

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](mailto: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<sup>A</sup>T<sub>E</sub>X and the TUFTE-LATEX book style.

The L<sup>A</sup>T<sub>E</sub>X 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 April 2019*

This book was developed from my set of notes for the Population Biology graduate group core class (PBGG) 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 R (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>	7
2	<i>Allele and Genotype Frequencies</i>	11
3	<i>Genetic Drift and Neutral Diversity</i>	45
4	<i>Phenotypic Variation and the Resemblance Between Relatives.</i>	89
5	<i>The Response to Phenotypic Selection</i>	109
6	<i>One-Locus Models of Selection</i>	123
7	<i>The Impact of Genetic Drift on Selected Alleles</i>	159
8	<i>The Effects of Linked Selection.</i>	171
9	<i>Bibliography</i>	187



# 1

## *2 Introduction*

BIOLOGICAL EVOLUTION IS THE CHANGE OVER TIME IN THE  
4 GENETIC COMPOSITION OF A POPULATION.<sup>1</sup> 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 of indi-  
8 viduals or the migration of individuals in or out of the population. If  
our individuals vary in the number of children they have, this also al-  
10 ters the genetic composition of the population in the next generation.  
Every new individual born into the population subtly changes the  
12 genetic composition of the population. Their genome is a unique com-  
bination of their parents' genomes, having been shuffled by segregation  
14 and recombination during meioses, and possibly changed by mutation.  
These individual events seem minor at the level of the population, but  
16 it is the accumulation of small changes in aggregate across individuals  
and generations that is the stuff of evolution. It is the compounding  
18 of these small changes over tens, hundreds, and millions of genera-  
tions that drives the amazing diversity of life that has emerged on this  
20 earth.

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

32 While the models we will develop will seem naïve, and indeed they  
are, they are nonetheless incredibly useful and powerful. Throughout

<sup>1</sup> DOBZHANSKY, T., 1951 *Genetics and the Origin of Species* (3rd Ed. ed.), pp. 16

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

<sup>34</sup> the course we will see that these simple models often yield accurate predictions, such that much of our understanding of the process of evolution is built on these models. We will also see how these models are incredibly useful for understanding real patterns we see in the evolution of phenotypes and genomes, such that much of our analysis of evolution, in a range of areas from human medical genetics to conservation, is based on these models. Therefore, population and quantitative genetics are key to understanding various applied questions, from how medical genetics identifies the genes involved in disease to how we preserve species from extinction.

<sup>44</sup> 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.

<sup>56</sup> 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.

<sup>68</sup> Population and quantitative genetics are a necessary but not a 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 do we need to know the precise selection pressures and the ordering of genetic

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

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

“DOBZHANSKY (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”

– LEWONTIN

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



# 2

## <sup>86</sup> *Allele and Genotype Frequencies*

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

<sup>90</sup> Loci and alleles are the basic currency of population genetics—and indeed of genetics. If all individuals in the population carry the same  
<sup>92</sup> allele, we say that the locus is *monomorphic*; at this locus there is no genetic variability in the population. If there are multiple alleles in  
<sup>94</sup> the population at a locus, we say that this locus is *polymorphic* (this is sometimes referred to as a segregating site).

<sup>96</sup> Table 2.1 show a small stretch orthologous sequence for the ADH locus from samples from *Drosophila melanogaster*, *D. simulans*, and <sup>98</sup> *D. yakuba*. *D. melanogaster* and *D. simulans* are sister species and *D. yakuba* is a close outgroup to the two. Each column represents a <sup>100</sup> single haplotype from an individual (the individuals are diploid but were inbred so they're homozygous for their haplotype). Only sites that differ among individuals of the three species are shown. Site 834 <sup>102</sup> is an example of a polymorphism; some *D. simulans* individuals carry a *C* allele while others have a *T*. Fixed differences are sites that differ between the species but are monomorphic within the species. Site 781 <sup>104</sup> is an example of a fixed difference between *D. melanogaster* and the other two species.

<sup>106</sup> We can also annotate the alleles and loci in various ways. For example, position 781 is a non-synonymous fixed difference. We call the <sup>108</sup> less common allele at a polymorphism the *minor allele* and the common allele the *major allele*, e.g. at site 1068 the *T* allele is the minor <sup>110</sup> allele in *D. melanogaster*. We call the more evolutionarily recent of the two alleles the *derived allele* and the older of the two the *ancestral allele*. The *T* allele at site 1068 is the derived allele as the *C* is found in <sup>112</sup> both the other species, suggesting that the *T* allele arose via a *C → T* mutation.

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*.

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

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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NS	
789	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
808	A	-	-	-	-	-	-	-	-	-	T	T	T	G	G	G	G	G	G	G	G	G	G	G	NS	
816	G	T	T	T	T	-	-	-	-	-	-	-	C	C	-	-	-	-	-	-	-	-	-	-	S	
834	T	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	S	
859	C	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS	
867	C	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	A	G	G	G	G	G	G	S	
870	C	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
950	G	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	S	
974	G	-	-	-	-	-	-	-	-	-	T	-	T	T	T	T	-	-	-	-	-	-	-	-	S	
983	T	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S	
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	S	
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S	
1034	T	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	-	C	-	C	C	C	S		
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S		
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S	
1089	C	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	-	-	-	-	-	-	NS		
1101	G	-	-	-	-	-	-	-	-	-	-	-	A	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	C	S	
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	S		
1160	T	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S	

- from *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 ( $\pi$ )** 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  $\pi$ . For example, we can calculate  $\pi$  for our ADH locus from Table 2.1 above: we have 6 sequences from *D. simulans* (a-f), there's a total of 15 ways of pairing these sequences, and

Table 2.1: Variable sites in exons 2 and 3 of the ADH gene in *Drosophila* McDONALD and KREITMAN (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

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

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

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

Number of segregating sites. Another measure of genetic variability  
 158 is the total number of sites that are polymorphic (segregating) in our  
 sample. One issue is that the number of segregating sites will grow  
 160 as we sequence more individuals (unlike  $\pi$ ). Later in the course, we'll  
 talk about how to standardize the number of segregating sites for the  
 162 number of individuals sequenced (see eqn (3.39)).

The frequency spectrum. We also often want to compile information  
 164 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  
 166 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  
 168 in the sample, and this is called the frequency spectrum. We'll want  
 to do this in some consistent manner, so we often calculate the minor  
 170 allele frequency spectrum, or the frequency spectrum of derived alleles.

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

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

<sup>1</sup> 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.

with vastly different census sizes. This observation represented a puzzle as Neutral theory predicts that levels of genetic diversity should scale 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.

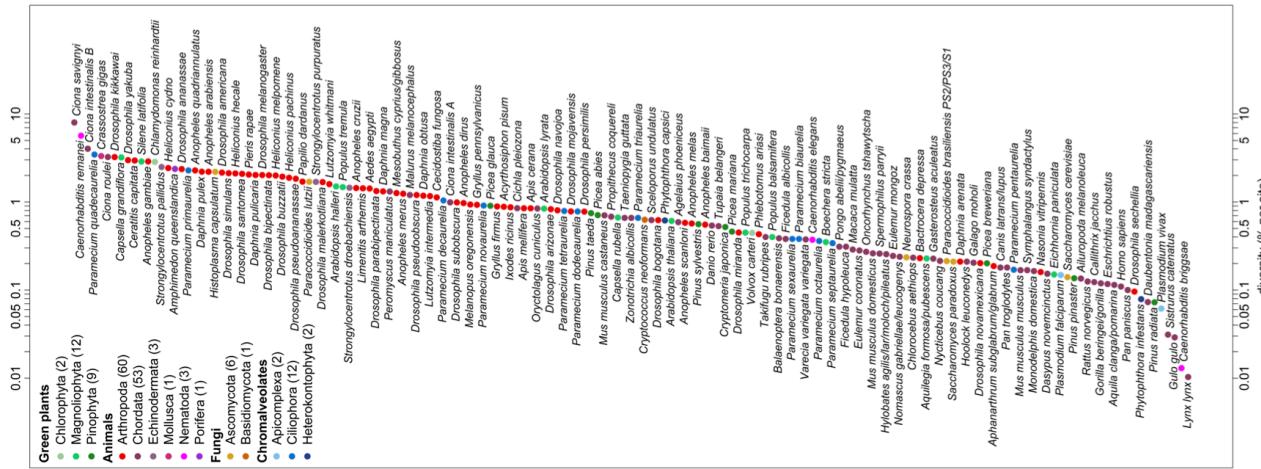
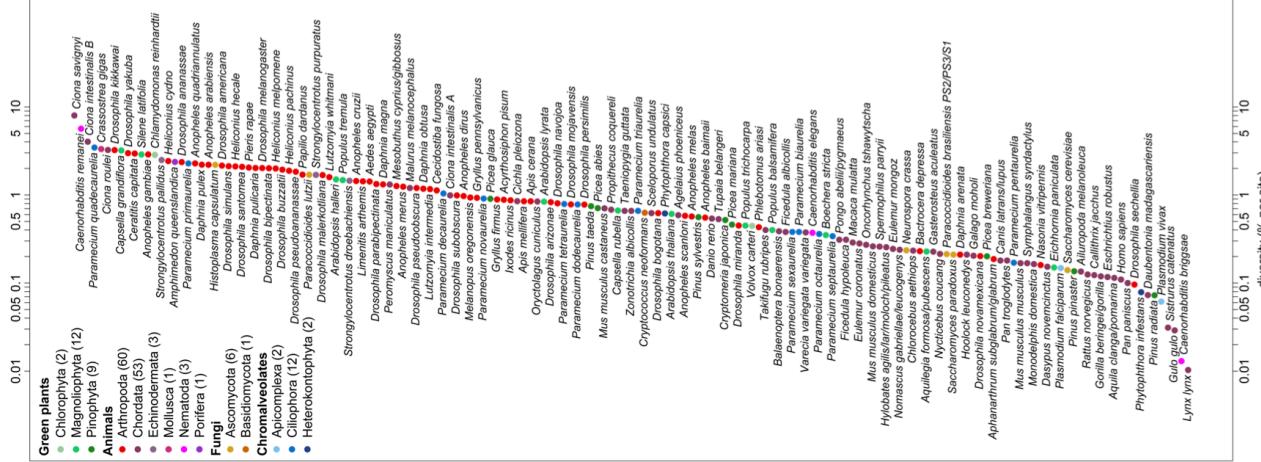


Figure 2.1: Sea Squirt (*Ciona intestinalis*).

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.



For example, LEFFLER *et al.* (2012) compiled data on levels of within-population, autosomal nucleotide diversity ( $\pi$ ) for 167 species across 14 phyla from non-coding and synonymous sites (Figure 2.2). The species with the lowest levels of  $\pi$  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 probability is also the expected frequencies of the  $A_1A_1$  homozygote in the popula-

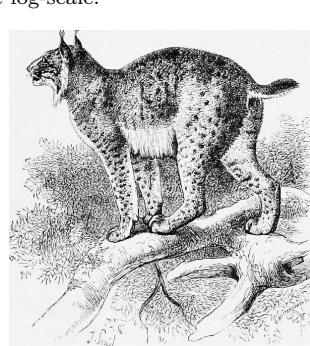


Figure 2.3: Eurasian Lynx (*Lynx lynx*).

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.

208 tion. The expected frequency of the three possible genotypes are

$$\begin{array}{ccc} f_{11} & f_{12} & f_{22} \\ \hline p^2 & 2pq & q^2 \end{array}$$

210 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  
212 zygotes forming the next generation. Evolutionary forces, such as  
selection, change allele frequencies within generations, but do not  
214 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.

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

Below are the genotype counts for the MC1R polymorphism in  
224 a sample of bears from British Columbia's island populations from  
RITLAND *et al.*.

	<i>AA</i>	<i>AG</i>	<i>GG</i>
226	42	24	21

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

See Figure 2.5 for a nice empirical demonstration of Hardy-Weinberg  
230 proportions. The mean frequency of each genotype closely match their  
HW expectations, and much of the scatter of the dots around the ex-  
232 pected line is due to our small sample size ( $\sim 60$  individuals). While  
HW often seems like a silly model, it often holds remarkably well  
234 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  
236 the loci in the genome.

238 **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?

240 Microsatellites are regions of the genome where individuals vary  
for the number of copies of some short DNA repeat that they carry.  
242 These regions are often highly variable across individuals, making  
them a suitable way to identify individuals from a DNA sample. This  
244 so-called DNA-fingerprinting has a range of applications from estab-  
lishing paternity, identifying human remains, to matching individuals  
246 to DNA samples from a crime scene. The FBI make use of the CODIS



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](#).

database<sup>2</sup>. 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

**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.

- 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)?
- B) The FBI uses  $\geq 13$  markers. Why is this higher number necessary to make the match statement convincing evidence in court?
- C) An early case that triggered debate among forensic geneticists was a crime among the Abenaki, a Native American community in Vermont (see LEWONTIN, 1994, for discussion). There was a DNA sample from the crime scene, and the perpetrator was thought likely

<sup>2</sup> CODIS: Combined DNA Index System

Table 2.2: Data for 155 Europeans at the D16S539 microsatellite from CODIS from ALGEE-HEWITT *et al.*. 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 to be a member of the Abenaki community. Given that allele frequencies vary among populations, why would people be concerned about  
 268 using data from a non-Abenaki population to compute a false match probability?

270 *2.2 Allele sharing among related individuals and Identity by Descent*

272 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 population these relationships are very distant (e.g. distant cousins),  
 274 while some individuals will be more closely related (e.g. sibling/first  
 276 cousins). All individuals are related to one another by varying levels  
 278 of relatedness, or *kinship*. Related individuals can share alleles that  
 have both descended from the shared common ancestor. To be shared,  
 280 these alleles must be inherited through all meioses connecting the two  
 individuals (e.g. surviving the  $1/2$  probability of segregation each meiosis). As closer relatives are separated by fewer meioses, closer relatives  
 282 share more alleles. In Figure 2.6 we show the sharing of chromosomal  
 284 regions between two cousins. As we'll see, many population and quantitative genetic concepts rely on how closely related individuals are,  
 and thus we need some way to quantify the degree of kinship among  
 286 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.

288 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 generations<sup>3</sup>. For the moment, we ignore mutation, and we will  
 290 be more precise about what we mean by 'past few generations' later  
 on. For example, parent and child share exactly one allele identical  
 292 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  
 294 show a pedigree demonstrating some configurations of IBD.

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

One summary of how related two individuals are is the probability  
 296 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  $r_2$  re-  
 298 spectively. See Table 2.4 for some examples. We can also interpret  
 300 these probabilities as genome-wide averages. For example, on aver-  
 age, at a quarter of all their autosomal loci full-sibs share zero alleles  
 identical by descent.

302 One summary of relatedness that will be important is the prob-  
 ability that two alleles picked at random, one from each of the two  
 304 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  
 306 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 dis-  
 308 cussion 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	$1/4$
full siblings	$1/4$	$1/2$	$1/4$	$1/4$
Monzygotic twins	0	0	1	$1/2$
1 <sup>st</sup> cousins	$3/4$	$1/4$	0	$1/16$

310 **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 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

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

Or, in our  $r_0$ ,  $r_1$ ,  $r_2$  notation:

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



Figure 2.7: A pair of diploid individu-  
 als (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.

312 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  
 314 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  
 316  $p$ , and then the other non-IBD allele, in both relatives, also needs to  
 be  $A_1$  which happens with probability  $p^2$ , so  $P(A_1A_1|1 \text{ alleles IBD}) =$   
 318  $p \times p^2$ . 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  
 320 homozygous for the  $A_1$  allele, both individuals will be. Putting this all  
 together our equation (2.5) becomes

$$P(A_1A_2) = p^4r_0 + p^3r_1 + p^2r_2 \quad (2.6)$$

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

326 We can write out terms like eq (2.6) for all of the possible configura-  
 tions of genotype sharing/non-sharing between a pair of individuals.  
 328 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.

330 **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  
 332 a match to the DNA from the crime scene; however, they could be a  
 sibling.

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

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

340 There's a variety of ways to estimate the relationships among in-  
 dividuals using genetic data. An example of using allele sharing to  
 identify relatives is offered by the work of Nancy Chen (in collabo-  
 342 ration with Stepfanie Aguillon, see CHEN *et al.*, 2016; AGUILLO  
*et al.*, 2017). CHEN *et al.* has collected genotyping data from thou-  
 344 sandes of Florida Scrub Jays at over ten thousand loci. These Jays  
 live at the Archbold field site, and have been carefully monitored for  
 346 many decades allowing the pedigree of many of the birds to be known.  
 Using these data she estimates allele frequencies at each locus. Then  
 348 by equating the observed number of times that a pair of individuals  
 share 0, 1, or 2 alleles to the theoretical expectation, she estimates  
 350 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  
 352 known from the pedigree.



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.



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.



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.

*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 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.

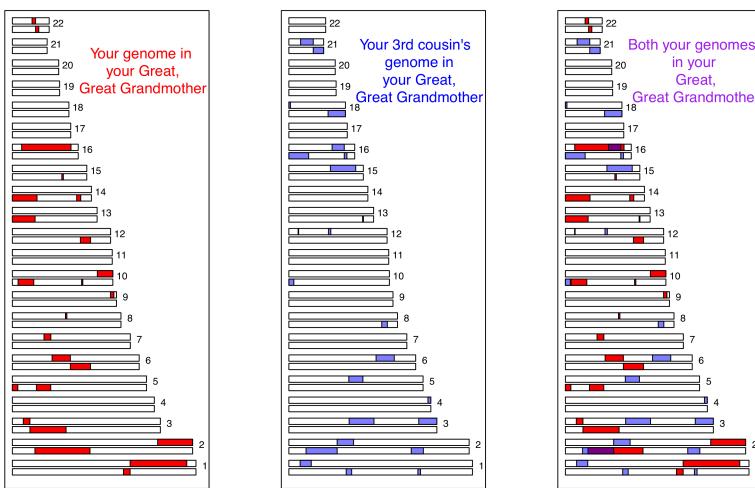


Figure 2.11: A simulation of sharing between third cousins, the details are the same as in Figure 2.10.

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.<sup>4</sup> More distant relations than third cousins, e.g. fourth cousins, start to have

<sup>4</sup> 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).

- 388 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  
 390 with some of them; however, it gets increasingly hard to identify the  
 degree of relatedness from genetic data the deeper in the family tree  
 392 this sharing goes.

### 2.2.1 Inbreeding

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

When two related individuals produce an offspring, that individual can receive two alleles that are identical by descent, i.e. they can be homozygous by descent (sometimes termed autozygous), due to the fact that they have two copies of an allele through different paths through the pedigree. This increased likelihood of being homozygous relative to an outbred individual is the most obvious effect of 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 cousins is homozygous by descent having received the same IBD allele via two different routes around an inbreeding loop.

As the offspring receives a random allele from each parent ( $i$  and  $j$ ), the probability that those two alleles are identical by descent is 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.

$f_{11}$	$f_{12}$	$f_{22}$
$(1 - F)p^2 + Fp$	$(1 - F)2pq$	$(1 - F)q^2 + Fq$

The only way the offspring can be heterozygous ( $A_1A_2$ ) is if their  
 414 two alleles at a locus are not IBD (otherwise they would necessarily be  
 homozygous). Therefore, the probability that they are heterozygous is

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

416 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.  
 418 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  
 420 inherited allele  $A_1$  by two different routes from the same ancestor.  
 Thus, the probability that an offspring is homozygous for  $A_1$  is

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



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

Note that the generalized Hardy–Weinberg proportions completely  
426 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).  
428 Therefore, any combination of genotype frequencies at a biallelic site  
can be specified by a combination of  $p$  and  $F$ .

430 **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  
432 genotype frequencies in the population follow from HW. We select two  
individuals at random to mate from this population. We then mate  
434 the children from this cross. What is the probability that the child  
from this full sib-mating is homozygous?

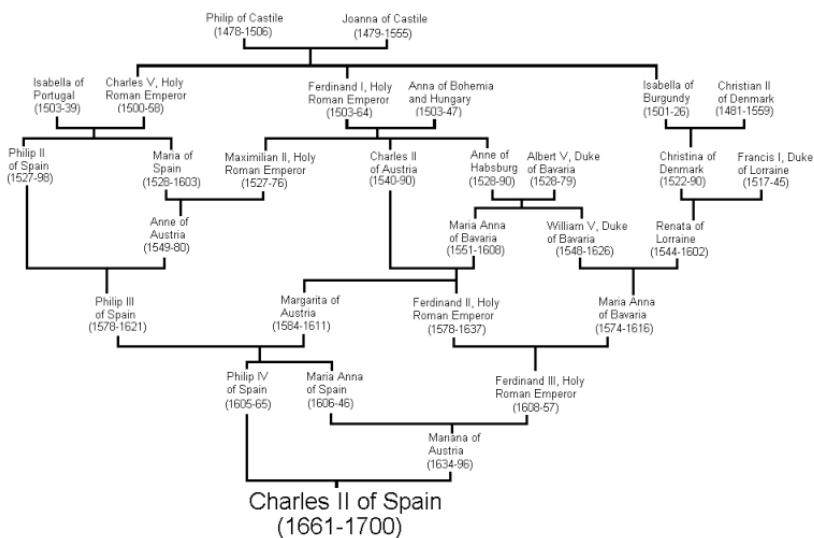
436 *Multiple inbreeding loops in a pedigree.* Up to this point we have as-  
sumed that there is at most one inbreeding loop in the recent family  
438 history of our individuals, i.e. the parents of our inbred individual  
have at most one recent genealogical connection. However, an indi-  
440 vidual who has multiple inbreeding loops in their pedigree can be  
homozygous by descent thanks to receiving IBD alleles via multiple  
442 different different loops. To calculate inbreeding in pedigrees of ar-  
bitrary complexity, we can extend beyond our original relatedness  
444 coefficients  $r_0$ ,  $r_1$ , and  $r_2$  to account for higher order sharing of alleles  
IBD among relatives. For example, we can ask, what is the probability  
446 that *both* of the alleles in the first individual are shared IBD with one  
allele in the second individual? There are nine possible relatedness  
448 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 keep track  
450 of. However, we will show how we can calculate the inbreeding coeffi-  
cient of an individual with multiple inbreeding loops more directly.

452 Let's say the parents of our inbred individual (B and C) have  $K$   
shared ancestors, i.e. individuals who appear in both B and C's recent  
454 family trees. We denote these shared ancestors by  $A_1, \dots, A_K$ , and  
we denote by  $n$  the total number of individuals in the chain from B  
456 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,  
458 equivalently, C's grandparents. In this case, there are n=4 individuals  
from B to C through each of these two shared ancestor. In the general  
460 case, the kinship coefficient of B and C, i.e. the inbreeding coefficient  
of their child, is

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

where  $f_{A_i}$  is the inbreeding coefficient of the ancestor  $A_i$ . What's 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 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 case the alleles are IBD.<sup>5</sup> 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.

<sup>5</sup> 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).



486 a child of an uncle-niece marriage.

488 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 490 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 492 homozygous for a number of recessive disease alleles, and indeed he was a very sickly man who left no descendants due to his infertility.<sup>6</sup> 494 Thus plausibly the end of one of the great European dynasties came about through inbreeding.

496 *2.2.2 Calculating inbreeding coefficients from genetic data*

If the observed heterozygosity in a population is  $H_O$ , and we assume 498 that the generalized Hardy–Weinberg proportions hold, we can set  $H_O$  equal to  $f_{12}$ , and solve Eq. (2.7) for  $F$  to obtain an estimate of the 500 inbreeding coefficient as

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

502 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 the heterozygosity expected in the absence of inbreeding,  $H_E = 2pq$ , as

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

504 Hence,  $\hat{F}$  quantifies the deviation due to inbreeding of the observed heterozygosity from the one expected under random mating, relative 506 to the latter.

**Question 9.** Suppose the following genotype frequencies were observed 508 for an esterase locus in a population of *Drosophila* (A denotes the “fast” allele and B denotes the “slow” allele):

	AA	AB	BB
510	0.6	0.2	0.2

What is the estimate of the inbreeding coefficient at the esterase locus? 512

If we have multiple loci, we can replace  $H_O$  and  $H_E$  by their means 514 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 average 516 across loci. However, this procedure is more prone to introducing a bias if sample sizes vary across loci, which is not unlikely when we 518 are dealing with real data.

Genetic markers are commonly used to estimate inbreeding for wild 520 and/or captive populations of conservation concern. As an example of

<sup>6</sup> Pedro Gargantilla, who performed Charles' 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.

this, consider the case of the Mexican wolf (*Canis lupus baileyi*), also known as the lobo, a sub-species of gray wolf.

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



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.

Figure 2.16: .

552 small blocks of autozygosity. For example, Carlos the second had an  
553 inbreeding coefficient that is equivalent to that of the child of full-sibs,  
554 with a quarter of his genome expected to homozygous by descent, but  
555 this would be made up of many shorter blocks.

556 We can hope to detect these blocks by looking for unusually long  
557 genomic runs of homozygosity (ROH) sites in an individual's genome.  
558 One way to estimate an individual's inbreeding coefficient is then to  
559 total up the proportion of an individual's genome that falls in such  
560 ROH regions. This estimate is called  $F_{ROH}$ .

561 An example of using  $F_{ROH}$  to study inbreeding comes from the  
562 work of SAMS and BOYKO (2018b), who identified runs of homozy-  
563 gosity in 2,500 dogs, ranging from 500kb up to many megabases. Fig-



Figure 2.17: English bulldog. The  
dogs of Boytowm. 1918. Dyer, W. A.



Figure 2.18: The distribution of  
 $F_{ROH}$  of individuals from various  
dog breeds from SAMS and BOYKO  
(2018a), licensed under CC BY 4.0.

564 ure 2.18 shows the distribution of  $F_{ROH}$  of individuals in each dog  
565 breed for the X and autosome. In Figure 2.19 this is broken down by  
566 the length of ROH segments.

567 Dog breeds have been subject to intense breeding that has resulted  
568 in high levels of inbreeding. Of the population samples examined,  
569 Doberman Pinschers have the highest levels of their genome in runs  
570 of homozygosity ( $F_{ROH}$ ), somewhat higher than English bulldogs.  
571 In 2.19 we can see that English bulldogs have more short ROH than  
572 Doberman Pinschers, but that Doberman Pinschers have more of their  
573 genome in very large ROH (> 16Mb). This suggests that English bull-  
574 dogs have had long history of inbreeding but that Doberman Pinschers  
575 have a lot of recent inbreeding in their history.



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

### *576 2.3 Summarizing population structure*

INDIVIDUALS RARELY MATE COMPLETELY AT RANDOM; your  
 578 parents weren't two Bilateria plucked at random from the tree of life.  
 Even within species, there's often geographically-restricted mating  
 580 among individuals. Individuals tend to mate with individuals from the  
 same, or closely related sets of populations. This form of non-random  
 582 mating is called population structure and can have profound effects  
 on the distribution of genetic variation within and among natural  
 584 populations.

#### *2.3.1 Inbreeding as a summary of population structure.*

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

WRIGHT<sup>7</sup> developed a set of ‘F-statistics’ (also 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  $Y$ . We will return to why  $F$ -statistics are statements about correlations between alleles in just a moment. One commonly used  $F$ -statistic is  $F_{IS}$ , which is the inbreeding coefficient between an individual ( $I$ ) and 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

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

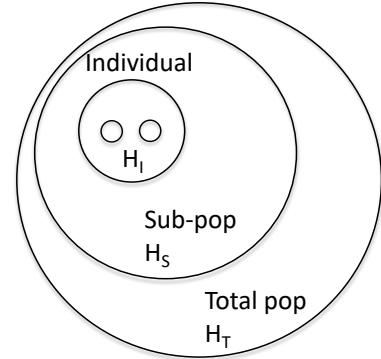


Figure 2.20: The hierarchical nature of F-statistics. The two dots within an individual represent the two alleles at a locus for an individual  $I$ . We can compare the heterozygosity on individuals ( $H_I$ ), to that found by randomly drawing alleles from the sub-population (S), to that found in the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

632  $K$  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  $\bar{H}_S \leq H_T$ ,<sup>8</sup> and so  $F_{ST} \geq 0$  and  $F_{IS} \leq F_{IT}$ . Furthermore, if we have multiple sites, we can replace  $H_I$ ,  $H_S$ , and  $H_T$  with their averages across loci (as above).<sup>9</sup>

638 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 golden-winged warblers (*Vermivora cyanoptera* and *V. chrysoptera* 1-2 & 5-6 o, Figure 2.21).

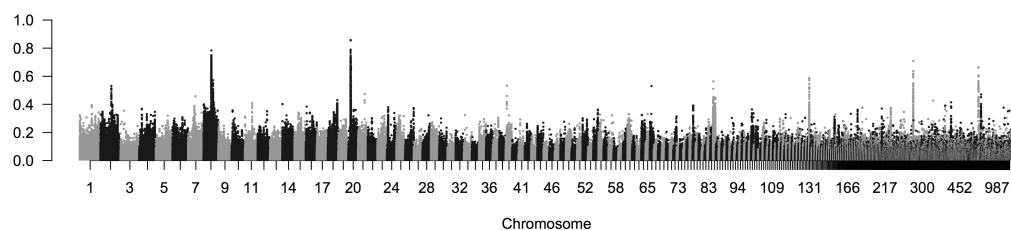
642 These two species are spread across eastern Northern America, with the golden-winged warbler having a smaller, more northerly range. 644 They're quite different in terms of plumage, but have long been known to have similar songs and ecologies. The two species hybridize readily 646 in the wild; in fact two other previously-recognized species, Brewster's and Lawrence's warbler (4 & 3 in 2.21), are actually found to just 648 be hybrids between these two species. The golden-winged warbler is listed as 'threatened' under the Canadian endangered species act. 650 The golden-winged warbler's habitat is under pressure from human activity and increased hybridization with the blue warbler, which is 652 moving north into its range, also poses a significant issue. TOEWS 654 *et al.* investigated the population genomics of these warblers, sequencing ten golden- and ten blue-winged warblers. They found very low divergence among these species, with a genome-wide  $F_{ST} = 0.0045$ . 656 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

<sup>8</sup> 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.

<sup>9</sup> Averaging heterozygosity across loci first, then calculating  $F_{ST}$ , rather than calculating  $F_{ST}$  for each locus individually and then taking the average, has better statistical properties as statistical noise in the denominator is averaged out.



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.



658 high  $F_{ST}$  stand out. Nearly all of these regions correspond to large 660 allele frequency difference at loci in, or close, to genes known to be 662 involved in plumage colouration difference in other birds. To illustrate these frequency differences TOEWS *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

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.

664 gene involved in feather development:

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

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

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

672 *Interpretations of F-statistics* Let us now return to Wright's definition of the  $F$ -statistics as correlations between random gametes, drawn from the same level  $X$ , relative to level  $Y$ . Without loss of generality, 674 we may think about  $X$  as individuals and  $S$  as the subpopulation. Rewriting  $F_{IS}$  in terms of the observed homozygote frequencies ( $f_{11}$ , 676  $f_{22}$ ) and expected homozygosities ( $p_S^2, q_S^2$ ) we find

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

678 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 of alleles found in an individual, divided by the expected variance 680 under binomial sampling. Thus,  $F$ -statistics can be understood as the correlation between alleles drawn from a population (or an individual) above that expected by chance (i.e. drawing alleles sampled at random 682 from some broader population).

684 We can also interpret  $F$ -statistics as proportions of variance explained by different levels of population structure. To see this, let 686 us think about  $F_{ST}$  averaged over  $K$  subpopulations, whose frequencies are  $p_1, \dots, p_K$ . The frequency in the total population is 688  $p_T = \bar{p} = \frac{1}{K} \sum_{i=1}^K p_i$ . Then, we can write

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

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

### 2.3.2 Other approaches to population structure

692 There is a broad spectrum of methods to describe patterns of population structure in population genetic datasets. We'll briefly discuss two 694 broad-classes of methods that appear often in the literature: assignment methods and principal components analysis.

## 696 2.3.3 Assignment Methods

Here we'll describe a simple probabilistic assignment to find the probability that an individual of unknown population comes from one of  $K$  predefined populations. For example, there are three broad populations of common chimpanzee (*Pan troglodytes*) in Africa: western, central, and eastern. Imagine that we have a chimpanzee, whose 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 of individuals representative of these populations, we can calculate the probability that our chimp comes from each of these populations.

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

710 *A simple assignment method* We have genotype data from unlinked  $S$  biallelic loci for  $K$  populations. The allele frequency of allele  $A_1$  at 712 locus  $l$  in population  $k$  is denoted by  $p_{k,l}$ , so that the allele frequencies in population 1 are  $p_{1,1}, \dots, p_{1,L}$  and population 2 are  $p_{2,1}, \dots, p_{2,L}$  and 714 so on.

You genotype a new individual from an unknown population at 716 these  $L$  loci. This individual's genotype at locus  $l$  is  $g_l$ , where  $g_l$  denotes the number of copies of allele  $A_1$  this individual carries at this 718 locus ( $g_l = 0, 1, 2$ ).

The probability of this individual's genotype at locus  $l$  conditional 720 on coming from population  $k$ , i.e. their alleles being a random HW draw from population  $k$ , is

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

722 Assuming that the loci are independent, the probability of the individual's genotype across all  $S$  loci, conditional on the individual 724 coming from population  $k$ , is

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

We wish to know the probability that this new individual comes 726 from population  $k$ , i.e.  $P(\text{pop } k | \text{ind.})$ . We can obtain this through Bayes' rule

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

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

728 where

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

is the normalizing constant. We interpret  $P(\text{pop } k)$  as the prior probability of the individual coming from population  $k$ , and unless we have some other prior knowledge we will assume that the new individual has an equal probability of coming from each population  $P(\text{pop } k) = 1/K$ .

734 We interpret

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

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

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 come from population  $k$  and estimate the set of  $q_k$ 's.

#### 740 Question 11.

Returning to our chimp example, imagine that we have genotyped a set of individuals from the Western and Eastern populations at two 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

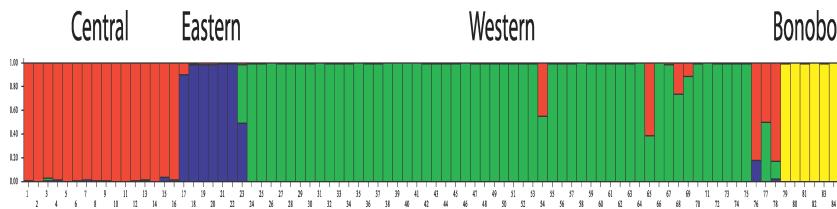
746 **A)** Our individual, whose origin is unknown, has the genotype  $AA$  at the first locus and  $bb$  at the second. What is the posterior probability 748 that our individual comes from the Western population versus Eastern chimp population?

750 **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 752 population and with probability  $q_C = 1 - q_W$  they draw an allele from the Eastern population. What is the probability of our individual's 754 genotype given  $q_C$ ?

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

*Clustering based on assignment methods* While it is great to be able 758 to assign our individuals to a particular population, these ideas can be pushed to learn about how best to describe our genotype data in 760 terms of discrete populations without assigning any of our individuals to populations *a priori*. We wish to cluster our individuals into  $K$  762 unknown populations. We begin by assigning our individuals at random to these  $K$  populations.

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



772 To do this in a full Bayesian scheme we need to place priors on  
774 the allele frequencies (for example, one could use a beta distribution  
prior). Technically we are using the joint posterior of our allele fre-  
776 quencies and assignments. Programs like STRUCTURE, use this type  
of algorithm to cluster the individuals in an “unsupervised” manner  
778 (i.e. they work out how to assign individuals to an unknown set of  
populations). See Figure 2.23 for an example of Becquet *et al* using  
780 STRUCTURE to determine the population structure of chimpanzees.

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

There is no right choice of K, the number of clusters to partition  
790 into. There are methods of judging the ‘best’ K by some statistical  
measure given some particular dataset, but that is not the same as  
792 saying this is the most meaningful level on which to summarize pop-  
ulation structure in data. For example, running STRUCTURE on  
794 world-wide human populations for low value of K will result in popula-  
tion clusters that roughly align with continental populations (ROSEN-  
796 BERG *et al.*, 2002). However, that does not tell us that assigning  
ancestral at the level of continents is a particularly meaningful way of  
798 partitioning individuals. Running the same data for higher value of K,

Figure 2.23: BECQUET *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 BECQUET *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.

or within continental regions, will result in much finer-scale partitioning of continental groups (ROSENBERG *et al.*, 2002; LI *et al.*, 2008).  
 800 No one of these layers of population structure identified is privileged  
 802 as being more meaningful than another.

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

### 2.3.4 Principal components analysis

810 Principal component analysis (PCA) is a common statistical approach  
 812 to visualize high dimensional data, and used by many fields. The idea  
 814 of PCA is to give a location to each individual data-point on each of  
 816 a small number principal component axes. These PC axes are chosen  
 818 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,  
 818 PCA has made come back.<sup>11</sup>

Consider a dataset consisting of  $N$  individuals at  $S$  biallelic SNPs.  
 820 The  $i^{th}$  individual's genotype data at locus  $\ell$  takes a value  $g_{i,\ell} =$   
 822 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  $\ell$  by  $p_\ell$ , it's  
 824 common to standardize the genotype in the following way

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

826 i.e. at each SNP we center the genotypes by subtracting the mean  
 828 genotype ( $2p_\ell$ ) and divide through by the square root of the expected  
 variance assuming that alleles are sampled binomially from the mean  
 830 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  
 832 components analysis of this data matrix to uncover the major axes  
 of genotype variance in our sample. Figure 2.24 shows a PCA from  
 BECQUET *et al.* (2007) using the same chimpanzee data as in Figure  
 834 2.23.

It is worth taking a moment to delve further into what we are doing  
 836 here. There's a number of equivalent ways of thinking about what  
 PCA is doing. One of these ways is to think that when we do PCA we  
 838 are building the individual by individual covariance matrix and per-

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



Figure 2.24: Principal Component Analysis by BECQUET *et al.* (2007) using the same chimpanzee data as in Figure 2.23. Here BECQUET *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). PCA, The 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) are 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.

forming 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. ALCAIDE *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 ALCAIDE *et al.* (2014). The samples are coloured by the subspecies. Code here.

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

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

We can then perform PCA on this kinship matrix to identify the  
 874 major axes of variation in the dataset. Figure 2.28 shows the sam-  
 ples plotted on the first two PCs. The two major routes of expansion  
 876 clearly occupy different parts of PC space. The first principal com-  
 ponent distinguishes populations running North to South along the  
 878 western route of expansion, while the second principal component  
 distinguishes among populations running North to South along the  
 880 Eastern route of expansion. Thus genetic data supports the hypoth-  
 esis that the Greenish warblers speciated as they moved around the  
 882 Himalayan plateau. However, as noted by ALCAIDE *et al.* (2014), it  
 also suggests additional complications to the traditional view of these



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.](#)

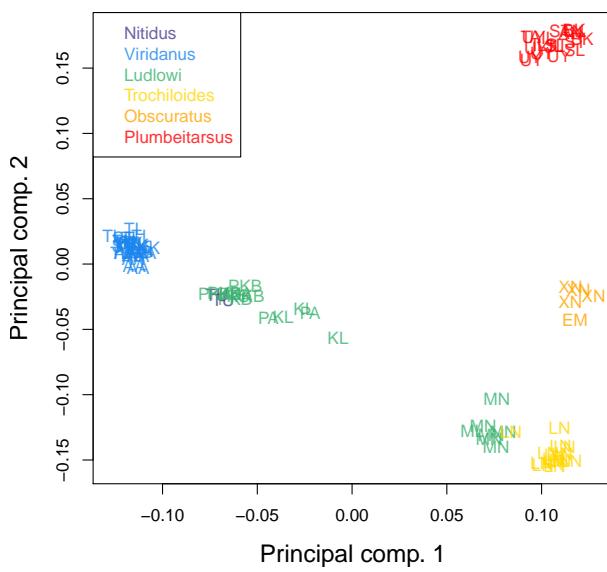


Figure 2.28: The 95 greenish warbler samples plotted on their locations on the first two principal components. The labels of individuals are the population labels from Figure 2.25, and coloured by subspecies label as in that figure. [Code here.](#)

884 warblers as an unbroken ring species, a case of speciation by continuous  
885 geographic isolation. The *Ludlowi* subspecies shows a significant  
886 genetic break, with the southern most MN samples clustering with the  
887 *Trochiloides* subspecies, in both the PCA and kinship matrix (Figures  
888 2.28 and 2.27), despite being much more geographically close to the  
889 other *Ludlowi* samples. This suggests that genetic isolation is not just  
890 a result of geographic distance, and other biogeographic barriers must  
891 be considered in the case of this broken ring species.

892 Finally, while PCA is a wonderful tool for visualizing genetic data,  
893 care must be taken in its interpretation. The U-like shape in the case  
894 of the Greenish warbler PC might be consistent with some low level  
895 of gene flow between the red and the blue populations, pulling them  
896 genetically closer together and helping to form a genetic ring as well  
897 as a geographic ring. However, U-like shapes are expected to appear in  
898 PCAs even if our populations are just arrayed along a line, and more  
899 complex geometric arrangements of populations in PC space can result  
900 under simple geographic models (NOVEMBRE and STEPHENS,  
901 2008). Inferring the geographical and population-genetic history of  
902 species requires the application of a range of tools; see ALCAIDE  
903 *et al.* (2014) and BRADBURD *et al.* (2016) for more discussion of the  
904 Greenish warblers.

906      *2.3.5 Correlations between loci, linkage disequilibrium, and recombination*

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

912    *Recombination* To understand correlations between loci we need  
 to understand recombination a bit more carefully. Let us consider  
 914    a heterozygous individual, containing  $AB$  and  $ab$  haplotypes. If no  
 recombination occurs between our two loci in this individual, then  
 916    these two haplotypes will be transmitted intact to the next genera-  
 tion. While if a recombination (i.e. an odd number of crossing over  
 918    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  
 920    an  $aB$  haplotype. Effectively, recombination breaks up the association  
 between loci. We'll define the recombination fraction ( $r$ ) to be the  
 922    probability of an odd number of crossing over events between our loci  
 in a single meiosis. In practice we'll often be interested in relatively  
 924    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-  
 926    nation rate (in Morgans) per base pair (typically  $\sim 10^{-8}$ ) and  $L$  is the  
 number of base pairs separating our two loci.

928    *Linkage disequilibrium* The (horrible) phrase linkage disequilibrium  
 (LD) refers to the statistical non-independence (i.e. a correlation)  
 930    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 diverse  
 932    topics, from sexual selection and speciation to the limits of genome-  
 wide association studies.

934    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 two  
 936    locus haplotype  $AB$  is  $p_{AB}$ , and likewise for our other three combi-  
 nations. If our loci were statistically independent then  $p_{AB} = p_A p_B$ ,  
 938    otherwise  $p_{AB} \neq p_A p_B$ . We can define a covariance between the  $A$  and  
 $B$  alleles at our two loci as

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

940    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)  
 942    are 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,

- 944 we can think of  $D$  as measuring the *excess* of coupling to repulsion  
 gametes. These  $D$  statistics are all closely related to each other as  
 946  $D_{AB} = -D_{Ab}$  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  
 948 handy result is 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  
 950  $D > 0$  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 0  
 952 before making this choice). You should be careful to keep the concepts  
 954 of linkage and linkage disequilibrium separate in your mind. Genetic  
 linkage refers to the linkage of multiple loci due to the fact that they  
 are transmitted through meiosis together (most often because the loci  
 956 are on the same chromosome). Linkage disequilibrium merely refers to  
 the covariance between the alleles at different loci; this may in part be  
 958 due to the genetic linkage of these loci but does not necessarily imply  
 this (e.g. genetically unlinked loci can be in LD due to population  
 960 structure).

**Question 12.** You genotype 2 bi-allelic loci (A & B) segregating  
 962 in two mouse subspecies (1 & 2) which mate randomly among them-  
 selves, but have not historically interbreed since they speciated. On  
 964 the basis of previous work you estimate that the two loci are separated  
 by a recombination fraction of 0.1. The frequencies of haplotypes in  
 966 each population are:

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

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

974 Our linkage disequilibrium statistic  $D$  depends strongly on the al-  
 lele frequencies of the two loci involved. One common way to partially  
 976 remove this dependence, and make it more comparable across loci, is  
 to divide  $D$  through by its the maximum possible value given the fre-  
 978 quency of the loci. This normalized statistic is called  $D'$  and varies be-  
 tween +1 and -1. In Figure 2.29 there's an example of LD across the  
 980 TAP2 region in human and chimp. Notice how physically close SNPs,  
 i.e. those close to the diagonal, have higher absolute values of  $D'$  as



Figure 2.29: LD across the TAP2 gene region in a sample of Humans and Chimps, from PTAK *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.

982 closely linked alleles are separated by recombination less often allowing  
983 high levels of LD to accumulate. Over large physical distances, away  
from the diagonal, there is lower  $D'$ . This is especially notable in  
985 humans as there is an intense, human-specific recombination hotspot in  
986 this region, which is breaking down LD between opposite sides of this  
region.

988 Another common statistic for summarizing LD is  $r^2$  which we write  
as

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

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

994 Figure 2.31 shows  $r^2$  for pairs of SNPs at various physical distances  
in two population samples of *Mus musculus domesticus*. Again LD  
is highest between physically close markers as LD is being generated  
996 faster than it can decay via recombination; more distant markers have  
much lower LD as here recombination is winning out. Note the decay  
998 of LD is much slower in the advanced-generation cross population than  
in the natural wild-caught population. This persistence of LD across  
1000 megabases is due to the limited number of generations for recombination  
since the cross was created.

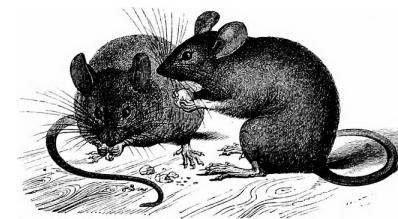
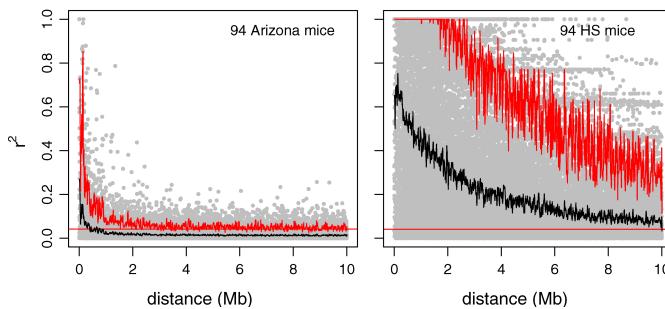


Figure 2.30: *Mus musculus*.  
A history of British quadrupeds, including the Cetacea. 1874. Bell T., Tomes, R. F. m Alston E. R. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. No known copyright restrictions.

Figure 2.31: The decay of LD for autosomal SNP 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  $\sim 3000$  SNPs. The solid black line gives the mean, the jagged the 95<sup>th</sup> percentile, and the flat red line a cutoff for significant LD. From LAURIE *et al.* (2007), licensed under CC BY 4.0.

1002 *The generation of LD.* Various population genetic forces can generate LD. Selection can generate LD by favouring particular combinations  
 1004 of alleles. Genetic drift will also generate LD, not because particular combinations of alleles are favoured, but simply because at random  
 1006 particular haplotypes can by chance drift up in frequency. Mixing between divergent populations can also generate LD, as we saw in the  
 1008 mouse question above.

1010 *The decay of LD due to recombination* We will now examine what happens to LD over the generations if we only allow recombination to occur in a very large population (i.e. no genetic drift, i.e. the frequencies of our loci follow their expectations). To do so, consider the frequency of our  $AB$  haplotype in the next generation,  $p'_{AB}$ . We lose  
 1012 a fraction  $r$  of our  $AB$  haplotypes to recombination ripping our alleles apart but gain a fraction  $rp_{APB}$  per generation from other haplotypes  
 1014 recombining together to form  $AB$  haplotypes. Thus in the next generation  
 1016

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

1018 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  
 1020 genotypes that could generate a  $p_{AB}$  haplotype.

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

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

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

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

where we can cancel out  $\Delta p_A$  and  $\Delta p_B$  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

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

1026 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  
 1028 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.

- Question 13.** You find a hybrid population between the two mouse subspecies described in the question above, which appears to be comprised of equal proportions of ancestry from the two subspecies.
- 1030
- You estimate LD between the two markers to be 0.0723. Assuming that this hybrid population is large and was formed by a single mixture event, can you estimate how long ago this population formed?
- 1032
- 1034

A particularly striking example of the decay of LD generated by the mixing of populations is offered by the LD created by the interbreeding between humans and Neanderthals. Neanderthals and modern Humans diverged from each other likely over half a million years ago, allowing time for allele frequency differences to accumulate 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 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 this interbreeding, SANKARARAMAN *et al.* (2012) looked at the LD in modern humans between pairs of alleles found to be derived from the Neanderthal genome (and nearly absent from African populations). In Figure 2.35 we show the average LD between these loci as a function of the genetic distance ( $r$ ) between them, from the work of SANKARARAMAN *et al.*

1036

1040

1042

1044

1046

1048

1050

1052

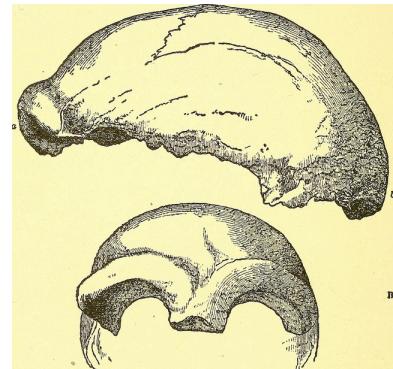
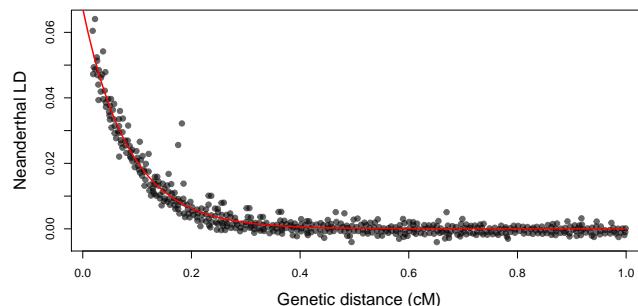


Figure 2.34: The earliest discovered fossil of a Neanderthal, fragments of skull found in a cave in the Neander Valley in Germany.  
Man's place in nature. 1890. Huxley, T. H.  
Image from the Internet Archive. Contributed by The Library of Congress. No known copyright restrictions.

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 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).

- Assuming a recombination rate  $r$ , we can fit the exponential decay 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, or about 35 thousand years (using a human generation time of 29 years). Thus the LD in modern Eurasians, between alleles derived from the interbreeding with Neanderthals, represents over thirty thousand years of recombination slowly breaking down these old associations.
- 1054
- 1056
- 1058

The calculation done by SANKARARAMAN *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.

## *Genetic Drift and Neutral Diversity*

1062 RANDOMNESS IS INHERENT TO EVOLUTION, from the lucky  
 birds blown of course to colonize some new oceanic island, to which  
 1064 mutations arise first in the HIV strain infecting an individual taking  
 anti-retroviral drugs. One major source of stochasticity in evolution-  
 1066 ary biology is genetic drift. Genetic drift occurs because more or less  
 copies of an allele by chance can be transmitted to the next genera-  
 1068 tion. This can occur because, by chance, the individuals carrying a  
 particular allele can leave more or less offspring in the next generation.  
 1070 In a sexual population, genetic drift also occurs because Mendelian  
 transmission means that only one of the two alleles in an individual,  
 1072 chosen at random at a locus, is transmitted to the offspring.

Genetic drift can play a role in the dynamics of all alleles in all  
 1074 populations, but it will play the biggest role for neutral alleles. A  
 neutral polymorphism occurs when the segregating alleles at a poly-  
 1076 morphic site have no discernible differences in their effect on fitness.  
 We'll make clear what we mean by "discernible" later, but for the  
 1078 moment think of this as "no effect" on fitness.

1080 *The neutral theory of molecular evolution.* The role of genetic drift  
 in molecular evolution has been hotly debated since the 60s when  
 he Neutral theory of molecular evolution was proposed (see OHTA  
 1082 and GILLESPIE, 1996, for a history).<sup>1</sup> The central premise of Neu-  
 tral theory theory is that patterns of molecular polymorphism within  
 1084 species and substitution between species can be well understood by  
 supposing that the vast majority of these molecular polymorphisms  
 1086 and substitutions were neutral alleles, whose dynamics were just sub-  
 ject to the vagaries of genetic drift and mutation. Early proponents of  
 1088 this view suggested that the vast majority of new mutations are either  
 neutral or highly deleterious (e.g. mutations that disrupt important  
 1090 protein functions). This latter class of mutations are too deleterious  
 to contribute much to common polymorphisms or substitutions be-

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

<sup>1092</sup> tween species, because they are quickly weeded out of the population by selection.

<sup>1094</sup> Neutral theory can sound strange given that much of the time our first brush with evolution often focuses of adaptation and phenotypic evolution. However, proponents of this world-view didn't deny the existence of advantageous mutations, they simply thought that beneficial mutations are rare enough that their contribution to the bulk of polymorphism or divergence can be largely ignored. They also often thought that much of phenotypic evolution may well be adaptive, but again the loci responsible for these phenotypes are a small fraction of all the molecular change that occur. The original neutral theory of molecular evolution was original proposed to explain protein polymorphism. However, we can apply it more broadly to think about neutral evolution genome-wide. With that in mind, what types of molecular changes could be neutral? Perhaps:

- <sup>1108</sup> 1. Changes in non-coding DNA that don't disrupt regulatory sequences. For example, in the human genome only about 2% of the genome codes for proteins. The rest is mostly made up of old transposable element and retrovirus insertions, repeats, pseudo-genes, and general genomic clutter. Current estimates suggesting that, even counting conserved, functional, non-coding regions that < 10% of our genome is subject to evolutionary constraint (RANDS *et al.*, 2014).
  - <sup>1116</sup> 2. Synonymous changes in coding regions, i.e. those that don't change the amino-acid encoded by a codon.
  - <sup>1118</sup> 3. Non-synonymous changes that don't have a strong effect on the functional properties of the amino acid encoded, e.g. changes that don't change the size, charge, or hydrophobic properties of the amino acid too much.
  - <sup>1122</sup> 4. An amino-acid change with phenotypic consequences, but little relevance to fitness, e.g. a mutation that causes your ears to be a slightly different shape, or that prevents an organism from living past 50 in a species where most individuals reproduce and die by their 20s.
- <sup>1126</sup> There are counter examples to all of these ideas, e.g. synonymous changes can affect the translation speed and accuracy of proteins and so are subject to selection. However, the list above hopefully convinces you that the general thinking that some portion of molecular change may not be subject to selection isn't as daft as it may have initially sounded.
- <sup>1132</sup> Various features of molecular polymorphism and divergence have been viewed as consistent with the neutral theory of molecular evo-

lution. The two we'll focus on in this chapter are the high level of molecular polymorphism in many species, see for example Figure 2.2, and the molecular clock. We'll see that various aspects of the original neutral theory have merit in describing some features and types of molecular change, but we'll also see that it is demonstrably wrong in some cases. We'll also see the primary utility of the neutral theory isn't whether it is right or wrong, but that it serves as a simple null model that can be tested and in some cases rejected, and subsequently built on. The broader debate currently in the field of molecular evolution is the balance of neutral, adaptive, and deleterious changes that drive different types of evolutionary change.

### 3.1 Loss of heterozygosity due to drift.

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

Imagine a randomly mating population of a constant size  $N$  diploid individuals, and that we are examining a locus segregating for two alleles that are neutral with respect to each other. This population 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 alleles within a small population.

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

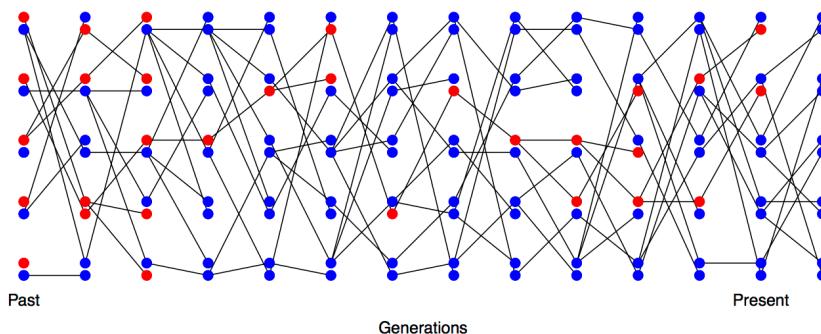


Figure 3.1: Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals over the generations, with lines showing transmission. In the first generation every individual is a heterozygote. Code here.

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



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.

<sup>1164</sup> alleles in generation  $t$ , are non-identical is  $H_t$ .

Conversely, if the two alleles in our pair had the same parental

<sup>1166</sup> allele in the proceeding generation (i.e. the alleles are identical by descent one generation back) then these two alleles must be identical  
<sup>1168</sup> (as we are not allowing for any mutation).

In a diploid population of size  $N$  individuals there are  $2N$  alleles.

<sup>1170</sup> The probability that our two alleles have the same parental allele in the proceeding generation is  $1/(2N)$  and the probability that they have  
<sup>1172</sup> different parental alleles is  $1 - 1/(2N)$ . So by the above argument, the expected heterozygosity in generation  $t + 1$  is

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

<sup>1174</sup> Thus, if the heterozygosity in generation 0 is  $H_0$ , our expected heterozygosity in generation  $t$  is

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

<sup>1176</sup> i.e. the expected heterozygosity within our population is decaying geometrically with each passing generation. If we assume that  $1/(2N) \ll 1$   
<sup>1178</sup> then we can approximate this geometric decay by an exponential decay (see Question 2 below), such that

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

<sup>1180</sup> i.e. heterozygosity decays exponentially at a rate  $1/(2N)$ .

In Figure 3.3 we show trajectories through time for 40 independently simulated loci drifting in a population of 50 individuals. Each 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.  
<sup>1184</sup> 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).  
<sup>1186</sup>



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.

**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?

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 offsprings' 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 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 popu-

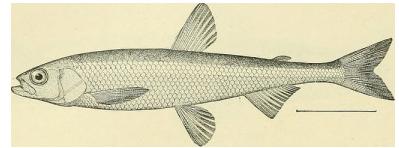


Figure 3.4: Pond smelt (*Hypomesus olidus*), a close relative of delta smelt. Bulletin of the United States Fish Commission, 1906. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

lation frequency  $p_t$  in the previous generation. If you've read to here,  
 1218 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  
 1220 JBS Haldane and to read more about his life; he's a true character and  
 one of the last great polymaths. )

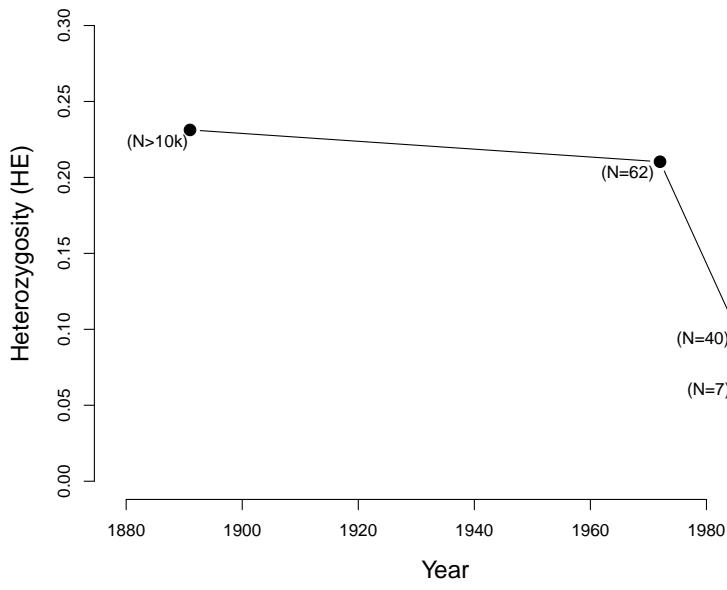


Figure 3.6: Loss of heterozygosity in the Black-footed Ferrets in their declining population. Numbers in brackets give estimated number of individuals alive at that time. Data from WISELY *et al.* (2002). Code here.



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.

1222 To see how a decline in population size can affect levels of heterozygosity, let's consider the case of black-footed ferrets (*Mustela*  
 1224 *nigripes*). The black-footed ferret population has declined dramatically through the twentieth century due to destruction of their habitat. In  
 1226 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  
 1228 rediscovered (40 individuals), but in 1985 this population suffered a number of disease outbreaks. All of the 18 remaining wild individuals  
 1230 were brought into captivity, 7 of which reproduced. Thanks to intense captive breeding efforts and conservation work, a wild population of  
 1232 over 300 individuals has been established since. However, because all of these individuals are descended from those 7 individuals who  
 1234 survived the bottleneck, diversity levels remain low. WISELY *et al.* measured heterozygosity at a number of microsatellites in individuals  
 1236 from museum collections, showing the sharp drop in diversity as population sizes crashed (see Figure 3.6).

1238 **Question 2.** In mathematical population genetics, a commonly used approximation is  $(1 - x) \approx e^{-x}$  for  $x \ll 1$  (formally, this

1240 follows from the Taylor series expansion of  $\exp(-x)$ , ignoring second  
 1242 order and higher terms of  $x$ ). This approximation is especially useful  
 1244 for approximating a geometric decay process by an exponential decay  
 1246 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  $t = 50$ . I.e. calculate both expressions for these values, hand  
 in your answers and briefly comment on your results.

1248 *3.1.1 Levels of diversity maintained by a balance between mutation  
 and drift*

1250 Next we're going to consider the amount of neutral polymorphism that  
 can be maintained in a population as a balance between genetic drift  
 1252 removing variation and mutation introducing new neutral variation,  
 see Figure 3.7 for an example. Note in our example, how no-one allele  
 1254 is maintained at a stable equilibrium, rather an equilibrium level of  
 polymorphism is maintained by a constantly shifting case of alleles.

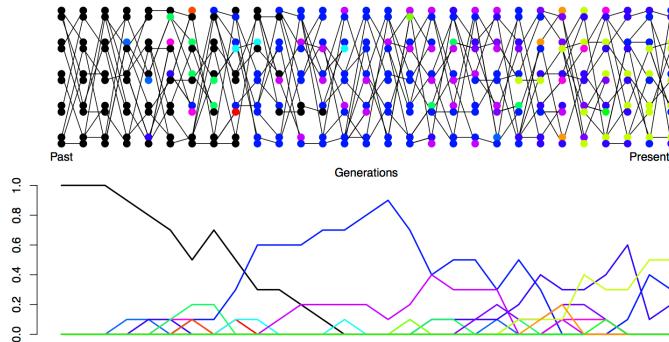


Figure 3.7: Mutation-drift balance. A diploid population of 5 individuals. In the first generation everyone has the same allele (black). Each generation the transmitted allele can mutate and we generate a new colour. In the bottom plot, I trace the frequency of alleles in our population over time. The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

1256 *The neutral mutation rate.* We'll first want to consider the rate at  
 which neutral mutations arise in the population. Thinking back to our  
 1258 discussion of the neutral theory of molecular evolution, let's suppose  
 that there are only two classes of mutation that can arise in our ge-  
 1260 nomic region of interest: neutral mutations and highly deleterious  
 1262 mutations. The total mutation rate at our locus is  $\mu_T$  per genera-  
 tion, i.e. per transmission from parent to child. A fraction  $C$  of our  
 1264 mutations are new alleles that are highly deleterious and so quickly  
 removed from the population. We'll call this  $C$  parameter the con-  
 straint, and it will differ according to the genomic region we consider.  
 1266 The remaining fraction  $(1 - C)$  are our neutral mutations, such that  
 our neutral mutation rate is

$$\mu = (1 - C)\mu_T \quad (3.4)$$

<sub>1268</sub> This is the per generation rate.

**Question 3.** It's worth taking a minute to get familiar with both  
<sub>1270</sub> how rare, and how common, mutation is. The per base pair mutation  
<sub>1272</sub> rate in humans is around  $1.5 \times 10^{-8}$  per generation. That means, on  
<sub>1274</sub> average, we have to monitor a site for  $\sim 66.6$  million transmissions  
<sub>1276</sub> from parent to child to see a mutation. Yet populations and genomes  
<sub>1278</sub> are big places, so mutations are common at these levels.

**A)** Your autosomal genome is  $\sim 3$  billion base pairs long ( $3 \times 10^9$ ).

<sub>1276</sub> You have two copies, the one you received from your mum and one  
<sub>1278</sub> from your dad. What is the average (i.e. the expected) number of  
<sub>1280</sub> mutations that occurred in the transmission from your mum and your  
<sub>1282</sub> dad to you?

**B)** The current human population size is  $\sim 7$  billion individuals.

<sub>1280</sub> How many times, at the level of the entire human population, is a  
<sub>1282</sub> single base-pair mutated in the transmission from one generation to  
<sub>1284</sub> the next?

<sub>1284</sub> *Levels of heterozygosity maintained as a balance between mutation  
<sub>1286</sub> and selection.* Looking backwards in time from one generation to  
<sub>1288</sub> the previous generation, we are going to say that two alleles which  
<sub>1290</sub> have the same parental allele (i.e. find their common ancestor) in  
<sub>1292</sub> the preceding generation have *coalesced*, and refer to this event as a  
<sub>1294</sub> *coalescent event*.

<sub>1290</sub> The probability that our pair of randomly sampled alleles have  
<sub>1292</sub> coalesced in the preceding generation is  $1/(2N)$ , the probability that our  
<sub>1294</sub> pair of alleles fail to coalesce is  $1 - 1/(2N)$ .

<sub>1294</sub> The probability that a mutation changes the identity of the trans-  
<sub>1296</sub> mitted allele is  $\mu$  per generation. So the probability of no mutation  
<sub>1298</sub> occurring is  $(1 - \mu)$ . We'll assume that when a mutation occurs it cre-  
<sub>1300</sub> ates some new allelic type which is not present in the population. This  
<sub>1302</sub> assumption (commonly called the infinitely-many-alleles model) makes  
<sub>1304</sub> the math slightly cleaner, and also is not too bad an assumption bi-  
<sub>1306</sub> logically. See Figure 3.7 for a depiction of mutation-drift balance in  
<sub>1308</sub> this model over the generations.

<sub>1302</sub> This model lets us calculate when our two alleles last shared a  
<sub>1304</sub> common ancestor and whether these alleles are identical as a result of  
<sub>1306</sub> failing to mutate since this shared ancestor. For example, we can work  
<sub>1308</sub> out the probability that our two randomly sampled alleles coalesce 2  
<sub>1310</sub> generations in the past (i.e. they fail to coalesce in generation 1 and  
<sub>1312</sub> then coalesce in generation 2), and that they are identical as

$$\left(1 - \frac{1}{2N}\right) \frac{1}{2N} (1 - \mu)^4 \quad (3.5)$$

Note the power of 4 is because our two alleles have to have failed to

<sup>1308</sup> 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

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

<sup>1312</sup> To make this slightly easier on ourselves let's further assume that  $t \approx t + 1$  and so rewrite this as:

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

<sup>1314</sup> This gives us the approximate probability that two alleles will coalesce in the  $(t + 1)^{\text{th}}$  generation. In general, we may not know <sup>1316</sup> when two alleles may coalesce: they could coalesce in generation  $t = 1, t = 2, \dots$ , and so on. Thus, to calculate the probability that <sup>1318</sup> two alleles coalesce in *any* generation before mutating, we can write:

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

which follows from basic probability and the fact that coalescing in a <sup>1320</sup> particular generation is mutually exclusive with coalescing in a different generation.

While we could calculate a value for this sum given  $N$  and  $\mu$ , 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:

$$P(\text{coal. in } t+1 \& \text{ no mutations}) \approx \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t (1 - \mu)^{2t} \quad (3.8)$$

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

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

<sup>1322</sup> Then we can approximate the summation by an integral, giving us:

$$\frac{1}{2N} \int_0^{\infty} e^{-t(2\mu+1/(2N))} dt = \frac{1/(2N)}{1/(2N) + 2\mu} = \frac{1}{1 + 4N\mu} \quad (3.11)$$

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

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

$$H = \frac{4N\mu}{1 + 4N\mu} \quad (3.12)$$

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

$$\theta = 4N\mu \quad (3.13)$$

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

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

### 3.1.2 The effective population size

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



This result was derived by KIMURA and CROW (1964) and MALÉCOT (1948) (see MALÉCOT, 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 EWENS (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 .

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.

1348 To cope with this discrepancy, population geneticists often invoke  
 the concept of an *effective population size* ( $N_e$ ). In many situations  
 1350 (but not all), departures from model assumptions can be captured by  
 substituting  $N_e$  for  $N$ .

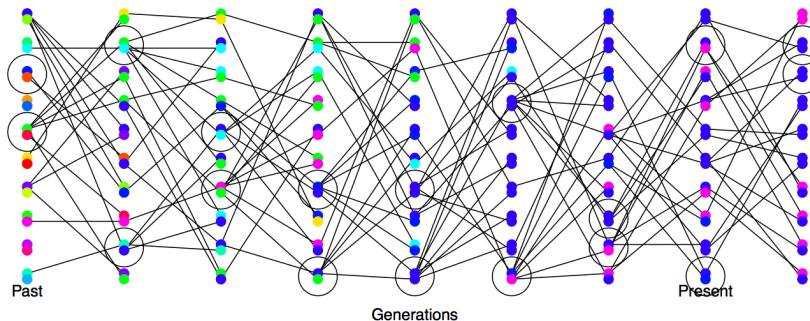
1352 If population sizes vary rapidly in size, we can (if certain conditions  
 are met) replace our population size by the harmonic mean population  
 1354 size. Consider a diploid population of variable size, whose size is  $N_t$   $t$   
 generations into the past. The probability our pairs of alleles have not  
 1356 coalesced by generation  $t$  is given by

$$\prod_{i=1}^t \left(1 - \frac{1}{2N_i}\right) \quad (3.14)$$

1358 Note that this simply collapses to our original expression  $(1 - \frac{1}{2N})^t$  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

$$N_e = \frac{1}{\frac{1}{t} \sum_{i=1}^t \frac{1}{N_i}}. \quad (3.15)$$

1360 This is the harmonic mean of the varying population size.<sup>2</sup>  
 Thus our effective population size, the size of an idealized constant  
 1362 population which matches the rate of genetic drift, is the harmonic  
 mean true population size over time. The harmonic mean is very  
 1364 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  
 1366 generations,  $N_e$  will be much closer to 1000 than a million.



<sup>2</sup> To see this, note that if  $1/(N_i)$  is small, then we can approximate (3.14) using the exponential approximation:

$$\prod_{i=1}^t \exp\left(-\frac{1}{2N_i}\right) = \exp\left(-\sum_{i=1}^t \frac{1}{2N_i}\right). \quad (3.16)$$

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

$$\frac{t}{2N_e} = \sum_{i=1}^t 1/2N_i \quad (3.17)$$

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 (3.15).

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.

1368 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 the  
 1370 rate of drift will reflect this much smaller number of reproducing individuals. See Figure 3.9 for a depiction of the higher rate of drift in a  
 1372 population where there is high variance in reproductive success.

To see one example of this, consider the case where  $N_F$  of females  
 1374 get to reproduce and  $N_M$  males get reproduce. While every individual

has a mother an a father, not every individual gets to be a parent. In  
 1376 practice, in many animal species far more females get to reproduce  
 than males, i.e.  $N_M < N_F$ , as a few males get many mating oppor-  
 1378 tunities and many males get no/few mating opportunities (see JAN-  
 ICKE *et al.*, 2016, for a broad analysis, and note that there are certainly  
 1380 many exceptions to this general pattern). When our two alleles pick  
 an ancestor, 25% of the time our alleles were both in a female ances-  
 1382 tor, in which case they are IBD with probability  $1/(2N_F)$ , and 25% of  
 the time they are both in a male ancestor, in which case they coalesce  
 1384 with probability  $1/(2N_M)$ . The remaining 50% of the time, our alleles  
 trace back to two individuals of different sexes in the prior generation  
 1386 and so cannot coalesce. Therefore, our probability of coalescence in  
 the preceding generation is

$$\frac{1}{4} \left( \frac{1}{2N_M} \right) + \frac{1}{4} \left( \frac{1}{2N_F} \right) \quad (3.18)$$

1388 i.e. the rate of coalescence is the harmonic mean of the two sexes'  
 population sizes, equating this to  $\frac{1}{2N_e}$  we find

$$N_e = \frac{4N_F N_M}{N_F + N_M} \quad (3.19)$$

1390 Thus if reproductive success is very skewed in one sex (e.g.  $N_M \ll N/2$ ), our effective population size will be much reduced as a re-  
 1392 sult. For more on how different evolutionary forces affect the rate  
 of genetic drift, and their impact on the effective population size, see  
 1394 CHARLESWORTH (2009).

**Question 5.** You are studying a population of 500 males and 500  
 1396 females Hamadryas baboons. Assume that all of the females but only  
 1/10 of the males get to mate: **A)** What is the effective population  
 1398 size for the autosome?

**B)** Do you expect the *ratio* of X-chromosome to autosomal diversity  
 1400 to be higher or lower in this species compared to a species where the  
 sexes have more similar variance in reproductive success? Explain the  
 1402 intuition behind your answer.

### 3.2 The Coalescent and patterns of neutral diversity

1404 "Life can only be understood backwards; but it must be lived for-  
 wards." – Kierkegaard

1406 *Pairwise Coalescent time distribution and the number of pairwise  
 differences.* Thinking back to our calculations we made about the  
 1408 loss of neutral heterozygosity and equilibrium levels of diversity (in  
 Sections 3.1 and 3.1.1), you'll note that we could first specify which



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

<sup>1410</sup> generation a pair of sequences coalesce in, and then calculate some properties of heterozygosity based on that. That's because neutral  
<sup>1412</sup> mutations do not affect the probability that an individual transmits an allele, and so don't affect the way in which we can trace ancestral  
<sup>1414</sup> lineages back through the generations.

<sup>1416</sup> 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  
<sup>1418</sup> example of this.

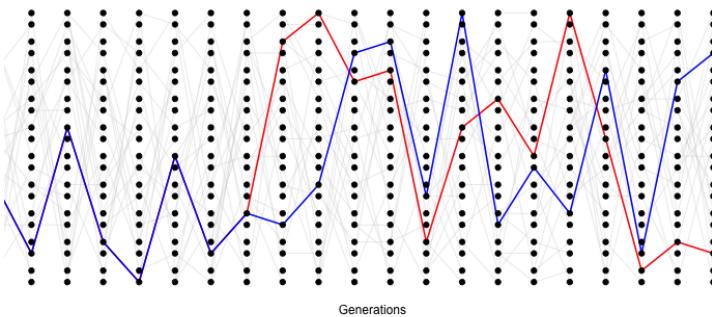


Figure 3.11: A simple simulation of the coalescent process. The simulation consists of a diploid population of 10 individuals (20 alleles). In each generation, each individual is equally likely to be the parent of an offspring (and the allele transmitted is indicated by a light grey line). We track a 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  $t$   
<sup>1420</sup> generations and then coalesce in the  $t + 1$  generation back is

$$P(T_2 = t + 1) = \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t \quad (3.20)$$

Thus the coalescent time of our pair of alleles is a Geometrically distributed random variable, where the probability of success is  $1/(2N)$ ;  
<sup>1422</sup> we denote this by  $T_2 \sim \text{Geo}(1/(2N))$ . The expected (i.e. the mean  
<sup>1424</sup> over many replicates) coalescent time of a pair of alleles is then

$$\mathbb{E}(T_2) = 2N \quad (3.21)$$

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  $\mu$ , then the expected number of mutations between a pair of alleles coalescing  $t$  generations ago is  $2t\mu$  (the alleles have gone through a total of  $2t$  meioses since they last shared a common ancestor). So we can write the expected number of mutations

Blurring our eyes a little, we can see that 3.20 is

$$\approx \frac{1}{2N} e^{-t/(2N)} \quad (3.22)$$

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.

$(S_2)$  separating two alleles drawn at random from the population as

$$\begin{aligned}\mathbb{E}(S_2) &= \sum_{t=0}^{\infty} \mathbb{E}(S_2|T_2 = t)P(T_2 = t) \\ &= \sum_{t=0}^{\infty} 2\mu t P(T_2 = t) \\ &= 2\mu \mathbb{E}(T_2) \\ &= 4\mu N\end{aligned}\tag{3.23}$$

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 <sup>3</sup> 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  $\pi$  (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

$$\mathbb{E}(\pi) = 4N\mu = \theta\tag{3.24}$$

So we can get an empirical estimate of  $\theta$  from  $\pi$ , 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  $\mu$ , 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's a boat load of 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.<sup>4</sup>

Lets take a moment to distinguish our expected heterozygosity (eqn. 3.12) from our expected number of pairwise differences ( $\pi$ ). 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  $\pi$  measure keeps track of the average total number of differences between our loci. As such,  $\pi$  is often a more useful measure, as it records the number of differences between

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

<sup>4</sup> 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  $\pi$  calculated for on 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  $\pi$ . But even then caution is warranted.

the sequences, not just whether they are different from each other  
 1460 (however, for certain types of loci, e.g. microsatellites, heterozygosity  
 is often used as we cannot usually count up the minimum number of  
 1462 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  
 1464  $H \approx \theta$  for small values of  $\theta$ . For example, comparing two sequences  
 at random in humans,  $\pi \approx 1/1000$  per basepair, and the probability  
 1466 that a specific base pair differs between two sequences is  $\approx 1/1000$ .  
 However, these two quantities start to differ from each other when  
 1468 we consider regions with higher mutation rates. For example, if we  
 consider a 10kb region, our mutation rate will 10,000 times larger than  
 1470 a single base pair. For this length of sequence the probability that two  
 randomly chosen haplotypes differ is quite different from the number  
 1472 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  $\mathbb{E}[\pi]$   
 1474 and  $H$  to see this.)

**Question 6.** ROBINSON *et al.* (2016) found that the endangered

1476 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  
 1478 California mainland gray fox ( $0.0012\text{bp}^{-1}$ ).

**A)** Assuming a mutation rate of  $2 \times 10^{-8}$  per bp, what effective

1480 population sizes do you estimate for these two populations?

**B)** Why is the effective population size of the Channel Island fox

1482 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.

1486

*More details on the pairwise coalescent and the randomness of mutation.* We've derived the expected number of differences between a  
 1488 pair of sequences and talked about how variable the coalescent time  
 is for a pair of sequences. The mutation process is also very variable;  
 1490 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  
 1492 that time.

1494 Conditional on the coalescent time  $t$ , the probability that our pair  
 of alleles are separated by  $S_2$  mutations since they last shared a com-  
 1496 mon ancestor is

$$P(S_2|T_2 = t) = \binom{2t}{j} \mu^j (1 - \mu)^{2t-j} \quad (3.25)$$

i.e. mutations happen in  $j$  generations and do not happen in  $2t - j$   
 1498 generations (with  $\binom{2t}{j}$  ways this combination of events can possibly



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.

happen). Assuming that  $\mu \ll 1$  and that  $2t - j \approx 2t$ , then we can  
 1500 approximate the probability that we have  $S_2$  mutations as a Poisson  
 distribution:

$$P(S_2|T_2 = t) = \frac{(2\mu t)^j e^{-2\mu t}}{j!} \quad (3.26)$$

1502 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  
 1504 mutation.

### 3.3 The coalescent process of a sample of alleles.

1506 Usually we are not just interested in pairs of alleles, or the average  
 pairwise diversity. Generally we are interested in the properties of di-  
 1508 versity in samples of a number of alleles drawn from the population.  
 Instead of just following a pair of lineages back until they coalesce, we  
 1510 can follow the history of a sample of alleles back through the popula-  
 tion.

1512 Consider first sampling three alleles at random from the population.  
 The probability that all three alleles choose exactly the same ancestral  
 1514 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  
 1516 alleles coalesce all at once, and in a moment we'll see that it is safe to  
 ignore such unlikely events.

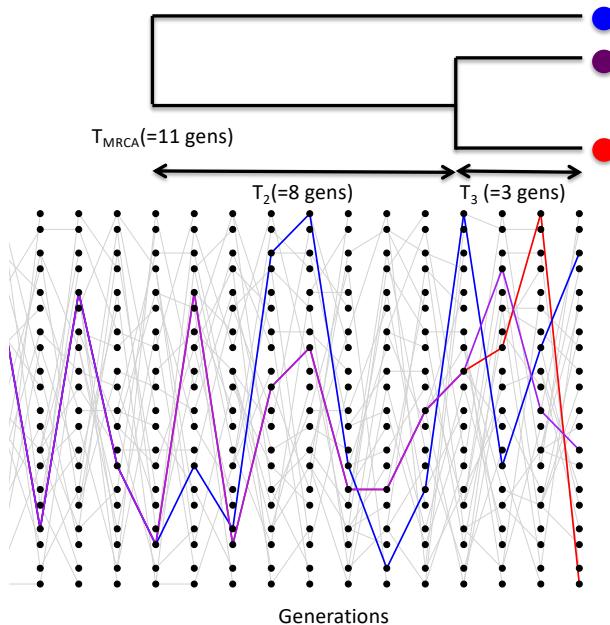


Figure 3.13: A simple simulation of the coalescent process for three lineages. We track the ancestry of three modern-day alleles, the first pair (blue and purple) coalesce four generations back, after which there are only two independent lineages we are tracking. This pair then coalesces twelve generations 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.

1518 The probability that a specific pair of alleles find a common ances-  
 tor in the preceding generation is still  $1/(2N)$ . There are three possible

<sup>1520</sup> pairs of alleles, so the probability that no pair finds a common ancestor in the preceding generation is

$$\left(1 - \frac{1}{2N}\right)^3 \approx \left(1 - \frac{3}{2N}\right) \quad (3.27)$$

<sup>1522</sup> 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 <sup>1524</sup> random realization of this process.

More generally, when we sample  $i$  alleles there are  $\binom{i}{2}$  pairs,<sup>5</sup> i.e.

<sup>1526</sup>  $i(i - 1)/2$  pairs. Thus, the probability that no pair of alleles in a sample of size  $i$  coalesces in the preceding generation is

$$\left(1 - \frac{1}{(2N)}\right)^{\binom{i}{2}} \approx \left(1 - \frac{\binom{i}{2}}{2N}\right) \quad (3.28)$$

<sup>1528</sup> 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. <sup>1530</sup> tripletons) simultaneously coalesce at once as terms of  $1/N^2$  and higher can be ignored as they are vanishingly rare. Obviously in reasonable <sup>1532</sup> 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 <sup>1534</sup> ignore these terms.

When there are  $i$  alleles, the probability that we wait until the  $t + 1$  <sup>1536</sup> generation before any pair of alleles coalesces is

$$P(T_i = t + 1) = \frac{\binom{i}{2}}{2N} \left(1 - \frac{\binom{i}{2}}{2N}\right)^t \quad (3.29)$$

Thus the waiting time to the first coalescent event while there are  $i$  <sup>1538</sup> lineages is a geometrically distributed random variable with probability of success  $\binom{i}{2}/2N$ , which we denote by

$$T_i \sim \text{Geo}\left(\frac{\binom{i}{2}}{2N}\right). \quad (3.30)$$

<sup>1540</sup> The mean waiting time till any of pair within our sample coalesces is

$$\mathbb{E}(T_i) = \frac{2N}{\binom{i}{2}} \quad (3.31)$$

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 <sup>1542</sup> 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 <sup>1544</sup> when a pair of these  $i - 1$  alleles coalesce, we then only have to follow  $i - 2$  alleles back. This process continues until we coalesce back <sup>1546</sup> to a sample of two, and from there to a single most recent common ancestor (MRCA).

<sup>5</sup> said as “i choose 2”

To see the continuous time version of this, note that (3.29) is

$$\approx \frac{\binom{i}{2}}{2N} \exp\left(-\frac{\binom{i}{2}}{2N} t\right) \quad (3.32)$$

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

1550 *Simulating a coalescent genealogy* To simulate a coalescent genealogy at a locus for a sample of  $n$  alleles we therefore simply follow the  
1552 following algorithm:

1. Set  $i = n$ .
- 1554 2. Simulate a random variable to be the time  $T_i$  to the next coalescent event from  $T_i \sim \text{Exp}((\frac{i}{2})/2N)$
- 1556 3. Choose a pair of alleles to coalesce at random from all possible pairs.
- 1558 4. Set  $i = i - 1$
- 1560 5. Continue looping steps 1-3 until  $i = 1$ , i.e. the most recent common ancestor of the sample is found.

1562 By following this algorithm we are generating realizations of the genealogy of our sample.

### 3.3.1 Expected properties of coalescent genealogies and mutations.

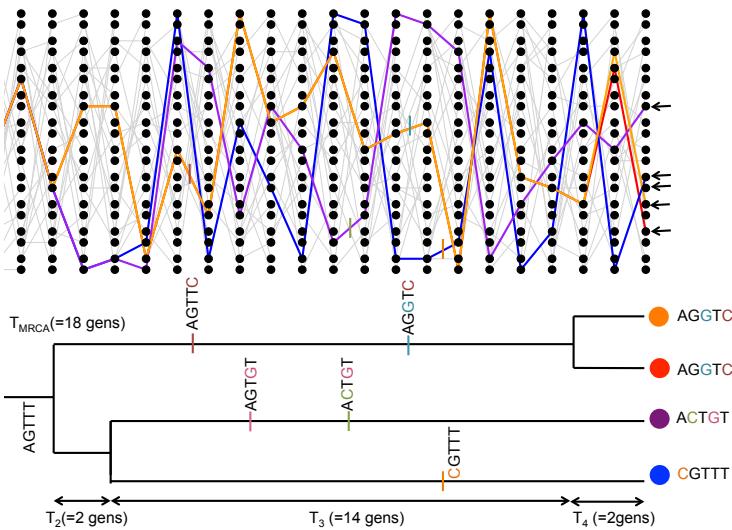


Figure 3.14: A simple coalescent tree from a single coalescent simulation, tracing the genealogy of 4 alleles with mutational changes marked with dashes showing transitions away from the MRCA sequence (AGTTT). The  $T_{MRCA}$  is  $T_4 + T_3 + T_2$ . The total time in the tree is  $T_{tot} = 4T_4 + 3T_3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

1564 *The expected time to the most recent common ancestor.* We will first consider the time to the most recent common ancestor of the entire  
1566 sample ( $T_{MRCA}$ ). This is

$$T_{MRCA} = \sum_{i=n}^2 T_i \quad (3.33)$$

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

$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^2 \mathbb{E}(T_i) = \sum_{i=n}^2 2N / \binom{i}{2} \quad (3.34)$$

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

$$\mathbb{E}(T_{MRCA}) = 4N \left( 1 - \frac{1}{n} \right) \quad (3.35)$$

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

$$T_{tot} = \sum_{i=n}^2 iT_i \quad (3.36)$$

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

$$\mathbb{E}(T_{tot}) = \sum_{i=n}^2 i \frac{2N}{\binom{i}{2}} = \sum_{i=n}^2 \frac{4N}{i-1} = \sum_{i=n-1}^1 \frac{4N}{i} \quad (3.37)$$

1598 we see that our expected total amount of time in the genealogy scales  
 linearly with our population size  $N$ . Our expected total amount of  
 1600 time is also increasing with sample size  $n$ , but is doing so very slowly.  
 This again follows from the fact that in large samples, the initial  
 1602 coalescence usually happens very rapidly, so that extra samples add  
 little to the total amount of time in the genealogical tree.

1604 We saw above that the number of mutational differences between  
 a pair of alleles that coalescence  $T_2$  generations ago was Poisson with  
 1606 a mean of  $2\mu T_2$ , where  $2T_2$  is the total branch length in this simple  
 2-sample genealogical tree. A mutation that occurs on any branch of  
 1608 our genealogy will cause a segregating polymorphism in the sample  
 (meeting our infinitely-many-sites assumption). Thus, if the total time  
 1610 in the genealogy is  $T_{tot}$ , there are  $T_{tot}$  generations for mutations. So  
 the total number of mutations segregating in our sample ( $S$ ) is Poisson  
 1612 with mean  $\mu T_{tot}$ . Thus the expected number of segregating sites in a  
 sample of size  $n$  is

$$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^1 \frac{4N\mu}{i} = \theta \sum_{i=n-1}^1 \frac{1}{i} \quad (3.38)$$

1614 Note that this is growing with the sample size  $n$ , albeit very slowly  
 (roughly at the rate of the log of the sample size). We can use this  
 1616 formula to derive another estimate of the population scaled mutation  
 rate  $\theta$ , by setting our observed number of segregating sites in a sample  
 1618 ( $S$ ) equal to this expectation. We'll call this estimator  $\hat{\theta}_W$ :

$$\hat{\theta}_W = \frac{S}{\sum_{i=n-1}^1 1/i} \quad (3.39)$$

This estimator of  $\theta$  was devised by WATTERSON (1975), hence the  
 1620  $W$ .

The neutral site-frequency spectrum. We can use our coalescent process to find the expected number of derived alleles present  $i$  times out of a sample size  $n$ , e.g. how many singletons ( $i = 1$ ) do we expect 1622 to find in our sample? For example, in Figure 3.14 in our sample of four sequences, there are 3 singletons and 2 doubletons. The number 1624 of sites with these different allele frequencies depends on the lengths of specific genealogical branches. A mutation that falls on a branch 1626 with  $i$  descendants will create a derived allele with frequency  $i$ . For example, in our example tree in Figure 3.14, the total number of generations where a mutation could arise and be a doubleton is  $T_3 + 2T_2$ , 1628 the total length of the branch ancestral to just the orange and red 1630 allele ( $T_3 + T_2$ ) plus the branch ancestral to just the blue and purple 1632 allele ( $T_2$ ).

To get a better sense of how  $T_{tot}$  grows with the sample size, we can approximate the sum 3.37 by an integral, which will work for large  $n$ . The result is  $\int_1^{n-1} \frac{4N}{i} di = 4N \log(n - 1)$ .

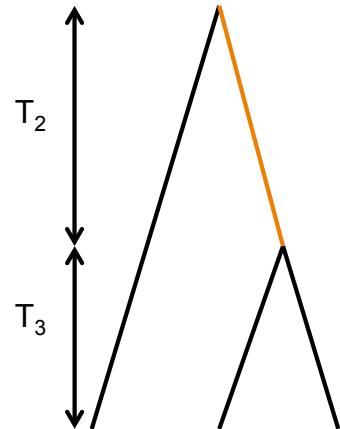


Figure 3.15: A tree for three samples; note that this is the only possible tree shape (treating the tips as unlabeled, i.e. I don't care which pair of sequences carry a doubleton, just that any two sequences carry a derived allele).

To see how we could go about working this out, lets start by considering the simple coalescent tree, shown in Figure 3.15, for sample of 3 alleles drawn from a population. Mutations that fall on the branches coloured in black will be derived singletons, while mutations 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 (3.31), is

$$\mathbb{E}(S_i) = \mu (3\mathbb{E}(T_3) + \mathbb{E}(T_2)) = \mu \left( 3 \frac{2N}{3} + 2N \right) = \theta \quad (3.40)$$

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

Extending this logic to larger samples might be doable, but is tedious (I mean really tedious: for 10 alleles there are thousands 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 HUDSON, but we won't give this here. The general form is:

$$\mathbb{E}(S_i) = \frac{\theta}{i} \quad (3.41)$$

i.e. there are twice as many singletons as doubletons, three times as many singletons as tripletons, and so on. The other thing that will be helpful for us to know is that neutral alleles at intermediate frequency tend to be old, and those that are rare in the sample are 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 four samples. Draw both of them and separately colour (or otherwise mark) the branches by where singletons, doubletons, and tripleton derived alleles could arise.

We can also ask the probability of observing a derived allele segregating at frequency  $i/n$  given that the site is polymorphic in our sample of size  $n$  (i.e. given that  $0 < i < n$ ). We can obtain this probability by dividing the expected number of sites segregating for an allele at frequency  $i$  by the expected number segregating at all of the possible allele frequencies for polymorphisms in our sample

$$P(i|0 < i < n) = \frac{\mathbb{E}(S_i)}{\sum_{j=1}^{n-1} \mathbb{E}(S_j)} = \frac{1/i}{\sum_{j=1}^{n-1} 1/j}. \quad (3.42)$$

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

1670 *tests based on the site frequency spectrum* Population geneticists have  
1672 proposed a variety of ways to test whether an observed site frequency  
1674 spectrum conforms to its neutral, constant-population expectations.  
1676 These tests are useful for detecting population size changes using data  
1678 across many loci, or for detecting the signal of selection at individual  
1680 loci. One of the first tests was proposed by TAJIMA, and is called  
1682 Tajima's  $D$ . Tajima's  $D$  is

$$D = \frac{\theta_\pi - \theta_W}{C} \quad (3.43)$$

1684 where the numerator is the difference between the estimate of  $\theta$  based  
1686 on pairwise differences and that based on segregating sites. As these  
1688 two estimators both have expectation  $\theta$  under the neutral, constant-  
1690 population model, the expectation of  $D$  is zero. The denominator  $C$  is  
1692 a positive constant; it's the square-root of an estimator of the variance  
1694 of this difference under the constant population size, neutral model.  
1696 This constant was chosen for  $D$  to have mean zero and variance 1  
1698 under the null model, so we can test for departures from this simple  
1700 null model.

1702 An excess of rare alleles compared to the constant-population,  
1704 neutral model will result in a negative Tajima's  $D$ , because each ad-  
1706ditional rare allele increases the number of segregating sites by 1, but  
1708 only has a small effect on the number of pairwise differences between  
1710 samples. In contrast, a positive Tajima's  $D$  reflects an excess of inter-  
1712mediate frequency alleles relative to the constant-population, neutral  
1714 expectation. Alleles at intermediate-frequency increase pairwise diver-  
1716 sity more per segregating site than typical, thus increasing  $\theta_\pi$  more  
1718 than  $\theta_W$ .

### 3.3.2 Demography and the coalescent

1718 We've already seen how changes in population size can change the rate  
1720 at which heterozygosity is lost from the population (see the discussion  
1722 around eqn. (3.14)). If the population size in generation  $i$  is  $N_i$ , the  
1724 probability that a pair of lineages coalesce is  $1/2N_i$ ; this conforms to  
1726 our intuition that if the population size is small, the rate at which  
1728 pairs of lineages find their common ancestor is faster. We can poten-  
1730tially accommodate rapid random fluctuations in population size by  
1732 simply using the effective population size  $N_e$  in place of  $N$ . However,  
1734 longer term more systematic changes in population size will distort  
1736 the coalescent genealogies, and hence patterns of diversity, in more  
1738 systematic ways.

1740 We can see how demography potentially distorts the observed fre-  
1742quency spectrum away from the neutral expectation in a very large  
1744 sample of humans shown in Figure 3.20. For comparison, the neu-

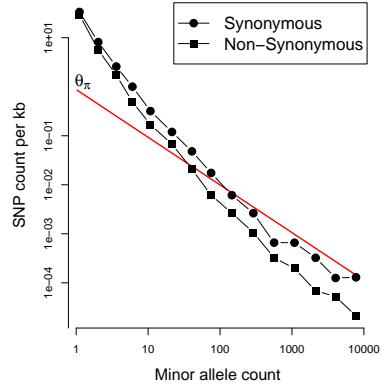
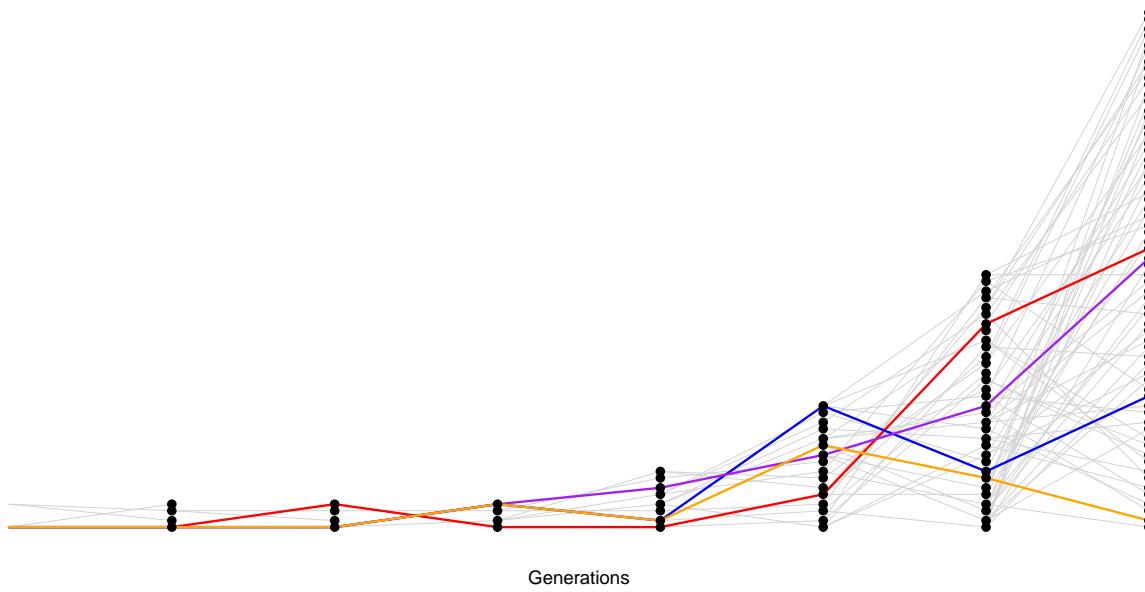


Figure 3.16: Data from 202 genes from 14002 people of European ancestry (28004 alleles). Note the double log-scale. The red line gives the neutral, constant population size estimate of the site frequency spectrum, our equation (3.41), using a  $\theta$  estimated from  $\pi$ . Note how the non-synonymous changes are even more skewed towards rare alleles, that's likely due to selection against non-synonymous alleles acting to push them towards rare frequency. Data from NELSON *et al.* (2012). Code here.

<sup>1710</sup> tral frequency spectrum, eqn (3.41), is shown as a red line. There are vastly more rare alleles than expected under our neutral, constant-  
<sup>1712</sup> population-size model, but the neutral prediction and reality agree somewhat more for alleles that are more common.



<sup>1714</sup> Why is this? Well, these patterns are likely the result of the very recent explosive growth in human populations. If the population has grown rapidly, then the pairwise-coalescent rate in the past may be much higher than the coalescent rate closer to the present. (see Figure 3.17).

<sup>1720</sup> One consequence of a recent population expansion is that there is much less genetic diversity in the population than you'd predict using the census population size. Humans are one example of this effect; <sup>1722</sup> 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. <sup>1724</sup> Our level of genetic diversity is very much lower than you'd predict given our census size, reflecting our much smaller ancestral population. <sup>1726</sup> A second consequence of recent population expansion is that the deeper coalescent branches are much more squished together in time, compared to those in a constant population. Mutations on deeper <sup>1728</sup> branches are the source of alleles at more intermediate frequencies, and so there are even fewer intermediate-frequency alleles in growing <sup>1730</sup> populations. That's why there are so many rare alleles, especially <sup>1732</sup> singletons, in this large sample of Europeans.

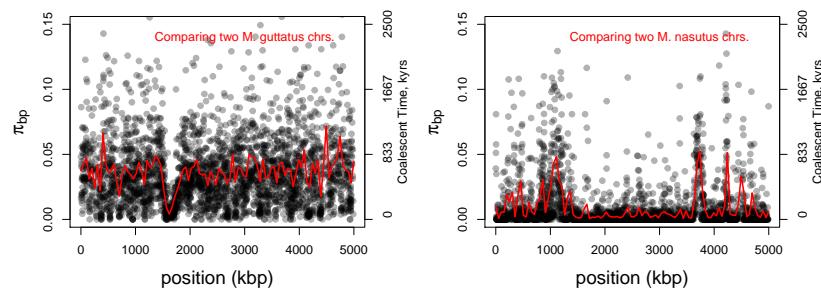
<sup>1734</sup> Another common demographic scenario is a population bottleneck. In a bottleneck, the population size crashes dramatically, and sub-

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.](#)

sequently recovers. For example, our population may have had size  
 1736  $N_{\text{Big}}$  and crashed down to  $N_{\text{Small}}$ . One example of a bottleneck is  
 shown in Figure 3.18. Looking at a sample of lineages drawn from the



1738 population today, if the bottleneck was somewhat recent ( $\ll N_{\text{Big}}$   
 generations in the past) many of our lineages will not have coalesced  
 1740 before reaching the bottleneck, moving backward in time. But during  
 the bottleneck our lineages coalesce at a much higher rate, such that  
 1742 many of our lineages will coalesce if the bottleneck lasts long enough  
 ( $\sim N_{\text{Small}}$  generations). If the bottleneck is very strong, then all of  
 1744 our lineages will coalesce during the bottleneck, and the resulting site  
 frequency spectrum may look very much like our population growth  
 1746 model (i.e. an excess of rare alleles). However, if some pairs of lineages  
 escape coalescing during the bottleneck, they will coalesce much more  
 deeply in time (e.g. the blue and orange ancestral lineages in 3.18).



1748 An example of this is shown Figure 3.19, data from BRANDVAIN  
 1750 et al.. *Mimulus nasutus* is a selfing species that arose recently from an  
 out-crossing progenitor *M. guttatus*, and experienced a strong bottle-  
 1752 neck. *M. guttatus* has a very high levels of genetic diversity ( $\pi = 4\%$   
 at synonymous sites), but *M. nasutus* has lost much of this diversity

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.

Figure 3.19: Diversity along the *Mimulus* genome. Black dots give  $\pi$  in 1kb windows between chromosomes sampled from two individuals, the red line is a moving average (data from BRANDVAIN et al.). Pairwise coalescent times ( $t$ ) estimated assuming  $t = \pi/2\mu$  using  $\mu_{BP} = 10^{-9}$ . Code here.



Figure 3.20: Yellow Monkeyflower *M. guttatus*.

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

1754 ( $\pi = 1\%$ ). Looking along the genome, between a pair of *M. guttatus* chromosomes, levels of diversity are fairly uniformly high.

1756 But in comparing two *M. nasutus* chromosomes, diversity is low because the pair of lineages generally coalesce recently. Yet in a few 1758 places we see levels of diversity comparable to *M. guttatus*; these regions correspond to genomic sites where our pair of lineages fail to 1760 coalesce during the bottleneck and subsequently coalesce much more deeply in the ancestral *M. guttatus* population.



Figure 3.21: Data for polymorphism from Maize and Teosinte: 774 genes from WRIGHT *et al.* (2005). **Left)** Genetic diversity levels in maize and Teosinte samples at each of these genes. 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. Code here.

1762 Mutations that arise on deeper lineages will be at intermediate frequency in our sample, and so mild bottlenecks can lead to an excess 1764 of intermediate frequency alleles compared to the standard constant-population model. This can skew Tajima's D, see eqn 3.43, towards 1766 positive values and away from its expectation of zero. One example 1768 of this skew is shown in Figure 3.21. Maize ((*Zea mays* subsp.*mays*) was domesticated from its wild progenitor teosinte ((*Zea mays* subsp. 1770 *parviglumis*) roughly ten thousand years ago. We can see how the 1772 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.

1774 **Question 10.** VOIGHT *et al.* (2005) sequenced 40 autosomal regions from 15 diploid samples of Hausa people from Yaounde, 1776 Cameroon. The average length of locus they sequenced for each region was 2365bp. They found that the average number of segregating 1778 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 1780 with a bottleneck or growth more consistent with this result?

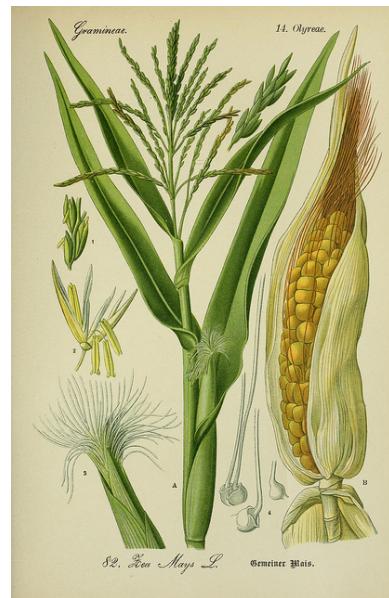


Figure 3.22: Maize (*Zea mays*). Prof. Dr. Thomé's Flora von Deutschland, 1886. Thomé, O. W. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden. Not in copyright.

### 3.4 Molecular Evolution and the fixation of neutral alleles

<sup>1782</sup> "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.

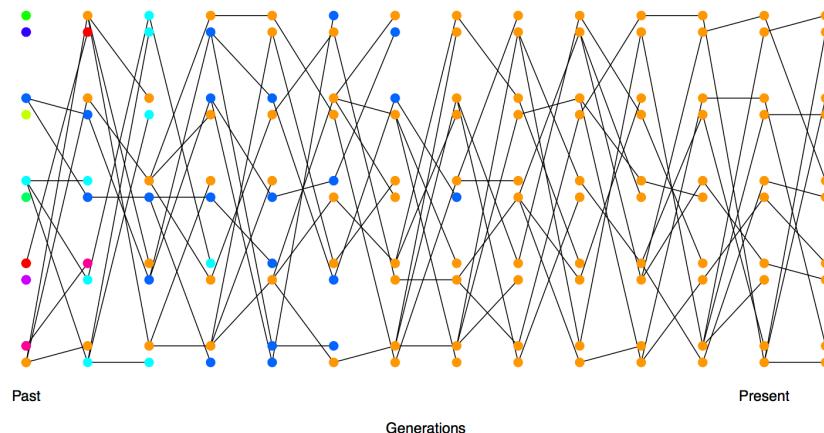


Figure 3.23: Each allele initially present in a small diploid population is given a different colour so we can track their descendants over time. By the 9th generation, all of the alleles present in the population can trace their ancestry back to the orange allele. [Code here.](#)

*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

<sup>1808</sup> in this simulation it does become the common ancestor, by the 9th generation).

<sup>1810</sup> More generally, if our neutral allele is present in  $i$  copies in the population, of  $2N$  alleles, the probability that this allele becomes fixed <sup>1812</sup> 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 <sup>1814</sup> derive this result by letting  $Ns \rightarrow 0$  in eqn. (7.11), a result we'll encounter later.)

<sup>1816</sup> A newly arisen mutation only becomes a fixed difference if it is lucky enough to be the ancestor of the entire population. As we saw <sup>1818</sup> above, this occurs with probability  $1/(2N)$ .

<sup>1820</sup> How long does it take on average for such an allele to fix within our population? Well, in developing equation (3.35) we've seen that it takes  $4N$  generations for a large sample of alleles to all trace their <sup>1822</sup> ancestry back to a single most recent common ancestral allele. Any single-base pair change which arose as a single mutation at a locus, <sup>1824</sup> and fixed in the population, must have been present in the sequence transmitted by the most recent common ancestor of the population <sup>1826</sup> at that locus. Thus it must take roughly  $4N$  generations for a neutral allele present in a single copy within the population to the ancestor <sup>1828</sup> of all alleles within our population. This argument can be made more precise, but in general we would still find that it takes  $\approx 4N$  <sup>1830</sup> generations for a neutral allele to go from its introduction to fixation with the population.

<sup>1832</sup> *Rate of substitution of neutral alleles* A substitution between populations that do not exchange gene flow is simply a fixation event within <sup>1834</sup> one population. The rate of substitution is therefore the rate at which new alleles fix in the population, so that the long-term substitution <sup>1836</sup> rate is the rate at which mutations arise that will eventually become fixed within our population.

<sup>1838</sup> Lets assume, based on our discussion of the neutral theory of molecular evolution, that there are only two classes of mutational changes <sup>1840</sup> that can occur with a region, highly deleterious mutations and neutral mutations. A fraction  $C$  of all mutational changes are highly deleterious, <sup>1842</sup> and cannot possibly contribute to substitution nor polymorphism. The other  $1 - C$  fraction of mutations are neutral. If our mutation rate <sup>1844</sup> is  $\mu$  per transmitted allele per generation, then a total of  $2N\mu(1 - C)$  neutral mutations enter our population each generation.

<sup>1846</sup> Each of these neutral mutations has a  $1/(2N)$  probability chance of eventually becoming fixed in the population. Therefore, the rate at <sup>1848</sup> which neutral mutations arise that eventually become fixed within our population is

$$2N\mu(1 - C) \frac{1}{2N} = \mu(1 - C) \quad (3.44)$$

<sub>1850</sub> Thus the rate of substitution, under a model where newly arising  
 alleles are either highly deleterious or neutral, is simply given by the  
<sub>1852</sub> mutation rate of neutral alleles, i.e.  $\mu(1 - C)$ .

<sub>1854</sub> Consider a pair of species that have diverged for  $T$  generations,  
 i.e. orthologous sequences shared between the species last shared a  
<sub>1856</sub> common ancestor  $T$  generations ago. If these species have maintained  
 a constant  $\mu$  over that time, they will have accumulated an average of

$$2\mu(1 - C)T \quad (3.45)$$

<sub>1858</sub> neutral substitutions. This assumes that  $T$  is a lot longer than the  
<sub>1860</sub> 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  
<sub>1862</sub> total number of substitutions.

<sub>1864</sub> This is a really pretty result as the population size has completely  
 canceled out of the neutral substitution rate. However, there is an-  
<sub>1866</sub> other way to see this in a more straight forward way. If I look at a  
<sub>1868</sub> 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  
<sub>1870</sub> 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 dif-  
<sub>1872</sub> ferences separating our pair of species is simply  $2\mu(1 - C)T$ .

<sub>1874</sub> A number of observations follow under this model, from equation  
<sub>1876</sub> (3.45), the first is that a primary determinant of patterns of molecu-  
<sub>1878</sub> lar evolution in a genomic region is the level of constraint ( $C$ ). This  
<sub>1880</sub> pattern generally seems to hold empirically: non-coding regions often  
<sub>1882</sub> evolve more rapidly than coding regions; synonymous substitutions ac-  
<sub>1884</sub> cumulate faster than nonsynonymous; nonsynonymous changes faster  
<sub>1886</sub> in less vital proteins than ones that are absolutely necessary for early  
<sub>1888</sub> development. Note that this is not a unique prediction of the neu-  
<sub>1890</sub> tral model, e.g. lower pleiotropy means that less constrained regions  
<sub>1892</sub> may be better able to evolve adaptively. However, it is a fantastically  
<sub>1894</sub> useful general insight, e.g. it allows us to spot putatively functional  
<sub>1896</sub> non-coding regions by looking for genomic regions that have very low  
<sub>1898</sub> levels of divergence among distantly related species.

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

- KIMURA and OHTA (1974)



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

1884 The second important insight, and critical for the development of  
 1886 the neutral theory, is that equation (3.45) is seemingly consistent with  
 ZUCKERKANDL and PAULING (1965)'s hypothesis of a surprisingly  
 constant, protein molecular clock. The protein molecular clock is the  
 1888 observation that for some proteins there's a linear relationship between  
 the number of non-synonymous substitutions and the time species last  
 1890 shared a common ancestor in the fossil record. DICKERSON (1971)  
 provided an early example of this observation (Figure 3.24), by  
 1892 comparing various organisms whose molecular sequences were available  
 to him. For example, he found that humans and rattlesnakes, who last  
 1894 share a common ancestor in the fossil record around 300 million years,  
 are separated by roughly 15 NS substitutions per 100 sites in the  
 1896 Cytochrome c protein. While, humans and dog fish, which diverged  
 around 400 million years, are separated by 19 NS substitutions per 100  
 1898 sites in this gene.

In equation (3.45) we double the amount of time separating a pair  
 1900 of species  $T$ , we double the number of substitutions predicted. Note  
 that for this to be true  $T$  must be measured in generations. To ex-  
 1902 plain a protein molecular clock between species that clearly differed  
 dramatically in generation time it was hypothesized that the muta-  
 1904 tion rate actually scaled with generation time, i.e. short-lived organ-  
 isms introduced less mutations per generation, e.g. as they had fewer  
 1906 rounds of mitosis. This generation-time assumption meant that the  
 mutation rate per year could be constant, such that  $\mu T$  would be a  
 1908 constant for pairs of species that had diverged for similar geological



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.

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 7 and OHTA (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 3.26, 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

$$T = T_s + 2N_A. \quad (3.46)$$

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  $5.5 \times 10^6$  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.

Nachman and Crowell sequenced 12 pseudogenes in human and chimp and found substitutions at 1.3% of sites.

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

**B)** All of the pseudogenes they sequenced are on the autosomes. What would your prediction be for pseudogenes on the X and Y chromosomes, given that there are fewer rounds of replication in the female germline than in the male germline.

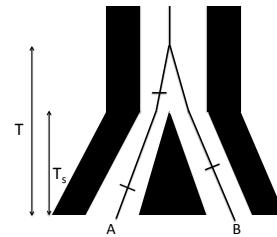


Figure 3.26: The genealogy of two alleles one sampled from population A and B. Mutations on the lineages are shown as dashes. The pair of alleles coalesce in the ancestral population of A and B. The two populations split  $T_s$  generations ago, with no subsequent gene flow, but the two lineages must coalesce deeper in time.

<sup>1948</sup> 3.5 *Tests of molecular evolution.*

<sup>1950</sup> 3.5.1 *Comparing the rates of non-synonymous to synonymous substitutions  $d_N/d_S$*

One common tool in molecular evolution is to compare the estimated number (or rates) of substitutions in different classes of genomic sites, for example the ratio of the number of non-synonymous to synonymous substitutions in a given gene. The simplest way to calculate  $d_N$  is to count up the non-synonymous changes and divide by the total number of positions in the gene where a non-synonymous point 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 to think about what  $d_N/d_S$  represents, however, this ignores the fact that some changes are more likely to occur by mutation than others and also does not account for multiple hits (multiple mutations at the same bp position). Therefore, in practice the ratio  $d_N/d_S$  is more typically calculated by model-based likelihood and bayesian methods that can account for these features.

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 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.)

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

$$d_N = 2T(1 - C)\mu \quad (3.47)$$

Dividing by  $d_S$ , we find

$$d_N/d_S = (1 - C) \quad (3.48)$$

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

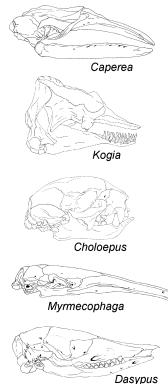
1986 We can test whether our gene is evolving in a constrained way at  
 1988 the protein level by estimating  $d_N/d_S$  and testing whether this is sig-  
 1990 nificantly less than 1. A  $d_N/d_S$  test can provide evolutionary evidence  
 1992 that a stretch of DNA proposed to be protein-coding is subject to se-  
 lective constraint, and so likely does encode for a functional protein.  
 We can also perform a  $d_N/d_S$  test on specific branches of a phylogeny  
 1992 for a gene, to test on which branches the gene is subject to constraint,  
 or to test for changes in the level of constraint across the phylogeny.

1994 *Loss of constraint at pseudogenes.* While most protein genes evolve  
 under constraint, we can find examples of genes that are evolving  
 1996 in a less constrained manner. The simplest example of this is where  
 the gene has lost function. Genes can lose function because of inac-  
 1998 tivating mutations that stop them being transcribed or translated  
 into functional proteins. Such genes are called 'pseudogenes'. When  
 2000 a gene completely loses function there is no longer selection against  
 non-synonymous changes and so such mutations are just as free to ac-  
 2002 cumulate as synonymous changes, and so  $d_N/d_S = 1$ . Pseudogenes  
 are a wonderful example of the extension of Darwin's ideas about  
 2004 vestigial traits ('Rudimentary organs') to the DNA level; we can still  
 recognize a once useful word (gene) whose spelling is slowly degrading.  
 2006 Our genomes are filled with old pseudogenes whose original meanings  
 (functional protein coding sequences) are slowly being eroded through  
 2008 the accumulation of neutral substitutions. One nice example of a  
 gene that has repeatedly lost function, i.e. become repeatedly psue-  
 2010 dogenized, is the Enamlin gene from the study of MEREDITH *et al.*  
 (2009).

C	818	827	1239	1247	2501	2512	2533	2542	4028	4039
<i>Sus</i>	...AAATCAA	CT	TGTTTACTA	..ACATGCC	ATGCA	..GGGGCACAGTTT				
<i>Hippopotamus</i>	...AAATCAA	CT	TGTTTACTA	..ACATGCC	ATGCA	..GGGGCACAGTTT				
<i>Eubalaena glacialis</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCACAGTTT		
<i>Eubalaena australis</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCACAGTTT		
<i>Megaptera</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCACAGTTT		
<i>Caperea</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCACAGTTT		
<i>Eschrichtius</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCACAGTTT		
<i>Kogia sima</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCA	GTTT	
<i>Kogia breviceps</i>	...AAATCAA	CT	TGTTTACTA	..ATA	TGCA	..CATC	TTAGATC	..AGGGCA	GTTT	

D	918	935	1584	1593	1614	1620	2499	2507	4017	4023
<i>Sus</i>	...GGGA	-GTCCAAAAGGCC	..ACCTTCC	..CAAAAC	..CAACATGGC	..GCT	-AGC			
<i>Bradypterus</i>	...???	???	???	???	???	???	???	???	???	???
<i>Choloepus didactylus</i>	...ACT	TCGC	..CAAAAC	..CAACATGGC	..GTT	-AGC				
<i>Choloepus hoffmanni</i>	...???	???	???	???	???	???	???	???	???	???
<i>Myrmecophaga</i>	...GTTG	-TTCAGGAGACTG	..ATT	TCGC	..CAAAAC	..CAACATGGC	..GTT	-AGC		
<i>Tamandua</i>	...GAGA	-TTCAGGAGAATC	..ATT	TCGC	..CAAAAC	..CAACATGGC	..GTT	-AGC		
<i>Cyclopes</i>	...GAGA	-TTCAGGAGAATC	..ATT	TCGC	..CAAAAC	..CAACATGGC	..GTT	-AGC		
<i>Dasypus</i>	...GAGA	-TTCAGGAGAAC	..ATT	TCGC	..CAAAAC	..CAACATGGC	..GTT	AGG		
<i>Tolypeutes</i>	...GAGA	-CTCCAGAGAGTC	..GTC	TCGC	..CAAAAC	..CAACATGGC	..GTT	AGA		
<i>Chaetophractus</i>	...GAGA	-	ATC	..ATCTTACCA	..CAAAAC	..CAATATGGC	..GTT	AGG		
<i>Euphractus</i>	...GAGA	-----	ATC	..ATCTTACCA	..CAAAAC	..CAATATGGC	..GTT	AGG		



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

Figure 3.27: Examples of frameshift mutations (insertions blue, deletions red) and premature stop codons in Enamlin in Cetacea and Xenarthra. Figure from MEREDITH *et al.* (2009), licensed under CC BY 4.0.



Figure 3.28: 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.

ample, 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 3.27 for examples). MEREDITH *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 3.29).



Figure 3.29: 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 pseudogenized (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 MEREDITH *et al.* (2009), licensed under CC BY 4.0.

The branches of the Enamlin phylogeny with a functional Enamlin gene (black) had an estimated  $d_N/d_S = 0.51$ , 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  
 2038 to neutral and deleterious mutations; beneficial mutations must also  
 arise and fix from time to time. Let's assume that a fraction  $B$  of  
 2040 non-synonymous mutations that arise are beneficial such that  $2N\mu B$   
 beneficial mutations arise per generation. Newly arisen beneficial  
 2042 alleles are not destined to fix in the population, as they may be lost to  
 genetic drift when they are rare in the population (we'll discuss how  
 2044 to calculate the fixation probability for beneficial alleles in Chapter  
 7). A newly arisen beneficial allele reaches fixation in the population  
 2046 with probability  $f_B$  from its initial frequency of  $1/2N$ . This fixation  
 probability may be much higher than that of neutral mutations, but  
 2048 still much less than 1. If  $2T$  generations of divergence have elapsed  
 between the two populations then a total of

$$dN = 2T(1 - C - B)\mu + 2T \times (2N\mu B) \times f_B \quad (3.49)$$

2050 non-synonymous substitutions will have accumulated. Then

$$d_N/d_S = (1 - C - B) + 2NBf_B \quad (3.50)$$

assuming again that all synonymous mutations are neutral. Note that  
 2052 this means that our estimates of  $C$  using  $1 - d_N/d_S$  will be a lower  
 bound on the true constraint if even a small fraction of mutations  
 2054 are beneficial. Those cases where the gene is evolving more rapidly  
 at the protein level than at synonymous sites, i.e.  $d_N/d_S > 1$ , are  
 2056 potentially strong candidates for positive selection rapidly driving  
 change at the protein level. We can identify genes that have  $d_N/d_S$   
 2058 significantly greater than one, either on the complete gene phylogeny,  
 or on particular branches. Note that is a very conservative test that  
 2060 few genes in the genome meet, as many genes that are fixing adaptive  
 non-synonymous substitutions will have  $d_N/d_S < 1$ ; even if adaptive  
 2062 mutations are common, genes may still evolve in a constrained way  
 (i.e.  $d_N/d_S < 1$ ) if the rapid fixation of beneficial mutations due to pos-  
 2064 itive selection is outweighed by the loss of non-synonymous mutations  
 to negative selection.

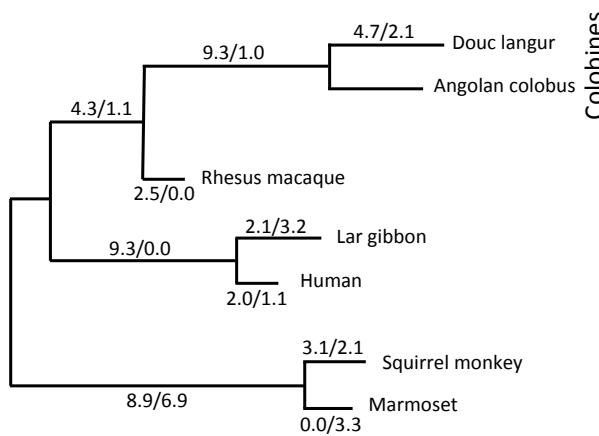


Figure 3.30: A phylogram for the primate lysozyme gene, data from YANG (1998). For each branch, the numbers give the estimated average number of non-synonymous to synonymous changes in the lysozyme protein.

2066 A classic example for looking at adaptive evolution using  $dN/dS$   
is the evolution of the lysozyme protein in primates (MESSIER and  
2068 STEWART, 1997; YANG, 1998), see the phylogeny in Figure 3.30. The  
lysozyme protein is a key component for the breakdown of bacterial  
2070 walls. It shows very fast protein evolution, notably on the lineages  
leading to apes (e.g. gibbons and humans) and Colobines (e.g. colobus  
2072 and langur monkeys). Colobines have leaf-based diets. They digest  
these leaves by bacterial fermentation in their foregut, and then use  
2074 lysozymes to break down the bacteria to extract energy from the  
leaves. In Colobines, the lysozyme protein has evolved to work well  
2076 in the high-PH environment of the stomach. Remarkably, the Colobine  
lysozyme has convergently evolved this activity via very similar amino-  
2078 acid changes at 5 key residuals in cows and Hoatzins (a leaf eating  
bird, KORNEGAY *et al.*, 1994)

2080 *The McDonald-Kreitman test* As noted above, a big issue with using  
 $dN/dS$  to detect adaptation is that it is very conservative. For a more  
2082 powerful test of rapid divergence, what we need to do is adjust for  
the level of constraint a gene experiences at non-synonymous sites.  
2084 One way to do this is to use polymorphism data as an internal control. If we see little non-synonymous polymorphism at a gene, but a  
2086 lot of synonymous polymorphism, we now know that there is likely  
strong constraint on the gene (i.e. high  $C$ ), thus we expect  $dN/dS$  to  
2088 be low. McDONALD and KREITMAN (1991) devised a simple test  
of the neutral theory of molecular evolution at a gene based on this  
2090 intuition (building on the conceptually similar HKA test HUDSON  
*et al.*, 1987). McDONALD and KREITMAN took the case where we  
2092 have polymorphism data at a gene for one species and divergence to

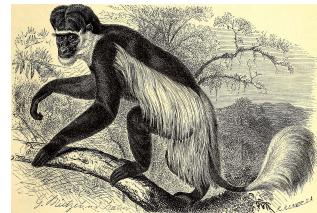


Figure 3.31: Abyssinian black-and-white colobus (*Colobus guereza*). A member of the leaf-eating Colobines. Brehm's Tierleben, Brehm, A.E. 1893. Image from the Biodiversity Heritage Library. Contributed by University of Illinois Urbana-Champaign. Not in copyright.



Figure 3.32: (hoatzin (*Opisthocomus hoazin*)). A leaf-eating bird. A history of birds (1910) Pycraft, W.P. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

a closely related species. They partitioned polymorphism and fixed differences in their sample into non-synonymous and synonymous changes:

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

Under neutral theory, we expect a smaller number of non-synonymous to synonymous fixed differences ( $D_N/D_S < 1$ ) and exactly the same expectation holds for polymorphism ( $P_N/P_S$ ). Let's consider a gene with  $L_S$  and  $L_N$  sites where synonymous and non-synonymous mutations could arise respectively. We can think of the underlying gene genealogy at our gene, see Figure 3.33, with the total time on the coalescent genealogy within the species as  $T_{tot}$  and the total time for fixed differences between our species as  $T'_{div}$ . Then under neutrality we expect  $\mu L_N(1 - C)T_{tot}$  non-synonymous polymorphisms (i.e. our number of segregating sites), and  $\mu L_N(1 - C)T'_{div}$  non-synonymous 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	$L_N(1 - C)/(L_S)$	$L_N(1 - C)/(L_S)$

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

As example of a McDonald-Kreitman table consider the work of FRENTIU *et al.* (2007) on the molecular evolution of L Photopigment opsin in Admiral (*Limenitis*) butterflies, responsible for colour vision in the long-wavelength part of the visual spectrum. FRENTIU *et al.* found that the sensitivity of this opsin had 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:



Figure 3.33: An example ogene genealogy for a set of alleles sampled within a population and a single allele sampled from a distantly-related species.

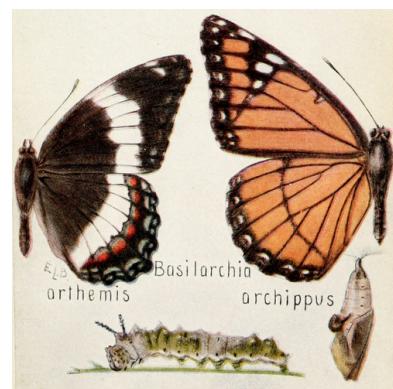


Figure 3.34: White admiral (*Limenitis arthemis*) and Viceroy (*Limenitis archippus*). *Basilarchia* is the old genus that these two species were originally placed in. Viceroy and Monarch butterflies are Müllerian mimics.  
Field book of insects (1918). Lutz, F.E. . illustrations by Edna L. Beutenmüller. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

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

2128 Note the strong excess of non-synonymous to synonymous diver-  
 2129 gence compared to polymorphism (p-value of 0.006, Fisher's exact  
 2130 test), which is consistent with the gene evolving in an adaptive man-  
 2131 ner among the two species. We would expect roughly only 3 non-  
 2132 synonymous substitutions out of 16 substitutions if the gene was  
 2133 evolving neutrally ( $16 \times 2/11$ ).

2134 *3.6 Neutral diversity and population structure*

We've considered alleles drawn from a randomly-mating population,  
 2135 and divergence among alleles drawn from two distantly-related pop-  
 2136 ulations. We'll now turn to consider divergence among more closely  
 2137 related populations. In thinking about the coalescent within pop-  
 2138 ulations we made the assumption that any pair of lineages is equally  
 2139 likely to coalesce with each other. However, when there is population  
 2140 structure this assumption is violated.

2141 We have previously written the measure of population structure  
 $F_{ST}$  as

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad (3.51)$$

2142 where  $H_S$  is the probability that two alleles sampled at random from  
 2143 a subpopulation differ, and  $H_T$  is the probability that two alleles  
 2144 sampled at random from the total population differ.

2145 *A simple population split model* Imagine a population of constant size  
 2146 of  $N_e$  diploid individuals that  $T$  generations in the past split into two  
 2147 daughter populations (sub-populations) each of size  $N_e$  individuals,  
 2148 which do not subsequently exchange migrants. In the current day we  
 2149 sample an equal number of alleles from both subpopulations.

2150 Consider a pair of alleles sampled within one of our sub-populations  
 2151 and think about their per site heterozygosity. These alleles have expe-  
 2152 rienced a population of size  $N_e$  and so the probability that they differ  
 2153 is  $H_S \approx 4N_e\mu$  (assuming that  $N_e\mu \ll 1$ , using our equation 3.12 for  
 2154 heterozygosity within a population ).

2155 The heterozygosity in our total population is a little more tricky  
 2156 to calculate. Assuming that we equally sample both sub-populations,  
 2157 when we draw two alleles from our total sample, 50% of the time  
 2158 they are drawn from the same subpopulation and 50% of the time  
 2159 they are drawn from different subpopulations. Therefore, our total

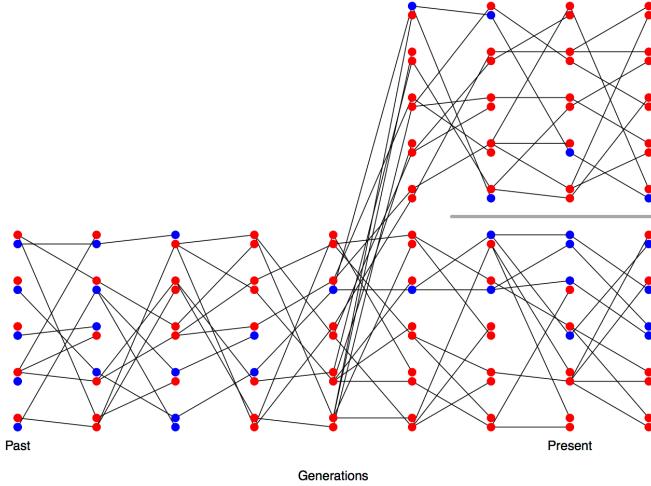


Figure 3.35: Change in allele frequencies following a population split. Code here.

2162 heterozygosity is given by

$$H_T = \frac{1}{2}H_S + \frac{1}{2}H_B \quad (3.52)$$

2164 where  $H_B$  is the probability that a pair of alleles drawn from our two different sub-populations differ from each other. A pair of alleles from different sub-populations cannot find a common ancestor with each 2166 other for at least  $T$  generations into the past as they are in distinct populations (not connected by migration). Once our alleles find them- 2168 selves back in the combined ancestral population it takes them on average  $2N$  generations to coalesce. So the total opportunity for mu- 2170 tation between our pair of alleles sampled from different populations is  $2(T + 2N)$  generations of meioses, such that the probability that our 2172 pairs of alleles is different is

$$H_B \approx 2\mu(T + 2N) \quad (3.53)$$

We can plug this into our expression for  $H_T$ , and then that in turn 2174 into  $F_{ST}$ . Doing so we find that

$$F_{ST} \approx \frac{\mu T}{\mu T + 4N_e \mu} = \frac{T}{T + 4N_e} \quad (3.54)$$

2176 Note that  $\mu$  cancels out of this equation. In this simple toy model,  $F_{ST}$  is increasing because the amount of between-population diversity 2178 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 Suma- 2182 tran orang-utan species samples (*Pongo pygmaeus* and *Pongo abelii*)

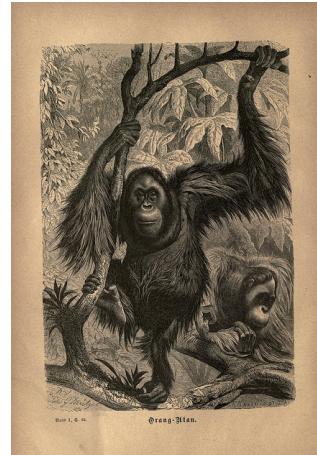


Figure 3.36: Orangutan (*Pongo*). Brehms thierleben, allgemeine kunde des thierreichs. Brehm, A. E. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

is  $\approx 0.37$  (LOCKE *et al.*, 2011), representing a deep population split  
 2184 between the species (potentially with little subsequent gene flow).

Within the populations the genome-wide average Watterson's  $\theta$  is  
 2186  $\theta_W = 1.4\text{kb}^{-1}$ , estimated from the number of segregating sites. Assume a generation time of 20 years, and a mutation rate of  $2 \times 10^{-8}$   
 2188 per base per generation. How far in the past did the two populations diverge?

2190 *A simple model of migration between an island and the mainland.* We can also use the coalescent to think about patterns of differentiation  
 2192 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.  
 2194 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  
 2198 to our mainland population. Each generation some low fraction  $m$  of our individuals on the island have migrant parents from the mainland  
 2200 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  
 2202 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  
 2204 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  
 2206 of alleles sampled on the island are descended from a shared recent common ancestral allele on the island is the probability that our pair  
 2208 of alleles coalesces before either lineage migrates. For example, the probability that our pair of alleles coalesces  $t + 1$  generations back on  
 2210 the island is  
 2212

$$\frac{1}{2N_I} (1-m)^{2(t+1)} \left(1 - \frac{1}{2N_I}\right)^t \approx \frac{1}{2N_I} \exp\left(-t\left(\frac{1}{2N_I} + 2m\right)\right), \quad (3.55)$$

with the approximation following from assuming that  $m \ll 1$  &  
 2214  $\frac{1}{(2N_I)} \ll 1$  (note that this is very similar to our derivation of heterozygosity above). The probability that our alleles coalesce before  
 2216 either one of them migrates off the island, irrespective of the time, is

$$\int_0^\infty \frac{1}{2N_I} \exp\left(-t\left(\frac{1}{2N_I} + 2m\right)\right) dt = \frac{1/(2N_I)}{1/(2N_I) + 2m}. \quad (3.56)$$

Let's assume that the mutation rate is very low such that it is very  
 2218 unlikely that the pair of alleles mutate before they coalesce on the island. Therefore, the only way that the alleles can be different from

<sup>2220</sup> each other is if one or other of them migrates to the mainland, which happens with probability

$$1 - \frac{1/(2N_I)}{1/(2N_I) + 2m} \quad (3.57)$$

<sup>2222</sup> Conditional on one or other of our alleles migrating to the mainland, both of our alleles represent independent draws from the mainland and <sup>2224</sup> so differ from each other with probability  $H_M$ . Therefore, the level of heterozygosity on the island is given by

$$H_I = \left(1 - \frac{1/(2N_I)}{1/(2N_I) + 2m}\right) H_M \quad (3.58)$$

<sup>2226</sup> So the reduction of heterozygosity on the island compared to the mainland is

$$F_{IM} = 1 - \frac{H_I}{H_M} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{1 + 4N_I m}. \quad (3.59)$$

<sup>2228</sup> The level of inbreeding on the island compared to the mainland will be high if the migration rate is low and the effective population size <sup>2230</sup> of the island is low, as allele frequencies on the island are drifting and diversity on the island is not being replenished by migration. The key <sup>2232</sup> parameter here is the number individuals on the island replaced by immigrants from the mainland each generation ( $N_I m$ ).

<sup>2234</sup> We have framed this problem as being about the reduction in genetic diversity on the island compared to the mainland. However, if we <sup>2236</sup> consider collecting individuals on the island and mainland in proportion to their population sizes, the total level of heterozygosity would <sup>2238</sup> be  $H_T = H_M$ , as samples from our mainland would greatly outnumber those from our island. Therefore, considering the island as our <sup>2240</sup> sub-population, we have derived another simple model of  $F_{ST}$ .

**Question 13.** You are investigating a small river population of <sup>2242</sup> sticklebacks, which receives infrequent migrants from a very large marine population. At a set of putatively neutral biallelic markers the <sup>2244</sup> freshwater population has frequencies:

0.2, 0.7, 0.8

<sup>2246</sup> at the same markers the marine population has frequencies:

0.4, 0.5 and 0.7.

<sup>2248</sup> From studying patterns of heterozygosity at a large collection of markers, you have estimated the long term effective size of your fresh-<sup>2250</sup> water population is 2000 individuals.

<sup>2252</sup> What is your estimate of the migration rate from the marine populations into the river?

*Incomplete lineage sorting* Because it can take a long time for an <sup>2254</sup> polymorphism to drift up or down in frequency, multiple population

splits may occur during the time an allele is still segregating. This  
2256 can lead to incongruence between the overall population tree and the  
information about relationships present at individual loci. In Figure  
2258 3.37 and 3.38 we show simulations of three populations where the  
bottom population splits off from the other two first, followed by  
2260 the subsequent splitting of the top and the middle populations.  
We start both simulations with a newly introduced red allele being  
2262 polymorphic in the combined ancestral population. The most likely  
fate of this allele is that it is quickly lost from the population, but  
2264 sometimes the allele can drift up in frequency and be polymorphic  
when the populations split, as the alleles in our two figures have done.  
2266 If the allele is lost/fixed in the descendant populations before the next  
population split, our allele configuration will agree with the population  
2268 tree, as it does in Figure 3.37, and so too the gene tree will agree with  
population tree (as shown in the left side of Figure 3.39). However,  
2270 if the allele persists as a polymorphism in the ancestral population  
till the top and the middle populations split, then the allele can fix in  
2272 one of these populations and not the other. Such an event can lead to  
a substitution pattern that disagrees with the population tree, as in  
2274 Figure 3.38. If we were to construct a phylogeny using the variation  
at this site we would see a disagreement between the gene tree and  
2276 population tree. In Figure 3.38 an allele drawn from the top and the  
bottom populations are necessarily more closely related to each other  
2278 than either is to an allele drawn from population 2; tracing our allelic  
lineages from the top and bottom populations back through time, they  
must coalesce with each other before we reach the point where  
2280 the red mutation arose; in contrast, a lineage from the middle population  
cannot have coalesced with either other lineage until past the time the  
red mutation arose. An example of this 'incomplete lineage sorting' in  
2282 terms of the underlying tree is shown on the right side of Figure 3.39 .  
2284

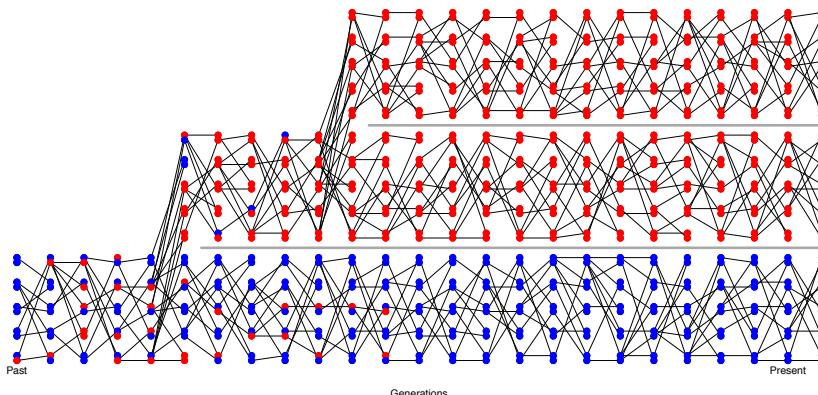


Figure 3.37: An example of alleles assorting among three populations such that there is no incomplete lineage sorting. Code here.

A natural pedigree analogy to incomplete lineage sorting is the fact



Figure 3.38: An example of alleles assorting among three populations leading to incomplete lineage sorting. Code here.



Figure 3.39: 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.

2286 that while two biological siblings are more closely related to each other  
 2287 genealogically than either is to their cousin, at any given locus one of  
 2288 the siblings can share an allele IBD with their cousin that they do not  
 2289 share with their own sibling, due to the randomness of Mendelian seg-  
 2290 regation down their pedigree. In these cases, the average relatedness of  
 2291 the individuals/populations disagrees with the patterns of relatedness  
 2292 at a particular locus.

As an empirical example of incomplete lineage sorting, let's consider  
 2294 the work of JENNINGS and EDWARDS who sequenced a single allele  
 2295 from three different species of Australian grass finches (*Poephila*): two  
 2296 sister species of long-tailed finches (*Poephila acuticauda* and *P. hecki*)  
 2297 and the black-throated finch (*Poephila cincta*, see Figure 3.40). They  
 2298 collected sequence data for 30 genes, and constructed phylogenetic  
 2299 gene trees at each of these loci, resulting in 28 well-resolved gene trees.  
 2300 16 of the gene trees showed *P. acuticauda* and *P. hecki* as sisters with  
*P. cincta* (the tree ((A,H),C)), while for twelve genes the gene tree  
 2301 was discordant with the population tree: for seven of their genes *P.*  
*hecki* fell as an outgroup to the other two and at five *P. acuticauda* fell  
 2302 as an outgroup (the trees ((A,C),H) and ((H,C),A) respectively).

Let's use the coalescent to understand this discordance between  
 2306 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



Figure 3.40: 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.

2308 third outgroup population (C)  $t_2$  generations in the past. We'll assume that there's no gene flow among our populations after each split.  
 2310 We can trace back the ancestral lineages of our three alleles. The first opportunity for the A & B lineages to coalesce is  $t_1$  generations ago.  
 2312 If they coalesce with each other in their shared ancestral population before  $t_2$  in the past (left side of Figure 3.39) their gene 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 & B lineages to fail to coalesce in their shared ancestral population between  $t_1$  and  $t_2$ ; this happens with probability  $(1 - 1/2N)^{t_2-t_1}$ . We'll get a discordant gene tree if A & B make it back to the shared ancestral population with C without coalescing, and then one or the other of them coalesces with the C lineage before they coalesce with each other. This happens with probability 2/3, as at the first pairwise-coalescent event there are 2318 are three possible pairs of lineages that could coalesce, two of which (A & C and B & C ) result in a discordant tree. So the probability 2320 that we get a coalescent tree that is discordant with the population tree is

$$\frac{2}{3} (1 - 1/2N)^{t_2-t_1}. \quad (3.60)$$

2326 Thus we should expect gene-tree population-tree discordance when populations split in rapid succession and/or population sizes are large.

2328

**Question 14.** Let's return to JENNINGS and EDWARDS's Australian grass finches example. They estimated that the ancestral population size of our two long-tailed finches was four hundred thousand. 2330 What is your best estimate of the inter-speciation time, i.e.  $t_2 - t_1$ ?

2334 *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 2336 occurred or demonstrate that gene flow occurred after two populations began to speciate. A broad range of methods have been designed to 2338 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 2340 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 2342 also lead to gene-tree discordance. While both ILS and gene flow can 2344 lead to discordance, under simplifying assumptions, ILS implies more symmetry in how these discordances manifest themselves.

2346 Take a look at Figure 3.41. In both cases the lineages from A and B fail to coalesce in their initial shared ancestral population, and one 2348 or the other of them coalesces with the lineage from C before they



Figure 3.41: 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 coalescence with D. The small dash on the branch indicates the mutation A  $\rightarrow$  B occurring, giving rise to the ABBA or BABA mutational pattern shown at the bottom.

coalesce with each other. Each option is equally likely; therefore the 2350 mutational patterns ABBA and BABA are equally likely to occur under ILS.<sup>6</sup>

2352 However, if gene flow occurs from population C into population B, in addition to ILS the lineage from B can more recently coalesce with 2354 the lineage from C, and so we should see more ABBAs than BABAs. To test for this effect of gene flow, we can sample a sequence from 2356 each of our 4 populations and count up the number of sites that show the two mutational patterns consistent with the gene-tree discordance 2358  $n_{ABBA}$  and  $n_{BABA}$  and calculate

$$\frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}} \quad (3.61)$$

This statistic will have expectation zero if the gene-tree discordance is 2360 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).

<sup>6</sup> here we have to assume no structure in the ancestral population.

## *Phenotypic Variation and the Resemblance Between Relatives.*

THE DISTINCTION BETWEEN GENOTYPE AND PHENOTYPE is one of the most useful ideas in Biology.<sup>1</sup> 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, WANG *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. fit a linear regression of phenotypes for our individuals on their geno-

<sup>1</sup> JOHANNSEN, W., 1911 The Genotype Conception of Heredity. *The American Naturalist* 45(531): 129–159



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

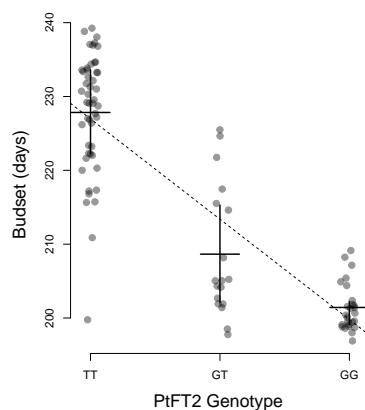


Figure 4.2: The effect of a flowering time gene (*PtFT2*) SNP on budset time in European aspen. Each dot gives the genotype-phenotype combination for an individual. The horizontal lines give the budset mean for each genotype and the vertical lines show the inter-quartile range. The dotted line gives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

<sup>2394</sup> types at a particular SNP ( $l$ ):

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

In the equation above,  $X$  is a vector of the phenotypes of a set of individuals and  $G_l$  is our vector of genotypes at locus  $l$ , with  $G_{i,l}$  taking the value 0, 1, or 2 depending on whether our individual  $i$  is homozygote, heterozygote, or the alternate homozygote at our locus of interest. Here  $\mu$  is our phenotypic mean. The slope of this regression line ( $a_l$ ) has the interpretation of being the average effect of substituting 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 budset forward by 13.6 days, such that the  $GG$  homozygote is predicted to set buds 27.2 days earlier than the  $TT$  homozygote.

As a measure of the significance of this genotype-phenotype relationship, we can calculate the p-value of our regression. To try and identify loci that are associated with our trait genome-wide, we can conduct this regression at each SNP we genotype in the genome. 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 logarithm of the p-value for each SNP along genome (a so-called Manhattan plot). Here's one from WANG *et al.* (2018) for their Aspen budset phenotype

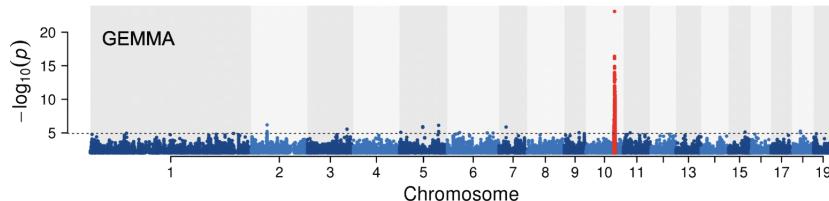


Figure 4.3: Manhattan plot of the p-value of the linear association between genotype and budset in Aspen. Each dot represents the test at a single SNP, plotted at its physical coordinate in the genome. Different chromosomes are plotted in alternating colours. The SNPs surrounding the PtFT2 gene are shown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

<sup>2414</sup> The SNP with the most significant p-value is the PtFT2 SNP. Note  
 that other SNPs in the surrounding region also light up as showing a  
<sup>2416</sup> significant association with budset timing. This is because loci that  
 are in LD with a functional locus may in turn show an association,  
<sup>2418</sup> not because they directly affect the phenotype, but simply because  
 the genotypes at the two loci are themselves non-randomly associated.  
<sup>2420</sup> Below is a zoomed in version (Figure 2 in WANG *et al.* (2018)) with  
 SNPs coloured by the strength of their LD with the putatively func-  
<sup>2422</sup> tional SNP. Note how SNPs in strong LD with the functional allele  
 (redder points) have more significant p-values.

<sup>2424</sup> Variation in some traits seems to have a relatively simple genetic basis. In our Aspen example there is one clear large-effect locus, which  
<sup>2426</sup> explains 62% of the variation in budset. Note that even in this case, where we have an allele with a very strong effect on a phenotype, this

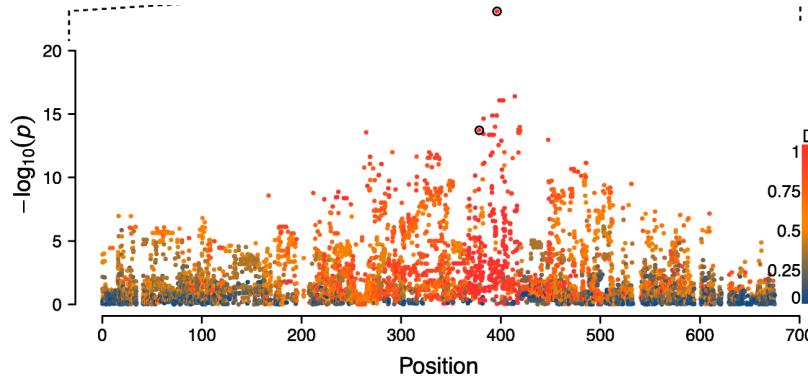


Figure 4.4: The Manhattan plot zoomed in on the top-hit (red SNPs from Figure 4.3). SNPs are now coloured by their  $D_f$  value with the most significant SNP.  $D_f$  is the LD covariance between a pair of loci ( $D$ ) normalized by the largest value  $D$  can take given the allele frequencies. Figure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

2428 is not an allele *for* budset, nor is PtFT2 a gene *for* budset. It is an  
 2429 allele that is associated with budset in the sampled environments and  
 2430 populations. In a different set of environments, this allele's effects  
 2431 may be far smaller, and a different set of alleles may contribute to  
 2432 phenotype variation. PtFT2, the gene our focal SNP falls close to, is  
 2433 just one of many genes and molecular pathways involved in budset.  
 2434 A mutant screen for budset may uncover many genes with larger ef-  
 2435 fects; this gene is just a locus that happens to be polymorphic in this  
 2436 particular set of genotyped individuals.

While phenotypic variation for some phenotypes has a relatively  
 2437 simple genetic basis, many phenotypes are likely much more genetically  
 2438 complex, involving the functional effect of many alleles at hun-  
 2439 dreds or thousands of polymorphic loci. For example hundreds of  
 2440 small effect loci affecting human height have been mapped in Euro-  
 2441 pean populations to date. Such genetically complex traits are called  
 2442 polygenic traits.

2444 In this chapter, we will use our understanding of the sharing of  
 2445 alleles between relatives to understand the phenotypic resemblance  
 2446 between relatives in quantitative phenotypes. This will allow us to  
 2447 understand the contribution of genetic variation to phenotypic varia-  
 2448 tion. In the next chapter, we will then use these results to understand  
 2449 the evolutionary change in quantitative phenotypes in response to  
 2450 selection.

#### 4.0.1 A simple additive model of a trait

2452 Let's imagine that the genetic component of the variation in our trait  
 2453 is controlled by  $L$  autosomal loci that act in an additive manner. The  
 2454 frequency of allele 1 at locus  $l$  is  $p_l$ , with each copy of allele 1 at this  
 2455 locus increasing your trait value by  $a_l$  above the population mean.  
 2456 The phenotype of an individual, let's call her  $i$ , is  $X_i$ . Her genotype  
 2457 at SNP  $l$  is  $G_{i,l}$ . Here  $G_{i,l} = 0, 1$ , or  $2$ , representing the number of

"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." - STURTEVANT (1915)

<sup>2458</sup> copies of allele 1 she has at this SNP. Her expected phenotype, given her genotype at all  $L$  SNPs, is then

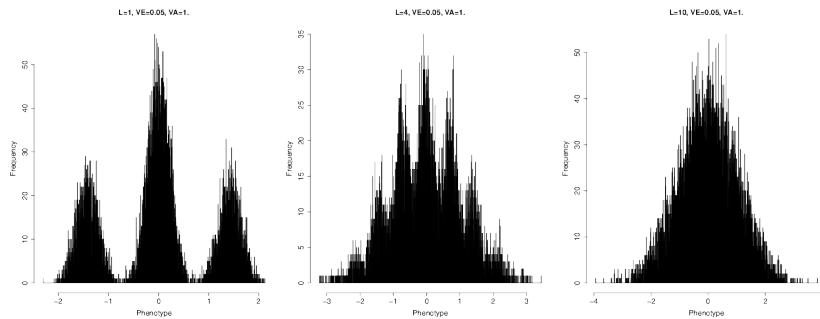
$$\mathbb{E}(X_i|G_{i,1}, \dots, G_{i,L}) = \mu + X_{A,i} = \mu + \sum_{l=1}^L G_{i,l}a_l \quad (4.2)$$

<sup>2460</sup> where  $\mu$  is the mean phenotype in our population, and  $X_{A,i}$  is the deviation away from the mean phenotype due to her genotype. Now <sup>2462</sup> in reality the phenotype is a function of the expression of those alleles in a particular environment. Therefore, we can think of this expected <sup>2464</sup> phenotype as being an average across a set of environments that occur in the population.

<sup>2466</sup> When we measure our individual's observed phenotype we see

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

<sup>2468</sup> where  $X_E$  is the deviation from the mean phenotype due to the environment. This  $X_E$  includes the systematic effects of the environment our individual finds herself in and all of the noise during development, <sup>2470</sup> growth, and the various random insults that life throws at our individual. If a reasonable number of loci contribute to variation in our <sup>2472</sup> trait then we can approximate the distribution of  $X_{A,i}$  by a normal distribution due to the central limit theorem (see Figure 4.5). Thus if we can approximate the distribution of the effect of environmental <sup>2474</sup> variation on our trait ( $X_{E,i}$ ) also by a normal distribution, which is reasonable as there are many small environmental effects, then the <sup>2476</sup> distribution of phenotypes within the population ( $X_i$ ) will be normally distributed (see Figure 4.5).



<sup>2480</sup> 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

$$X_i = \mu + X_{iM} + X_{iP} + X_{iE} \quad (4.4)$$

<sup>2482</sup> where  $X_{iM}$  and  $X_{iP}$  are the contribution to the phenotype of the alleles that our individual received from her mother (maternal alleles) and

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 ( $L = 1, 4, \text{ and } 10$ , 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.

father (paternal alleles) respectively. This will come in handy in just  
 2484 a moment when we start thinking about the phenotypic covariance of  
 relatives.

2486 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  
 2488 same loci (dominance) and at different loci (epistasis). Later we'll  
 relax this assumption, however, we'll find that if we are interested  
 2490 in evolutionary change over short time-scales it is actually only the  
 "additive component" of genetic variation that will (usually) concern  
 2492 us. We will define this more formally later on, but for the moment  
 we can offer the intuition that parents only get to pass on a single  
 2494 allele at each locus on to the next generation. As such, it is the effect  
 of these transmitted alleles, averaged over possible matings, that is  
 2496 an individual's average contribution to the next generation (i.e. the  
 additive effect of the alleles that their genotype consists of).

#### 2498 4.0.2 Additive genetic variance and heritability

As we are talking about an additive genetic model, we'll talk about  
 2500 the additive genetic variance ( $V_A$ ), the phenotypic variance due to the  
 additive effects of segregating genetic variation. This is a subset of the  
 2502 total genetic variance if we allow for non-additive effects.

The variance of our phenotype across individuals ( $V$ ) we can write  
 2504 as

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

In writing the phenotypic variance as a sum of the additive and environmental contributions, we are assuming that there is no covariance  
 2506 between  $X_{G,i}$  and  $X_{E,i}$  i.e. there is no covariance between genotype  
 2508 and environment.

Our additive genetic variance can be written as

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

2510 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  
 2512 are in linkage equilibrium, we can write our additive genetic variance  
 as

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

2514 where the  $2p_l(1 - p_l)$  term follows from the binomial sampling of two  
 alleles per individual at each locus.

2516 **Question 1.** You have two biallelic SNPs contributing to variance  
 in human height. At the first SNP you have an allele with an additive

2518 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  $-0.5\text{cm}$  segregat-  
 2520 ing at 50% frequency. Which SNP contributes more to the additive  
 genetic variance? Explain the intuition of your answer.

2522 *An example of calculating polygenic scores.* Now we don't usually  
 get to see the individual loci contributing to highly polygenic traits.  
 2524 Instead, we only get to see the distribution of the trait in the popu-  
 lation. However, with the advent of GWAS in human genetics we can  
 2526 see some of the underlying genetics using the many trait-associated  
 loci identified to date. Using the estimated effect sizes at each locus,  
 2528 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  
 2530 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  
 2532 the strongest signal of association with height in 1700 roughly inde-  
 pendent bins spaced across the genome. The effects of these SNPs are  
 2534 tiny; the medium, absolute additive effect size is 0.07cm. Figure 4.6  
 shows the distribution of a thousand individuals' polygenic scores cal-  
 2536 culated using these 1700 SNPs (simulated genotypes using the UKBB  
 frequencies). The standard deviation of these polygenic scores  $\sim 2\text{cm}$ .  
 2538 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.

2540 *The narrow sense heritability* We would like a way to think about  
 what proportion of the variation in our phenotype across individuals  
 2542 is due to genetic differences as opposed to environmental differences.  
 Such a quantity will be key in helping us think about the evolution of

<sup>2544</sup> phenotypes. For example, if variation in our phenotype had no genetic basis, then no matter how much selection changes the mean phenotype <sup>2546</sup> within a generation the trait will not change over generations.

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

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

<sup>2545</sup> Remember that we are thinking about a trait where all of the alleles act in a perfectly additive manner. In this case our heritability  $h^2$  <sup>2550</sup> is referred to as the *narrow sense heritability*, the proportion of the variance explained by the additive effect of our loci. When we allow <sup>2552</sup> dominance and epistasis into our model, we'll also have to define the <sup>2554</sup> *broad sense heritability* (the total proportion of the phenotypic variance attributable to genetic variation).

<sup>2555</sup> The narrow sense heritability of a trait is a useful quantity; indeed we'll see shortly that it is exactly what we need to understand the <sup>2558</sup> evolutionary response to selection on a quantitative phenotype. We can calculate the narrow sense heritability by using the resemblance <sup>2560</sup> between relatives. For example, if the phenotypic differences between individuals in our population were solely determined by environmental <sup>2562</sup> differences experienced by these different individuals, we should not expect relatives to resemble each other any more than random individuals drawn from the population. Now the obvious caveat here is that <sup>2564</sup> relatives also share an environment, so may resemble each other due to shared environmental effects.

<sup>2566</sup> Note that the heritability is a property of a sample from the population in a particular set of environments at a particular time.

<sup>2568</sup> Changes in the environment may change the phenotypic variance.

<sup>2570</sup> Changes in the environment may also change how our genetic alleles are expressed through development and so change  $V_A$ . Thus estimates <sup>2572</sup> of heritability are not transferable across environments or populations.

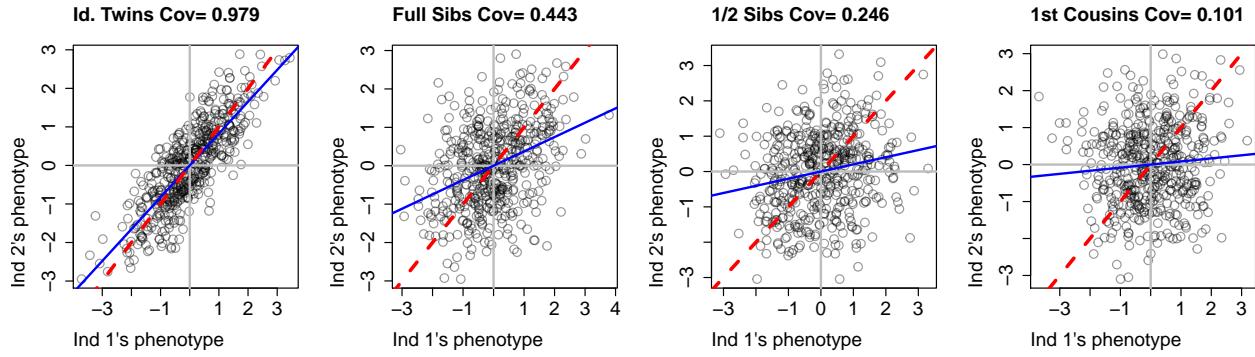
#### 4.0.3 The covariance between relatives

<sup>2574</sup> So we'll go ahead and calculate the covariance in phenotype between two individuals (1 and 2) who have phenotypes  $X_1$  and  $X_2$  respectively. To think about imagine plotting the phenotypes of, say, sisters <sup>2576</sup> against each other. The x and y coordinates of each point will be <sup>2578</sup> the, 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 <sup>2580</sup> their phenotypes covary. If some of the variation in our phenotype is genetic we expect identical twins to resemble each other more than full siblings, who in turn will resemble each other more than half-sibs <sup>2582</sup> and so on out (see Figure 4.7). Under our simple additive model of

2584 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 various  
2586 components in these sums.



To make our task easier, we will make two commonly made assumptions:  
2588

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

The failure of these assumptions to hold can undermine our estimates of heritability, but we'll return to that later. Moving forward with these assumptions, we can simplify our original expression above  
2592 and write our phenotypic covariance between our pair of individuals as  
2594

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

This equation is saying that, under our simple additive model, we can see the covariance in phenotypes between individuals as the covariance  
2600 between the maternal and paternal allelic effects in our individuals.  
2602 We can use our results about the sharing of alleles between relatives to obtain these covariance terms. But before we write down the general  
2604 case, let's quickly work through some examples.

*2606 The covariance between identical twins* Let's first consider the case of a pair of identical twins from two unrelated parents. Our pair of

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.

<sup>2608</sup> twins share their maternal and paternal allele identical by descent ( $X_{1M} = X_{2M}$  and  $X_{1P} = X_{2P}$ ). As their maternal and paternal alleles  
<sup>2610</sup> are not correlated draws from the population, i.e. have no probability  
<sup>2612</sup> of being *IBD* as we've said the parents are unrelated, the covariance  
<sup>2614</sup> between their effects on the phenotype is zero (i.e.  $Cov(X_{1P}, X_{2M}) = Cov(X_{1M}, X_{2P}) = 0$ ). In that case, eqn. 4.10 is

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

<sup>2614</sup> Now in general identical twins are not going to be super helpful for us in estimating  $h^2$ , because under models with non-additive effects,  
<sup>2616</sup> identical twins will have higher covariance than we'd expect just based on the alleles they share. This is because identical twins don't just  
<sup>2618</sup> share alleles, they share their entire genotypes, and thus resemble each other in phenotype also because of shared dominance effects.

<sup>2620</sup> *The covariance in phenotype between mother and child* If a mother and father are unrelated individuals (i.e. are two random draws from  
<sup>2622</sup> the population) then this mother and her child share one allele IBD at each locus (i.e.  $r_1 = 1$  and  $r_0 = r_2 = 0$ ). Half the time our  
<sup>2624</sup> mother (ind 1) transmits her paternal allele to the child (ind 2), in which case  $X_{P1} = X_{M2}$ , and so  $Cov(X_{P1}, X_{M2}) = Var(X_{P1})$ ,  
<sup>2626</sup> and all the other covariances in eqn. 4.10 are zero. The other half of the time she transmits her maternal allele to the child, in which  
<sup>2628</sup> case  $Cov(X_{M1}, X_{M2}) = Var(X_{M1})$  and all the other terms are zero. By this argument,  $Cov(X_1, X_2) = \frac{1}{2}Var(X_{M1}) + \frac{1}{2}Var(X_{P1}) = \frac{1}{2}V_A$ .

<sup>2630</sup> *The covariance between general pairs of relatives under an additive model* The two examples above make clear that to understand  
<sup>2632</sup> the covariance between phenotypes of relatives, we simply need to think about 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 respectively. When they share zero alleles  
<sup>2634</sup>  $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = 0$ , when they share one allele  
<sup>2636</sup>  $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = Var(X_{1M}) = \frac{1}{2}V_A$ , and when they  
<sup>2638</sup> share two alleles  $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = V_A$ . Therefore, the general covariance between two relatives is

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

<sup>2640</sup> So under a simple additive model of the genetic basis of a phenotype, to measure the narrow sense heritability we need to measure the  
<sup>2642</sup> 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  $h^2$ .

2646 **Question 2. A)** In polygynous red-winged blackbird populations  
 (i.e. males mate with several females), paternal half-sibs can be iden-  
 2648 tified. Suppose that the covariance of tarsus lengths among half-sibs  
 is  $0.25 \text{ cm}^2$  and that the total phenotypic variance is  $4 \text{ cm}^2$ . Use these  
 2650 data to estimate  $h^2$  for tarsus length in this population.

2652 **B)** Why might paternal half-sibs be preferable for measuring heri-  
 tability than maternal half-sibs?

2654 *Parent-midpoint offspring regression* Another way that we can esti-  
 mate the narrow sense heritability is through the regression of child's  
 2656 phenotype on the parental mid-point phenotype. The parental mid-  
 point phenotype is simply the average of the mum and dad's pheno-  
 2658 type. We denote the child's phenotype by  $X_{kid}$  and mid-point phe-  
 notype by  $X_{mid}$ , so that if we take the regression  $X_{kid} \sim X_{mid}$  this  
 2660 regression has slope  $\beta = \text{Cov}(X_{kid}, X_{mid})/\text{Var}(X_{mid})$ . The covari-  
 ance 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  
 2662 slope of the regression is

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

2664 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 environ-  
 2666 mental correlations between relatives. But it's a useful way to think  
 about heritability and will be directly relevant to our discussion of the  
 2668 response to selection in the next chapter.

2670 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  
 2672 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  
 2674 guess at the child's phenotype.

2676 More formally, the expected phenotype of the child given the  
 parental phenotypes is

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

2678 which follows from the definition of linear regression. So to find the  
 child's predicted phenotype, we simply take the mean phenotype and  
 add on the difference between our parental mid-point and the popula-  
 2680 tion mean, multiplied by our narrow sense heritability.

2682 **Question 3.** Briefly explain what Galton meant by 'regression  
 towards mediocrity', and why he observed this pattern in light of  
 Mendelian inheritance.



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.

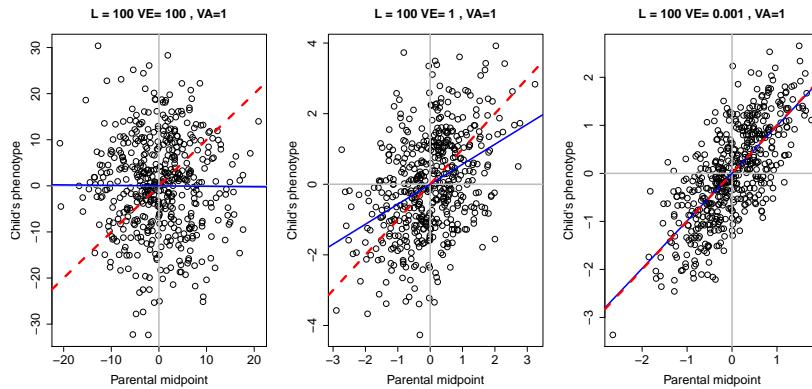


Figure 4.9: Regression of child's phenotype of the parental mid-point phenotype. The three panels show decreasing levels of environmental variance ( $V_E$ ) holding the additive genetic variance constant ( $V_A = 1$ ). In these figures, we simulate 100 loci, as described in the caption of Figure 4.5. 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.

2684 *Estimating additive genetic variance across a variety of different relationships.* In many natural populations we may have access to  
 2686 individuals with a range of different relationships to each other (e.g. through monitoring of the paternity of individuals), but relatively few  
 2688 pairs of individuals for a specific relationship (e.g. sibs). We can try and use this information on various relatives as fully as possible in a  
 2690 mixed model framework. Building from equation 4.3, we can write an individual's phenotype  $X_i$  as

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

2692 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  
 2694 of individuals i and j are  $A_{ij} = 2F_{i,j}$  and  $A_{ii} = 1$ . Given the matrix  $A$  we can estimate  $V_A$ . We can also add fixed effects into this model  
 2696 to account for generation effects, additional mixed effects could also be included to account for shared environments between particular  
 2698 individuals (e.g. a shared nest). This approach is sometimes called the “animal model”.

#### 2700 4.1 Multiple traits

Traits often covary with each other, both due to environmentally induced effects (e.g. due to the effects of diet on multiple traits) and due to the expression of underlying genetic covariance between traits.  
 2702 Genetic covariance, in turn, can reflect pleiotropy, a mechanistic effect of an allele on multiple traits (e.g. variants that affect skin pigmentation often affect hair color), the genetic linkage of loci independently  
 2704 affecting multiple traits, or the effects of assortative mating.  
 2706

2708 Consider two traits  $X_{1,i}$  and  $X_{2,i}$  in an individual  $i$ . These traits  
 could be, say, the individual's leg length and nose length. As before,  
 2710 we can write these as

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

As before we can talk about the total phenotypic variance ( $V_1, V_2$ ),  
 2712 environmental variance ( $V_{1,E}$  and  $V_{2,E}$ ), and the additive genetic  
 variance for trait one and two ( $V_{1,A}, V_{2,A}$ ). But now we also have  
 2714 to consider the total covariance between trait one and trait two,  
 $V_{1,2} = \text{Cov}(X_1, X_2)$ , as well as the environmentally induced covariance  
 2716 ( $V_{E,1,2} = \text{Cov}(X_{1,E}, X_{2,E})$ ) and the additive genetic covariance  
 $(V_{A,1,2} = \text{Cov}(X_{1,A}, X_{2,A}))$ . To better understand the covariance arising  
 2718 due to pleiotropy, let's think about a set of  $L$  SNPs contributing  
 to our two traits. If the additive effect of an allele at the  $i^{th}$  SNP is  
 2720  $\alpha_{i,1}$  and  $\alpha_{i,2}$  on traits 1 and 2, then the additive covariance between  
 our traits is

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

2722 assuming our loci are in linkage disequilibrium. Thus a genetic correlation arises due to pleiotropy, because loci that tend to affect trait 1  
 2724 also systematically affect trait 2. For example, alleles associated with later Age at Menarche (AAM) in European females also tend to be  
 2726 positively associated with height (see Figure 4.10), thereby creating a genetic correlation between AAM and height.

2728 We can store our variance and covariance values in matrices, a way of gathering these terms that will be useful when we discuss selection:

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

2730 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  
 2732 this to an arbitrary number of traits.

We can estimate these quantities, in a similar way as before, by  
 2734 studying the covariance in different traits between relatives:

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

We can also talk about the genetic correlation between two phenotypes  
 2736

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

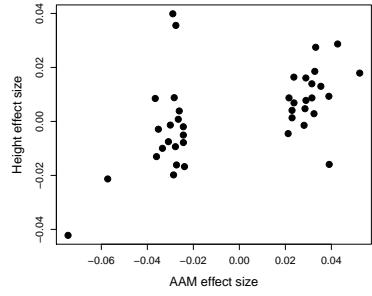


Figure 4.10: The additive effect sizes of loci associated with female Age at Menarche (AAM) and their effect size on Height in a European population. Data from PICKRELL *et al.* (2016). Code here.

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.

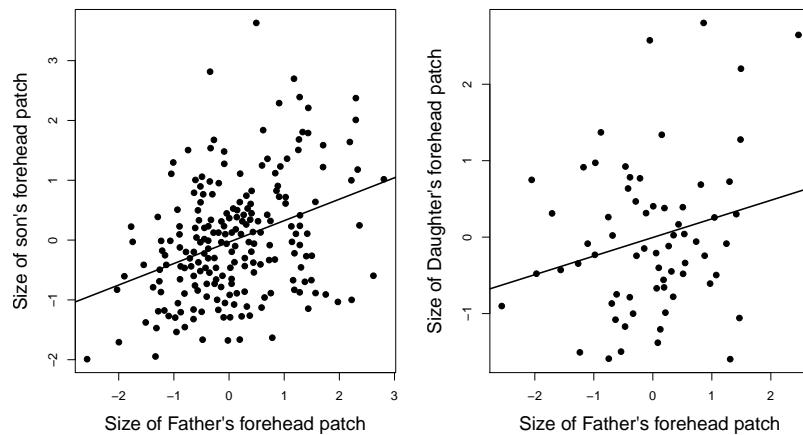


Figure 4.11: Relationship of standardized forehead patch size between fathers and sons and daughters in Pied fly-catchers. Data from POTTI and CANAL. Code here.



Figure 4.12: *Ficedula hypoleuca*, Pied fly-catcher.  
Coloured illustrations of British birds, and their eggs (1842-1850). London :G.W. Nickisson. 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.

#### 2754 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 locus (*dominance*) or among alleles at different loci (*epistasis*). We can accommodate these complications into our models. We do this by partitioning our total genetic variance into independent variance components.

2762 *Dominance.* To understand the effect of dominance, let's consider  
2764 how the allele that a parent transmits influences their offspring's phe-  
2766 notype. A parent transmits one of their two alleles at a locus to their  
2768 offspring. Assuming that individuals mate at random, this allele is  
2770 paired with another allele drawn at random from the population. For  
2772 example, assume your mother transmitted an allele 1 to you: with  
2774 probability  $p$  it would be paired with another allele 1, and you would  
2776 be a homozygote; and with probability  $q$  it's paired with a 2 allele and  
2778 you're a heterozygote.

2772 Now consider an autosomal biallelic locus  $\ell$ , with frequency  $p$  for  
2774 allele 1, and genotypes 0, 1, and 2 corresponding to how many copies  
2776 of allele 1 individuals carry. We'll denote the mean phenotype of an  
2778 individual with genotype 0, 1, and 2 as  $\bar{X}_{\ell,0}$ ,  $\bar{X}_{\ell,1}$ ,  $\bar{X}_{\ell,2}$  respectively.  
2780 This mean is taking an average phenotype over all the environments  
2782 and genetic backgrounds the alleles are present on. We'll mean center  
2784 (MC) these phenotypic values, setting  $\bar{X}'_{\ell,0} = \bar{X}_{\ell,0} - \mu$ , and likewise  
2786 for the other genotypes.

2780 We can think about the average (marginal) MC phenotype of an  
2782 individual who received an allele 1 from their parent as the average  
2784 of the MC phenotype for heterozygotes and 11 homozygotes, weighted  
2786 by the probability that the individual has these genotypes, i.e. the  
2788 probability they receive an additional allele 1 or an allele 2 from their  
2790 other parent:

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

2792 Similarly, if your parent transmitted an 2 allele to you, your average  
2794 MC phenotype would be

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

2796 Let's now consider the average phenotype of an offspring of each of  
2798 our three genotypes

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

2796 i.e. the mean phenotype of each genotypes' offspring averaged over  
2798 all possible matings to other individuals in the population (assuming  
2800 individuals mate at random). These are the additive MC genetic  
2802 values (breeding values) of our genotypes. Here we are simply adding  
2804 up the additive contributions of the alleles present in each genotype  
2806 and ignoring any non-additive effects of genotype.

2808 To illustrate this, in Figure 4.13 we plot two different cases of dom-  
2810 inance relationships; in the top row an additive polymorphism and in  
2812 the second row a fully dominant allele. The additive genetic values of  
2814 the genotypes are shown as red dots. Note that the additive values of  
2816 the genotypes line up with the observed MC phenotypic means in the



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.

top row, when our alleles interact in a completely additive manner.

2802 Our additive genetic values always fall along a linear line (the red line in our figure). The additive values are falling along the best fitting line  
2804 of linear regression for our population, when phenotype is regressed  
2806 against the additive genotype (0, 1, 2 copies of allele 1) across all in-  
2808 dividuals in our population. Note in the dominant case the additive  
genetic values differ from the observed phenotypic means, and are  
closer to the observed values for the genotypes that are most common  
in the population.

2810 The difference in the additive effect of the two alleles  $a_{\ell,2} - a_{\ell,1}$   
2812 can be interpreted as an average effect of swapping an allele 1 for an  
allele 2; we'll call this difference  $\alpha_{\ell} = a_{\ell,2} - a_{\ell,1}$ . Our  $\alpha_{\ell}$  is also the  
2814 slope of the regression of phenotype against genotype (the red line  
on genotype ( $\alpha_{\ell}$ ) does not depend on the population allele frequency  
2816 for our completely additive locus (top row of 4.13). In contrast, when  
there is dominance, the slope between genotype and phenotype ( $\alpha_{\ell}$ )  
2818 is a function of allele frequency (bottom row of 4.13). When a domi-  
nant allele (1) is rare there is a strong slope of phenotype on genotype,  
2820 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  
2822 average phenotype, as it will most likely move an individual from be-  
ing a 22 homozygote to being a 12 heterozygote. In contrast, when the  
2824 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,

2826 leading to a weak slope bottom right Figure 4.13. This small effect is  
 because as we are mainly turning heterozygotes into homozygotes (11),  
 2828 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 (AYLLON *et al.*, 2015; BARSON *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 (COUS-MINER *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).



The variance in the population phenotype due to these additive breeding values at locus  $\ell$ , assuming HW proportions, is

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

2848 The total additive variance for the whole genotype can be found by

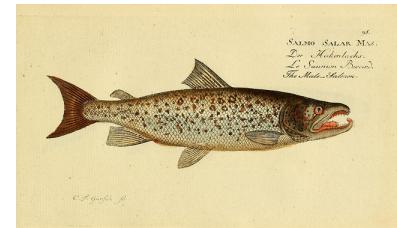


Figure 4.14: Atlantic Salmon (*Salmo salar*).

*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 BARSON *et al.* (2015). Code here.

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

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

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 dominance variance at our locus as the genotype-frequency weighted sum of our squared dominance deviations

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

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 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 additive and dominant deviations for the genotypes is zero. We can define the narrow sense heritability as before  $h^2 = V_A/V_P = V_A/(V_G + V_E)$ , which is the proportion of phenotypic variance due to additive genetic variance. We can also define the total proportion of the phenotypic variance due to genetic differences among individuals, as the broad-sense heritability  $H^2 = V_G/(V_G + V_E)$ .

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 <sup>st</sup> 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).

When dominance is present in the loci influencing our trait ( $V_D >$

0), we need to modify our phenotype covariance among relatives to

account for this non-additivity. Specifically, our equation for the  
 2874 covariance among a general pair of relatives (eqn. 4.12 for additive variation) becomes

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

2876 where  $r_2$  is the probability that the pair of individuals share 2 alleles identical by descent, making the same assumptions (other than  
 2878 additivity) that we made in deriving eqn. 4.12. In table 4.1 we show the phenotypic covariance for some common pairs of relatives. The  
 2880 regression of offspring phenotype on parental midpoint still has a slope  $V_A/V_P$ .

2882 Full sibs and parent-offspring have the same covariance if there  
 2884 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-  
 2886 semble each other more than parent-offspring pairs. That's because parents and offspring share precisely one allele, while full-sibs can  
 2888 share both alleles (i.e. the full genotype at a locus) identical by de-  
 2890 scent. We can attempt to estimate  $V_D$  by comparing different sets of relationships. For example, non-identical twins (full sibs born at same  
 2892 time) should have 1/2 the phenotypic covariance of identical twins if  $V_D = 0$ . Therefore, we can attempt to estimate  $V_D$  by looking at whether identical twins have more than twice the phenotypic covari-  
 2894 ance than non-identical twins.

2894 The most important aspect of this discussion for thinking about evolutionary genetics is that the parent-offspring covariance is still  
 2896 only a function of  $V_A$ . This is because our parent (e.g. the mother) transmits only a single allele, at each locus, to its offspring. The other  
 2898 allele the offspring receives is random (assuming random mating), as it comes from the other unrelated parent (the father). Therefore, the  
 2900 average effect on the child's phenotype of an allele the child receives from their mother is averaged over all possible random alleles the child  
 2902 could receive from their father (weighted by their frequency in the population). Thus we only care about the additive effect of the allele,  
 2904 as parents transmit only alleles (not genotypes) to their offspring.  
 This means that the short-term response to selection, as described by  
 2906 the breeder's equation, depends only on  $V_A$  and the additive effect  
 2908 of alleles. Therefore, if we can estimate the narrow-sense heritability  
 2910 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  
 2912 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 ex-

ample among pairs of loci). This gets a little tricky to think about,  
2916 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  
2918 over their possible genetic backgrounds (including the other interacting  
alleles they are possibly paired with), just as before. We can then  
2920 calculate the additive genetic variance from this. We can estimate the  
dominance variance, by calculating the residual variance among geno-  
2922 types at a locus unexplained by the additive effect of the loci. We can  
then estimate the epistatic variance by estimating the residual vari-  
2924 ance left unexplained among the two locus genotypes after accounting  
for the additive and dominant deviations calculated from each locus  
2926 separately. In practice these high variance components are hard to  
estimate, and usually small as much of our variance is assigned to the  
2928 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  
2930 the additive genetic variance will depend on the epistatic relationships  
among loci.

2932     **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.  
2934 non-identical twins better suited for this?

2936     **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).



## The Response to Phenotypic Selection

- 2940 Evolution by natural selection requires:
1. Variation in a phenotype

2942 2. That survival is non-random with respect to this phenotypic variation.

2944 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.

<sup>1</sup> It is the heritable nature of variation that couples change within a generation due to natural selection to change across generations (evolutionary change).

2950 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  $\mu_S$ , i.e. after selection has acted, and our mean phenotype before selection acts by  $\mu_{BS}$ . 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  $\mu_{NG}$ . The change across generations we'll call the response to selection  $R$  and put this equal to  $\mu_{NG} - \mu_{BS}$ .

2968 The mean phenotype in the next generation is

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

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

<sup>1</sup> 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.

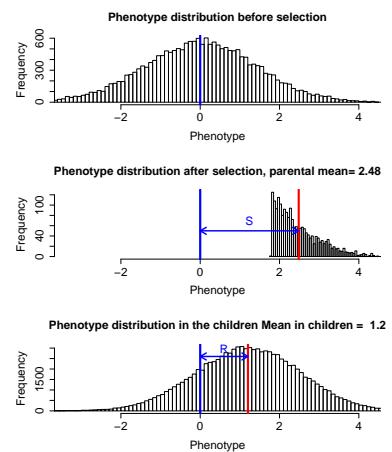


Figure 5.1: **Top.** Distribution of a phenotype in the parental population prior to selection,  $V_A = V_E = 1$ . **Middle.** Only individuals in the top 10% of the phenotypic distribution are selected to reproduce; the resulting shift in the phenotypic mean is  $S$ . **Bottom.** Phenotypic distribution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

where the outer expectation is over possible pairs of randomly mating 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  $\mu_{NG}$  we need to compute  $\mathbb{E}(X_{mid})$ , the expected mid-point phenotype of pairs of individuals who survive to reproduce. Well this is just the expected phenotype in the individuals who survived to reproduce ( $\mu_S$ ), so

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

So we can write our response to selection as

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

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. 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 ( $h^2$ ).



Figure 5.2: A visual representation of the Breeder's equation. Regression of child's phenotype on parental midpoint phenotype ( $V_A = V_E = 1$ ). Under truncation selection, only individuals with phenotypes  $> 1$  (red) are bred. [Code here.](#)

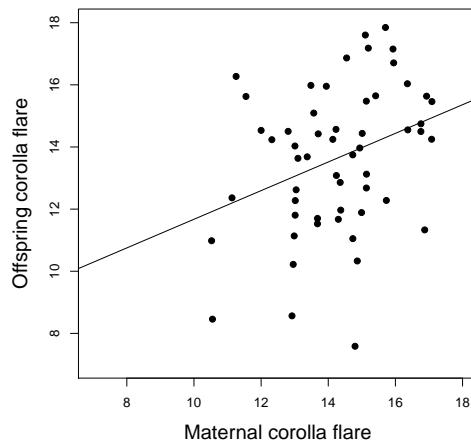


Figure 5.3: The relationship between maternal and offspring corolla flare (flower width) in *P. viscosum*. From GALEN's data the covariance of mother and child is 1.3, while the variance of the mother is 2.8. Data from GALEN (1996). [Code here.](#)



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.** GALEN (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?

2990 To understand the genetic basis of the response to selection take a look at Figure 5.5. The setup is the same as in our previous simulation figures. The individuals who are selected to form our next

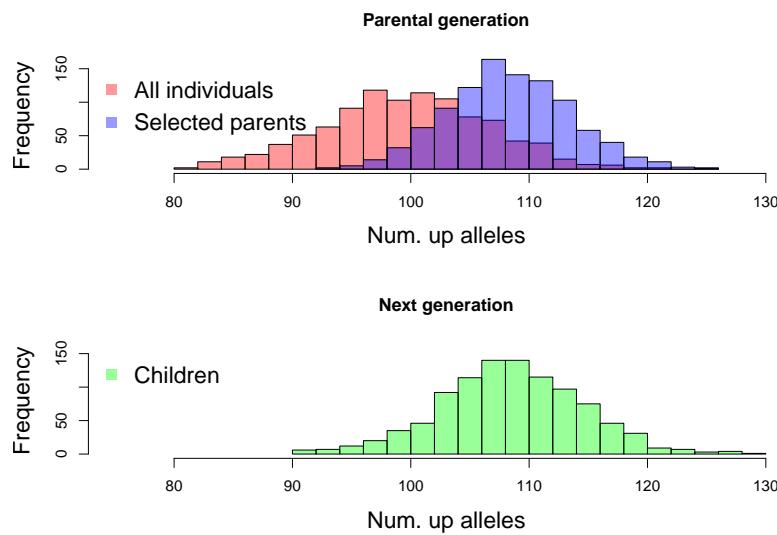


Figure 5.5: **Top.** Distribution of the number of up alleles in the parental population prior to selection (red), for the selected individuals the top 10% of the population (blue) **Bottom.** The same distribution for the offspring of the selected parents in the next generation (green). Code here.

2992 generation carry more alleles that increase the phenotype, in the current range of environments currently experienced by the population.  
 2994 The average individual before selection carried 100 of these 'up' alleles, the average individual surviving selection 108 'up' alleles. As  
 2996 individuals faithfully transmit their alleles to the next generation the  
 2998 average child of the selected parents carries 108 up alleles. Note that  
 3000 the variance has changed little, the children have plenty of variation in  
 3002 their genotype, such that selection can readily drive evolution in future  
 3004 generations. The average frequency of an 'up' allele has changed from  
 3006 50% to 54%. Our gains due to selection will be stably inherited to  
 3008 future generations.

3004 *The long-term response to selection* If our selection pressure is sustained over many generations, we can use our breeder's equation to  
 3006 predict the response. If we are willing to assume that our heritability does not change and we maintain a constant selection gradient, then  
 3008 after  $n$  generations our phenotype mean will have shifted

$$nh^2S \quad (5.5)$$

i.e. our population will keep up a linear response to selection.

3010 **Question 2.** A population of red deer were trapped on Jersey (an island off of England) during the last inter-glacial period. From the

<sup>3012</sup> fossil record <sup>2</sup> we can see that the population rapidly adapted to their new conditions. Within 6,000 years they evolved from an estimated  
<sup>3014</sup> mean weight of the population of 200kg to an estimated mean weight of 36kg (a 6 fold reduction)! You estimate that the generation time  
<sup>3016</sup> of red deer is 5 years and, from a current day population, that the narrow sense heritability of the phenotype is 0.5.

<sup>3018</sup> **A)** Estimate the mean change per generation in the mean body weight.

<sup>3020</sup> **B)** Estimate the change in mean body weight caused by selection within a generation. State your assumptions.

<sup>3022</sup> **C)** Assuming we only have fossils from the founding population and the population after 6000 years, should we assume that the calculations accurately reflect what actually occurred within our population?

<sup>3026</sup> *Alternative formulations of the Breeder's equation.* A change in mean phenotype within a generation occurs because of the differential fitness 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

$$\mu_{BS} = \mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx \quad (5.6)$$

<sup>3034</sup> 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  
<sup>3036</sup> organism. The probability distribution of phenotypes in those who do survive to reproduce is

$$\mathbb{P}(X|\text{survive}) = \frac{p(x)w(x)}{\int_{-\infty}^{\infty} p(x)w(x)dx}. \quad (5.7)$$

<sup>3038</sup> where the denominator is a normalization constant which ensures that our phenotypic distribution integrates to one. The denominator also  
<sup>3040</sup> has the interpretation of being the mean fitness of the population, which we'll call  $\bar{w}$ , i.e.

$$\bar{w} = \int_{-\infty}^{\infty} p(x)w(x)dx. \quad (5.8)$$

<sup>3042</sup> Therefore, we can write the mean phenotype in those who survive to reproduce as

$$\mu_S = \frac{1}{\bar{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx \quad (5.9)$$

<sup>2</sup> LISTER, A., 1989 Rapid dwarfing of red deer on Jersey in the last interglacial. *Nature* 342(6249): 539

3044 If we mean center our population, i.e. set the phenotype before selection to zero, then

$$S = \frac{1}{\bar{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx \quad (5.10)$$

- 3046 Inspecting this more closely, we can see that  $S$  has the form of a covariance between our phenotype  $X$  and our relative fitness  $w(X)$  ( $S = Cov(X, w(X)/\bar{w})$ ). Thus our change in mean phenotype is directly a measure of the covariance of our phenotype and our fitness.
- 3048
- 3050 Rewriting our breeder's equation using this observation we see

$$R = \frac{V_A}{V} Cov(X, w(X)/\bar{w}) \quad (5.11)$$

3052 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.

3054 The phenotype-fitness covariance divided by the phenotypic variance,  $Cov(X, w(X)/\bar{w})/V$ , is the slope of the linear regression of phenotype on fitness. Let's call this slope the fitness gradient and denote it by  $\beta$ . Then, equivalently, we can write the breeder's equation as

$$R = V_A \beta \quad (5.12)$$

3058 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.

3062 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 ( $\beta$ ).

### 3066 5.0.1 The response of multiple traits to selection, the multivariate breeder's equation.

3068 We can generalize these results for multiple traits, to ask how selection on multiple phenotypes plays out over short time intervals.<sup>3</sup> Considering two traits we can write our responses in both traits as

$$\begin{aligned} R_1 &= V_{A,1}\beta_1 + V_{A,1,2}\beta_2 \\ R_2 &= V_{A,2}\beta_2 + V_{A,1,2}\beta_1 \end{aligned} \quad (5.13)$$

3072 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,  $\beta_1$ , represents the change in fitness changing trait 1 alone holding



Figure 5.6: 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.

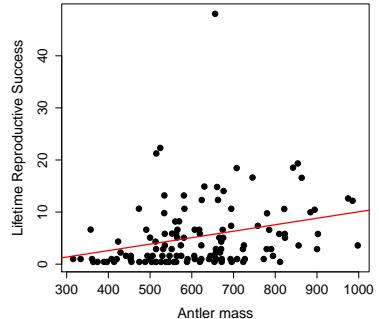


Figure 5.7: Lifetime reproductive success (LRS) of male Red Deer as a function of their antler mass. Data from KRUUK *et al.* (2002), see the paper for discussion of the complexities of equating this selection gradient with the evolutionary response. Code here..

<sup>3</sup> LANDE, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution 33(1Part2): 402–416

<sup>3074</sup> everything else constant. This is a statement that our response in any  
<sup>3075</sup> one phenotype is modified by selection on other traits that covary with  
<sup>3076</sup> that trait. This offers a good way to think about how genetic trade  
<sup>3077</sup> offs play out over short-term evolution.

<sup>3078</sup> We can also write this in matrix form. We can write our change in  
<sup>3079</sup> the mean of our multiple phenotypes within a generation as the vector  
<sup>3080</sup>  $\mathbf{S}$  and our response across multiple generations as the vector  $\mathbf{R}$ . These  
<sup>3081</sup> two quantities are related by

$$\mathbf{R} = \mathbf{G}\mathbf{V}^{-1}\mathbf{S} = \mathbf{G}\boldsymbol{\beta} \quad (5.14)$$

<sup>3082</sup> where  $\mathbf{V}$  and  $\mathbf{G}$  are our matrices of the variance-covariance of pheno-  
<sup>3083</sup> types and additive genetic values (eqn. (4.19) (4.18)) and  $\boldsymbol{\beta}$  is a vector  
<sup>3084</sup> of selection gradients (i.e. the change within a generation as a fraction  
<sup>3085</sup> of the total phenotypic variance).

<sup>3086</sup> **Question 3.** You collect observations of red deer within a genera-  
<sup>3087</sup> tion, recording an individual's number of offspring and phenotypes for  
<sup>3088</sup> a number of traits which are known to have additive genetic variation.  
<sup>3089</sup> Using your data, you construct the plots shown in Figure 5.10 (stan-  
<sup>3090</sup> dardizing the phenotypes). Answer the following questions by choosing  
<sup>3091</sup> one of the bold options. Briefly justify each of your answers with refer-  
<sup>3092</sup> ence to the breeder's equation and multi-trait breeder's equation.

<sup>3093</sup> **A)** Looking just at figure 5.10 A, in what direction do you expect  
<sup>3094</sup> male antler size to evolve?

**Insufficient information, increase, decrease.**

<sup>3095</sup> **B)** Looking just at figures 5.10 B and C, in what direction do you  
<sup>3096</sup> expect male antler size to evolve?

**Insufficient information, increase, decrease.**

<sup>3097</sup> **C)** Looking at figures 5.10 A, B, and C, in what direction do you  
<sup>3098</sup> expect male antler size to evolve?

**Insufficient information, increase, decrease.**

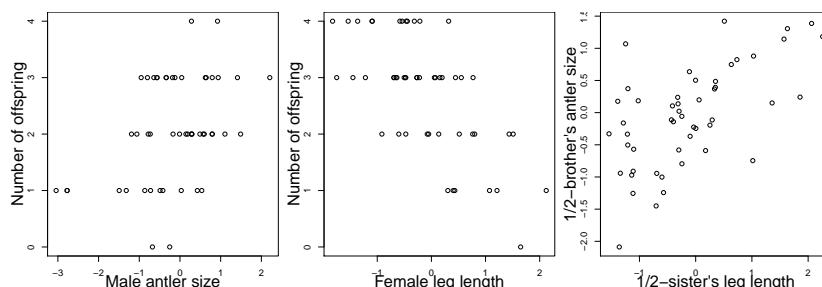


Figure 5.8: Observations of red deer within a generation; recording an individual's number of offspring and phenotypes (simulated data), which are known to have additive genetic variation. The figures left to right are A-C. (Data are simulated. Code here.)

<sup>3102</sup> As an example of correlated responses to selection, consider the  
<sup>3103</sup> WILKINSON (1993) selection experiment on Stalk-eyed flies (*Cyr-*  
<sup>3104</sup> *todiopsis dalmanni*). Stalk-eyed flies have evolved amazingly long

eye-stalks. In the lab, WILKINSON established six populations of wild-caught flies and selected up and down on males eye-stalk to body size ratio for 10 generations (left plot in Figure 5.11). Despite the fact that he did not select on females, he saw a correlated response in the females from each of the lines (right plot), because of the genetic correlation between male and female body proportions.

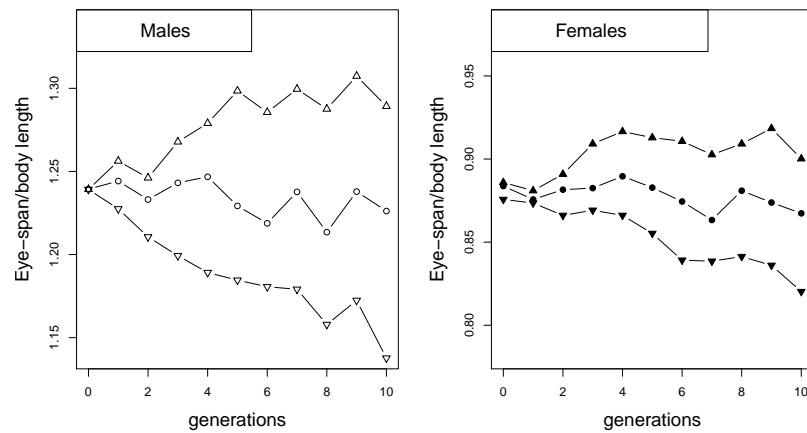


Figure 5.9: WILKINSON selected two of populations for flies for increased and eye-stalk to body length ratio in males (mean shown as up triangles), and two for a decreased ratio (down triangles), by taking the top 10 males with the highest (lowest) ratio out of 50 measures. He also established two control populations (circles). He constructed each generation of females by sampling 10 at random from each population. Data from WILKINSON (1993). Code here.

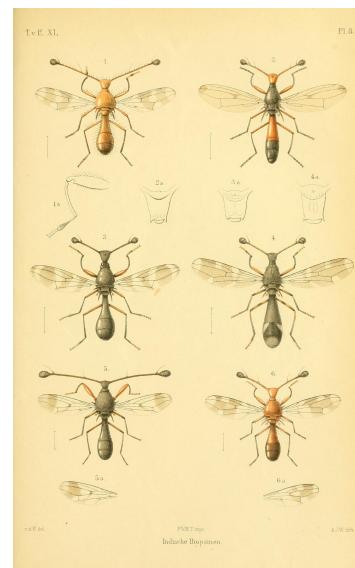


Figure 5.10: Stalk-eyed Flies (*Diopsidae*).  
Diptera. van der Wulp. 1898. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

#### Question 4.

At the end of ten generations in WILKINSON's experiment (Figure 5.11), the males from the up- and down-selected 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) WILKINSON estimated that by selecting the top/bottom 10 males, he had on average shifted the mean body 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 phenotypes are equal and that there is no direct selection on female 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?

### 5.1 Some applications of the multivariate trait breeder's equation

The multivariate breeders equation has a lot of different uses in understanding the response of multiple traits to selection. It also offers some insights into kin selection and sexual selection. We'll discuss these

3130 next.

*Hamilton's Rule and the evolution of altruistic and selfish behaviours*

3132 Individuals frequently behave in ways that sacrifice their own fitness  
 for the benefit of others. That selection favours such apparent acts of  
 3134 altruism is puzzling at first sight. HAMILTON (1964a,b) supplied the  
 first general evolutionary explanation of such altruism. His intuition  
 3136 was that while an individual is losing out of some reproductive output,  
 the alleles underlying an altruistic behaviour can still spread in the  
 3138 population if this cost is outweighed by benefits gained through the  
 transmission of these alleles through a related individual. Note that  
 3140 this means that the allele is not acting in an self-sacrificing manner,  
 even though individuals may as a result.

3142 Altruism reflects social interactions. So as a simple model let's  
 imagine that individuals interact in pairs, with our focal individual  
 3144  $i$  being paired with an individual  $j$ . This could be pairs of siblings  
 interacting. Imagine that individuals have two possible phenotypes  
 3146  $X = 1$  or  $0$ , corresponding to providing or withholding some small  
 act of 'altruism' (we could just as easily flip these labels and call them  
 3148 an unselfish act and a selfish act respectively). Our pairs of individ-  
 uals interacting could, for example, be siblings sharing a nest. The  
 3150 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  
 3152 complicated acts of altruism such as children foregoing their own re-  
 production so as to help their parents raise their siblings.

3154 Providing the altruistic act has a cost  $C$  to the fitness of our in-  
 dividual and failing to provide this act has no cost. Receiving this  
 3156 altruistic act confers a fitness benefit  $B$  over individuals who did not  
 receive this act. HAMILTON's rule states that such a trait will spread  
 3158 through the population if

$$2FB > C \quad (5.15)$$

where  $F$  is the average kinship coefficient between the interacting  
 3160 individuals ( $i$  and  $j$ ). In the usual formulation of Hamilton's Rule  
 our  $2F$  is replaced by the 'Coefficient of relationship', which is the  
 3162 proportion of alleles shared between the individuals. Here we use two  
 times the kinship coefficient to keep things inline with our notation for  
 3164 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  
 3166 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.

3168 As one example of kin-selection consider KRAKAUER (2005)'s  
 work on co-operative courtship in wild turkeys (*Meleagris gallopavo*).  
 3170 Male turkeys often form display partnerships, with a subordinate male

MAYNARD SMITH (1964) coined the name kin selection to describe 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.

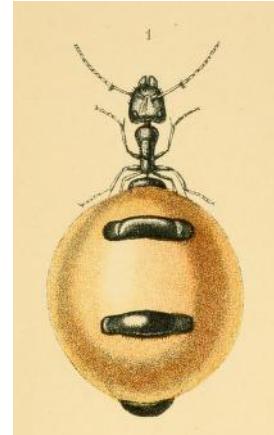


Figure 5.11: Australian Honey-pot Ant (*Camponotus inflatus*). Honey ants are gorged with honeydew collected by their nest mates, till they swell to the size of grapes, and used as a food storage device.

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.

helping a dominant male with displaying to females and defending the  
 3172 females from other groups of males.

These pairs are often full brothers ( $F = 0.25$ ), with the subordinate  
 3174 male often being the younger of the two. The subordinate male  
 often loses out on mating opportunities over their entire lifetime by  
 3176 acting as a wingman to their older brothers. KRAKAUER (2005)  
 estimated that dominant males gained an extra 6.1 offspring when  
 3178 they display with a partner than males who display alone. While  
 the subordinate males lose out on fathering 0.9 offspring compared  
 3180 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  
 3182 in co-operative courtship is potentially well explained by kin-selection  
 3184 (see AKÇAY and VAN CLEVE, 2016, for more analysis).

**Question 5.** How would this answer be changed if the male  
 3186 Turkey partnerships were only  $1/2$  sibs, or first cousins?

Where does this result come from? Well, we can use our quantitative  
 3188 genetics framework to gain some intuition by deriving a simple  
 version of Hamilton's Rule by thinking about the phenotypes of an  
 3190 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  
 3192 written as

$$W(i, j) = W_0 + W_i + W_j \quad (5.16)$$

where  $W_i$  is the contribution of the fitness of the individual  $i$  due to  
 3194 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  
 3196 Figure 5.15.

Following our multivariate breeder's equation, we can write the  
 3198 expected change of our behavioural phenotype as

$$R = \beta_i V_A + \beta_j V_{A,i,j}, \quad (5.17)$$

3200 Our altruistic phenotype is increasing in the population if  $R > 0$ , i.e. if

$$\beta_i V_A + \beta_j V_{A,i,j} > 0 \quad (5.18)$$

3202 The slope  $\beta_i$  of the regression of our focal individual's behavioural  
 phenotype on fitness is proportional to  $-C$ . The slope  $\beta_j$  of the re-  
 3204 gression of our interacting partner's phenotype on our focal individ-  
 ual's fitness is proportional to  $B$  (with the same constant of propor-  
 3206 tionality). Therefore, our altruistic phenotype is increasing in the



Figure 5.12: Turkey (*Meleagris gallopavo*).  
 Bilder-atlas zur Wissenschaftlich-populären  
 Naturgeschichte der Vögel in ihren sämtlichen  
 Hauptformen (1864). Wien,K.K. Hof Image  
 from the Biodiversity Heritage Library.  
 Contributed by Smithsonian Libraries. Not in  
 copyright.

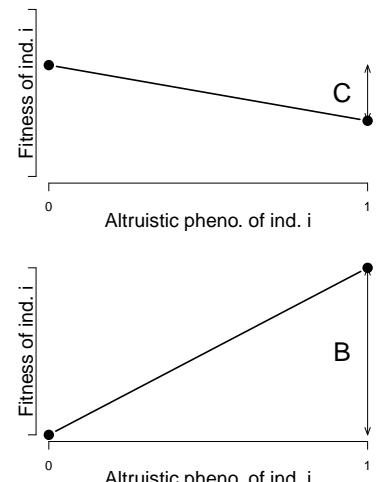


Figure 5.13: **Top**) The fitness of individual  $i$  as a function of their behavioural phenotype, where altruistic/non-altruistic behavioural phenotypes are encoded as 1 and 0 respectively. The direct fitness cost of behaving altruistically is  $C$ . **Bottom**) The fitness of our focal individual  $i$  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've following a simplified version of QUELLER (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).

population if

$$\begin{aligned} \beta_i V_A + \beta_j V_{A,i,j} &> 0 \\ B \frac{V_{A,i,j}}{V_A} &> C \end{aligned} \quad (5.19)$$

- 3208 So what's the average genetic covariance between individual  $i$  and  
 $j$ 's altruistic phenotype? Well it's the same behavioural phenotype  
 3210 in both individuals, so the phenotypes are genetically correlated if  
 our individuals are related to each other. The covariance of the same  
 3212 phenotype between two individuals is just  $2F_{i,j}V_A$  (see (4.12)). So our  
 altruistic phenotype is increasing in the population if

$$\begin{aligned} B \frac{2F_{i,j}V_A}{V_A} &> C \\ 2F_{i,j}B &> C \end{aligned} \quad (5.20)$$

- 3214 Seen from this perspective, HAMILTON's rule is simply a statement  
 that altruistic behaviours can spread via kin-selection, if the average  
 3216 cost to an individual of carrying altruistic alleles is paid back through  
 the average benefit of interacting with altruistic relatives (kin)

- 3218 *Sexual selection and the evolution of mate preference by indirect benefits.* Organisms often put an enormous effort into finding and at-  
 3220 tracting mates, sometimes at a considerable cost to their chances of survival. Why are individuals so choosy about who they mate with,  
 3222 particularly when their choice seems to be based on elaborate characters and arbitrary displays that surely lower the viability of their  
 3224 mates?

One major reason why individuals evolve to be choosy about who  
 3226 they mate with is that it can directly impact their fitness. By choosing a mate with particular characteristics, individuals can gain more  
 3228 parental care for their offspring, avoid parasites, or be choosing a mate with higher fertility. For example, female glow-worms flash at night  
 3230 to attract males flying by. Females with larger, brighter lanterns have higher fecundity, so males with a preference for brighter flashes will  
 3232 gain a direct benefit to their own fitness. (Note that males will benefit even if these differences in female fecundity are entirely driven by  
 3234 differences in environment, and so non-heritable.) Indeed male glow worms have evolved to be attracted to brighter flashing lures.

3236 However, even in the absence of direct benefits of choice, selection can still indirectly favour the evolution of choosiness. These indirect  
 3238 benefits occur because individuals can have higher fitness offspring by choosing a mate whose phenotype indicates high viability (the so-  
 3240 called good genes hypothesis), or by choosing a mate whose phenotype is simply attractive, and likely to produce similarly attractive offspring  
 3242 (the 'runaway' or sexy sons hypothesis).

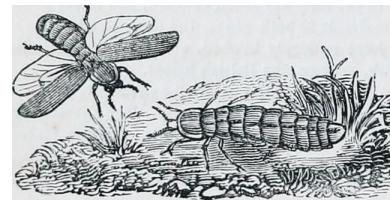


Figure 5.14: Male (left) and female (right) common glow worm (*Lampyris noctiluca*).  
 The animal kingdom : arranged after its organization; forming a natural history of animals, and an introduction to comparative anatomy. (1863) Cuvier, G. Image from the Biodiversity Heritage Library. Contributed by University of Toronto - Gerstein Science Information Centre. Not in copyright.

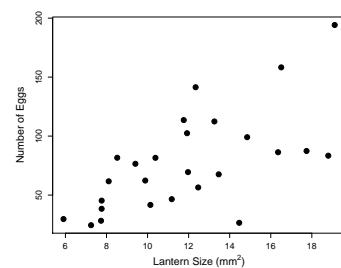


Figure 5.15: Female Glow worms who have the largest, and therefore brightest, lanterns have the highest fecundity. Data from HOPKINS *et al.* (2015). Code here.

We'll denote a display trait, e.g. tail length, in males by  $\sigma$  and  
 3244 a preference trait in females by  $\varphi$ . Our display trait is under direct  
 selection in males, such that its response to selection can be written as

3246

$$R_\sigma = \beta_\sigma V_{A,\sigma} \quad (5.21)$$

Let's assume that the female preference trait, the degree to which  
 3248 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 as

$$R_\varphi = \beta_\varphi V_{A,\varphi} + \beta_\sigma V_{A,\varphi\sigma} = \beta_\sigma V_{A,\varphi\sigma} \quad (5.22)$$

3250 So the female preference will respond to selection if it is genetically  
 correlated with the male trait, i.e. if  $V_{A,\varphi\sigma}$  is not zero. There's a  
 3252 number of different ways this genetic correlation could arise; the sim-  
 plest is that the loci underlying the male trait may have a pleiotropic  
 3254 effect on female preference. However, female preference may often  
 have quite a distinct genetic basis from male display traits.

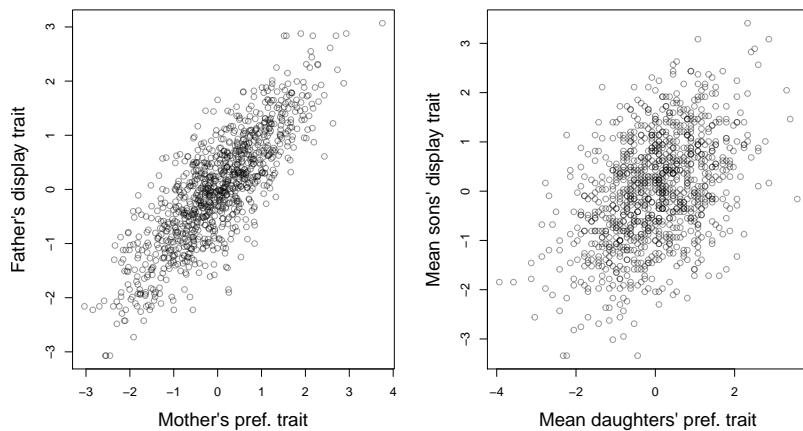


Figure 5.16: **Left**) Assortative mating between males and females. Males vary in a display trait (e.g. tail length), females vary in their preference for this trait. We see evidence of assortative mating as females with a preference for a particular value of the male trait tend to mate with those males. **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.

3256 A more general way in which trait-preference genetic correlations  
 may arise is through assortative mating. As females vary in their tail-  
 3258 length preference, the ones with a preference for longer tails will mate  
 with long-tailed males and the opposite for females with a preference  
 3260 for shorter-tails. Therefore, a genetic correlation between mates dis-  
 play and preference traits will become established (see Figure 5.18).

3262 The males with the longer tails will also carry the alleles associated  
 with the preference for longer tails, as their long-tailed dads tended  
 3264 to mate with females with a genetic preference for long tails. Simi-  
 larly, the the males with shorter tails will carry alleles associated with  
 3266 the preference for shorter tails. Thus if there is direct selection for

males with longer tails, then the female preference for longer tails will  
 3268 increase too, as it is genetically correlated via assortative mating.

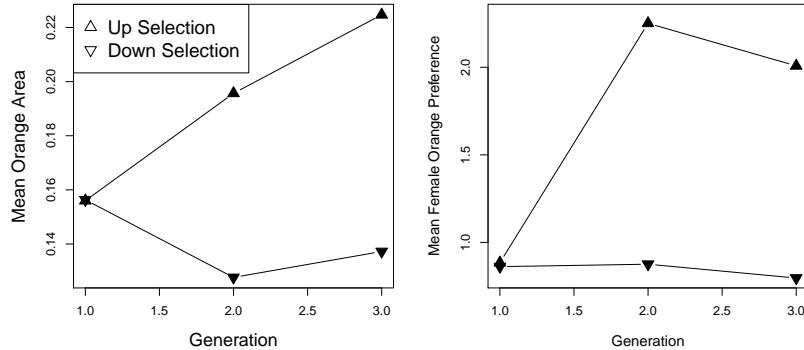


Figure 5.17: Mean phenotypes for the two up- and two down-selected populations of Guppies. Left panel: A response to selection was seen due to the direct selection on male colouration. Right panel: An indirect, correlated response was also seen in female preference. Data from HOUDE (1994). Code here.

As an example of how direct selection on display traits can drive  
 3270 the evolution of preference traits, let's consider some data from guppies. Guppies (*Poecilia reticulata*) are a classic system for studying  
 3272 the interplay of natural and sexual selection. In some populations  
 3274 of guppies, females show a preference for males with more orange  
 colouration.

HOUDE established four replicate population pairs of guppies and  
 3276 selected one of each pair for an increased or decreased orange col-  
 oration in males, selecting the top/bottom 20 out of 50 males. She  
 3278 randomly chose females from each population to form the next genera-  
 tion, and so did not exert direct selection on females. She measured  
 3280 the response to selection on male colouration and on female prefer-  
 ence for orange (left and right panels of Figure 5.1 respectively). In  
 3282 the lines that were selected for more orange males females showed an  
 increased preference for orange. While in those lines that she selected  
 3284 males for less orange in their display females showed a decreased pref-  
 erence for orange. This is consistent with indirect selection on female  
 3286 orange preference as a response to selection on male colouration, due  
 to a genetic correlation between female preference and male trait. It is  
 3288 *a priori* unlikely that pleiotropy is the source of the genetic correlation  
 between these traits, rather it is likely caused by females assortative  
 3290 mating with males that match their colour preference.

Returning to our bird tail example, what could drive the direct  
 3292 selection on male tail length? The selection for longer tails in males  
 could come about because longer tails are genetic correlated with  
 3294 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  
 3296 long tail is an honest signal. This would be a good genes explanation



Figure 5.18: Guppy (*Poecilia reticulata*).

From a set of 1962 stamps of Hungary.  
 Contributed to wikimedia by Darjac, not  
 covered by copyright

of female mate choice evolution.

3298 There's another subtler way that selection could favour our male trait. Imagine that the variation in female preference trait is because  
3300 some females have no strong preference for the male-tail length, but some females have a strong preference for males with longer tails.  
3302 Males with longer tails would then have higher fecundity than the short-tailed males as there's a subset of females who are strongly at-  
3304 tracted to long tails, and these males also get to mate with the other females. Thus selection favours long-tailed males, and so indirectly  
3306 favours female preference for longer tails; females with a preference for longer-tails have sons who in turn who are more attractive. This  
3308 model is sometimes called the sexy-son model. It is also called the Fisherian runaway model (FISHER, 1915), as female preference and  
3310 male trait can coevolve in an escalating fashion driving more and more extreme preferences for arbitrary traits. Thus many extravagant dis-  
3312 play traits in males and females may exist purely because individuals find them beautiful and are attracted to them.

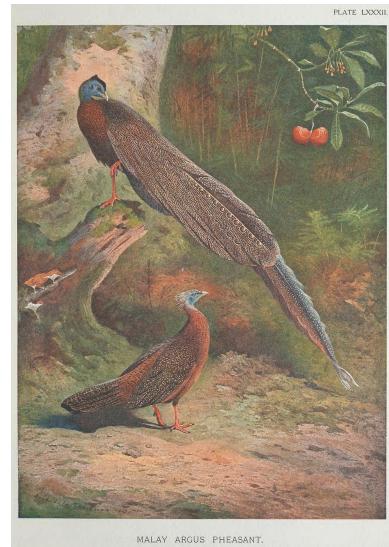


Figure 5.19: Argus Pheasant.  
A monograph of the pheasants. (1918). Beebe, W. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Institution Libraries. Licensed under CC BY-2.0.

“The case of the male Argus Pheasant is eminently interesting, because it affords good evidence that the most refined beauty may serve as a sexual charm, and for no other purpose.” – DARWIN (1888)



## *One-Locus Models of Selection*

3316     ”Socrates consisted of the genes his parents gave him, the experiences  
3318     they and his environment later provided, and a growth and development  
   mediated by numerous meals. For all I know, he may have been  
3320     very successful in the evolutionary sense of leaving numerous offspring.  
   His phenotype, nevertheless, was utterly destroyed by the hemlock  
   and has never since been duplicated. The same argument holds also  
3322     for genotypes. With Socrates’ death, not only did his phenotype dis-  
   appear, but also his genotype.[...] The loss of Socrates’ genotype is  
3324     not assuaged by any consideration of how prolifically he may have  
   reproduced. Socrates’ genes may be with us yet, but not his genotype,  
3326     because meiosis and recombination destroy genotypes as surely as  
   death.” –WILLIAMS (1966)

3328     Individuals are temporary, their phenotypes are temporary, and  
   their genotypes are temporary. However, the alleles that individuals  
3330     transmit across generations have permanence. Sustained phenotypic  
   evolutionary change due to natural selection occurs because of changes  
3332     in the allelic composition of the population. To understand these  
   changes, we need to understand how the frequency of alleles (genes)  
3334     changes over time due to natural selection.

As we have seen, natural selection occurs when there are differences  
3336     between individuals in fitness. We may define fitness in various ways.  
   Most commonly, it is defined with respect to the contribution of a  
3338     phenotype or genotype to the next generation. Differences in fitness  
   can arise at any point during the life cycle. For instance, different  
3340     genotypes or phenotypes may have different survival probabilities from  
   one stage in their life to the stage of reproduction (viability), or they  
3342     may differ in the number of offspring produced (fertility), or both.  
   Here, we define the absolute fitness of a genotype as the expected  
3344     number of offspring of an individual of that genotype. Differences in  
   fitness among genotypes drive allele frequency change. In this chapter  
3346     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 the determin-  
3348     istic dynamics of selection. We’ll return to discuss the interaction of

selection and drift in the next chapter.

3350 *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.

3358 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 3360 mean absolute fitness of the population at time  $t$  is

$$\bar{W}_t = W_1 \frac{P_t}{P_t + Q_t} + W_2 \frac{Q_t}{P_t + Q_t} = W_1 p_t + W_2 q_t, \quad (6.1)$$

3362 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.

3364 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. 3366 FEDER *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. 3368 The viral load of the macaque's blood plasma is shown as a black line in Figure 6.1. Twelve weeks after infection, the macaque was treated 3370 with an anti-retroviral drug that targeted the the virus' reverse transcriptase protein. Note how the viral load initially starts to drop once 3372 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 3374 their numbers are decreasing). However, the viral population rebounds as a mutation that confers drug resistance to the anti-retroviral drug 3376 arises in the SHIV and starts to spread. Viruses carrying this mutation 3378 (let's call them allele 1) likely have absolute fitness  $W_1 > 1$ . 3380 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 3382 SHIV population in week 20.

3384 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 FEDER *et al.*'s work was modeling the complicated spatial dynamics of drug-resistant SHIV adaptation in different organ systems.

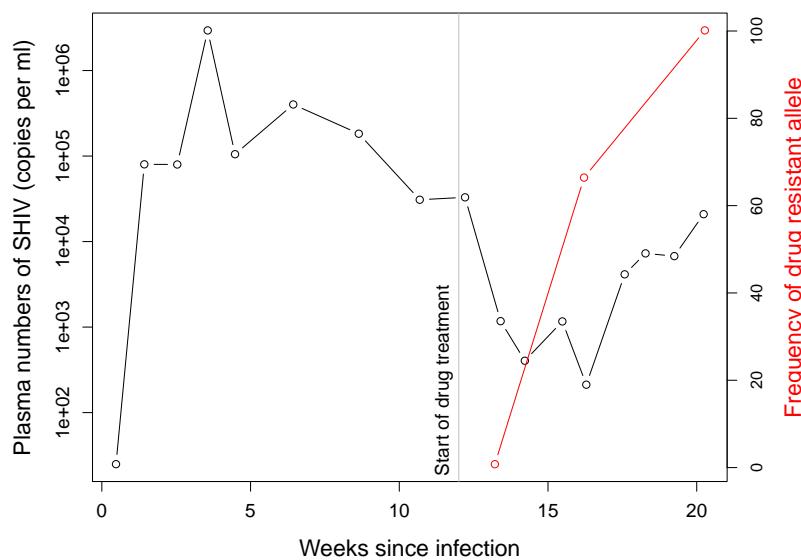


Figure 6.1: The rapid evolution of drug-resistant SHIV. The viral load of SHIV in the blood of a macaque (black line), the frequency of a drug resistance mutation (red line). Data from FEDER *et al.* (2017). Code here.

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  $\tau$  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  
 3406 relative frequency of allele  $A_1$  after  $\tau$  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  
 3408 switching from absolute to relative allele frequencies.

Rearranging eqn. (6.5) and setting  $t = 0$ , we can work out the time  
 3410  $\tau$  for the frequency of  $A_1$  to change from  $p_0$  to  $p_\tau$ . 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

$$\frac{p_\tau}{q_\tau} = \frac{p_0}{q_0} \left( \frac{w_1}{w_2} \right)^\tau. \quad (6.7)$$

3412 Solving this for  $\tau$  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  
 3414  $w_i$  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,  
 3416  $s$  is simply the difference in relative fitnesses between the two alleles.

Equation (6.5) becomes

$$p_{t+\tau} = \frac{p_t}{p_t + q_t(1-s)^\tau}, \quad (6.9)$$

3418 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

$$p_{t+\tau} \approx \frac{p_t}{p_t + q_t e^{-s\tau}}. \quad (6.10)$$

3420 This equation takes the form of a logistic function. That is because we  
 are looking at the relative frequencies of two ‘populations’ (of alleles  
 3422  $A_1$  and  $A_2$ ) that are growing (or declining) exponentially, under the  
 constraint that  $p$  and  $q$  always sum to 1.

3424 Moreover, eqn. (6.7) for the number of generations  $\tau$  it takes for a  
 certain change in frequency to occur becomes

$$\tau = -\log \left( \frac{p_\tau q_0}{q_\tau p_0} \right) / \log(1 - s). \quad (6.11)$$

3426 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 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 very close to fixation. Then, plugging these values into eqn. (6.12), we obtain

$$\begin{aligned}\tau &= \frac{1}{s} \log \left( \frac{1 - 2/N + 1/N^2}{1/N^2} \right) \\ &\approx \frac{1}{s} (\log(N) + \log(N - 2)) \\ &\approx \frac{2}{s} \log(N)\end{aligned}\tag{6.13}$$

where we make the approximations  $N^2 - 2N + 1 \approx N^2 - 2N$  and later  $N - 2 \approx N$ .

**Question 1.** In our example of the evolution of drug resistance, the drug-resistant SHIV virus spread from undetectable frequencies to  $\sim 65\%$  frequency by 16 weeks post infection. An estimated effective population size of SHIV is  $1.5 \times 10^5$ , and its generation time is  $\sim 1$  day. Assuming that the mutation arose as a single copy allele very shortly the start of drug treatment at 12 weeks, what is the selection coefficient favouring the drug resistance allele?

*Haploid model with fluctuating selection* Selection pressures may change while a polymorphism persists in the population due to environmental changes. We can use our haploid model to consider this case where the fitnesses depend on time (DEMPSTER, 1955), and say that  $w_{1,t}$  and  $w_{2,t}$  are the fitnesses of the two types in 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 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 frequencies  $p_t$  and  $q_t$ , then  $\tau$  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 frequency 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  $\tau^{\text{th}}$  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  $\sqrt[\tau]{\prod_{i=t}^{\tau-1} w_{1,i}}$  is the geometric mean fitness of allele  $A_1$  over the  $\tau$  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).

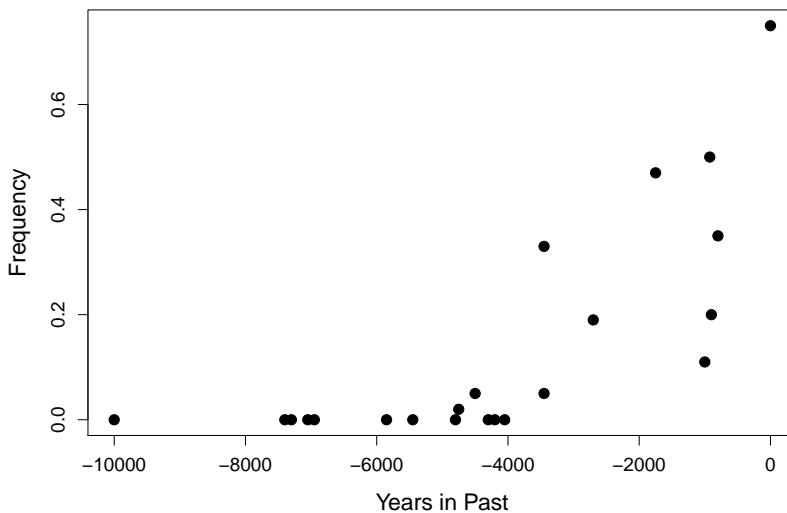


Figure 6.2: Frequency of the Lactase persistence allele in ancient and modern samples from Central Europe. Data compiled by MARCINIAK and PERRY (2017) from various sources. Thanks to Stephanie Marciniak for sharing these data. Code here.



Figure 6.3: 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.

### 6.0.2 Diploid model

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 6.2. 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 mate at random, the number of zygotes that are of the three genotypes and form generation  $t$  are

$$Np_t^2, \quad N2p_tq_t, \quad Nq_t^2. \quad (6.18)$$

The mean fitness of the population of zygotes is then

$$\bar{W}_t = W_{11}p_t^2 + W_{12}2p_tq_t + W_{22}q_t^2. \quad (6.19)$$

Again, this is simply the weighted mean of the genotypic fitnesses.

How many zygotes of each of the three genotypes survive to reproduce? An individual of genotype  $A_1A_1$  has a probability of  $W_{11}$  of surviving to reproduce, and similarly for other genotypes. Therefore, the expected number of  $A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  individuals who survive to reproduce is

$$NW_{11}p_t^2, \quad NW_{12}2p_tq_t, \quad NW_{22}q_t^2. \quad (6.20)$$

It then follows that the total number of individuals who survive to reproduce is

$$N(W_{11}p_t^2 + W_{12}2p_tq_t + W_{22}q_t^2). \quad (6.21)$$

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. Therefore, the relative frequency of individuals with the three different genotypes at reproduction is

$$\frac{NW_{11}p_t^2}{N\bar{w}}, \quad \frac{NW_{12}2p_tq_t}{N\bar{w}}, \quad \frac{NW_{22}q_t^2}{N\bar{w}} \quad (6.22)$$

(see Table 6.1).

As there is no difference in the fecundity of the three genotypes, the allele frequencies in the zygotes forming the next generation are simply

	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$
Absolute no. at birth	$Np_t^2$	$N2p_t q_t$	$Nq_t^2$
Fitnesses	$W_{11}$	$W_{12}$	$W_{22}$
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_t q_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{W_{11}}{W} p_t^2$	$\frac{W_{12}}{W} 2p_t q_t$	$\frac{W_{22}}{W} q_t^2$

Table 6.1: Relative genotype frequencies after one episode of viability selection.

- 3506 the allele frequency among the reproducing individuals of the previous generation. Hence, the frequency of  $A_1$  in generation  $t + 1$  is

$$p_{t+1} = \frac{W_{11}p_t^2 + W_{12}p_t q_t}{\bar{W}}. \quad (6.23)$$

- 3508 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 3510 absolute fitnesses with the relative fitnesses. That is, we may replace  $W_{ij}$  by  $w_{ij} = W_{ij}/\bar{W}_{11}$ , for instance.

3512 Each of our genotype frequencies is responding to selection in a manner that depends just on its fitness compared to the mean fitness 3514 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, 3516 we can estimate this fitness ratio for each genotype by comparing the frequency at birth compared to adults. As an example of this 3518 calculation, we'll look at some data from sticklebacks.

3520 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 system for studying the genetics of adaptation. In marine habitats, most 3522 of the stickleback have armour plates to protect them from predation, but freshwater populations repeatedly evolve the loss of armour plates 3524 due to selection on an allele at the Ectodysplasin gene (EDA). This 3526 allele is found as a standing variant at very low frequency in marine populations; BARRETT *et al.* took advantage of this fact and collected 3528 and bred a population of marine individuals carrying both the low- (L) and completely-plated (C) alleles. They introduced the offspring 3530 of this cross into four freshwater ponds and monitored genotype frequencies<sup>1</sup> 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

3532 The heterozygotes have increased in frequency dramatically in the 3534 population as their fitness is more than double the mean fitness of the population. We can also calculate the relative fitness of each genotype by dividing through by the fitness of the fittest genotype, the 3536

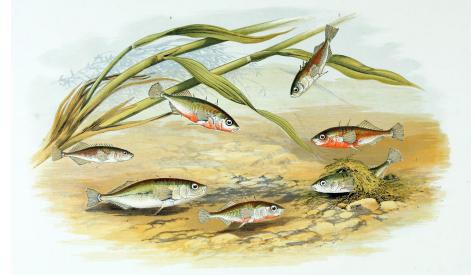


Figure 6.4: Freshwater threespine Stickleback (*G. aculeatus*). British fresh-water fishes. Houghton W 1879. Image from the Biodiversity Heritage Library. Contributed by Ernst Mayr Library, Harvard.. Not in copyright.

<sup>1</sup> The actual dynamics observed by BARRETT *et al.* are more complicated as in the very young fish selection reverses direction.

heterozygote in this case (doing this cancels through  $\bar{W}$ ). The 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 HWE in the juveniles.

**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?

**Question 3.** For many generations you have been studying an annual wildflower that has two color morphs, orange and white. You have discovered that a single bi-allelic locus controls flower color, with the white allele being recessive. The pollinator of these plants is an almost blind bat, so individuals are pollinated at random with respect to flower color. Your population census of 200 individuals showed that the population consisted of 168 orange-flowered individuals, and 32 white-flowered individuals.

Heavy February rainfall creates optimal growing conditions for an exotic herbivorous beetle with a preference for orange-flowered individuals. This year it arrives at your study site with a ravenous appetite. Only 50% of orange-flowered individuals survive its wrath, while 90% of white-flowered individuals survive until the end of the growing season.

A What is the initial frequency of the white allele, and what do you have to assume to obtain this?

B What is the frequency of the white allele in the seeds forming the next generation?

The change in frequency from generation  $t$  to  $t + 1$  is

$$\Delta p_t = p_{t+1} - p_t = \frac{w_{11}p_t^2 + w_{12}p_tq_t}{\bar{w}} - p_t. \quad (6.24)$$

To simplify this equation, we will first define two variables  $\bar{w}_1$  and  $\bar{w}_2$  as

$$\bar{w}_1 = w_{11}p_t + w_{12}q_t, \quad (6.25)$$

$$\bar{w}_2 = w_{12}p_t + w_{22}q_t. \quad (6.26)$$

These are called the marginal fitnesses of allele  $A_1$  and  $A_2$ , respectively. 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 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 note that the mean relative fitness can be expressed in terms of the marginal fitnesses as

$$\bar{w} = \bar{w}_1p_t + \bar{w}_2q_t, \quad (6.27)$$

3572 where, for notational simplicity, we have omitted subscript t for the  
dependence of mean and marginal fitnesses on time.

3574 We can then rewrite eqn. (6.24) using  $\bar{w}_1$  and  $\bar{w}_2$  as

$$\Delta p_t = \frac{(\bar{w}_1 - \bar{w}_2)}{\bar{w}} p_t q_t. \quad (6.28)$$

The sign of  $\Delta p_t$ , i.e. whether allele  $A_1$  increases or decreases in frequency, depends only on the sign of  $(\bar{w}_1 - \bar{w}_2)$ . The frequency of  $A_1$  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. (6.28)  
3578 and the respective expression for the haploid model in eqn. (6.4). (We  
3580 will return to the special case where  $\bar{w}_1 = \bar{w}_2$  shortly).

3582 We can also rewrite (6.24) as

$$\Delta p_t = \frac{1}{2} \frac{p_t q_t}{\bar{w}} \frac{d\bar{w}}{dp}, \quad (6.29)$$

the demonstration of which we leave to the reader. This form shows  
3584 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  
3586 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  $\frac{d\bar{w}}{dp} < 0$ ).  
3588 Thus, although selection acts on individuals, under this simple model,  
selection is acting to increase the mean fitness of the population. The  
3590 rate of this increase is proportional to the variance in allele frequencies  
within the population ( $p_t q_t$ ).

3592 **Question 4.** Show that eqns. (6.29) and (6.28) are equivalent.

(Trickier question.)

3594 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  
3596 parametrizations to gain insight about two specific modes of selection:  
directional selection and heterozygote advantage.

### 3598 6.0.3 Diploid directional selection

Directional selection means that one of the two alleles always has  
3600 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  
3602 we are interested in changes in allele frequencies, we may use relative  
fitnesses. We parameterize the reduction in relative fitness in terms  
3604 of a selection coefficient, similar to the one we met in the haploid  
selection section, as follows:

3606

genotype	$A_1A_1$	$A_1A_2$	$A_2A_2$
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	$1 - sh$	$1 - s.$

3608 Here, the selection coefficient  $s$  is the difference in relative fitness  
 between the two homozygotes, and  $h$  is the dominance coefficient. For  
 3610 selection to be directional, we require that  $0 \leq h \leq 1$  holds. The  
 dominance coefficient allows us to move between two extremes. One  
 3612 is when  $h = 0$ , such that allele  $A_1$  is fully dominant and  $A_2$  fully  
 recessive. In this case, the heterozygote  $A_1A_2$  is as fit as the  $A_1A_1$   
 3614 homozygote genotype. The inverse holds when  $h = 1$ , such that allele  
 $A_1$  is fully recessive and  $A_2$  fully dominant.

3616 We can then rewrite eqn. (6.28) as

$$\Delta p_t = \frac{p_t h s + q_t s(1-h)}{\bar{w}} p_t q_t, \quad (6.30)$$

where

$$\bar{w} = 1 - 2p_t q_t sh - q_t^2 s. \quad (6.31)$$

3618 **Question 5.** Throughout the Californian foothills are old copper and gold-mines, which have dumped out soils that are polluted  
 3620 with heavy metals. While these toxic mine tailings are often depauperate of plants, *Mimulus guttatus* and a number of other plant species  
 3622 have managed to adapt to these harsh soils. WRIGHT *et al.* (2015)  
 3624 have mapped one of the major loci contributing to the adaptation to  
 soils at two mines near Copperopolis, CA. WRIGHT *et al.* planted  
 3626 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.

3628 **A)** What is the selection coefficient acting against the non-copper-tolerant allele on the mine tailing?

3630 **B)** The copper-tolerant allele is fairly dominant in its action on fitness. If we assume that  $h = 0.1$ , what percentage of heterozygotes  
 3632 should survive to flower?

3634 **Question 6.** Comparing the red ( $h = 0$ ) and black ( $h = 0.5$ ) trajectories in Figure 6.5, provide an explanation for why  $A_1$  increases faster initially if  $h = 0$ , but then approaches fixation more slowly  
 3636 compared to the case of  $h = 0.5$ .

3638 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:  
 3640 silver, cross, and red (see Figure 6.8), with this difference primarily

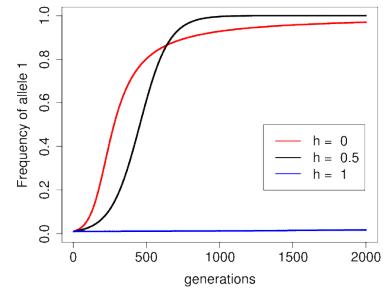


Figure 6.5: The trajectory of the frequency of allele  $A_1$ , starting from  $p_0 = 0.01$ , for a selection coefficient  $s = 0.01$  and three different dominance coefficients. The recessive beneficial allele ( $h = 1$ ) will eventually fix in the population, but it takes a long time. Code here.

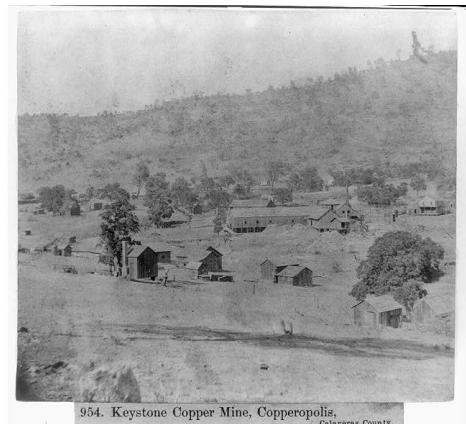


Figure 6.6: Keystone Copper Mine 1866, Copperopolis, Calaveras County. Image from picryl. Source Library of Congress, Public Domain.

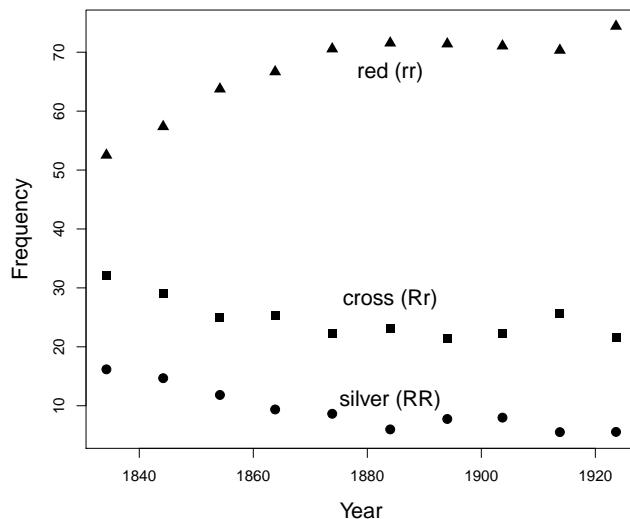


Figure 6.7: The frequency of red, cross, and silver fox morphs over the decades in Eastern Canada. These data are well described by recessive selection acting against the silver fox morph. Data from ELTON (1942), compiled by ALLENDORF and HARD (2009). [Code here.](#)

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 6.7, as documented by ELTON, from 16% to 5% from 1834 to 1937. HALDANE reanalyzed these data and showed that they were consistent with recessive selection acting against the silver morph alone.

Note how the heterozygotes (cross) decline somewhat as a result of selection on the silver homozygotes, but overall the R allele is slow to respond to selection as it is ‘hidden’ from selection in the heterozygote state.

*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 alleles with respect to fitness is strictly additive. Then, eqn. (6.30) simplifies to

$$\Delta p_t = \frac{1}{2} \frac{s}{\bar{w}} p_t q_t. \quad (6.32)$$

If selection is very weak, i.e.  $s \ll 1$ , the denominator ( $\bar{w}$ ) is close to 1 and we have

$$\Delta p_t = \frac{1}{2} s p_t q_t. \quad (6.33)$$



Figure 6.8: Three colour morphs in red fox *V. vulpes*, cross, red, and silver foxes from left to right.  
The previous black and silver gray foxes are mostly color prints, according to prints of the author's old colored fox book, page 410.

Nelson, E.W., Fuertes, L.A. 1916. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. No known copyright restrictions.

3662 It is instructive to compare eqn. (6.33) to the respective expression  
 under the haploid model. To this purpose, start from the generic term  
 3664 for  $\Delta p_t$  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) becomes  
 3666  $\Delta p_t = sp_t q_t$ . Hence, if  $s$  is small, the diploid model of directional  
 selection without dominance is identical to the haploid model, up to a  
 3668 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  $w_{22} = 1 - 2s$  in our diploid  
 3670 model instead, in which case the agreement with the haploid model  
 would be perfect.

3672 From this analogy, we can borrow some insight we gained from the  
 haploid model. Specifically, the trajectory of the frequency of allele  
 3674  $A_1$  in the diploid model without dominance follows a logistic growth  
 curve similar to (6.10). From this similarity, we can extrapolate from  
 3676 Equation (6.12) to find the time it takes for our diploid, beneficial,  
 additive allele ( $A_1$ ) to move from frequency  $p_0$  to  $p_\tau$ :

$$\tau \approx \frac{2}{s} \log \left( \frac{p_\tau q_0}{q_\tau p_0} \right) \quad (6.34)$$

3678 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,  
 3680 additive allele ( $A_1$ ) to transit from its entry into the population ( $p_0 =$   
 $1/(2N)$ ) to close to fixation ( $p_\tau = 1 - 1/(2N)$ ):

$$\tau \approx \frac{4}{s} \log(2N) \quad (6.35)$$

3682 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  
 3684 (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 our  
 3686 allele to move through the population.

**Question 7.** Gulf killifish (*Fundulus grandis*) have rapidly adapted  
 3688 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 dele-  
 3690 tion of their aryl hydrocarbon receptor (AHR) gene. Oziolor et al.  
 estimated that individuals who were homozygote for the intact AHR  
 3692 gene had a relative fitness of 20% of that of homozygotes for the dele-  
 tion. Assuming an effective population size of 200 thousand individu-  
 3694 als, how long would it take for the deletion to reach fixation, starting  
 as a single copy in this population?

3696 Directional selection on genotypes is expected to remove variation  
 from populations, yet we see plentiful phenotypic and genetic variation  
 3698 in every natural population. Why is this? Three broad explanations  
 for the maintenance of polymorphisms are

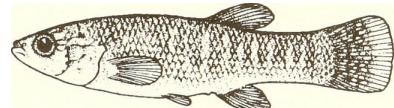


Figure 6.9: Gulf killifish (*Fundulus grandis*).

Distribution and abundance of fishes and invertebrates in Gulf of Mexico estuaries. Nelson D M and Pattillo M E Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. No known copyright restrictions.

- 3700 1. Variation is maintained by a balance of genetic drift and mutation  
 (we discussed this explanation in Chapter 3).
- 3702 2. Selection can sometimes act to maintain variation in populations  
 (balancing selection).
- 3704 3. Deleterious variation can be maintained in the population as a bal-  
 ance between selection removing variation and mutation constantly  
 3706 introducing new variation into the population.

We'll turn to these latter two explanations through the rest of the  
 3708 chapter. Note that these explanations are not mutually exclusive, and  
 each of them will explain some proportion of the variation.

3710 *6.0.4 Heterozygote advantage*

One form of balancing selection occurs when the heterozygotes are  
 3712 fitter than either of the homozygotes. In this case, it is useful to pa-  
 rameterize the relative fitnesses as follows:

genotype	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$
absolute fitness	$w_{11}$	$\langle w_{12} \rangle$	$w_{22}$
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	$1 - s_2$

3716 Here,  $s_1$  and  $s_2$  are the differences between the relative fitnesses  
 of the two homozygotes and the heterozygote. Note that to obtain  
 3718 relative fitnesses we have divided absolute fitness by the heterozygote  
 fitness. We could use the same parameterization as in the model of  
 3720 directional selection, but the reparameterization we have chosen here  
 makes the math easier.

3722 In this case, when allele  $A_1$  is rare, it is often found in a heterozy-  
 gous state, while the  $A_2$  allele is usually in the homozygous state, and  
 3724 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, while  
 3726 the allele  $A_2$  is often found in a heterozygous state; thus it is now al-  
 lele  $A_2$  that increases in frequency at the expense of allele  $A_1$ . Thus,  
 3728 at least in the deterministic model, neither allele can reach fixation  
 and both alleles will be maintained at an equilibrium frequency as a  
 3730 balanced polymorphism in the population.

We can solve for this equilibrium frequency by setting  $\Delta p_t = 0$   
 3732 in eqn. (6.28), 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  
 3734 interesting ( $p = 0$  or  $q = 0$ ), but the third one is the polymorphic equi-  
 librium, where  $\bar{w}_1 - \bar{w}_2 = 0$  holds. Using our  $s_1$  and  $s_2$  parametrization

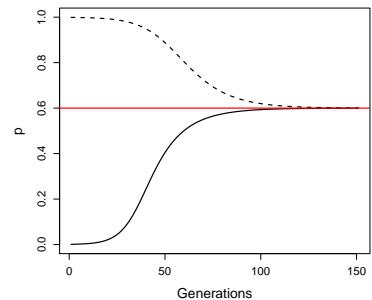
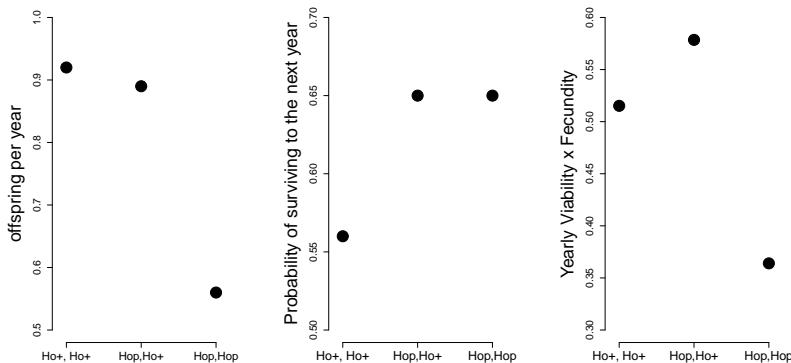


Figure 6.10: Two allele frequency trajectories of the  $A_1$  allele subject to heterozygote advantage ( $w_{11} = 0.9$ ,  $w_{12} = 1$ , and  $w_{22} = 0.85$ ). In one simulation the allele is started from being rare in the population ( $p = 1/1000$ , solid line) and increases in frequency/ In the other 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.

above, we see that the marginal fitnesses of the two alleles are equal when

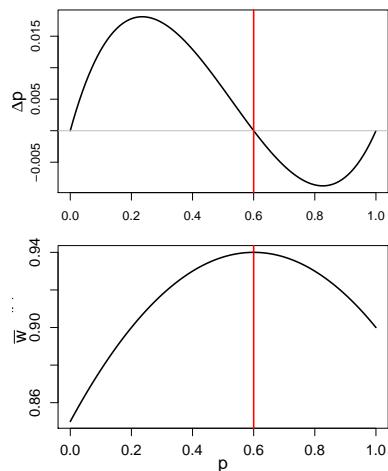
$$p_e = \frac{s_2}{s_1 + s_2} \quad (6.36)$$

for the equilibrium frequency of interest. This is also the frequency 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 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 equilibrium frequency represents an evolutionary compromise between the advantages of the heterozygote and the comparative costs of the two homozygotes.



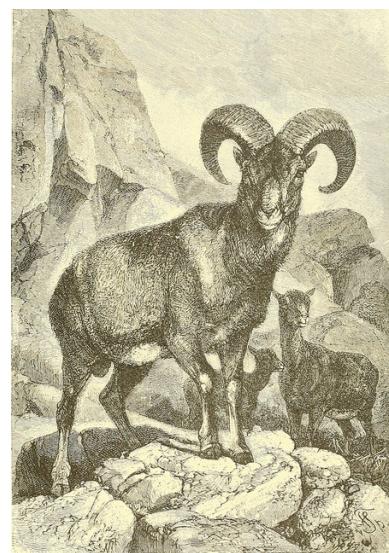
One example of a polymorphism maintained by heterozygote advantage is a horn-size polymorphism found in Soay sheep, a population of feral sheep on the island of Soay (about 40 miles off the coast of Scotland). The horns of the soay sheep resemble those of the wild Mouflon sheep, and the male Soay sheep use their horns to defend females during the rut. JOHNSTON *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. JOHNSTON *et al.* (2013) found that the  $Ho$  locus had substantial effects on male, but not female, fitness (see Figure 6.13).

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 6.13). Conversely, the  $Ho^+$  has a mostly recessive effect on viability, with  $Ho^+$  homozygotes having lower yearly



**Figure 6.11:** **Top)** The change in frequency of an allele with heterozygote advantage within a generation ( $\Delta p$ ) as a function of the allele frequency. Fitnesses as in Figure 6.10. Note how the frequency change is positive below the equilibrium frequency ( $p_e$ ) and negative above. **Bottom)** Mean fitness ( $\bar{w}$ ) as a function of the allele frequency. The red line shows the equilibrium frequency ( $p_e$ ). Code here.

Figure 6.12: 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 JOHNSTON *et al.* (2013). Code here.



**Figure 6.13:** Mouflon (*Ovis orientalis orientalis*). Animate creation. (1898). Wood, J. G. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

survival (middle plot Figure 6.13), 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  $\text{Ho}^P\text{Ho}^+$  heterozygotes have the highest fitness (right plot Figure 6.13). 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  $\text{Ho}^P$  allele is 10%, that there are 1000 males at birth, and that individual adults mate at random.

A) What is the expected number of males with each of the three genotypes in the population at birth?

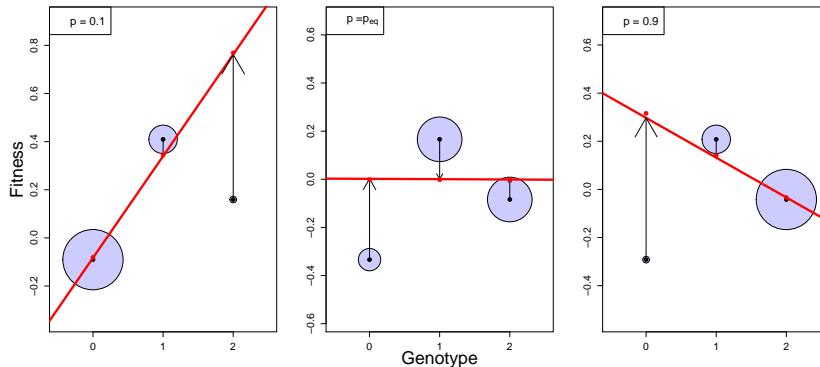
B) Assume that a typical male individual of each genotypes has the following probability of surviving to adulthood:

$\text{Ho}^+$	$\text{Ho}^P$	$\text{Ho}^+ \text{Ho}^P$	$\text{Ho}^P \text{Ho}^P$
0.5	0.8	0.8	0.5

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  $\text{Ho}^+ \text{Ho}^+$  and  $\text{Ho}^P \text{Ho}^P$  genotype have on average 2.5 offspring, while  $\text{Ho}^P \text{Ho}^P$  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  $\text{Ho}^+$  allele in the sperm that will form this next generation?

E) How would your answers to B-D change if the  $\text{Ho}^P$  allele was at 90% frequency?



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

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.

Figure 6.14: The deviation of the fitness of each genotype away from the mean population fitness (0) 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 genetic value 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.

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  $\Delta p_t$  is negative and allele  $A_1$  will be lost, while if  $p > p_U$ , allele  $A_1$  will become fixed.

While underdominant alleles might not spread within populations (if  $p_U \gg 0$  and selection is reasonably strong), they are of 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, if isolation of the subpopulations is not complete.

**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



Figure 6.15: 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 (OWEN and CHANTER, 1972). Illustrations of new species of exotic butterflies (1868) Hewitson. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

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.

- 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.

**Question 10.** OPTIONAL trickier question.

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  $r$  of the gametes of heterozygotes (D/d) to carry the D allele ( $r > 0.5$ ). The D allele causes viability problems in the homozygous state, such that the relative fitnesses are  $w_{dd} = 1$ ,  $w_{Dd} = 1$ ,  $w_{DD} = 1 - e$ . The D allele is currently at frequency  $p$  in the population at birth. Assume that the population is very large and no mutation occurs:

- A)** What is the frequency of the D allele in the next generation, before selection has had a chance to act?  
**B)** What conditions do you need for a polymorphic equilibrium to be maintained? What is the equilibrium frequency of this balanced polymorphism?  
**C)** Imagine the cost of the driver were additive:  $w_{dd} = 1$ ,  $w_{Dd} = 1 - e$ ,  $w_{DD} = 1 - 2e$ . Under what conditions can the driver invade the population? Can a polymorphic equilibrium be maintained?

*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 6.16; 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. DOBZHANSKY (1943) and WRIGHT and DOBZHANSKY (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 6.16). 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*. PAABY *et al.* (2014) tracked the frequency of this allele over time and found it oscillated with the seasons (right side of 6.16). 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.

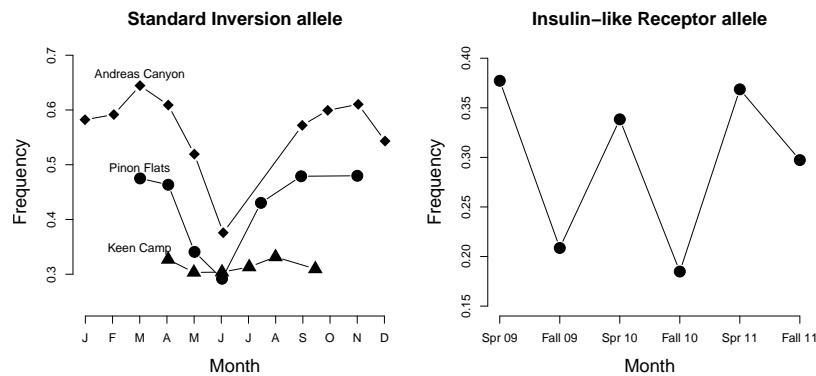


Figure 6.16: **Left**) Seasonal variation in the frequency of the ‘Standard’ inversion allele in *Drosophila pseudoobscura* for three populations from Mount San Jacinto, CA. These frequencies are an average over four years. Data from WRIGHT and DOBZHANSKY (1946). **Right**) The frequency of an allele at the Insulin-like Receptor gene over three years in *Drosophila melanogaster* samples from an Orchard in Pennsylvania. Data from PAABY *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 HALDANE and JAYAKAR, 1963).

When  $A_1$  is rare, i.e.  $p_t \ll 1$ , the frequency of  $A_1$  in the next generation (6.23) can be approximated as

$$p_{t+1} \approx \frac{w_{12}}{\bar{w}} p_t. \quad (6.37)$$

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

$$\frac{p_{t+1}}{q_{t+1}} = \frac{w_{12,t}}{w_{22,t}} \frac{p_t}{q_t}. \quad (6.38)$$

Starting from out from  $p_0$  and  $q_0$  in generation 0, then  $t+1$  generations later we have

$$\frac{p_{t+1}}{q_{t+1}} = \left( \prod_{i=0}^t \frac{w_{12,i}}{w_{22,i}} \right) \frac{p_0}{q_0}. \quad (6.39)$$

From this we can see, following our haploid argument from above, that  
 3900 the frequency of allele  $A_1$  will increase when rare only if

$$\frac{\sqrt[t]{\prod_{i=0}^t w_{12,i}}}{\sqrt[t]{\prod_{i=0}^t w_{22,i}}} > 1, \quad (6.40)$$

i.e. if the heterozygote has higher geometric mean fitness than the  
 3902  $A_2A_2$  homozygote.

The question now is whether allele  $A_1$  will approach fixation in  
 3904 the population, or whether there are cases in which we can obtain a  
 balanced polymorphism. To investigate that, we can simply repeat our  
 3906 analysis for  $q \ll 1$ , and see that in that case

$$\frac{p_{t+1}}{q_{t+1}} = \left( \prod_{i=0}^t \frac{w_{11,i}}{w_{12,i}} \right) \frac{p_0}{q_0}. \quad (6.41)$$

Now, for allele  $A_1$  to carry on increasing in frequency and to approach  
 3908 fixation, the  $A_1A_1$  genotype has to be out-competing the heterozy-  
 gotes. For allele  $A_1$  to approach fixation, we need the geometric mean  
 3910 of  $w_{11,i}$  to be greater than the geometric mean fitness of heterozy-  
 gotes ( $w_{12,i}$ ). At the same time, if heterozygotes have higher geometric  
 3912 mean fitness than the  $A_1A_1$  homozygotes, then the  $A_2$  allele will in-  
 crease in frequency when it is rare.

Intriguingly, we can thus have a balanced polymorphism even if the  
 3914 heterozygote is never the fittest genotype in any generation, as long  
 3916 as the heterozygote has a higher geometric mean fitness than either of  
 the homozygotes. In this case, the heterozygote comes out ahead when  
 3918 we think about long-term fitness across heterogeneous environmental  
 conditions, despite never being the fittest genotype in any particular  
 3920 environment.

As a toy example of this type of balanced polymorphism, consider a  
 3922 plant population found in one of two different environments each gen-  
 eration. These occur randomly;  $1/2$  of time the population experiences  
 3924 the dry environment and with probability  $1/2$  it experiences the wet  
 environment. The absolute fitnesses of the genotypes in the different  
 3926 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 ho-  
 3928 mozygote in the two environments. Then, if the two environments are  
 3930 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  $t$  genera-  
 3932 tions, we can take the  $t^{th}$  root to obtain the geometric mean fitness.

This example is loosely based on the work of SCHEMSKE and BIERZY-  
 CHUDEK (2001) on *Linanthus par-ryae*, a desert annual, endemic to California. There are blue- and a white-flowered colour morphs poly-  
 morphic 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 TURELLI *et al.* (2001) for a more detailed analysis.

Taking the  $t^{th}$  root, we find the geometric mean fitness of the AA allele is  $w_{AA,\text{dry}}^{1/2} w_{AA,\text{wet}}^{1/2}$ . 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) 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 polli-



Figure 6.17: Elderflower orchid (*Dactylorhiza sambucina*).

Abbildungen der in Deutschland und den angrenzenden gebieten vorkommenden grundformen der orchideenarten (1904). Müller, W. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden. Not in copyright.

nated 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 (SMITHSON and MACNAIR, 1997). GIGORD *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 6.18 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.

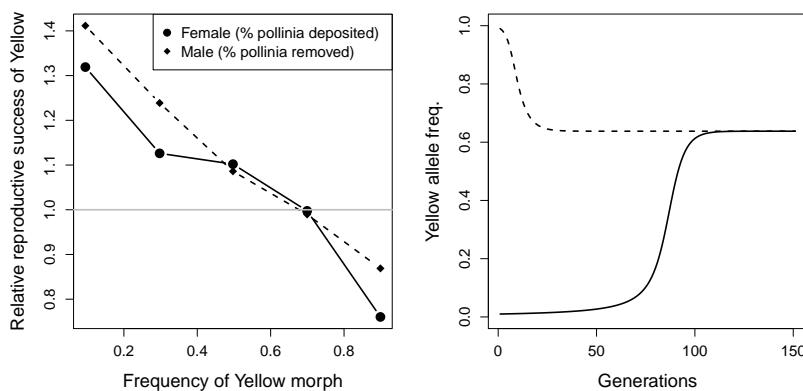


Figure 6.18: **Left)** Measures of the relative male- and female- reproductive success of the yellow Elderflower orchid morph as a function of the yellow morph in experimental plots. **Right)** Two allele frequency trajectories of the Yellow allele subject to negative frequency scheme given in the left 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 GIGORD *et al.* (2001). Code here.

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 large majority of males are ‘Independent’, with black or chestnut ruff plumage, and try to defend and display on small territories. ‘Satellite’ males, with white ruff plumage, make up ~ 16% of males and do not defend territories, but rather join in displays with Independent males and opportunistically mate with females visiting the lek. Finally, the rare ‘Faeder’

morph was only discovered in 2006 (JUKEMA and PIERSMA, 2006)  
4004 and makes up less than 1% of males. These Faeder males are female  
mimics who hang around the territories of Independents and try to  
4006 'sneak' in matings with females. Faeder males have plumage closely  
resembling that of females and a smaller body size than other males,  
4008 but with larger testicles (presumably to take advantage of rare mating  
opportunities). All three of these morphs, with their complex be-  
4010 havioural and morphological differences, are controlled by three alleles  
at a single autosomal locus, with the Satellite and Faeder alleles being  
genetically dominant over the high frequency Independent allele. The



Figure 6.19: 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.

4012 genetic variation for these three morphs is potential maintained by  
4014 negative frequency-dependent selection, as all three male strategies are  
likely at an advantage when they are rare in the population. For ex-  
4016 ample, while the Satellites mostly lose out on mating opportunities to  
Independents, they may have longer life-spans and so may have equal  
4018 life-time reproductive success (WIDEMO, 1998). However, Satellite  
and Faeder males are totally reliant on the lekking Independent males,  
4020 and so both of these alternative strategies cannot become overly com-  
mon in the population. The locus controlling these differences has  
4022 been mapped, and the underlying alleles have persisted for roughly  
four million years (KÜPPER *et al.*, 2016; LAMICHHANEY *et al.*,  
4024 2016). While this mating system is bizarre, the frequency dependent  
dynamics mean that it has been around longer than we've been using  
4026 stone tools.

4028 While these examples may seem somewhat involved, they must be  
simple compared to the complex dynamics that maintain the hundreds  
of alleles present at the genes in the Major histocompatibility complex  
4030 (MHC). MHC genes are key to the coordination of the vertebrate  
immune system in response to pathogens, and are likely caught in an  
4032 endless arms race with pathogens adapting to common MHC alleles,

allowing rare MHC alleles to be favoured. Balancing selection at the  
 4034 MHC locus has maintained some polymorphisms for tens of millions  
 of years, such that some of your MHC alleles may be genetically more  
 4036 closely related to MHC alleles in other primates than they are to  
 alleles in your close human friends.

4038 *6.0.5 Mutation-selection balance*

Mutation is constantly introducing new alleles into the population.  
 4040 Therefore, variation can be maintained within a population not only  
 if selection is balancing (e.g. through heterozygote advantage or fluctu-  
 4042 ating selection over time, as we have seen in the previous section),  
 but also due to a balance between mutation introducing deleterious  
 4044 alleles and selection acting to purge these alleles from the population  
 (HALDANE, 1937). To study mutation-selection balance, we return to  
 4046 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	$w_{11} = 1$	$w_{12} = 1 - sh$	$w_{22} = 1 - s$ .

4048 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 dis-  
 4050 advantage. 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 re-  
 4052 verse 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  
 4054 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 fol-  
 4056 lowing 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 of  
 4058  $A_2$  due to selection can be written as

$$\Delta_S q_t = \frac{\bar{w}_2 - \bar{w}_1}{\bar{w}} p_t q_t \approx -hsq_t. \quad (6.42)$$

This approximation can be found by assuming that  $q^2 \approx 0$ ,  $p \approx 1$ ,  
 4060 and that  $\bar{w} \approx w_1$ . All of these assumptions make sense if  $q \ll 1$ .  
 From eqn. (6.42) we see that selection acts to reduce the frequency of  
 4062  $A_2$  (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  
 4064 frequency at time  $t$  is approximately

$$q_t = q_0(1 - hs)^t. \quad (6.43)$$

We will now consider the change in frequency induced by mutation.

4066 Recalling that  $\mu$  is the mutation rate from  $A_1$  to  $A_2$  per generation,  
 the frequency of  $A_2$  after mutation is

$$q' = \mu p_t + q_t = \mu(1 - q_t) + q_t. \quad (6.44)$$

<sup>4068</sup> 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

$$\Delta_M q_t = q' - q_t = \mu. \quad (6.45)$$

<sup>4070</sup> Hence, when  $A_2$  is rare and the mutation rate is low, mutation acts to linearly increase the frequency of the deleterious allele  $A_2$ .

<sup>4072</sup> If selection is to balance deleterious mutation, their combined effect over one generation has to be zero. Therefore, to find the mutation–<sup>4074</sup> selection equilibrium, we set

$$\Delta_M q_t + \Delta_S q_t = 0, \quad (6.46)$$

insert eqns. (6.42) and (6.45), and solve for  $q$  to obtain

$$q_e = q_t = \frac{\mu}{hs}. \quad (6.47)$$

<sup>4076</sup> We see that the frequency of the deleterious allele  $A_2$  is balanced at a frequency equal to the mutation rate ( $\mu$ ) divided by the reduction in <sup>4078</sup> relative fitness in the heterozygote ( $hs$ ).

<sup>4080</sup> 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. <sup>4082</sup> 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 <sup>4084</sup> 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 <sup>4086</sup> assumption that both  $\mu$  and  $s$  are small.

<sup>4088</sup> If an allele is truly recessive (although few likely are), we have  $h = 0$ , and so eqn. (6.47) is not valid. However, we can make an argument similar to the one above to show that, for truly recessive <sup>4090</sup> alleles,

$$q_e = \sqrt{\frac{\mu}{s}}. \quad (6.48)$$

**Question 11.** Oblong-winged katydids (*Amblycorypha oblongifolia*) <sup>4094</sup> are usually green. However, some are bright pink, thanks to an erythrism mutation (a nice example of early Mendelian reasoning in a <sup>4096</sup> wonderfully titled paper<sup>2</sup>). This pink condition is thought to be due to a dominant mutation (Crew, 2013). Assume that roughly one in <sup>4098</sup> 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 <sup>4100</sup> heterozygotes for the pink mutation?



Figure 6.20: Oblong-winged katydid. Field book of insects (1918). Lutz, F.E. Illustrations by Edna L. Beutemüller. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

<sup>2</sup> WHEELER, W. M., 1907 Pink Insect Mutants. The American Naturalist 41(492): 773–780

*The genetic load of deleterious alleles* What effect do such deleterious mutations at mutation-selection balance have on the population? 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 single site at which a deleterious mutation is segregating at frequency  $q_e = \mu/(hs)$ , the population mean relative fitness is reduced to

$$\bar{w} = 1 - 2p_e q_e hs - q_e^2 s \approx 1 - 2\mu. \quad (6.49)$$

Somewhat remarkably, the drop in mean fitness due to a site segregating at mutation-selection balance is independent of the selection coefficient against the heterozygote; it depends only on the mutation rate. Intuitively this is because, given a fixed mutation rate, less deleterious alleles can rise to a higher equilibrium frequency, and thus contribute the same total load as more deleterious (rarer) alleles, but 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. if  $h > 0$ .

A fitness reduction of  $2\mu$  is very small, given that the mutation rate of a gene is likely  $< 10^{-5}$ . However, if there are many loci segregating 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, the human genome contains over twenty thousand genes, and many other functional regions, the vast majority of which will be subject to purifying selection against mutations that disrupt their function. In humans, most loss of function (LOF) variants, which severely disrupt a protein-coding gene, are found at low frequencies. However, each human genome typically carries over a hundred LOF variants (MACARTHUR *et al.*, 2012; LEK *et al.*, 2016). Not every LOF allele will be deleterious; some could even be advantageous. However, the combined load of these LOF alleles must on average lower our fitness, otherwise selection wouldn't be removing them from the population. Each one of us carries a unique set of these LOF alleles, usually in a heterozygous state. We differ slightly in how many of these alleles we carry. For example, the left side of Figure 6.21 shows the distribution of the number of LOF alleles carried by 769 individuals of Dutch ancestry. The individuals who carry fewer of these LOF alleles will on average have higher fitness than those individuals with more.

How do these differences across individuals in total LOF mutations mount up? Well, if we are willing to assume that the fitness costs of deleterious alleles interact multiplicatively, we can make some progress. If an individual who carries one LOF mutation has a fitness  $1 - hs$ , then an individual who's heterozygote for two LOF mutations would have fitness  $(1 - hs)^2$ , and an individual who is heterozygote

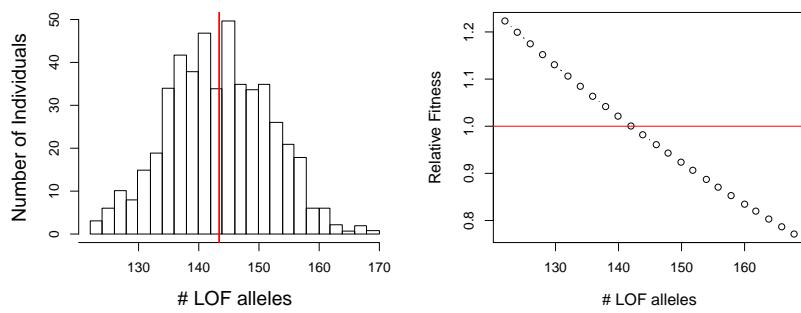


Figure 6.21: **Left)** The distribution of LOF alleles in 769 individuals from the Genome of the Netherlands project. Data from FRANCIOLI *et al.*. The average individual (red line) carries 144 LOF alleles. **Right).** The relative fitness of individuals carrying these varying numbers of LOF alleles, assuming multiplicative selection and a selection coefficient of  $sh = 10^{-2}$  acting against these alleles (CASSA *et al.*, 2017). Code here.

for  $L$  LOF alleles would have fitness  $(1 - hs)^L$ . The right-hand side of Figure 6.21 shows the predicted fitness of individuals carrying varying number of LOF alleles, relative to the mean fitness of the sample, using this multiplicative model. We don't yet know how much lower the fitness of these individuals really is, nor do we know how most of these LOF alleles manifest their fitness consequences through disease and other mechanisms. However, it's a reasonable guess that this variation in LOF alleles, presumably maintained by mutation-selection balance, is a major source of variation in fitness.

#### 6.0.6 Inbreeding depression

All else being equal, eqn. (6.47) suggests that mutations that have a smaller effect in the heterozygote can segregate at higher frequency under mutation-selection balance. As a consequence, alleles that have strongly deleterious effects in the homozygous state can still segregate at low frequencies in the population, as long as they do not have too strong a deleterious effect in heterozygotes. Thus, outbred populations may have many alleles with recessive deleterious effects segregating within them.

**Question 12.** 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 considering the fitness effects of this locus, and measuring fitness relative to the most fit genotype, answer the following questions:

- 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?

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 in-

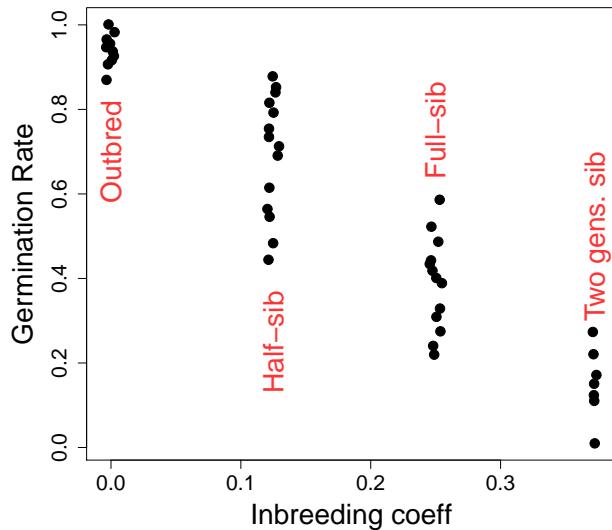


Figure 6.22: Data showing inbreeding depression over different degrees of inbreeding in *S. latifolia*. Each point is the mean seed germination rates for different family crosses. Data from RICHARDS. Code here.

breeding coefficient, i.e. so-called 'inbreeding depression' is a common observation. This wide-spread observation dates back to systematic surveys of inbreeding depression by DARWIN (1876). Inbreeding depression is likely primarily a consequence of being homozygous at many loci for alleles with recessive deleterious effects.

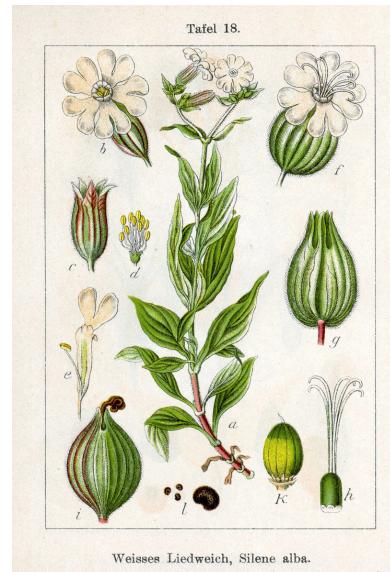
One example of inbreeding depression is shown in Figure 6.22. White campion (*Silene latifolia*) is a dioecious flowering plant; dioecious means that the males and females are separate individuals. RICHARDS performed crosses to create offspring who were outbred, the offspring of half-sibs, full-sibs, and of two generations of full-sib mating. He measured their germination success, which is plotted in Figure 6.22. Note how the fitness of individuals declines with increased inbreeding.

*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

$$q_e = \frac{\mu}{(h(1 - F) + F)s} \quad (6.50)$$

The frequency of the deleterious allele is decreased due to the al-



Tafel 18.  
Weisses Liedweich, *Silene alba*.  
Deutschlands Flora in Abbildungen (1796).  
Johann Georg Sturm (Painter: Jacob Sturm).  
Public Domain, wikimedia.

4196 allele now being expressed in homozygotes, and therefore exposed to selection, more often due to inbreeding. Thus, all else being equal,  
4198 populations with a high degree of inbreeding will purge their load.

### 6.0.7 Migration-selection balance

4200 Another reason for the persistence of deleterious alleles in a population  
4202 is that there is a constant influx of maladaptive alleles from other pop-  
4204ulations where these alleles are locally adaptive. Migration-selection  
4206 balance seems unlikely to be as broad an explanation for the persis-  
tence 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  
4208 adaptation is relatively unimpeded by migration at broad geo-  
graphically scales, where selection pressures change more slowly than  
4210 distances over which individuals typically migrate over a number of  
generations. Adaptation can, however, potentially occur on much finer  
4212 geographic scales, from kilometers down to meters in some species. On  
such small scales, dispersal is surely rapidly moving alleles between  
4214 environments, but local adaptation is maintained by the continued  
action of selection. An example of adaptation at fine-scales is shown in  
4216 Figure 6.25. JAIN and BRADSHAW (1966) studied the patterns of  
heavy-metal resistance in plants on mine tailings and in nearby mead-  
4218 ows, a set of classic studies of population differences maintained by  
local adaptation to different soils. Even at these very short geo-



Figure 6.24: Sweet vernal grass (*Anthoxanthum odoratum*).  
Billeder af nordens flora (1917). Mentz, A &  
Ostenfeld, C H. Image from the Biodiversity  
Heritage Library. Contributed by New York  
Botanical Garden. Not in copyright.

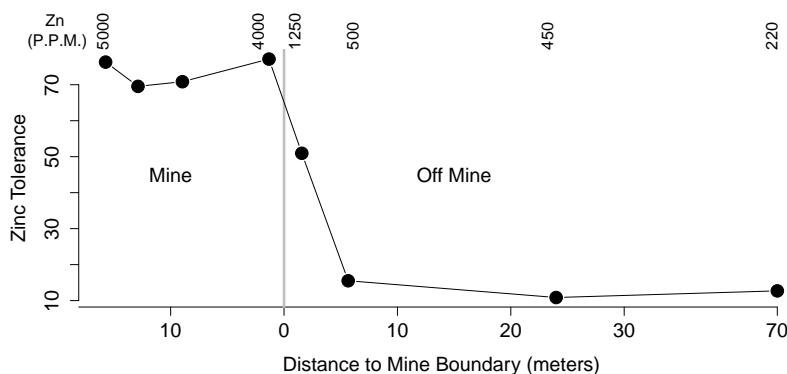


Figure 6.25: Data showing the Zinc tolerance of *Anthoxanthum odoratum* on and off of the Trelogan Mine, Flintshire, North Wales. The numbers along the top give the soil contamina-  
tion of Zinc in parts per million. Data from JAIN and BRADSHAW (1966).  
Code here.

4220 ically scales, over which seed and pollen will definitely move, we see  
strong local adaptation. Zinc-intolerant alleles are nearly absent from  
4222 the mine tailings because they prevent plants from growing on these  
zinc-heavy soils; conversely, zinc-tolerant alleles do not spread into  
4224 the meadow populations, likely due to some trade-off or fitness cost of

zinc-tolerance.

4226 As a first pass at developing a model of local adaptation, let's consider a haploid two-allele model with two different populations, see  
4228 Figure 6.26, where the relative fitnesses of our alleles are as follows

allele	1	2
population 1	1	1-s
population 2	1-s	1

4230 As a simple model of migration, let's suppose within a population a fraction of  $m$  individuals are migrants from the other population, and  
4232  $1 - m$  individuals are from the same population.

To quickly sketch an equilibrium solution to this scenario, we'll take  
4234 an approach analogous to our mutation-selection balance model. To do this, let's assume that selection is strong compared to migration ( $s \gg$   
4236  $m$ ), 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  
4238 changes the frequency of allele 2 in population 1 ( $q_1$ ) by

$$\Delta_{Mig.} q_1 \approx m \quad (6.51)$$

while as noted above  $\Delta_S q_1 = -sq_1$ , so that migration and selection  
4240 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

$$q_{e,1} = \frac{m}{s} \quad (6.52)$$

4242 Here, migration is playing the role of mutation and so migration–  
selection balance (at least under strong selection) is analogous to  
4244 mutation–selection balance.

We can use this same model by analogy for the case of migration–  
4246 selection balance in a diploid model. For the diploid case, we replace  
our haploid  $s$  by the cost to heterozygotes  $hs$  from our directional  
4248 selection model, resulting in a diploid migration–selection balance  
equilibrium frequency of

$$q_{e,1} = \frac{m}{hs} \quad (6.53)$$

4250 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.  
4252 Throughout the deserts of the American Southwest there are old lava flows, where the rocks and soils are much dark than the surrounding  
4254 desert.

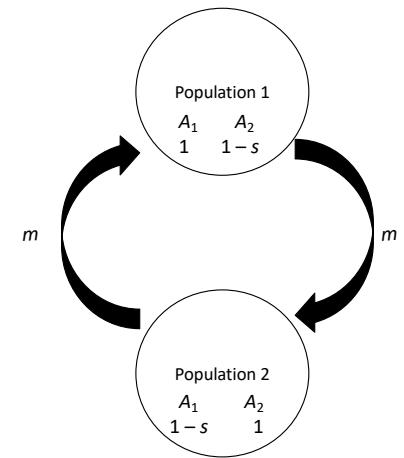


Figure 6.26: Setup of a two-population haploid model of local adaptation.

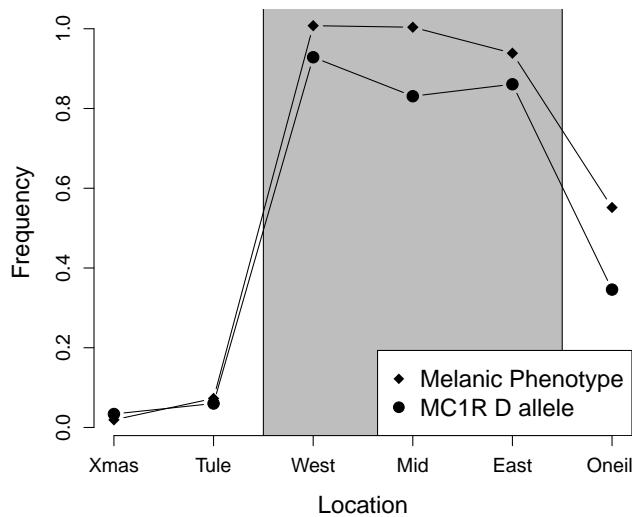


Figure 6.27: Frequency of melanic mice on the lava flow, and at nearby locations (diamonds). Frequency of MC1R melanic allele at same locations. Data from HOEKSTRA *et al.* (2004). Code here.

Many populations of small animals that live on these flows have evolved darker pigmentation to be cryptic against this dark substrate and better avoid visual predators. One example of such a locally adapted population are the rock pocket mice (*Chaetodipus intermedius*) who live on the Pinacate lava flow on the Arizona-Mexico border, studied by HOEKSTRA *et al.* (2004). These mice have much darker, more melanic pelts than the mice who live on nearby rocky outcrops (see Figure 6.27). NACHMAN *et al.* (2003) determined that a dominant allele (*D*) at MC1R is the primary determinant of this melanic phenotype. The frequency of this allele across study sites is shown in Figure 6.27. HOEKSTRA *et al.* (2004) found that other, unlinked markers showed little differentiation over these populations, suggesting that the migration rate is high.

**Question 13.** HOEKSTRA *et al.* (2004) found that 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?

4274

*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 land-



Figure 6.28: Two species from the genus *Chaetodipus*, pocket mice, formerly 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.

scape 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  $\sigma$  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  $\sigma$  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 6.29). 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

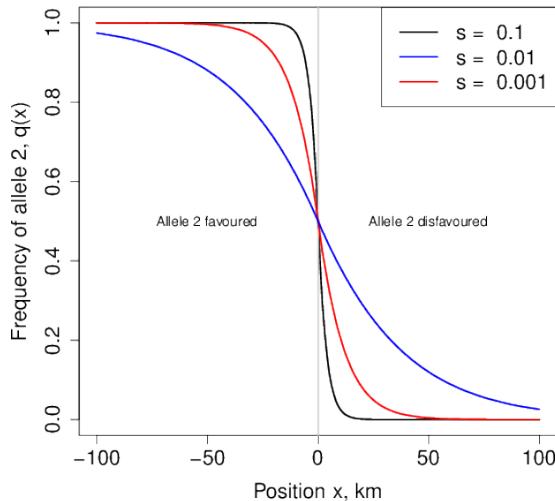


Figure 6.29: An equilibrium cline in allele frequency (the frequency of allele 2,  $q(\cdot)$ ) is shown). Our individuals disperse an average distance of  $\sigma = 1$  miles per generation, and our allele 2 has a relative fitness of  $1 + s$  and  $1 - s$  on either side of the environmental change 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,  $q(\cdot)$ , is shown as a function of location along the cline for a variety of selection coefficients ( $s$ ) in Figure 6.29. 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 6.30. Under this definition, the cline width is approximately

$$0.6\sigma/\sqrt{s} \text{ miles}, \quad (6.54)$$

note that the units are miles here just because we defined the average dispersal distance ( $\sigma$ ) in miles above. Thus the cline will be wider if individuals disperse further, higher  $\sigma$ , and if selection is weaker, smaller  $s$ . The appendix below talks through the math underlying these ideas in more detail.

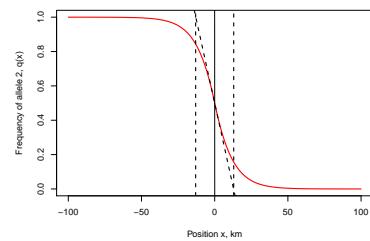
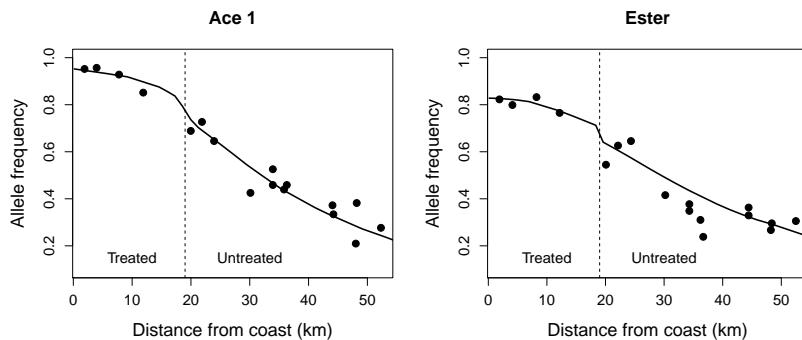


Figure 6.30: An equilibrium cline in allele frequency from Figure 6.29,  $s = 0.01$ . Vertical lines show the cline width. The diagonal line show the tangent to the cline at its midpoint. Code here.

Figure 6.31: 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 LENORMAND *et al.* (1999). Code here.

LENORMAND *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 mosquitoes have evolved resistance, but areas just a few tens of kilometers from the coast were untreated. LENORMAND *et al.* estimated the frequency of two unlinked, pesticide-resistance alleles, and found them at high frequency near the coast but found that their frequencies declined rapidly moving inland. LENORMAND *et al.* fit migration-selection cline models to their data, similar to those in Figure 6.29, 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

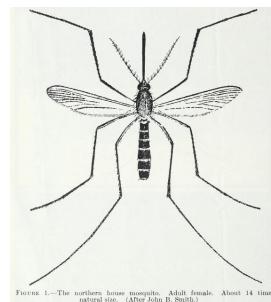


Figure 6.32: mosquito (*Culex pipiens*). Domestic mosquitoes (1939), Bishopp, F. C. Image from the Biodiversity Heritage Library. Contributed by U.S. Department of Agriculture, National Agricultural Library. Not in copyright.

higher cost to the Ace 1 allele than Ester allele in untreated areas  
 4334 ( $c = 0.11$  and  $c = 0.7$  respectively) potentially explaining the less  
 extreme cline for Ester allele than the Ace 1 allele. Despite these  
 4336 strong selection pressures we still see a cline over tens of kilometers  
 because dispersal is relatively high ( $\sigma = 6.6\text{km}$  per generation).

4338 *Appendix: Some theory of the spatial distribution of allele frequencies under deterministic models of selection*

4340 Imagine a continuous haploid population spread out along a line. Each  
 individual disperses a random distance  $\Delta x$  from its birthplace to the  
 4342 location where it reproduces, where  $\Delta x$  is drawn from the probability  
 density  $g(\cdot)$ . To make life simple, we will assume that  $g(\Delta x)$  is  
 4344 normally distributed with mean zero and standard deviation  $\sigma$ , i.e.  
 migration is unbiased and individuals migrate an average distance of  
 4346  $\sigma$ .

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 our allele frequency change in the next generation? Our allele frequency in  
 4348 the next generation at location  $x$  reflects the migration from different locations in the proceeding generation. Our population at location  $x$   
 4350 receives a contribution  $g(\Delta x)q(x + \Delta x, t)$  of allele 2 from the population at location  $x + \Delta x$ , such that the frequency of our allele at  $x$  in  
 4352 the next generation is  
 4354

$$q(x, t+1) = \int_{-\infty}^{\infty} g(\Delta x)q(x + \Delta x, t)d\Delta x. \quad (6.55)$$

To obtain  $q(x + \Delta x, t)$ , let's take a Taylor series expansion of  $q(x, t)$ :

$$q(x + \Delta x, t) = q(x, t) + \Delta x \frac{dq(x, t)}{dx} + \frac{1}{2}(\Delta x)^2 \frac{d^2q(x, t)}{dx^2} + \dots \quad (6.56)$$

4356 then

$$q(x, t+1) = q(x, t) + \left( \int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x \right) \frac{dq(x, t)}{dx} + \frac{1}{2} \left( \int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x \right) \frac{d^2q(x, t)}{dx^2} + \dots \quad (6.57)$$

Because  $g(\cdot)$  has a mean of zero,  $\int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x = 0$ , and has

4358 because  $g(\cdot)$  has variance  $\sigma^2$ ,  $\int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x = \sigma^2$ . All higher  
 order terms in our Taylor series expansion cancel out (as all high  
 4360 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

$$\Delta q(x, t) = \frac{\sigma^2}{2} \frac{d^2q(x, t)}{dx^2} \quad (6.58)$$

4362 This is a diffusion equation, so that migration is acting to smooth  
 4364 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  
 4366 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).

4368 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  $1 + s\gamma(x)$ .

4370 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  
 4372 allele 2 obtained within a generation due to selection is

$$q'(x, t) - q(x, t) \approx s\gamma(x)q(x, t)(1 - q(x, t)) \quad (6.59)$$

i.e. logistic growth of our favoured allele at location  $x$ . Putting our  
 4374 selection and migration terms together, we find the total change in  
 allele frequency at location  $x$  in one generation is

$$q(x, t+1) - q(x, t) = s\gamma(x)q(x, t)(1 - q(x, t)) + \frac{\sigma^2}{2} \frac{d^2q(x, t)}{dx^2} \quad (6.60)$$

4376 In deriving this result, we have essentially assumed that migration  
 acted upon our original allele frequencies before selection, and in doing  
 4378 so have ignored terms of the order of  $\sigma s$ .

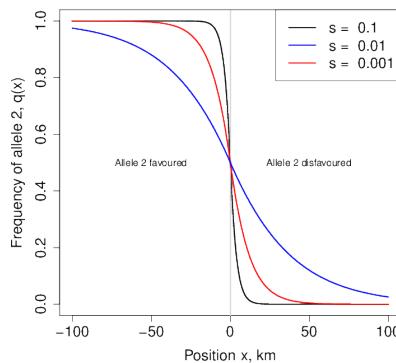


Figure 6.33: An equilibrium cline in allele frequency. Our individuals disperse an average distance of  $\sigma = 1$  km per generation, and our allele 2 has a relative fitness of  $1 + s$  and  $1 - s$  on either side of the environmental change at  $x = 0$ .

4380 The cline in allele frequency associated with a sharp environmental  
 transition. To make progress, let's consider a simple model of local  
 4382 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  
 4384 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

4386 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,  
 4388 the frequency of our allele 2 decreases in a smooth cline.

4390 Our equilibrium spatial distribution of allele frequencies can be  
 found by setting the left-hand side of eqn. (6.60) to zero to arrive at

$$s\gamma(x)q(x)(1-q(x)) = -\frac{\sigma^2}{2} \frac{d^2q(x)}{dx^2} \quad (6.61)$$

We then could solve this differential equation with appropriate bound-  
 4392 ary 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  
 4394 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  
 4396 equation. This compound parameter  $\ell$  is the characteristic length of  
 our cline, and it is this parameter which determines over what geo-  
 4398 graphic scale we change from allele 2 predominating to allele 1 pre-  
 dominatiing as we move across our environmental shift.

## *The Impact of Genetic Drift on Selected Alleles*

4402 “Natural selection is a mechanism for generating an exceedingly high  
degree of improbability.” –R.A. Fisher

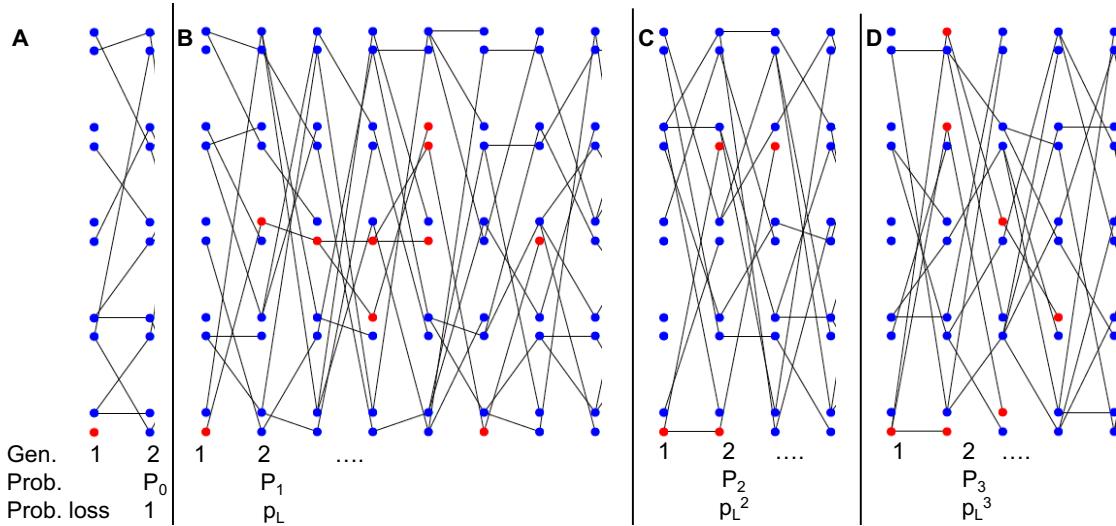
4404 In the previous chapter we assumed that the selection acting on our  
alleles was strong enough that we could ignore the action of genetic  
4406 drift in shaping allele frequencies. However, genetic drift affects all al-  
leles, and so in this chapter we explore the interaction of selection and  
4408 drift. Strongly selected alleles can be lost from the population via drift  
when they are rare in the population, while both weakly beneficial and  
4410 weakly deleterious alleles are subject to the random whims of genetic  
drift throughout their entire time in the population. Understanding  
4412 the interaction of selection and genetic drift is key to understand-  
ing the extent to which small populations may be mutation-limited  
4414 in their rates of adaptation, and how rates of molecular and genome  
evolution may differ across taxa.

4416 *7.1 Stochastic loss of strongly selected alleles*

Even strongly beneficial alleles can be lost from the population when  
4418 they are sufficiently rare. This is because the number of offspring left  
by individuals to the next generation is fundamentally stochastic. A  
4420 selection coefficient of  $s=1\%$  is a strong selection coefficient, which can  
drive an allele through the population in a few hundred generations  
4422 once the allele is established. However, if individuals have on average a  
small number of offspring per generation, the first individual to carry  
4424 our beneficial allele, who has on average 1% more children than their  
peers, could easily have zero offspring, leading to the loss of our allele  
4426 before it ever gets a chance to spread.

To take a first stab at this problem, let’s think of a very large hap-  
4428 loid population in which a single individual starts with the selected  
allele, and ask about the probability of eventual loss of our selected  
4430 allele starting from this single copy. To derive this probability of loss  
( $p_L$ ), we’ll make use of a simple argument (derived from branching

<sup>4432</sup> processes FISHER, 1923; HALDANE, 1927). Our selected allele will be eventually lost from the population if every individual with the allele fails to leave descendants. Well we can think about different cases:

<sup>4434</sup>

1. In our first generation, with probability  $P_0$  our individual allele leaves no copies of itself to the next generation, in which case our allele is lost (Figure 7.1A).
- <sup>4436</sup> 2. Alternatively, our allele could leave one copy of itself to the next generation (with probability  $P_1$ ), in which case with probability  $p_L$  this copy eventually goes extinct (Figure 7.1B).
- <sup>4440</sup> 3. Our allele could leave two copies of itself to the next generation (with probability  $P_2$ ), in which case with probability  $p_L^2$  both of these copies eventually go extinct (Figure 7.1C).
- <sup>4442</sup> 4. More generally, our allele could leave could leave  $k$  copies ( $k > 0$ ) of itself to the next generation (with probability  $P_k$ ), in which case with probability  $p_L^k$  all of these copies eventually go extinct (e.g. Figure 7.1D).

<sup>4448</sup> Summing over these probabilities, we see that

$$p_L = \sum_{k=0}^{\infty} P_k p_L^k \quad (7.1)$$

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

Figure 7.1: 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.

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

$$P_i = \frac{(1+s)^i e^{-(1+s)}}{i!} \quad (7.2)$$

Substituting  $P_k$  into the equation above, we see

$$\begin{aligned} p_L &= \sum_{k=0}^{\infty} \frac{(1+s)^k e^{-(1+s)}}{k!} p_L^k \\ &= e^{-(1+s)} \left( \sum_{k=0}^{\infty} \frac{(p_L(1+s))^k}{k!} \right) \end{aligned} \quad (7.3)$$

The term in the brackets is itself an exponential expansion, so we can rewrite this equation as

$$p_L = e^{(1+s)(p_L - 1)} \quad (7.4)$$

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. (7.4) as

$$1 - p_F = e^{-p_F(1+s)} \quad (7.5)$$

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  $p_F^2$ ):

$$1 - p_F \approx 1 - p_F(1 + s) + p_F^2(1 + s)^2/2 \quad (7.6)$$

Solving this we find that

$$p_F = 2s. \quad (7.7)$$

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.

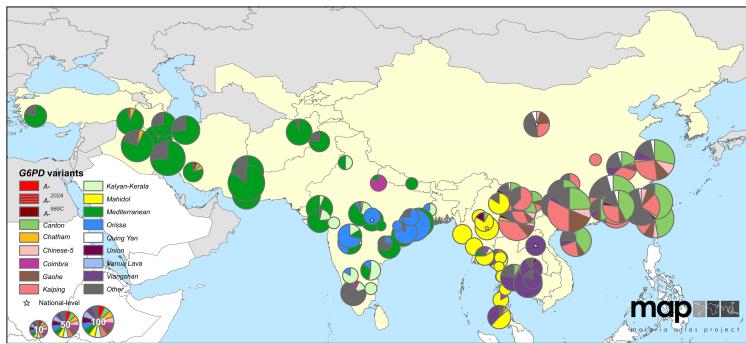
If the mutation rate towards our advantageous allele is  $\mu$ , and there are  $N$  individuals in our haploid population, then  $N\mu$  advantageous mutations arise per generation. Each of these new beneficial mutations has a probability  $p_F$  of fixing. Thus the number of advantageous mutations 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  
 4480 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 un-  
 4482 likely that the population will manage to fix a second drug-resistant  
 allele.

4484 *Diploid model of stochastic loss of strongly selected alleles.* We can  
 also adapt this result to a diploid setting. Assuming that heterozy-  
 4486 gotes 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

$$p_F = 2hs \quad (7.8)$$

4488 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  
 4490 positively selected allele, with  $h = 1$  corresponding to the full se-  
 lective advantage expressed in an individual with only a single copy.  
 4492 Thus the probability that a beneficial allele is not lost depends just  
 on the relative fitness advantage of the heterozygote; this is because  
 4494 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 indi-  
 4496 viduals compared to homozygotes for the ancestral allele (assuming an  
 outbred population).



**Figure 7.2: 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 HOWES *et al.* (2013), licensed under CC BY 4.0.

4498 Over roughly the past ten thousand years, adaptive alleles con-  
 ferring resistance to malaria have arisen in a number of genes and  
 4500 spread through human populations in areas where malaria is en-  
 demic (KWIATKOWSKI, 2005). One particularly impressive case of  
 4502 convergent evolution in response to selection pressures imposed by  
 malaria are the numerous changes throughout the G6PD gene, which  
 4504 include at least 15 common variants in Central and Eastern Asia  
 alone that lower the activity of the enzyme (HOWES *et al.*, 2013).  
 4506 These alleles are now found at a combined frequency of around 8%  
 frequency in malaria endemic areas, rarely exceeding 20% (HOWES

4508 *et al.*, 2012). Whether these variants *all* confer resistance to malaria  
 4510 is unknown, but a number of these alleles have demonstrated effects  
 4512 against malaria and are thought to have a selective advantage to heterozygotes  $sh > 5\%$  where malaria is endemic (RUWENDE *et al.*,  
 1995; TISHKOFF *et al.*, 2001; LOUICHAROEN *et al.*, 2009).

With a 5% advantage in heterozygotes, a G6PD allele present  
 4514 as a single copy would only have a 10% probability of fixing in the  
 4516 population. If that's so, how come malaria adaptation has repeatedly  
 4518 occurred via changes at G6PD? Well, maybe adaptation didn't start from a single copy of the selected allele. How many copies of the  
 4520 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  
 4522 carriers suffering from G6PD deficiency, leading to hemolytic anemia  
 4524 when individuals are exposed to a variety of different compounds,  
 notably those present in fava beans. There's upward of one hundred  
 4526 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  
 4528 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  
 4530 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,  
 4532 roughly five to ten thousand years ago that means that there  
 4534 would have been forty to four hundred copies of the G6PD-deficiency  
 4536 allele present in the population when selection pressures shifted at  
 the introduction of malaria. The chance that one of these newly  
 4538 adaptive alleles is lost is 90% but the chance that they're all lost is  
 4540  $< (0.9)^{40} \approx 0.02$ , i.e. there would have been a greater than 98% chance  
 4542 that adaptation would occur via one or more alleles at G6PD. How  
 4544 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  
 4546 mutations that contribute to adaptation are likely to be dominant or  
 at least co-dominant.

**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

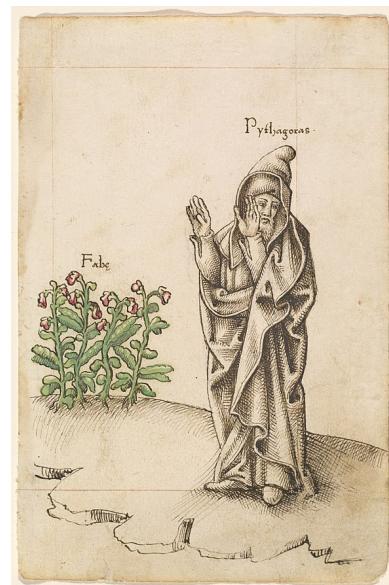


Figure 7.3: Pythagoras’s “just say no to fava beans” campaign. Pythagoras prohibited the consumption of fava beans by his followers; perhaps because favism, the anemia induced in G6PD-deficient individuals by fava beans, is relatively common in the Mediterranean due to adaptation to endemic malaria. French early 16th Century. Woodner Collection, National Gallery of Art. Public Domain, wikimedia.

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 (RALPH and COOP, 2015).



Figure 7.4: 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, wikimedia.

4552 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.  
4554 selfing, populations. Why is this?

**Question 2.** Melanic squirrels suffer a higher rate of predation  
4556 (due to hawks) than normally pigmented squirrels. Melanism is due to  
a dominant, autosomal mutation. The frequency of melanic squirrels  
4558 at birth is  $4 \times 10^{-5}$ .

A) If the mutation rate to new melanic alleles is  $10^{-6}$ , assuming  
4560 the melanic allele is at mutation-selection equilibrium, what is the  
reduction in fitness of the heterozygote?

4562 Suddenly levels of pollution increase dramatically in our population,  
and predation by hawks now offers an equal (and opposite) advantage  
4564 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  
4566 just once in the population) is lost?

C) If the population size of our squirrels is a million individuals,  
4568 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?

## 4572 7.2 The interaction between genetic drift and weak selection.

For strongly selected alleles, once the allele has escaped initial loss at  
4574 low frequencies, its path will be determined deterministically by its  
selection coefficients. However, if selection is weak compared to genetic  
4576 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  
4578 is sufficiently weak compared to genetic drift, then genetic drift will  
dominate the dynamics of alleles and they will behave like they're  
4580 effectively neutral. Thus, the extent to which selection can shape  
patterns of molecular evolution will depend on the relative strengths  
4582 of selection and genetic drift. But how weak must selection on an  
allele be for drift to overpower selection? And do these interactions  
4584 between selection and drift have longterm consequences for genome-  
wide patterns evolution?

To model selection and drift each generation, we can first calculate  
4586 the deterministic change in our allele frequency due to selection using  
our deterministic formula. Then, using our newly calculated expected  
4588 allele frequency, we can binomially sample two alleles for each of our  
offspring to construct the next generation. This approach to jointly  
4590 modeling genetic drift and selection is called the Wright-Fisher model.

4592 Under the Wright-Fisher model, we will calculate the expected



Figure 7.5: 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.

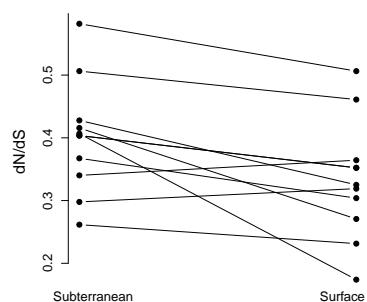


Figure 7.6: Asellid isopods have repeatedly invaded subterranean, ground-water habitats from surface-water habitats, and leading to a genome-wide increase in  $dN/dS$  and larger genomes (Data from LEFÉBURE *et al.*, 2017, comparing independent isopod species pairs). One possible explanation of this is that the longterm effective population sizes of the subterranean species are lower and so these species are less able to prevent mildly deleterious

change in allele frequency due to selection and the variance around  
 4594 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  $\bar{w} \approx 1$ .  
 4596 Using our directional selection deterministic model, from Chapter 6,  
 and these approximations gives us our deterministic change due to  
 4598 selection

$$\Delta_{sp} = \mathbb{E}(\Delta p) = \frac{s}{2}p(1-p) \quad (7.9)$$

To obtain our new frequency in the next generation,  $p_1$ , we binomially  
 4600 sample from our new deterministic frequency  $p' = p + \Delta_{sp}$ , so the  
 variance in our allele frequency change from one generation to the  
 4602 next is given by

$$Var(\Delta p) = Var(p_1 - p) = Var(p_1) = \frac{p'(1-p')}{2N} \approx \frac{p(1-p)}{2N}. \quad (7.10)$$

where the previous allele frequency  $p$  drops out because it is a constant  
 4604 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 we  
 4608 can simply look at when our change in allele frequency caused by selection within a generation is reasonably faithfully passed down through  
 4610 the generations. In particular, if our expected change in allele frequency is much greater than the variance around this change, genetic  
 4612 drift will play little role in the fate of our selected allele (once the allele is not at low copy number within the population). 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  
 4614 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 selection coefficients of the order  $1/N$  requires more careful modeling.  
 4620 However, under our diploid model, with an additive selection coefficient  $s$ , we can obtain the probability that allele 1 fixes within the  
 4622 population, starting from a frequency  $p$ :

$$p_F(p) = \frac{1 - e^{-2Ns}}{1 - e^{-2Ns}} \quad (7.11)$$

The proof of this result is sketched out below (see Section 7.2.1). A  
 4626 new allele that arrives in the population at frequency  $p = 1/(2N)$  has a probability of reaching fixation of

$$p_F\left(\frac{1}{2N}\right) = \frac{1 - e^{-s}}{1 - e^{-2Ns}} \quad (7.12)$$

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

$$Var(\Delta p) = Var(p' - p) \approx p(1-p)/2N.$$

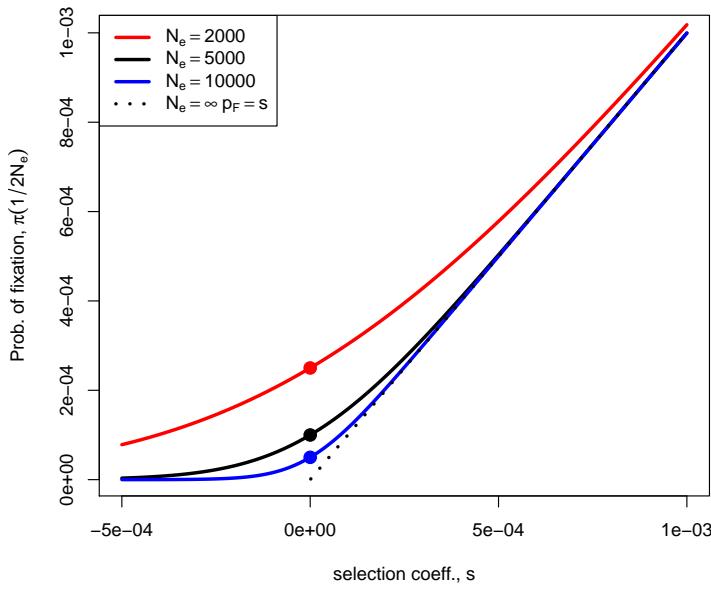


Figure 7.7: The probability of the fixation of a new mutation with selection coefficient  $s$  ( $h = 1/2$ ) in a diploid population of effective size  $N_e$ . The dashed line gives the infinite population solution. The dots give the solution for  $s \rightarrow 0$ , i.e. the neutral case, where the probability of fixation is  $1/(2N_e)$ . Code here.

- 4628 If  $s \ll 1$  but  $Ns \gg 1$  then  $p_F(\frac{1}{2N}) \approx s$ , which nicely gives us back the  
 result that we obtained above for an allele under strong selection (eqn.  
 4630 (7.8)). Our probability of fixation (eqn. (7.12)) is plotted as a function  
 of  $s$  and  $N$  in Figure 7.8. To recover our neutral result, we can take  
 4632 the limit  $s \rightarrow 0$  to obtain our neutral fixation probability,  $1/(2N)$ .

In the case where  $Ns$  is close to 1, then

$$p_F\left(\frac{1}{2N}\right) \approx \frac{s}{1 - e^{-2Ns}} \quad (7.13)$$

- 4634 This is greater than our earlier result  $p_F = s$  from the branching  
 process argument (using our additive model of  $h = 1/2$ ), increasingly  
 4636 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.  
 4638 So to persist indefinitely, the allele has to escape loss permanently,  
 by never being absorbed by the zero state. When the population size  
 4640 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  
 4642 probability, as they only have to reach  $2N$  to never be lost.

- 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

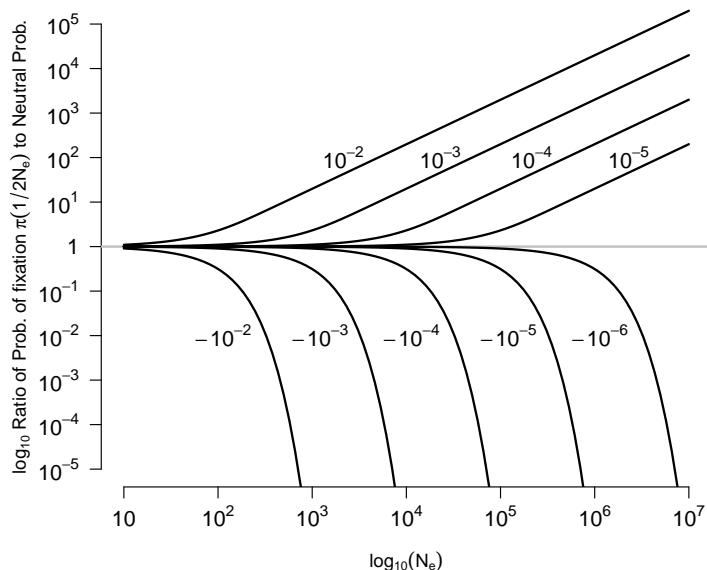


Figure 7.8: The probability of the fixation of a new mutation with selection coefficient  $s$  relative to the neutral fixation probability ( $1/2N_e$ ) as a function of the effective size  $N_e$ . The selection coefficient is shown next to the line. Note how quickly the probabilities move away from the neutral expectation as  $N_e s$  moves passed 1. Code here.

terms of the number of offspring they leave to the next generation.

4650 While we are incapable of detecting measuring all but the large fitness effect sizes, except in some elegant experiments (e.g. in microbes),  
 4652 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.

4654 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 differences in translation rates likely often correspond to tiny differences in fitness, but do they matter?

4666 In many organisms there is a strong bias in the codons to encode particular amino-acids, see Figure 7.9, with the most abundant codon matching the most abundant tRNA in cells. This 'codon bias' likely reflects the combined action of weak selection and mutational pressure, pushing the codon composition of the genome and tRNA abundances towards an adaptive compromise. These selection pressures have acted over long time periods, as codon usage patterns are often very simi-

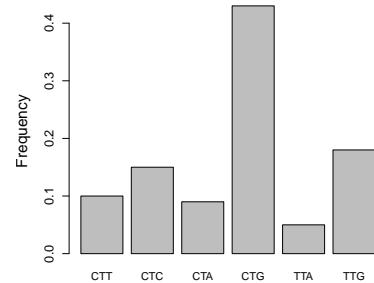


Figure 7.9: Data from *Drosophila melanogaster* on the frequency of different codons for Leucine. Data from Genscript. Code here.

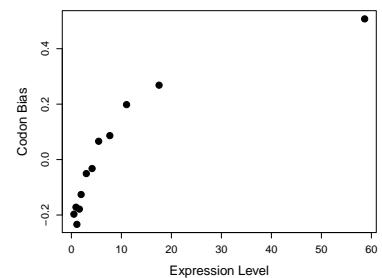


Figure 7.10: A measure of unequal codon frequencies ( $F$ ) plotted in bins of gene expression ( $E$ ) for genes across the *Drosophila melanogaster* genome. Data from HEY and KLIMAN (2002). Code here.

4672 lar for species that diverged over many tens of millions of years ago.  
 Compared to other genes, highly expressed genes show a strong bias  
 4674 towards using codons matching abundant tRNAs, consistent with the  
 idea that the synonymous codon content of highly expressed genes  
 4676 is evolving to optimize their translation (see Figure 7.10 for an early  
 example). These patterns likely represent the action of selection pres-  
 4678 sures that are incredibly weak on average, but that have played out  
 over vast time-periods.

4680 *The fixation of slightly deleterious alleles.* From Figure 7.8 we can  
 see that weakly deleterious alleles can also fix, especially in small  
 4682 populations. To understand how likely it is that deleterious alleles by  
 chance reach fixation by genetic drift, let's assume a diploid model  
 4684 with additive selection (with a selection coefficient of  $-s$  against our  
 allele 2).

4686 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

$$p_F \left( \frac{1}{2N} \right) \approx \frac{s}{e^{2Ns} - 1} \quad (7.14)$$

4688 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  
 4690 above, this is because while deleterious mutations will never escape  
 loss in infinite populations, they can become fixed in finite population  
 4692 by reaching  $2N$  copies.

**Question 3.** An additive mutation arises that lowers the relative  
 4694 fitness of heterozygotes by  $10^{-5}$ . What is the probability that this  
 mutation fixes in a diploid population with effective size of  $10^4$ ? What  
 4696 is the probability it fixes in a population of effective size  $10^6$ ? By  
 comparing both to their neutral probability describe the intuition  
 4698 behind this result.

OHTA proposed the ‘nearly-neutral’ theory of molecular evolu-  
 4700 tion in a series of papers<sup>1</sup>. She suggested that a reasonable fraction  
 of newly arising functional mutations may have very weak selection  
 4702 coefficients, such that species with smaller effective population sizes  
 may have higher rates of fixation of these very weakly deleterious al-  
 4704 leles. In effect, her suggestion is that the constraint parameter  $C$  of  
 a functional region is not a fixed property, but rather depends on the  
 4706 ability of the population to resist the influx of very weakly deleterious  
 mutations.

<sup>1</sup> OHTA, T., 1972 Population size and rate of evolution. *Journal of Molecular Evolution* 1(4): 305–314; OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428): 96; and OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution* 26(1-2): 1–6

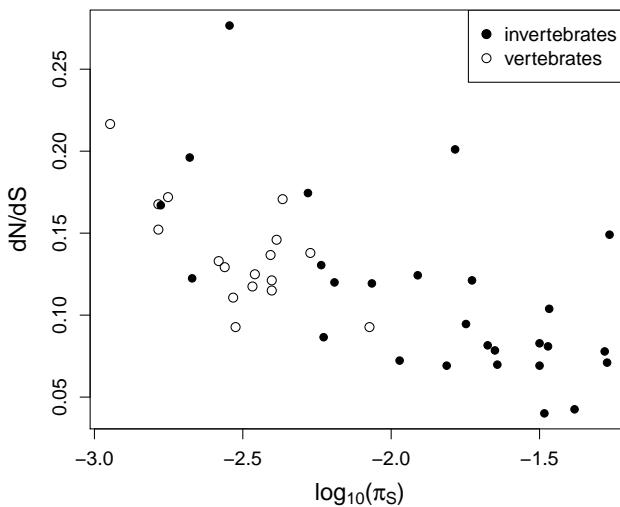


Figure 7.11: Data from 44 metazoan species from Cuttlefish to Sifakas. Each dot represents the average of over many genes plotting  $d_N/d_S$  against synonymous diversity ( $\pi_S$ ). Data from GALTIER (2016). Code here.

4708 Across species, genome-wide averages of  $d_N/d_S$  do seem to be correlated with measures of the effective population size (such as synonymous diversity), see Figure 7.11. This evidence supports the idea that 4710 in species with smaller effective population sizes (lower  $\pi_S$ ), proteins 4712 may be subject to lower degrees of constraint, as very weakly deleterious mutations are able to fix. Thus, some reasonable proportion of 4714 functional substitutions in populations with small effective population sizes, such as humans, may be mildly deleterious.

#### 4716 7.2.1 Appendix: The fixation probability of weakly selected alleles

What is the probability a weakly beneficial or deleterious additive 4718 allele fixes in our population? We'll let  $P(\Delta p)$  be the probability that our allele frequency shifts by  $\Delta p$  in the next generation. Using this, we 4720 can write our probability  $p_F(p)$  in terms of the probability of achieving fixation averaged over the frequency in the next generation

$$p_F(p) = \int p_F(p + \Delta p)P(\Delta p)d(\Delta p) \quad (7.15)$$

4722 This is very similar to the technique that we used when deriving our probability of escaping loss in a very large population above.

4724 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  $\Delta p$  is small:

$$p_F(p + \Delta p) \approx p_F(p) + \Delta p \frac{dp_F(p)}{dp} + (\Delta p)^2 \frac{d^2 p_F(p)}{dp^2}(p) \quad (7.16)$$



Figure 7.12: 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.

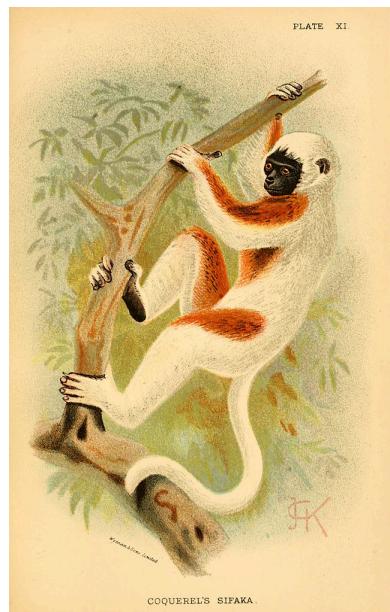


Figure 7.13: Coquerel's Sifaka (*Propithecus coquereli*). A hand-book to the primates (1894). Forbes, H. O. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

<sup>4726</sup> ignoring higher order terms.

Taking the expectation over  $\Delta p$  on both sides, as in eqn. 7.15, we  
<sup>4728</sup> obtain

$$p_F(p) = p_F(p) + \mathbb{E}(\Delta p) \frac{dp_F(p)}{dp} + \mathbb{E}((\Delta p)^2) \frac{d^2 p_F(p)}{dp^2} \quad (7.17)$$

Well,  $\mathbb{E}(\Delta p) = \frac{s}{2}p(1-p)$  and  $Var(\Delta p) = \mathbb{E}((\Delta p)^2) - \mathbb{E}^2(\Delta p)$ , so if  
<sup>4730</sup>  $s \ll 1$  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

<sup>4732</sup> us with

$$0 = \frac{s}{2}p(1-p) \frac{dp_F(p)}{dp} + \frac{p(1-p)}{2N} \frac{d^2 p_F(p)}{dp^2} \quad (7.18)$$

and we can specify the boundary conditions to be  $p_F(1) = 1$  and  
<sup>4734</sup>  $p_F(0) = 0$ . Solving this differential equation is a somewhat involved  
process, but in doing so we find that

$$p_F(p) = \frac{1 - e^{-2Ns p}}{1 - e^{-2Ns}} \quad (7.19)$$

<sup>4736</sup> This proof can be extended to alleles with arbitrary dominance, how-  
ever, this does not lead to a analytically tractable expression so we do  
<sup>4738</sup> not pursue this here.

# 8

## *4740 The Effects of Linked Selection.*

GENETIC DRIFT IS NOT THE ONLY SOURCE OF RANDOMNESS

4742 in the dynamics of alleles. Alleles also experience random fluctua-  
tions in frequency due to the fact that they present on a set of random  
4744 genetic backgrounds with different fitnesses. For example, when a  
beneficial allele arises via a single mutation, it arises on a particular  
4746 genetic background, i.e. a particular haplotype (Figure 8.1A). Imagine  
this mutation arising in a region with no recombination, or in an or-  
4748 ganism where genetic exchange is rare. If our beneficial allele becomes  
established in the population, i.e. escapes loss by genetic drift in those  
4750 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  
4752 on the haplotype that the mutation arose on (if those other alleles are  
neutral or at least not too deleterious). These other alleles are get-  
4754 ting to 'hitchhiking' along. The alleles that are not on that particular  
background are swept out of the population, so the net effect of this  
4756 selective sweep is to remove genetic diversity from the population. Di-  
versity will eventually recover, as new mutations arise and some slowly  
4758 drift up in frequency. But in the short-term, selective sweeps remove  
genetic variation from populations.

4760 WILLIAMS and PENNINGS (2019) have visualized selective  
sweeps in HIV. In Figure 8.1B) we see a set of HIV haplotypes sam-  
4762 pled from a patient before and after of a selective sweep of a drug-  
resistant mutation. The patient is taking a retrotransposase inhibitor  
4764 (Efavirenz), but sadly within 161 days a drug-resistant mutation that  
changes the HIV retrotransposase protein has arisen and spread. Note  
4766 how a particular haplotype is now fixed in the sample, and little ge-  
netic diversity remains, due to the hitchhiking effect of the strong  
4768 selective sweep of this allele.

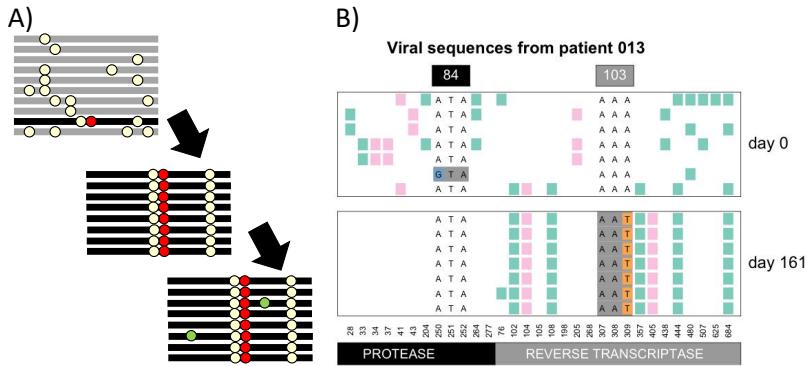


Figure 8.1: **A)** In the top panel, a selected mutation (red dot) arises on a particular haplotype in the population. It sweeps to fixation, carrying with it the haplotype on which it arose, middle panel, erasing the standing genetic diversity in the region. The bottom panel is some time after the selective 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 103<sup>rd</sup> codon of retrotransposase. Each row is a haplotype, with the alleles present shown as coloured blocks. Figure B from WILLIAMS and PENNINGS (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 8.6). This is because that background neutral allele, which existed  $\tau$  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. (6.35)). 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 recombination breaks up their association with the selected allele very

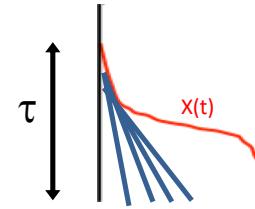


Figure 8.2: The coalescent of 4 lineages, marked in blue, at a locus completed linked to our selected allele. The frequency trajectory of the selected allele  $X(t)$  is shown in red.

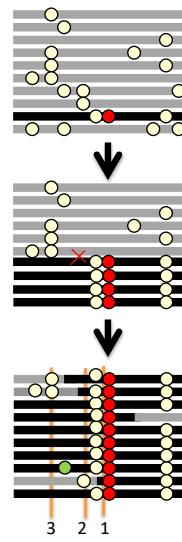


Figure 8.3: A cartoon depiction of a sweep of a red beneficial allele over three time points. The haplotype that the beneficial arose on by mutation is shown in black. The three vertical orange lines mark the loci shown in Figure 8.4. Neutral alleles segregating prior to the sweep appear as white circles, new mutations after the sweep as green circles.

rapidly.

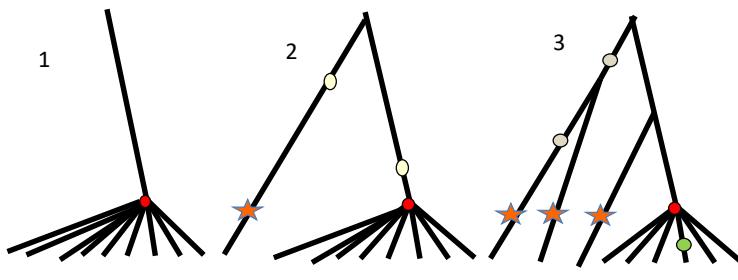


Figure 8.4: Coalescent genealogies at three loci different distances along the genome from a selective sweep. The locations of these three loci along the genome are marked in Figure 8.3. The selected mutation is shown in red. Lineages descended from recombination events during the sweep are marked in stars. Neutral mutations close to each of the loci are shown on the genealogy.

4800 What do the coalescent genealogies look like at loci various dis-  
 4802 tances away from the selected site? Well, close to the selected site all  
 4804 our alleles in the present day trace back to a most recent common an-  
 4806 cestral allele present on that selected haplotype, and so are all forced  
 4808 to coalesce around  $\tau$  generations ago (locus 1). Slightly further out  
 4810 from the selected site (locus 2), we have lineages that don't trace their  
 4812 ancestry back to the original selected haplotype, but instead are de-  
 4814 scended from recombinant haplotypes that recombined onto the sweep  
 (the haplotype 2 from the bottom). These lineages can coalesce neu-  
 trally with the other ancestral lineages over far deeper time scales and  
 mutations on these deeper lineages correspond to the standing diver-  
 sity 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  $\tau$ , allowing diversity to recover to  
 background levels as we move away from the selected site.

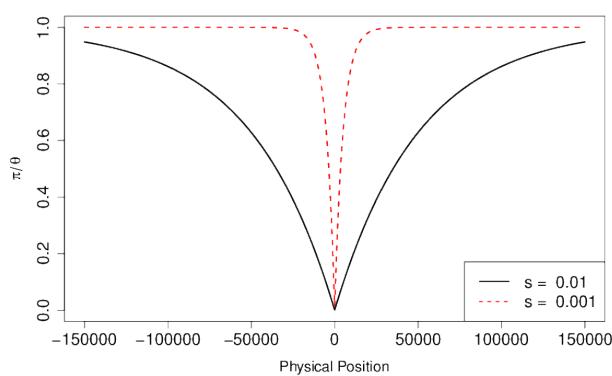


Figure 8.5: The expected reduction in diversity compared to its neutral expectation as a function of the distance away from a site where a selected allele has just gone to fixation. The sweeps associated with two different strengths of selection are shown, corresponding to a short timescale ( $\tau$ ) for the sweep and long one. The recombination rate is  $r_{BP} = 1 \times 10^{-8}$ . Code here.

4816 To model the expected pattern of diversity surrounding a selected  
 4818 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  
 4820 are descended from recombinant haplotypes.

We know that in the present day our neutral lineage is linked to the  
 4822 selected allele. The probability that our lineage, in some generation  
 4824  $t$  back in time, is in a heterozygote is  $1 - X(t)$ , and the probability  
 4826 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

$$r(1 - X(t)) \quad (8.1)$$

So the probability ( $p_{NR}$ ) that our lineage is not descended from a re-  
 4828 combinatorial haplotype from a recombination event in the  $\tau$  generations  
 it takes our selected allele to move through the population is

$$p_{NR} = \prod_{t=1}^{\tau} (1 - r(1 - X(t))) \quad (8.2)$$

4830 Assuming that  $r$  is small, then  $(1 - r(1 - X(t))) \approx e^{-r(1-X(t))}$ , such  
 that

$$p_{NR} = \prod_{t=1}^{\tau} (1 - r(1 - X(t))) \approx \exp \left( -r \sum_{t=1}^{\tau} 1 - X(t) \right) = \exp \left( -r\tau(1 - \hat{X}) \right) \quad (8.3)$$

4832 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  
 4834 our allele is additive, its trajectory for frequencies  $< 0.5$  is the mirror  
 4836 image of its trajectory for frequencies  $> 0.5$ , therefore its average  
 frequency  $\hat{X} = 0.5$ . This simplifies our expression to

$$p_{NR} = e^{-r\tau/2}. \quad (8.4)$$

The probability that neither of our lineages is descended from a re-  
 4838 combinatorial haplotype, and hence are forced to coalesce, is  $p_{NR}^2$  (as-  
 suming that they coalesce at a time close to  $\tau$  so that they recombine  
 4840 independently of each other for times  $< \tau$ ).

If one or other of our lineages is descended from a recombinant  
 4842 haplotype, it will take them on average  $\approx 2N$  generations to find a  
 4844 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

$$\mathbb{E}(T_2) = \tau \times p_{NR}^2 + (1 - p_{NR}^2)(\tau + 2N) \approx (1 - p_{NR}^2) 2N \quad (8.5)$$

4846 where this last approximation assumes that  $\tau \ll 2N$ . So the expected  
 4848 pairwise diversity for neutral alleles at a recombination distance  $r$   
 away from the selected sweep ( $\pi_r$ ) is

$$\mathbb{E}(\pi_r) = 2\mu\mathbb{E}(T_2) \approx \pi_0 (1 - e^{-r\tau}) \quad (8.6)$$

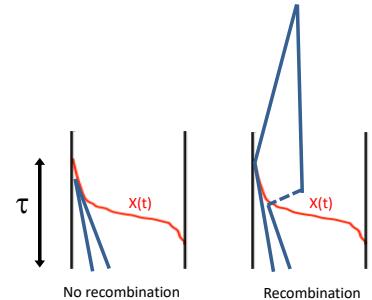
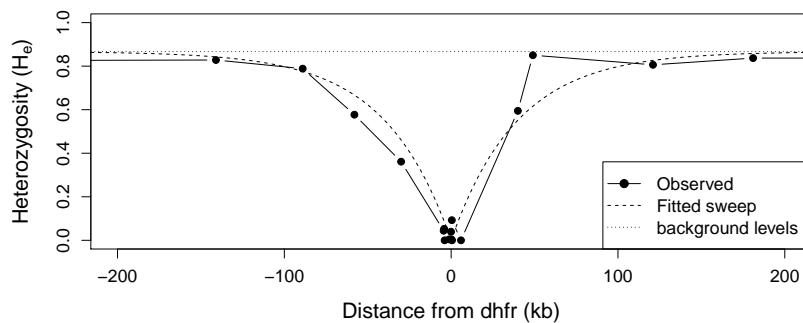


Figure 8.6:

So diversity increases as we move away from the selected site, slowly and exponentially plateauing to its neutral expectation  $\pi_0$ .

The malaria pathogen (*Plasmodium falciparum*) has evolved drug resistance to anti-malaria drugs, often by changes at the dhfr gene. Figure 8.8 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 (NASH *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  $\ell$  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 8.5 we show the reduction in diversity, given by eqn. (8.6), for two different selection coefficients.

For our expected diversity level to recover to 50% of its neutral expectation  $\mathbb{E}(\pi_r)/\theta = 0.5$ , requires a physical distance  $\ell^*$  such that  $\log(0.5) = -r_{BP}\ell^*\tau$ , and by re-arrangement,

$$\ell^* = \frac{-\log(0.5)}{r_{BP}\tau}. \quad (8.7)$$

As  $\tau$  depends inversely on the selection  $s$  (eqn. (6.35)), 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  $r_{BP} = 10^{-8}\text{bp}^{01}$  a selection coefficient of  $s = 0.1\%$  would reduce diversity over 10's of kb, while a sweep of  $s = 1\%$  would affect  $\sim 100\text{kb}$ .

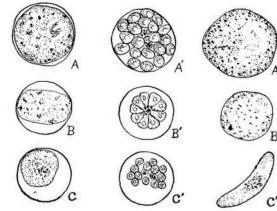


FIG. 47. Comparison of three species of malaria parasites  $\times 2000$  (figures selected largely from Manson). A, A' and A'', *Plasmodium vivax*; B, B' and B'', *Plasmodium vivax*; C, C' and C'', *Plasmodium falciparum*. A, B and C, mature parasites in red corpuscles. A', B' and C', segmented parasites ready to leave corpuscles. A'', B'' and C'', mature gametocytes.

A, B and C, mature parasites in red corpuscles.

A', B' and C', segmented parasites ready to leave corpuscles.

A'', B'' and C'', mature gametocytes.

Figure 8.7: 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 8.8: 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  $\pi_0$ . The dashed line shows our fitted hitchhiking model from equation 8.6 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 NASH *et al.* (2005). Code here.

**Question 1.** VAN'T 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 case of adaptation to industrial pollution (made famous by the work of KETTLEWELL, see MAJERUS (2009) and COOK *et al.* (2012)). The genetic basis of melanism is a transposable element (TE) inserted into a pigmentation gene. VAN'T HOF *et al.* found that diversity is suppressed in a broad region around the TE. Specifically, on the background of the TE, it takes roughly 200 kb in either direction for diversity levels to recover to 50% of genome-wide levels.

Random facts: In all moths and butterflies only males recombine; chromosomes are transmitted without recombination in females. The recombination rate in males is 2.9 cM/Mb. Peppered moths have an effective population size of roughly a hundred thousand individuals. Kettlewell used to eat moths when out collecting them in the field (personal communication, Art. Shapiro).

**A)** Briefly explain how this pattern offers further evidence that the melanic allele was favoured by selection.

**B)** Using this information, and assuming the allele's effects on fitness are additive, what is your estimate of the age of the allele?

**C)** What is your estimate of the selection coefficient favouring this melanic allele?

*Other signals of selective sweeps* The primary signal of a recently completed selective sweep is the characteristic reduction in diversity surrounding the selected site. However, sweeps do leave other signals and these have also often been used to identify loci undergoing selection. For example, neutral alleles further away from the selected site may hitchhiking only part of the way to fixation if recombination occurs during the sweep, which can lead to an excess of high-frequency derived alleles at intermediate distances away from the selected site, a pattern lasting for a short time after a sweep (FAY and WU, 2000; PRZEWORSKI, 2002; KIM, 2006). Also, as neutral diversity levels slowly recover through an influx of new mutations after a sweep, there is a strong skew towards low frequency derived alleles, a pattern that persists for many generations (BRAVERMAN *et al.*, 1995; PRZEWORSKI, 2002; KIM, 2006). The excess of rare alleles, compared to a neutral model, can be captured by statistics such as Tajima's D (which we encountered back in our discussion of the neutral site frequency eqn 3.43). Thus one way to look for loci that have undergone selective sweeps is to calculate Tajima's D from data in windows along the genome and look for strong departures from the null distribution.



Figure 8.9: 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.

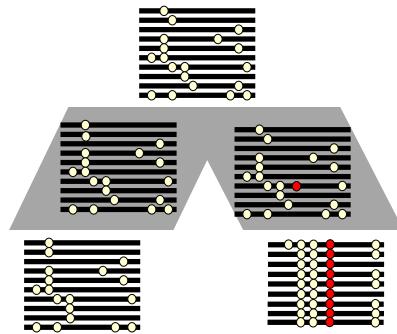


Figure 8.10: Two populations descended from a common ancestral population. A beneficial mutation has occurred in population and swept to fixation.

4920 We can also use comparisons among multiple populations to look  
 for evidence of sweeps occurring in one of the populations, for example  
 4922 to identify alleles involved in local adaptation (see 8.11). A selective  
 sweep will decrease the within-population diversity ( $H_S$ ) surrounding  
 4924 the selected site, without affecting the diversity between different  
 populations. Thus local sweeps create peaks of  $F_{ST}$  between weakly  
 4926 differentiated populations.

HOHENLOHE *et al.* (2010) studied genome-wide patterns of  $F_{ST}$   
 4928 between marine and freshwater populations of threespine stickleback  
 $(Gasterosteus aculeatus)$ . Between different marine populations, they  
 4930 found no strong peaks of  $F_{ST}$ ; however, between the marine and fresh-  
 water comparisons they found a number of high  $F_{ST}$  peaks that were  
 4932 replicated over a number of freshwater-marine comparisons. They  
 identified a number of novel regions responsible for the adaptation  
 4934 of sticklebacks to freshwater environments and also a number of loci  
 previously identified in crosses between marine and freshwater pop-  
 4936ulations. For example, the first peak of Linkage Group IV includes  
 Ectodysplasin A (Eda), a gene involved in the adaptive loss of armour  
 plating in freshwater environments.

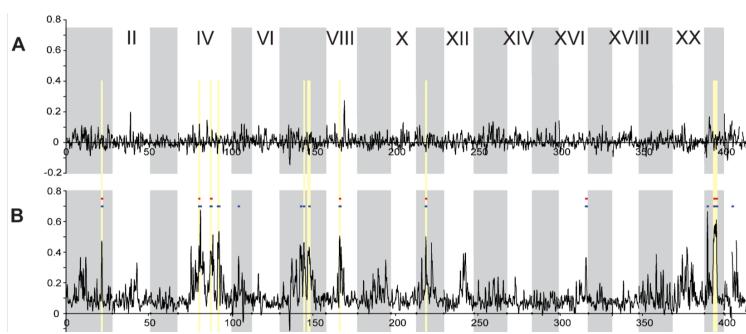


Figure 8.11:  $F_{ST}$  across the stickleback genome, with colored bars indicating significantly elevated ( $p \leq 10^{-5}$ , blue;  $p \leq 10^{-7}$ , red) and reduced ( $p \leq 10^{-5}$ , green) values. The alternating white and grey panels indicate different linkage groups. **A)**  $F_{ST}$  between two oceanic populations **B)** Average  $F_{ST}$  between a freshwater population and the two marine populations. Figure and caption text from HOHENLOHE *et al.* (2010), licensed under CC BY 4.0.

*Soft Sweeps from multiple mutations and standing variation.* In our sweep model above, we assumed that selection favoured a beneficial allele from the moment it entered the population as a single copy mutation (left panel, Figure 8.12). However, when a novel selection pressure switches on, multiple mutations at the same gene may start to sweep, such that no one of these alleles sweeps to fixation (middle panel, Figure 8.12). These sweeps involving multiple mutations significantly soften the impact of selection on genomic diversity, and so are called 'soft sweeps'.

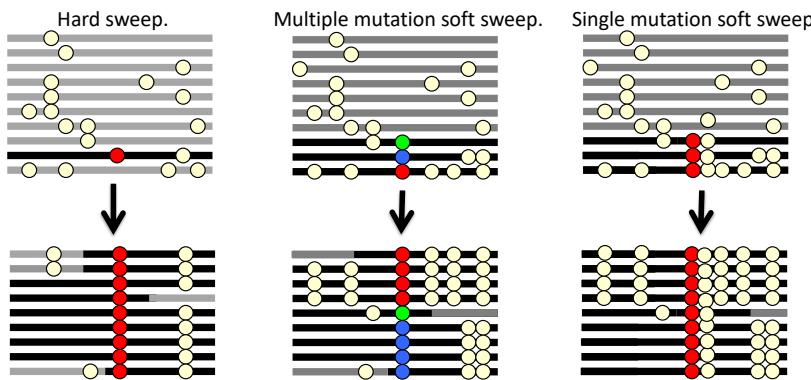


Figure 8.12: Three types of sweeps.

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 8.12). Detecting and differentiating these different types of sweeps is an active area of empirical research and theory in population genomics (see HERMISSON and PENNINGS (2017) for an overview of developments in this area).

### 8.1 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

4968 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 LEFFLER *et al.* (2012) for a recent review). As highlighted by LEWONTIN (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 MAYNARD SMITH and HAIGH (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 (KAPLAN *et al.*, 1989; GILLESPIE, 2000).

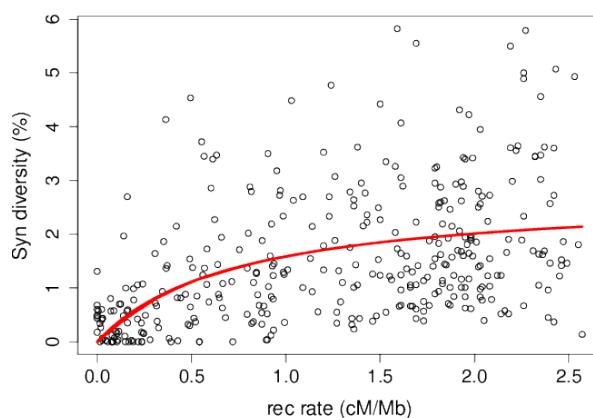


Figure 8.13: The relationship between (sex-averaged) recombination rate and synonymous site pairwise diversity ( $\pi$ ) in *Drosophila melanogaster*. The curve is the predicted relationship between  $\pi$  and recombination rate, obtained by fitting the recurrent hitchhiking equation (8.13) to this data using non-linear least squares via the `nls()` function in R. Data from (SHAPIRO *et al.*, 2007), kindly provided by Peter Andolfatto, see SELLA *et al.* (2009) for details. Code here.

4988 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 4990 species, as, all else being equal, linked selection should remove diversity more quickly in regions of low recombination (AGUADÉ *et al.*, 4992 1989; BEGUN and AQUADRO, 1992; WIEHE and STEPHAN, 1993b; CUTTER and CHOI, 2010; CAI *et al.*, 2009). For example, 4994 *Drosophila melanogaster* diversity levels are much lower in genomic regions of low recombination (see Figure 8.13). This pattern can not

4996 be explained by differences in mutation rate between low and high re-  
4998 combination regions as this pattern is not seen strongly in divergence  
 data among species.

5000 These patterns could reflect the action of selective sweeps happen-  
5002 ing recurrently along the genome. In the next section we'll present a  
5004 model for how levels of genetic diversity should depend on recombi-  
5006 nation and the density of functional sites under a model of recurrent  
5008 selective sweeps. However, other forms of linked selection can impact  
 genetic diversity in similar ways. For example, linked genetic diversity  
 is continuously lost from natural populations due to the removal of  
5006 haplotypes that carry deleterious alleles (CHARLESWORTH *et al.*,  
 1995; HUDSON and KAPLAN, 1995b); this is called the 'background  
5008 selection' model. Below we'll discuss the background selection model  
 and its basic predictions.

5010 More generally, a wide range of models of selection predict the  
 removal of neutral diversity linked to selected sites. This is because  
5012 the diversity-reducing effects of high variance in reproductive success  
 are compounded over the generations when there is heritable variance  
5014 in fitness (ROBERTSON, 1961; SANTIAGO and CABALLERO, 1995,  
 1998; BARTON, 2000). Many different modes of linked selection likely  
5016 contribute to these genome-wide patterns of diversity; the present  
 challenge is how to differentiate among these different modes.

### 5018 8.1.1 A simple recurrent model of selective sweeps

To explain how a constant influx of sweeps could impact levels of  
5020 diversity, here we will develop a model of recurrent selective sweeps.

5022 Imagine we sample a pair of neutral alleles at a locus a genetic  
5024 distance  $r$  away from a locus where sweeps are initiated within the  
 population at some very low rate  $\nu$  per generation. The waiting time  
5026 between sweeps at our locus is exponentially distributed  $\sim \text{Exp}(\nu)$ .  
 Each sweep rapidly transits through the population in  $\tau$  generations,  
5028 such that each sweep is finished long before the next sweep ( $\tau \ll 1/\nu$ ).

5028 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  
5030 a very low probability

$$\nu e^{-r\tau} \tag{8.8}$$

5032 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  
5034 between our neutral pair of lineages due to either a sweep or a neutral  
 coalescent event is

$$\mathbb{E}(T_2) = \frac{1}{\nu e^{-r\tau} + 1/2N} \tag{8.9}$$

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  $\nu_{BP}$  per basepair per generation. The rate of coalescence due to sweeps at a locus  $\ell$  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

$$2 \int_0^L \nu_{BP} e^{-r_{BP}\ell\tau} d\ell = \frac{2\nu_{BP}}{r_{BP}\tau} (1 - e^{-r_{BP}\tau L}) \quad (8.10)$$

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

$$\frac{2\nu_{BP}}{r_{BP}\tau} + \frac{1}{2N} \quad (8.11)$$

So our average time till a pair of lineages coalesce is

$$\mathbb{E}(T_2) = \frac{1}{2\nu_{BP}/r_{BP}\tau + 1/2N} = \frac{r_{BP}2N}{4N\nu_{BP}/\tau + r_{BP}} \quad (8.12)$$

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  $\nu_{BP}$  is

$$\mathbb{E}(\pi) = \pi_0 \frac{r_{BP}}{4N\nu_{BP}/\tau + r_{BP}} \quad (8.13)$$

where  $\pi_0$  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  $\nu_{BP}$ , 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  $\nu_{BP}$  doesn't increase with  $r_{BP}$ , our level of diversity should plateau to  $\theta$ , the level of genetic diversity of a neutral site completely unlinked to any selected loci. If we assume that our genome experiences a constant rate of sweeps of a given strength, i.e. that  $4N\nu_{BP}/\tau$  is a constant, we can fit the variation in  $\pi$  across regions that vary in their recombination rate ( $r_{BP}$ ) to estimate a population's rate of recurrent sweeps per basepair. An example of fitting this curve to data from *Drosophila melanogaster* is shown in Figure 8.13; see WIEHE and STEPHAN (1993a) for an early example of fitting a similar recurrent hitchhiking model to such data. The parameter giving us this best-fitting curve is  $4N\nu_{BP}/\tau \approx 7 \times 10^{-9}$ . With an effect population size of a million and assuming that the sweeps take a thousand 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 megabase every million generations, is all we need to explain the profound dip in diversity seen in regions of the genome with low recombination. However, sweeps from positively selected alleles are not the only cause of genome-wide signals of linked selection. Selection against deleterious alleles can also drive these patterns.

### 5076 8.1.2 Background selection

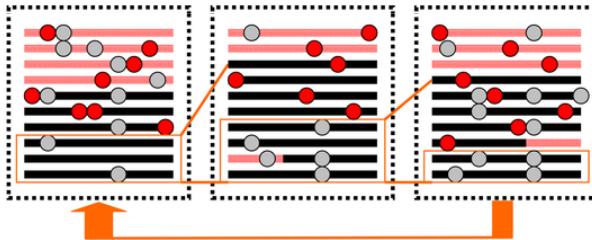
Populations experience a constant influx of deleterious mutations at functional loci while selection acts to purge them from the population, thus preventing deleterious substitutions and maintaining function at these loci. As we discussed in Chapter 6, this balance between mutation and selection results in a constant level of deleterious variation in the population. The constant selection against this deleterious variation has effects on diversity at linked sites. Each deleterious mutation arises at random on a haplotype in the population, and as selection purges this mutation, it removes with it any neutral alleles that were also on this haplotype. This constant removal of linked alleles from the population acts to reduce diversity in regions surrounding functional loci (HUDSON and KAPLAN, 1995a; NORDBORG *et al.*, 1996), an effect known as background selection (BGS).

What proportion of our haplotypes are free of deleterious mutations in any given generation, and so free to contribute to future generations? Well, under mutation-selection balance, a constrained locus with a mutation rate  $\mu$  towards deleterious alleles that experience a selection coefficient  $sh$  against them in heterozygotes, will result in  $\mu/sh$  chromosomes carrying the deleterious allele. Some of 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

$$\mathbb{E}[\pi] = 2\mu \times 2N(1 - \mu/sh) = \pi_0(1 - \mu/sh) \quad (8.14)$$

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.



5110 For a neutral locus that is a recombination fraction  $r$  away from a locus subject to constraint, the level of diversity is

$$\mathbb{E}[\pi] = \pi_0 \left(1 - \frac{\mu sh}{2(r+sh)^2}\right) \quad (8.15)$$

5112 As we move away from a locus experiencing purifying selection, we increase  $r$ , and diversity should recover. For example, moving away 5114 from genic regions in the maize genome we see the average level of diversity recover. This occurs in both maize and teosinte, the wild 5116 progenitor of maize. The dip in diversity around non-synonymous sites is stronger in teosinte, perhaps because the accelerated drift due to 5118 the bottleneck in maize may have somewhat released constraint on sites where very weakly deleterious alleles segregated previously at 5120 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 5122 compounding equation (8.16) across these loci, the expected reduced 5124 diversity is approximately

$$\mathbb{E}[\pi] = \pi_0 \prod_{i=1}^L \left(1 - \frac{\mu sh}{2(r_i+sh)^2}\right) \approx \exp\left(\sum_{i=1}^L \frac{\mu sh}{2(r_i+sh)^2}\right) \quad (8.16)$$

To model an average neutral locus in a genomic region with a given 5126 recombination rate, we can imagine that our neutral locus is situated in the center of a large region with total recombination rate  $R$  and 5128 total deleterious mutation rate  $U$ , where  $U = \mu L$ . Then our expression for diversity, equation (8.16), simplifies to

$$\mathbb{E}[\pi] \approx \pi_0 \exp(-U/(sh+R)) \approx \pi_0 \exp(-U/R). \quad (8.17)$$

5130 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 5132 (6.49), this expression depends only on the total deleterious mutation rate. Any dependence on the selection coefficient drops out, as weakly 5134 selected mutations segregate in the population at higher frequencies, but are also removed from the population more slowly, allowing more 5136 of the genome to recombine off the deleterious background.

Figure 8.14: 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 SELLA *et al.* (2009), licensed under CC BY 4.0.

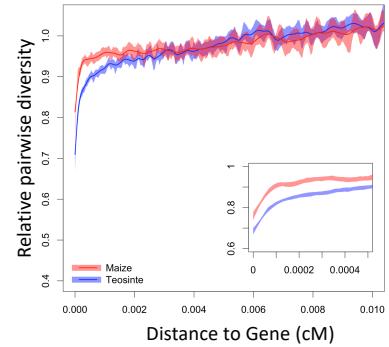


Figure 8.15: Relative diversity compared to the mean diversity in windows  $\geq 0.01$  cM as a function of the distance to the nearest gene. See (BEISSINGER *et al.*, 2016) for details. Figure licensed under CC BY 4.0 by Jeff Ross-Ibarra.

For a first go at fitting this to genome-wide data, we could look  
 5138 at diversity in windows of length  $W$  bp (as in Figure 8.16). If we  
 assume that there is a constant rate of deleterious mutation per base  
 5140 pair,  $\mu_{BP}$ , then  $U = \mu_{BP}W$ . Furthermore, if our genomic window  
 has a recombination rate  $r_{BP}$  per base-pair, our total genetic length  
 5142 is  $R = r_{BP}W$ . Making these substitutions in equation (8.17), our  
 window size cancels out to give

$$\mathbb{E}[\pi] \approx \pi_0 \exp(-\mu_{BP}/r_{bp}) \quad (8.18)$$

5144 Looking across windows that vary in their recombination rate, i.e.  
 $r_{BP}$ , we can fit equation (8.18) to data to estimate  $\mu_{BP}$ . An example  
 5146 of doing this to data from *D. melanogaster* is shown in Figure 8.16,  
 yielding an estimate of the deleterious mutation rate of  $\mu_{BP} \approx 3.2 \times$   
 5148  $10^{-9}$ . This is roughly on the same order as the mutation rate per  
 base pair in *D. melanogaster*, and so this deleterious mutation rate  
 5150 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  
 5152 how similar the fit is to a model of hitchhiking, suggesting that both  
 BGS and hitchhiking are capable of explaining the broad relationship  
 5154 between diversity and recombination seen in *D. melanogaster* and  
 other species.

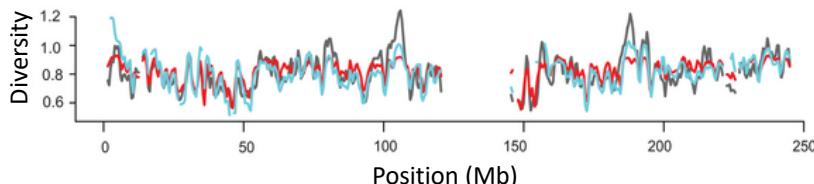


Figure 8.16: The relationship between recombination rate and synonymous site pairwise diversity ( $\pi$ ) in *D. melanogaster*, as in Figure 8.13. The red curve is the predicted relationship between  $\pi$  and recombination rate, obtained by fitting the BGS equation (8.17) to this data using non-linear least squares via the `nls()` function in R. The blue line is the recurrent hitchhiking equation line from Figure 8.13. Code here.

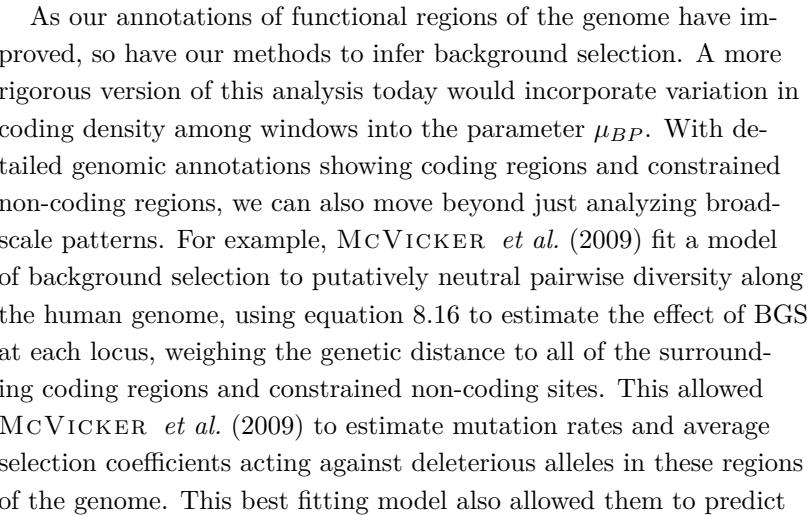


Figure 8.17: Observed (black line) and predicted pairwise diversity across chromosome 1, from a background selection model that assumes a uniform mutation rate (red line) or a mutation rate that varies with local human/dog divergence (blue line). Figure from (MCVICKER *et al.*, 2009), licensed under CC BY 4.0.

5170 diversity levels along the genome, a section of which is shown in figure  
 8.17. Thus, broad-scale features of polymorphism along the genome  
 5172 are well described by background selection (or by linked selection more  
 generally).

5174 The deleterious mutation rates estimated by MC VICKER *et al.*  
 (2009) from fitting a model of BGS were again too high, as in the  
 5176 *Drosophila* example above, suggesting the BGS alone is not sufficient  
 to explain all of the effect of linked selection. But how then do we go  
 5178 about distinguishing the impact of BGS from hitchhiking?

*Distinguishing the impact of hitchhiking from background selection*

5180 *in genome-wide data* A variety of approaches have been taken to  
 start to separate the effects of hitchhiking from background selection.  
 5182 Much of the strongest evidence showing the effects of both comes from  
*Drosophila melanogaster* and we review some of that evidence here.  
 5184 Hitchhiking is expected to have systematic effects on the neutral site  
 frequency spectrum, distorting it towards rare minor alleles, (reflecting  
 5186 the slow recovery of diversity following a sweep). Therefore, we should  
 expect a distortion of summary statistics such as Tajima's D in regions  
 5188 of low recombination if hitchhiking is contributing to the reduction in  
 diversity in these regions (BRAVERMAN *et al.*, 1995; PRZEWORSKI,  
 5190 KIM, 2006). In *D. melanogaster*, there is a greater skew towards  
 rare alleles at putatively neutral sites in regions of low recombination  
 5192 (ANDOLFATTO and PRZEWORSKI, 2001; SHAPIRO *et al.*, 2007),  
 see left panel of Figure 8.18. However, while this skew isn't expected  
 5194 under simple models of strong background selection, other models of  
 background selection can lead to such patterns.

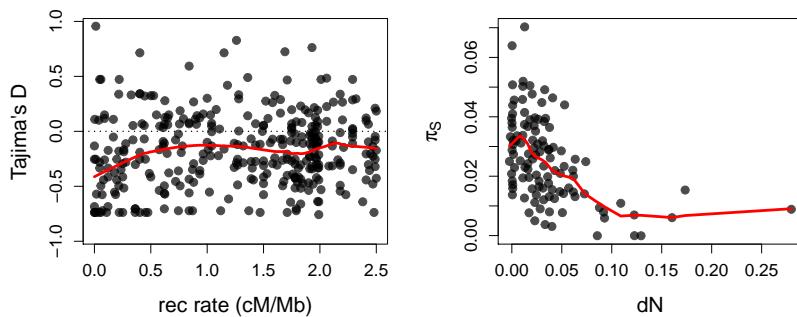


Figure 8.18: **Left)** Average Tajima's D in genomic windows plotted against their recombination rate in *D. melanogaster*. Data from SHAPIRO *et al.* (2007). **Right)** Synonymous pairwise diversity in genomic windows as a function of the density of non-synonymous substitutions in the window. Data from ANDOLFATTO (2007). Code here.

5196 Another prediction of the hitchhiking model, where an allele sweeps  
 to fixation, is that there should be a functional substitution associ-  
 5198 ated with each sweep. Or, to flip that around, we might expect to  
 see a greater impact of hitchhiking where there are more functional

5200 substitutions. For example, regions surrounding non-synonymous substitutions should have lower levels of diversity, if a high fraction of  
 5202 non-synonymous substitutions are adaptive. Again, this pattern is seen in *D. melanogaster* (ANDOLFATTO, 2007; MACPHERSON *et al.*,  
 5204 2007; SATTATH *et al.*, 2011b), right side of Figure 8.18.

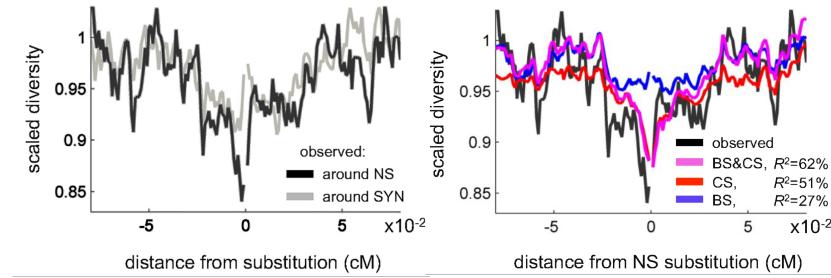


Figure 8.19: **Left)** Scaled synonymous pairwise diversity levels around non-synonymous (NS) and synonymous (SYN) substitutions in *D. melanogaster*. **Right)** Predicted scaled diversity levels around non-synonymous substitutions based on models including background selection (BS), classic sweeps (CS) and both (BS & CS). Figure from ELYASHIV *et al.* (2016), licensed under CC BY 4.0.

Pushing this idea further, we can look at the dip in diversity surrounding a non-synonymous substitution averaged across all the substitutions in the genome. ELYASHIV *et al.* (2016) found a stronger dip in diversity around non-synonymous substitutions than synonymous substitutions (see also SATTATH *et al.*, 2011a). Extending the model of MCVICKER *et al.* (2009) to fit a model of background selection and hitchhiking to putative neutral diversity along the genome, they found that the dip in diversity around synonymous substitutions comes mostly from BGS. But to fully explain the dip in diversity around non-synonymous substitutions, a reasonable proportion of these non-synonymous substitutions have to have been accompanied by a classic (hard) sweep. The majority of these sweeps are estimated to be due to very weak selection, with selection coefficients  $< 10^{-4}$ . Furthermore, ELYASHIV *et al.* (2016) estimated a 77 - 89% reduction in neutral diversity due to selection on linked sites, and concluded that no genomic window was entirely free of the effects of selection. Thus linked selection has a profound effect in some species such as  
 5222 *Drosophila melanogaster*.

## Bibliography

- 5224 AGUADÉ, M., N. MIYASHITA, and C. H. LANGLEY, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. *Genetics* **122**: 607–615.
- 5228 AGUILLO, S. M., J. W. FITZPATRICK, R. BOWMAN, S. J. SCHOECH, A. G. CLARK, G. COOP, and N. CHEN, 2017, 08) Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLOS Genetics* **13**(8): 1–27.
- 5232 AKÇAY, E. and J. VAN CLEVE, 2016 There is no fitness but fitness, and the lineage is its bearer. *Phil. Trans. R. Soc. B* **371**(1687): 20150085.
- 5236 ALCAIDE, M., E. S. SCORDATO, T. D. PRICE, and D. E. IRWIN, 2014 Genomic divergence in a ring species complex. *Nature* **511**(7507): 83.
- 5238 ALEXANDER, D. H., J. NOVEMBRE, and K. LANGE, 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**(9): 1655–1664.
- 5242 ALGEE-HEWITT, B. F., M. D. EDGE, J. KIM, J. Z. LI, and N. A. ROSENBERG, 2016 Individual identifiability predicts population identifiability in forensic microsatellite markers. *Current Biology* **26**(7): 935–942.
- 5246 ALLENDORF, F. W. and J. J. HARD, 2009 Human-induced evolution caused by unnatural selection through harvest of wild animals. *Proceedings of the National Academy of Sciences* **106**(Supplement 1): 9987–9994.
- 5250 ALVAREZ, G., F. C. CEBALLOS, and C. QUINTEIRO, 2009 The role of inbreeding in the extinction of a European royal dynasty. *PLoS One* **4**(4): e5174.

- 5252 ANDOLFATTO, P., 2007 Hitchhiking effects of recurrent beneficial  
amino acid substitutions in the *Drosophila melanogaster* genome.  
5254 *Genome Res.* **17**: 1755–1762.
- 5256 ANDOLFATTO, P. and M. PRZEWORSKI, 2001 Regions of lower  
crossing over harbor more rare variants in African populations of  
*Drosophila melanogaster*. *Genetics* **158**: 657–665.
- 5258 AYLLON, F., E. KJÆRNER-SEMB, T. FURMANEK, V. WEN-  
NEVIK, M. F. SOLBERG, G. DAHLE, G. L. TARANGER,  
5260 K. A. GLOVER, M. S. ALMÉN, C. J. RUBIN, and OTHERS,  
2015 The vgl3 locus controls age at maturity in wild and domesti-  
5262 cated Atlantic salmon (*Salmo salar* L.) males. *PLoS genetics* **11**(11):  
e1005628.
- 5264 BARRETT, R. D. H., S. M. ROGERS, and D. SCHLUTER,  
2008 Natural Selection on a Major Armor Gene in Threespine  
5266 Stickleback. *Science* **322**(5899): 255–257.
- 5268 BARSON, N. J., T. AYKANAT, K. HINDAR, M. BARANSKI,  
G. H. BOLSTAD, P. FISKE, C. JACQ, A. J. JENSEN, S. E.  
JOHNSTON, S. KARLSSON, and OTHERS, 2015 Sex-dependent  
5270 dominance at a single locus maintains variation in age at maturity  
in salmon. *Nature* **528**(7582): 405.
- 5272 BARTON, N. H., 2000 Genetic hitchhiking. *Philos. Trans. R. Soc.  
Lond., B, Biol. Sci.* **355**: 1553–1562.
- 5274 BECQUET, C., N. PATTERSON, A. C. STONE, M. PRZE-  
WORSKI, and D. REICH, 2007 Genetic structure of chimpanzee  
5276 populations. *PLoS genetics* **3**(4): e66.
- 5278 BEGUN, D. J. and C. F. AQUADRO, 1992 Levels of naturally  
occurring DNA polymorphism correlate with recombination rates in  
*D. melanogaster*. *Nature* **356**: 519–520.
- 5280 BEISSINGER, T. M., L. WANG, K. CROSBY, A. DURVA-  
SULA, M. B. HUFFORD, and J. ROSS-IBARRA, 2016 Recent  
5282 demography drives changes in linked selection across the maize  
genome. *Nature plants* **2**(7): 16084.
- 5284 BOX, G. E., 1979 Robustness in the strategy of scientific model  
building. In *Robustness in statistics*, pp. 201–236. Elsevier.
- 5286 BRADBURD, G. S., P. L. RALPH, and G. M. COOP, 2016  
A spatial framework for understanding population structure and  
5288 admixture. *PLoS genetics* **12**(1): e1005703.

- BRANDVAIN, Y., A. M. KENNEY, L. FLAGEL, G. COOP, and  
5290 A. L. SWEIGART, 2014 Speciation and introgression between  
Mimulus nasutus and Mimulus guttatus. *PLoS Genetics* **10**(6):  
5292 e1004410.
- BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H.  
5294 LANGLEY, and W. STEPHAN, 1995 The hitchhiking effect on  
the site frequency spectrum of DNA polymorphisms. *Genetics* **140**:  
5296 783–796.
- CAI, J. J., J. M. MACPHERSON, G. SELLA, and D. A.  
5298 PETROV, 2009 Pervasive hitchhiking at coding and regulatory  
sites in humans. *PLoS Genet.* **5**: e1000336.
- CASSA, C. A., D. WEGHORN, D. J. BALICK, D. M. JOR-  
5300 DAN, D. NUSINOW, K. E. SAMOCHA, A. O'DONNELL-  
5302 LURIA, D. G. MACARTHUR, M. J. DALY, D. R. BEIER,  
5304 and OTHERS, 2017 Estimating the selective effects of heterozy-  
gous protein-truncating variants from human exome data. *Nature  
genetics* **49**(5): 806.
- CHARLESWORTH, B., 2009 Effective population size and pat-  
5306 terns of molecular evolution and variation. *Nature Reviews Ge-  
5308 netics* **10**(3): 195.
- CHARLESWORTH, D., B. CHARLESWORTH, and M. T. MOR-  
5310 GAN, 1995 The pattern of neutral molecular variation under the  
background selection model. *Genetics* **141**: 1619–1632.
- CHEN, N., E. J. COSGROVE, R. BOWMAN, J. W. FITZ-  
5312 PATRICK, and A. G. CLARK, 2016 Genomic Consequences of  
5314 Population Decline in the Endangered Florida Scrub-Jay. *Current  
Biology* **26**(21): 2974 – 2979.
- COOK, L. M., B. S. GRANT, I. J. SACCHERI, and J. MAL-  
5316 LET, 2012 Selective bird predation on the peppered moth: the last  
5318 experiment of Michael Majerus. *Biology Letters* **8**(4): 609–612.
- COTTERMAN, C. W., 1940 A calculus for statistico-genetics. Ph.  
5320 D. thesis, The Ohio State University.
- COUSMINER, D. L., D. J. BERRY, N. J. TIMPSON, W. ANG,  
5322 E. THIERING, E. M. BYRNE, H. R. TAAL, V. HUIKARI,  
J. P. BRADFIELD, M. KERKHOF, and OTHERS, 2013  
5324 Genome-wide association and longitudinal analyses reveal genetic  
loci linking pubertal height growth, pubertal timing and childhood  
5326 adiposity. *Human molecular genetics* **22**(13): 2735–2747.

- CUTTER, A. D. and J. Y. CHOI, 2010 Natural selection shapes  
 5328 nucleotide polymorphism across the genome of the nematode  
*Caenorhabditis briggsae*. *Genome Res.* **20**: 1103–1111.
- DARWIN, C., 1859 *On the Origin of Species by Means of Natural Selection*. London: Murray. or the Preservation of Favored Races in  
 5330 the Struggle for Life.
- DARWIN, C., 1876 The effect of cross and self fertilization in the  
 5334 vegetable kingdom: Murray. London, UK.
- DARWIN, C., 1888 *The descent of man and selection in relation to sex*, Volume 1. Murray.
- DEMPSTER, E., 1955 Maintenance of genetic heterogeneity. *Cold Spring Harb Symp Quant Biol* **20**: 25–32.
- DICKERSON, R. E., 1971 The structure of cytochrome c and the  
 5340 rates of molecular evolution. *Journal of Molecular Evolution* **1**(1):  
 26–45.
- DOBZHANSKY, T., 1943 Genetics of natural populations IX. Temporal changes in the composition of populations of *Drosophila pseudoobscura*. *Genetics* **28**(2): 162.
- DOBZHANSKY, T., 1951 *Genetics and the Origin of Species* (3rd  
 5346 Ed. ed.), pp. 16.
- ELTON, C., 1942 *Voles, mice and lemmings. Problems in population dynamics*. Oxford: Clarendon Press.
- ELYASHIV, E., S. SATTATH, T. T. HU, A. STRUTSOVSKY,  
 5350 G. MCVICER, P. ANDOLFATTO, G. COOP, and G. SELLA,  
 2016 A genomic map of the effects of linked selection in *Drosophila*.  
 5352 *PLoS genetics* **12**(8): e1006130.
- EWENS, W. J., 2016 Motoo Kimura and James Crow on the Infinitely Many Alleles Model. *Genetics* **202**(4): 1243–1245.
- FAY, J. C. and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- FEDER, A. F., C. KLINE, P. POLACINO, M. COTTRELL,  
 5358 A. D. KASHUBA, B. F. KEELE, S.-L. HU, D. A. PETROV,  
 P. S. PENNINGS, and Z. AMBROSE, 2017 A spatio-temporal  
 5360 assessment of simian/human immunodeficiency virus (SHIV) evolution reveals a highly dynamic process within the host. *PLoS*  
 5362 *pathogens* **13**(5): e1006358.

- FISHER, R. A., 1915 The evolution of sexual preference. The  
5364 Eugenics Review 7(3): 184.
- FISHER, R. A., 1923 XXI.—on the dominance ratio. Proceedings of  
5366 the royal society of Edinburgh 42: 321–341.
- FRANCIOLI, L. C., A. MENELAOU, S. L. PULIT,  
5368 F. VAN DIJK, P. F. PALAMARA, C. C. ELBERS, P. B.  
NEERINCX, K. YE, V. GURYEV, W. P. KLOOSTERMAN,  
5370 and OTHERS, 2014 Whole-genome sequence variation, population  
structure and demographic history of the Dutch population. Nature  
5372 genetics 46(8): 818.
- FRENTIU, F. D., G. D. BERNARD, C. I. CUEVAS, M. P.  
5374 SISON-MANGUS, K. L. PRUDIC, and A. D. BRISCOE, 2007  
Adaptive evolution of color vision as seen through the eyes of but-  
5376 terflies. Proceedings of the National Academy of Sciences 104(suppl  
1): 8634–8640.
- GALEN, C., 1996 Rates of floral evolution: adaptation to bumblebee  
5378 pollination in an alpine wildflower, *Polemonium viscosum*. Evolution  
5380 50(1): 120–125.
- GALTIER, N., 2016 Adaptive protein evolution in animals and the  
5382 effective population size hypothesis. PLoS genetics 12(1): e1005774.
- GIGORD, L. D., M. R. MACNAIR, and A. SMITHSON, 2001  
5384 Negative frequency-dependent selection maintains a dramatic  
flower color polymorphism in the rewardless orchid *Dactylorhiza*  
5386 *sambucina* (L.) Soo. Proceedings of the National Academy of Sci-  
ences 98(11): 6253–6255.
- GILLESPIE, J. H., 2000 Genetic drift in an infinite population. The  
5388 pseudohitchhiking model. Genetics 155: 909–919.
- HALDANE, J., 1942 The selective elimination of silver foxes in east-  
5390 ern Canada. Journal of Genetics 44(2-3): 296–304.
- HALDANE, J. and S. JAYAKAR, 1963 Polymorphism due to selec-  
5392 tion of varying direction. Journal of Genetics 58(2): 237–242.
- HALDANE, J. B. S., 1927 A mathematical theory of natural and  
5394 artificial selection, part V: selection and mutation. In *Mathematical*  
5396 *Proceedings of the Cambridge Philosophical Society*, Volume 23, pp.  
838–844. Cambridge University Press.
- HALDANE, J. B. S., 1937 The Effect of Variation of Fitness. The  
5398 American Naturalist 71(735): 337–349.

- 5400 HAMILTON, W. D., 1964a The genetical evolution of social behaviour. II. Journal of theoretical biology 7(1): 17–52.
- 5402 HAMILTON, W. D., 1964b The genetical evolution of social behaviour. II. Journal of theoretical biology 7(1): 17–52.
- 5404 HERMISSON, J. and P. S. PENNINGS, 2017 Soft sweeps and beyond: understanding the patterns and probabilities of selection 5406 footprints under rapid adaptation. Methods in Ecology and Evolution 8(6): 700–716.
- 5408 HEY, J. and R. M. KLIMAN, 2002 Interactions between natural selection, recombination and gene density in the genes of 5410 *Drosophila*. Genetics 160(2): 595–608.
- 5412 HOEKSTRA, H. E., K. E. DRUMM, and M. W. NACHMAN, 2004 Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. Evolution 5414 58(6): 1329–1341.
- 5416 HOHENLOHE, P. A., S. BASSHAM, P. D. ETTER, N. STIFFLER, E. A. JOHNSON, and W. A. CRESKO, 2010 5418 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS genetics 6(2): e1000862.
- 5420 HOLLISTER, J. D., S. GREINER, W. WANG, J. WANG, Y. ZHANG, G. K.-S. WONG, S. I. WRIGHT, and M. T. 5422 JOHNSON, 2014 Recurrent loss of sex is associated with accumulation of deleterious mutations in *Oenothera*. Molecular biology and evolution 32(4): 896–905.
- 5424 HOPKINS, J., G. BAUDRY, U. CANDOLIN, and A. KAITALA, 2015 I'm sexy and I glow it: female ornamentation in a nocturnal capital breeder. Biology letters 11(10): 20150599.
- 5428 HOUDE, A. E., 1994 Effect of artificial selection on male colour 5430 patterns on mating preference of female guppies. Proc. R. Soc. Lond. B 256(1346): 125–130.
- 5432 HOWES, R. E., M. DEWI, F. B. PIEL, W. M. MONTEIRO, K. E. BATTLE, J. P. MESSINA, A. SAKUNTABHAI, A. W. 5434 SATYAGRAHA, T. N. WILLIAMS, J. K. BAIRD, and S. I. HAY, 2013 Spatial distribution of G6PD deficiency variants across malaria-endemic regions. Malar. J. 12: 418.
- 5436 HOWES, R. E., F. B. PIEL, A. P. PATIL, O. A. NYANGIRI, P. W. GETHING, M. DEWI, M. M. HOGG, K. E. BATTLE, C. D. PADILLA, J. K. BAIRD, and S. I. HAY, 2012

- 5438 G6PD deficiency prevalence and estimates of affected populations in  
malaria endemic countries: a geostatistical model-based map. PLoS  
5440 Medicine 9(11): e1001339.
- 5442 HUDSON, R. R., 2015, 07)A New Proof of the Expected Frequency  
Spectrum under the Standard Neutral Model. PLOS ONE 10(7):  
1–5.
- 5444 HUDSON, R. R. and N. L. KAPLAN, 1995a Deleterious back-  
ground selection with recombination. Genetics 141: 1605–1617.
- 5446 HUDSON, R. R. and N. L. KAPLAN, 1995b The coalescent pro-  
cess and background selection. Philos. Trans. R. Soc. Lond., B, Biol.  
5448 Sci. 349: 19–23.
- 5450 HUDSON, R. R., M. KREITMAN, and M. AGUADÉ, 1987 A test  
of neutral molecular evolution based on nucleotide data. Genet-  
ics 116(1): 153–159.
- 5452 JAIN, S. and A. D. BRADSHAW, 1966 Evolutionary divergence  
among adjacent plant populations I. The evidence and its theoreti-  
5454 cal analysis. Heredity 21(3): 407.
- 5456 JANICKE, T., I. K. HÄDERER, M. J. LAJEUNESSE, and  
N. ANTHES, 2016 Darwinian sex roles confirmed across the an-  
imal kingdom. Science advances 2(2): e1500983.
- 5458 JENNINGS, W. B. and S. V. EDWARDS, 2005 Speciational his-  
tory of Australian grass finches (Poephila) inferred from thirty gene  
5460 trees. Evolution 59(9): 2033–2047.
- JOHANNSEN, W., 1911 The Genotype Conception of Heredity. The  
5462 American Naturalist 45(531): 129–159.
- 5464 JOHNSTON, S. E., J. GRATTON, C. BERENOS, J. G. PILK-  
INGTON, T. H. CLUTTON-BROCK, J. M. PEMBERTON, and  
5466 J. SLATE, 2013 Life history trade-offs at a single locus maintain  
sexually selected genetic variation. Nature 502(7469): 93.
- JUKEMA, J. and T. PIERSMA, 2006 Permanent female mimics in  
5468 a lekking shorebird. Biology letters 2(2): 161–164.
- KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989  
5470 The hitchhiking effect revisited. Genetics 123: 887–899.
- KETTLEWELL, H. B. D., 1955 Selection experiments on industrial  
5472 melanism in the Lepidoptera. Heredity 9(3): 323.
- KIM, Y., 2006 Allele frequency distribution under recurrent selective  
5474 sweeps. Genetics 172: 1967–1978.

- KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature 217(5129): 624–626.
- KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.
- KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics 49(4): 725.
- KIMURA, M. and T. OHTA, 1974 On some principles governing molecular evolution. Proceedings of the National Academy of Sciences 71(7): 2848–2852.
- KING, J. L. and T. H. JUKES, 1969 Non-darwinian evolution. Science 164(3881): 788–798.
- KORNEGAY, J. R., J. W. SCHILLING, and A. C. WILSON, 1994 Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. Molecular Biology and Evolution 11(6): 921–928.
- KRAKAUER, A. H., 2005 Kin selection and cooperative courtship in wild turkeys. Nature 434(7029): 69.
- KRUUK, L. E., J. SLATE, J. M. PEMBERTON, S. BROTH-ERSTONE, F. GUINNESS, and T. CLUTTON-BROCK, 2002 Antler size in red deer: heritability and selection but no evolution. Evolution 56(8): 1683–1695.
- KÜPPER, C., M. STOCKS, J. E. RISSE, N. DOS REMEDIOS, L. L. FARRELL, S. B. MCRAE, T. C. MORGAN, N. KAR-LIONOVA, P. PINCHUK, Y. I. VERKUIL, and OTHERS, 2016 A supergene determines highly divergent male reproductive morphs in the ruff. Nature Genetics 48(1): 79.
- KWIATKOWSKI, D. P., 2005, August)How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77(2): 171–192.
- LAMICHHANEY, S., G. FAN, F. WIDEMO, U. GUNNARSSON, D. S. THALMANN, M. P. HOEPPNER, S. KERJE, U. GUSTAFSON, C. SHI, H. ZHANG, and OTHERS, 2016 Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). Nature Genetics 48(1): 84.
- LANDE, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution 33(1Part2): 402–416.

- LAURIE, C. C., D. A. NICKERSON, A. D. ANDERSON, B. S.  
5512 WEIR, R. J. LIVINGSTON, M. D. DEAN, K. L. SMITH,  
E. E. SCHADT, and M. W. NACHMAN, 2007, 08)Linkage Dise-  
5514 quilibrium in Wild Mice. PLOS Genetics 3(8): 1–9.
- LAWSON, D. J., L. VAN DORP, and D. FALUSH, 2018 A tuto-  
5516 rial on how not to over-interpret STRUCTURE and ADMIXTURE  
bar plots. Nature communications 9(1): 3258.
- LEFÉBURE, T., C. MORVAN, F. MALARD, C. FRANÇOIS,  
5518 L. KONECNY-DUPRÉ, L. GUÉGUEN, M. WEISS-GAYET,  
5520 A. SEGUIN-ORLANDO, L. ERMINI, C. DER SARKISSIAN,  
and OTHERS, 2017 Less effective selection leads to larger genomes.  
5522 Genome research: gr–212589.
- LEFFLER, E. M., K. BULLAUGHEY, D. R. MATUTE, W. K.  
5524 MEYER, L. SEGUREL, A. VENKAT, P. ANDOLFATTO, and  
M. PRZEWORSKI, 2012 Revisiting an old riddle: what deter-  
5526 mines genetic diversity levels within species? PLoS biology 10(9):  
e1001388.
- LEK, M., K. J. KARCZEWSKI, E. V. MINIKEL, K. E.  
5528 SAMOCHA, E. BANKS, T. FENNELL, A. H. O'DONNELL-  
5530 LURIA, J. S. WARE, A. J. HILL, B. B. CUMMINGS, and  
OTHERS, 2016 Analysis of protein-coding genetic variation in  
5532 60,706 humans. Nature 536(7616): 285.
- LENORMAND, T., D. BOURGUET, T. GUILLEMAUD, and  
5534 M. RAYMOND, 1999 Tracking the evolution of insecticide resis-  
tance in the mosquito *Culex pipiens*. Nature 400(6747): 861.
- LEWONTIN, R. C., 1970 The units of selection. Annual review of  
5536 ecology and systematics 1(1): 1–18.
- LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary  
5538 Change*. Columbia University Press, New York.
- LEWONTIN, R. C., 1994, 05)[DNA Fingerprinting: A Review of  
5540 the Controversy]: Comment: The Use of DNA Profiles in Forensic  
5542 Contexts. Statist. Sci. 9(2): 259–262.
- LEWONTIN, R. C., 2001 *Thinking about evolution: historical, philo-  
5544 sophical, and political perspectives*, Chapter Natural History and  
Formalism in Evolutionary Genetics, pp. 7–20. Cambridge Univer-  
5546 sity Press.
- LI, J. Z., D. M. ABSHER, H. TANG, A. M. SOUTHWICK,  
5548 A. M. CASTO, S. RAMACHANDRAN, H. M. CANN, G. S.

- 5550 BARSH, M. FELDMAN, L. L. CAVALLI-SFORZA, and OTHERS, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. *science* **319**(5866): 1100–1104.
- 5552 LISTER, A., 1989 Rapid dwarfing of red deer on Jersey in the last interglacial. *Nature* **342**(6249): 539.
- 5554 LOCKE, D. P., L. W. HILLIER, W. C. WARREN, K. C. WORLEY, L. V. NAZARETH, D. M. MUZNY, S.-P. YANG, 5556 Z. WANG, A. T. CHINWALLA, P. MINX, and OTHERS, 2011 Comparative and demographic analysis of orang-utan genomes. 5558 *Nature* **469**(7331): 529.
- 5560 LOUICHAROEN, C., E. PATIN, R. PAUL, I. NUCHPRAY- 5562 OON, B. WITOONPANICH, C. PEERAPITTAYAMONGKOL, I. CASADEMONT, T. SURA, N. M. LAIRD, P. SINGHASI- 5564 VANON, L. QUINTANA-MURCI, and A. SAKUNTABHAI, 2009, December)Positively selected G6PD-Mahidol mutation reduces *Plasmodium vivax* density in Southeast Asians. *Science* **326**(5959): 1546–1549.
- 5566 MACARTHUR, D. G., S. BALASUBRAMANIAN, A. FRANK- 5568 ISH, N. HUANG, J. MORRIS, K. WALTER, L. JOSTINS, L. HABEGGER, J. K. PICKRELL, S. B. MONTGOMERY, and OTHERS, 2012 A systematic survey of loss-of-function variants in 5570 human protein-coding genes. *Science* **335**(6070): 823–828.
- 5572 MACPHERSON, J. M., G. SELLA, J. C. DAVIS, and D. A. PETROV, 2007 Genomewide spatial correspondence between non-synonymous divergence and neutral polymorphism reveals extensive 5574 adaptation in *Drosophila*. *Genetics* **177**: 2083–2099.
- 5576 MAJERUS, 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.
- 5578 MALÉCOT, G., 1948 Les mathématiques de l'hérédité.
- 5580 MALÉCOT, G., 1969 The Mathematics of Heredity (Revised, edited and translated by Yermanos, DM).
- 5582 MARCINIAK, S. and G. H. PERRY, 2017 Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* **18**(11): 659.
- 5584 MAYNARD SMITH, J., 1964 Group selection and kin selection. *Nature* **201**(4924): 1145.

- 5586 MAYNARD SMITH, J. and J. HAIGH, 1974 The hitch-hiking  
effect of a favourable gene. *Genet. Res.* **23**: 23–35.
- 5588 McDONALD, J. H. and M. KREITMAN, 1991 Adaptive protein  
evolution at the Adh locus in *Drosophila*. *Nature* **351**(6328): 652.
- 5590 MCVICKER, G., D. GORDON, C. DAVIS, and P. GREEN, 2009  
Widespread genomic signatures of natural selection in hominid  
5592 evolution. *PLoS Genet.* **5**: e1000471.
- MENOZZI, P., A. PIAZZA, and L. CAVALLI-SFORZA, 1978  
5594 Synthetic maps of human gene frequencies in Europeans. *Sci-  
ence* **201**(4358): 786–792.
- 5596 MEREDITH, R. W., J. GATESY, W. J. MURPHY, O. A. RY-  
DER, and M. S. SPRINGER, 2009, 09)Molecular Decay of the  
5598 Tooth Gene Enamelin (ENAM) Mirrors the Loss of Enamel in the  
Fossil Record of Placental Mammals. *PLOS Genetics* **5**(9): 1–12.
- 5600 MESSIER, W. and C.-B. STEWART, 1997 Episodic adaptive  
evolution of primate lysozymes. *Nature* **385**(6612): 151.
- 5602 NACHMAN, M. W., H. E. HOEKSTRA, and S. L.  
D'AGOSTINO, 2003 The genetic basis of adaptive melanism  
5604 in pocket mice. *Proceedings of the National Academy of Sci-  
ences* **100**(9): 5268–5273.
- 5606 NASH, D., S. NAIR, M. MAYXAY, P. N. NEWTON, J.-P.  
GUTHMANN, F. NOSTEN, and T. J. ANDERSON, 2005 Selec-  
5608 tion strength and hitchhiking around two anti-malarial resistance  
genes. *Proceedings of the Royal Society of London B: Biological  
5610 Sciences* **272**(1568): 1153–1161.
- NELSON, M. R., D. WEGMANN, M. G. EHM, D. KESSNER,  
5612 P. S. JEAN, C. VERZILLI, J. SHEN, Z. TANG, S.-A. BA-  
CANU, D. FRASER, and OTHERS, 2012 An abundance of rare  
5614 functional variants in 202 drug target genes sequenced in 14,002  
people. *Science*: 1217876.
- 5616 NORDBORG, M., B. CHARLESWORTH, and  
D. CHARLESWORTH, 1996 The effect of recombination on back-  
5618 ground selection. *Genet. Res.* **67**: 159–174.
- NOVEMBRE, J. and M. STEPHENS, 2008 Interpreting principal  
5620 component analyses of spatial population genetic variation. *Nature  
genetics* **40**(5): 646.
- 5622 OHTA, T., 1972 Population size and rate of evolution. *Journal of  
Molecular Evolution* **1**(4): 305–314.

- 5624 OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. *Nature* *246*(5428): 96.
- 5626 OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution* *26*(1-2): 1–6.
- 5628 OHTA, T. and J. H. GILLESPIE, 1996 Development of neutral and nearly neutral theories. *Theoretical population biology* *49*(2): 5630 128–142.
- OWEN, D. and D. CHANTER, 1972 Polymorphic mimicry in a 5632 population of the African butterfly, *Pseudacraea eurytus* (L.) (Lep. Nymphalidae). *Insect Systematics & Evolution* *3*(4): 258–266.
- 5634 PAABY, A. B., A. O. BERGLAND, E. L. BEHRMAN, and P. S. SCHMIDT, 2014 A highly pleiotropic amino acid polymorphism 5636 in the *Drosophila insulin* receptor contributes to life-history adaptation. *Evolution* *68*(12): 3395–3409.
- 5638 PATTERSON, N., A. L. PRICE, and D. REICH, 2006 Population structure and eigenanalysis. *PLoS genetics* *2*(12): e190.
- 5640 PICKRELL, J. K., T. BERISA, J. Z. LIU, L. SÉGUREL, J. Y. TUNG, and D. A. HINDS, 2016 Detection and interpretation of 5642 shared genetic influences on 42 human traits. *Nature genetics* *48*(7): 709.
- 5644 POTTI, J. and D. CANAL, 2011 Heritability and genetic correlation between the sexes in a songbird sexual ornament. *Heredity* *106*(6): 945.
- PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 5648 Inference of population structure using multilocus genotype data. *Genetics* *155*(2): 945–959.
- 5650 PROVINE, W. B., 2001 *The origins of theoretical population genetics: with a new afterword*. University of Chicago Press.
- 5652 PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. *Genetics* *160*: 1179–1189.
- 5654 PTAK, S. E., A. D. ROEDER, M. STEPHENS, Y. GILAD, S. PÄÄBO, and M. PRZEWORSKI, 2004 Absence of the TAP2 5656 human recombination hotspot in chimpanzees. *PLoS biology* *2*(6): e155.
- 5658 QUELLER, D. C., 1992 Quantitative genetics, inclusive fitness, and group selection. *The American Naturalist* *139*(3): 540–558.
- 5660 R, 2018 R: A Language and Environment for Statistical Computing.

- 5662 RALPH, P. L. and G. COOP, 2015 The role of standing variation in geographic convergent adaptation. *The American Naturalist* 186(S1): S5–S23.
- 5664 RANDS, C. M., S. MEADER, C. P. PONTING, and G. LUNTER, 2014 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS genetics* 10(7): e1004525.
- 5668 RICHARDS, C. M., 2000 Inbreeding depression and genetic rescue in a plant metapopulation. *The American Naturalist* 155(3): 383–394.
- 5672 RITLAND, K., C. NEWTON, and H. MARSHALL, 2001 Inheritance and population structure of the white-phased “Kermode” black bear. *Current Biology* 11(18): 1468 – 1472.
- 5674 ROBERTSON, A., 1961 Inbreeding in artificial selection programmes. *Genet. Res.* 2: 189—194.
- 5676 ROBINSON, J. A., D. ORTEGA-DEL VECCHYO, Z. FAN, B. Y. KIM, C. D. MARSDEN, K. E. LOHMUELLER, R. K. WAYNE, and OTHERS, 2016 Genomic flatlining in the endangered island fox. *Current Biology* 26(9): 1183–1189.
- 5680 ROBINSON, L. M., J. R. BOLAND, and J. M. BRAVERMAN, 2016 Revisiting a Classic Study of the Molecular Clock. *Journal of molecular evolution* 82(2-3): 110–116.
- 5684 ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD, L. A. ZHIVOTOVSKY, and M. W. FELDMAN, 2002 Genetic structure of human populations. *science* 298(5602): 2381–2385.
- 5688 RUWENDE, C., S. C. KHOO, R. W. SNOW, S. N. YATES, D. KWIATKOWSKI, S. GUPTA, P. WARN, C. E. ALLSOPP, S. C. GILBERT, and N. PESCHU, 1995, July)Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* 376(6537): 246–249.
- 5692 SAMS, A. J. and A. R. BOYKO, 2018a Fine-scale resolution and analysis of runs of homozygosity in domestic dogs. *bioRxiv*.
- 5694 SAMS, A. J. and A. 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.

- 5698 SANKARARAMAN, S., N. PATTERSON, H. LI, S. PÄÄBO, and  
 5700 D. REICH, 2012, 10)The Date of Interbreeding between Neander-  
 tals and Modern Humans. *PLOS Genetics* **8**(10): 1–9.
- SANTIAGO, E. and A. CABALLERO, 1995 Effective size of popu-  
 5702 lations under selection. *Genetics* **139**: 1013–1030.
- SANTIAGO, E. and A. CABALLERO, 1998 Effective size and  
 5704 polymorphism of linked neutral loci in populations under directional  
 selection. *Genetics* **149**: 2105–2117.
- 5706 SATTATH, S., E. ELYASHIV, O. KOLODNY, Y. RINOTT, and  
 5708 G. SELLA, 2011a Pervasive adaptive protein evolution apparent  
 in diversity patterns around amino acid substitutions in *Drosophila*  
*simulans*. *PLoS genetics* **7**(2): e1001302.
- 5710 SATTATH, S., E. ELYASHIV, O. KOLODNY, Y. RINOTT, and  
 5712 G. SELLA, 2011b Pervasive adaptive protein evolution apparent  
 in diversity patterns around amino acid substitutions in *Drosophila*  
*simulans*. *PLoS Genet.* **7**: e1001302.
- 5714 SCHEMSKE, D. W. and P. BIERZYCHUDEK, 2001 Perspective:  
 evolution of flower color in the desert annual *Linanthus parryae*:  
 5716 Wright revisited. *Evolution* **55**(7): 1269–1282.
- SELLA, G., D. A. PETROV, M. PRZEWORSKI, and P. AN-  
 5718 DOLFATTO, 2009 Pervasive natural selection in the *Drosophila*  
 genome? *PLoS genetics* **5**(6): e1000495.
- 5720 SHAPIRO, J. A., W. HUANG, C. ZHANG, M. J. HUBISZ,  
 5722 J. LU, D. A. TURISSINI, S. FANG, H. Y. WANG, R. R.  
 HUDSON, R. NIELSEN, Z. CHEN, and C. I. WU, 2007 Adap-  
 tive genic evolution in the *Drosophila* genomes. *Proc. Natl. Acad.  
 5724 Sci. U.S.A.* **104**: 2271–2276.
- SMITHSON, A. and M. R. MACNAIR, 1997 Negative frequency-  
 5726 dependent selection by pollinators on artificial flowers without re-  
 wards. *Evolution* **51**(3): 715–723.
- 5728 STURTEVANT, A. H., 1915 The behavior of the chromosomes as  
 studied through linkage. *Zeitschrift für induktive Abstammungs-und  
 5730 Vererbungslehre* **13**(1): 234–287.
- TAJIMA, F., 1989 Statistical method for testing the neutral muta-  
 5732 tion hypothesis by DNA polymorphism. *Genetics* **123**(3): 585–595.
- TISHKOFF, S. A., R. VARKONYI, N. CAHINHINAN,  
 5734 S. ABBES, G. ARGYROPOULOS, G. DESTRO-BISOL,  
 A. DROUSIOTOU, B. DANGERFIELD, G. LEFRANC,

- 5736 J. LOISELET, A. PIRO, M. STONEKING, A. TAGARELLI,  
G. TAGARELLI, E. H. TOUMA, S. M. WILLIAMS, and  
5738 A. G. CLARK, 2001 Haplotype Diversity and Linkage Disequi-  
librium at Human G6PD: Recent Origin of Alleles That Confer  
5740 Malarial Resistance. *Science* 293(5529): 455–462.
- 5742 TOEWS, D. P., S. A. TAYLOR, R. VALLENDER, A. BRELS-  
FORD, B. G. BUTCHER, P. W. MESSEY, and I. J.  
5744 LOVETTE, 2016 Plumage Genes and Little Else Distinguish the  
Genomes of Hybridizing Warblers. *Current Biology* 26(17): 2313 –  
2318.
- 5746 TURELLI, M., D. W. SCHEMSKE, and P. BIERZYCHUDEK,  
2001 Stable two-allele polymorphisms maintained by fluctuating  
5748 fitnesses and seed banks: protecting the blues in *Linanthus parryae*.  
*Evolution* 55(7): 1283–1298.
- 5750 VAN'T HOF, A. E., N. EDMONDS, M. DALÍKOVÁ, F. MAREC,  
and I. J. SACCHERI, 2011 Industrial melanism in British pep-  
5752 pered moths has a singular and recent mutational origin. *Sci-  
ence* 332(6032): 958–960.
- 5754 VOIGHT, B. F., A. M. ADAMS, L. A. FRISSE, Y. QIAN,  
R. R. HUDSON, and A. DI RIENZO, 2005 Interrogating mul-  
5756 tiple aspects of variation in a full resequencing data set to infer hu-  
man population size changes. *Proceedings of the National Academy  
of Sciences* 102(51): 18508–18513.
- VONHOLDT, B. M., J. P. POLLINGER, D. A. EARL,  
5760 J. C. KNOWLES, A. R. BOYKO, H. PARKER, E. GEF-  
FEN, M. PILOT, W. JEDRZEJEWSKI, B. JEDRZEJEJEW-  
5762 SKA, V. SIDOROVICH, C. GRECO, E. RANDI, M. MU-  
SIANI, R. KAYS, C. D. BUSTAMANTE, E. A. OSTRANDER,  
5764 J. NOVEMBRE, and R. K. WAYNE, 2011 A genome-wide per-  
spective on the evolutionary history of enigmatic wolf-like canids.  
5766 *Genome Research*.
- WANG, J., J. DING, B. TAN, K. M. ROBINSON, I. H.  
5768 MICHELSON, A. JOHANSSON, B. NYSTEDT, D. G.  
SCOFIELD, O. NILSSON, S. JANSSON, and OTHERS, 2018  
5770 A major locus controls local adaptation and adaptive life history  
variation in a perennial plant. *Genome biology* 19(1): 72.
- 5772 WATTERSON, G., 1975 On the number of segregating sites in ge-  
netical models without recombination. *Theoretical population  
5774 biology* 7(2): 256–276.

- WHEELER, W. M., 1907 Pink Insect Mutants. *The American Naturalist* 41(492): 773–780.
- 5776
- WIDEMO, F., 1998 Alternative reproductive strategies in the ruff, *Philomachus pugnax*: a mixed ESS? *Animal Behaviour* 56(2): 329–336.
- 5778
- 5780 WIEHE, T. and W. 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.
- 5782
- 5784 WIEHE, T. H. and W. STEPHAN, 1993b Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10: 842–854.
- 5786
- 5788 WILKINSON, G. S., 1993 Artificial sexual selection alters allometry in the stalk-eyed fly *Cyrtodiopsis dalmanni* (Diptera: Diopsidae). *Genetics Research* 62(3): 213–222.
- 5790 WILLIAMS, G. C., 1966 *Adaptation and Natural Selection*. Princeton.
- 5792 WILLIAMS, K.-A. and P. 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.
- 5794
- 5796 WISELY, S. M., S. W. BUSKIRK, M. A. FLEMING, D. B. McDONALD, and E. A. OSTRANDER, 2002 Genetic Diversity and Fitness in Black-Footed Ferrets Before and During a Bottleneck. *Journal of Heredity* 93(4): 231–237.
- 5798
- 5800 WRIGHT, K. M., U. HELLSTEN, C. XU, A. L. JEONG, A. SREEDASYAM, J. A. CHAPMAN, J. SCHMUTZ, G. COOP, D. S. ROKHSAR, and J. H. WILLIS, 2015 Adaptation to heavy-metal contaminated environments proceeds via selection on pre-existing genetic variation. bioRxiv: 029900.
- 5802
- 5804
- WRIGHT, S., 1943 Isolation by Distance. *Genetics* 28(2): 114–138.
- 5806 WRIGHT, S., 1949 The Genetical Structure of Populations. *Annals of Eugenics* 15(1): 323–354.
- 5808 WRIGHT, S. and T. DOBZHANSKY, 1946 Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. *Genetics* 31(2): 125.
- 5810

- 5812 WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI,  
J. F. DOBLEY, M. D. McMULLEN, and B. S. GAUT,  
5814 2005 The Effects of Artificial Selection on the Maize Genome.  
Science 308(5726): 1310–1314.
- 5816 YANG, Z., 1998 Likelihood ratio tests for detecting positive selection  
and application to primate lysozyme evolution. Molecular Biology  
5818 and Evolution 15(5): 568–573.
- ZUCKERKANDL, E. and L. PAULING, 1965 Evolutionary diver-  
gence and convergence in proteins. In *Evolving genes and proteins*,  
5820 pp. 97–166. Elsevier.