. The sweeps associated with two different strengths of selection are shown, corresponding to a short timescale (τ) for the sweep and long one. The recombination rate is $r_{BP} = 1 \times 10^{-8}$. Code here.

To model the expected pattern of diversity surrounding a selected

site, we can think about a pair of alleles sampled at a neutral locus

a recombination distance r away from our selected site. Our pair of alleles will be forced to coalesce $\approx \tau$ generations if neither of them of are descended from recombinant haplotypes.

We know that in the present day our neutral lineage is linked to the selected allele. The probability that our lineage, in some generation t back in time, is in a heterozygote is $1 - X(t)$, and the probability that a recombination occurs in that individual is r . So the probability that our neutral lineage is descended from a recombinant haplotype t generations back is

$$r(1 - X(t)) \quad (9.1)$$

So the probability (p_{NR}) that our lineage is not descended from a recombinant haplotype from a recombination event in the τ generations it takes our selected allele to move through the population is

$$p_{NR} = \prod_{t=1}^{\tau} (1 - r(1 - X(t))) \quad (9.2)$$

Assuming that r is small, then $(1 - r(1 - X(t))) \approx e^{-r(1-X(t))}$, such that

$$p_{NR} = \prod_{t=1}^{\tau} (1 - r(1 - X(t))) \approx \exp \left(-r \sum_{t=1}^{\tau} 1 - X(t) \right) = \exp \left(-r\tau(1 - \hat{X}) \right) \quad (9.3)$$

where \hat{X} is the average frequency of the derived beneficial allele across its trajectory as it sweeps up in frequency, $\hat{X} = \frac{1}{\tau} \sum_{t=1}^{\tau} X(t)$. As our allele is additive, its trajectory for frequencies < 0.5 is the mirror image of its trajectory for frequencies > 0.5 , therefore its average frequency $\hat{X} = 0.5$. This simplifies our expression to

$$p_{NR} = e^{-r\tau/2}. \quad (9.4)$$

The probability that neither of our lineages is descended from a recombinant haplotype, and hence are forced to coalesce, is p_{NR}^2 (assuming that they coalesce at a time close to τ so that they recombine independently of each other for times $< \tau$).

If one or other of our lineages is descended from a recombinant haplotype, it will take them on average $\approx 2N$ generations to find a common ancestor, as we are back to our neutral coalescent probabilities. Thus, the expected time till our pair of lineages find a common ancestor is

$$\mathbb{E}(T_2) = \tau \times p_{NR}^2 + (1 - p_{NR}^2)(\tau + 2N) \approx (1 - p_{NR}^2)2N \quad (9.5)$$

where this last approximation assumes that $\tau \ll 2N$. So the expected pairwise diversity for neutral alleles at a recombination distance r away from the selected sweep (π_r) is

$$\mathbb{E}(\pi_r) = 2\mu\mathbb{E}(T_2) \approx \pi_0 (1 - e^{-r\tau}) \quad (9.6)$$

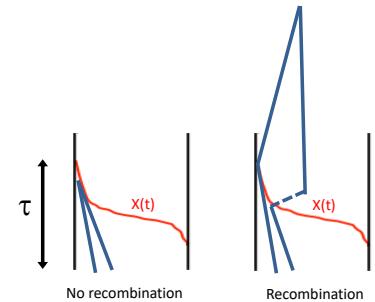
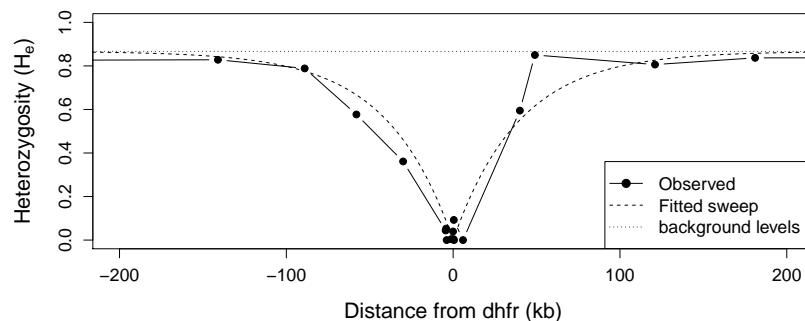


Figure 9.6:

So diversity increases as we move away from the selected site, slowly and exponentially plateauing to its neutral expectation π_0 .

The malaria pathogen (*Plasmodium falciparum*) has evolved drug resistance to anti-malaria drugs, often by changes at the dhfr gene. Figure 9.8 shows levels of genetic diversity (heterozygosity) at a set of markers moving out from the dhfr gene in a set of drug resistant malaria sequences collected in Thailand (LetterSpace=10NASH *et al.*, 2005). We see the characteristic dip in diversity around the gene, with zero diversity at a number of the loci very close to the gene, suggesting a strong selective sweep. Fitting our simple model of a sweep to this data, we estimate that $\tau \approx 40$ generations, corresponding to the drug-resistance allele fixing in very short time period.



To get a sense of the physical scale over which diversity is reduced, consider a region where recombination occurs at a rate r_{BP} per base pair per generation, and a locus ℓ base pairs away from the selected site, such that $r = r_{BP}\ell$ (where $r_{BP}\ell \ll 1$ so we don't need to worry about more than one recombination event occurring per generation). Typical recombination rates are on the order of $r_{BP} = 10^{-8}$. In Figure 9.5 we show the reduction in diversity, given by eqn. (9.6), for two different selection coefficients.

For our expected diversity level to recover to 50% of its neutral expectation $E(\pi_r)/\theta = 0.5$, requires a physical distance ℓ^* such that $\log(0.5) = -r_{BP}\ell^*\tau$, and by re-arrangement,

$$\ell^* = \frac{-\log(0.5)}{r_{BP}\tau}. \quad (9.7)$$

As τ depends inversely on the selection s (eqn. (6.39)), the width of our trough of reduced diversity depends on s/r_{BP} . All else being equal, we expect stronger sweeps or sweeps in regions of low recombination to have a larger hitchhiking effect. For example, in a genomic region with a recombination rate $r_{BP} = 10^{-8}\text{bp}^{-1}$ a selection coefficient of $s = 0.1\%$ would reduce diversity over 10's of kb, while a sweep of $s = 1\%$ would affect $\sim 100\text{kb}$.

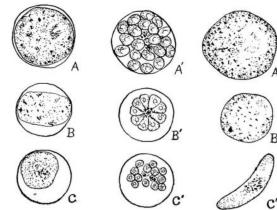


FIG. 9.7. Comparison of three species of malaria parasites $\times 2000$ (figures selected largely from Manson, G., C. and A', Plasmodium vivax; B, B' and B'', Plasmodium malariae; C, C' and C'', Plasmodium falciparum).

A, B and C, mature parasites in red corpuscles.

A', B' and C', segmented parasites ready to leave corpuscles.

A'', B'' and C'', mature gameteocytes.

Figure 9.7: Three species of malaria parasites (*Plasmodium*) in red blood cells.

Animal parasites and human disease (1918). Chandler, A.C. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. Not in copyright.

Figure 9.8: Levels of heterozygosity at a set of microsatellite markers surrounding the dhfr gene in samples of drug-resistant malaria (*Plasmodium falciparum*) from Thailand. The dotted horizontal line gives the average level of heterozygosity found at these markers in a set of drug-resistant malaria; we take this background as our π_0 . The dashed line shows our fitted hitchhiking model from equation 9.6 with $\tau \approx 40$, fitted by non-linear least squares. The recombination rate in *P. falciparum* is $r_{BP} \approx 10^{-6}\text{bp}^{-1}$. Data from LetterSpace=10NASH *et al.* (2005). Code here.

5464 **Question 1.** LetterSpace=10VANĀŽT Hof *et al.* (2011) identified
 the genetic basis of melanism in the peppered moth (*Biston betularia*).
 This allele swept to fixation in northern parts of the UK; a classic
 5466 case of adaptation to industrial pollution (made famous by the work
 of LetterSpace=10KETTLEWELL, see LetterSpace=10MAJERUS (2009)
 5468 and LetterSpace=10COOK *et al.* (2012)). The genetic basis of melanism
 is a transposable element (TE) inserted into a pigmentation gene.
 5470 LetterSpace=10VANĀŽT Hof *et al.* found that diversity is suppressed
 in a broad region around the TE. Specifically, on the background of
 5472 the TE, it takes roughly 200 kb in either direction for diversity levels
 to recover to 50% of genome-wide levels.

5474 Random facts: In all moths and butterflies only males recombine;
 chromosomes are transmitted without recombination in females. The
 5476 recombination rate in males is 2.9 cM/Mb. Peppered moths have an
 effective population size of roughly a hundred thousand individuals.
 5478 Kettlewell used to eat moths when out collecting them in the field
 (personal communication, Art. Shapiro).

5480 **A)** Briefly explain how this pattern offers further evidence that the
 melanic allele was favoured by selection.

5482 **B)** Using this information, and assuming the allele's effects on
 fitness are additive, what is your estimate of the age of the allele?

5484 **C)** What is your estimate of the selection coefficient favouring this
 melanic allele?

5486 **Other signals of selective sweeps** The primary signal of a recently
 completed selective sweep is the characteristic reduction in diversity
 5488 surrounding the selected site. However, sweeps do leave other signals
 and these have also often been used to identify loci undergoing selec-
 5490 tion. For example, neutral alleles further away from the selected site
 may hitchhiking only part of the way to fixation if recombination oc-
 5492 curs during the sweep, which can lead to an excess of high-frequency
 derived alleles at intermediate distances away from the selected site,
 5494 a pattern lasting for a short time after a sweep (LetterSpace=10FAY
 and LetterSpace=10WU, 2000; LetterSpace=10PRZEWORSKI, 2002;
 5496 LetterSpace=10KIM, 2006). Also, as neutral diversity levels slowly
 recover through an influx of new mutations after a sweep, there is
 5498 a strong skew towards low frequency derived alleles, a pattern that
 persists for many generations (LetterSpace=10BRAVERMAN *et al.*, 1995;
 5500 LetterSpace=10PRZEWORSKI, 2002; LetterSpace=10KIM, 2006). The ex-
 cess of rare alleles, compared to a neutral model, can be captured by
 5502 statistics such as Tajima's D (which we encountered back in our dis-
 cussion of the neutral site frequency eqn 3.42). Thus one way to look



Figure 9.9: peppered moth (*Biston betularia*), non-melanic morph
Les papillons dans la nature (1934). Robert, P.-A. Image from the Biodiversity Heritage Library. Contributed by University of Illinois Urbana-Champaign. Not in copyright.

5504 for loci that have undergone selective sweeps is to calculate Tajima's
5505 D from data in windows along the genome and look for strong de-
partures from the null distribution.

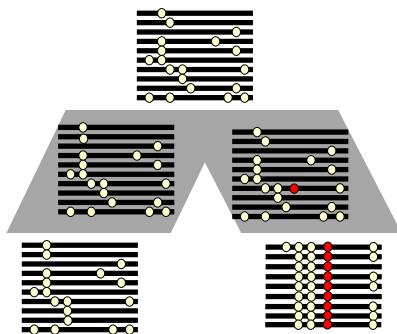


Figure 9.10: Two populations de-
scended from a common ancestral
population. A beneficial mutation has
occurred in population and swept to
fixation.

We can also use comparisons among multiple populations to look
5508 for evidence of sweeps occurring in one of the populations, for ex-
ample to identify alleles involved in local adaptation (see 9.11). A
5510 selective sweep will decrease the within-population diversity (H_S)
surrounding the selected site, without affecting the diversity between
5512 different populations. Thus local sweeps create peaks of F_{ST} between
weakly differentiated populations.

5514 LetterSpace=10HOHENLOHE *et al.* (2010) studied genome-wide pat-
terns of F_{ST} between marine and freshwater populations of threespine
5516 stickleback (*Gasterosteus aculeatus*). Between different marine pop-
ulations, they found no strong peaks of F_{ST} ; however, between the
5518 marine and freshwater comparisons they found a number of high F_{ST}
peaks that were replicated over a number of freshwater-marine com-
5520 parisons. They identified a number of novel regions responsible for
the adaptation of sticklebacks to freshwater environments and also a
5522 number of loci previously identified in crosses between marine and
freshwater populations. For example, the first peak of Linkage Group
5524 IV includes Ectodysplasin A (Eda), a gene involved in the adaptive
loss of armour plating in freshwater environments.

5526 *Soft Sweeps from multiple mutations and standing variation.* In our
sweep model above, we assumed that selection favoured a beneficial
5528 allele from the moment it entered the population as a single copy
mutation (left panel, Figure 9.12). However, when a novel selection
5530 pressure switches on, multiple mutations at the same gene may start
to sweep, such that no one of these alleles sweeps to fixation (middle
5532 panel, Figure 9.12). These sweeps involving multiple mutations sig-
nificantly soften the impact of selection on genomic diversity, and so
5534 are called 'soft sweeps'.

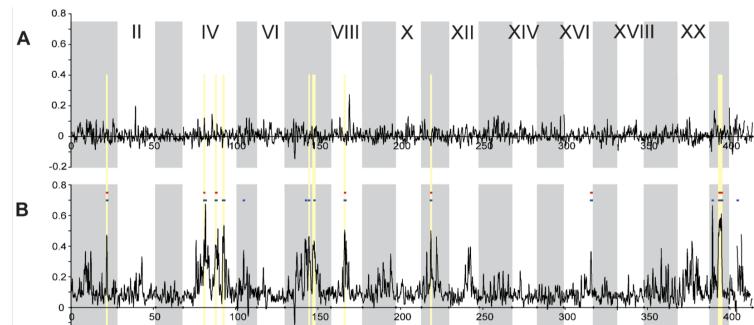


Figure 9.11: F_{ST} across the stickleback genome, with colored bars indicating significantly elevated ($p \leq 10^5$, blue; $p \leq 10^7$, red) and reduced ($p \leq 10^5$, green) values. The alternating white and grey panels indicate different linkage groups. **A)** F_{ST} between two oceanic populations **B)** Average F_{ST} between a freshwater population and the two marine populations. Figure and caption text from LetterSpace=10HOHENLOHE *et al.* (2010), licensed under CC BY 4.0.

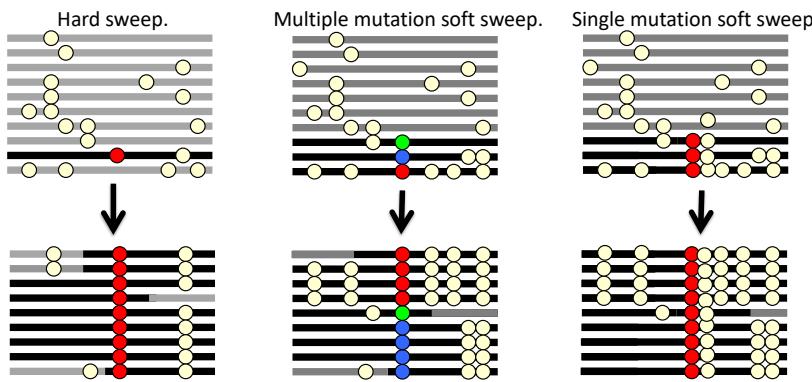


Figure 9.12: Three types of sweeps.

Another way that the impact of a sweep can be softened is if our allele was segregating in the population for some time before it became beneficial. That additional time means that our allele can have recombined onto various haplotype backgrounds, such that when selection pressures switch, the selected allele sweeps up in frequency on multiple different haplotypes (right panel, Figure 9.12). Detecting and differentiating these different types of sweeps is an active area of empirical research and theory in population genomics (see LetterSpace=10HERMISSON and LetterSpace=10PENNINGS (2017) for an overview of developments in this area).

9.1 The genome-wide effects of linked selection.

- To what extent are patterns of variation along the genome and among species shaped by linked selection, such as selective sweeps? We can hope to identify individual cases of strong selective sweeps along the genome, but how do they contribute to broader patterns of variation?

Two observations have puzzled population geneticists since the inception of molecular population genetics. The first is the relatively

high level of genetic variation observed in most obligately sexual species. The neutral theory of molecular evolution was developed in part to explain these high levels of diversity. As we saw in Chapter 3, under a simple neutral model, with constant population size, we should expect the amount of neutral genetic diversity to scale with the product of the population size and mutation rate. The second observation, however, is the relatively narrow range of polymorphism across species with vastly different census sizes (see Figure 2.2 and LetterSpace=10LEFFLER *et al.* (2012) for a recent review). As highlighted by LetterSpace=10LEWONTIN (1974) in his discussion of the paradox of variation, this observation seemingly contradicts the prediction of the neutral theory that genetic diversity should scale with the census population size. There are a number of explanations for the discrepancy between genetic diversity levels and census population sizes. The first is that the effective size of the population (N_e) is often much lower than the census size, due to high variance in reproductive success and frequent bottlenecks (as discussed in Chapter 3). The second major explanation, put forward by LetterSpace=10MAYNARD SMITH and LetterSpace=10HAIGH (1974), is that neutral levels of diversity are also systematically reduced by the effects of linked selection. In large populations, selective sweeps and other forms of linked selection may come to dominate over genetic drift as a source of stochasticity in allele frequencies, potentially establishing an upper limit to levels of diversity (LetterSpace=10KAPLAN *et al.*, 1989; LetterSpace=10GILLESPIE, 2000).

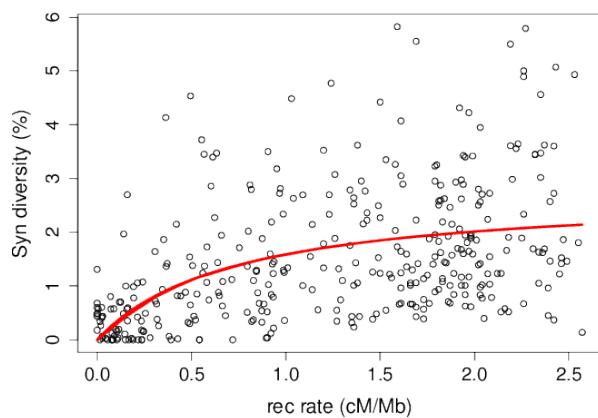


Figure 9.13: The relationship between (sex-averaged) recombination rate and synonymous site pairwise diversity (π) in *Drosophila melanogaster*. The curve is the predicted relationship between π and recombination rate, obtained by fitting the recurrent hitchhiking equation (9.13) to this data using non-linear least squares via the `nls()` function in R. Data from (LetterSpace=10SHAPIRO *et al.*, 2007), kindly provided by Peter Andolfatto, see LetterSpace=10SELLA *et al.* (2009) for details. Code here.

One strong line of evidence for the action of linked selection in reducing levels of polymorphism is the positive correlation between putatively neutral diversity and recombination seen in a

number of species, as, all else being equal, linked selection should
 5582 remove diversity more quickly in regions of low recombination
 (LetterSpace=10AGUADÉ *et al.*, 1989; LetterSpace=10BEGUN and
 5584 LetterSpace=10AQUADRO, 1992; LetterSpace=10WIEHE and Let-
 terSpace=10STEPHAN, 1993b; LetterSpace=10CUTTER and LetterSpace=10CHOI,
 5586 2010; LetterSpace=10CAI *et al.*, 2009). For example, *Drosophila melanogaster*
 diversity levels are much lower in genomic regions of low recombi-
 5588 nation (see Figure 9.13). This pattern can not be explained by differ-
 ences in mutation rate between low and high recombination regions
 5590 as this pattern is not seen strongly in divergence data among species.

These patterns could reflect the action of selective sweeps happen-
 5592 ing recurrently along the genome. In the next section we'll present
 a model for how levels of genetic diversity should depend on re-
 5594 combination and the density of functional sites under a model of
 recurrent selective sweeps. However, other forms of linked selec-
 5596 tion can impact genetic diversity in similar ways. For example,
 linked genetic diversity is continuously lost from natural popula-
 5598 tions due to the removal of haplotypes that carry deleterious alleles
 (LetterSpace=10CHARLESWORTH *et al.*, 1995; LetterSpace=10HUDSON
 5600 and LetterSpace=10KAPLAN, 1995b); this is called the 'background
 selection' model. Below we'll discuss the background selection model
 5602 and its basic predictions.

More generally, a wide range of models of selection predict the re-
 5604 moval of neutral diversity linked to selected sites. This is because the
 diversity-reducing effects of high variance in reproductive success are
 5606 compounded over the generations when there is heritable variance
 in fitness (LetterSpace=10ROBERTSON, 1961; LetterSpace=10SANTIAGO
 5608 and LetterSpace=10CABALLERO, 1995, 1998; LetterSpace=10BARTON,
 2000). Many different modes of linked selection likely contribute to
 5610 these genome-wide patterns of diversity; the present challenge is how
 to differentiate among these different modes.

5612 9.1.1 A simple recurrent model of selective sweeps

To explain how a constant influx of sweeps could impact levels of
 5614 diversity, here we will develop a model of recurrent selective sweeps.

Imagine we sample a pair of neutral alleles at a locus a genetic
 5616 distance r away from a locus where sweeps are initiated within the
 population at some very low rate v per generation. The waiting time
 5618 between sweeps at our locus is exponentially distributed $\sim \text{Exp}(v)$.
 Each sweep rapidly transits through the population in τ generations,
 5620 such that each sweep is finished long before the next sweep ($\tau \ll$
 $1/v$).

5622 As before, the chance that our neutral lineage fails to recombine

off the sweep is p_{NR} , such that the probability that our pair of lineages are forced to coalesce by a sweep is $e^{-r\tau}$. Our lineages therefore have a very low probability

$$\nu e^{-r\tau} \quad (9.8)$$

of being forced to coalesce by a sweep per generation. If our lineages do not coalesce due to a sweep, they coalesce at a neutral rate of $1/2N$ per generation. Thus the average waiting time till a coalescent event between our neutral pair of lineages due to either a sweep or a neutral coalescent event is

$$\mathbb{E}(T_2) = \frac{1}{\nu e^{-r\tau} + 1/2N} \quad (9.9)$$

Now imagine that the sweeps don't occur at a fixed location with respect to our locus of interest, but now occur uniformly at random across our genome. The sweeps are initiated at a very low rate of ν_{BP} per basepair per generation. The rate of coalescence due to sweeps at a locus ℓ basepairs away from our neutral loci is $\nu_{BP}e^{-r_{BP}\ell\tau}$. If our neutral locus is in the middle of a chromosome that stretches L basepairs in either direction, the total rate of sweeps per generation that could force our pair of lineages to coalesce is

$$2 \int_0^L \nu_{BP}e^{-r_{BP}\ell\tau} d\ell = \frac{2\nu_{BP}}{r_{BP}\tau} \left(1 - e^{-r_{BP}\tau L}\right) \quad (9.10)$$

so that if L is very large ($r_{BP}\tau L \gg 1$), the rate of coalescence per generation due to sweeps is $2\nu_{BP}/r_{BP}\tau$. The total rate of coalescence for a pair of lineages per generation is then

$$\frac{2\nu_{BP}}{r_{BP}\tau} + \frac{1}{2N} \quad (9.11)$$

So our average time till a pair of lineages coalesce is

$$\mathbb{E}(T_2) = \frac{1}{2\nu_{BP}/r_{BP}\tau + 1/2N} = \frac{r_{BP}2N}{4N\nu_{BP}/\tau + r_{BP}} \quad (9.12)$$

such that our expected pairwise diversity ($\pi = 2\mu\mathbb{E}(T_2)$) in a region with recombination rate r_{BP} that experiences sweeps at rate ν_{BP} is

$$\mathbb{E}(\pi) = \pi_0 \frac{r_{BP}}{4N\nu_{BP}/\tau + r_{BP}} \quad (9.13)$$

where π_0 is our expected diversity without any selective sweeps, ($p_{i0} = \theta = 4N\mu$). The expected diversity increases with r_{BP} , as higher recombination rates decrease the likelihood a neutral allele hitchhikes along with a sweep and is thus forced to coalesce by the sweep. Expected diversity decreases with ν_{BP} , as a greater density of functional sites experiencing sweeps increases the chance of being

linked to a nearby sweep. As we move to high r_{BP} , assuming that ν_{BP}
5652 doesn't increase with r_{BP} , our level of diversity should plateau to θ ,
the level of genetic diversity of a neutral site completely unlinked to
5654 any selected loci. If we assume that our genome experiences a con-
stant rate of sweeps of a given strength, i.e. that $4N\nu_{BP}/\tau$ is a constant,
5656 we can fit the variation in π across regions that vary in their recom-
bination rate (r_{BP}) to estimate a population's rate of recurrent sweeps
5658 per basepair. An example of fitting this curve to data from *Drosophila*
melanogaster is shown in Figure 9.13; see LetterSpace=10WIEHE and
5660 LetterSpace=10STEPHAN (1993a) for an early example of fitting a sim-
ilar recurrent hitchhiking model to such data. The parameter giving
5662 us this best-fitting curve is $4N\nu_{BP}/\tau \approx 7 \times 10^{-9}$. With an effect popula-
tion size of a million and assuming that the sweeps take a thousand
5664 generations to reach fixation, we find this implies $\nu_{BP} \approx 10^{-12}$. Thus,
a really low rate of moderately strong sweeps, roughly one every
5666 megabase every million generations, is all we need to explain the
profound dip in diversity seen in regions of the genome with low re-
5668 combination. However, sweeps from positively selected alleles are not
the only cause of genome-wide signals of linked selection. Selection
5670 against deleterious alleles can also drive these patterns.

9.1.2 Background selection

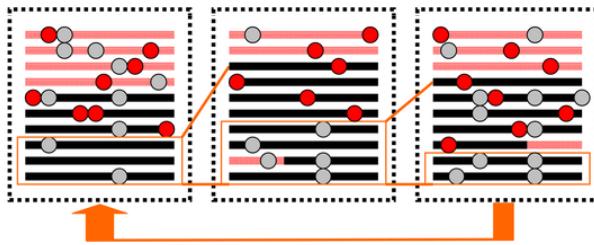
5672 Populations experience a constant influx of deleterious mutations
at functional loci while selection acts to purge them from the pop-
5674 ulation, thus preventing deleterious substitutions and maintaining
function at these loci. As we discussed in Chapter 6, this balance
5676 between mutation and selection results in a constant level of dele-
terious variation in the population. The constant selection against
5678 this deleterious variation has effects on diversity at linked sites. Each
deleterious mutation arises at random on a haplotype in the popu-
5680 lation, and as selection purges this mutation, it removes with it any
neutral alleles that were also on this haplotype. This constant re-
5682 moval of linked alleles from the population acts to reduce diversity
in regions surrounding functional loci (LetterSpace=10HUDSON and
5684 LetterSpace=10KAPLAN, 1995a; LetterSpace=10NORDBORG *et al.*, 1996),
an effect known as background selection (BGS).

5686 What proportion of our haplotypes are free of deleterious mu-
tations in any given generation, and so free to contribute to future
5688 generations? Well, under mutation-selection balance, a constrained
locus with a mutation rate μ towards deleterious alleles that expe-
5690 rience a selection coefficient sh against them in heterozygotes, will
result in μ/sh chromosomes carrying the deleterious allele. Some of
5692 these haplotypes may be passed on to the next generation, but if they

are fully linked to the deleterious locus they will all eventually be lost because they carry a deleterious mutation at a site under constraint. Thus, for a neutral polymorphism completely linked to a constrained locus, only $2N(1 - \mu/sh)$ alleles get to contribute to future generations. Therefore, the level of pairwise diversity in a constant population due to BGS at such a locus will be

$$\mathbb{E}[\pi] = 2\mu \times 2N(1 - \mu/sh) = \pi_0(1 - \mu/sh) \quad (9.14)$$

where $\pi_0 = 4N\mu$, the level of neutral pairwise diversity in the absence of linked selection.



The effects of background selection are more pronounced in regions of low recombination, where neutral alleles are less able to recombine off the background of deleterious alleles. Thus, under background selection, we also expect to see reduced diversity in regions of lower recombination.

For a neutral locus that is a recombination fraction r away from a locus subject to constraint, the level of diversity is

$$\mathbb{E}[\pi] = \pi_0 \left(1 - \frac{\mu sh}{2(r + sh)^2} \right) \quad (9.15)$$

As we move away from a locus experiencing purifying selection, we increase r , and diversity should recover. For example, moving away from genic regions in the maize genome we see the average level of diversity recover. This occurs in both maize and teosinte, the wild progenitor of maize. The dip in diversity around non-synonymous sites is stronger in teosinte, perhaps because the accelerated drift due to the bottleneck in maize may have somewhat released constraint on sites where very weakly deleterious alleles segregated previously at mutation-selection balance.

More generally, if a neutral locus is surrounded by L loci experiencing purifying selection at recombination distances r_1, \dots, r_L , then compounding equation (9.16) across these loci, the expected reduced diversity is approximately

$$\mathbb{E}[\pi] = \pi_0 \prod_{i=1}^L \left(1 - \frac{\mu sh}{2(r_i + sh)^2} \right) \approx \exp \left(\sum_{i=1}^L \frac{\mu sh}{2(r_i + sh)^2} \right) \quad (9.16)$$

Figure 9.14: A cartoon depiction of a region for 10 haplotypes experiencing background selection. Neutral mutations are shown as gray circles, and deleterious mutations in red. Over time, chromosomes carrying deleterious mutations are removed from the population, such that most individuals are descended from a subset of chromosomes free of deleterious alleles (highlighted here by orange boxes). Mutation is constantly generating new deleterious alleles on the background of chromosomes previously free of deleterious alleles. Figure modified from LetterSpace=10SELLA *et al.* (2009), licensed under CC BY 4.0.

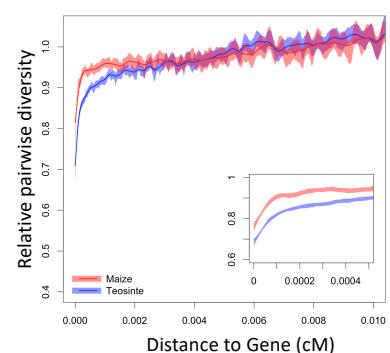


Figure 9.15: Relative diversity compared to the mean diversity in windows ≥ 0.01 cM as a function of the distance to the nearest gene. See (LetterSpace=10BEISSINGER *et al.*, 2016) for details. Figure licensed under CC BY 4.0 by Jeff Ross-Ibarra.

To model an average neutral locus in a genomic region with a given recombination rate, we can imagine that our neutral locus is situated in the center of a large region with total recombination rate R and total deleterious mutation rate U , where $U = \mu L$. Then our expression for diversity, equation (9.16), simplifies to

$$\mathbb{E}[\pi] \approx \pi_0 \exp(-U/(sh+R)) \approx \pi_0 \exp(-U/R). \quad (9.17)$$

In this last approximation, we assume that we're looking at a large region, with $R \gg sh$. Note that much like genetic load, equation (7.8), this expression depends only on the total deleterious mutation rate. Any dependence on the selection coefficient drops out, as weakly selected mutations segregate in the population at higher frequencies, but are also removed from the population more slowly, allowing more of the genome to recombine off the deleterious background.

For a first go at fitting this to genome-wide data, we could look at diversity in windows of length W bp (as in Figure 9.16). If we assume that there is a constant rate of deleterious mutation per base pair, μ_{BP} , then $U = \mu_{BP}W$. Furthermore, if our genomic window has a recombination rate r_{BP} per base-pair, our total genetic length is $R = r_{BP}W$. Making these substitutions in equation (9.17), our window size cancels out to give

$$\mathbb{E}[\pi] \approx \pi_0 \exp(-\mu_{BP}/r_{bp}) \quad (9.18)$$

Looking across windows that vary in their recombination rate, i.e. r_{BP} , we can fit equation (9.18) to data to estimate μ_{BP} . An example of doing this to data from *D. melanogaster* is shown in Figure 9.16, yielding an estimate of the deleterious mutation rate of $\mu_{BP} \approx 3.2 \times 10^{-9}$. This is roughly on the same order as the mutation rate per base pair in *D. melanogaster*, and so this deleterious mutation rate estimate is somewhat high as it would require most of the genome to be constrained, but as a first approximation it's not terrible. Note how similar the fit is to a model of hitchhiking, suggesting that both BGS and hitchhiking are capable of explaining the broad relationship between diversity and recombination seen in *D. melanogaster* and other species.

As our annotations of functional regions of the genome have improved, so have our methods to infer background selection. A more rigorous version of this analysis today would incorporate variation in coding density among windows into the parameter μ_{BP} . With detailed genomic annotations showing coding regions and constrained non-coding regions, we can also move beyond just analyzing broad-scale patterns. For example, LetterSpace=10McVICER *et al.* (2009) fit a model of background selection to putatively neutral pairwise

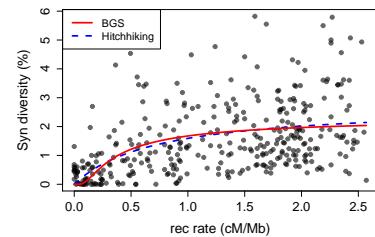


Figure 9.16: The relationship between recombination rate and synonymous site pairwise diversity (π) in *D. melanogaster*, as in Figure 9.13. The red curve is the predicted relationship between π and recombination rate, obtained by fitting the BGS equation (9.17) to this data using non-linear least squares via the `nls()` function in R. The blue line is the recurrent hitchhiking equation line from Figure 9.13. Code here.

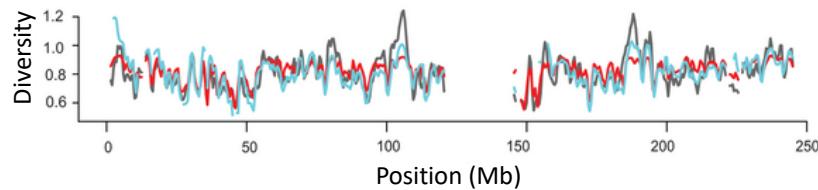


Figure 9.17: Observed (black line) and predicted pairwise diversity across chromosome 1, from a background selection model that assumes a uniform mutation rate (red line) or a mutation rate that varies with local human/dog divergence (blue line). Figure from (LetterSpace=10McVICER *et al.*, 2009), licensed under CC BY 4.0.

diversity along the human genome, using equation 9.16 to estimate
 5762 the effect of BGS at each locus, weighing the genetic distance to all
 of the surrounding coding regions and constrained non-coding sites.
 5764 This allowed LetterSpace=10McVICER *et al.* (2009) to estimate mutation
 rates and average selection coefficients acting against deleterious
 5766 alleles in these regions of the genome. This best fitting model also al-
 lowed them to predict diversity levels along the genome, a section of
 5768 which is shown in figure 9.17. Thus, broad-scale features of polymor-
 phism along the genome are well described by background selection
 5770 (or by linked selection more generally).

The deleterious mutation rates estimated by LetterSpace=10McVICER
 5772 *et al.* (2009) from fitting a model of BGS were again too high, as in the
Drosophila example above, suggesting the BGS alone is not sufficient
 5774 to explain all of the effect of linked selection. But how then do we go
 about distinguishing the impact of BGS from hitchhiking?

5776 *Distinguishing the impact of hitchhiking from background selection in*
genome-wide data A variety of approaches have been taken to start to
 5778 separate the effects of hitchhiking from background selection. Much
 of the strongest evidence showing the effects of both comes from
 5780 *Drosophila melanogaster* and we review some of that evidence here.
 Hitchhiking is expected to have systematic effects on the neutral site
 5782 frequency spectrum, distorting it towards rare minor alleles, (reflect-
 ing the slow recovery of diversity following a sweep). Therefore, we
 5784 should expect a distortion of summary statistics such as Tajima's D
 in regions of low recombination if hitchhiking is contributing to the
 5786 reduction in diversity in these regions (LetterSpace=10BRAVERMAN
et al., 1995; LetterSpace=10PRZEWORSKI, 2002; LetterSpace=10KIM,
 5788 2006). In *D. melanogaster*, there is a greater skew towards rare alle-
 les at putatively neutral sites in regions of low recombination (Let-
 5790 terSpace=10ANDOLFATTO and LetterSpace=10PRZEWORSKI, 2001; Let-
 terSpace=10SHAPIRO *et al.*, 2007), see left panel of Figure 9.18. How-
 5792 ever, while this skew isn't expected under simple models of strong
 background selection, other models of background selection can lead
 5794 to such patterns.

Another prediction of the hitchhiking model, where an allele

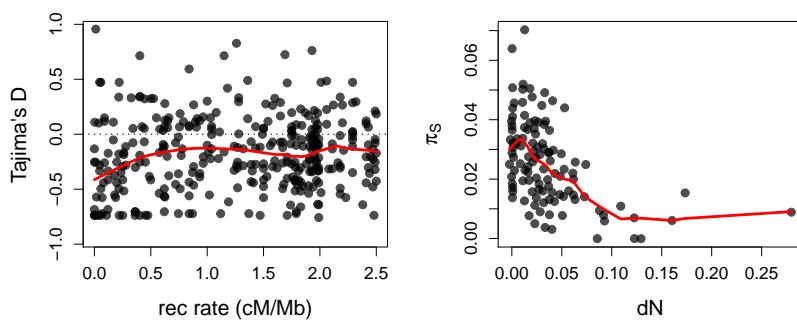


Figure 9.18: **Left)** Average Tajima's D in genomic windows plotted against their recombination rate in *D. melanogaster*. Data from LetterSpace=10SHAPIRO *et al.* (2007). **Right)** Synonymous pairwise diversity in genomic windows as a function of the density of non-synonymous substitutions in the window. Data from LetterSpace=10ANDOLFATTO (2007). Code here.

5796 sweeps to fixation, is that there should be a functional substitution
 5798 associated with each sweep. Or, to flip that around, we might ex-
 pect to see a greater impact of hitchhiking where there are more
 5800 functional substitutions. For example, regions surrounding non-
 synonymous substitutions should have lower levels of diversity, if a
 5802 high fraction of non-synonymous substitutions are adaptive. Again,
 5804 this pattern is seen in *D. melanogaster* (LetterSpace=10ANDOLFATTO,
 2007; LetterSpace=10MACPHERSON *et al.*, 2007; LetterSpace=10SATTATH
et al., 2011b), right side of Figure 9.18.

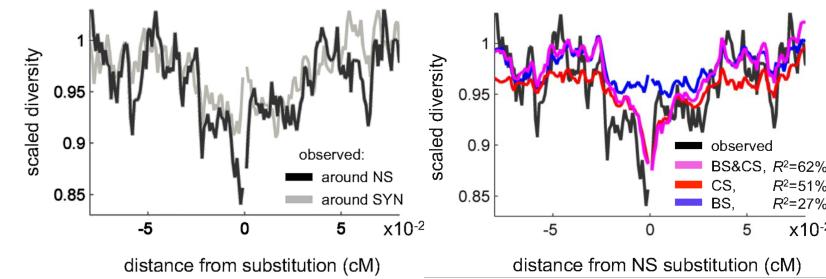


Figure 9.19: **Left)** Scaled synonymous pairwise diversity levels around non-synonymous (NS) and synonymous (SYN) substitutions in *D. melanogaster*. **Right)** Predicted scaled diversity levels around non-synonymous substitutions based on models including background selection (BS), classic sweeps (CS) and both (BS & CS). Figure from LetterSpace=10ELYASHIV *et al.* (2016), licensed under CC BY 4.0.

Pushing this idea further, we can look at the dip in diversity sur-
 5806 rounding a non-synonymous substitution averaged across all the sub-
 stitutions in the genome. LetterSpace=10ELYASHIV *et al.* (2016) found
 5808 a stronger dip in diversity around non-synonymous substitutions
 than synonymous substitutions (see also LetterSpace=10SATTATH
 5810 *et al.*, 2011a). Extending the model of LetterSpace=10McVICER *et al.*
 (2009) to fit a model of background selection and hitchhiking to pu-
 5812 tative neutral diversity along the genome, they found that the dip in
 diversity around synonymous substitutions comes mostly from BGS.
 5814 But to fully explain the dip in diversity around non-synonymous
 substitutions, a reasonable proportion of these non-synonymous sub-
 5816 stitutions have to have been accompanied by a classic (hard) sweep.

The majority of these sweeps are estimated to be due to very weak
5818 selection, with selection coefficients $< 10^{-4}$. Furthermore, Let-
terSpace=10ELYASHIV *et al.* (2016) estimated a 77 - 89% reduction
5820 in neutral diversity due to selection on linked sites, and concluded
that no genomic window was entirely free of the effects of selection.
5822 Thus linked selection has a profound effect in some species such as
Drosophila melanogaster.

Interaction of Multiple Selected Loci

- 5826 Consider two biallelic loci segregating for A/a and B/b . There are
 four haplotypes, AB , Ab , aB , ab , which for simplicity we label 1-4.
 5828 The frequency of our four haplotypes are x_1 , x_2 , x_3 , and x_4 . Each in-
 dividual has a genotype consisting of two haplotypes; we label w_{ij}
 5830 the fitness of an individual with the genotype made up of haplotype
 i and j (we assume that $w_{ij} = w_{ji}$, i.e. there are no parent of origin
 5832 effects). Assuming that these fitnesses reflect differences due to via-
 bility selection, and that individuals mate at random, we can write
 5834 the following table of our genotype proportions after selection:

	AB	Ab	aB	ab
AB	$w_{11}x_1^2$	$w_{12}2x_1x_2$	$w_{13}2x_1x_3$	$w_{14}2x_1x_4$
Ab	•	$w_{22}x_2^2$	$w_{23}2x_2x_3$	$w_{24}2x_2x_4$
aB	•	•	$w_{33}x_3^2$	$w_{34}2x_3x_4$
ab	•	•	•	$w_{44}x_4^2$

- 5836 This follows from assuming that our haplotypes are brought together
 at random (HWE), then discounted by their fitnesses. Our mean
 5838 fitness \bar{w} is the sum of all the entries in the table, so dividing by \bar{w}
 normalizes the complete table to sum to one. The frequency of the
 5840 AB haplotype (1) in the next generation of gametes is

$$x'_1 = \frac{(w_{11}x_1^2 + \frac{1}{2}w_{12}2x_1x_2 + \frac{1}{2}w_{13}2x_1x_3 + \frac{1}{2}(1-r)w_{14}2x_1x_4 + \frac{1}{2}rw_{23}2x_2x_3)}{\bar{w}} \quad (10.1)$$

- This is a bit of a mouthful, but each of the terms is easy to un-
 5842 derstand. Each of the HWE genotype frequencies (e.g. $2x_1x_2$) is
 weighted by its fitness relative to the mean fitness (w_{ij}/\bar{w}), and by
 5844 its probability of transmitting the AB haplotype to the next gen-
 eration. For example, AB/Ab individuals ($1/2$) transmit the AB
 5846 haplotype only half the time. The final two terms include the recom-
 bination fraction (r). The first term involving recombination refers to
 5848 the AB/ab genotype ($1/4$), who with probability $(1-r)/2$ transmits
 a non-recombinant AB haplotype to the gamete. Similarly, the second

5850 term refers to the Ab/aB genotype; a proportion $r/2$ of its gametes carry the recombinant AB haplotype.

5852 In the single locus case, we defined the marginal fitness of an allele. Here it will help us to define the marginal fitness of the i^{th} 5854 haplotype:

$$\bar{w}_i = \sum_{j=1}^4 w_{ij}x_j \quad (10.2)$$

This is the fitness of the i^{th} haplotype averaged over all of the *diploid* 5856 genotypes it could occur in, weighted by their probability under random mating. Using this notation, and with some rearrangement 5858 of equation (10.1), we obtain

$$x'_1 = \frac{x_1\bar{w}_1 - w_{14}rD}{\bar{w}} \quad (10.3)$$

Here we have assumed that $w_{23} = w_{14}$, i.e. that the fitness of AB/ab 5860 individuals is the same as Ab/aB individuals (i.e. that fitness depends only on the alleles carried by an individual, and not on which 5862 chromosome they are carried; this assumption is sometimes called no *cis*-epistasis).

5864 We can then write the change in the frequency of our 1 haplotype as

$$\Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}} \quad (10.4)$$

5866 Generalizing this result, we write the change in *any haplotype i from* our set of four haplotypes as

$$\Delta x_i = \frac{x_i(\bar{w}_i - \bar{w}) \pm rw_{14}D}{\bar{w}} \quad (10.5)$$

5868 where the coupling haplotypes 1 and 4 use $+D$ and repulsion haplotypes 2 and 3 use $-D$. Note that the sum of these four Δx_i is zero, as 5870 our haplotype frequencies sum to one.

So the change in the frequency of a haplotype (e.g. AB, haplotype 5872 1) is determined by the interplay of two factors: First, the extent 5874 to which the marginal fitness of our haplotype is higher (or lower) than the mean fitness of the population (the magnitude and sign 5876 of $(\bar{w}_1 - \bar{w})/\bar{w}$). Second, whether there is a deficit or any excess of our haplotype compared to linkage equilibrium (the magnitude and 5878 sign of D), modified by the strength of recombination. This tension between selection promoting particular haplotypic combinations, and recombination breaking up overly common haplotypes is the key to a 5880 lot of interesting dynamics and evolutionary processes.

10.1 Types of interaction between selection and recombination

5882 To illustrate these ideas we make use of Muller diagrams (LetterSpace=10MULLER, 1932), where we visualize the allele dynamics in
5884 terms of a plot of the stack frequencies over time.

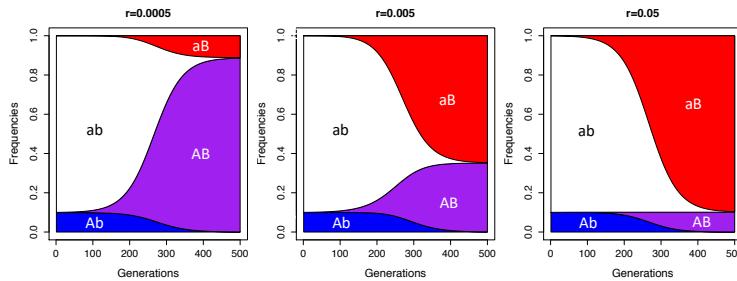
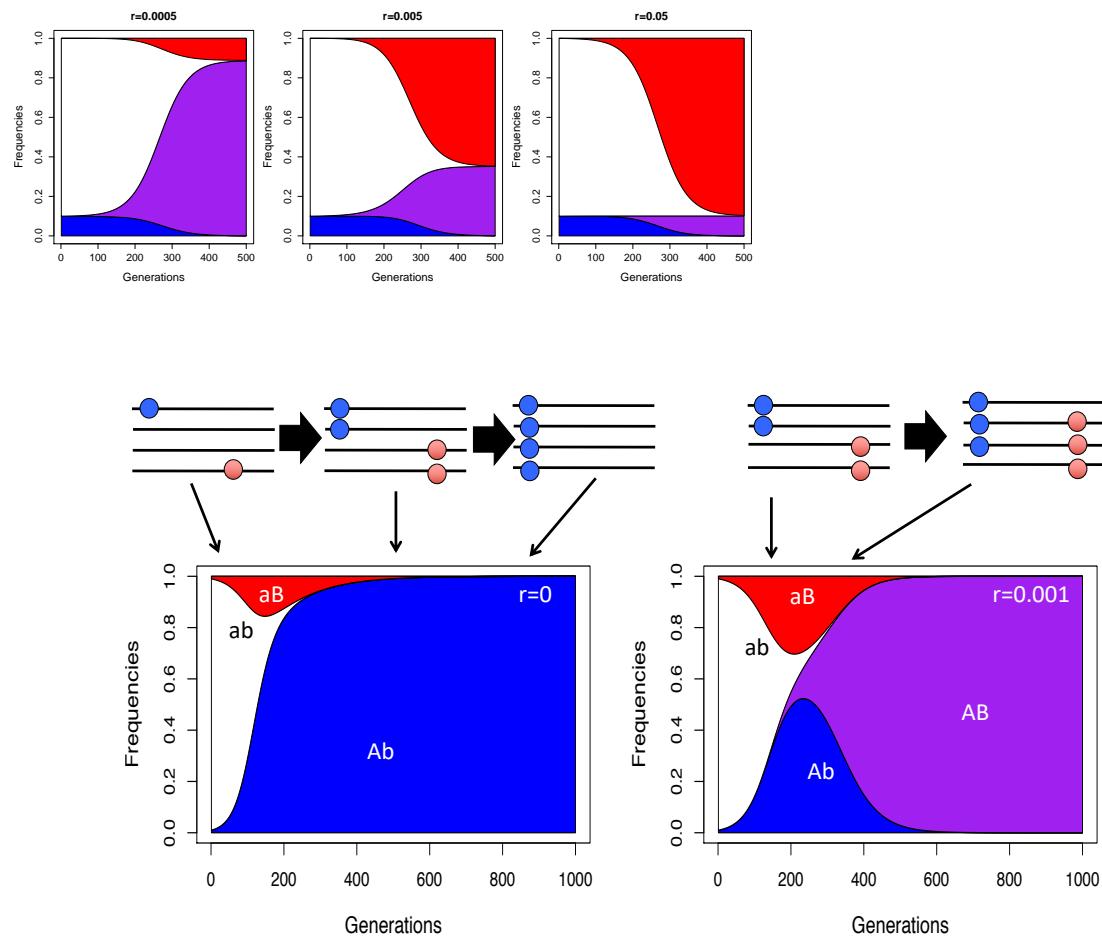


Figure 10.1: A beneficial mutation B arises on the background of a neutral allele whose initial frequency is $p_A = 10\%$. The beneficial allele has a strong, additive selection coefficient of $hs = 0.05$.

5886 *The hitchhiking of deleterious alleles* Let's start by revisiting our neutral hitchhiking in this two locus setting in the previous chapter we saw that neutral alleles can hitchhike along with our selected allele
5888 if they are tightly linked enough. Figure 10.1 shows the frequency trajectories of the various haplotypes for neutral allele (A) that is
5890 present at 10% frequency in the population when our beneficial allele
5892 (B) arises on its background. When the recombination rate (r) is low
5894 between the loci, A gets to hitchhike to high frequency, but for higher
5896 recombination rates it only gets dragged to intermediate frequencies.
5898 For the highest recombination rate shown ($r \approx s$) the neutral allele's
5900 dynamics ($p_{Ab} + p_{AB}$) are barely changed at all, as it recombines on
5902 and off the sweeping allele frequently and so barely perceives the
5904 sweep.

5906 *The hitchhiking of deleterious alleles* Deleterious alleles can also hitchhike
5908 along with beneficial mutations if they are not too deleterious com-
5910 pared to the benefits offered by the selected allele. Again our allele
5912 A is at 10% frequency in the population in Figure ??, but this time it
is deleterious and so initially decreasing in frequency across the genera-
tions when the beneficial mutation (B) arises on its background.
5914 If the loci are tightly linked, and A were too deleterious, B would
5916 never get to take off in the population. However, if the benefits of
5918 B outweighs the cost of A , even in the case of no recombination be-
5920 tween our loci, allele A gets to hitchhike to fixation and merely slows
5922 down B 's rate of increase and their combined fitness is reduced. With
5924 moderate amounts of recombination between the loci, our deleterious
5926 starts to hitchhike but before it can get to fixation the beneficial allele
5928 manages to recombine off its background. This recombinant aB hap-
5930 lotype, which has higher fitness as it lacks the deleterious allele, now
5932 sweeps through the population displacing the AB haplotype. For

higher recombination events we have to wait less long for a recombination to breakup the hitchhiking deleterious allele, so the adaptive allele easily escapes its background. For the purposes of illustration here we've used a relatively common deleterious allele, but in reality these alleles will likely be often be rare in the population and at mutation selection balance. If they are rare it is likely that a beneficial mutation arises on a specific deleterious allele's background, but as we have seen there are likely going to be many rare deleterious alleles in the population so it is likely that a beneficial mutations may often have to contend with deleterious hitchhikers.



Clonal interference between favourable alleles. When rates of sex and recombination are zero, or very low, positively selected alleles can prevent each other reach fixation and so the rate of adaptation can be slowed. In the absence of sex and recombination, when two positively selected alleles arise on different genetic backgrounds in the

Figure 10.2: Interference between two positively selected alleles. **Left)** the red and blue (A and B) beneficial alleles arise on different haplotypes. They rise in frequency, but in the absence of recombination only one can fix. This is shown in a Muller diagram, where p_{AB} is initially set to zero. **Right)** In the presence of recombination the population can generate the recombinant (AB) haplotype, which can subsequently fix.

population they cannot both fix (left side of Figure 10.2). They can initially increase in frequency, but necessarily compete with each other when they become common. This is called selective interference, or sometime clonal interference. If one of the alleles has a much larger selection coefficient it will fix, forcing the other allele from the population, but when they are relatively equally matched it may take some time for this situation to resolve itself resulting in a traffic jam in the population. Thus in an asexual adaptive alleles necessarily have to fix sequentially. However, with even a small amount of recombination beneficial alleles can recombine on to each others background, allowing them to fix in parallel (right side of Figure 10.2).

Given the rapid evolution of HIV we can see interference taking place over very short time periods indeed. HIV uses its reverse transcriptase (RT) gene to write itself from an RNA virus into its host's DNA, allowing HIV to hijack the hosts regulatory machinery, a critical part of its life cycle. One of the early HIV drugs was Efavirenz, which inhibits HIV's RT protein. Sadly, mutations are common in the RT HIV gene, and these mutations, in the presence of the drug, confer a profound fitness advantage, allowing them to spread through the HIV population in patients undergoing anti-HIV treatment. In Figure 10.3 we see that by day 224 after the start of drug treatment two different drug-resistance amino-acid changes beginning to spread within a patient (also shown as a Muller diagram in Figure 10.4). Because these alleles occur on different genetic backgrounds, with little chance for genetic exchange between them, they interfere in each other progress as they compete to fix within the population. Eventually the amino acid change at site 188 wins out.

Figure 10.3: HIV sequences from a patient over the course of drug treatment in the retrotransposase coding region. Figure cropped from LetterSpace=10WILLIAMS and LetterSpace=10PENNINGS (2019), licensed under CC BY 4.0.

An example of the costs of asexuality.

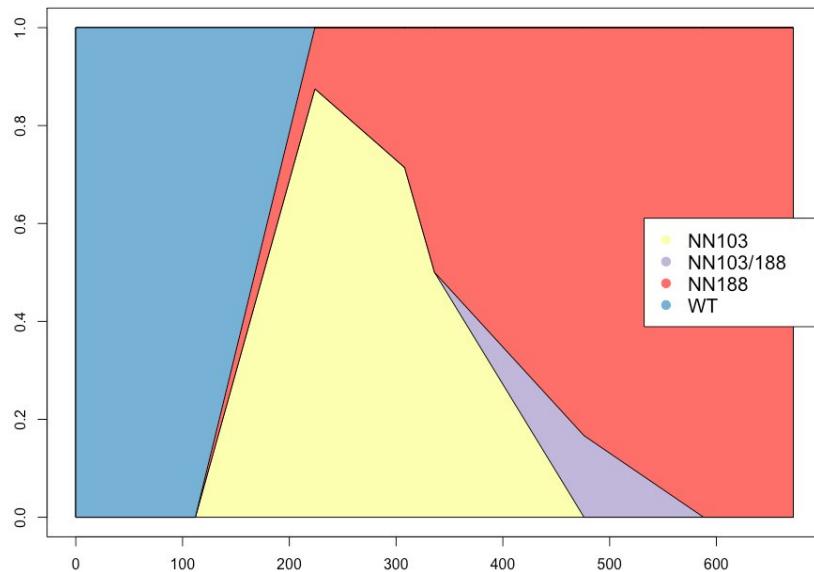


Figure 10.4: Muller plot of the drug resistance interference dynamics from Figure 10.3. Figure from LetterSpace=10WILLIAMS and LetterSpace=10PENNINGS (2019), licensed under CC BY 4.0.

5958 In the Evening primrose genus (*Oenothera*), there are a number of
 5959 young, independently-derived, asexual species. In each species this
 5960 asexuality is due to a complicated series of reciprocal translocations
 5961 which prevent recombination and segregation and ensure that every
 5962 plant is permanently-heterozygote for these rearrangements due
 5963 to lethality. This system is quite complicated, and super cool. We
 5964 don't need to worry about the details but importantly each species is
 5965 functionally asexual. LetterSpace=10HOLLISTER *et al.* (2014) sampled
 5966 transcriptome data from across the Evening primrose clade, and took
 5967 advantage of 7 independent, asexual-sexual sister pairs of species
 5968 to examine the impact of the evolution of asexuality for molecular
 5969 evolution.

5970 The d_N/d_S for the sexual and asexual species for each of the seven
 5971 pairs (C1-C7) is shown in Figure 10.6. In every pair d_N/d_S is higher in
 5972 the asexual species. The genomes of the asexual species are evolving
 5973 in a less constrained fashion, likely due to weakly deleterious muta-
 5974 tions accumulating due to hitchhiking with beneficial alleles and the
 slow crank of Muller's ratchet.

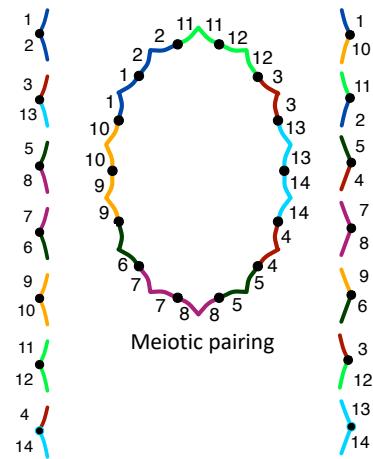


Figure 10.5: A schematic diagram of the karotype of an evening primrose. The two columns show a heterozygote individual's diploid chromosomal complement. Each chromosome is heterozygote for two different translocations. For example both the top-most chromosomes has one arm from chromosome 1, but the other arm is heterozygote for a large translocation from the ancestral chromosome 2 and 10. Due to these translocations the meiotic pairing form a complete ring of chromosomes, which prevent crossing over and independent segregation. Thanks to Jesse Hollister for this image.

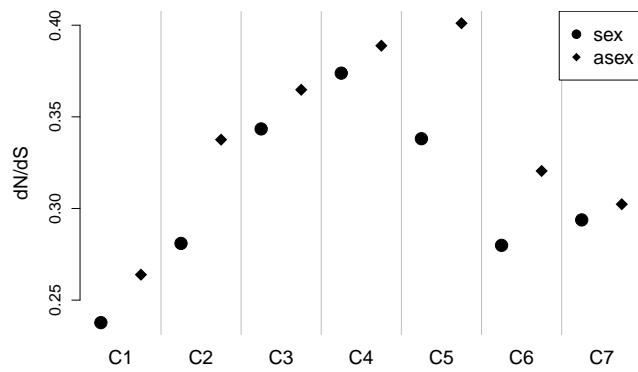


Figure 10.6: d_N/d_S calculated on sexual (circles) and asexual (diamonds) lineages of each of seven sister pairs of species. Data from LetterSpace=10 HOLLISTER *et al.* (2014). Code here. ↴

5976 *The maintainance of combinations of alleles in the face of recombination.*
 In some cases balancing selection may be attempting to maintain
 5978 multiple combinations of alleles in the population that work well
 together. However, recombination may be constantly ripping those
 5980 alleles away from each other making it difficult to maintain these alleles.
 This can select for the suppression of recombination. Some of the
 5982 most dramatic demonstrations of this tension involve the evolution of
 so-called super genes. We'll first consider the evolution of a mimicry
 5984 supergene in *Heliconius numata* as an example of this.

Some of the most spectacular examples of Müllerian mimicry in
 5986 the world are found in *Heliconius* butterflies. These butterflies are
 unpalatable to predators, and different species mimic each other so
 5988 benefiting from not being eaten by predators, which rapidly learn to
 avoid all these species). In many of these species multiple mimicry
 5990 morphs are found as we move across geographic space. In *Heliconius*
numata a number of different morphs mimic morphs from a distantly
 5992 related *Melinaea* species.

To keep things relatively simple lets focus on two differences between *silvana* and *bicoloratus*, the yellow stripe on the top wing of *silvana* and the black bottom wing of *bicoloratus*. Lets imagine that these
 5994 two differences are due to a simple two locus system. The first locus
 segregates for Y/y, where the Y allele encodes for a top-wing yellow
 5996 band, and y encodes for the absence of the yellow band. The second
 locus segregates for B/b where B encodes for the bottom-wing being
 5998 black, and b for the absence of black on the bottom wing. If Y is re-
 cessive and B is dominant, then the *silvana* phenotype corresponds
 6000 to a YY bb genotype. Due to the dominance of the y and B alleles the
 6002

illustration_images/multiple_sel_loci/Evening_prime

Figure 10.7: Showy evening primrose (*Oenothera speciosa*), the sexual species in the clade C2 from Figure 10.6.
 Favourite flowers of garden and greenhouse (1896). Step E.
 Image from the Biodiversity Heritage Library. Contributed by Missouri Botanical Garden. Licensed under CC BY-2.0.



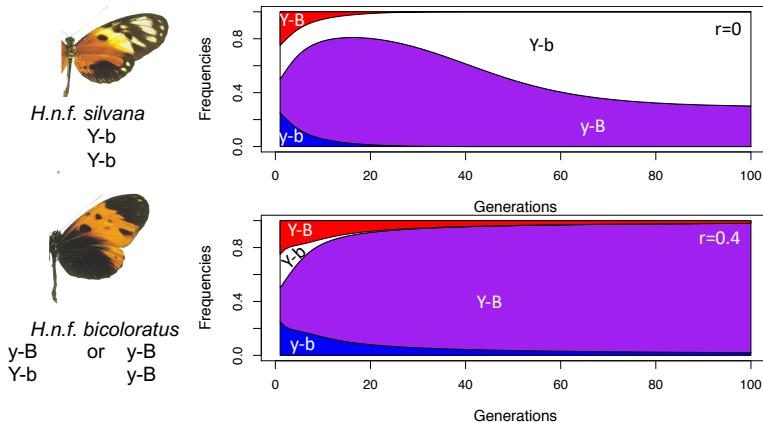
Figure 10.8: Five sympatric forms of *H. numata* from northern Peru, and their distantly related comimetic *Melinaea* species. First row: *M. menophilus* ssp. nov., *M. ludovica ludovica*, *M. marsaeus rileyi*, *M. marsaeus mothone*, and *M. marsaeus phasiana*. Second row, *H. n. f. tarapotensis*, *H. n. f. silvana*, *H. n. f. aurora*, *H. n. f. bicoloratus*, and *H. n. f. arcuella*. Figure and caption from LetterSpace=10JONON et al. (2006) cropped, licensed under CC BY 4.0.

bicoloratus phenotype can be achieved by various genotypes ($Yy Bb$, $yy BB$, $Yy BB$, $yy Bb$). Both of these phenotypes offer an advantage as they mimic a *M. menophilus* model. But there are also genotypes that don't do as well; $YY BB$ individuals have a yellow band and a black bottom and so don't do a great job mimicing anything and so will be eaten. Thinking about the four possible haplotypes, $y-B$ has high marginal fitness as due to its combo of dominant alleles it'll always produce a bicoloratus phenotype. Likewise the $Y-b$ haplotype has high marginal fitness, as it does well in the homozygous state (*silvana* phenotype), and when it is paired with the $y-B$ allele. However, the $Y-B$ and $y-b$ haplotypes fair less well as they carry two alleles that don't work well with each other and so are often individuals who suffer high rates of predation.

If no recombination occurs between these loci ($r = 0$, Figure 10.9), then the $Y-B$ and $y-b$ are selected out of the population, and the $y-B$ and $Y-b$ can be stably maintained. However, when there's too much recombination between our loci (e.g. $r = 0.4$, Figure 10.9) the high-fitness haplotypes keep getting ripped apart by recombination and the $Y-b$ is lost from the population as it's recessive advantage is lost as it's too often being broken up by recombination in heterozygotes.

Supergenes to the rescue! So our polymorphisms can only be maintained if they are tightly linked, i.e. if these alleles arose at loci that are genetically close to each other. But how is it possible that these alleles arose close to each other? Well the trick is that they don't necessarily have to arise very close to each other. If such a system is polymorphic but being regularly broken up by recombination, a chromosomal inversion—the flipping around of a whole section of chromosome—can arise and will suppress recombination. Imagine that our two loci are far apart genetically, and a chromosomal inversion arises on the

Figure 10.9:



Y-b background forming the b-Y haplotype. This inverted haplotype will not recombine with the y-B haplotype when it is present in a heterozygote, thus it is not broken down by recombination. This inverted haplotype, which enjoys the fitness benefits of the Y-b, can therefore replace the Y-b haplotype in the population. The two other low fitness haplotypes will disappear as they are no longer being generated by recombination, leaving just the y-B and b-Y. The polymorphism system now behaves like alleles at a single locus, a super gene (e.g. like $r = 0$ in Figure 10.9).

Now the *H. numata* system is vastly more complicated than our toy two locus system, presumably involving many changes and refinements, but the same principle holds (LetterSpace=10JORON *et al.*, 2011). The differences between the different *H. numata* mimics morphs is found on a single chromosome, and the inheritance behaves as if controlled by a single locus (albeit with many alleles). The *H. n. f. silvana* individuals carry a recessive haplotype of alleles that which is known to be locked together by a $\sim 400\text{kb}$ inversion, that is a different chromosomal orientation from the *bicoloratus* allele (haplotype) which acts as a dominant allele. Other alleles at this same chromosomal region provide the genetic basis of the other morphs, and sometimes correspond to further inversions with a range of dominance relationships.

"coadapted combinations of several or many genes locked in inverted sections of chromosomes and therefore inherited as single units."
 LetterSpace=10DOBZHANSKY (1970) on supergenes.

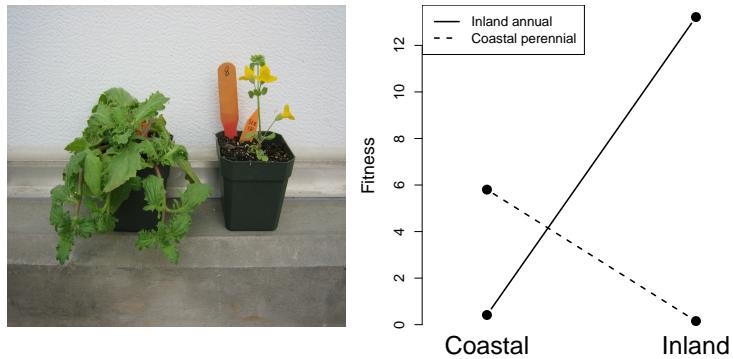


Figure 10.10: **Left)** A coastal perennial and an Inland annuals *Mimulus guttatus* LetterSpace=10LOWRY and LetterSpace=10WILLIS (2010), image from LetterSpace=10LOWRY and LetterSpace=10WILLIS (2010) licensed under CC BY 4.0. **Right)** A reciprocal transplant experiment showing that coastal perennial and an Inland annuals are locally adapted to their respective habitats. Data from LetterSpace=10LOWRY and LetterSpace=10WILLIS (2010), Code here..

Local Adaptation, Speciation, and Inversions. Inversions have long been thought to play an important role in local adaptation and speciation. One example of an inversion underlying local adaptation occurs in *Mimulus guttatus*, in Western North America, where there are annual and perennial ecomorphs. The perennial form grows in many places along the Pacific coast, and in other places with year around moisture; it invests a lot of resources in achieving large size and laying down resources for the next year, and as a result flowers late. The annual form grows inland, e.g. the California central valley, where it has to invest all its effort in flowering rapidly before the long, hot, dry summer. Neither ecomorph does well in the other's environment. The perennials get crisped before they have a chance to flower, while the annuals suffer from high rates of herbivory and cannot tolerate the salt spray. LetterSpace=10LOWRY and LetterSpace=10WILLIS (2010) found that large inversion controlled a lot of the phenotypic variation in flowering time and a range of other morphological differences between these two morphs. They also showed that the inversion controlled a reasonable proportion of the differences in fitness in the field, consistent with it underlying the fitness tradeoffs involved in local adaptation.

Why would an inversion be involved in locking together local adapted alleles? The basic idea, like above, is an inversion can be selected for we have two (or more) loci segregating for locally adapted alleles. Locally advantageous haplotypes are in danger of being broken up by recombination with maladapted haplotypes, which are constantly being introduced into each population by migration from the other. If an inversion arises that locks these alleles together in one population, it can be selected for as does not suffer the ill effects from recombination with migrating maladaptive haplotype.

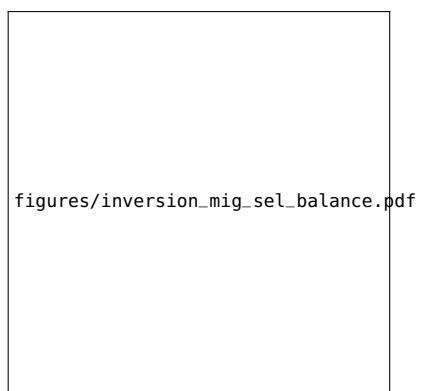


Figure 10.11: A two locus, two population migration-selection balance system. Two loci A and B segregate for an Inland and Coastal adapted alleles.

6084 10.1.1 Sex Chromosomes and the dynamics of selection and recombination.

6086 The production of different sized gametes (anisogamy) has arisen a number of times in multi-cellular life, with male and female gametes 6088 are defined by their relative sizes. The smaller, and often more mobile, gametes are defined male gametes (e.g. sperm), while the larger, 6090 well provisioned, and often less mobile are defined as female gametes (e.g. egg cell). The evolution of anisogamy is thought to be due to 6092 disruptive selection due to a tradeoff pulling in opposite directions towards mobile gametes able to move further and in the opposite direction towards better provisioned gametes better able to build larger 6094 zygotes. In many organisms individuals can produce both male and 6096 female gametes, while some species have evolved separate sexes, likely in part as an inbreeding avoidance mechanism. There is huge 6098 diversity in sex determination mechanisms across the eukaryotic tree (Figure 10.13). This is all to say, that biology is wonderfully complicated.

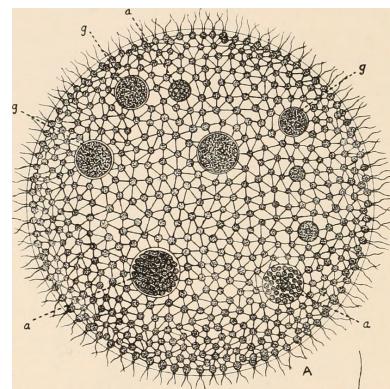


Figure 10.12: *Volvox aureus*, Volvox are spherical, multicellular green algae. The surface is made up of a single layer of somatic cells (up to 50k cells) beating their flagella. Some species of Volvox have male and female gametes, being made in the germ cells (a and g respectively) in the middle of the sphere. Some Volvox have separate sexes, where different individuals

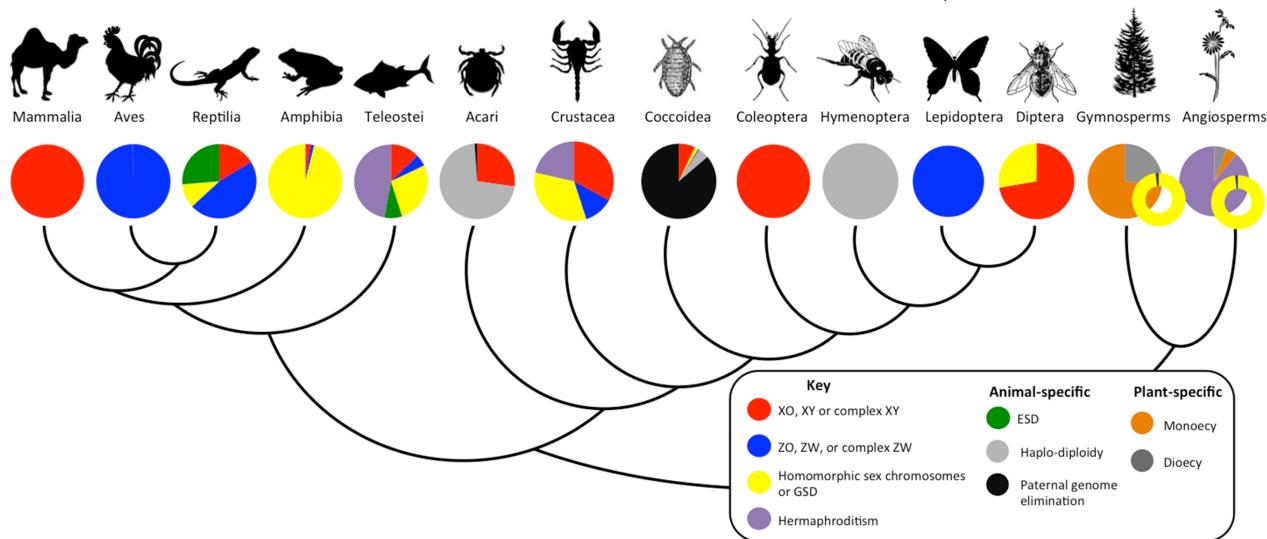


Figure 10.13: Diversity of sex determination systems for representative plant and animal clades. Figure and caption from LetterSpace=10BACHTROG *et al.* (2014), licensed under CC BY 4.0.

6100 In mammals, and many other systems with genetic sex determination, the genes responsible for sex determination lie on a pair of heteromorphic sex chromosomes, i.e. pair of chromosomes that are 6102 quite different in size. In mammals it is the male determining Y chromosome that has a very small gene content compared to the X chromosome (Figure 10.14). But in other groups such as birds, and 6106 some snakes, females carry a gene poor W with males being the homogametic sex, carrying two Zs. If you are still reading send Graham 6108 a picture of Nettie Stevens, she discovered sex chromosomes in 1905

6110 (LetterSpace=10STEVENS, 1905). These examples of heteromorphic
 6111 sex chromosomes, and many others like them, are thought to have
 6112 arisen from an ancestral pair of autosomes? What then explains their
 6113 evolution?

6114 A broad explanation for the evolution of sex chromosome goes as
 follows:

6116 In lake Malawi there are many very closely related cichlids species.
 6117 In many of these species the males are brightly coloured to attract
 6118 females, while the females are often brown to help them avoid predators.
 6119 In some of these species there is an alternative orange morph,
 6120 called the marmalade cat morph, which are cryptic against the rocky
 6121 bottom of the lake. This morph is due to a dominant (?) mutation
 6122 called OB at the *pax7* (?), and the allele appears to shared across
 6123 many of these species. This OB allele works well in females, however,
 6124 in the males the OB allele disrupts their bright colouration. Thus
 6125 the OB polymorphism is sexually antagonistic, i.e. it works well in
 6126 females and poorly in males.

6127 Males carrying the male-deleterious OB allele are rarely found,
 6128 despite the allele being common in females. Why is that? Well be-
 6129 cause the OB allele is tightly linked to a newly emerged female-
 6130 determining allele (W), with males carrying two copies of the Z al-
 6131 lele. Males usually are homozygous for the ob-Z haplotype, while
 6132 females can be either orange (OB-W/ob-Z) or brown (ob-W/ob-
 6133 Z). Recombination between these two loci seems to be very rare, and
 6134 so the sexually antagonistic allele OB appears to be mainly female
 specific. An inversion on the Z background would lock together the

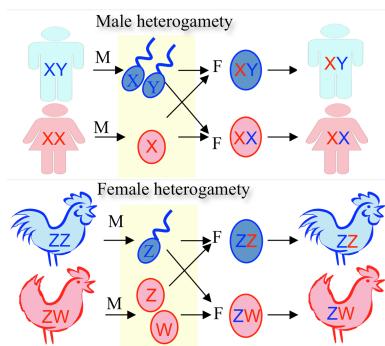


Figure 10.14: Figure from LetterSpace=10BACHTROG *et al.* (2014), licensed under CC BY 4.0cropped from original.

figures/selection_recom_interaction/Cichlid_0B_sex_linkage.pdf

Figure 10.15:
Image credits: Blue mbuna Male *L. fuelleborni* by Chmee2; OB Male *L. fuelleborni* by Doronenko; Brown ob *Tropheops* female by Alexandra Tyers; Female *L. fuelleborni* orange morph, by Mikko Stenberg

Bibliography

- 6138 LetterSpace=10AGUADÉ, M., LetterSpace=10N. MIYASHITA, and
LetterSpace=10C. H. LANGLEY, 1989 Reduced variation in the
6140 yellow-achaete-scute region in natural populations of *Drosophila*
melanogaster. *Genetics* **122**: 607–615.
- 6142 LetterSpace=10AGUILLO, S. M., LetterSpace=10J. W. FITZPATRICK,
LetterSpace=10R. BOWMAN, LetterSpace=10S. J. SCHOECH, Let-
6144 terSpace=10A. G. CLARK, LetterSpace=10G. COOP, and Let-
terSpace=10N. CHEN, 2017, 08) Deconstructing isolation-by-distance:
6146 The genomic consequences of limited dispersal. *PLOS Genet-*
ics **13**(8): 1–27.
- 6148 LetterSpace=10AKÇAY, E. and LetterSpace=10J. VAN CLEVE, 2016
There is no fitness but fitness, and the lineage is its bearer. *Phil.*
6150 *Trans. R. Soc. B* **371**(1687): 20150085.
- 6152 LetterSpace=10ALCAIDE, M., LetterSpace=10E. S. SCORDATO, Let-
terSpace=10T. D. PRICE, and LetterSpace=10D. E. IRWIN, 2014 Ge-
6154 nomic divergence in a ring species complex. *Nature* **511**(7507): 83.
- 6156 LetterSpace=10ALEXANDER, D. H., LetterSpace=10J. NOVEMBRE,
and LetterSpace=10K. LANGE, 2009 Fast model-based estimation of
ancestry in unrelated individuals. *Genome research* **19**(9): 1655–
1664.
- 6158 LetterSpace=10ALGEE-HEWITT, B. F., LetterSpace=10M. D. EDGE, Let-
terSpace=10J. KIM, LetterSpace=10J. Z. LI, and LetterSpace=10N. A.
6160 ROSENBERG, 2016 Individual identifiability predicts population
identifiability in forensic microsatellite markers. *Current Biol-*
6162 *ogy* **26**(7): 935–942.
- 6164 LetterSpace=10ALLENDORF, F. W. and LetterSpace=10J. J. HARD, 2009
Human-induced evolution caused by unnatural selection through
harvest of wild animals. *Proceedings of the National Academy of*
6166 *Sciences* **106**(Supplement 1): 9987–9994.

- 6168 LetterSpace=10ALVAREZ, G., LetterSpace=10F. C. CEBALLOS, and
 LetterSpace=10C. QUINTERO, 2009 The role of inbreeding in the
 extinction of a European royal dynasty. PLoS One 4(4): e5174.
- 6170 LetterSpace=10ANDOLFATTO, P., 2007 Hitchhiking effects of re-
 current beneficial amino acid substitutions in the *Drosophila*
 6172 *melanogaster* genome. Genome Res. 17: 1755–1762.
- 6174 LetterSpace=10ANDOLFATTO, P. and LetterSpace=10M. PRZEWORSKI,
 2001 Regions of lower crossing over harbor more rare variants in
 African populations of *Drosophila melanogaster*. Genetics 158: 657–
 6176 665.
- 6178 LetterSpace=10AYLLON, F., LetterSpace=10E. KJÆRNER-SEMB,
 LetterSpace=10T. FURMANEK, LetterSpace=10V. WENNEVIK,
 LetterSpace=10M. F. SOLBERG, LetterSpace=10G. DAHLE, Let-
 6180 terSpace=10G. L. TARANGER, LetterSpace=10K. A. GLOVER, Let-
 terSpace=10M. S. ALMÉN, LetterSpace=10C. J. RUBIN, and Let-
 6182 terSpace=10OTHERS, 2015 The vgl3 locus controls age at maturity in
 wild and domesticated Atlantic salmon (*Salmo salar* L.) males. PLoS
 6184 genetics 11(11): e1005628.
- 6186 LetterSpace=10BACHTROG, D., LetterSpace=10J. E. MANK, Let-
 terSpace=10C. L. PEICHEL, LetterSpace=10M. KIRKPATRICK,
 LetterSpace=10S. P. OTTO, LetterSpace=10T.-L. ASHMAN, Let-
 6188 terSpace=10M. W. HAHN, LetterSpace=10J. KITANO, Let-
 terSpace=10I. MAYROSE, LetterSpace=10R. MING, and Let-
 6190 terSpace=10OTHERS, 2014 Sex determination: why so many ways
 of doing it? PLoS biology 12(7): e1001899.
- 6192 LetterSpace=10BARRETT, R. D. H., LetterSpace=10S. M. ROGERS, and
 LetterSpace=10D. SCHLUTER, 2008 Natural Selection on a Major
 6194 Armor Gene in Threespine Stickleback. Science 322(5899): 255–257.
- 6196 LetterSpace=10BARSON, N. J., LetterSpace=10T. AYKANAT, Let-
 terSpace=10K. HINDAR, LetterSpace=10M. BARANSKI, Let-
 terSpace=10G. H. BOLSTAD, LetterSpace=10P. FISKE, Let-
 6198 terSpace=10C. JACQ, LetterSpace=10A. J. JENSEN, LetterSpace=10S. E.
 JOHNSTON, LetterSpace=10S. KARLSSON, and LetterSpace=10OTHERS,
 6200 2015 Sex-dependent dominance at a single locus maintains varia-
 tion in age at maturity in salmon. Nature 528(7582): 405.
- 6202 LetterSpace=10BARTON, N. and LetterSpace=10G. HEWITT, 1981 A
 chromosomal cline in the grasshopper *Podisma pedestris*. Evolu-
 6204 tion: 1008–1018.
- 6206 LetterSpace=10BARTON, N. H., 2000 Genetic hitchhiking. Philos.
 Trans. R. Soc. Lond., B, Biol. Sci. 355: 1553–1562.

- 6208 LetterSpace=10BASOLO, A. L., 1994 The dynamics of Fisherian sex-
ratio evolution: theoretical and experimental investigations. *The
American Naturalist* **144**(3): 473–490.
- 6210 LetterSpace=10BAZYKIN, A., 1969 Hypothetical mechanism of speci-
ation. *Evolution* **23**(4): 685–687.
- 6212 LetterSpace=10BECQUET, C., LetterSpace=10N. PATTERSON, Let-
terSpace=10A. C. STONE, LetterSpace=10M. PRZEWORSKI, and Let-
terSpace=10D. REICH, 2007 Genetic structure of chimpanzee popu-
lations. *PLoS genetics* **3**(4): e66.
- 6216 LetterSpace=10BEGUN, D. J. and LetterSpace=10C. F. AQUADRO, 1992
Levels of naturally occurring DNA polymorphism correlate with
recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- 6220 LetterSpace=10BEISSINGER, T. M., LetterSpace=10L. WANG, Let-
terSpace=10K. CROSBY, LetterSpace=10A. DURVASULA, Let-
terSpace=10M. B. HUFFORD, and LetterSpace=10J. ROSS-IBARRA,
6222 2016 Recent demography drives changes in linked selection across
the maize genome. *Nature plants* **2**(7): 16084.
- 6224 LetterSpace=10BELL, M. A., LetterSpace=10M. P. TRAVIS, and Let-
terSpace=10D. M. BLOUW, 2006 Inferring natural selection in a fossil
6226 threespine stickleback. *Paleobiology* **32**(4): 562–577.
- 6228 LetterSpace=10Box, G. E., 1979 Robustness in the strategy of sci-
entific model building. In *Robustness in statistics*, pp. 201–236. Elsevier.
- 6230 LetterSpace=10BRADBURD, G. S., LetterSpace=10P. L. RALPH, and
LetterSpace=10G. M. COOP, 2016 A spatial framework for under-
standing population structure and admixture. *PLoS genetics* **12**(1):
6232 e1005703.
- 6234 LetterSpace=10BRANDVAIN, Y., LetterSpace=10A. M. KENNEY,
LetterSpace=10L. FLAGEL, LetterSpace=10G. COOP, and Let-
terSpace=10A. L. SWEIGART, 2014 Speciation and introgression
6236 between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genet-*
ics **10**(6): e1004410.
- 6238 LetterSpace=10BRAVERMAN, J. M., LetterSpace=10R. R. HUDSON,
LetterSpace=10N. L. KAPLAN, LetterSpace=10C. H. LANGLEY, and
6240 LetterSpace=10W. STEPHAN, 1995 The hitchhiking effect on the site
frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783–796.
- 6242 LetterSpace=10BRODIE III, E. D., 1992 Correlational selection
for color pattern and antipredator behavior in the garter snake
6244 *Thamnophis ordinoides*. *Evolution* **46**(5): 1284–1298.

- 6246 LetterSpace=10CAI, J. J., LetterSpace=10J. M. MACPHERSON, Let-
 terSpace=10G. SELLA, and LetterSpace=10D. A. PETROV, 2009 Per-
 vasive hitchhiking at coding and regulatory sites in humans. PLoS
 6248 Genet. **5**: e1000336.
- 6250 LetterSpace=10CASSA, C. A., LetterSpace=10D. WEGHORN, Let-
 terSpace=10D. J. BALICK, LetterSpace=10D. M. JORDAN, Let-
 terSpace=10D. NUSINOW, LetterSpace=10K. E. SAMOCHA,
 6252 LetterSpace=10A. O'DONNELL-LURIA, LetterSpace=10D. G.
 MACARTHUR, LetterSpace=10M. J. DALY, LetterSpace=10D. R. BEIER,
 6254 and LetterSpace=10OTHERS, 2017 Estimating the selective effects of
 heterozygous protein-truncating variants from human exome data.
 6256 Nature genetics **49**(5): 806.
- 6258 LetterSpace=10CHARLESWORTH, B., 2009 Effective population size
 and patterns of molecular evolution and variation. Nature Reviews
 Genetics **10**(3): 195.
- 6260 LetterSpace=10CHARLESWORTH, D., Let-
 terSpace=10B. CHARLESWORTH, and LetterSpace=10M. T.
 6262 MORGAN, 1995 The pattern of neutral molecular variation under the
 background selection model. Genetics **141**: 1619–1632.
- 6264 LetterSpace=10CHARLESWORTH, D. and LetterSpace=10V. LAPORTE,
 1998 The male-sterility polymorphism of *Silene vulgaris*: analysis
 6266 of genetic data from two populations and comparison with *Thymus*
vulgaris. Genetics **150**(3): 1267–1282.
- 6268 LetterSpace=10CHEN, N., LetterSpace=10E. J. COSGROVE, Let-
 terSpace=10R. BOWMAN, LetterSpace=10J. W. FITZPATRICK, and
 6270 LetterSpace=10A. G. CLARK, 2016 Genomic Consequences of Pop-
 ulation Decline in the Endangered Florida Scrub-Jay. Current Biol-
 ogy **26**(21): 2974 – 2979.
- 6274 LetterSpace=10COHEN, D., 1966 Optimizing reproduction in a ran-
 domly varying environment. Journal of theoretical biology **12**(1):
 119–129.
- 6276 LetterSpace=10COOK, L. M., LetterSpace=10B. S. GRANT, Let-
 terSpace=10I. J. SACCHERI, and LetterSpace=10J. MALLET, 2012 Se-
 lective bird predation on the peppered moth: the last experiment of
 6278 Michael Majerus. Biology Letters **8**(4): 609–612.
- 6280 LetterSpace=10COTTERMAN, C. W., 1940 A calculus for statistico-
 genetics. Ph. D. thesis, The Ohio State University.
- 6282 LetterSpace=10COUSMINER, D. L., LetterSpace=10D. J. BERRY,
 LetterSpace=10N. J. TIMPSON, LetterSpace=10W. ANG, Let-

- 6284 terSpace=10E. THIERING, LetterSpace=10E. M. BYRNE, Let-
terSpace=10H. R. TAAL, LetterSpace=10V. HUIKARI, Let-
terSpace=10J. P. BRADFIELD, LetterSpace=10M. KERKHOF, and Let-
terSpace=10OTHERS, 2013 Genome-wide association and longitu-
6288 dinal analyses reveal genetic loci linking pubertal height growth,
pubertal timing and childhood adiposity. *Human molecular genet-
ics* 22(13): 2735–2747.
- 6292 LetterSpace=10CUTTER, A. D. and LetterSpace=10J. Y. CHOI, 2010
Natural selection shapes nucleotide polymorphism across the
genome of the nematode *Caenorhabditis briggsae*. *Genome Res.* 20:
6294 1103–1111.
- 6296 LetterSpace=10DARWIN, C., 1859 *On the Origin of Species by Means of
Natural Selection*. London: Murray. or the Preservation of Favored
Races in the Struggle for Life.
- 6298 LetterSpace=10DARWIN, C., 1876 The effect of cross and self fertil-
ization in the vegetable kingdom: Murray. London, UK.
- 6300 LetterSpace=10DARWIN, C., 1888 *The descent of man and selection in
relation to sex*, Volume 1. Murray.
- 6302 LetterSpace=10DEMPSTER, E., 1955 Maintenance of genetic hetero-
geneity. *Cold Spring Harb Symp Quant Biol* 20: 25–32.
- 6304 LetterSpace=10DICKERSON, R. E., 1971 The structure of cytochrome c
and the rates of molecular evolution. *Journal of Molecular Evolu-
6306 tion* 1(1): 26–45.
- 6308 LetterSpace=10DOBZHANSKY, T., 1943 Genetics of natural popula-
tions IX. Temporal changes in the composition of populations of
Drosophila pseudoobscura. *Genetics* 28(2): 162.
- 6310 LetterSpace=10DOBZHANSKY, T., 1951 *Genetics and the Origin of
Species* (3rd Ed. ed.), pp. 16.
- 6312 LetterSpace=10DOBZHANSKY, T., 1970 *Genetics of the evolutionary
process*, Volume 139. Columbia University Press.
- 6314 LetterSpace=10EDWARDS, A., 1961 The population genetics of
sex-ratio in *Drosophila pseudoobscura*. *Heredity* 16(3):
6316 291.
- 6318 LetterSpace=10EDWARDS, A. W., 1998 Natural selection and the sex
ratio: Fisher's sources. *The American Naturalist* 151(6): 564–569.
- 6320 LetterSpace=10ELTON, C., 1942 *Voles, mice and lemmings. Problems in
population dynamics*. Oxford: Clarendon Press.

LetterSpace=10ELYASHIV, E., LetterSpace=10S. SATTATH, LetterSpace=10T. T. HU, LetterSpace=10A. STRUTSOVSKY, LetterSpace=10G. McVICKER, LetterSpace=10P. ANDOLFATTO, LetterSpace=10G. COOP, and LetterSpace=10G. SELLA, 2016 A genomic map of the effects of linked selection in *Drosophila*. PLoS genetics 12(8): e1006130.

LetterSpace=10EWENS, W. J., 2010 What is the gene trying to do? British Journal for the Philosophy of Science 62(1): 155–176.

LetterSpace=10EWENS, W. J., 2016 Motoo Kimura and James Crow on the Infinitely Many Alleles Model. Genetics 202(4): 1243–1245.

LetterSpace=10FAY, J. C. and LetterSpace=10C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

LetterSpace=10FEDER, A. F., LetterSpace=10C. KLINE, LetterSpace=10P. POLACINO, LetterSpace=10M. COTTRELL, LetterSpace=10A. D. KASHUBA, LetterSpace=10B. F. KEELE, LetterSpace=10S.-L. HU, LetterSpace=10D. A. PETROV, LetterSpace=10P. S. PENNINGS, and LetterSpace=10Z. AMBROSE, 2017 A spatio-temporal assessment of simian/human immunodeficiency virus (SHIV) evolution reveals a highly dynamic process within the host. PLoS pathogens 13(5): e1006358.

LetterSpace=10FERREE, P. M. and LetterSpace=10D. A. BARBASH, 2007 Distorted sex ratios: a window into RNAi-mediated silencing. PLoS biology 5(11): e303.

LetterSpace=10FISHER, R. A., 1915 The evolution of sexual preference. The Eugenics Review 7(3): 184.

LetterSpace=10FISHER, R. A., 1923 XXI. On the dominance ratio. Proceedings of the royal society of Edinburgh 42: 321–341.

LetterSpace=10FISHER, R. A., 1930 *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

LetterSpace=10FRANCIOLI, L. C., LetterSpace=10A. MENELAOU, LetterSpace=10S. L. PULIT, LetterSpace=10F. VAN DIJK, LetterSpace=10P. F. PALAMARA, LetterSpace=10C. C. ELBERS, LetterSpace=10P. B. NEERINCKX, LetterSpace=10K. YE, LetterSpace=10V. GURYEV, LetterSpace=10W. P. KLOOSTERMAN, and LetterSpace=10OTHERS, 2014 Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature genetics 46(8): 818.

- 6358 LetterSpace=10FRENTIU, F. D., LetterSpace=10G. D. BERNARD, LetterSpace=10C. I. CUEVAS, LetterSpace=10M. P. SISON-MANGUS, LetterSpace=10K. L. PRUDIC, and LetterSpace=10A. D. BRISCOE, 2007
6360 Adaptive evolution of color vision as seen through the eyes of butterflies. *Proceedings of the National Academy of Sciences* 104(suppl 1): 8634–8640.
- 6364 LetterSpace=10GALEN, C., 1996 Rates of floral evolution: adaptation to bumblebee pollination in an alpine wildflower, *Polemonium viscosum*. *Evolution* 50(1): 120–125.
- 6366 LetterSpace=10GALTIER, N., 2016 Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS genetics* 12(1): e1005774.
- 6370 LetterSpace=10GIGORD, L. D., LetterSpace=10M. R. MACNAIR, and LetterSpace=10A. SMITHSON, 2001 Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid *Dactylorhiza sambucina* (L.) Soo. *Proceedings of the National Academy of Sciences* 98(11): 6253–6255.
- 6376 LetterSpace=10GILLESPIE, J. H., 1973 Natural selection with varying selection coefficients—a haploid model. *Genetics Research* 21(2): 115–120.
- 6378 LetterSpace=10GILLESPIE, J. H., 1977 Natural selection for variances in offspring numbers: a new evolutionary principle. *The American Naturalist* 111(981): 1010–1014.
- 6382 LetterSpace=10GILLESPIE, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155: 909–919.
- 6384 LetterSpace=10GREMER, J. R. and LetterSpace=10D. L. VENABLE, 2014 Bet hedging in desert winter annual plants: optimal germination strategies in a variable environment. *Ecology Letters* 17(3): 380–387.
- 6388 LetterSpace=10HALDANE, J., 1942 The selective elimination of silver foxes in eastern Canada. *Journal of Genetics* 44(2-3): 296–304.
- 6390 LetterSpace=10HALDANE, J. and LetterSpace=10S. JAYAKAR, 1963 Polymorphism due to selection of varying direction. *Journal of Genetics* 58(2): 237–242.
- 6392 LetterSpace=10HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: selection and mutation. In 6394 *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 23, pp. 838–844. Cambridge University Press.

- 6396 LetterSpace=10HALDANE, J. B. S., 1937 The Effect of Variation of Fitness. *The American Naturalist* 71(735): 337–349.
- 6398 LetterSpace=10HAMILTON, W. D., 1964a The genetical evolution of social behaviour. II. *Journal of theoretical biology* 7(1): 17–52.
- 6400 LetterSpace=10HAMILTON, W. D., 1964b The genetical evolution of social behaviour. II. *Journal of theoretical biology* 7(1): 17–52.
- 6402 LetterSpace=10HERMISSON, J. and LetterSpace=10P. S. PENNINGS, 2017 Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution* 8(6): 700–716.
- 6406 LetterSpace=10HEY, J. and LetterSpace=10R. M. KLIMAN, 2002 Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* 160(2): 595–608.
- 6410 LetterSpace=10HOEKSTRA, H. E., LetterSpace=10K. E. DRUMM, and LetterSpace=10M. W. NACHMAN, 2004 Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. *Evolution* 58(6): 1329–1341.
- 6414 LetterSpace=10HOHENLOHE, P. A., LetterSpace=10S. BASSHAM, LetterSpace=10P. D. ETTER, LetterSpace=10N. STIFFLER, LetterSpace=10E. A. JOHNSON, and LetterSpace=10W. A. CRESKO, 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics* 6(2): e1000862.
- 6418 LetterSpace=10HOLLISTER, J. D., LetterSpace=10S. GREINER, LetterSpace=10W. WANG, LetterSpace=10J. WANG, LetterSpace=10Y. ZHANG, LetterSpace=10G. K.-S. WONG, LetterSpace=10S. I. WRIGHT, and LetterSpace=10M. T. JOHNSON, 2014 Recurrent loss of sex is associated with accumulation of deleterious mutations in *Oenothera*. *Molecular biology and evolution* 32(4): 896–905.
- 6426 LetterSpace=10HOPKINS, J., LetterSpace=10G. BAUDRY, LetterSpace=10U. CANDOLIN, and LetterSpace=10A. KAITALA, 2015 I'm sexy and I glow it: female ornamentation in a nocturnal capital breeder. *Biology letters* 11(10): 20150599.
- 6430 LetterSpace=10HOUDE, A. E., 1994 Effect of artificial selection on male colour patterns on mating preference of female guppies. *Proc. R. Soc. Lond. B* 256(1346): 125–130.
- 6432 LetterSpace=10HOWES, R. E., LetterSpace=10M. DEWI, LetterSpace=10F. B. PIEL, LetterSpace=10W. M. MONTEIRO, LetterSpace=10K. E. BATTLE, LetterSpace=10J. P. MESSINA, LetterSpace=10A. SAKUNTABHAI, LetterSpace=10A. W. SATYAGRAHA,

- 6436 LetterSpace=10T. N. WILLIAMS, LetterSpace=10J. K. BAIRD, and LetterSpace=10S. I. HAY, 2013 Spatial distribution of G6PD deficiency variants across malaria-endemic regions. *Malar. J.* **12**: 418.
- 6438
- 6440 LetterSpace=10Howes, R. E., LetterSpace=10F. B. PIEL, LetterSpace=10A. P. PATIL, LetterSpace=10O. A. NYANGIRI, LetterSpace=10P. W. GETHING, LetterSpace=10M. DEWI, LetterSpace=10M. M. HOGG, LetterSpace=10K. E. BATTLE, LetterSpace=10C. D. PADILLA, LetterSpace=10J. K. BAIRD, and LetterSpace=10S. I. HAY, 2012 G6PD deficiency prevalence and estimates of affected populations in malaria endemic countries: a geostatistical model-based map. *PLoS Medicine* **9**(11): e1001339.
- 6442
- 6444
- 6446
- 6448 LetterSpace=10HUDSON, R. R., 2015, 07)A New Proof of the Expected Frequency Spectrum under the Standard Neutral Model. *PLOS ONE* **10**(7): 1–5.
- 6450 LetterSpace=10HUDSON, R. R. and LetterSpace=10N. L. KAPLAN, 1995a Deleterious background selection with recombination. *Genetics* **141**: 1605–1617.
- 6452
- 6454 LetterSpace=10HUDSON, R. R. and LetterSpace=10N. L. KAPLAN, 1995b The coalescent process and background selection. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **349**: 19–23.
- 6456 LetterSpace=10HUDSON, R. R., LetterSpace=10M. KREITMAN, and LetterSpace=10M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**(1): 153–159.
- 6458
- 6460 LetterSpace=10HUNT, G., LetterSpace=10M. A. BELL, and LetterSpace=10M. P. TRAVIS, 2008 Evolution toward a new adaptive optimum: phenotypic evolution in a fossil stickleback lineage. *Evolution* **62**(3): 700–710.
- 6462
- 6464 LetterSpace=10JAIN, S. and LetterSpace=10A. D. BRADSHAW, 1966 Evolutionary divergence among adjacent plant populations I. The evidence and its theoretical analysis. *Heredity* **21**(3): 407.
- 6466 LetterSpace=10JANICKE, T., LetterSpace=10I. K. HÄDERER, LetterSpace=10M. J. LAJEUNESSE, and LetterSpace=10N. ANTHES, 2016 Darwinian sex roles confirmed across the animal kingdom. *Science advances* **2**(2): e1500983.
- 6468
- 6470 LetterSpace=10JENNINGS, W. B. and LetterSpace=10S. V. EDWARDS, 2005 Speciation history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**(9): 2033–2047.
- 6472
- 6474 LetterSpace=10JOHANNSEN, W., 1911 The Genotype Conception of Heredity. *The American Naturalist* **45**(531): 129–159.

- 6476 LetterSpace=10JOHNSTON, S. E., LetterSpace=10J. GRATTEN, Let-
 terSpace=10C. BERENOS, LetterSpace=10J. G. PILKINGTON, Let-
 terSpace=10T. H. CLUTTON-BROCK, LetterSpace=10J. M. PEMBERTON,
 6478 and LetterSpace=10J. SLATE, 2013 Life history trade-offs at a single
 locus maintain sexually selected genetic variation. *Nature* 502(7469):
 6480 93.
- 6482 LetterSpace=10JORON, M., LetterSpace=10L. FREZAL, Let-
 terSpace=10R. T. JONES, LetterSpace=10N. L. CHAMBERLAIN,
 LetterSpace=10S. F. LEE, LetterSpace=10C. R. HAAG, Let-
 terSpace=10A. WHIBLEY, LetterSpace=10M. BECUWE, Let-
 terSpace=10S. W. BAXTER, LetterSpace=10L. FERGUSON, and Let-
 terSpace=10OTHERS, 2011 Chromosomal rearrangements main-
 6486 tain a polymorphic supergene controlling butterfly mimicry. *Na-
 ture* 477(7363): 203.
- 6488 LetterSpace=10JORON, M., LetterSpace=10R. PAPA, Let-
 terSpace=10M. BELTRÁN, LetterSpace=10N. CHAMBERLAIN,
 LetterSpace=10J. MAVÁREZ, LetterSpace=10S. BAXTER, Let-
 terSpace=10M. ABANTO, LetterSpace=10E. BIRMINGHAM, Let-
 terSpace=10S. J. HUMPHRAY, LetterSpace=10J. ROGERS, and Let-
 terSpace=10OTHERS, 2006 A conserved supergene locus controls
 6494 colour pattern diversity in *Heliconius* butterflies. *PLoS biol-*
 ogy 4(10): e303.
- 6496 LetterSpace=10JUKEMA, J. and LetterSpace=10T. PIERSMA, 2006 Per-
 manent female mimics in a lekking shorebird. *Biology letters* 2(2):
 161–164.
- 6500 LetterSpace=10KAPLAN, N. L., LetterSpace=10R. R. HUDSON, and
 LetterSpace=10C. H. LANGLEY, 1989 The hitchhiking effect revisited.
 6502 *Genetics* 123: 887–899.
- 6504 LetterSpace=10KARN, M. N. and LetterSpace=10L. PENROSE, 1951
 Birth weight and gestation time in relation to maternal age, parity
 and infant survival. *Annals of eugenics* 16(1): 147–164.
- 6506 LetterSpace=10KETTLEWELL, H. B. D., 1955 Selection experiments
 on industrial melanism in the Lepidoptera. *Heredity* 9(3): 323.
- 6508 LetterSpace=10KIM, Y., 2006 Allele frequency distribution under
 recurrent selective sweeps. *Genetics* 172: 1967–1978.
- 6510 LetterSpace=10KIMURA, M., 1968 Evolutionary rate at the molecular
 level. *Nature* 217(5129): 624–626.
- 6512 LetterSpace=10KIMURA, M., 1983 *The neutral theory of molecular
 evolution*. Cambridge University Press.

- 6514 LetterSpace=10KIMURA, M. and LetterSpace=10J. F. CROW, 1964
The number of alleles that can be maintained in a finite population.
6516 Genetics 49(4): 725.
- 6518 LetterSpace=10KIMURA, M. and LetterSpace=10T. OHTA, 1974 On
some principles governing molecular evolution. Proceedings of the
National Academy of Sciences 71(7): 2848–2852.
- 6520 LetterSpace=10KING, J. L. and LetterSpace=10T. H. JUKES, 1969
Non-darwinian evolution. Science 164(3881): 788–798.
- 6522 LetterSpace=10KORNEGAY, J. R., LetterSpace=10J. W. SCHILLING, and
LetterSpace=10A. C. WILSON, 1994 Molecular adaptation of a leaf-
6524 eating bird: stomach lysozyme of the hoatzin. Molecular Biology
and Evolution 11(6): 921–928.
- 6526 LetterSpace=10KRAKAUER, A. H., 2005 Kin selection and coopera-
tive courtship in wild turkeys. Nature 434(7029): 69.
- 6528 LetterSpace=10KRUUK, L. E., LetterSpace=10J. SLATE, Let-
terSpace=10J. M. PEMBERTON, LetterSpace=10S. BROTHERSTONE,
6530 LetterSpace=10F. GUINNESS, and LetterSpace=10T. CLUTTON-BROCK,
2002 Antler size in red deer: heritability and selection but no evolu-
6532 tion. Evolution 56(8): 1683–1695.
- 6534 LetterSpace=10KÜPPER, C., LetterSpace=10M. STOCKS, Let-
terSpace=10J. E. RISSE, LetterSpace=10N. DOS REMEDIOS, Let-
terSpace=10L. L. FARRELL, LetterSpace=10S. B. MCRAE, Let-
6536 terSpace=10T. C. MORGAN, LetterSpace=10N. KARLIONOVA, Let-
terSpace=10P. PINCHUK, LetterSpace=10Y. I. VERKUIL, and Let-
terSpace=10OTHERS, 2016 A supergene determines highly divergent
6538 male reproductive morphs in the ruff. Nature Genetics 48(1): 79.
- 6540 LetterSpace=10KWIATKOWSKI, D. P., 2005, August)How malaria has
affected the human genome and what human genetics can teach us
6542 about malaria. Am. J. Hum. Genet. 77(2): 171–192.
- 6544 LetterSpace=10LABERGE, A.-M., LetterSpace=10M. JOMPHE,
LetterSpace=10L. HOUDE, LetterSpace=10H. VÉZINA, Let-
terSpace=10M. TREMBLAY, LetterSpace=10B. DESJARDINS, Let-
6546 terSpace=10D. LABUDA, LetterSpace=10M. ST-HILAIRE, Let-
terSpace=10C. MACMILLAN, LetterSpace=10E. A. SHOUBRIDGE, and
6548 LetterSpace=10OTHERS, 2005 A àAJFille du RoyâÀI introduced
the T14484C Leber hereditary optic neuropathy mutation in French
6550 Canadians. The American Journal of Human Genetics 77(2): 313–
317.

- 6552 LetterSpace=10LAMICHHANEY, S., LetterSpace=10G. FAN, Let-
 terSpace=10F. WIDEMO, LetterSpace=10U. GUNNARSSON, Let-
 terSpace=10D. S. THALMANN, LetterSpace=10M. P. HOEPP-
 NER, LetterSpace=10S. KERJE, LetterSpace=10U. GUSTAFSON,
 LetterSpace=10C. SHI, LetterSpace=10H. ZHANG, and Let-
 terSpace=10OTHERS, 2016 Structural genomic changes underlie
 alternative reproductive strategies in the ruff (*Philomachus pugnax*).
 Nature Genetics 48(1): 84.
- 6556 LetterSpace=10LANDE, R., 1976 Natural selection and random ge-
 netic drift in phenotypic evolution. Evolution 30(2): 314–334.
- 6560 LetterSpace=10LANDE, R., 1979 Quantitative genetic analysis of
 multivariate evolution, applied to brain: body size allometry. Evolu-
 tion 33(1Part2): 402–416.
- 6564 LetterSpace=10LAURIE, C. C., LetterSpace=10D. A. NICKERSON,
 LetterSpace=10A. D. ANDERSON, LetterSpace=10B. S. WEIR, Let-
 terSpace=10R. J. LIVINGSTON, LetterSpace=10M. D. DEAN, Let-
 terSpace=10K. L. SMITH, LetterSpace=10E. E. SCHADT, and Let-
 terSpace=10M. W. NACHMAN, 2007, 08)Linkage Disequilibrium in
 Wild Mice. PLOS Genetics 3(8): 1–9.
- 6568 LetterSpace=10LAWSON, D. J., LetterSpace=10L. VAN DORP, and Let-
 terSpace=10D. FALUSH, 2018 A tutorial on how not to over-interpret
 STRUCTURE and ADMIXTURE bar plots. Nature communica-
 tions 9(1): 3258.
- 6572 LetterSpace=10LEFÉBURE, T., LetterSpace=10C. MORVAN, Let-
 terSpace=10F. MALARD, LetterSpace=10C. FRANÇOIS, Let-
 terSpace=10L. KONECNY-DUPRÉ, LetterSpace=10L. GUÉGUEN, Let-
 terSpace=10M. WEISS-GAYET, LetterSpace=10A. SEGUIN-ORLANDO,
 LetterSpace=10L. ERMINI, LetterSpace=10C. DER SARKISSIAN, and
 LetterSpace=10OTHERS, 2017 Less effective selection leads to larger
 genomes. Genome research: gr–212589.
- 6576 LetterSpace=10LEFFLER, E. M., LetterSpace=10K. BULLAUGHEY,
 LetterSpace=10D. R. MATUTE, LetterSpace=10W. K. MEYER,
 LetterSpace=10L. SECUREL, LetterSpace=10A. VENKAT, Let-
 terSpace=10P. ANDOLFATTO, and LetterSpace=10M. PRZEWORSKI,
 2012 Revisiting an old riddle: what determines genetic diversity
 levels within species? PLoS biology 10(9): e1001388.
- 6580 LetterSpace=10LEK, M., LetterSpace=10K. J. KARCZEWSKI, Let-
 terSpace=10E. V. MINIKEL, LetterSpace=10K. E. SAMOCHA,
 LetterSpace=10E. BANKS, LetterSpace=10T. FENNELL, Let-
 terSpace=10A. H. O'DONNELL-LURIA, LetterSpace=10J. S. WARE,

- 6592 LetterSpace=10A. J. HILL, LetterSpace=10B. B. CUMMINGS, and
6594 LetterSpace=10OTHERS, 2016 Analysis of protein-coding genetic
variation in 60,706 humans. *Nature* 536(7616): 285.
- 6596 LetterSpace=10LENORMAND, T., LetterSpace=10D. BOURGUET, Let-
terSpace=10T. GUILLEMAUD, and LetterSpace=10M. RAYMOND, 1999
6598 Tracking the evolution of insecticide resistance in the mosquito
Culex pipiens. *Nature* 400(6747): 861.
- 6600 LetterSpace=10LEWONTIN, R. C., 1970 The units of selection. *An-
nual review of ecology and systematics* 1(1): 1–18.
- 6602 LetterSpace=10LEWONTIN, R. C., 1974 *The Genetic Basis of Evolution-
ary Change*. Columbia University Press, New York.
- 6604 LetterSpace=10LEWONTIN, R. C., 1994, 05)[DNA Fingerprinting: A
Review of the Controversy]: Comment: The Use of DNA Profiles in
Forensic Contexts. *Statist. Sci.* 9(2): 259–262.
- 6606 LetterSpace=10LEWONTIN, R. C., 2001 *Thinking about evolution: his-
torical, philosophical, and political perspectives*, Chapter Natural History
6608 and Formalism in Evolutionary Genetics, pp. 7–20. Cambridge
University Press.
- 6610 LetterSpace=10LI, J. Z., LetterSpace=10D. M. ABSHER, Let-
terSpace=10H. TANG, LetterSpace=10A. M. SOUTHWICK, Let-
6612 terSpace=10A. M. CASTO, LetterSpace=10S. RAMACHANDRAN,
LetterSpace=10H. M. CANN, LetterSpace=10G. S. BARSH, Let-
6614 terSpace=10M. FELDMAN, LetterSpace=10L. L. CAVALLI-SFORZA,
and LetterSpace=10OTHERS, 2008 Worldwide human relationships
6616 inferred from genome-wide patterns of variation. *science* 319(5866):
1100–1104.
- 6618 LetterSpace=10LIN, C.-J., LetterSpace=10F. HU, Let-
terSpace=10R. DUBRUILLE, LetterSpace=10J. VEDANAYAGAM,
6620 LetterSpace=10J. WEN, LetterSpace=10P. SMIBERT, Let-
terSpace=10B. LOPPIN, and LetterSpace=10E. C. LAI, 2018 The
6622 hpRNA/RNAi pathway is essential to resolve intragenomic con-
flict in the Drosophila male germline. *Developmental cell* 46(3):
6624 316–326.
- 6626 LetterSpace=10LISTER, A., 1989 Rapid dwarfing of red deer on
Jersey in the last interglacial. *Nature* 342(6249): 539.
- 6628 LetterSpace=10LOCKE, D. P., LetterSpace=10L. W. HILLIER, Let-
terSpace=10W. C. WARREN, LetterSpace=10K. C. WORLEY, Let-
terSpace=10L. V. NAZARETH, LetterSpace=10D. M. MUZNY, Let-
6630 terSpace=10S.-P. YANG, LetterSpace=10Z. WANG, LetterSpace=10A. T.

- CHINWALLA, LetterSpace=10P. MINX, and LetterSpace=10OTHERS, 2011 Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331): 529.
- LetterSpace=10Losos, J. B., LetterSpace=10S. J. ARNOLD, LetterSpace=10G. BEJERANO, LetterSpace=10E. BRODIE III, LetterSpace=10D. HIBBETT, LetterSpace=10H. E. HOEKSTRA, LetterSpace=10D. P. MINDELL, LetterSpace=10A. MONTEIRO, LetterSpace=10C. MORITZ, LetterSpace=10H. A. ORR, and LetterSpace=10OTHERS, 2013 Evolutionary biology for the 21st century. *PLoS Biology* 11(1): e1001466.
- LetterSpace=10LOUICHAROEN, C., LetterSpace=10E. PATIN, LetterSpace=10R. PAUL, LetterSpace=10I. NUCHPRAYOON, LetterSpace=10B. WITOONPANICH, LetterSpace=10C. PEERAPITTAYA-MONGKOL, LetterSpace=10I. CASADEMONT, LetterSpace=10T. SURA, LetterSpace=10N. M. LAIRD, LetterSpace=10P. SINGHASIVANON, LetterSpace=10L. QUINTANA-MURCI, and LetterSpace=10A. SAKUNTABHAI, 2009, December)Positively selected G6PD-Mahidol mutation reduces *Plasmodium vivax* density in Southeast Asians. *Science* 326(5959): 1546–1549.
- LetterSpace=10LOWRY, D. B. and LetterSpace=10J. H. WILLIS, 2010 A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS biology* 8(9): e1000500.
- LetterSpace=10MACARTHUR, D. G., LetterSpace=10S. BALASUBRAMANIAN, LetterSpace=10A. FRANKISH, LetterSpace=10N. HUANG, LetterSpace=10J. MORRIS, LetterSpace=10K. WALTER, LetterSpace=10L. JOSTINS, LetterSpace=10L. HABEGGER, LetterSpace=10J. K. PICKRELL, LetterSpace=10S. B. MONTGOMERY, and LetterSpace=10OTHERS, 2012 A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070): 823–828.
- LetterSpace=10MACPHERSON, J. M., LetterSpace=10G. SELLA, LetterSpace=10J. C. DAVIS, and LetterSpace=10D. A. PETROV, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* 177: 2083–2099.
- LetterSpace=10MAJERUS, M. E., 2009 Industrial melanism in the peppered moth, *Biston betularia*: an excellent teaching example of Darwinian evolution in action. *Evolution: Education and Outreach* 2(1): 63.
- LetterSpace=10MALÉCOT, G., 1948 Les mathématiques de l'hérédité.

- 6672 LetterSpace=10MALÉCOT, G., 1969 The Mathematics of Heredity
(Revised, edited and translated by Yermanos, DM).
- 6674 LetterSpace=10MARCINIAK, S. and LetterSpace=10G. H. PERRY,
2017 Harnessing ancient genomes to study the history of human
adaptation. *Nature Reviews Genetics* 18(11): 659.
- 6676 LetterSpace=10MAYNARD SMITH, J., 1964 Group selection and kin
selection. *Nature* 201(4924): 1145.
- 6678 LetterSpace=10MAYNARD SMITH, J. and LetterSpace=10J. HAIGH,
1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* 23:
23–35.
- 6682 LetterSpace=10McDONALD, J. H. and LetterSpace=10M. KREITMAN,
1991 Adaptive protein evolution at the Adh locus in Drosophila.
Nature 351(6328): 652.
- 6684 LetterSpace=10MC VICKER, G., LetterSpace=10D. GORDON,
LetterSpace=10C. DAVIS, and LetterSpace=10P. GREEN, 2009
6686 Widespread genomic signatures of natural selection in hominid
evolution. *PLoS Genet.* 5: e1000471.
- 6688 LetterSpace=10MENOZZI, P., LetterSpace=10A. PIAZZA, and Let-
terSpace=10L. CAVALLI-SFORZA, 1978 Synthetic maps of human
6690 gene frequencies in Europeans. *Science* 201(4358): 786–792.
- 6692 LetterSpace=10MEREDITH, R. W., LetterSpace=10J. GATESY, Let-
terSpace=10W. J. MURPHY, LetterSpace=10O. A. RYDER, and Let-
terSpace=10M. S. SPRINGER, 2009, 09)Molecular Decay of the Tooth
6694 Gene Enamelin (ENAM) Mirrors the Loss of Enamel in the Fossil
Record of Placental Mammals. *PLOS Genetics* 5(9): 1–12.
- 6696 LetterSpace=10MESSIER, W. and LetterSpace=10C.-B. STEWART,
1997 Episodic adaptive evolution of primate lysozymes. *Na-
ture* 385(6612): 151.
- 6700 LetterSpace=10MILOT, E., LetterSpace=10C. MOREAU, Let-
terSpace=10A. GAGNON, LetterSpace=10A. A. COHEN, Let-
terSpace=10B. BRAIS, and LetterSpace=10D. LABUDA, 2017 Moth-
6702 erâžs curse neutralizes natural selection against a human genetic
disease over three centuries. *Nature ecology & evolution* 1(9): 1400.
- 6704 LetterSpace=10MULLER, H. J., 1932 Some genetic aspects of sex. *The
American Naturalist* 66(703): 118–138.
- 6706 LetterSpace=10NACHMAN, M. W., LetterSpace=10H. E. HOEKSTRA,
and LetterSpace=10S. L. D'AGOSTINO, 2003 The genetic basis of
6708 adaptive melanism in pocket mice. *Proceedings of the National
Academy of Sciences* 100(9): 5268–5273.

- 6710 LetterSpace=10NASH, D., LetterSpace=10S. NAIR, Let-
terSpace=10M. MAYXAY, LetterSpace=10P. N. NEWTON, Let-
6712 terSpace=10J.-P. GUTHMANN, LetterSpace=10F. NOSTEN, and Let-
terSpace=10T. J. ANDERSON, 2005 Selection strength and hitchhiking
6714 around two anti-malarial resistance genes. *Proceedings of the Royal
Society of London B: Biological Sciences* 272(1568): 1153–1161.
- 6716 LetterSpace=10NELSON, M. R., LetterSpace=10D. WEGMANN,
LetterSpace=10M. G. EHM, LetterSpace=10D. KESSNER, Let-
6718 terSpace=10P. S. JEAN, LetterSpace=10C. VERZILLI, Let-
terSpace=10J. SHEN, LetterSpace=10Z. TANG, LetterSpace=10S.-A.
6720 BACANU, LetterSpace=10D. FRASER, and LetterSpace=10OTHERS, 2012
An abundance of rare functional variants in 202 drug target genes
6722 sequenced in 14,002 people. *Science*: 1217876.
- 6724 LetterSpace=10NORDBORG, M., LetterSpace=10B. CHARLESWORTH,
and LetterSpace=10D. CHARLESWORTH, 1996 The effect of recombi-
nation on background selection. *Genet. Res.* 67: 159–174.
- 6726 LetterSpace=10NOVEMBRE, J. and LetterSpace=10M. STEPHENS, 2008
Interpreting principal component analyses of spatial population
6728 genetic variation. *Nature genetics* 40(5): 646.
- 6730 LetterSpace=10OHTA, T., 1972 Population size and rate of evolution.
Journal of Molecular Evolution 1(4): 305–314.
- 6732 LetterSpace=10OHTA, T., 1973 Slightly deleterious mutant substitu-
tions in evolution. *Nature* 246(5428): 96.
- 6734 LetterSpace=10OHTA, T., 1987 Very slightly deleterious mutations
and the molecular clock. *Journal of Molecular Evolution* 26(1-2):
1–6.
- 6736 LetterSpace=10OHTA, T. and LetterSpace=10J. H. GILLESPIE, 1996
Development of neutral and nearly neutral theories. *Theoretical
6738 population biology* 49(2): 128–142.
- 6740 LetterSpace=10OWEN, D. and LetterSpace=10D. CHANTER, 1972
Polymorphic mimicry in a population of the African butterfly, *Pseu-*
dacraea eurytus (L.) (Lep. Nymphalidae). *Insect Systematics &*
6742 *Evolution* 3(4): 258–266.
- 6744 LetterSpace=10PAABY, A. B., LetterSpace=10A. O. BERGLAND, Let-
terSpace=10E. L. BEHRMAN, and LetterSpace=10P. S. SCHMIDT, 2014
A highly pleiotropic amino acid polymorphism in the *Drosophila*
6746 insulin receptor contributes to life-history adaptation. *Evolu-*
tion 68(12): 3395–3409.

- 6748 LetterSpace=10PATTERSON, N., LetterSpace=10A. L. PRICE, and LetterSpace=10D. REICH, 2006 Population structure and eigenanalysis.
6750 PLoS genetics 2(12): e190.
- 6752 LetterSpace=10PICKRELL, J. K., LetterSpace=10T. BERISA, LetterSpace=10J. Z. LIU, LetterSpace=10L. SÉGUREL, LetterSpace=10J. Y. TUNG, and LetterSpace=10D. A. HINDS, 2016 Detection and interpretation of shared genetic influences on 42 human traits. Nature genetics 48(7): 709.
- 6754 LetterSpace=10POTTI, J. and LetterSpace=10D. CANAL, 2011 Heritability and genetic correlation between the sexes in a songbird sexual ornament. Heredity 106(6): 945.
- 6756 LetterSpace=10PRITCHARD, J. K., LetterSpace=10M. STEPHENS, and LetterSpace=10P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics 155(2): 945–959.
- 6758 LetterSpace=10PROVINE, W. B., 2001 *The origins of theoretical population genetics: with a new afterword*. University of Chicago Press.
- 6760 LetterSpace=10PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.
- 6762 LetterSpace=10PTAK, S. E., LetterSpace=10A. D. ROEDER, LetterSpace=10M. STEPHENS, LetterSpace=10Y. GILAD, LetterSpace=10S. PÄÄBO, and LetterSpace=10M. PRZEWORSKI, 2004 Absence of the TAP2 human recombination hotspot in chimpanzees.
6764 PLoS biology 2(6): e155.
- 6766 LetterSpace=10QUELLER, D. C., 1992 Quantitative genetics, inclusive fitness, and group selection. The American Naturalist 139(3): 540–558.
- 6768 LetterSpace=10R, 2018 R: A Language and Environment for Statistical Computing.
- 6770 LetterSpace=10RALPH, P. L. and LetterSpace=10G. COOP, 2015 The role of standing variation in geographic convergent adaptation. The American Naturalist 186(S1): S5–S23.
- 6772 LetterSpace=10RANDS, C. M., LetterSpace=10S. MEADER, LetterSpace=10C. P. PONTING, and LetterSpace=10G. LUNTER, 2014 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage.
6774 PLoS genetics 10(7): e1004525.
- 6776 LetterSpace=10RICHARDS, C. M., 2000 Inbreeding depression and genetic rescue in a plant metapopulation. The American Naturalist 155(3): 383–394.

- LetterSpace=10RITLAND, K., LetterSpace=10C. NEWTON, and LetterSpace=10H. MARSHALL, 2001 Inheritance and population structure of the white-phased àÁIJKermodeâÁÍ black bear. *Current Biology* 11(18): 1468 – 1472.
- LetterSpace=10ROBERTSON, A., 1961 Inbreeding in artificial selection programmes. *Genet. Res.* 2: 189âÁS–194.
- LetterSpace=10ROBINSON, J. A., LetterSpace=10D. ORTEGA-DEL VECCHYO, LetterSpace=10Z. FAN, LetterSpace=10B. Y. KIM, LetterSpace=10C. D. MARDEN, LetterSpace=10K. E. LOHMUELLER, LetterSpace=10R. K. WAYNE, and LetterSpace=10OTHERS, 2016 Genomic flatlining in the endangered island fox. *Current Biology* 26(9): 1183–1189.
- LetterSpace=10ROBINSON, L. M., LetterSpace=10J. R. BOLAND, and LetterSpace=10J. M. BRAVERMAN, 2016 Revisiting a Classic Study of the Molecular Clock. *Journal of molecular evolution* 82(2-3): 110–116.
- LetterSpace=10ROSENBERG, N. A., LetterSpace=10J. K. PRITCHARD, LetterSpace=10J. L. WEBER, LetterSpace=10H. M. CANN, LetterSpace=10K. K. KIDD, LetterSpace=10L. A. ZHIVOTOVSKY, and LetterSpace=10M. W. FELDMAN, 2002 Genetic structure of human populations. *science* 298(5602): 2381–2385.
- LetterSpace=10RUWENDE, C., LetterSpace=10S. C. KHOO, LetterSpace=10R. W. SNOW, LetterSpace=10S. N. YATES, LetterSpace=10D. KWIATKOWSKI, LetterSpace=10S. GUPTA, LetterSpace=10P. WARN, LetterSpace=10C. E. ALLSOPP, LetterSpace=10S. C. GILBERT, and LetterSpace=10N. PESCHU, 1995, July)Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* 376(6537): 246–249.
- LetterSpace=10SAMS, A. J. and LetterSpace=10A. R. BOYKO, 2018a Fine-scale resolution and analysis of runs of homozygosity in domestic dogs. *bioRxiv*.
- LetterSpace=10SAMS, A. J. and LetterSpace=10A. R. BOYKO, 2018b Fine-Scale Resolution of Runs of Homozygosity Reveal Patterns of Inbreeding and Substantial Overlap with Recessive Disease Genotypes in Domestic Dogs. *G3: Genes, Genomes, Genetics*: g3–200836.
- LetterSpace=10SANKARARAMAN, S., LetterSpace=10N. PATTERSON, LetterSpace=10H. LI, LetterSpace=10S. PÄÐÄBO, and LetterSpace=10D. REICH, 2012, 10)The Date of Interbreeding between Neandertals and Modern Humans. *PLOS Genetics* 8(10): 1–9.

- 6828 LetterSpace=10SANTIAGO, E. and LetterSpace=10A. CABALLERO,
1995 Effective size of populations under selection. *Genetics* **139**:
1013–1030.
- 6830 LetterSpace=10SANTIAGO, E. and LetterSpace=10A. CABALLERO,
1998 Effective size and polymorphism of linked neutral loci in
6832 populations under directional selection. *Genetics* **149**: 2105–2117.
- 6834 LetterSpace=10SATTATH, S., LetterSpace=10E. ELYASHIV, Let-
terSpace=10O. KOLODNY, LetterSpace=10Y. RINOTT, and Let-
terSpace=10G. SELLA, 2011a Pervasive adaptive protein evolution
6836 apparent in diversity patterns around amino acid substitutions in
Drosophila simulans. *PLoS genetics* **7**(2): e1001302.
- 6838 LetterSpace=10SATTATH, S., LetterSpace=10E. ELYASHIV, Let-
terSpace=10O. KOLODNY, LetterSpace=10Y. RINOTT, and Let-
terSpace=10G. SELLA, 2011b Pervasive adaptive protein evolution
6840 apparent in diversity patterns around amino acid substitutions in
6842 *Drosophila simulans*. *PLoS Genet.* **7**: e1001302.
- 6844 LetterSpace=10SCHEMSKE, D. W. and LetterSpace=10P. BIERZY-
CHUDEK, 2001 Perspective: evolution of flower color in the desert
annual *Linanthus parryae*: Wright revisited. *Evolution* **55**(7): 1269–
1282.
- 6846 LetterSpace=10SEGER, J. and LetterSpace=10H. BROCKMANN, 1987
6848 *Oxford Surveys in Evolutionary Biology*, Volume 4, Chapter What is
bet-hedging, pp. 182–211. Oxford University Press.
- 6850 LetterSpace=10SELLA, G., LetterSpace=10D. A. PETROV, Let-
terSpace=10M. PRZEWORSKI, and LetterSpace=10P. ANDOLFATTO,
6852 2009 Pervasive natural selection in the *Drosophila* genome? *PLoS*
genetics **5**(6): e1000495.
- 6854 LetterSpace=10SHAPIRO, J. A., LetterSpace=10W. HUANG,
LetterSpace=10C. ZHANG, LetterSpace=10M. J. HUBISZ,
6856 LetterSpace=10J. LU, LetterSpace=10D. A. TURISSINI, Let-
terSpace=10S. FANG, LetterSpace=10H. Y. WANG, Let-
terSpace=10R. R. HUDSON, LetterSpace=10R. NIELSEN, Let-
terSpace=10Z. CHEN, and LetterSpace=10C. I. WU, 2007 Adaptive
6860 genetic evolution in the *Drosophila* genomes. *Proc. Natl. Acad. Sci.*
U.S.A. **104**: 2271–2276.
- 6862 LetterSpace=10SMITH, J. M., 1982 *Evolution and the Theory of Games*.
Cambridge university press.
- 6864 LetterSpace=10SMITH, T. B., 1993 Disruptive selection and the ge-
netic basis of bill size polymorphism in the African finch Pyren-
6866 estes. *Nature* **363**(6430): 618.

- 6868 LetterSpace=10SMITHSON, A. and LetterSpace=10M. R. MACNAIR,
 1997 Negative frequency-dependent selection by pollinators on
 artificial flowers without rewards. *Evolution* 51(3): 715–723.
- 6870 LetterSpace=10STEVENS, N. M., 1905 *Studies in Spermatogenesis: With*
especial reference to the "accessory chromosome", Volume 36. Carnegie
 6872 Institution of Washington.
- 6874 LetterSpace=10STUMPF, M. P., LetterSpace=10Z. LAIDLAW, and Let-
 terSpace=10V. A. JANSEN, 2002 Herpes viruses hedge their bets.
 Proceedings of the National Academy of Sciences 99(23): 15234–
 6876 15237.
- 6878 LetterSpace=10STURTEVANT, A. H., 1915 The behavior of the chro-
 mosomes as studied through linkage. *Zeitschrift für induktive
 Abstammungs- und Vererbungslehre* 13(1): 234–287.
- 6880 LetterSpace=10TAJIMA, F., 1989 Statistical method for testing the
 neutral mutation hypothesis by DNA polymorphism. *Genet-*
 6882 *ics* 123(3): 585–595.
- 6884 LetterSpace=10TAO, Y., LetterSpace=10L. ARARIPE, Let-
 terSpace=10S. B. KINGAN, LetterSpace=10Y. KE, Let-
 6886 terSpace=10H. XIAO, and LetterSpace=10D. L. HARTL, 2007 A sex-
 ratio meiotic drive system in *Drosophila simulans*. II: an X-linked
 distorter. *PLoS biology* 5(11): e293.
- 6888 LetterSpace=10TAO, Y., LetterSpace=10J. P. MASLY, Let-
 terSpace=10L. ARARIPE, LetterSpace=10Y. KE, and Let-
 6890 terSpace=10D. L. HARTL, 2007 A sex-ratio meiotic drive system
 in *Drosophila simulans*. I: an autosomal suppressor. *PLoS biol-*
 6892 *ogy* 5(11): e292.
- 6894 LetterSpace=10TISHKOFF, S. A., LetterSpace=10R. VARKONYI,
 LetterSpace=10N. CAHINHINAN, LetterSpace=10S. ABBES, Let-
 terSpace=10G. ARGYROPOULOS, LetterSpace=10G. DESTRO-
 6896 BISOL, LetterSpace=10A. DROUSIOTOU, LetterSpace=10B. DAN-
 GERFIELD, LetterSpace=10G. LEFRANC, LetterSpace=10J. LOISE-
 6898 LET, LetterSpace=10A. PIRO, LetterSpace=10M. STONEKING, Let-
 terSpace=10A. TAGARELLI, LetterSpace=10G. TAGARELLI, Let-
 6900 terSpace=10E. H. TOUMA, LetterSpace=10S. M. WILLIAMS, and
 LetterSpace=10A. G. CLARK, 2001 Haplotype Diversity and Link-
 6902 age Disequilibrium at Human G6PD: Recent Origin of Alleles That
 Confer Malarial Resistance. *Science* 293(5529): 455–462.
- 6904 LetterSpace=10TOEWS, D. P., LetterSpace=10S. A. TAYLOR, Let-
 terSpace=10R. VALLENDER, LetterSpace=10A. BRELSFORD, Let-
 6906 terSpace=10B. G. BUTCHER, LetterSpace=10P. W. MESSER, and

- 6908 LetterSpace=10I. J. LOVETTE, 2016 Plumage Genes and Little Else
Distinguish the Genomes of Hybridizing Warblers. *Current Biology* 26(17): 2313 – 2318.
- 6910 LetterSpace=10TURELLI, M., LetterSpace=10D. W. SCHEMSKE, and
LetterSpace=10P. BIERZYCHUDEK, 2001 Stable two-allele polymor-
6912 phisms maintained by fluctuating fitnesses and seed banks: protect-
ing the blues in *Linanthus parryae*. *Evolution* 55(7): 1283–1298.
- 6914 LetterSpace=10ULIZZI, L. and LetterSpace=10L. TERRENATO, 1992
Natural selection associated with birth weight. VI. Towards the
6916 end of the stabilizing component. *Annals of human genetics* 56(2):
113–118.
- 6918 LetterSpace=10VANĀŽT HOF, A. E., LetterSpace=10N. EDMONDS,
LetterSpace=10M. DALÍKOVÁ, LetterSpace=10F. MAREC, and Let-
terSpace=10I. J. SACCHERI, 2011 Industrial melanism in British
6920 peppered moths has a singular and recent mutational origin. *Sci-
ence* 332(6032): 958–960.
- 6922 LetterSpace=10VOIGHT, B. F., LetterSpace=10A. M. ADAMS,
LetterSpace=10L. A. FRISSE, LetterSpace=10Y. QIAN, Let-
terSpace=10R. R. HUDSON, and LetterSpace=10A. DI RIENZO, 2005
6924 Interrogating multiple aspects of variation in a full resequencing
data set to infer human population size changes. *Proceedings of the
6926 National Academy of Sciences* 102(51): 18508–18513.
- 6930 LetterSpace=10VONHOLDT, B. M., LetterSpace=10J. P. POLLINGER,
LetterSpace=10D. A. EARL, LetterSpace=10J. C. KNOWLES,
LetterSpace=10A. R. BOYKO, LetterSpace=10H. PARKER,
6932 LetterSpace=10E. GEFFEN, LetterSpace=10M. PILOT, Let-
terSpace=10W. JEDRZEJEWSKI, LetterSpace=10B. JEDRZEJEW-
6934 SKA, LetterSpace=10V. SIDOROVICH, LetterSpace=10C. GRECO,
LetterSpace=10E. RANDI, LetterSpace=10M. MUSIANI, Let-
terSpace=10R. KAYS, LetterSpace=10C. D. BUSTAMANTE, Let-
terSpace=10E. A. OSTRANDER, LetterSpace=10J. NOVEMBRE, and
6936 LetterSpace=10R. K. WAYNE, 2011 A genome-wide perspective on
the evolutionary history of enigmatic wolf-like canids. *Genome
6938 Research*.
- 6940 LetterSpace=10WANG, J., LetterSpace=10J. DING, Let-
terSpace=10B. TAN, LetterSpace=10K. M. ROBINSON, Let-
terSpace=10I. H. MICHELSON, LetterSpace=10A. JOHANSSON,
6944 LetterSpace=10B. NYSTEDT, LetterSpace=10D. G. SCOFIELD, Let-
terSpace=10O. NILSSON, LetterSpace=10S. JANSSON, and Let-
terSpace=10OTHERS, 2018 A major locus controls local adaptation

- and adaptive life history variation in a perennial plant. *Genome biology* 19(1): 72.
- 6948 LetterSpace=10WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 7(2): 256–276.
- 6950 LetterSpace=10WEIS, A. E. and LetterSpace=10W. L. GORMAN, 1990 Measuring selection on reaction norms: an exploration of the *Eurosta-Solidago* system. *Evolution* 44(4): 820–831.
- 6952 LetterSpace=10WHEELER, W. M., 1907 Pink Insect Mutants. *The American Naturalist* 41(492): 773–780.
- 6954 LetterSpace=10WIDEMO, F., 1998 Alternative reproductive strategies in the ruff, *Philomachus pugnax*: a mixed ESS? *Animal Behaviour* 56(2): 329–336.
- 6956 LetterSpace=10WIEHE, T. and LetterSpace=10W. STEPHAN, 1993a Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution* 10(4): 842–854.
- 6958 LetterSpace=10WIEHE, T. H. and LetterSpace=10W. STEPHAN, 1993b Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10: 842–854.
- 6960 LetterSpace=10WILKINSON, G. S., 1993 Artificial sexual selection alters allometry in the stalk-eyed fly *Cyrtodiopsis dalmanni* (Diptera: Diopsidae). *Genetics Research* 62(3): 213–222.
- 6962 LetterSpace=10WILLIAMS, G. C., 1966 *Adaptation and Natural Selection*. Princeton.
- 6964 LetterSpace=10WILLIAMS, K.-A. and LetterSpace=10P. S. PENNINGS, 2019 Drug resistance evolution in HIV in the late 1990s: hard sweeps, soft sweeps, clonal interference and the accumulation of drug resistance mutations. *bioRxiv*.
- 6966 LetterSpace=10WISELY, S. M., LetterSpace=10S. W. BUSKIRK, LetterSpace=10M. A. FLEMING, LetterSpace=10D. B. McDONALD, and LetterSpace=10E. A. OSTRANDER, 2002 Genetic Diversity and Fitness in Black-Footed Ferrets Before and During a Bottleneck. *Journal of Heredity* 93(4): 231–237.
- 6968 LetterSpace=10WRIGHT, K. M., LetterSpace=10U. HELLSTEN, LetterSpace=10C. XU, LetterSpace=10A. L. JEONG, LetterSpace=10A. SREEDASYAM, LetterSpace=10J. A. CHAPMAN,
- 6970

- 6986 LetterSpace=10J. SCHMUTZ, LetterSpace=10G. COOP, Let-
terSpace=10D. S. ROKHSAR, and LetterSpace=10J. H. WILLIS, 2015
Adaptation to heavy-metal contaminated environments proceeds
6988 via selection on pre-existing genetic variation. bioRxiv: 029900.
- 6990 LetterSpace=10WRIGHT, S., 1943 Isolation by Distance. Genet-
ics 28(2): 114–138.
- 6992 LetterSpace=10WRIGHT, S., 1949 The Genetical Structure of Popula-
tions. Annals of Eugenics 15(1): 323–354.
- 6994 LetterSpace=10WRIGHT, S. and LetterSpace=10T. DOBZHANSKY, 1946
Genetics of natural populations. XII. Experimental reproduction of
some of the changes caused by natural selection in certain popula-
6996 tions of *Drosophila pseudoobscura*. Genetics 31(2): 125.
- 6998 LetterSpace=10WRIGHT, S. I., LetterSpace=10I. V. BI, Let-
terSpace=10S. G. SCHROEDER, LetterSpace=10M. YAMASAKI, Let-
terSpace=10J. F. DOEBLEY, LetterSpace=10M. D. McMULLEN, and
7000 LetterSpace=10B. S. GAUT, 2005 The Effects of Artificial Selection on
the Maize Genome. Science 308(5726): 1310–1314.
- 7002 LetterSpace=10YANG, Z., 1998 Likelihood ratio tests for detecting
positive selection and application to primate lysozyme evolution.
7004 Molecular Biology and Evolution 15(5): 568–573.
- 7006 LetterSpace=10ZUCKERKANDL, E. and LetterSpace=10L. PAULING,
1965 Evolutionary divergence and convergence in proteins. In
Evolving genes and proteins, pp. 97–166. Elsevier.