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: gmcoop@ucdavis.edu

Typeset using LaTeX and the TUFTE-LATEX book style.

Updated on February 2019

Contents

1
Introduction

Evolution is change over time. Biological evolution is the change over time in the genetic composition of a population.[1] Our population is made up of a set of interbreeding individuals, the genetic composition of which is made up of the genomes that each individual carries. While at first this definition of evolution seems at odds with the common textbook view of the evolution of phenotypes, such as the changing shape of finch beaks over generations, it is genetic changes that underpin these phenotypic changes.

[1] Dobzhansky, T., 1951 *Genetics and the Origin of Species* (3rd Ed. ed.)., pp. 16

The genetic composition of the population can alter due to the death of individuals or the migration of individuals in or out of the population. If our individuals vary in the number of children they have, this also alters the genetic composition of the population in the next generation. Every new individual born into the population subtly changes the genetic composition of the population. Their genome is a unique combination of their parents' genomes, having been shuffled by segregation and recombination during meioses, and possibly changed by mutation. These individual events seem minor at the level of the population, but it is the accumulation of small changes in aggregate across individuals and generations that is the stuff of evolution. It is the compounding of these small changes over tens, hundreds, and millions of generations that drives the amazing diversity of life that has emerged on this earth.

Population genetics is the study of the genetic composition of natural populations and its evolutionary causes and consequences. Quantitative genetics is the study of the genetic basis of phenotypic variation and how phenotypic changes can evolve. Both fields are closely conceptually aligned as we'll see throughout these notes. They seek to describe how the genetic and phenotypic composition of populations can be changed over time by the forces of mutation, recombination, selection, migration, and genetic drift. To understand how these forces interact, it is helpful to develop simple theoretical models to help our

intuition. In these notes we will work through these models and summarize the major areas of population- and quantitative-genetic theory.

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

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

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.

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

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.

"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

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

2
Allele and Genotype Frequencies

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

Loci and alleles are the basic currency of population genetics–and indeed of genetics. If all individuals in the population carry the same 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 the population at a locus, we say that this locus is *polymorphic* (this is sometimes referred to as a segregating site).

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*.

Table 2.1 show a small stretch orthologous sequence for the ADH locus from samples from *Drosophila melanogaster*, *D. simulans*, and *D. yakuba*. *D. melanogaster* and *D. simulans* are sister species and *D. yakuba* is a close outgroup to the two. Each column represents a 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 is an example of a polymophism; some *D. simulans* individuals carry a C allele while others have a T. *F*ixed differences are sites that differ between the species but are monomorphic within the species. Site 781 is an example of a fixed difference between *D. melanogaster* and the other two species.

We can also annotate the alleles and loci in various ways. For example, position 781 is a non-synonymous fixed difference. We call the less common allele at a polymorphism the *m*inor allele and the common allele the *m*ajor allele, e.g. at site 1068 the T allele is the minor allele in *D. melanogaster*. We call the more evolutionarily recent of the two alleles the *d*erived allele and the older of the two the *a*ncestral allele. The T allele at site 1068 is the derived allele as the C is found in both the other species, suggesting that the T allele arose via a $C \rightarrow T$ mutation.

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	a	b	c	d	e	f	g	h	i	j	k	l	NS/S
781	G	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NS
789	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
808	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
816	G	T	T	T	T	-	-	-	-	-	-	-	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
834	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	S
859	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	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	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	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

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.

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

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

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

2.1.1 Measures of genetic variability

Nucleotide diversity (π) One common measure of genetic diversity is the average number of single nucleotide differences between haplotypes chosen at random from a sample. This is called *n*ucleotide diversity and is often denoted by π. For example, we can calculate π for our ADH locus from Table 2.1 above: we have 6 sequences from *D. simulans* (a-f), there's a total of 15 ways of pairing these sequences, and

$$\pi = \frac{1}{15}\big((2+1+1+1+0)+(3+3+3+2)+(0+0+1)+(0+1)+(1)\big) = 1.2\overline{6} \tag{2.2}$$

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

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

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

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

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

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

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

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.

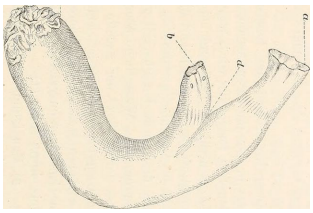

Figure 2.1: Sea Squirt (*Ciona intestinalis*).
Einleitung in die vergleichende gehirnphysiologie und Vergleichende psychologie. Loeb, J. 1899. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. No known copyright restrictions.

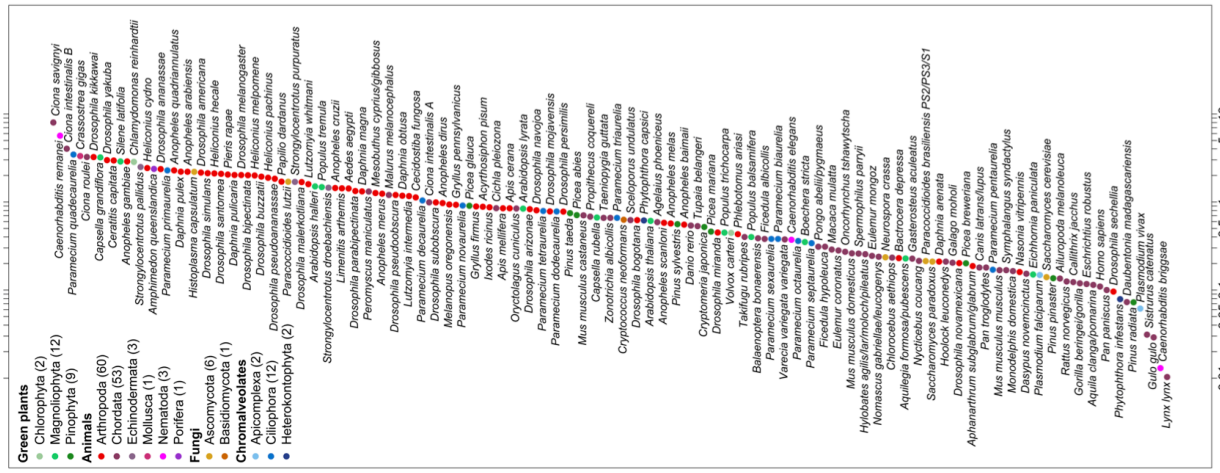

Figure 2.2: Levels of autosomal nucleotide diversity for 167 species across 14 phyla. Figure 1 from LEFFLER *et al.* (2012), licensed under CC BY 4.0. Points are ranked by their π, and coloured by their phylum. Note the log-scale.

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

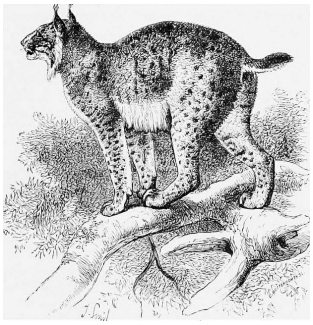

Figure 2.3: Eurasian Lynx (*Lynx lynx*).
An introduction to the study of mammals living and extinct. Flower, W.H. and Lydekker, R. 1891. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. No known copyright restrictions.

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-

tion. The expected frequency of the three possible genotypes are

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

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

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

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

AA	*AG*	*GG*
42	24	21

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

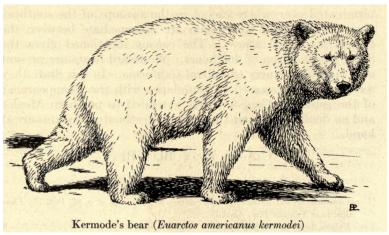

Figure 2.4: Kermode's bear.
Extinct and vanishing mammals of the western hemisphere. 1942. Glover A. Image from the Biodiversity Heritage Library. Contributed by Prelinger Library. Not in copyright.

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

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

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

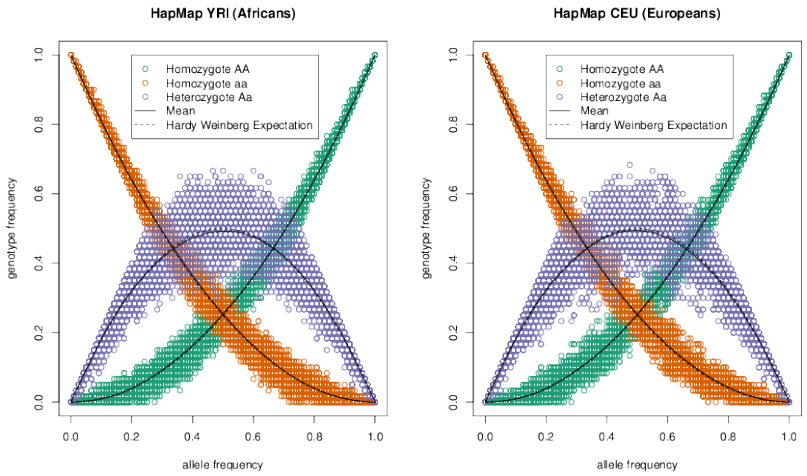

Figure 2.5: Demonstrating Hardy–Weinberg proportions using 10,000 SNPs from the HapMap European (CEU) and African (YRI) populations. Within each of these populations the allele frequency against the frequency of the 3 genotypes; each SNP is represented by 3 different coloured points. The solid lines 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[2]. 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).

[2] CODIS: Combined DNA Index System

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

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.

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

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

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 ≥ 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

to be a member of the Abenaki community. Given that allele frequencies vary among populations, why would people be concerned about using data from a non-Abenaki population to compute a false match probability?

2.2 Allele sharing among related individuals and Identity by Descent

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), while some individuals will be more closely related (e.g. sibling/first cousins). All individuals are related to one another by varying levels of relatedness, or *kinship*. Related individuals can share alleles that have both descended from the shared common ancestor. To be shared, 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 share more alleles. In Figure 2.6 we show the sharing of chromosomal 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 individuals.

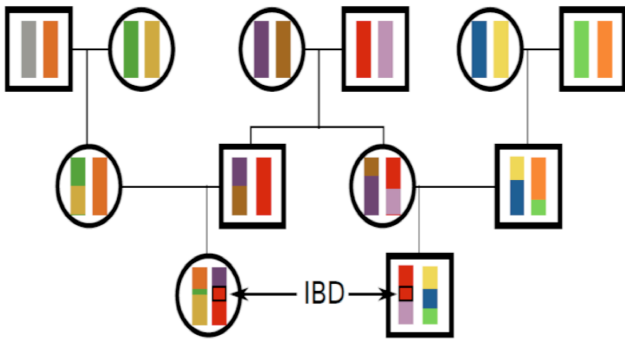

Figure 2.6: First cousins sharing a stretch of chromosome identical by descent. The different grandparental diploid chromosomes are coloured so we can track them and recombinations between them across the generations. Notice that the identity by descent between the cousins persists for a long stretch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

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[3]. For the moment, we ignore mutation, and we will be more precise about what we mean by 'past few generations' later on. For example, parent and child share exactly one allele identical by descent at a locus, assuming that the two parents of the child are randomly mated individuals from the population. In Figure 2.12, I show a pedigree demonstrating some configurations of IBD.

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

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

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

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

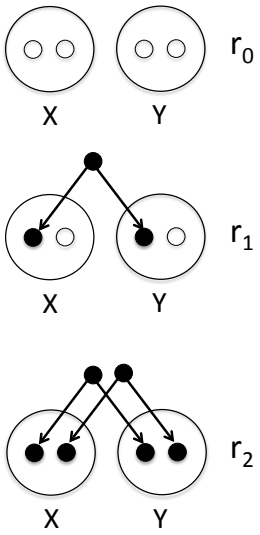

Figure 2.7: A pair of diploid individuals (X and Y) sharing 0, 1, or 2 alleles IBD where lines show the sharing of alleles by descent (e.g. from a shared ancestor).

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^{st} cousins	3/4	1/4	0	1/16

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.

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

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

Or, in our r_0, r_1, r_2 notation:

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

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

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

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

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

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

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

A) Assuming that they share no close relationship.

B) Assuming that they are full sibs.

C) Briefly explain your findings.

There's a variety of ways to estimate the relationships among individuals using genetic data. An example of using allele sharing to identify relatives is offered by the work of Nancy Chen (in collaboration with Stepfanie Aguillon, see Chen *et al.*, 2016; Aguillon *et al.*, 2017). Chen *et al.* has collected genotyping data from thousands of Florida Scrub Jays at over ten thousand loci. These Jays live at the Archbold field site, and have been carefully monitored for many decades allowing the pedigree of many of the birds to be known. Using these data she estimates allele frequencies at each locus. Then by equating the observed number of times that a pair of individuals share 0, 1, or 2 alleles to the theoretical expectation, she estimates the probability of r_0, r_1, and r_2 for each pair of birds. A plot of these are shown in Figure 2.9, showing how well the estimates match those known from the pedigree.

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

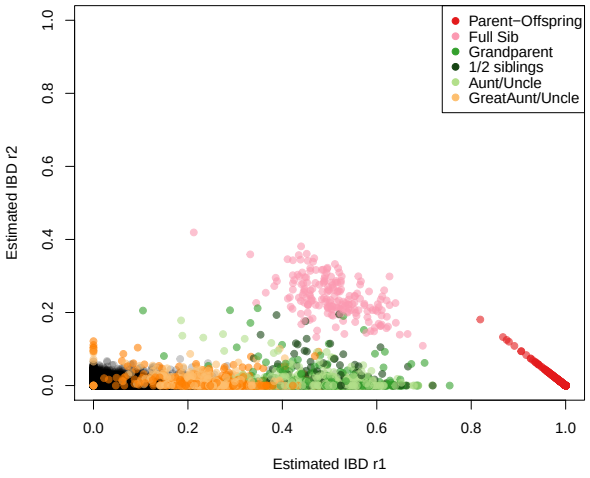

Figure 2.9: Estimated coefficient of kinship from Florida Scrub Jays. Each point is a pair of individuals, plotted by their estimated IBD (r_1 and r_2) from their genetic data. The points are coloured by their known pedigree relationships. Note that most pairs have low kinship, and no recent genealogical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

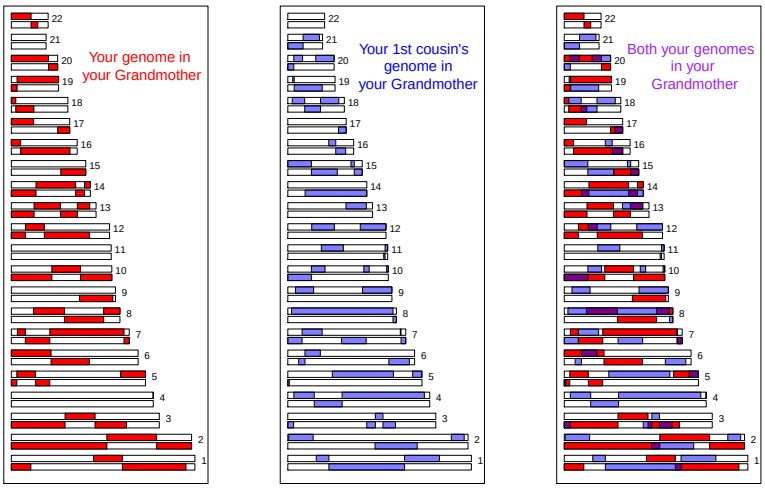

Figure 2.10: A simulation of sharing between first cousins. The regions of your grandmother's 22 autosomes that you inherited are coloured red, those that your cousins inherited are coloured blue. In the third panel we show the overlapping genomic regions in purple, these regions will be IBD in you and 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.

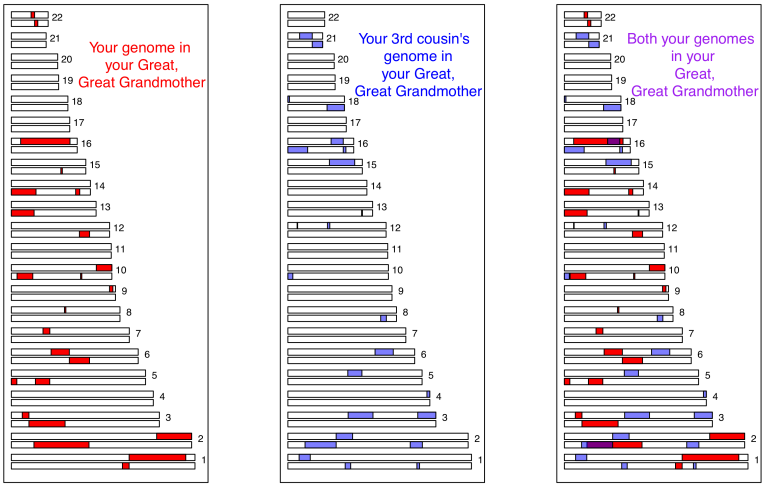

Figure 2.11: A simulation of sharing between third cousins, the details are the same as in Figure 2.10.

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. [4] More distant relations than third cousins, e.g. fourth cousins, start to have

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

a significant probability of sharing none of their genome IBD. But you have many fourth cousins, so you will share some of your genome IBD with some of them; however, it gets increasingly hard to identify the degree of relatedness from genetic data the deeper in the family tree this sharing goes.

2.2.1 Inbreeding

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

When two related individuals produce an offspring, that individual 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.

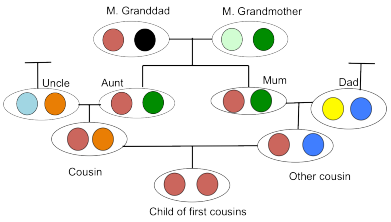

Figure 2.12: Alleles being transmitted through an inbred pedigree. The two sisters (mum and aunt) share two alleles identical by descent (IBD). The cousins share one allele IBD. The offspring of first cousins is homozygous by descent at this locus.

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}
$(1-F)p^2+Fp$	$(1-F)2pq$	$(1-F)q^2+Fq$

Table 2.5: **Generalized Hardy–Weinberg**

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

$$(1-F)2pq, \tag{2.7}$$

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

$$(1-F)p^2+Fp. \tag{2.8}$$

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

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

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

Multiple inbreeding loops in a pedigree. Up to this point we have assumed that there is at most one inbreeding loop in the recent family history of our individuals, i.e. the parents of our inbred individual have at most one recent genealogical connection. However, an individual who has multiple inbreeding loops in their pedigree can be homozygous by descent thanks to receiving IBD alleles via multiple different different loops. To calculate inbreeding in pedigrees of arbitrary complexity, we can extend beyond our original relatedness 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 that *both* of the alleles in the first individual are shared IBD with one allele in the second individual? There are nine possible relatedness 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 of. However, we will show how we can calculate the inbreeding coefficient of an individual with multiple inbreeding loops more directly.

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 family trees. We denote these shared ancestors by $A_1, \ldots, A_K$, and we denote by n the total number of individuals in the chain from B to C via ancestor A_i, including B, C, and A_i. For example, if B is C's aunt, then B and C share two ancestors, which are B's parents and, equivalently, C's grandparents. In this case, there are n=4 individuals from B to C through each of these two shared ancestor. In the general case, the kinship coefficient of B and C, i.e. the inbreeding coefficient of their child, is

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

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. [5] However, even if B inherits the maternal allele and C inherits the paternal allele of shared ancestor A_i, if A_i was themselves inbred, with probability f_{A_i} those two alleles are themselves IBD. Thus a shared *inbred* ancestor further increases the kinship of B and C.

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

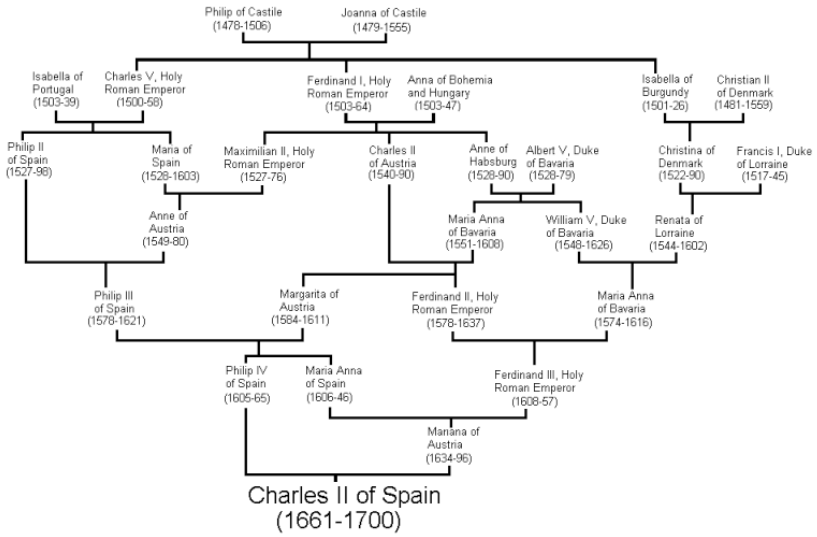

Figure 2.13: The pedigree of King Charles II of Spain. Pedigree from wikimedia drawn by Lec CRP1, public domain.

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

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

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

a child of an uncle-niece marriage.

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

[6] 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.

2.2.2 Calculating inbreeding coefficients from genetic data

If the observed heterozygosity in a population is H_O, and we assume 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 inbreeding coefficient as

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

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

$$\hat{F} = \frac{H_E - H_O}{H_E} = 1 - \frac{H_O}{H_E}. \tag{2.11}$$

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

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

AA	AB	BB
0.6	0.2	0.2

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

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

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

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 lobos 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.

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.

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 predigree 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

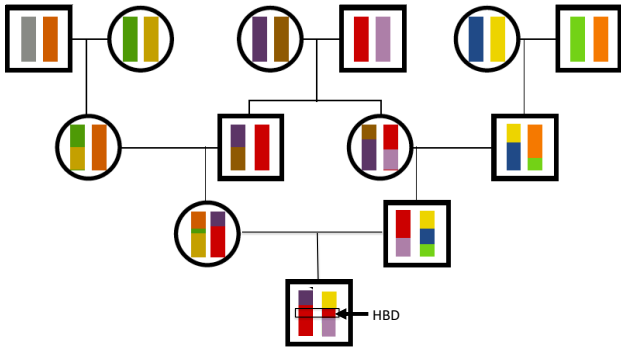

Figure 2.16: .

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

Individuals with multiple inbreeding loops in their family tree can have a high inbreeding coefficient due to the combined effect of many

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

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

An example of using F_{ROH} to study inbreeding comes from the work of Sams and Boyko (2018b), who identified runs of homozygosity in 2,500 dogs, ranging from 500kb up to many megabases. Figure 2.18 shows the distribution of F_{ROH} of individuals in each dog breed for the X and autosome. In Figure 2.19 this is broken down by the length of ROH segments.

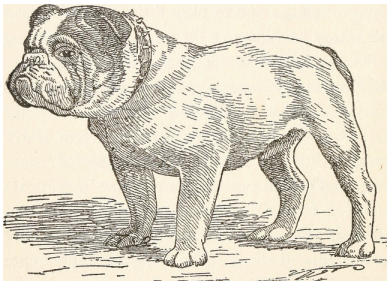

Figure 2.17: English bulldog. The dogs of Boytown. 1918. Dyer, W. A.

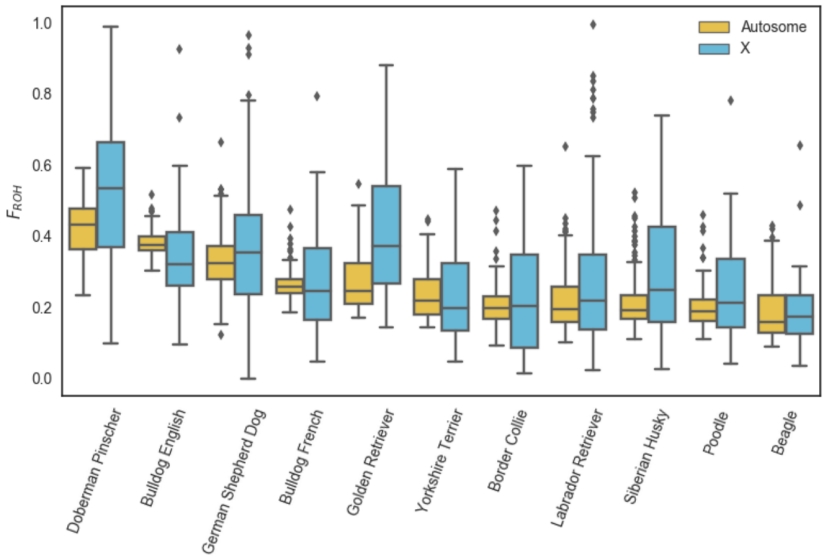

Figure 2.18: The distribution of F_{ROH} of individuals from various dog breeds from Sams and Boyko (2018a), licensed under CC BY 4.0.

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

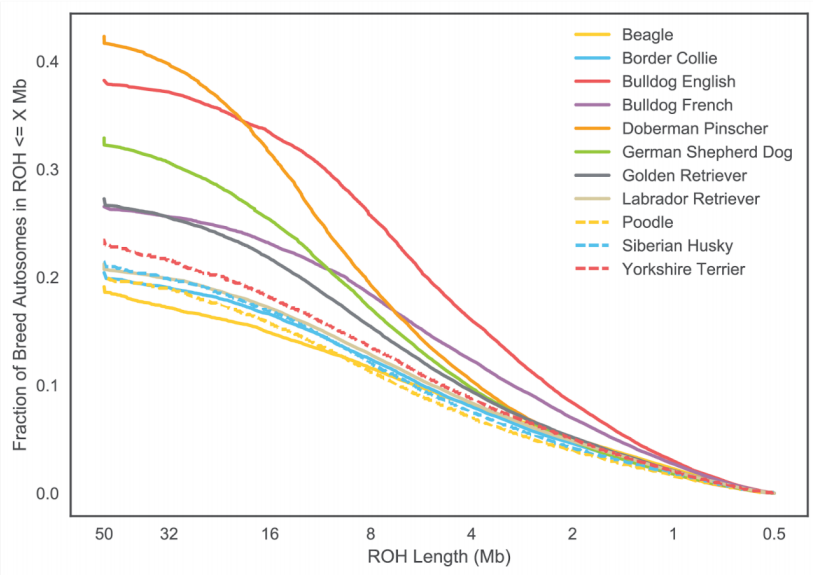

Figure 2.19: Cumulative density of length of ROH length, measured in megabases (Mb) from SAMS and BOYKO (2018a) for various dog breeds (licensed under CC BY 4.0). Note that longer lengths of ROH are on the left of the plot.

2.3 Summarizing population structure

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

2.3.1 Inbreeding as a summary of population structure.

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

WRIGHT[7] 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

$$F_{\mathrm{IS}} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_S q_S}, \tag{2.12}$$

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

$$F_{\mathrm{IT}} = 1 - \frac{H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T}, \tag{2.13}$$

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}:

$$F_{\mathrm{ST}} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T}. \tag{2.14}$$

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

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

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

[7] 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

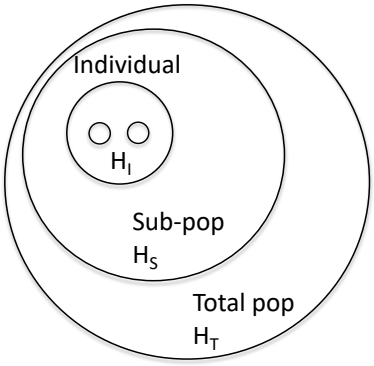

Figure 2.20: The hierarchical nature of F-statistics. The two dots within an individual represent the two alleles at a locus for an individual I. We can compare the heterozygosity on individuals (H_I), to that found by randomly drawing alleles from the sub-population (S), to that found in the total population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

K subpopulations (sampled with equal effort) is

$$F_{\mathrm{ST}} = 1 - \frac{\bar{H}_S}{H_T}, \tag{2.16}$$

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. If the total population contains the subpopulation then $2p_S q_S \leq 2p_T q_T$, and so $F_{\mathrm{IS}} \leq F_{\mathrm{IT}}$ and $F_{\mathrm{ST}} \geq 0$. Furthermore, if we have multiple sites, we can replace H_I, H_S, and H_T with their averages across loci (as above). [8]

[8] 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.

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).

These two species are spread across eastern Northern America, with the golden-winged warbler having a smaller, more northernly range. They're quite different in terms of plumage, but have long been known to have similar songs and ecologies. The two species hybridize readily in the wild; in fact two other previously-recognized species, Brewster's and Lawrence's warbler (4 & 3 in 2.21), are actually found to just be hybrids between theses two species. The golden-winged warbler is listed as 'threatened' under the Canadian endangered species act. The golden-winged warbler's habitat is under pressure from human activity and increased hybridization with the blue warbler, which is moving north into its range, also poses a significant issue. Toews *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 $F_{ST} = 0.0045$. 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

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.

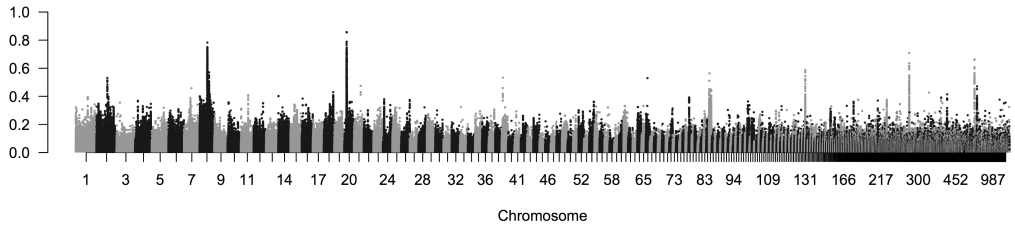

Figure 2.22: FST between blue- and golden-winged warbler population samples at SNPs across the genome. Each dot is a SNP, and SNPs are coloured alternating by scaffold. Thanks to David Toews for the figure.

high F_{ST} stand out. Nearly all of these regions correspond to large allele frequency difference at loci in, or close, to genes known to be 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

gene involved in feather development:

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

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

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, we may think about X as individuals and S as the subpopulation. Rewriting F_{IS} in terms of the observed homozygote frequencies (f_{11}, f_{22}) and expected homozygosities (p_S^2, q_S^2) we find

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

using the fact that $p^2 + 2pq + q^2 = 1$, and $f_{12} = 1 - f_{11} - f_{12}$. 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 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 from some broader population).

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

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

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

2.3.2 Other approaches to population structure

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

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. [9]

[9] 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

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

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

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

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

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

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

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

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

where

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

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 a equal probability of coming from each population $P(\text{pop k}) = 1/K$.

We interpret

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

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

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

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 that our individual comes from the Western population versus Eastern chimp population?

B) Lets assume that our individual is a hybrid. At each locus, with probability q_W our individual draws an allele from the Western 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 genotype given q_C?

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 to assign our individuals to a particular population, these ideas can be pushed to learn about how best to describe our genotype data in terms of discrete populations without assigning any of our individuals to populations *a priori*. We wish to cluster our individuals into K unknown populations. We begin by assigning our individuals at random to these K populations.

1. Given these assignments we estimate the allele frequencies at all of our loci in each population.

2. Given these allele frequencies we chose to reassign each individual to a population k with a probability given by eqn. (2.20).

We iterate steps 1 and 2 for many iterations (technically, this approach is known as *Gibbs Sampling*). If the data is sufficiently informative, the assignments and allele frequencies will quickly converge on a set of likely population assignments and allele frequencies for these populations.

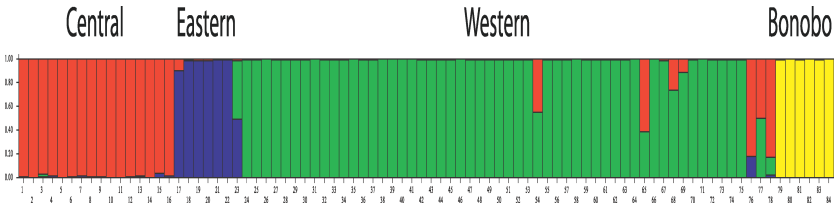

Figure 2.23: Becquet *et al.* (2007) genotyped 78 common chimpanzee and 6 bonobo at over 300 polymorphic markers (in this case microsatellites). They ran STRUCTURE to cluster the individuals using these data into $K = 4$ populations. In Becquet *et al.* (2007) above figure they show each individual as 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.

To do this in a full Bayesian scheme we need to place priors on the allele frequencies (for example, one could use a beta distribution prior). Technically we are using the joint posterior of our allele frequencies and assignments. Programs like STRUCTURE, use this type of algorithm to cluster the individuals in an "unsupervised" manner (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 STRUCTURE to determine the population structure of chimpanzees.

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

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

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

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

2.3.4 Principal components analysis

Principal component analysis (PCA) is a common statistical approach to visualize high dimensional data, and used by many fields. The idea of PCA is to give a location to each individual data-point on each of a small number principal component axes. These PC axes are chosen to reflect major axes of variation in the data, with the first PC being that which explains largest variance, the second the second most, and so on. The use of PCA in population genetics was pioneered by Cavalli-Sforza and colleagues and now with large genotyping datasets, PCA has made come back. [10]

[10] 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

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

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

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

i.e. at each SNP we center the genotypes by subtracting the mean genotype ($2p_\ell$) and divide through by the square root of the expected variance assuming that alleles are sampled binomially from the mean frequency ($\sqrt{2p_\ell(1-p_\ell)}$). Doing this to all of our genotypes, we form a data matrix (of dimension $N \times S$). We can then perform principal 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 2.23.

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

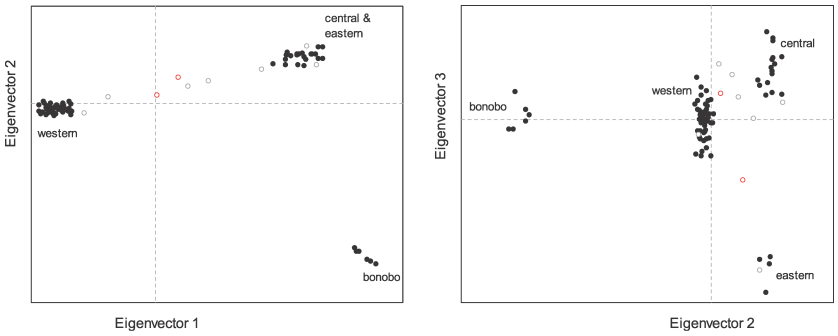

Figure 2.24: Principal Component Analysis by BECQUET *et al.* (2007) using the same chimpanzee data as in Figure 2.23. Here BECQUET *et al.* (2007) plot the location of each individual on the first two principal components (called eigenvectors) in the left panel, and on the second and third principal 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

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

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.

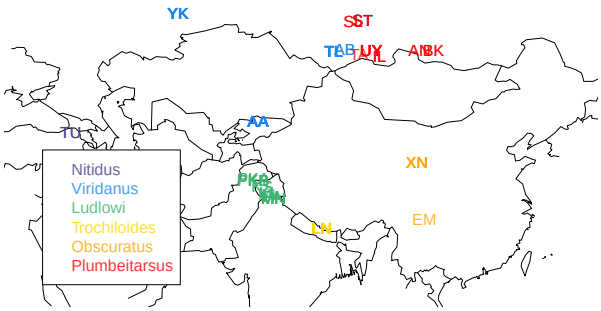

Figure 2.25: The sampling locations of 22 populations of Greenish warblers from Alcaide *et al.* (2014). The samples are coloured by the subspecies. Code here.

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

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

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

Figure 2.26: Greenish warbler, subspp. 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.)

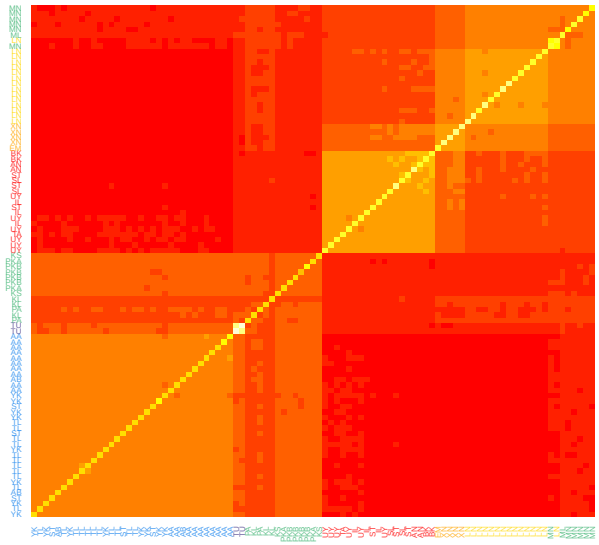

Figure 2.27: The matrix of kinship coefficients calcuated for the 95 samples of Greenish warblers. Each cell in the matrix gives the pairwise kinship coefficient calculated for a particular pair. Hotter colours indicating higher kinship. The x and y labels of individuals are the population labels from 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.

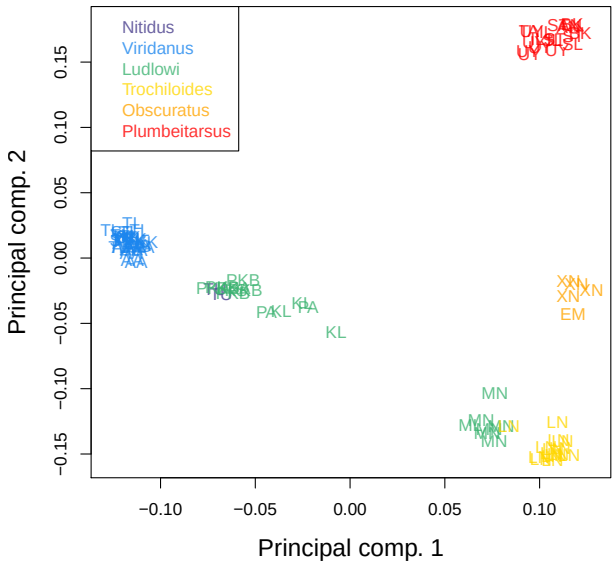

Figure 2.28: The 95 greenish warbler samples plotted on their locations on the first two principal components. The labels of individuals are the population labels from Figure 2.25, and coloured by subspecies label as in that figure. Code here.

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

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

2.3.5 Correlations between loci, linkage disequilibrium, and recombination

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 between parents affects the extent to which their offspring is inbred. We now turn to correlations between alleles at different loci.

Recombination To understand correlations between loci we need to understand recombination a bit more carefully. Let us consider a heterozygous individual, containing AB and ab haplotypes. If no recombination occurs between our two loci in this individual, then these two haplotypes will be transmitted intact to the next generation. While if a recombination (i.e. an odd number of crossing over 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 an aB haplotype. Effectively, recombination breaks up the association between loci. We'll define the recombination fraction (r) to be the 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 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 recombination rate (in Morgans) per base pair (typically $\sim 10^{-8}$) and L is the number of base pairs separating our two loci.

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

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

$$D_{AB} = p_{AB} - p_A p_B \tag{2.26}$$

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

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

$$p_{AB} = p_A p_B + D. \tag{2.27}$$

If $D = 0$ we'll say the two loci are in linkage equilibrium, while if $D > 0$ 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 before making this choice). You should be careful to keep the concepts 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 are on the same chromosome). Linkage disequilibrium merely refers to the covariance between the alleles at different loci; this may in part be due to the genetic linkage of these loci but does not necessarily imply this (e.g. genetically unlinked loci can be in LD due to population structure).

Question 12. You genotype 2 bi-allelic loci (A & B) segregating in two mouse subspecies (1 & 2) which mate randomly among themselves, but have not historically interbreed since they speciated. On 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 each population are:

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

A) How much LD is there within species? (i.e. estimate D)

B) If we mixed individuals from the two species together in equal proportions, we could form a new population with p_{AB} equal to the average frequency of p_{AB} across species 1 and 2. What value would D take in this new population before any mating has had the chance to occur?

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

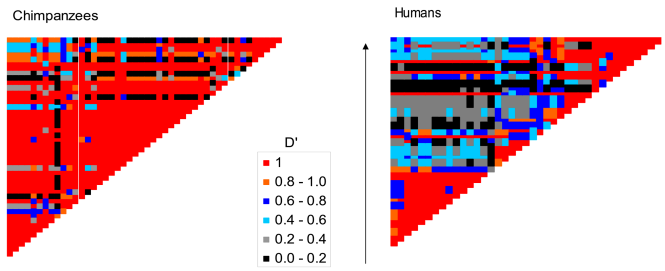

Figure 2.29: LD across the TAP2 gene region in a sample of Humans and Chimps, from PTAK *et al.* (2004), licensed under CC BY 4.0. The rows and columns are consecutive SNPs, with each cell giving the absolute D' value between a pair of SNPs. Note that these are different sets of SNPs in the two species,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

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

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

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

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 that this r in r^2 is NOT the recombination fraction.

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 faster than it can decay via recombination; more distant markers have much lower LD as here recombination is winning out. Note the decay of LD is much slower in the advanced-generation cross population than in the natural wild-caught population. This persistence of LD across megabases is due to the limited number of generations for recombination since the cross was created.

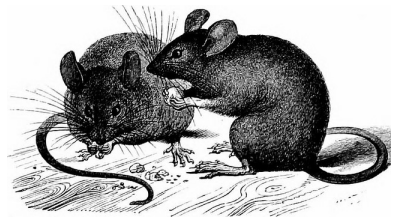

Figure 2.30: *Mus musculus*.
A history of British quadrupeds, including the Cetacea. 1874. Bell T., Tomes, R. F.m Alston E. R. Image from the Biodiversity Heritage Library. Contributed by Cornell University Library. No known copyright restrictions.

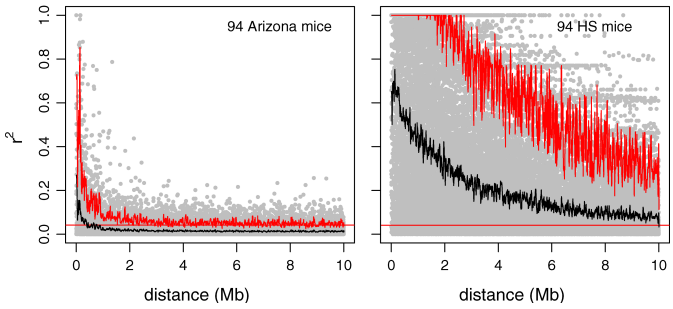

Figure 2.31: The decay of LD for autosomal SNP in *Mus musculus domesticus*, as measured by r^2, in a wild-caught mouse population from Arizona and a set of advanced-generation crosses between inbred lines of lab mice. Each dot gives the r^2 for a pair of SNPs a given physical distance apart, for a total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

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

The decay of LD due to recombination We will now examine what happens to LD over the generations if 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 a fraction r of our AB haplotypes to recombination ripping our alleles apart but gain a fraction rp_Ap_B per generation from other haplotypes recombining together to form AB haplotypes. Thus in the next generation

$$p'_{AB} = (1-r)p_{AB} + rp_Ap_B \tag{2.29}$$

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

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

$$\Delta p_{AB} = p'_{AB} - p_{AB} = -rp_{AB} + rp_Ap_B = -rD \tag{2.30}$$

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

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

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

$$D_t = (1-r)^t D_0 \tag{2.32}$$

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

$$D_t \approx D_0 e^{-rt} \tag{2.33}$$

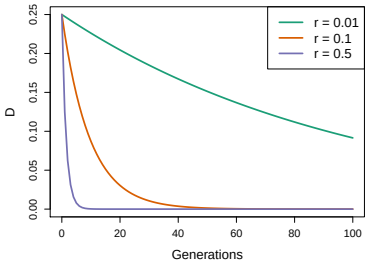

Figure 2.32: The decay of LD from an initial value of $D_0 = 0.25$ over time (Generations) for a pair of loci a recombination fraction r apart. Code here.

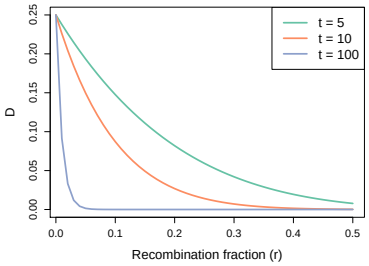

Figure 2.33: The decay of LD from an initial value of $D_0 = 0.25$ due to recombination over t generations, plotted across possible recombination fractions (r) between our pair of loci. Code here.

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. 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?

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.*.

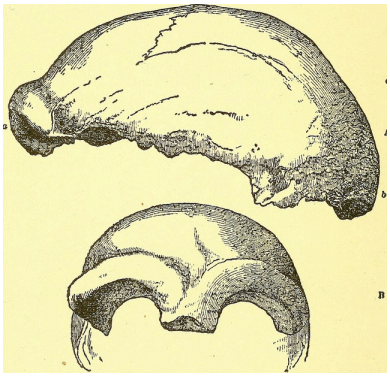

Figure 2.34: The earliest discovered fossil of a Neanderthal, fragments of a skull found in a cave in the Neander Valley in Germany.

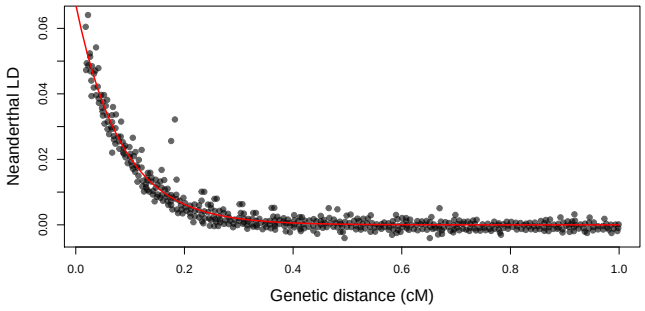

Figure 2.35: The LD between putative-Neanderthal alleles in a modern European population (the CEU sample from the 1000 Genomes Project). Each point represents the average D statistic between a pair of alleles at loci at a given genetic distance apart (as given on the x-axis and measured in centiMorgans (cM)). The putative Neanderthal alleles are alleles where the Neanderthal genome has a derived allele that is at very low frequency in an 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.

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.

3
Genetic Drift and Neutral Diversity

Randomness is inherent to evolution, from the lucky birds blown of course to colonize some new oceanic island, to which mutations arise first in the HIV strain infecting an individual taking anti-retroviral drugs. One major source of stochasticity in evolutionary biology is genetic drift. Genetic drift occurs because more or less copies of an allele by chance can be transmitted to the next generation. This can occur because, by chance, the individuals carrying a particular allele can leave more or less offspring in the next generation. In a sexual population, genetic drift also occurs because Mendelian transmission means that only one of the two alleles in an individual, 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 populations, but it will play the biggest role for neutral alleles. A neutral polymorphism occurs when the segregating alleles at a polymorphic site have no discernible differences in their effect on fitness. We'll make clear what we mean by "discernible" later, but for the moment think of this as "no effect" on fitness.

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 and Gillespie, 1996, for a history).[1]. The central premise of Neutral theory theory is that patterns of molecular polymorphism within species and substitution between species can be well understood by supposing that the vast majority of these molecular polymorphisms and substitutions were neutral alleles, whose dynamics were just subject to the vagaries of genetic drift and mutation. Early proponents of this view suggested that the vast majority of new mutations are either neutral or highly deleterious (e.g. mutations that disrupt important protein functions). This latter class of mutations are too deleterious to contribute much to common polymorphisms or substitutions be-

[1] 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

tween species, because they are quickly weeded out of the population by selection.

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:

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).

2. Synonymous changes in coding regions, i.e. those that don't change the amino-acid encoded by a codon.

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.

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.

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.

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 $t+1$?

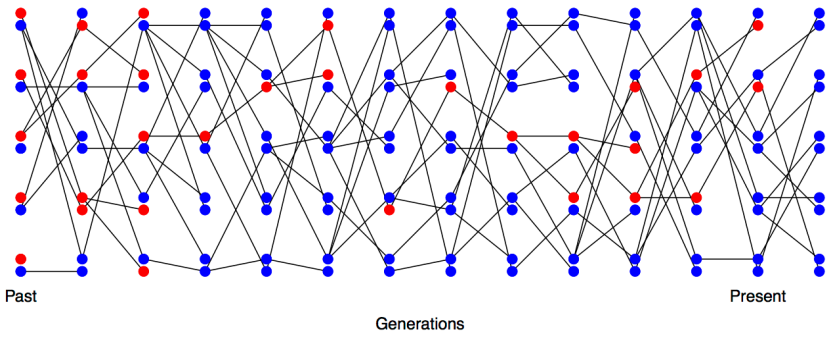

Figure 3.1: Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals over the generations, with lines showing transmission. In the first generation every individual is a heterozygote. Code here.

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

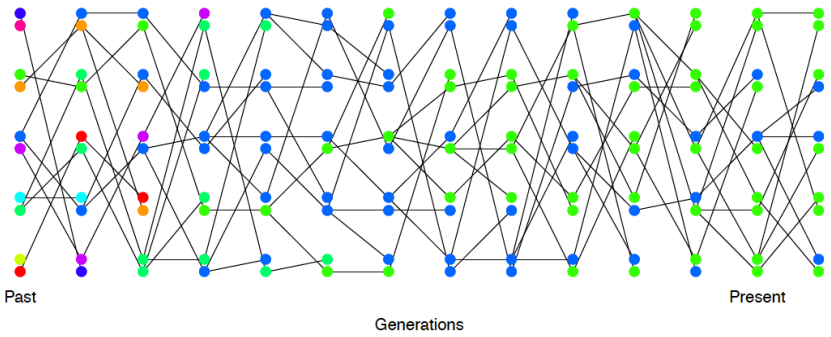

Figure 3.2: Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals. In the first generation I colour every allele a different colour so we can track their descendants. Code here.

alleles in generation t, are non-identical is H_t.

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

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

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

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

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

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

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

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

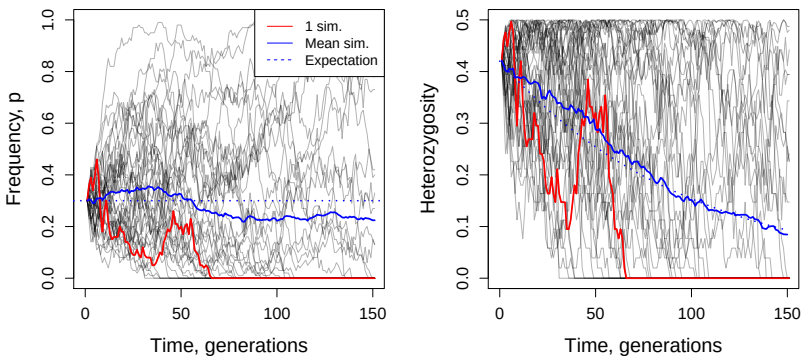

Figure 3.3: Change in allele frequency and loss of heterozygosity over time for 40 replicates. Simulations of genetic drift in a diploid population of 50 individuals, in the absence of new mutations. We start 40 independent, biallelic loci each with an initial allele at 30% frequency. The left panel 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?

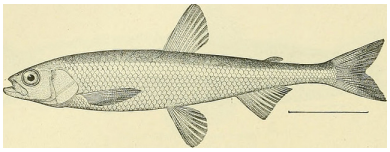

Figure 3.4: Pond smelt (*Hypomesus olidus*), a close relative of delta smelt. Bulletin of the United States Fish Commission. 1906. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

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 changed. 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-

lation frequency p_t in the previous generation. If you've read to here, please email Prof Coop a picture of JBS Haldane in a striped suit with the title "I'm reading the chapter 3 notes". (It's well worth googling JBS Haldane and to read more about his life; he's a true character and one of the last great polymaths.)

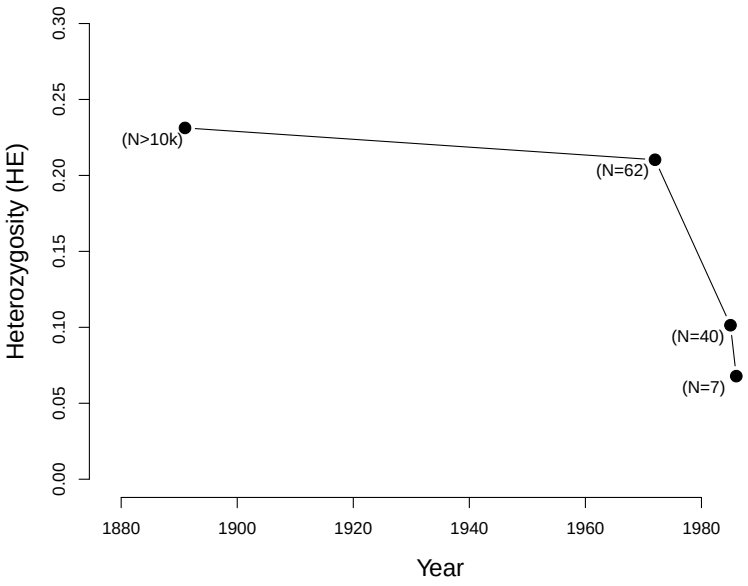

Figure 3.6: Loss of heterozygosity in the Black-footed Ferrets in their declining population. Numbers in brackets give estimated number of individuals alve at that time. Data from WISELY *et al.* (2002). Code here.

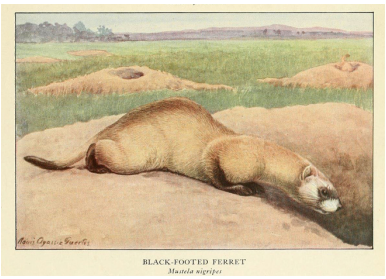

Figure 3.5: The black-footed ferret (*M. nigripes*).
Wild animals of North America, The National geographical society, 1918. Image from the Biodiversity Heritage Library. Contributed by American Museum of Natural History Library. Not in copyright.

To see how a decline in population size can affect levels of heterozygosity, let's consider the case of black-footed ferrets (*Mustela nigripes*). The black-footed ferret population has declined dramatically through the twentieth century due to destruction of their habitat. In 1979, when the last known black-footed ferret died in captivity, they were thought to be extinct. In 1981, a very small wild population was rediscovered (40 individuals), but in 1985 this population suffered a number of disease outbreaks. All of the 18 remaining wild individuals were brought into captivity, 7 of which reproduced. Thanks to intense captive breeding efforts and conservation work, a wild population of over 300 individuals has been established since. However, because all of these individuals are descended from those 7 individuals who survived the bottleneck, diversity levels remain low. WISELY *et al.* measured heterozygosity at a number of microsatellites in individuals from museum collections, showing the sharp drop in diversity as population sizes crashed (see Figure 3.6).

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

follows from the Taylor series expansion of $\exp(-x)$, ignoring second order and higher terms of x). This approximation is especially useful for approximating a geometric decay process by an exponential decay process, e.g. $(1-x)^t \approx e^{-xt}$. Using your calculator, or R, check how good of an approximation this is compared to the exact expression for two values of x, $x = 0.1$, and 0.01, across two different values of t, $t = 5$ and $t = 50$. I.e. calculate both expressions for these values, hand in your answers and briefly comment on your results.

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

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

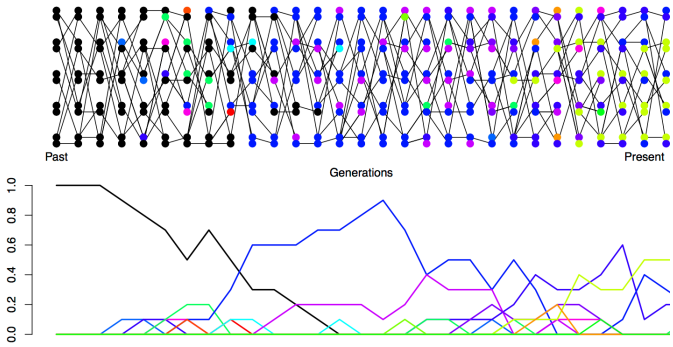

Figure 3.7: Mutation-drift balance. A diploid population of 5 individuals. In the first generation everyone has the same allele (black). Each generation the transmitted allele can mutate and we generate a new colour. In the bottom plot, I trace the frequency of alleles in our population over time. The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

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

$$\mu = (1-C)\mu_T \tag{3.4}$$

This is the per generation rate.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} P(\text{coal. in t+1 \& no mutations}) &\approx \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{t}(1-\mu)^{2t} && (3.8)\\ &\approx \frac{1}{2N}e^{-t/(2N)}e^{-2\mu t} && (3.9)\\ &= \frac{1}{2N}e^{-t(2\mu+1/(2N))} && (3.10) \end{aligned}$$

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:

$$H = \frac{4N\mu}{1 + 4N\mu} \tag{3.12}$$

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.

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

$$\theta = 4N\mu \tag{3.13}$$

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

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 .

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.

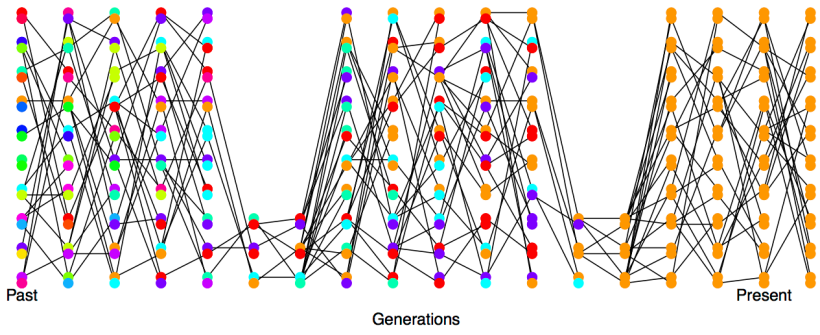

Figure 3.8: Loss of heterozygosity over time in a bottlenecking population. A diploid population of 10 individuals, that bottlenecks down to three individuals repeatedly. In the first generation, I colour every allele a different colour so we can track their descendants. There are no new mutations. Code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

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

If population sizes vary rapidly in size, we can (if certain conditions are met) replace our population size by the harmonic mean population 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 coalesced by generation t is given by

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

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

$$N_e = \frac{1}{\frac{1}{t}\sum_{i=1}^{t}\frac{1}{N_i}}. \tag{3.15}$$

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

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

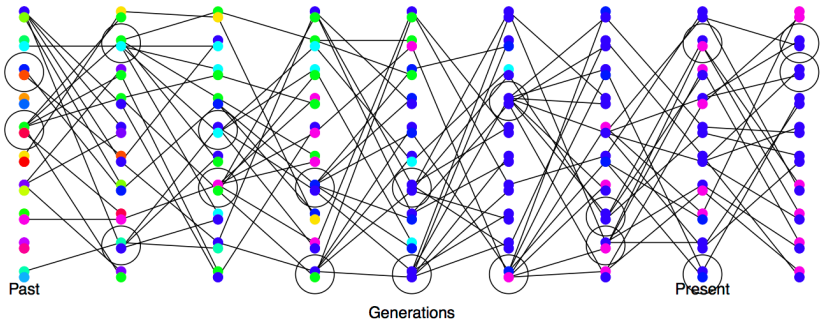

Figure 3.9: High variance on reproductive success increases the rate of genetic drift. A diploid population of 10 individuals, where the circled individuals have much higher reproductive success. In the first generation I colour every allele a different colour so we can track their descendants, there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

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

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

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

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

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

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

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).

has a mother an a father, not every individual gets to be a parent. In practice, in many animal species far more females get to reproduce than males, i.e. $NM < N_F$, as a few males get many mating opportunities and many males get no/few mating opportunities (see JANICKE *et al.*, 2016, for a broad analysis, and note that there a certainly many exceptions to this general pattern). When our two alleles pick an ancestor, 25% of the time our alleles were both in a female ancestor, 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 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 and so cannot coalesce. Therefore, our probability of coalescence in the preceding generation is

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

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

$$N_e = \frac{4N_F N_M}{N_F + N_M} \tag{3.19}$$

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 result. For more on how different evolutionary forces affect the rate of genetic drift, and their impact on the effective population size, see CHARLESWORTH (2009).

Figure 3.10: Male Hamadryas baboons. 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.

Question 5. You are studying a population of 500 males and 500 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 size for the autosome?

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

3.2 The Coalescent and patterns of neutral diversity

> "Life can only be understood backwards; but it must be lived forwards." – Kierkegaard

Pairwise Coalescent time distribution and the number of pairwise differences. Thinking back to our calculations we made about the 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

generation a pair of sequences coalesce in, and then calculate some properties of heterozygosity based on that. That's because neutral 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 lineages back through the generations.

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

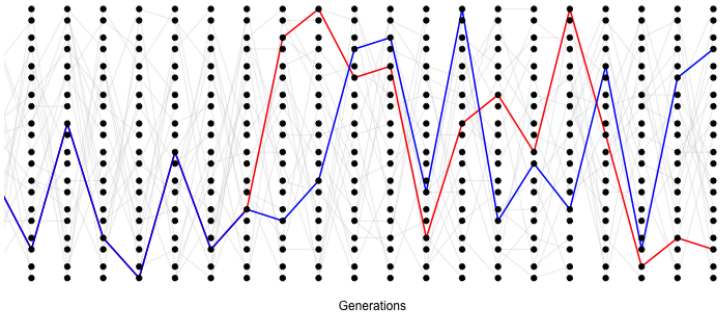

Figure 3.11: A simple simulation of the coalescent process. The simulation consists of a diploid population of 10 individuals (20 alleles). In each generation, each individual is equally likely to be the parent of an offspring (and the allele transmitted is indicated by a light grey line). We track a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

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

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

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

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 μ, 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

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

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 [3] Thus the number of mutations between a pair of sites is the observed number of differences between a pair of sequences. In the previous chapter we denote the observed number of pairwise differences at putatively neutral sites separating a pair of sequences as π (we usually average this over a number of pairs of sequences for a region). Therefore, under our simple, neutral, constant population-size model we expect

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

So we can get an empirical estimate of θ from π, let's call this $\widehat{\theta}_\pi$, by setting $\widehat{\theta}_\pi = \pi$., i.e. our observed level of pairwise genetic diversity. If we have an independent estimate of μ, then from setting $\pi = \widehat{\theta}_\pi = 4N\mu$ 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. [4]

Lets take a moment to distinguish our expected heterozygosity (eqn. 3.12) from our expected number of pairwise differences (π). Our expected heterozygosity is the probability that two alleles at a locus, sampled from a population at random, are different from each other. If one or more mutations have occurred since a pair of alleles last shared a common ancestor, then our sequences will be different from each other. On the other hand, our π measure keeps track of the average total number of differences between our loci. As such, π is often a more useful measure, as it records the number of differences between

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

[4] Up to this point we've been describing only neutral processes, however, selection can also alter levels of polymorphism. For example, if some synonymous sites directly experience selection, then even if we use π calculated for 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 π. But even then caution is warranted.

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

Figure 3.12: Gray Fox, *Urocyon cinereoargenteiis*.
Diseases and enemies of poultry. Pearson and Warren. (1897) Image from the Biodiversity Heritage Library. Contributed by University of California Libraries. Not in copyright.

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

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

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

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

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

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

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

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

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

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

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

3.3 The coalescent process of a sample of alleles.

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

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

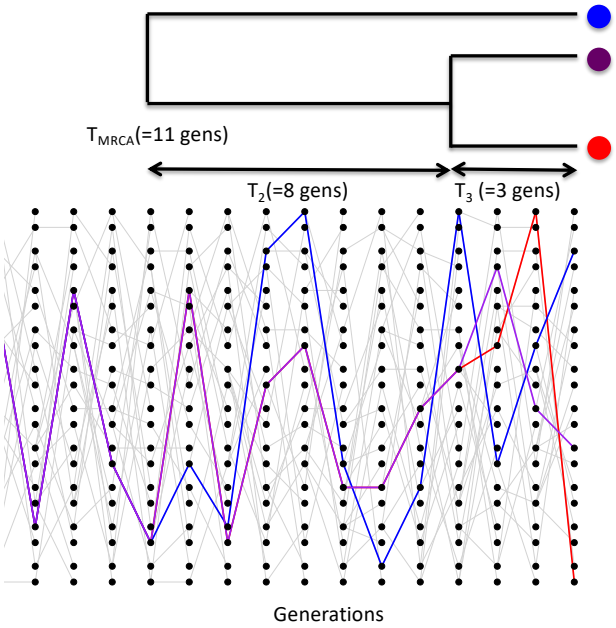

Figure 3.13: A simple simulation of the coalescent process for three lineages. We track the ancestry of three modern-day alleles, the first pair (blue and purple) coalesce four generations back, after which there are only two independent lineages we are tracking. This pair then coalesces twelve generations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

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

pairs of alleles, so the probability that no pair finds a common ancestor in the preceding generation is

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

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

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

[5] said as "i choose 2"

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

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

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

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

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

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

$$T_i \sim \text{Geo}\left({}^{\binom{i}{2}}/_{2N}\right). \tag{3.30}$$

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

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

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 past. Thus when a pair of alleles in our sample of i alleles coalesces, we then switch to having to follow $i-1$ alleles back in time. Then when a pair of these $i-1$ alleles coalesce, we then only have to follow $i-2$ alleles back. This process continues until we coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).

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

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

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

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

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

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

3.3.1 Expected properties of coalescent genealogies and mutations.

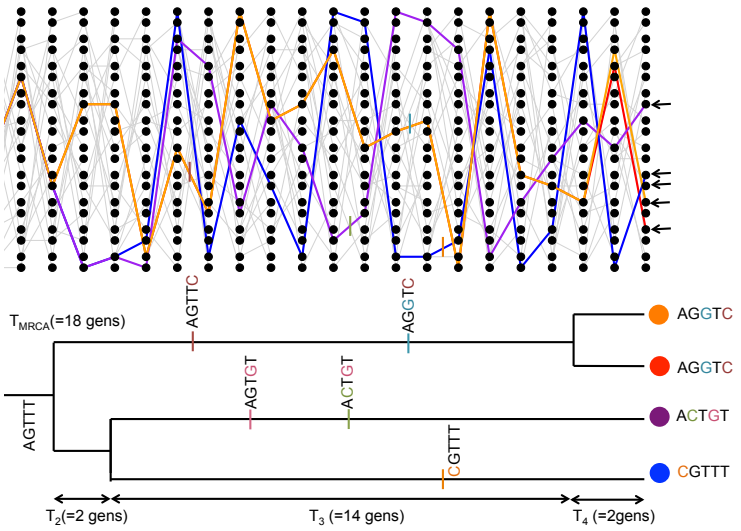

Figure 3.14: A simple coalescent tree from a single coalescent simulation, tracing the genealogy of 4 alleles with mutational changes marked with dashes showing transitions away from the MRCA sequence (AGTTT) . The T_{MRCA} is $T_4+T_3+T_2$. The total time in the tree is $T_{tot} = 4T_4 + 3T_3 + 2T_2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

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

$$T_{MRCA} = \sum_{i=n}^{2} T_i \tag{3.33}$$

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

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

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

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

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

$$T_{tot} = \sum_{i=n}^{2} iT_i \tag{3.36}$$

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,

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

we see that our expected total amount of time in the genealogy scales linearly with our population size N. Our expected total amount of 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 coalescence usually happens very rapidly, so that extra samples add little to the total amount of time in the genealogical tree.

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 $\int_1^{n-1} \frac{4N}{i} di = 4N \log(n-1)$.

We saw above that the number of mutational differences between a pair of alleles that coalescence T_2 generations ago was Poisson with 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 our genealogy will cause a segregating polymorphism in the sample (meeting our infinitely-many-sites assumption). Thus, if the total time 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 with mean μT_{tot}. Thus the expected number of segregating sites in a sample of size n is

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

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 formula to derive another estimate of the population scaled mutation rate θ, by setting our observed number of segregating sites in a sample (S) equal to this expectation. We'll call this estimator $\widehat{\theta}_W$:

$$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} 1/i} \tag{3.39}$$

This estimator of θ was devised by Watterson (1975), hence the 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 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 of sites with these different allele frequencies depends on the lengths of specific genealogical branches. A mutation that falls on a branch 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 $T_3 + 2T_2$, the total length of the branch ancestral to just the orange and red allele ($T_3 + T_2$) plus the branch ancestral to just the blue and purple allele (T_2).

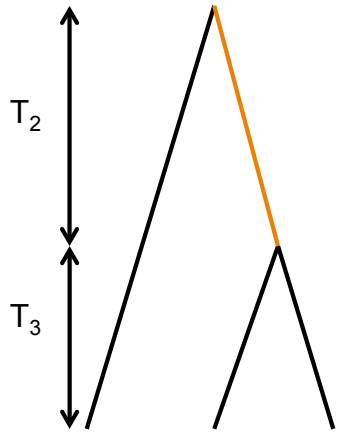

Figure 3.15: A tree for three samples; note that this is the only possible tree shape (treating the tips as unlabeled, i.e. I don't care which pair of sequences carry a doubleton, just that any two sequences carry a derived allele).

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

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

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:

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

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

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

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

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

$$D = \frac{\theta_\pi - \theta_W}{C} \tag{3.43}$$

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

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

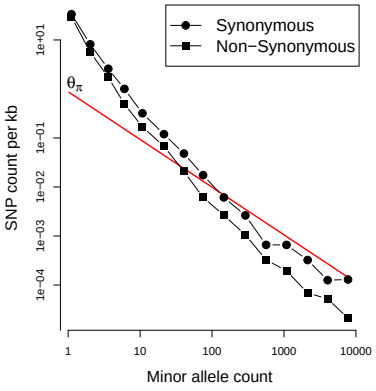

Figure 3.16: Data from 202 genes from 14002 people of European ancestry (28004 alleles). Note the double log-scale. Redrawn from NELSON *et al.*. The red line gives the neutral, constant population size estimate of the site frequency spectrum, our equation (3.41), using a θ estimated from π. Note how 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. Code here.

3.3.2 Demography and the coalescent

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

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

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

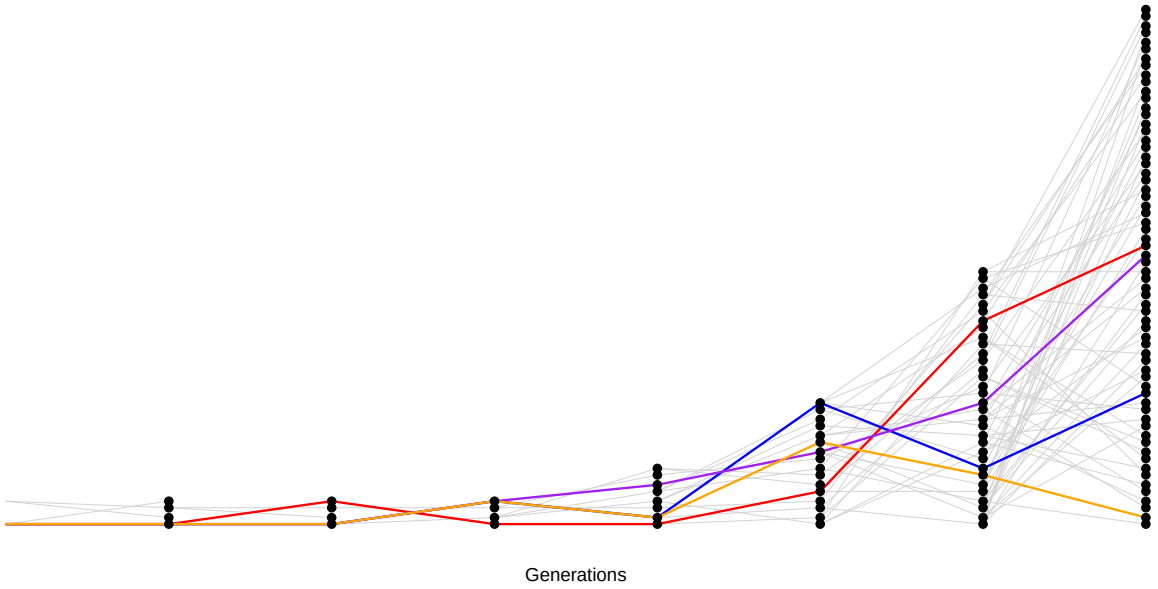

Figure 3.17: A realization of the coalescent process in a growing population. The population underwent a period of doubling every generation. The initial population size of just two individuals, maintained for a number of generations, is obviously highly unrealistic but serves our purpose. Code here.

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).

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; there are 7 billion of us alive today, but this is due to very rapid population growth over the past thousand to tens of thousands of years. Our level of genetic diversity is very much lower than you'd predict given our census size, reflecting our much smaller ancestral population. 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 branches are the source of alleles at more intermediate frequencies, and so there are even fewer intermediate-frequency alleles in growing populations. That's why there are so many rare alleles, especially singletons, in this large sample of Europeans.

Another common demographic scenario is a population bottleneck. In a bottleneck, the population size crashes dramatically, and sub-

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

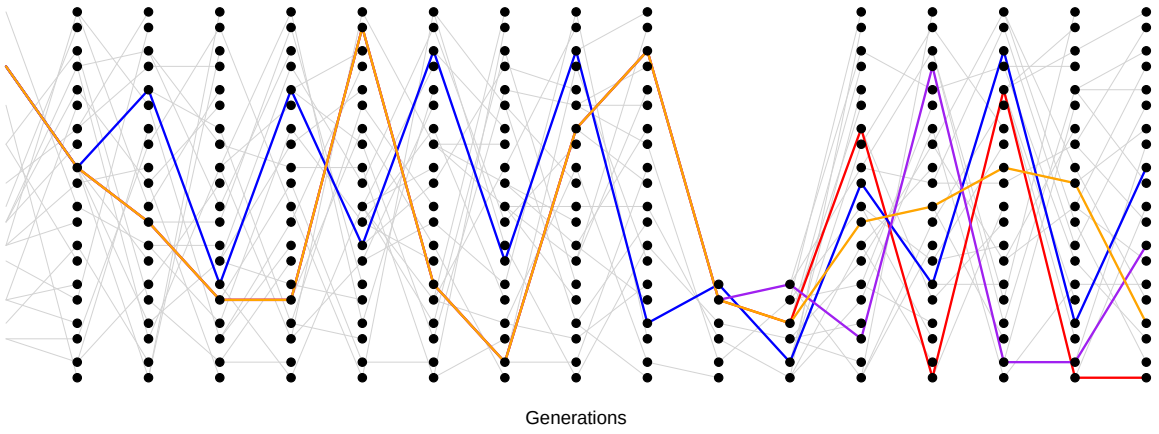

Figure 3.18: A realization of the coalescent process in a bottlenecked population. Our population under went a bottleneck eight generations in the past. Code here.

population today, if the bottleneck was somewhat recent ($\ll N_{\text{Big}}$ generations in the past) many of our lineages will not have coalesced before reaching the bottleneck, moving backward in time. But during the bottleneck our lineages coalesce at a much higher rate, such that many of our lineages will coalesce if the bottleneck lasts long enough ($\sim N_{\text{Small}}$ generations). If the bottleneck is very strong, then all of our lineages will coalesce during the bottleneck, and the resulting site frequency spectrum may look very much like our population growth 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).

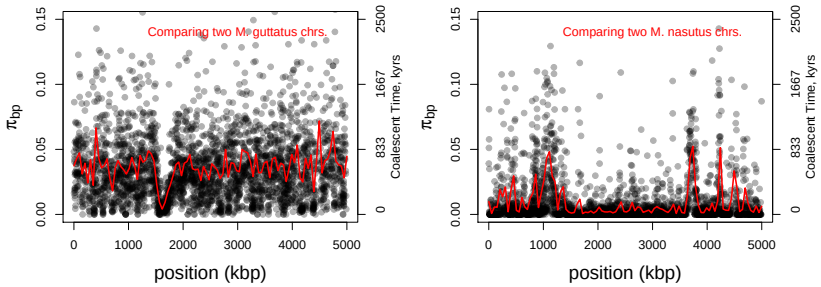

Figure 3.19: Diversity along the Mimulus genome. Black dots give π in 1kb windows between chromosomes sampled from two individuals, the red line is a moving average (data from Brandvain *et al.*). Pairwise coalescent times (t) estimated assuming $t = \pi/2\mu$ using $\mu_{BP} = 10^{-9}$. Code here.

An example of this is shown Figure 3.19, data from Brandvain *et al.*. *Mimulus nasutus* is a selfing species that arose recently from an out-crossing progenitor *M. guttatus*, and experienced a strong bottleneck. *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 3.20: 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.

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

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

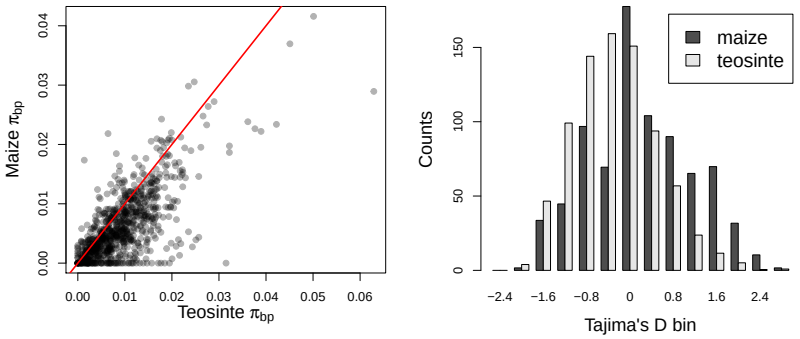

Figure 3.21: Data for polymorphism from Maize and Teosinite: 774 genes redrawn from WRIGHT *et al.* (2005). **Left)** Genetic diversity levels in maize and and Teosinte samples at each of these genes. Note how diversity levels are lower in maize than teosinte, i.e. most points are below the red $x =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

Mutations that arise on deeper lineages will be at intermediate frequency in our sample, and so mild bottlenecks can lead to an excess of intermediate frequency alleles compared to the standard constant-population model. This can skew Tajima's D, see eqn 3.43, towards positive values and away from its expectation of zero . One example of this skew is shown in Figure 3.21. Maize (*(Zea mays* subsp.*mays*) was domesticated from its wild progenitor teosinte (*(Zea mays subsp. parviglumis*) roughly ten thousand years ago. We can see how the bottleneck associated with domestication has resulted in a loss of genetic diversity in maize, compared to teosinte, and the polymorphism that remains is somewhat skewed towards intermediate frequencies resulting in more positive values of Tajima's D.

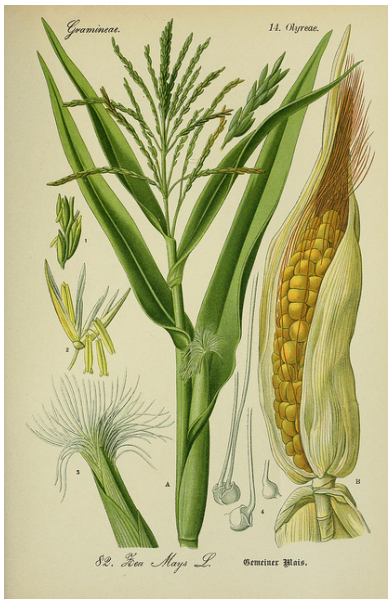

Figure 3.22: Maize (*Zea mays.*)
Prof. Dr. Thomé's Flora von Deutschland. 1886. Thomé, O. W. Image from the Biodiversity Heritage Library. Contributed by New York Botanical Garden. Not in copyright.

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

3.4 Molecular Evolution and the fixation of neutral alleles

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

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

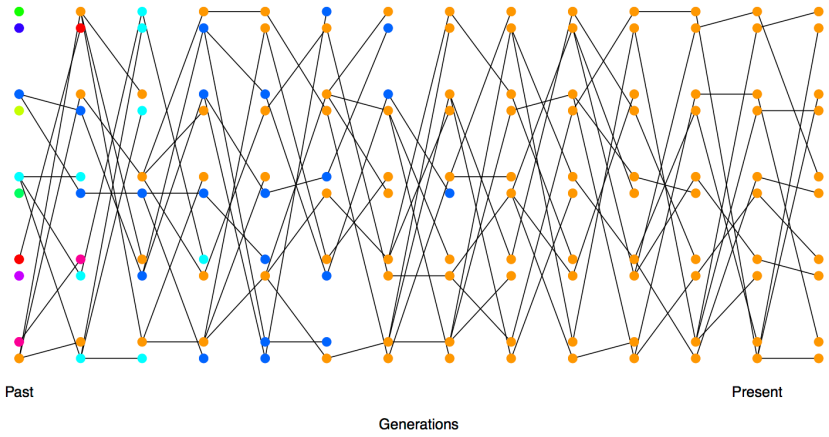

Figure 3.23: Each allele initially present in a small diploid population is given a different colour so we can track their descendants over time. By the 9th generation, all of the alleles present in the population can trace their ancestry back to the orange allele. Code here.

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

in this simulation it does become the common ancestor, by the 9th generation).

More generally, if our neutral allele is present in i copies in the population, of $2N$ alleles, the probability that this allele becomes fixed is $i/(2N)$, i.e. the probability that a neutral allele is eventually fixed is simply given by its frequency (p) in the population. (We can also derive this result by letting $Ns \rightarrow 0$ in eqn. (7.11), a result we'll encounter later.)

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 above, this occurs with probability $1/(2N)$.

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 ancestry back to a single most recent common ancestral allele. Any single-base pair change which arose as a single mutation at a locus, and fixed in the population, must have been present in the sequence transmitted by the most recent common ancestor of the population at that locus. Thus it must take roughly $4N$ generations for a neutral allele present in a single copy within the population to the ancestor of all alleles within our population. This argument can be made more precise, but in general we would still find that it takes $\approx 4N$ generations for a neutral allele to go from its introduction to fixation with the population.

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

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

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

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

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

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

$$2\mu(1-C)T \tag{3.45}$$

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

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

"functionally less important molecules or parts of a molecule evolve faster than more important ones."
– Kimura and Ohta (1974)

A number of observations follow under this model, from equation (3.45), the first is that a primary determinant of patterns of molecular evolution in a genomic region is the level of constraint (C). This pattern generally seems to hold empirically: non-coding regions often evolve more rapidly than coding regions; synonymous substitutions accumulate faster than nonsynonymous; nonsynonymous changes faster in less vital proteins than ones that are absolutely necessary for early development. Note that this is not a unique prediction of the neutral model, e.g. lower pleiotropy means that less constrained regions may be better able to evolve adaptively. However, it is a fantastically useful general insight, e.g. it allows us to spot putatively functional non-coding regions by looking for genomic regions that have very low levels of divergence among distantly related species.

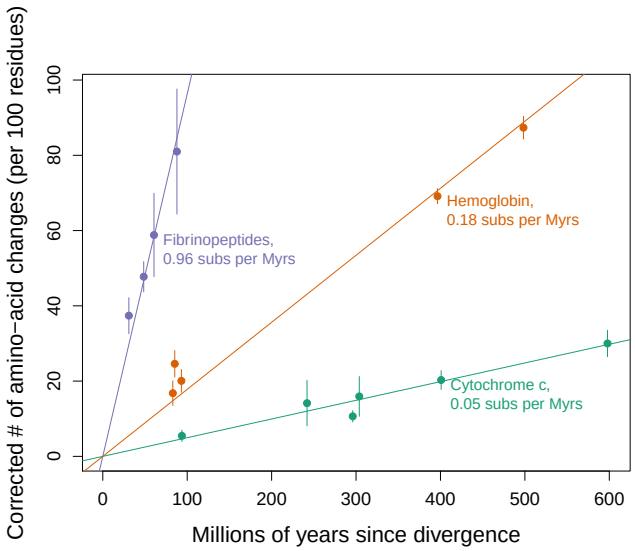

Figure 3.24: The numbers of substitutions between various pairs of groups, for three proteins, plotted against the time these groups shared a common ancestor in the fossil record. Data from DICKERSON (1971). The number of observed amino-acid differences is corrected for multiple hits to obtain the corrected 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.

The second important insight, and critical for the development of 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 observation that for some proteins there's a linear relationship between the number of non-synonymous substitutions and the time species last shared a common ancestor in the fossil record. DICKERSON (1971) provided an for early example of this observation (Figure 3.24), by comparing various organisms whose molecular sequences were available to him. For example, he found that humans and rattlesnakes, who last share a common ancestor in the fossil record around 300 million years, are separated by roughly 15 NS substitutions per 100 sites in the Cytochrome c protein. While, humans and dog fish, which diverged around 400 million years, are separated by 19 NS substitutions per 100 sites in this gene.

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

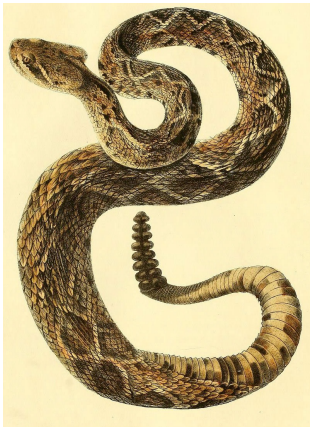

Figure 3.25: Eastern diamondback rattlesnake (*Crotalus adamanteus*. North American herpetology. Holbrook, J. E. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Licensed under CC BY-2.0.

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

$$T = T_s + 2N_A. \tag{3.46}$$

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

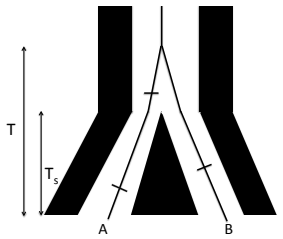

Figure 3.26: The genealogy of two alleles one sampled from population A and B. Mutations on the lineages are shown as dashes. The pair of alleles coalesce in the ancestral population of A and B. The two populations split T_S generations ago, with no subsequent gene flow, but the two lineages must coalesce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

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

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.

3.5 Tests of molecular evolution.

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

$$d_N = 2T(1 - C)\mu \tag{3.47}$$

Dividing by d_S, we find

$$d_N/d_S = (1 - C) \tag{3.48}$$

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.

We can test whether our gene is evolving in a constrained way at the protein level by estimating $^{d_N}/_{d_S}$ and testing whether this is significantly less that 1. A $^{d_N}/_{d_S}$ test can provide evolutionary evidence that a stretch of DNA proposed to be protein-coding is subject to selective 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 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.

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

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

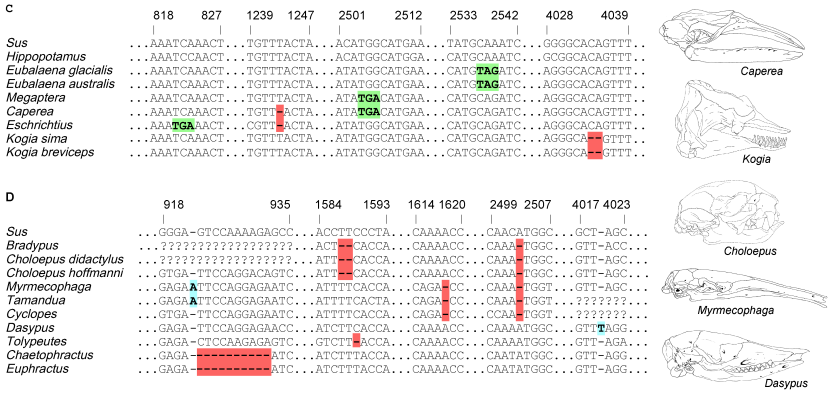

Figure 3.27: Examples of frameshift mutations (insertions blue, deletions red) and premature stop codons in Enamlin in Cetacea and Xenarthra. Figure from MEREDITH *et al.* (2009), licensed under CC BY 4.0.

The protein Enamlin is a key structural protein involved in the outer cap of enamel on teeth. Various mammals have secondarily evolved diets that do not require hard teeth, and so greatly reduced the selection pressure for hard enamel, or even teeth at all. For ex-

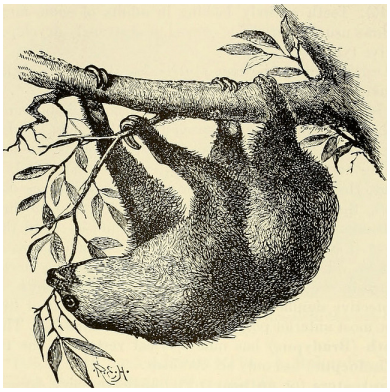

Figure 3.28: Two-toed sloth (*Choloepus hoffmanni*).
An introduction to the study of mammals, living and extinct. 1891. Flower W. H. and Lydekker R. Image from the Biodiversity Heritage Library. Contributed by University of Toronto. Not in copyright.

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).

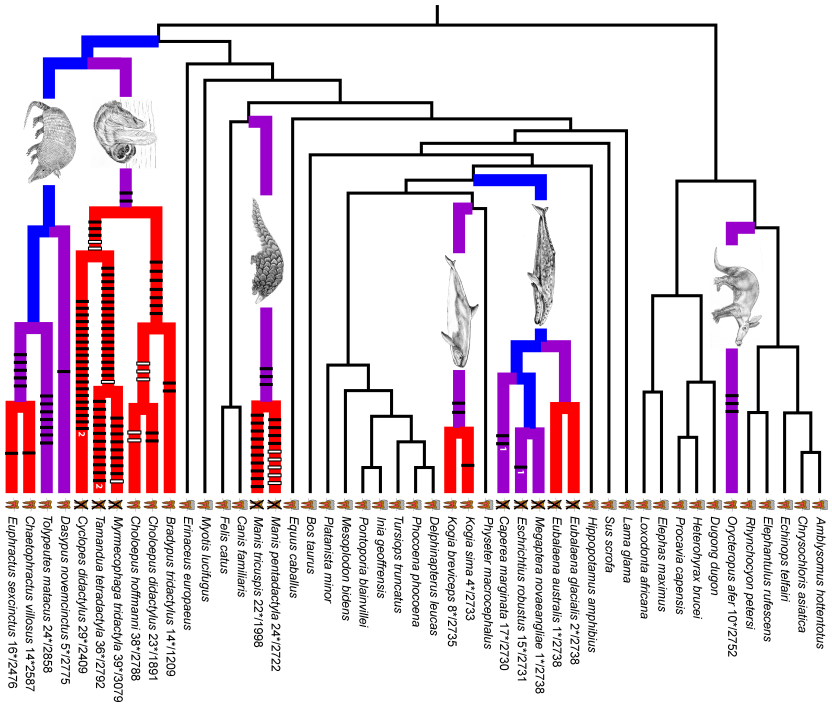

Figure 3.29: The tooth symbol next to each taxon shows whether they have teeth with enamel, lack enamel, or lack teeth. Branches of the phylogeny are coloured by whether their Enamlin is functional (black), pre-mutation (blue), mixed (purple), or pseudogenic (red). The black and white vertical bars on branches show frameshift mutations. The numbers after taxon names indicate minimum number of stop codons in the sequence divided by the length of the sequence. Figure from 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 to neutral and deleterious mutations; beneficial mutations must also arise and fix from from time to time. Let's assume that a fraction B of non-synonymous mutations that arise are beneficial such that $2N\mu B$ beneficial mutations arise per generation. Newly arisen beneficial 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 to calculate the fixation probability for beneficial alleles in Chapter 7). A newly arisen beneficial allele reaches fixation in the population with probability f_B from its initial frequency of $1/2N$. This fixation probability may be much higher than that of neutral mutations, but still much less than 1. If $2T$ generations of divergence have elapsed between the two populations then a total of

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

non-synonymous substitutions will have accumulated. Then

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

assuming again that all synonymous mutations are neutral. Note that this means that our estimates of C using $1 - d_N/d_S$ will be a lower bound on the true constraint if even a small fraction of mutations 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 potentially strong candidates for positive selection rapidly driving change at the protein level. We can identify genes that have d_N/d_S significantly greater than one, either on the complete gene phylogeny, or on particular branches. Note that is a very conservative test that 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 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 positive selection is outweighed by the loss of non-synonymous mutations to negative selection.

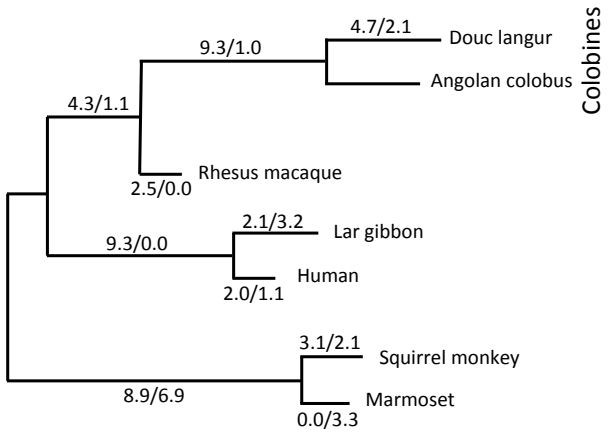

Figure 3.30: A phylogram for the primate lysozyme gene, data from YANG (1998). For each branch, the numbers give the estimated average number of non-synonymous to synonymous changes in the lysozyme protein.

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

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.

The Mcdonald-Kreitman test As noted above, a big issue with using $^{d_N}/_{d_S}$ to detect adaptation is that it is very conservative. For a more powerful test of rapid divergence, what we need to do is adjust for the level of constraint a gene experiences at non-synonymous sites. 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 lot of synonymous polymorphism, we now know that there is likely strong constraint on the gene (i.e. high C), thus we expect $^{d_N}/_{d_S}$ to be low. MCDONALD and KREITMAN (1991) devised a simple test of the neutral theory of molecular evolution at a gene based on this intuition (building on the conceptually similar HKA test HUDSON *et al.*, 1987). MCDONALD and KREITMAN took the case where we have polymorphism data at a gene for one species and divergence to

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:

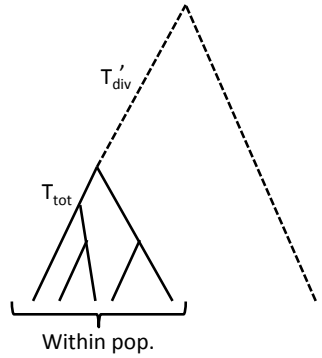

Figure 3.33: An example ogene genealogy for a set of alleles sampled within a population and a single allele sampled from a distantly-related species.

	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 2×2 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 3.34: White admiral (*Limenitis arthemis*).
American entomology : a description of the insects of North American (1859). Say, T. Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

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

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

3.6 Neutral diversity and population structure

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

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

$$F_{\mathrm{ST}} = \frac{H_T - H_S}{H_T} \tag{3.51}$$

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

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

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

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

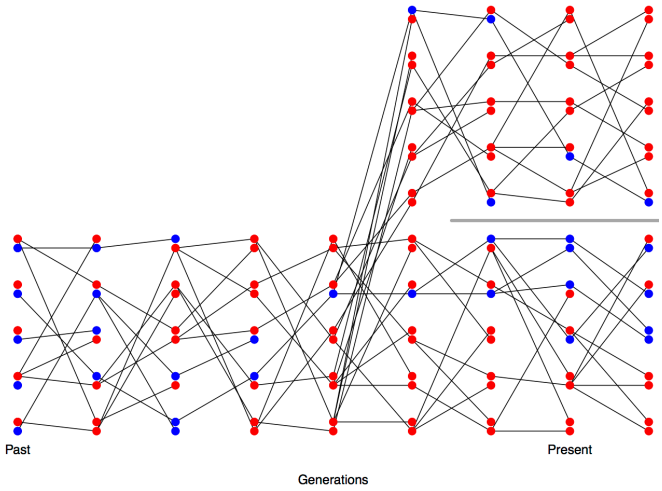

Figure 3.35: Change in allele frequencies following a population split. Code here.

heterozygosity is given by

$$H_T = \tfrac{1}{2}H_S + \tfrac{1}{2}H_B \tag{3.52}$$

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 other for at least T generations into the past as they are in distinct populations (not connected by migration). Once our alleles find themselves back in the combined ancestral population it takes them on average $2N$ generations to coalesce. So the total opportunity for mutation between our pair of alleles sampled from different populations is $2(T + 2N)$ generations of meioses, such that the probability that our pairs of alleles is different is

$$H_B \approx 2\mu(T + 2N) \tag{3.53}$$

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

$$F_{\mathrm{ST}} \approx \frac{\mu T}{\mu T + 4N_e\mu} = \frac{T}{T + 4N_e} \tag{3.54}$$

Note that μ cancels out of this equation. In this simple toy model, F_{ST} is increasing because the amount of between-population diversity increases with the divergence time of the two populations (initially linearly with T). F_{ST} grows at a rate give by $T/(4N_e)$ so that differentiation will be higher between populations separated by long divergence times or with small effective population sizes.

Question 12. The genome-wide F_{ST} between Bornean and Sumatran orang-utan species samples (*Pongo pygmaeus* and *Pongo abelii*)

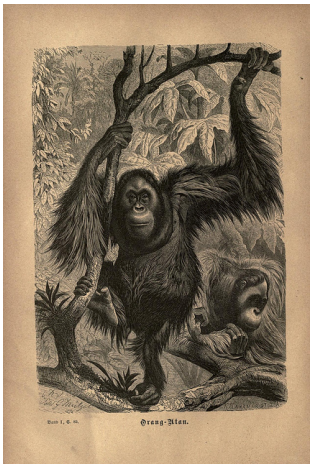

Figure 3.36: Orangutan (*Pongo*). Brehms thierleben, allgemeine kunde des thierreichs. Brehm, A. E. Image from the Biodiversity Heritage Library. Contributed by MBLWHOI Library. Not in copyright.

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

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

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

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

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

with the approximation following from assuming that $m \ll 1$ & $\frac{1}{(2N_I)} \ll 1$ (note that this is very similar to our derivation of heterozygosity above). The probability that our alleles coalesce before either 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 unlikely that the pair of alleles mutate before they coalesce on the island. Therefore, the only way that the alleles can be different from

each other is if one or other of them migrates to the mainland, which happens with probability

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

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

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

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

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

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 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 parameter here is the number individuals on the island replaced by immigrants from the mainland each generation ($N_I m$).

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

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

0.2, 0.7, 0.8

at the same markers the marine population has frequencies:

0.4, 0.5 and 0.7.

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

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 polymorphism to drift up or down in frequency, multiple population

splits may occur during the time an allele is still segregating. This can lead to incongruence between the overall population tree and the information about relationships present at individual loci. In Figure 3.37 and 3.38 we show a simulations of three populations where the bottom population splits off from the other two first, followed by the subsequent splitting of the the top and the middle populations. We start both simulations with a newly introduced red allele being polymorphic in the combined ancestral population. The most likely fate of this allele is that it is quickly lost from the population, but sometimes the allele can drift up in frequency and be polymorphic when the populations split, as the alleles in our two figures have done. If the allele is lost/fixed in the descendant populations before the next population split, our allele configuration will agree with the population 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, 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 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 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 population tree. In Figure 3.38 an allele drawn from the top and the bottom populations are necessarily more closely related to each other 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 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 terms of the underlying tree is shown on the right side of Figure 3.39 .

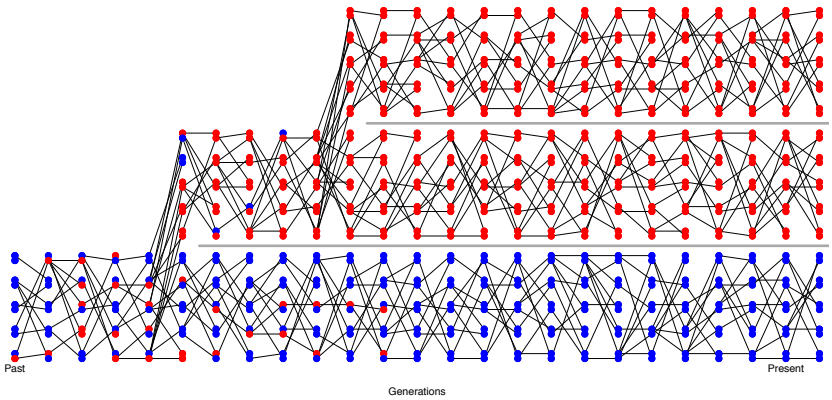

Figure 3.37: An example of alleles assorting among three populations such that there is no incomplete lineage sorting. Code here.

A natural pedigree analogy to incomplete lineage sorting is the fact

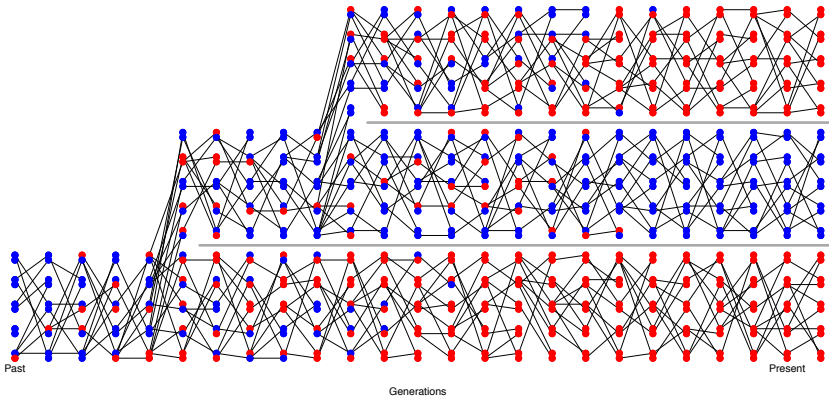

Figure 3.38: An example of alleles assorting among three populations leading to incomplete lineage sorting. Code here.

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.

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

As an empirical example of incomplete lineage sorting, let's consider the work of Jennings and Edwards who sequenced a single allele from three different species of Australian grass finches (Poephila): two sister species of long-tailed finches (*Poephila acuticauda* and *P. hecki*) and the black-throated finch (*Poephila cincta*, see Figure 3.40). They collected sequence data for 30 genes, and constructed phylogenetic gene trees at each of these loci, resulting in 28 well-resolved gene trees. 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 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 as an outgroup (the trees ((A,C),H) and ((H,C),A) respectively).

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

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. 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. 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 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 that we get a coalescent tree that is discordant with the population tree is

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

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

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. What is your best estimate of the inter-speciation time, i.e. $t_2 - t_1$?

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

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 or the other of them coalesces with the lineage from C before they

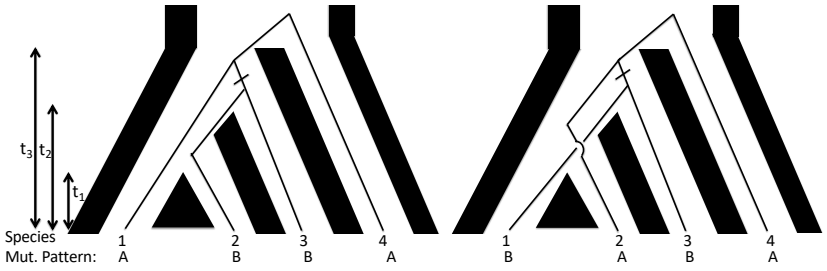

Figure 3.41: In both the left and and right trees ILS has occurred between our single lineages sampled from populations A, B, and C. Imagine that population D is a somewhat distant outgroup such that the lineages from A through C (nearly) always coalesce with each other before any coalesce with D. The small dash on the branch indicates the mutation A→B occurring, giving rise to the ABBA or BABA mutational pattern shown at the bottom.

coalesce with each other. Each option is equally likely; therefore the mutational patterns ABBA and BABA are equally likely to occur under ILS. [6]

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

$$\frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}} \tag{3.61}$$

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

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

4
Phenotypic Variation and the Resemblance Between Relatives.

The distinction between genotype and phenotype is one of the most useful ideas in Biology.[1] The genotype of an individual (the genome), for most purposes, is decided when the sperm fertilizes egg. The phenotype of an individual represents any measurable aspect of an organism.

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

There are many different ways to think about studying the path from genotype through to phenotype. The one we will take here is to think about how phenotypic variation among individuals in a population arises as a result of genetic variation in the population. One simple way to measure this genotype-phenotype relationship is to calculate the phenotypic mean for each genotype at a locus. For example, 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 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-

[1] Johannsen, W., 1911 The Genotype Conception of Heredity. The American Naturalist *45*(531): 129–159

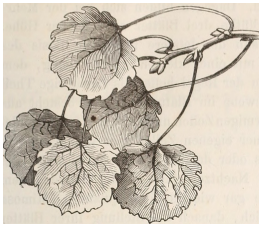

Figure 4.1: European aspen *P. tremula*.
Der baum. H. Schacht. 1860. BHL Image from the Biodiversity Heritage Library. Contributed by The Library of Congress. Not in copyright.

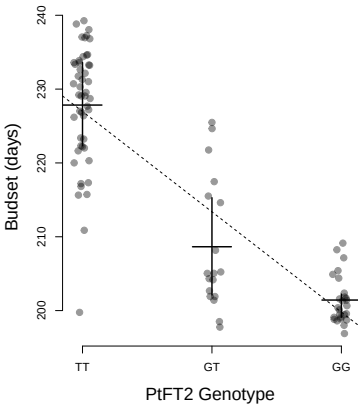

Figure 4.2: The effect of a flowering time gene (PtFT2) SNP on budset time in European aspen. Each dot gives the genotype-phenotype combination for an individual. The horizontal lines give the budset mean for each genotype and the vertical lines show the inter-quartile range. The dotted line gives the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

types at a particular SNP (l):

$$X \sim \mu + a_l G_l \tag{4.1}$$

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 μ 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

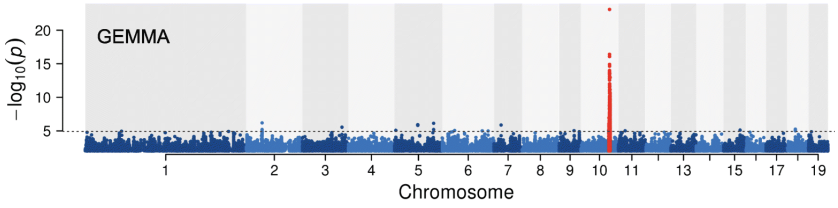

Figure 4.3: Manhattan plot of the p-value of the linear association between genotype and budset in Aspen. Each dot represents the test at a single SNP, plotted at its physical coordinate in the genome. Different chromosomes are plotted in alternating colours. The SNPs surrounding the PtFT2 gene are shown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

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 significant association with budset timing. This is because loci that are in LD with a functional locus may in turn show an association, not because they directly affect the phenotype, but simply because the genotypes at the two loci are themselves non-randomly associated. 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 functional SNP. Note how SNPs in strong LD with the functional allele (redder points) have more significant p-values.

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

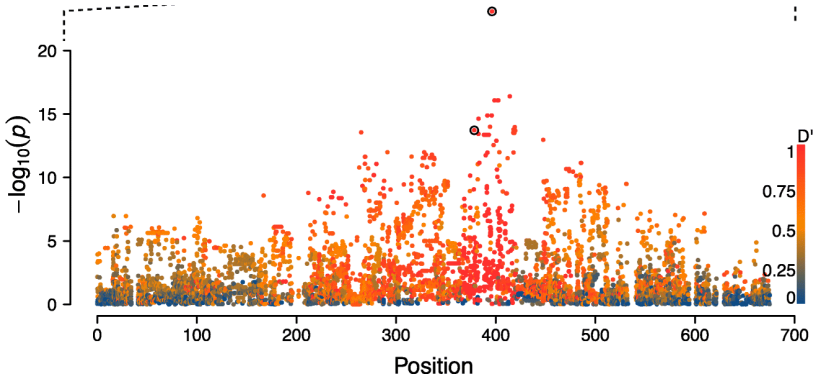

Figure 4.4: The Manhattan plot zoomed in on the top-hit (red SNPs from Figure 4.3). SNPs are now coloured by their $D\prime$ value with the most significant SNP. $D\prime$ is the LD covariance between a pair of loci (D) normalized by the largest value D can take given the allele frequencies. Figure from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

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

"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)

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

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

4.0.1 A simple additive model of a trait

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

copies of allele 1 she has at this SNP. Her expected phenotype, given her genotype at all L SNPs, is then

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

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

When we measure our individual's observed phenotype we see

$$X_i = \mu + X_{A,i} + X_{E,i} \tag{4.3}$$

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, growth, and the various random insults that life throws at our individual. If a reasonable number of loci contribute to variation in our trait then we can approximate the distribution of $X_{A,i}$ by a normal distribution due to the central limit theorem (see Figure 4.5). Thus if we can approximate the distribution of the effect of environmental variation on our trait ($X_{E,i}$) also by a normal distribution, which is reasonable as there are many small environmental effects, then the distribution of phenotypes within the population (X_i) will be normally distributed (see Figure 4.5).

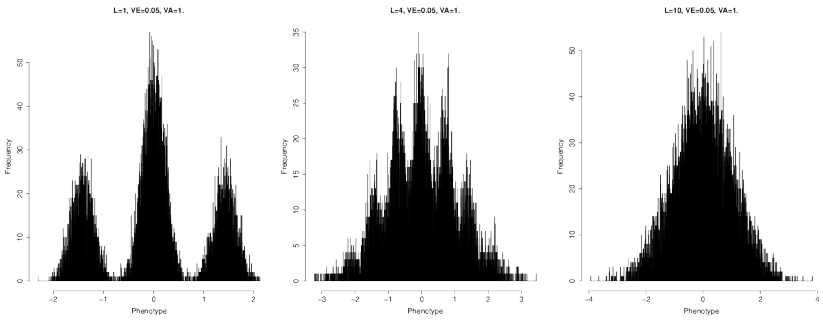

Figure 4.5: The convergence of the phenotypic distribution to a normal distribution. Each of the three histograms shows the distribution of the phenotype in a large sample, for increasingly large numbers of loci ($L = 1$, 4, and 10, with the proportion of variance explained held at $V_A = 1$). I have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

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

$$X_i = \mu + X_{iM} + X_{iP} + X_{iE} \tag{4.4}$$

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

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

Now obviously this model seems silly at first sight as alleles don't only act in an additive manner, as they interact with alleles at the same loci (dominance) and at different loci (epistasis). Later we'll relax this assumption, however, we'll find that if we are interested in evolutionary change over short time-scales it is actually only the "additive component" of genetic variation that will (usually) concern 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 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 an individual's average contribution to the next generation (i.e. the additive effect of the alleles that their genotype consists of).

4.0.2 Additive genetic variance and heritability

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

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

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

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

Our additive genetic variance can be written as

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

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

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

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

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

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

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

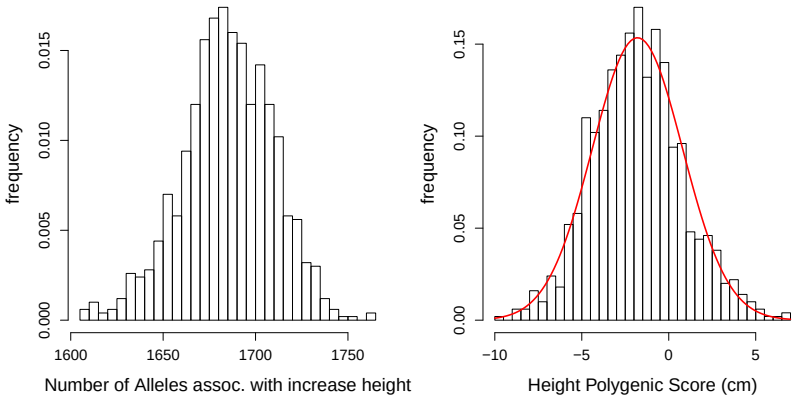

Figure 4.6: **Left)** The distribution of the number of height-increasing alleles that individuals carry at 1700 SNPs associated with height in the UK Biobank, for a sample of 1000 individuals. **right)** The distribution of the polygenic scores for these 1000 individuals. Plotted on top is a normal 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.

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

phenotypes. For example, if variation in our phenotype had no genetic basis, then no matter how much selection changes the mean phenotype 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

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

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 is referred to as the *narrow sense heritability*, the proportion of the variance explained by the additive effect of our loci. When we allow dominance and epistasis into our model, we'll also have to define the *broad sense heritability* (the total proportion of the phenotypic variance attributable to genetic variation).

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 evolutionary response to selection on a quantitative phenotype. We can calculate the narrow sense heritability by using the resemblance between relatives. For example, if the phenotypic differences between individuals in our population were solely determined by environmental 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 relatives also share an environment, so may resemble each other due to shared environmental effects.

Note that the heritability is a property of a sample from the population in a particular set of environments at a particular time. Changes in the environment may change the phenotypic variance. Changes in the environment may also change how our genetic alleles are expressed through development and so change V_A. Thus estimates of heritability are not transferable across environments or populations.

4.0.3 The covariance between relatives

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 against each other. The x and y coordinates of each point will be 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 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 and so on out (see Figure 4.7). Under our simple additive model of

phenotypes we can write the covariance as

$$Cov(X_1, X_2) = Cov\left((X_{1M} + X_{1P} + X_{1E}), ((X_{2M} + X_{2P} + X_{2E})\right) \tag{4.9}$$

We can expand this out in terms of the covariance between the various components in these sums.

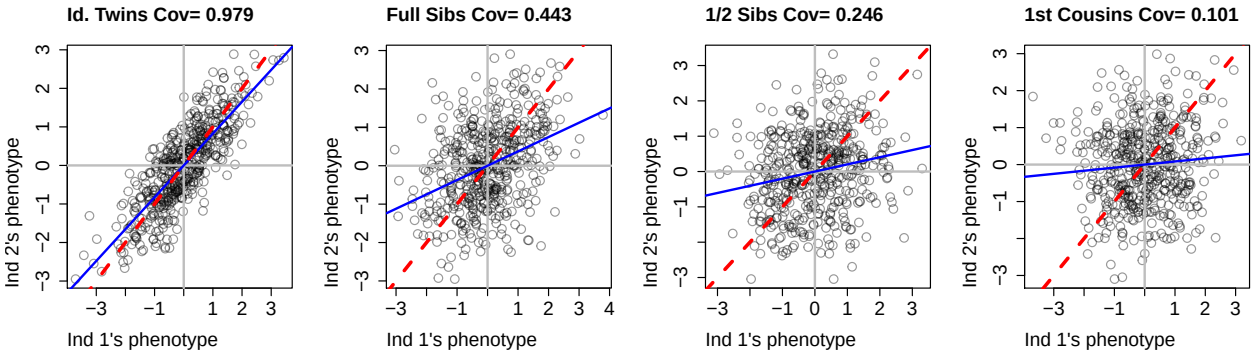

Figure 4.7: Covariance of phenotypes between pairs of individuals of a given relatedness. Each point gives the phenotypes of a different pair of individuals. The additive genetic variance is held constant at $V_A = 1$, such that the expected covariances ($2F_{1,2}V_A$) should be 1, 0.5, 0.25, and 0.125 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.

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

1. We can ignore the covariance of the environments between individuals (i.e. $Cov(X_{1E}, X_{2E}) = 0$)
2. We can ignore the covariance between the environment of one individual and the genetic variation in another individual (i.e. $Cov(X_{1E}, (X_{2M} + X_{2P})) = 0$). (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 and write our phenotypic covariance between our pair of individuals as

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

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

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

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 are not correlated draws from the population, i.e. have no probability of being IBD as we've said the parents are unrelated, the covariance 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

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

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, 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 share alleles, they share their entire genotypes, and thus resemble each other in phenotype also because of shared dominance effects.

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

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

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

So under a simple additive model of the genetic basis of a phenotype, to measure the narrow sense heritability we need to measure the covariance between pairs of relatives (assuming that we can remove the effect of shared environmental noise). From the covariance between relatives we can calculate V_A, and we can then divide this by the total phenotypic variance to get h^2.

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

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

Figure 4.8: Red-winged blackbird and Tricoloured Red-winged blackbirds (Ãgelaius 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.

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

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

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

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

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

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

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 population mean, multiplied by our narrow sense heritability.

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

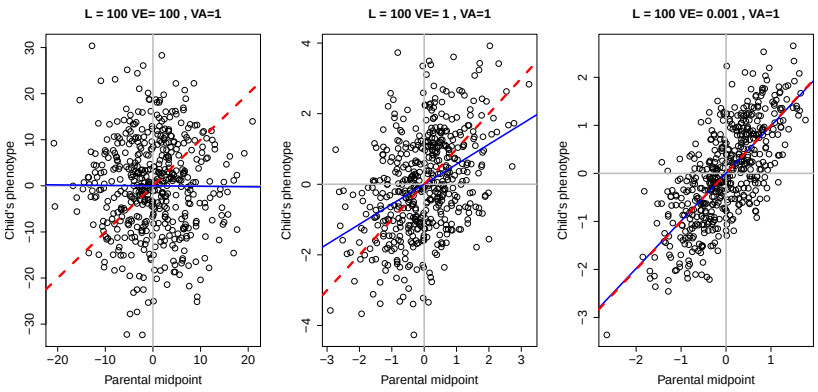

Figure 4.9: Regression of child's phenotype of the parental mid-point phenotype. The three panels show decreasing levels of environmental variance (V_E) holding the additive genetic variance constant ($V_A = 1$). In these figures, we simulate 100 loci, as described in the caption of Figure 4.5.We simulate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

Estimating additive genetic variance across a variety of different relationships. In many natural populations we may have access to individuals with a range of different relationships to each other (e.g. through monitoring of the paternity of individuals), but relatively few 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 mixed model framework. Building from equation 4.3, we can write an individual's phenotype X_i as

$$X_i = \mu + X_{A,i} + X_{E,i} \tag{4.15}$$

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 the entries for a pair 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 to account for generation effects, additional mixed effects could also be included to account for shared environments between particular individuals (e.g. a shared nest). This approach is sometimes called the "animal model".

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. 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 affecting multiple traits, or the effects of assortative mating.

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

we can write these as

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

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

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

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

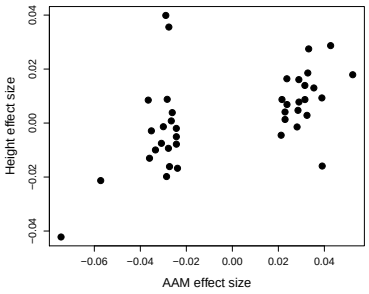

Figure 4.10: The additive effect sizes of loci associated with female Age at Menarche (AAM) and their effect size on Height in a European population. Data from Pickrell *et al.* (2016). Code here.

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

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

and

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

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

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

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

We can also talk about the genetic correlation between two phenotypes

$$r_g = \frac{V_{A,1,2}}{\sqrt{V_{A,1}V_{A,2}}} \tag{4.21}$$

where $V_{A,1}$ and $V_{A,2}$ are the additive genetic variance for trait 1 and 2 respectively. Here, r_g tells us to what extent the additive genetic variance in two traits is correlated.

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

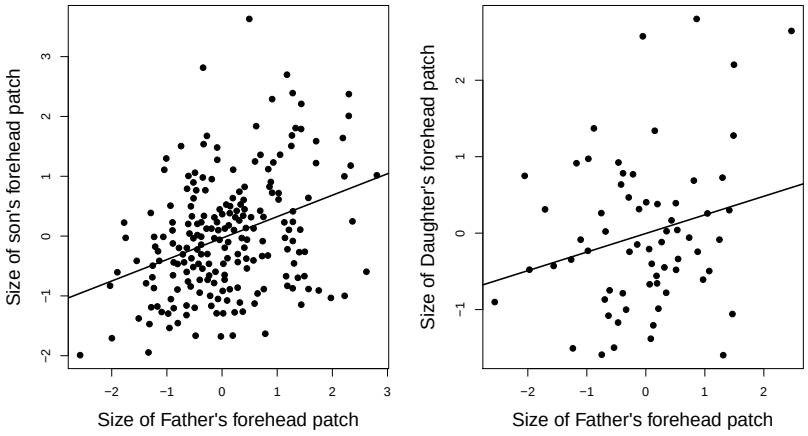

Figure 4.11: Relationship of standardized forehead patch size between fathers and sons and daughters in Pied fly-catchers. Data from Potti and Canal. Code here.

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.

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.

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.

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

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

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

$$a_{\ell,1} = p\overline{X}'_{\ell,2} + q\overline{X}'_{\ell,1}, \tag{4.22}$$

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

$$a_{\ell,2} = p\overline{X}'_{\ell,1} + q\overline{X}'_{\ell,0} \tag{4.23}$$

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

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

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

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

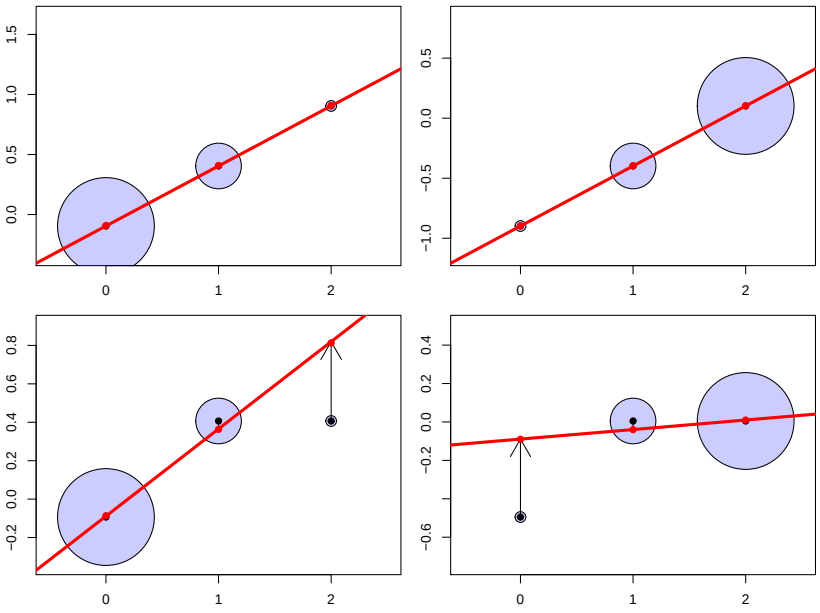

Figure 4.13: The average mean-centered (MC) phenotypes of each genotype. **Top Row:** Additive relationship between genotype and phenotype. **Bottom Row:** Allele 1 is dominant over allele 2, such that the heterozygote has the same phenotype as the 22 genotype (2). The area of each circle is proportion 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. 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 of linear regression for our population, when phenotype is regressed against the additive genotype (0, 1, 2 copies of allele 1) across all individuals 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.

The difference in the additive effect of the two alleles $a_{\ell,2} - a_{\ell,1}$ 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 α_ℓ is also the slope of the regression of phenotype against genotype (the red line in Figure 4.13). Note that the slope of our regression of phenotype on genotype (α_ℓ) does not depend on the population allele frequency for our completely additive locus (top row of 4.13). In contrast, when there is dominance, the slope between genotype and phenotype (α_ℓ) is a function of allele frequency (bottom row of 4.13). When a dominant allele (1) is rare there is a strong slope of phenotype on genotype, bottom left Figure 4.13. This strong slope is because replacing a single copy of the 2 allele with a 1 allele in an individual has a big effect on average phenotype, as it will most likely move an individual from being a 22 homozygote to being a 12 heterozygote. In contrast, when the dominant allele (1) is common in the population, replacing a 2 allele by a 1 allele in an individual on average has little phenotypic effect,

leading to a weak slope bottom right Figure 4.13. This small effect is because as we are mainly turning heterozygotes into homozygotes (11), who have the same mean phenotype as each other.

As 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 (COUSMINER *et al.*, 2013, variation at this gene in humans also influences the timing of puberty). The left side of Figure 4.15 shows the age at sexual maturity in males. The allele (E) associated with slower sexual maturity is recessive in males. While the LL homozygotes mature on average a whole year later, the additive effect of the allele is weak while the L allele is rare in the population. The right panel shows the effect of the L allele in females. Note how the allele is much more dominant in females, and has a much more pronounced additive effect. The dominance of an allele is not a fixed property of the allele but rather a statement of the relationship of genotype to phenotype, such that the dominance relationship between alleles may vary across phenotypes and contexts (e.g. sexes).

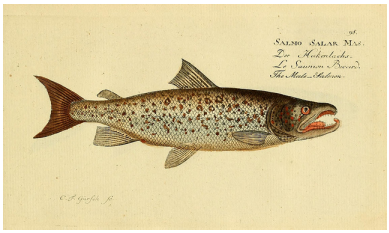

Figure 4.14: Atlantic Salmon (*Salmo salar*).
Histoire naturelle des poissons. 1796. Bloch, M. E. Image from the Biodiversity Heritage Library. Contributed by Ernst Mayr Library , Museum of Comparative Zoology. Not in copyright.

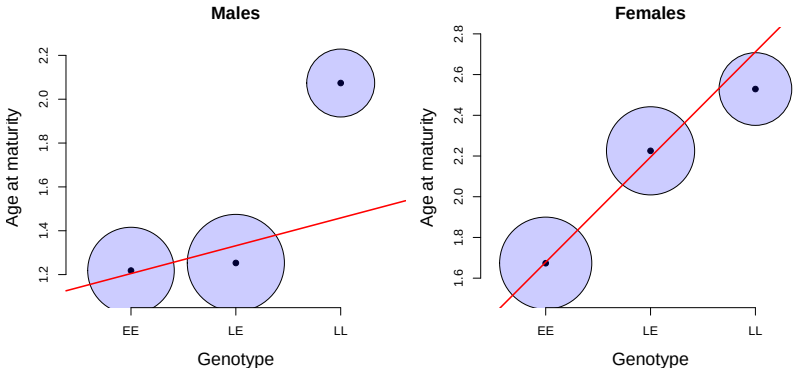

Figure 4.15: The average age at sexual maturity for each genotype, broken down by sex. The area of each circle is proportional to the fraction of the population in each genotypic class. The regression between phenotype and additive genotype is shown as a red line. Data from BARSON *et al.* (2015). Code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

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

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

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

summing the individual additive genetic variances over loci

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

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 ℓ (the length of the arrows in Figure 4.13). For example, the heterozygote deviates

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

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

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

Writing our total dominance variance as the sum across loci

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

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

$$V_G = V_A + V_D. \tag{4.29}$$

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	$1/2 V_A$
full siblings	$1/2 V_A + 1/4 V_D$
identical (monzygotic) twins	$V_A + V_D$
1^{st} cousins	$1/8 V_A$

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 covariance among a general pair of relatives (eqn. 4.12 for additive variation) becomes

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

where r_2 is the probability that the pair of individuals share 2 alleles identical by descent, making the same assumptions (other than 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 regression of offspring phenotype on parental midpoint still has a slope V_A/V_P.

Full sibs and parent-offspring have the same covariance if there is no dominance variance (as they have the same kinship coefficient $F_{1,2}$). However, when dominance is present ($V_D > 0$), full-sibs resemble each other more than parent-offspring pairs. That's because parents and offspring share precisely one allele, while full-sibs can share both alleles (i.e. the full genotype at a locus) identical by descent. We can attempt to estimate V_D by comparing different sets of relationships. For example, non-identical twins (full sibs born at same 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 covariance than non-identical twins.

The most important aspect of this discussion for thinking about evolutionary genetics is that the parent-offspring covariance is still 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 allele the offspring receives is random (assuming random mating), as it comes from the other unrelated parent (the father). Therefore, the 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 could receive from their father (weighted by their frequency in the population). Thus we only care about the additive effect of the allele, as parents transmit only alleles (not genotypes) to their offspring. This means that the short-term response to selection, as described by the breeder's equation, depends only on V_A and the additive effect of alleles. Therefore, if we can estimate the narrow-sense heritability we can predict the short-term response. However, if alleles display dominance, our value of V_A will change as alleles at our loci change in frequency, e.g. as dominant alleles become common in the population their contribution to V_A decreases. Therefore, if there is dominance our value of V_A will not be constant across generations.

Up to this point we have only considered dominance and not epistasis. However, we can include epistasis in a similar manner (for ex-

ample among pairs of loci). This gets a little tricky to think about, so we will only briefly explain it. We can first estimate the additive effect of the alleles by considering the effect of the alleles averaging over their possible genetic backgrounds (including the other interacting alleles they are possibly paired with), just as before. We can then calculate the additive genetic variance from this. We can estimate the dominance variance, by calculating the residual variance among genotypes at a locus unexplained by the additive effect of the loci. We can then estimate the epistatic variance by estimating the residual variance left unexplained among the two locus genotypes after accounting for the additive and dominant deviations calculated from each locus separately. In practice these high variance components are hard to estimate, and usually small as much of our variance is assigned to the additive effect. Again we would find that we mostly care about V_A for predicting short-term evolution, but that the contribution of loci to the additive genetic variance will depend on the epistatic relationships among loci.

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

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

5

The Response to Phenotypic Selection

Evolution by natural selection requires:

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

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

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

[1] Some people consider natural selection to only operate on heritable phenotype varation 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.

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

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

The mean phenotype in the next generation is

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

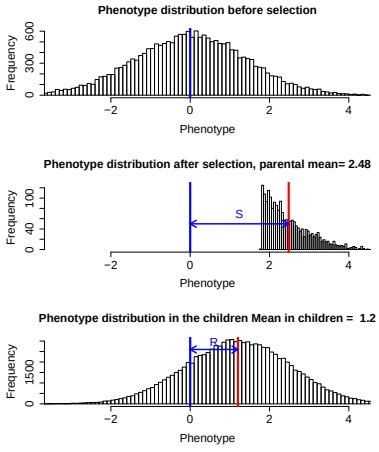

Figure 5.1: **Top.** Distribution of a phenotype in the parental population prior to selection, $V_A = V_E = 1$. **Middle.** Only individuals in the top 10% of the phenotypic distribution are selected to reproduce; the resulting shift in the phenotypic mean is S. **Bottom.** Phenotypic distribution of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

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

So to obtain μ_{NG} we need to compute $\mathbb{E}(X_{mid})$, the expected midpoint phenotype of pairs of individuals who survive to reproduce. Well this is just the expected phenotype in the individuals who survived to reproduce (μ_S), so

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

So we can write our response to selection as

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

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).

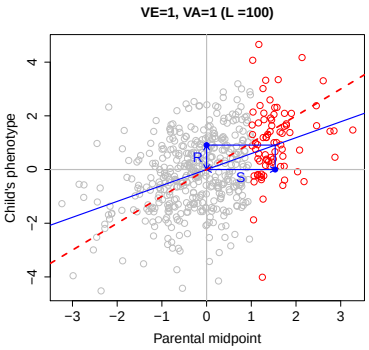

Figure 5.2: A visual representation of the Breeder's equation. Regression of child's phenotype on parental midpoint phenotype ($V_A = V_E = 1$). Under truncation selection, only individuals with phenotypes > 1 (red) are bred. Code here.

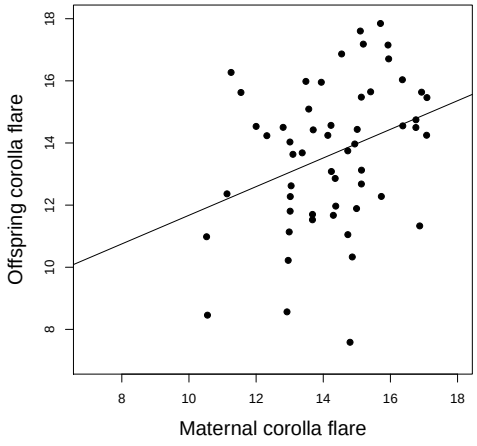

Figure 5.3: The relationship between maternal and offspring corolla flare (flower width) in P. viscosum. From Galen's data the covariance of mother and child is 1.3, while the variance of the mother is 2.8. Data from Galen (1996). Code here.

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?

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.

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

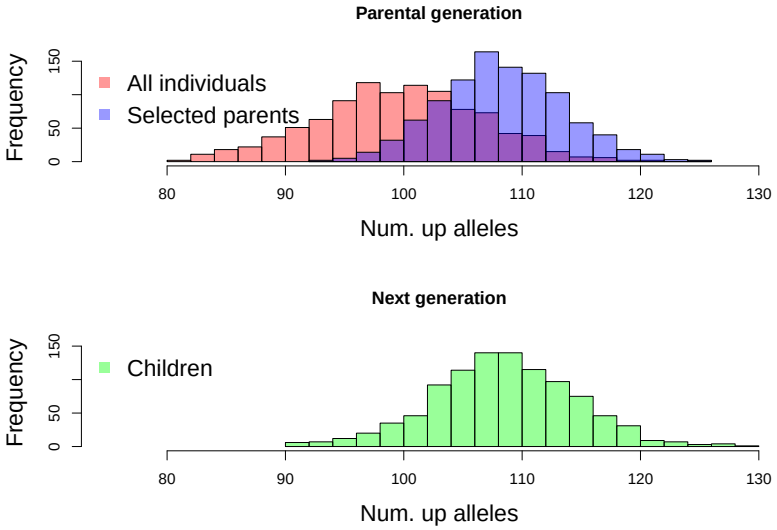

Figure 5.5: **Top.** Distribution of the number of up alleles in the parental population prior to selection (red), for the selected individuals the top 10% of the population (blue) **Bottom.** The same distribution for the offspring of the selected parents in the next generation (green). Code here.

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

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

$$nh^2S \tag{5.5}$$

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

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

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

A) Estimate the mean change per generation in the mean body weight.

B) Estimate the change in mean body weight caused by selection within a generation. State your assumptions.

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?

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

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

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

The probability that an organism with a phenotype X survives to reproduce is $w(X)$, and we'll think about this as the fitness of our organism. The probability distribution of phenotypes in those who do survive to reproduce is

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

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

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

Therefore, we can write the mean phenotype in those who survive to reproduce as

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

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

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

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

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

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.

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

$$R = V_A \beta \tag{5.12}$$

i.e. we'll see a directional response to selection if there is a linear relationship of phenotype on fitness, and if there is additive genetic variance for the phenotype.

As one example of a fitness gradient, in Figure 5.7 the lifetime reproductive success (LRS) of male Red Deer is plotted against the weight of their antlers. The red line gives the linear regression of fitness (LRS) on antler mass and the slope of this line is the fitness gradient (β).

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.

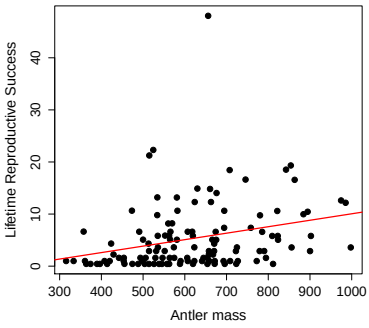

Figure 5.7: Lifetime reproductive success (LRS) of male Red Deer as a function of their antler mass. Data from Kruuk *et al.* (2002), see the paper for discussion of the complexities of equating this selection gradient with the evolutionary response. Code here..

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

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

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

where the 1 and 2 index our two different traits. Here $V_{A,1,2}$ is our additive covariance between our traits. Our selection gradient for trait 1, β_1, represents the change in fitness changing trait 1 alone holding

[3] Lande, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution *33*(1Part2): 402–416

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

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

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

where $\mathbf{V}$ and $\mathbf{G}$ are our matrices of the variance-covariance of phenotypes and additive genetic values (eqn. (4.19) (4.18)) and $\boldsymbol{\beta}$ is a vector of selection gradients (i.e. the change within a generation as a fraction of the total phenotypic variance).

Question 3. You collect observations of red deer within a generation, recording an individual's number of offspring and phenotypes for a number of traits which are known to have additive genetic variation. Using your data, you construct the plots shown in Figure 5.8 (standardizing the phenotypes). Answer the following questions by choosing one of the bold options. Briefly justify each of your answers with reference to the breeder's equation and multi-trait breeder's equation.

A) Looking just at figure 5.8 A, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

B) Looking just at figures 5.8 B and C, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

C) Looking at figures 5.8 A, B, and C, in what direction do you expect male antler size to evolve?

Insufficient information, increase, decrease.

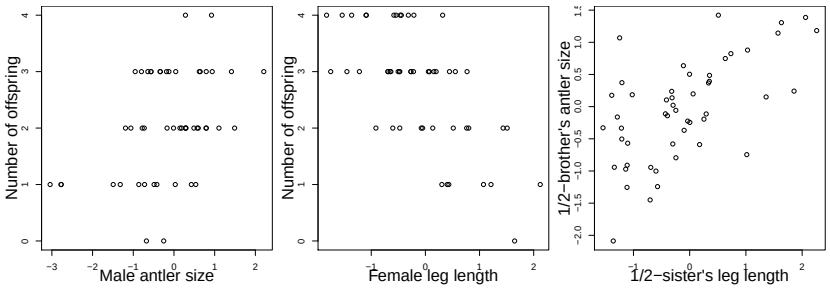

Figure 5.8: Observations of red deer within a generation; recording an individual's number of offspring and phenotypes (simulated data), which are known to have additive genetic variation. The figures left to right are A-C. (Data are simulated.)

As an example of correlated responses to selection, consider the WILKINSON (1993) selection experiment on Stalk-eyed flies (*Cyrtodiopsis dalmanni*). Stalk-eyed flies have evolved amazingly long