2.4 Genetic drift in structured populations.

So far, our models have ignored population structure. But of course, individuals do not choose their reproductive partners at random from the entire world's population. This nonrandom mating is referred to as ***population structure****, and over time it leads to differences in allele frequencies.*

Human populations are structured at all levels: between continents and geographic regions, and often between nearby ethnic groups, towns or villages. Here we discuss basic models of structure; we'll revisit these themes in Section 3 with a specific focus on human history.

Humans share a recent African origin. Spoiler Alert! We first need the briefest overview of human genetic history to set the stage.

Humans are descended from populations in sub-Saharan Africa. Around 80,000 years ago part of this population spread out of Africa, and eventually colonized most of the world's land masses. As a result of our shared ancestry, all human populations share much of our genetic variation.

Here's a schematic overview of the relationships among human populations; see Section 3 for more about this topic:

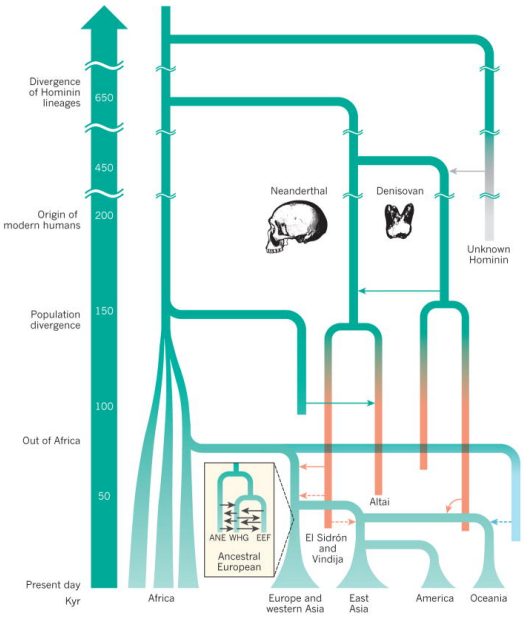

Figure 2.61: **Schematic overview of relationships among human populations.** *Most human populations descend from an ancestral population in sub-Saharan Africa. Non-African populations also briefly contacted archaic humans (Neanderthals and Denisovans) when they reached Eurasia.* *This overview is highly simplified: for example, there has been frequent migration among groups, and many populations have mixed ancestry across branches. Time estimates are approximate. Credit: Figure 2 from Rasmus Nielsen et al (2017) [Link].*

As we shall see in this chapter, the separation time of human populations is actually quite recent in terms of population genetic timescales so that most common genetic variants are shared among all human populations.

Allele frequency variation across populations. As we'll discuss in this chapter, population structure (non-random mating) allows alleles and haplotypes to drift independently in different populations. This leads to differences in allele and haplotype frequencies across populations.

To give you a sense of what this looks like, the next plot shows the allele frequencies in different human populations for a single common SNP. As is typical for intermediate-frequency SNPs, both alleles are present in all sampled populations, but at varying frequencies:

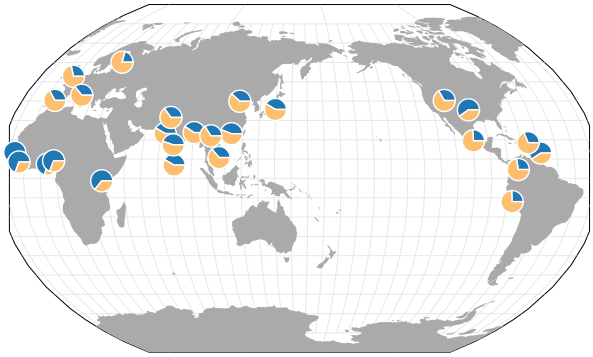

Figure 2.62: **Population allele frequencies at an arbitrary common SNP.** *Each pie chart shows allele frequencies for a 1000 Genomes population sampled at that location. The blue allele at this SNP is ancestral, and yellow derived.* *Credit: Plot made using the Geography of Genetic Variants browser: [Link]. SNP: rs7148516, Blue: A; Yellow: T. A is likely ancestral. Note that the populations plotted in the Americas are not primarily native populations.*

And here's a different visualization, showing allele frequencies for 100 randomly chosen SNPs from the 1000 Genomes Project data [208]. SNPs are sorted by global allele frequency (highest frequencies at the top) and populations from the same continent are show in adjacent columns:

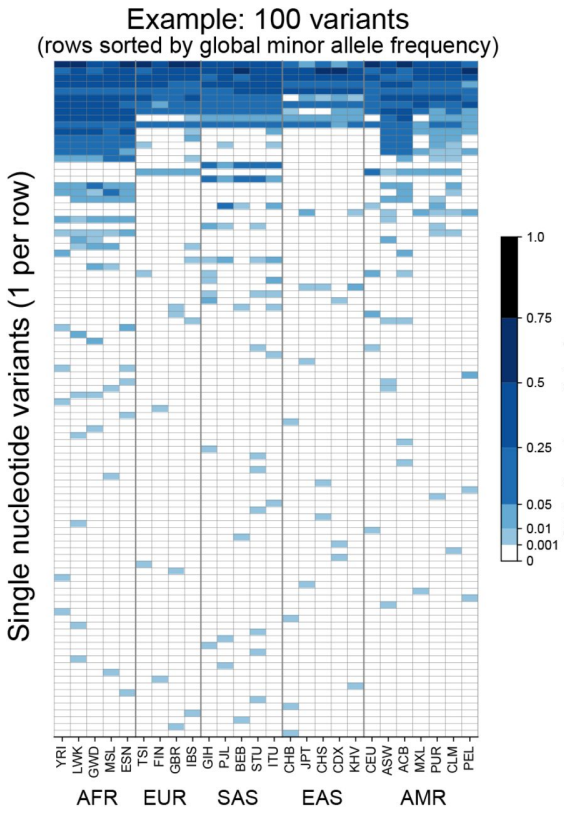

Figure 2.63: **Geographic distribution of 100 random SNPs**. *Rows are Single Nucleotide Variants, columns are populations, grouped into continental groups: Africa; Europe; South Asia; East Asia; Americas. White boxes mean that the derived allele is absent from a particular population sample;* ***the blue color scale indicates allele frequency in each population where the allele is present.*** *Credit: Figure 1 from Arjun Biddanda et al (2020) [Link]. CC-BY 4.0.*

The plots above suggest two key features of the data:

(1) Alleles that are common in one population tend to be common everywhere – notice the solid blue rows at the top of the plot that indicate SNPs that are segregating in most, or all, populations. As we will explain

shortly, this pattern arises because most common alleles arose in sub-Saharan Africa before the human diaspora and were carried everywhere as humans spread around the globe.

(2) Alleles that are rare are usually restricted to a single population or continent – lower down in the plot, the blue bars in each row are usually found in just one or a few populations. This pattern occurs because most rare alleles arose much more recently, after the separation of human populations, and are only found within the populations where they first occurred (or were carried by later migrants).

In the remainder of the chapter we'll discuss models for genetic drift with population structure to try to understand these observations.

Models of population separation and drift. To start thinking about models for allele frequency variation, consider the allele frequencies in two populations. For example, we might compare a pair of closely related populations such as Japanese and Korean; or more distantly related populations such as Japanese and Yoruba (from Nigeria). In each case, we can ask questions such as:

- How do allele frequencies differ between these pairs of populations?
- Are two samples from the same population more closely related (in a coalescent sense) than samples from different populations?

The most basic model for thinking about this is to consider a pair of populations that separated T generations ago from an ancestral population, as you can see in the sketch to the right. We'll start by assuming no migration between the two populations after the split (we'll introduce migration shortly).

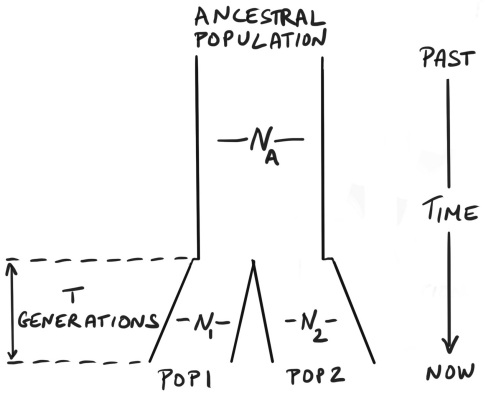

Figure 2.64: **A basic population-split model:** *An ancestral population of size N_A split at time T generations before the present, instantaneously creating two descendant populations, of sizes N_1 and N_2.*

We've discussed two different approaches to understanding drift: the Wright-Fisher forward-in-time approach, and the coalescent approach. Let's use each of these in turn to understand what happens after the population split.

First, in the forward-in-time framework, consider an allele that drifts in the ancestral population to a frequency p_A at the time of the split. Immediately following the split, it is at p_A in each of the descendant populations, *but after that it drifts independently in each population* [a].

This is illustrated in the plot below, which uses the Wright-Fisher model to simulate drift of a single allele with $N = 10,000$: first in an ancestral population (black), and then in two descendant populations (red and blue):

[a] *If Wright-Fisher drift is like a drunk man stumbling aimlessly between 0 and 1, this is now like two drunk men stumbling independently from the same starting position. The allele frequency difference between two populations is analogous to how far apart they get after T steps.*

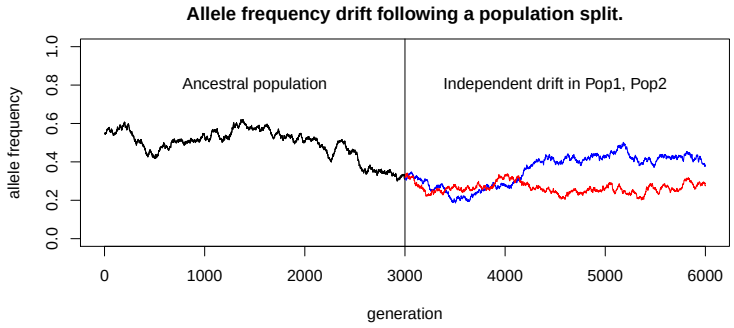

Figure 2.65: **Simulated drift of a single variant.** *Drift in the ancestral population is in black, and drift in the descendant populations is blue and red. Here time is measured forward from left to right; the amount of time plotted after the population split (3000 generations) is similar to the divergence of African and non-African populations.* $N_A = N_1 = N_2 = 10,000$.

That's just one random outcome from this process: what is the overall distribution of allele frequency differences under this model?

There is not a simple, exact mathematical formula for this, but we can get useful insight using the **Nicholson-Donnelly approximation** [209]. This provides a simple model for the present day frequency of an allele in a population (denoted p_T), given that the ancestral frequency was p_A at a time T generations before the present, assuming effective population size N. Nicholson *et al.* suggested that we can approximate this using a normal distribution [210]:

$$p_T \sim \text{Normal}(p_A, \frac{T}{2N} p_A(1-p_A)). \tag{2.53}$$

If p_T falls outside the range $[0,1]$ we think of this as equivalent to loss or fixation of the allele and set the frequency to 0 or 1.

Equation 2.53 may look complicated but is actually pretty intuitive. First, the mean of p_T is simply equal to the starting frequency p_A, since we're assuming no selection.

Meanwhile, the variance of the distribution is $T \cdot p_A(1-p_A)/2N$; this uses an approximation that the variance across T generations is simply T times the WF sampling variance $p_A(1-p_A)/2N$ per generation.

Here's what the model looks for an ancestral allele frequency of 0.55:

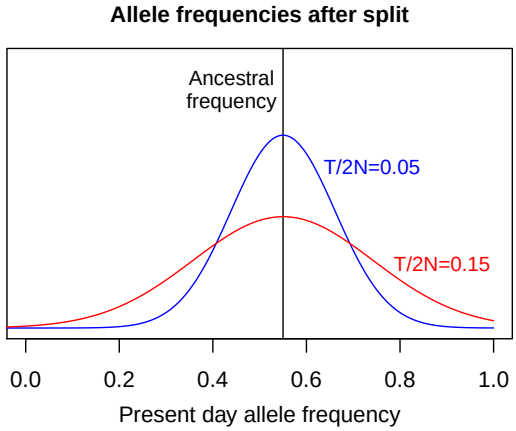

Figure 2.66: **Genetic drift after a population split.** *The plot shows the distributions of possible allele frequencies in two populations of different sizes, both starting from an ancestral frequency of* 0.55. *The population shown in red has larger* $T/2N$ *and shows more drift from the ancestral frequency. The red line approximates the amount of drift in non-African populations since the out-of-Africa migration.*

Example: Tibetans and Han. This basic model predicts the drift of each population from an ancestral population. But in practice we don't know

the ancestral allele frequencies, so we infer drift by comparing frequencies in different modern populations.

One example of this is shown in the plot below, which compares allele frequencies between 50 Tibetans and 40 Han Chinese for about 100,000 SNPs [211]. As you can see, **the allele frequencies are generally close to the diagonal, implying that frequencies are similar in the two populations.** Indeed, about half the scatter around the line actually comes from the limited sample sizes rather than from drift alone (the standard deviations of allele frequency estimates are up to 5% at these sample sizes).

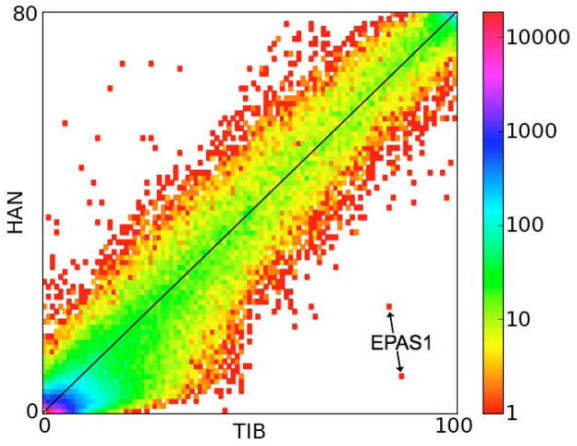

Figure 2.67: **SNP allele frequencies in Tibetans and Han Chinese.** *The plot shows Tibetan and Han allele frequencies for ~100K exonic SNPs. Colors indicate the density of points; notice the high color density along the main diagonal indicating that the vast majority of SNPs have very similar frequencies in the two populations.* *Credit: Figure 1 from Xin Ye et al 2010. [Link] Used with permission.*

While allele frequencies for most SNPs are very similar in the two populations, you'll notice that two SNPs in the **EPAS1 gene** are notable outliers in Tibetans. These outliers were the first indication of a remarkable evolutionary story.

Tibetans, of course, live at high elevations in the Himalayas, and it turns out that the EPAS1 SNPs tag a haplotype involved in local adaptation of Tibetans to altitude. EPAS1 is a transcription factor that plays a central role in regulating red blood cell production, and the haplotype that is common in Tibet increases fitness at high altitude. Natural selection has driven this haplotype to high frequency in Tibet, thus causing it to be an outlier against the genome-wide background of genetic drift [212].

A coalescent interpretation of population splits. So far we have been thinking about drift of allele frequencies forward in time, but it's also helpful to think about how population structure affects coalescence of samples. As before, we'll assume the basic split model shown in Figure 2.64, and to keep things simple, we'll assume that the effective population size is simply N at all times ($N_A = N_1 = N_2 = N$).

Consider two samples: either both from the same population, or both from different populations. When they both come from the same population, they are eligible to start coalescing immediately and, as before, the average coalescence time is simply $2N$ generations (Panels A and B, below).

But if the samples come from different populations, they cannot coalesce during the first T generations (looking backwards in time) until the lineages merge back into the ancestral population (Panel C). At that point, the usual coalescent process starts. Hence, for 2 samples from different populations, the average coalescence time is $2N + T$:

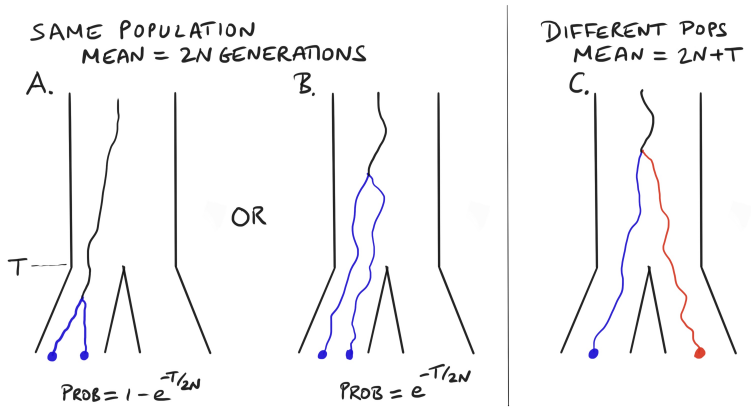

Figure 2.68: **Coalescent times for pairs of samples within and between populations.** *When both samples are from the same population, they either coalesce within their own population* **(A)**, *or back in the ancestral population* **(B)**. *If two samples are drawn from different populations they are not eligible to coalesce until both move into the ancestral population, starting T generations ago* **(C)**.

Under this model, what is the probability that two samples from the same population coalesce in the ancestral population? Using properties of the exponential distribution [213] we can show that this probability is $e^{-T/2N}$. So for example, if we take a person with recent European ancestry [214], then *at a typical locus in their genome there is a very high chance (∼85%) that their two alleles go back as independent lineages into the ancestral African population (Panel B, above).*

How does this look if we consider larger samples? Remember that in a large sample, the first coalescent events occur very quickly, while a few lineages take a long time to coalesce. This means that in a large sample, many lineages coalesce within the population, but the deeper lineages go back into the ancestral population (Panel A):

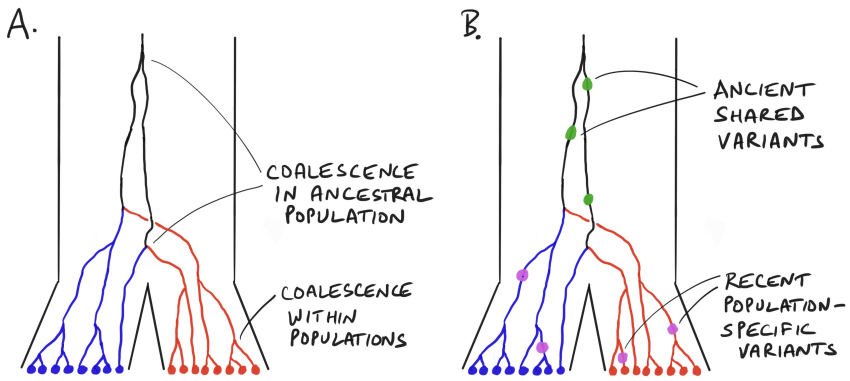

Figure 2.69: **Coalescence of larger samples within and between populations. A.** *Recent coalescences occur within populations, while deeper coalescences are in the ancestral population.* **B.** *Common variants (green) are generally older, and occur in the ancestral population; rare variants (purple) are generally younger, and usually population specific. Note: blue and red lineages are ancestral to samples in one population only; black lineages are ancestral to both.*

This has clear implications for genetic variation (Panel B, above): mutations that occur in the upper parts of the coalescent tree (i.e., older mutations) are usually common, and shared among populations. In contrast, mutations in the lower parts of the tree (younger mutations) are usually rare, and much more likely to be population-specific.

To give you some very rough numbers on this: suppose we sequence m

samples from the same population, then the expected time until the number of lineages goes down to K distinct lineages ($1 \leq K < m$) is

$$2N \sum_{k=m}^{K+1} \frac{2}{k(k-1)}. \tag{2.54}$$

If we start with $m = 1000$ samples in the present day, most of these coalesce very quickly – i.e., within populations. For example, at a time $0.15 \cdot 2N$ generations before the present (i.e., roughly the time of the out-of-Africa dispersal), only ~13 distinct lineages would survive back into the human ancestral population [215] [216]. In other words, each lineage that goes back into the ancestral population is ancestral to a bit less than 10% of the modern sample (on average), so mutations that occur since population splitting would usually be below ~10% frequency.

Meanwhile, the dozen or so deepest lineages would then take a very long time to finally reach an MRCA: almost another $4N$ generations, or ~2 million years. This is why most common genetic variation is old, predates the human diaspora, and is found in all modern populations.

Migration and other complications. I've been describing a highly simplified model in which two populations split at a fixed time T in the past. This simple model is helpful for understanding the main forces at work.

But in truth, real populations are far more complicated. Human structure is somewhat hierarchical, with many populations splitting at different times within and between continents, as in Figure 2.61. Furthermore, as we shall see in Section 3 of the book, populations don't always stay separated: populations exchange migrants or very often undergo major mixing events with other populations.

One important process is **migration** which refers to the movement of individuals (and their alleles) between populations. We can incorporate this into the Wright-Fisher model by defining a migration rate m, per generation.

Then to simulate a new generation, each new allele copy is sampled from its own population with probability $1 - m$, and from another population with probability m:

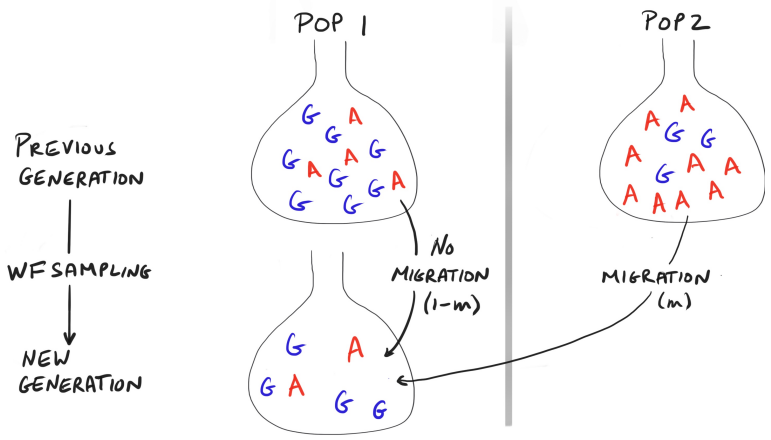

Figure 2.70: **Migration in the Wright Fisher model.** *To simulate a new generation (at bottom), alleles are drawn randomly from one of the parent populations: from the same population with probability from $1 - m$, and from a different population with probability m.*

Equivalently, in the coalescent, lineages switch between populations at rate m per generation. Coalescence can only take place between lineages that are currently in the same population [217] [218]:

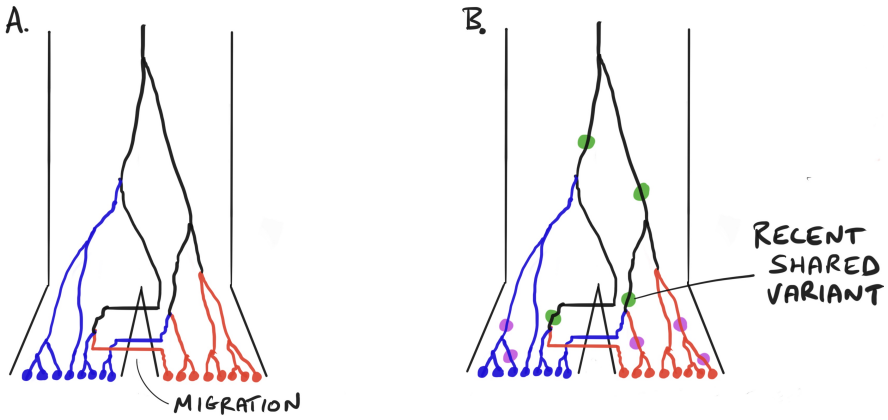

Figure 2.71: **Coalescence in a split model with migration. A.** *In the presence of recent migration, lineages can move between populations at a rate m per generation. They are then eligible to coalesce with lineages in their new population.* **B.** *In the presence of migration, it is possible for recent mutations to be shared among populations, as indicated.*

By moving alleles among populations, migration tends to reduce allele frequency differences among populations, and enables young mutations to move between populations in a way that is not possible in the pure-split model.

Together, these and other related conceptual models help us to understand the effects of a wide range of demographic processes on genetic variation. Using modern software it is now possible to simulate extremely complex models of population histories, including spatial structure, migration, population movements and splits [219].

Measuring population structure: F_{ST}. So far we have been talking about models but, for data analysis, how should we measure the extent of allele frequency differences between populations?

The most widely-used measure of differences between populations is known as F_{ST} (pronounced "F-S-T") [220]. The concept of F_{ST} was developed in the 1930s by Sewall Wright to measure the degree to which random alleles from the same subpopulation are more similar to one another than are random alleles drawn from the total population.

F_{ST} is defined to range from 0 to 1, where **F_{ST}=0 implies no population structure and a value of 1 implies perfect structure**, i.e., that subpopulations are completely fixed for different variants.

Optional: Estimation of F_{ST}

Wright's original formulation referred to F_{ST} as "the correlation between random gametes, drawn from the same subpopulation, relative to the total" [221]. This may sound precise, but there is no unique way to apply Wright's definition to data analysis, and so this idea has spawned many estimators, and many review articles. It's such a mess that I'm tempted to skip the concept entirely, but you can hardly shake a stick around in the population genetics literature without banging into F_{ST}.

One ambiguity is whether the "total" population should refer to the ancestral population, or to an average of modern populations (and if so, which populations to include). Secondly, it's unclear whether our goal should be to estimate an evolutionary parameter that depends on demographic history, or to estimate a simple arithmetic function of the allele frequencies, that can be computed even for individual SNPs. We won't go too far down the F_{ST} rabbit hole here, but I'll sketch out some main ideas [222].

First, consider a situation where multiple populations diverged from a common ancestral population. Let p_k be the present-day allele frequency of a SNP in the kth population; then we can define F_{ST} in terms of *the extent of drift relative to the ancestral population* as follows:

$$F_{ST} = \frac{\text{Var}(p_k)}{p_A(1-p_A)}. \tag{2.55}$$

This expression focuses on the variance in allele frequencies across subpopulations; the denominator $p_A(1-p_A)$ is the maximum possible variance if all subpopulations are fixed for one allele or the other[223]. **This version of F_{ST} measures the variance in allele frequency across subpopulations as a fraction of the maximum possible give the ancestral allele frequencies.**

We can interpret this further using the Nicholson-Donnelly expression $\text{Var}(p_k) \approx (T/2N)p_A(1-p_A)$. Plugging this into Equation 2.55 gives us

$$F_{ST} \approx \frac{T}{2N} \tag{2.56}$$

for small $T/2N$. In contrast, at very large divergence times (for example between species), when all the ancestral variation is either fixed or lost within populations, F_{ST} converges to 1 [224]. Equation 2.55 is not immediately useful for data analysis as it depends on the ancestral frequency p_A, which we cannot observe directly; but it's not hard to estimate this with Bayesian methods [225].

An alternative formulation (and closer to Wright's original framing) is to write F_{ST} in terms of the probability of identity of pairs of alleles between and within subpopulations [226]:

$$F'_{ST} = \frac{H_b - H_w}{H_b} \tag{2.57}$$

where H_w and H_b are the probabilities that two random samples from within a subpopulation, or between subpopulations, are different. The notation H is used here because this is analogous to *heterozygosity*. Although it's not evident at a first glance, this version of F_{ST} is actually a rearrangement of Equation 2.55, but using the total frequency p_t in modern populations instead of the ancestral frequency p_A [227].

Importantly, the expected number of differences between random samples is proportional to their coalescent times, so this expression can be related to average coalescent times within and between populations [228]. **This interpretation of F_{ST} measures the fractional reduction in coalescent times for a pair of samples from the same population compared to a random pair from the total population** [229].

Computing F_{ST} from Equation 2.57 has the advantage that it doesn't depend on the unknown ancestral allele frequency, but it arguably makes estimation *more* difficult because in real applications there is sampling error in both the numerator and the denominator which makes estimation a bit painful. For a helpful summary of moment estimators of F_{ST}, with recommendations, see Bhatia et al (2013).

Turning to data, Bhatia et al (2013) [230] estimated values of F_{ST} between human continental groups. The F_{ST} values are roughly centered around the value of 0.15 that we used above to illustrate our models:

Populations	F_{ST}
Yoruba and Han	0.161
Yoruba and European	0.139
Han and European	0.106

Table 2.6: F_{ST} between human populations. *The data include samples from three populations: Yoruba (from Nigeria), CEU (a sample of individuals from Utah of northwest European descent), and Han Chinese.* Modified from Bhatia et al (2013). [Link]

As you can see, F_{ST} in humans is fairly small (up to about 0.15), even between the most distantly related populations. This reflects the fact that most common variation is shared among all populations.

A second interesting point is that F_{ST} is a bit higher between the African population Yoruba and Han Chinese (0.161), than between Yoruba and Europeans (0.139), even though Europeans and Chinese are descended from the same out-of-Africa migration event. This is because east Asians underwent a stronger bottleneck than Europeans after the out-of-Africa event, resulting in a smaller effective population size and higher F_{ST}.

Third, F_{ST} between populations from the same continent is usually much lower, reflecting more-recent split times and subsequent migration. For example, in the Tibet-Han data set discussed above, the authors estimated that F_{ST} between the two populations is just 0.026.

It would be tempting to interpret F_{ST} values to estimate population split times, using the models described above. But in practice, F_{ST} values depend on a complex mixture of population split times, bottlenecks and migrations. F_{ST} provides a useful summary of the combined impact of all these processes but it's very difficult to untangle the contributions of all these distinct forces in real data [b].

[b] *Ancient DNA has been a game-changer for reconstructing complex population histories, far beyond what is possible using only modern genomes (Chapter 3.3).*

Example: Coalescence between species. To close this chapter, I'll show an example where we can use the coalescent to understand evolutionary splits in a very different context: *between species*.

Before DNA sequencing, the main way that we knew about the evolutionary histories of species was from fossil evidence. But interpretation of the fossil record is often based on just a few fragmentary specimens. It may be unclear how the fossils relate to one another, and to modern populations or species. Fossil evidence continues to be important, but genetic data gives us a powerful complementary type of information for studying our evolutionary history. The accumulation of sequence differences over time, due to mutation, is often called a **molecular clock**.

Figure 2.72: **Our closest relatives:** *female chimpanzee with infant.* Credit: Alain Houle CC BY 4.0 [Link]

Here we'll use genome sequence data to understand the evolutionary relationships among the **great apes**: humans and our closest living relatives: chimpanzees, gorillas, and orangutans.

Until the late 1990s, it was still debated whether humans are more closely related to chimpanzees or to gorillas (orangutans are more distantly related to all three) [231]. DNA sequence data now show that in fact we are most closely related to chimpanzees, with the human and chimpanzee genomes differing at 1.37% of aligned nucleotides compared to 1.75% for

human versus gorilla [232].

This tree shows the evolutionary relationships among the great apes, including that humans and chimpanzees are most-closely related:

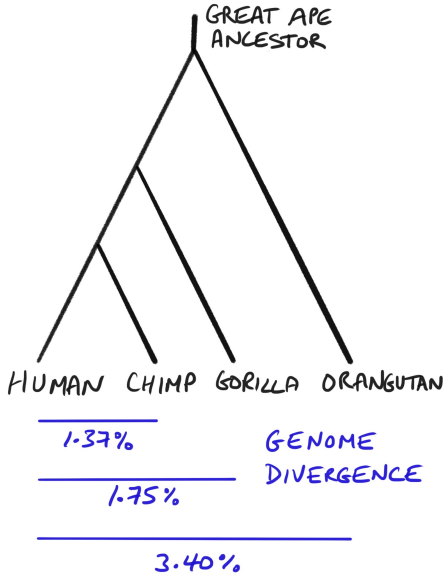

Figure 2.73: **Species tree for the great apes.** *This figure simplifies additional complexity within the nonhuman clades, as there are two recognized chimpanzee species, two gorilla species, and three orangutan species, and additional subspecies of each.* *After Figure 1a from Aylwyn Scally et al (2012) [Link].*

The picture above shows the relationships among the ancestral populations that gave rise to humans and the other great apes (this depiction is known as a **species tree**). But if you look at individual regions of the genome, a very interesting pattern emerges. The branching order for the human, chimpanzee and gorilla sequences vary from region to region across the genome [233]:

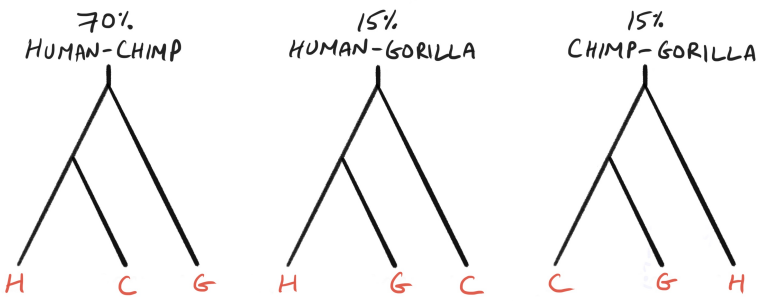

Figure 2.74: **Different parts of the genome support different trees.** *About 70% of the genome supports human and chimpanzee as closest to each other, while the rest supports grouping either human with gorilla or chimpanzee with gorilla.*

About 30% of the genome shows gorilla closer to either human, or chimpanzee. How should we interpret this?

The key to understanding this is to think about the relationships among the different genomes as a coalescent process. First, think about the ancestral lineages for a segment of the human and chimpanzee genomes.

As in the split model we described above, human and chimpanzee cannot coalesce immediately because they come from different species. But unlike our human examples, it is around 6 million years until the human and chimpanzee lineages flow back into an ancestral population. At that

point, the coalescent process you're already familiar with starts: the human and chimpanzee lineages have the opportunity to coalesce, and the average waiting time is an additional $2N_{HC}$ generations, where N_{HC} is the effective population size in the human-chimp ancestral population.

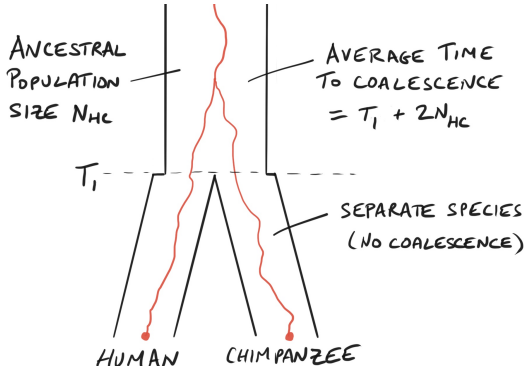

Figure 2.75: **Coalescence of human and chimpanzee lineages**. *Moving backward in time from the present, the lineages are ineligible to coalesce until they flow into the human-chimp ancestral population about 6 million years ago.*

If the human and chimp lineages coalesce quickly, then this always results in the "correct" tree. But if the lineages don't coalesce quickly, they flow back into the human-chimp-gorilla ancestral population. If this happens, all three possible branching patterns are equally likely:

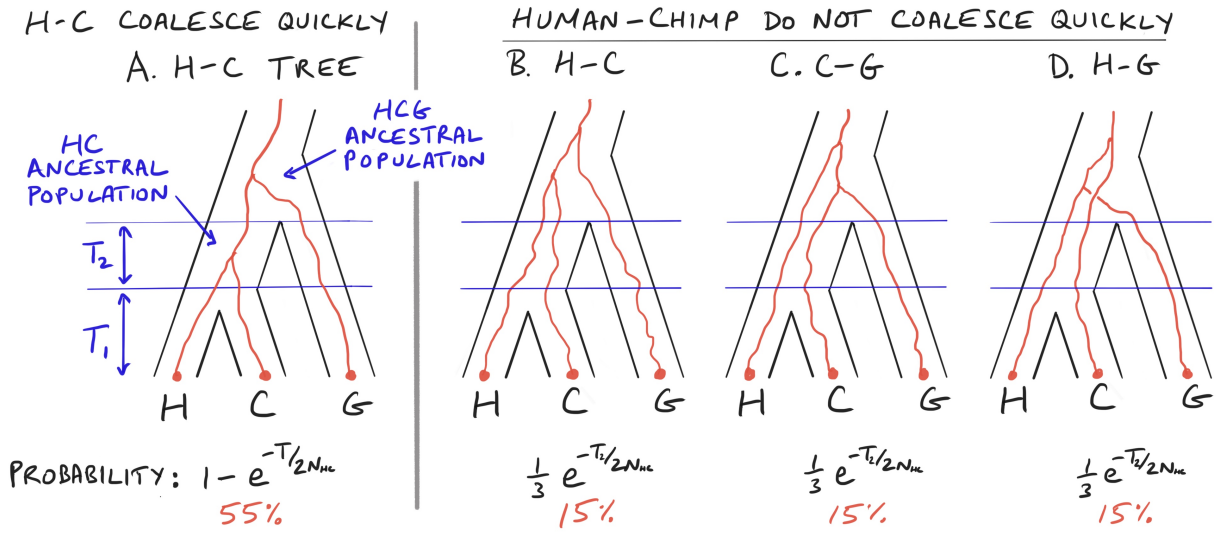

Figure 2.76: Possible coalescent trees relating human, chimpanzee. and gorilla. A. *Human and chimpanzee coalesce in the human-chimp ancestral population, and this ensures that the tree topology matches the overall "correct" relationship among the populations.* **B-D.** *Human and chimpanzee do not coalesce until they flow back into the human-chimp-gorilla ancestral population. When that happens, all three possible trees (with human-chimp, chimp-gorilla, or human-gorilla joining first) are equally likely. The theoretical and actual probabilities for each outcome are shown at the bottom.*

So to summarize, for about 55% of the genome, human and chimp coalesce in the H-C ancestral population. This ensures the "correct" genealogy – meaning that the genealogy matches the species relationships. However, for 45% of the genome the genome, the human and chimp lineages fail to coalesce within the H-C ancestral population and, instead,

flow separately into the H-C-G ancestral population. When that happens, all three possible trees are equally likely, and occur with about 15% probability each [c].

I told you before that 70% of the genome shows a human-chimp pairing: this is the sum of Tree A (55%) and Tree B (15%), while Trees C and D contribute about 15% of the genome each.

[c] *This situation where the local genealogies often differ from the species tree is known as* ***incomplete lineage sorting****.*

Optional math: Probabilities for the four H-C-G tree topologies.

We start by computing the probability of Tree A: i.e., that human and chimp coalesce within the H-C ancestral population. For this we will assume the simplest possible model: constant population size and no population structure. N_{HC} is the effective population size in the human-chimp ancestral population, and we assume that this population existed for T_{HC} generations.

To compute the probability of Tree A we need to compute the probability of a coalescent event for two samples within T_{HC} generations. Using properties of the exponential distribution we can write the probability of a coalescent event at time t as

$$\frac{1}{2N_{HC}} \exp(\frac{-t}{2N_{HC}}) \tag{2.58}$$

where $\exp(x)$ indicates e^x. Then the probability of Tree A equals the probability of $t < T_2$, which we compute by integration:

$$\int_0^{T_2} \frac{1}{2N_{HC}} \exp(\frac{-t}{2N_{HC}}) = 1 - \exp(\frac{-T_2}{2N_{HC}}). \tag{2.59}$$

The remaining probability, $\exp(\frac{-T_{HC}}{2N_{HC}})$, gives us the probability that the human and chimp lineages go back into the H-C-G ancestral population. At that point, there are three lineages (H, C, and G), and any pair of these are equally likely to make the first merger. So the probability of each of these three trees (B, C, D) is simply

$$\frac{1}{3} \exp(\frac{-T_{HC}}{2N_{HC}}). \tag{2.60}$$

It's beyond our scope here, but there has been some fascinating work on the structure of the ancestral great ape populations. While there's still uncertainty in the models, one main result is that the ancestral population sizes were huge: $\sim$120,000 for the human-chimpanzee ancestral population, which is $>$ 6-fold the current human effective size. Consequently, coalescence within that ancestral population was very slow. The human-chimpanzee population split is estimated at $5.5 - 7$ million years ago, and the split from gorilla at $8.5 - 12$ million years ago [234].

Well done! In these last few chapters we have covered the main forces of neutral population genetics! In the remainder of this section of the book we turn our attention to selection. As we shall see, selected alleles are still subject to all the processes we've covered already, but also subject to the guiding hand of natural selection.

Notes and References.

[208] Biddanda A, Rice DP, Novembre J. A variant-centric perspective on geographic patterns of human allele frequency variation. Elife. 2020;9:e60107

[209] Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(4):695-715

[210] Motivation for the Nicholson-Donnelly Approximation. The variance due to drift in a single generation of the WF model is $p(1-p)/2N$ (using standard properties of binomial sampling). For a sum of independent random variables, the variance of the sum equals the sum of the variances. This rule doesn't really apply here, because the drift is a function of p_t, which depends on the drift in the previous generations. However, if we make the approximation that the drift variance in each generation is constant, and determined by the ancestral frequency, p_A, then the variance over T generations is simply T times the variance in the first generation. This approximation works best for small values of $T/2N$ (for which the allele frequencies don't drift very far from p_A.

[211] Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-8

[212] There's also a second fascinating aspect to this story: the selected EPAS1 haplotype is highly divergent from other human haplotypes at this locus, and is believed to have entered the human population by gene flow from a species of archaic hominid known as the Denisovans, which were related to Neanderthals:

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-7), in a process known as *adaptive introgression*. We'll come back to this when we cover human history.

[213] Recall that coalescent times are exponentially distributed with parameter $1/2N$. The cumulative distribution of the exponential at time T is therefore given by $1 - e^{-T/2N}$; see e.g., [Link].

[214] Here I'm assuming that $T/2N$ since the out-of-Africa migration is around 0.15 time units.

[215] This is calculated using the formula above to compute the expected time to go from $m = 1000$ lineages down to $K = 13$ lineages. You can compute this formula in R using

f <- function(n) { 2/(n*(n-1))}

sum(f(14:1000)).

For simplicity I'm ignoring recent population growth and the out-of-Africa bottleneck. Both events would change the distribution of times but not the overall intuition.

[216] My treatment of this problem is a bit simplistic, for ease of exposition. However there is an extensive literature on the number of lineages at time t, for example:

Jewett EM, Rosenberg NA. Theory and applications of a deterministic approximation to the coalescent model. Theoretical population biology. 2014;93:14-29

Slatkin M. Allele age and a test for selection on rare alleles. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2000;355(1403):1663-8

and references therein.

[217] When there is migration, we can keep track of the number of lineages in each population at any given time (let's call this k_1 and k_2, respectively). Then, going backward in time, migration events from population 1 to population 2 are exponentially distributed at rate mk_1, and mk_2 for the reverse direction. A migration event from 1 to 2 decreases k_1 by one, and increases k_2 by one. Meanwhile, coalescent events occur within populations: e.g., within population 1 at rate $k_1(k_1 - 1)/2$, as usual. We can simulate the next event (coalescence in population 1 or 2, or migration from 1 or from 2) as a process of competing exponentials. Lastly, we can generalize this model to include more populations with an arbitrary matrix of migration rates between populations i and j in each generation.

[218] I'm illustrating the split-plus-migration model here because this is relevant to many human populations. But there's a simpler, classic, model in population genetics called *island migration* in which the populations never merge together, and are subject to migration going back infinitely far in time. In this model, provided that the migration rate is >0 it's guaranteed that eventually the ancestral lineages will happen to collect in one population so that they can merge together. You could motivate the island model by considering populations (for example birds on islands, or butterflies on disconnected systems of serpentine grasslands) that have occupied the same geographic space for a very long time – since long before the joint MRCA of all the populations.

[219] Such as SLiM [Link].

[220] F_{ST} was one of three measures of genetic structure known as Wright's F-statistics. Wright's other F statistics, F_{IS} and

F_{IT}, measure inbreeding of individuals relative to the sub- and total populations, and are less widely used nowadays.

[221] Wright S. The genetical structure of populations. Annals of eugenics. 1949;15(1):323-54

[222] There are various reviews of F_{ST}. I suggest Nicholson et al (2002, cited above) and Bhatia et al (2013), which I relied on for this section

Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Research. 2013;23(9):1514-21;

as well as:

Barton N. Identity and coalescence in structured populations: a commentary on 'Inbreeding coefficients and coalescence times' by Montgomery Slatkin. Genetics Research. 2007;89(5-6):475-7

Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics. 2009;10(9):639-50

[223] To be more precise, this is the variance if there are many subpopulations, each fixed for allele 0 or 1 with probability $1 - p_A$ and p_A respectively or, equivalently, the expected squared difference for each population between its actual allele frequency and the expected value p_A.

[224] We can see that F_{ST} converges to 1 as follows. Eventually every subpopulation either loses the allele (with probability $1 - p_A$) or fixes (with probability p_A). So eventually $\mathrm{Var}(p_k)$ is given by $(1 - p_A)p_A^2 + p_A(1 - p_A)^2 = p_A(1 - p_A)(p_A + 1 - p_A) = p_A(1 - p_A)$. This cancels with the denominator implying that F_{ST} ultimately converges to 1.

[225] Nicholson et al 2002

[226] One advantage of this framing is that it doesn't assume a particular evolutionary model (i.e., population splitting), and is equally applicable for any scenario with structure, such as migration-only models.

[227] To keep this simple we'll consider the frequency in a particular subpopulation p_s as a random variable, and the ancestral or total frequency p_A and p_T, respectively, as fixed parameters. The numerator of Equation 2.55 is $\mathrm{E}[(p_s - p_A)^2]$ by the definition of a variance. Then, noting that $\mathrm{E}[p_s] = p_A$ we have:

$$F_{ST} = \frac{\mathrm{E}[(p_s - p_A)^2]}{p_A(1 - p_A)} = \frac{\mathrm{E}[p_s^2] - 2\mathrm{E}[p_s p_A] + \mathrm{E}[p_A^2]}{p_A(1 - p_A)} = \frac{E[p_s^2] - \mathrm{E}[p_A^2]}{p_A(1 - p_A)}.$$

For Equation 2.57 we note that $H_b = 2p_t(1 - p_t)$ and $H_s = 2p_s(1 - p_s)$, similar to the logic for Hardy-Weinberg. Then

$$F'_{ST} = \frac{2p_t(1 - p_t) - 2\mathrm{E}[2p_s(1 - p_s)]}{2p_t(1 - p_t)} = \frac{\mathrm{E}[p_s^2] - \mathrm{E}[p_t^2] - \mathrm{E}[p_s - p_t])]}{p_t(1 - p_t)} = \frac{\mathrm{E}[p_s^2] - \mathrm{E}[p_t^2]}{p_t(1 - p_t)}$$

[228] See Equations 6 and 8 in Slatkin, M. (1991):

Slatkin M. Inbreeding coefficients and coalescence times. Genetics Research. 1991;58(2):167-75

[229] From Slatkin (1991):

$$F_{ST} = \frac{\bar{t} - \overline{t_w}}{\bar{t}}$$

where $\bar{t}$ is the mean coalescent time for two random samples from the total population and $\overline{t_w}$ is the mean coalescent time for two random samples from the same subpopulation.

[230] Bhatia et al (2013)

[231] A classic paper by Maryellen Ruvolo (1997) discussed incomplete lineage sorting in the human-chimpanzee-gorilla divergence, reporting that 11 out of 14 genomic data sets support the (human, chimpanzee) grouping (see her Table 1):

Ruvolo M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Molecular biology and evolution. 1997;14(3):248-65

[232] This section draws heavily on work by

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169-75

See also

Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS genetics. 2007;3(2):e7

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research. 2011;21(3):349-56

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471-5

[233]The trees at individual genomic regions are known as **gene trees** (although this is a misnomer, since the trees don't correspond to genes *per se*).

[234]There's still quite a bit of uncertainty in these models. One issue is potential changes in mutation rate over time:
Amster G, Sella G. Life history effects on the molecular clock of autosomes and sex chromosomes. Proceedings of the National Academy of Sciences. 2016;113(6):1588-93