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 October 2018

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 (3.37)).

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.

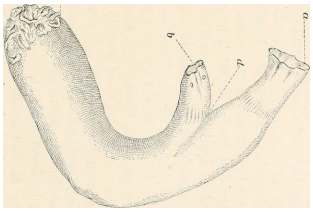

Figure 2.1: Sea Squirt (*Ciona intestinalis*). Einleitung in die vergleichende gehirnphysiologie und Vergleichende psychologie. Loeb, J. 1899.

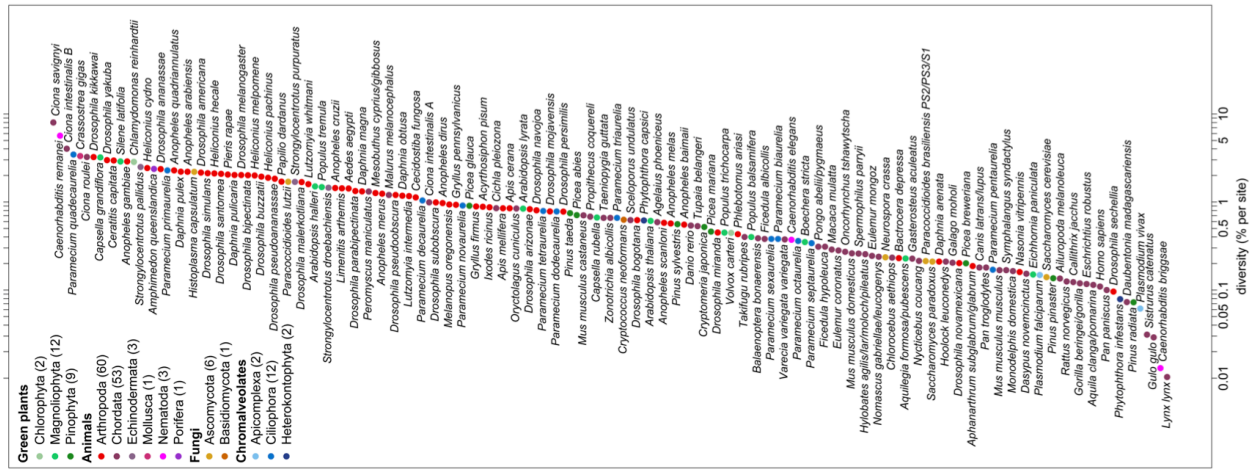

Figure 2.2: Levels of autosomal nucleotide diversity for 167 species across 14 phyla. Figure 1 from LEFFLER *et al.* (2012). Points are ranked by their π, and coloured by their phylum. Note the log-scale.

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.

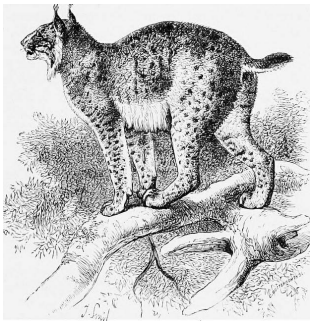

Figure 2.3: Eurasian Lynx (*Lynx lynx*). An introduction to the study of mammals living and extinct. Flower, W.H. and Lydekker, R. 1891.

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?

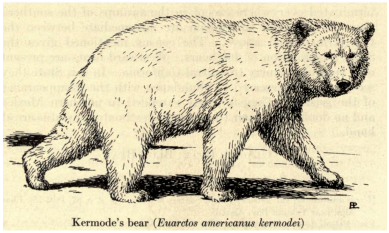

Figure 2.4: Kermode's bear. Extinct and vanishing mammals of the western hemisphere. 1942. Glover A.

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

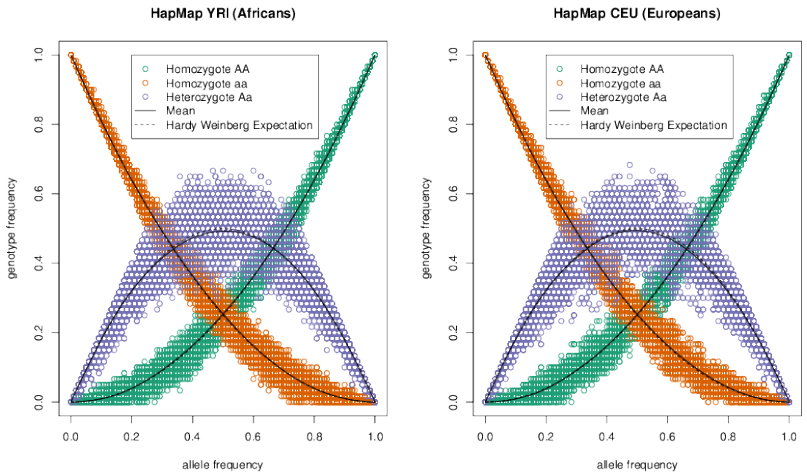

Figure 2.5: Demonstrating Hardy–Weinberg proportions using 10,000 SNPs from the HapMap European (CEU) and African (YRI) populations. Within each of these populations the allele frequency against the frequency of the 3 genotypes; each SNP is represented by 3 different coloured points. The solid lines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

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.

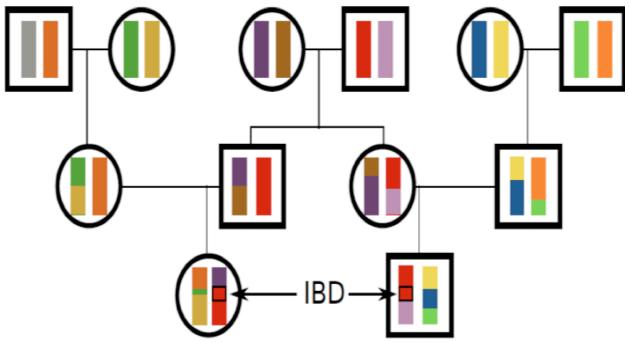

Figure 2.6: First cousins sharing a stretch of chromosome identical by descent. The different grandparental diploid chromosomes are coloured so we can track them and recombinations between them across the generations. Notice that the identity by descent between the cousins persists for a long stretch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

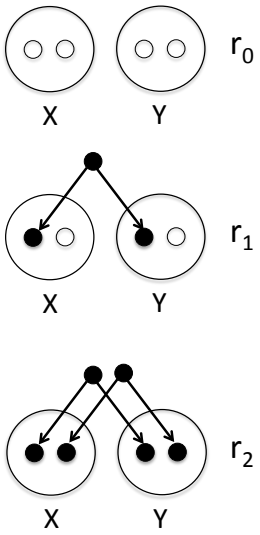

Figure 2.7: A pair of diploid individuals (X and Y) sharing 0, 1, or 2 alleles IBD where lines show the sharing of alleles by descent (e.g. from a shared ancestor).

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*). The birds of America : from drawings made in the United States and their territories. 1880. Audubon J.J.

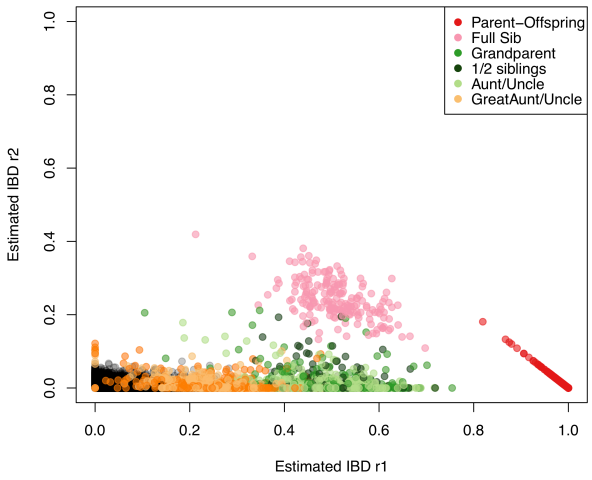

Figure 2.9: Estimated coefficient of kinship from Florida Scrub Jays. Each point is a pair of individuals, plotted by their estimated IBD (r_1 and r_2) from their genetic data. The points are coloured by their known pedigree relationships. Note that most pairs have low kinship, and no recent genealogical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

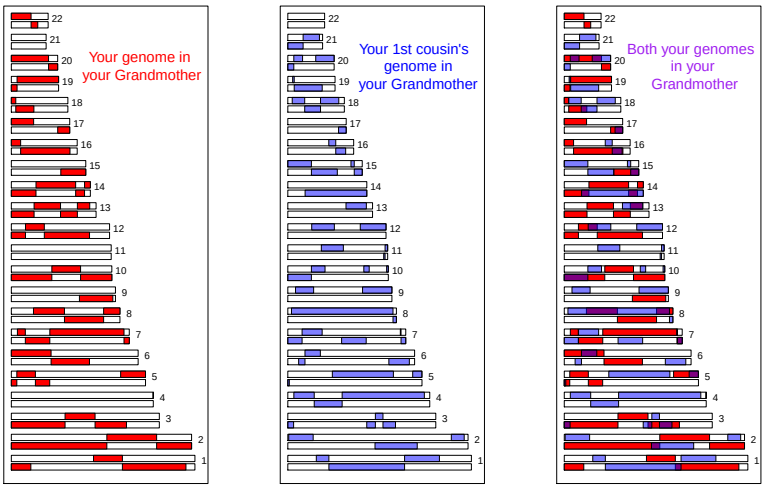

Figure 2.10: A simulation of sharing between first cousins. The regions of your grandmother's 22 autosomes that you inherited are coloured red, those that your cousins inherited are coloured blue. In the third panel we show the overlapping genomic regions in purple, these regions will be IBD in you and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

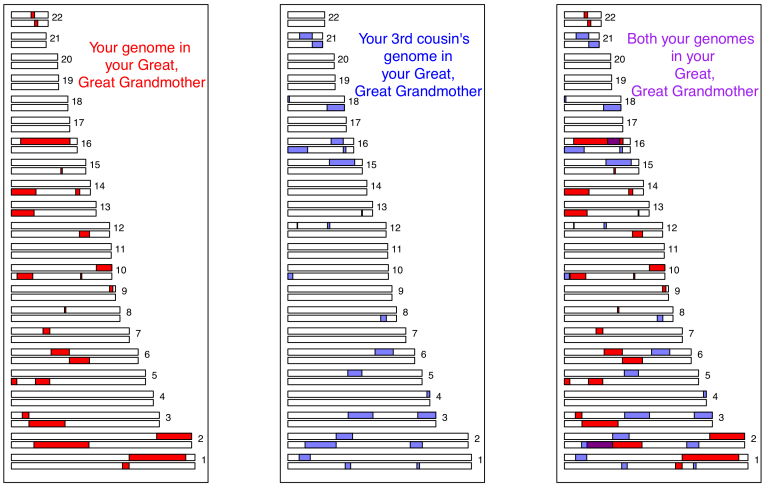

Figure 2.11: A simulation of sharing between third cousins, the details are the same as in Figure 2.10.

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.

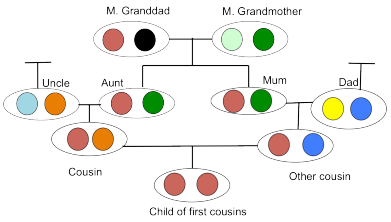

Figure 2.12: Alleles being transmitted through an inbred pedigree. The two sisters (mum and aunt) share two alleles identical by descent (IBD). The cousins share one allele IBD. The offspring of first cousins is homozygous by descent at this locus.

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).

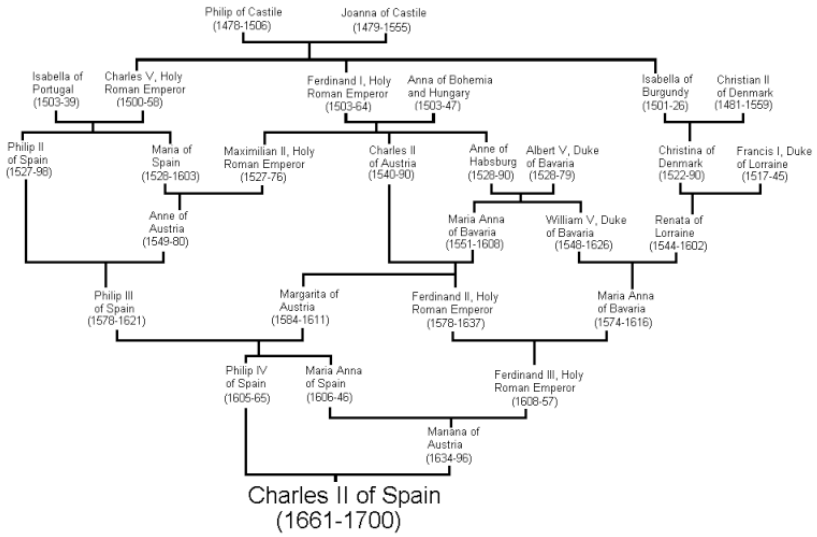

Figure 2.13: The pedigree of King Charles II of Spain

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).

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.

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.

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

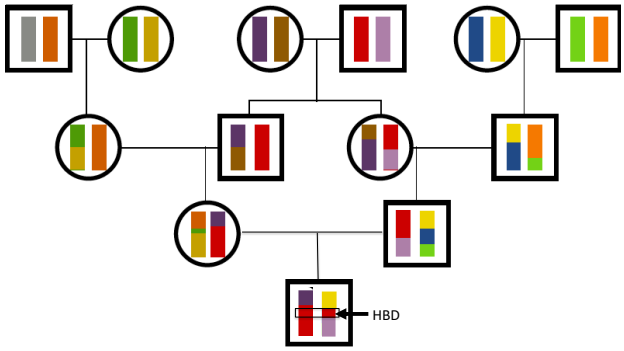

Figure 2.16: .

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 (2018), 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.

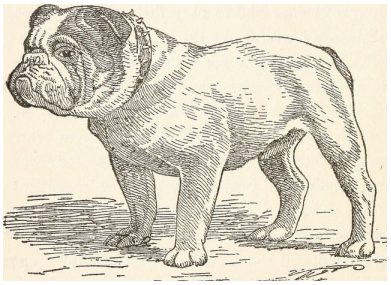

Figure 2.17: English bulldog. The dogs of Boytown. 1918. Dyer, W. A.

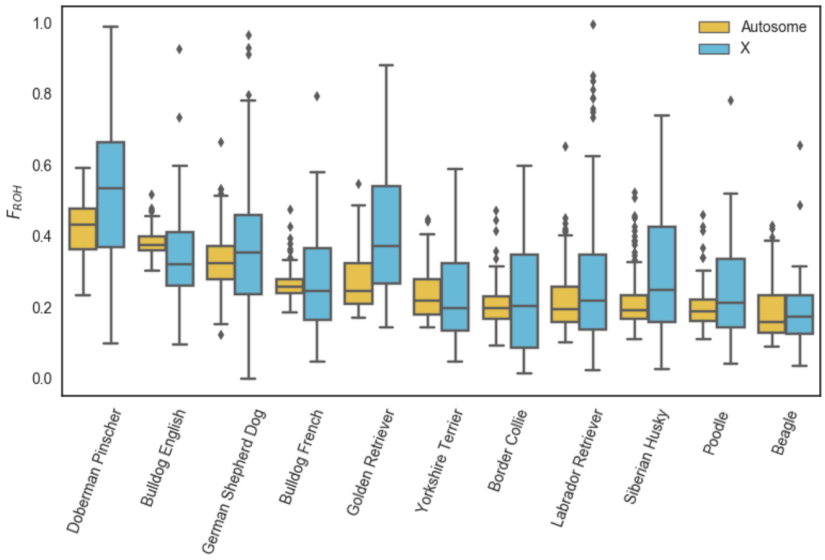

Figure 2.18: The distribution of F_{ROH} of individuals from various dog breeds from Sams and Boyko (2018).

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.

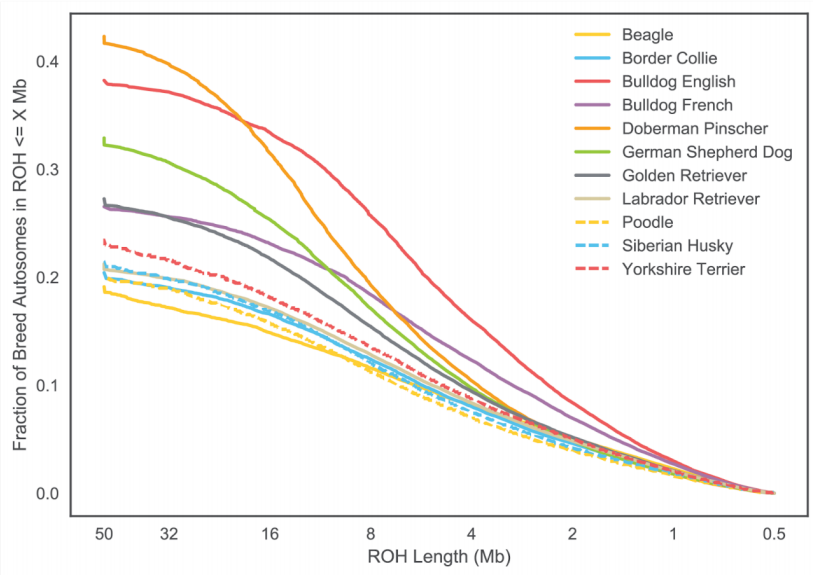

Figure 2.19: Cumulative density of length of ROH length, measured in megabases (Mb) from Sams and Boyko (2018) for various dog breeds. Note that longer lengths of ROH are on the left of the plot.

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

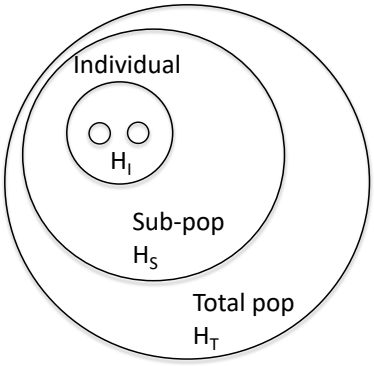

Figure 2.20: The hierarchical nature of F-statistics. The two dots within an individual represent the two alleles at a locus for an individual I. We can compare the heterozygosity on individuals (H_I), to that found by randomly drawing alleles from the sub-population (S), to that found in the total population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

Figure 2.21: The warblers of North America. Chapman, F.M. 1907.

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 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 in the *Wnt* region, a key regulatory gene involved in feather development:

Add ref to Bateson book

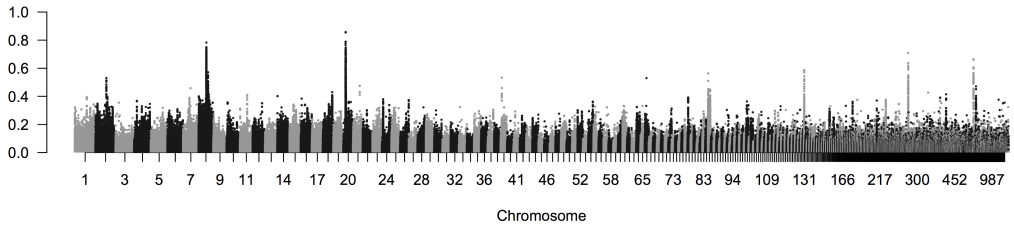

Figure 2.22: FST between blue- and golden-winged warbler population samples at SNPs across the genome. Each dot is a SNP, and SNPs are coloured alternating by scaffold. Thanks to David Toews for the figure.

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.

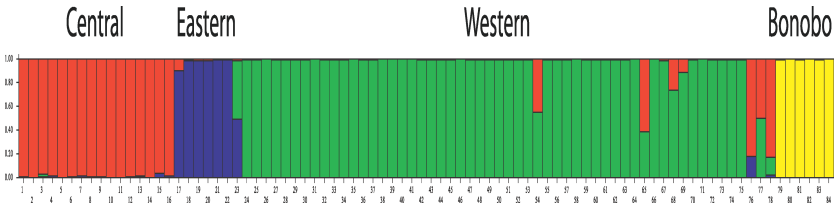

Figure 2.23: Becquet *et al.* (2007) genotyped 78 common chimpanzee and 6 bonobo at over 300 polymorphic markers (in this case microsatellites). They ran STRUCTURE to cluster the individuals using these data into $K = 4$ populations. In the above figure they show each individual as a vertical bar divided into four colours depicting the estimate of the fraction of ancestry that each individual draws from each of the four estimated populations. 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-

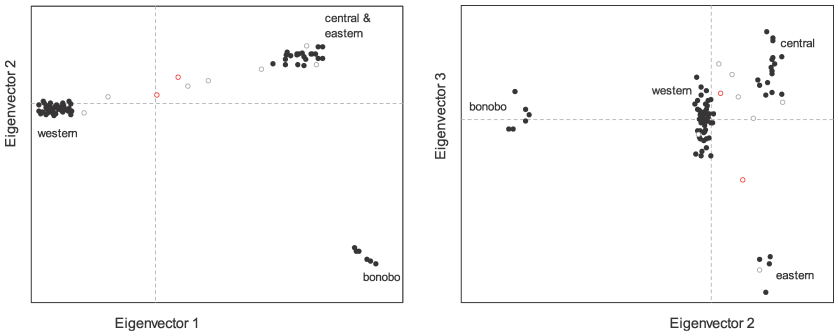

Figure 2.24: Principal Component Analysis by BECQUET *et al.* (2007) using the same chimpanzee data as in Figure 2.23. Here they plot the location of each individual on the first two principal components (called eigenvectors) in the left panel, and on the second and third principal components (eigenvectors)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

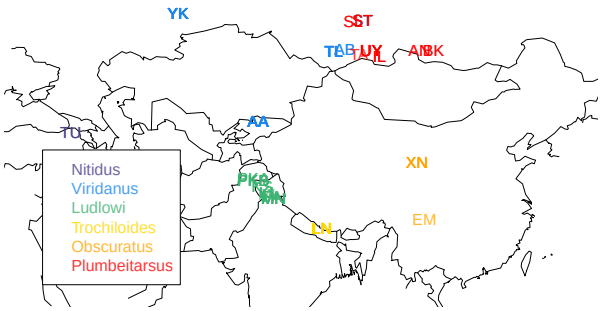

Figure 2.25: The sampling locations of 22 populations of Greenish warblers from Alcaide *et al.* (2014). The samples are coloured by the subspecies

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.. Greenish warblers are rare visitors to the UK.

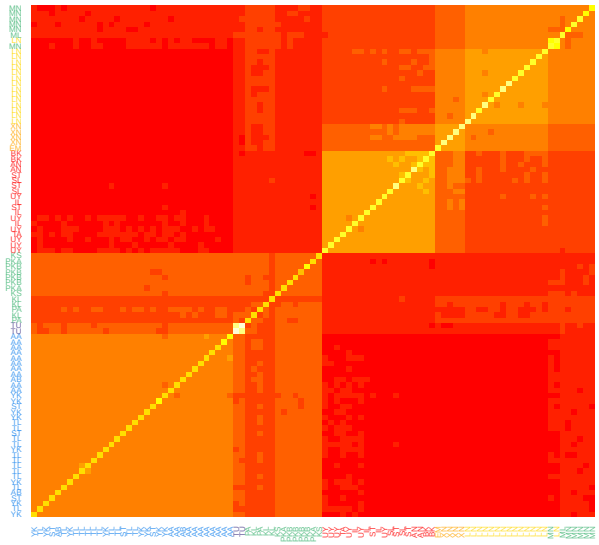

Figure 2.27: The matrix of kinship coefficients calcuated for the 95 samples of Greenish warblers. Each cell in the matrix gives the pairwise kinship coefficient calculated for a particular pair. Hotter colours indicating higher kinship. The x and y labels of individuals are the population labels from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

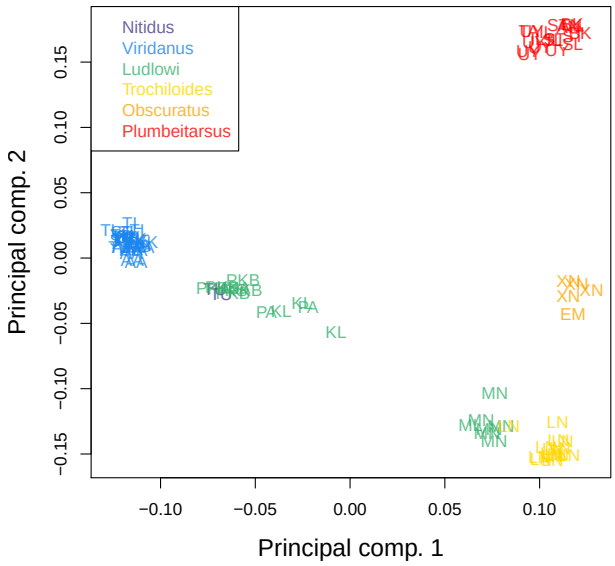

Figure 2.28: The 95 greenish warbler samples plotted on their locations on the first two principal components. The labels of individuals are the population labels from Figure 2.25, and coloured by subspecies label as in that figure.

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

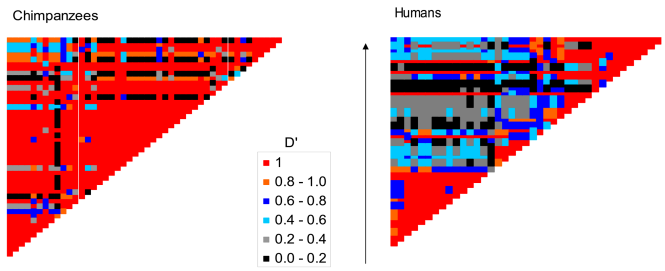

Figure 2.29: LD across the TAP2 gene region in a sample of Humans and Chimps. The rows and columns are consecutive SNPs, with each cell giving the absolute D' value between a pair of SNPs. Note that these are different sets of SNPs in the two species, as shared polymorphisms are very rare.

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 2.32 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.

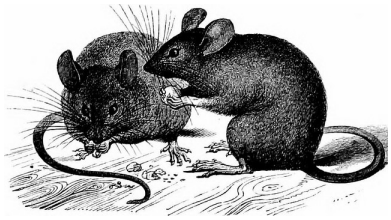

Figure 2.30: *Mus musculus*. A history of British quadrupeds, including the Cetacea. 1874. Bell T., Tomes, R. F.m Alston E. R.

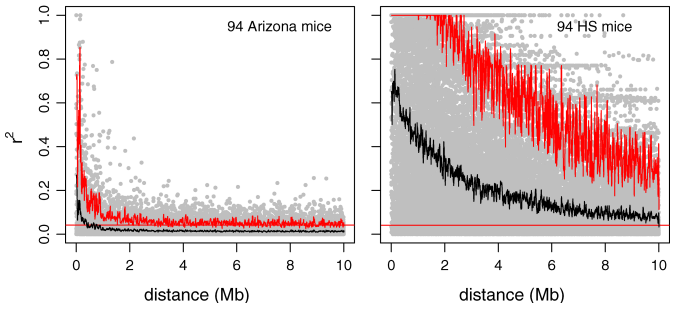

Figure 2.31: The decay of LD for autosomal SNP in *Mus musculus domesticus*, as measured by r^2, in a wild-caught mouse population from Arizona and a set of advanced-generation crosses between inbred lines of lab mice. Each dot gives the r^2 for a pair of SNPs a given physical distance apart, for a total 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.*.

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 **??**, 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

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

add units?

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

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

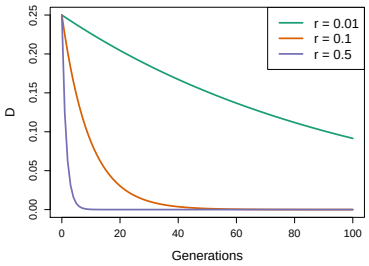

Figure 2.32: The decay of LD from an initial value of $D_0 = 0.25$ over time (Generations) for a pair of loci r apart.

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.* 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 2.33 we show the average LD between these loci as a function of the genetic distance (r) between them, from the works of SANKARARAMAN *et al.*.

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.

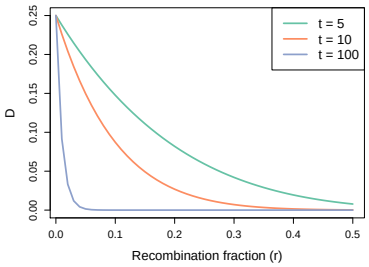

Figure 2.33: *The decay of LD from an initial value of $D_0 = 0.25$ due to recombination over t generations, plotted across possible recombination fractions (r) between our pair of loci.*

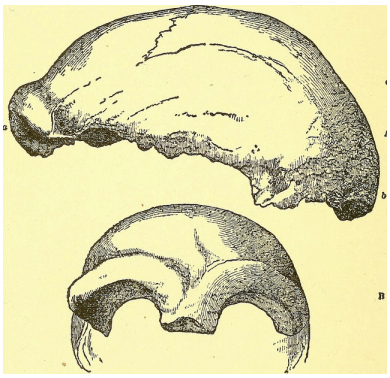

Figure 2.34: The earliest discovered fossil of a Neanderthal, fragments of a skull found in a cave in the Neander Valley in Germany. Man's place in nature. 1890. Huxley, T. H.

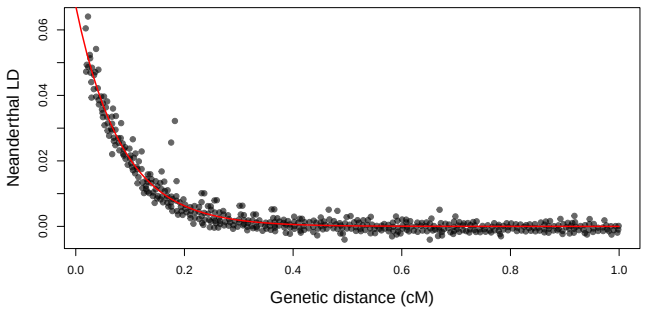

Figure 2.35: The LD between putative-Neanderthal alleles in a modern European population (the CEU sample *from the 1000 Genomes Project*). Each point represents the average D statistic between a pair of alleles at loci at a given genetic distance apart (as given on the x-axis and measured in centiMorgans (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).

The calculation done by SANKARARAMAN *et al.* 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.