2.2 More on genetic drift: The coalescent.

Here we introduce a different way of understanding the Wright-Fisher model, called the coalescent, but now looking backward in time. The coalescent may seem confusing at first but is incredibly powerful for understanding genetic variation and for data analysis [a].

[a] *The 19th Century Danish philosopher Søren Kierkegaard quipped that "Life can only be understood backwards, but it must be lived forwards." This quote encapsulates the difference between coalescent models (backward-in-time) and the Wright-Fisher model (forward-in-time)* [154].

A short history. In the early 20th Century, when people first started studying population genetics, it was natural to think about evolutionary models forward in time, and these ideas were developed into the Wright-Fisher model during the 1920s. For 50 years forward-in-time models were the main tools for understanding evolutionary processes.

But it turns out that forward models are not easily adapted for use in data analysis. When the first molecular data started to arrive at the end of the 1960s, this drove the development of new questions and models in population genetics [155]. One huge innovation was the coalescent, developed independently by three scientists in 1982 and 1983: John Kingman, Richard Hudson, and Fumio Tajima [156].

Like many breakthroughs in science, the coalescent stands the conventional thinking on its head. Instead of thinking about evolution forward in time to reach the present day, we look backward at the ancestors of modern samples. Many problems in population genetics, especially for neutral models, suddenly become far easier [157].

Inheritance of genetic material from a shared ancestor. The central concept of the coalescent is that the DNA sequences carried by present-day individuals – you or me, for example – are copies of DNA sequences carried by individuals in the distant past. Your genome and my genome are descended from many shared ancestors that lived hundreds of thousands of years ago.

To train your intuition, we start by thinking about inheritance of DNA within families. Imagine comparing your own genome with that of a second cousin (second cousins share great-grandparents). In some parts of the genome you, and that second cousin, inherited the exact same chunk of chromosome from one of your great-grandparents (marked in red, below). On average you share 1/32nd (about 3%) of your genome with that second cousin:

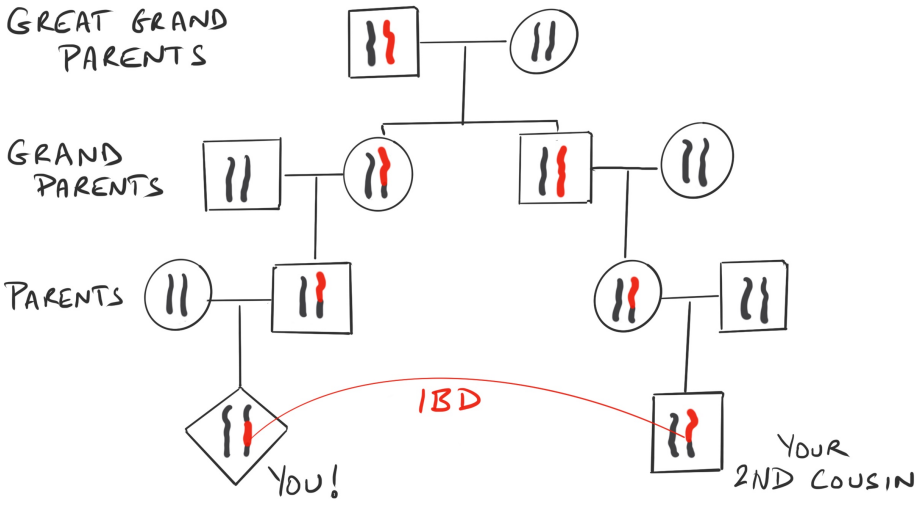

Figure 2.14: **Shared ancestry between second cousins.** *Inheritance of one chromosome (i.e., one homolog) from a great grandparent shown in red. For the two cousins the overlapping segment is said to have* ***coalesced*** *in their great-grandfather. (With such recent ancestry, the overlapping part of the red segments in the two cousins is also said to be* ***identical by descent (IBD)****.)*

We say that this part of your genome **coalesced** with the corresponding part of your cousin's genome 3 generations ago; the great-grandparent is your **common ancestor** at this locus. Coalescence means that this part of both your, and your cousin's genomes, are descended as copies of this ancestral genome [b].

Here we're focusing on coalescence within a family pedigree, between two people who are "related" in the usual sense of the word. But as I shall explain next, in fact *everyone* in a population is related in the same way, although coalescence is usually far more ancient [c].

[b] *With such recent shared ancestry we expect the two copies of this region to be identical, aside from any new mutations. Regions shared within ~10 generations are referred to as* ***identical by descent (IBD)****.*

[c] *There is an important distinction between* ***pedigree ancestors*** *(e.g., you have 8 great-grandparents) and* ***genetic ancestors, which are the focus here****. As in the picture above, you have two copies of any small region of your genome,* ***each of which comes from just a single parent, grandparent, great-grandparent and so on.***

The coalescent refers to ancient shared ancestry within populations.

Let's pick an arbitrary location in the human genome. You have two homologous copies of this locus. Pick one of those two copies at random. You inherited this copy from one of your parents – your mum, say – who got it from one of her parents, and so on backwards in time.

Now do the same thing for one of your friends. Pick one copy of this locus in your friend. Do these two copies have a common ancestor? Perhaps surprisingly, the answer is yes, although that common ancestor probably lived hundreds of thousands of years ago.

To see this, we're going to use the Wright-Fisher model again. Remember that going forward in time, the WF model generates each generation by random sampling with replacement from the generation before. We can think of this in terms of drawing colored balls out of bags. Each time we pull out a ball we write down its color and toss it back into the bag. Here, two red balls in the present generation are both copies of the same "ancestor" red ball two generations ago:

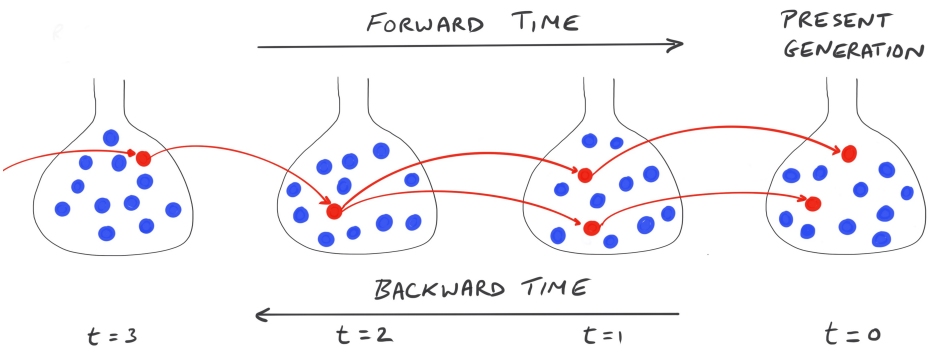

Figure 2.15: **Coalescence in the WF model.** *Two copies of this locus in the present generation are marked by red balls. These descend from a common ancestor (i.e., they coalesce) two generations ago.* ***In coalescent models it is most natural to measure time backward from the present.***

Measuring time forward or backward. The illustration above also shows that we can measure time either forward or backward. For the WF model it's natural to count generations forward from some arbitrary starting point, as we did in the previous chapter. But in the coalescent we will define the present day as $t = 0$ and count generations backward in time.

The genealogy of a sample. So far we have been talking about the ancestry for a pair of copies of this locus. Can we extend this to think about the ancestry of m copies of this locus? For example, we could sequence this locus in $m/2$ diploid individuals – how should we think about the ancestry of these m sequences?

You can think of the ancestry of the samples as coming from a **coalescent genealogy** (or just **genealogy**, or **tree**) that represents the relationships of all m sequences. This genealogy is embedded within the forward WF process:

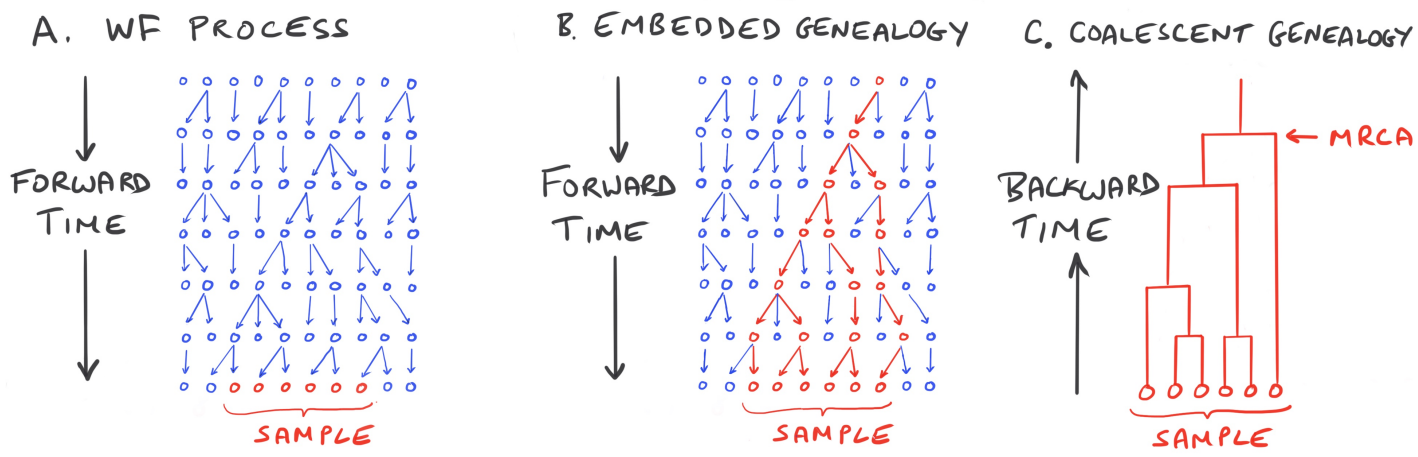

Figure 2.16: The WF history contains an embedded coalescent genealogy. A. *WF genealogy for a small population. This includes six chromosomes sampled at the present day, in red.* **B.** *Red circles and arrows indicate the ancestors of the sampled chromosomes, embedded within the WF process.* **C.** *The coalescent genealogy abstracts away all irrelevant details of the WF process, showing only the ancestral relationships of the 6 samples and the coalescent times.*

Notice that although the WF process runs forward in time, we can only reconstruct the genealogy backward in time, after we are told which six present-day samples are relevant. At that point we can find the genealogy by tracing backward through the ancestors of the sample. Looking

forward in time, there is nothing particularly remarkable about the chromosomes that wind up being ancestors, versus those that do not, and their relevance to the present-day sample only becomes clear in retrospect.

Eventually, we reach a single common ancestor of the entire sample, known as the **most recent common ancestor (MRCA)**, which we will return to shortly.

Time to coalescence. For the sake of simplicity, the pictures above show coalescence within a few generations. But how long would coalescence take in real populations. In fact, how sure can we be that any two copies of this locus ever find a common ancestor – i.e., that they ever coalesce?

Looking backwards in time, each copy of this locus has a random parent from among the $2N$ possible chromosomes in the previous generation. So the probability that they both descend from the *same* parent is $1/2N$.

Conversely, the probability that they do *not* have a common ancestor in the last generation is $1 - 1/2N$. What is the probability that we go back at least t generations without a common ancestor? Assuming this process is independent from one generation to the next, we can multiply the probabilities, giving us

$$(1 - \frac{1}{2N})^t. \tag{2.14}$$

Now the important thing here is that $(1 - 1/2N)$ is < 1, so if we multiply it by itself many times, this number steadily approaches zero. This means that **if we go far enough back in time we can guarantee that any pair of copies of this locus have a common ancestor**.

Ok, so any two copies are guaranteed to eventually coalesce, but how long will this take? To answer this we need to take a short detour:

Understanding waiting-time distributions: the geometric distribution. To understand coalescent models you should know a bit about mathematical models of waiting times. To make this more concrete, suppose that I have a 20-sided die. I keep rolling the die until it lands with the '20' face up (and then stop). How many times do I need to roll the die?

Obviously, the waiting time is random: there is a 1 in 20 chance that the '20' comes up on the first roll – or I might need to roll many times. But we can calculate the average number of rolls, and we can also write down what is called the **probability distribution** which in this case is a general formula for the probability that the '20' first comes up on any specific roll.

First of all, we consider the probability of getting a '20' on any particular roll. We'll call this probability p, and it is simply $1/20$ since we have a 20-side die. Then the probability of NOT getting a 20, is 1-p, or 1-(1/20). One important property of probabilities is that the probability of multiple independent events is the product of the probability of observing each separately, so the probability of NOT

getting a '20' in the first t rolls is

$$(1-p)^t. \tag{2.15}$$

The probability of getting a '20' on the next roll is p, so the total probability that the first '20' occurs on roll number $t+1$ is

$$p \times (1-p)^t \tag{2.16}$$

This function describes the waiting times for events and it is called the **geometric distribution** [Link]. We can get a sense of how long you have to wait to roll a '20' by computing Equation 2.16 for different values of t. For example, there is a 0.4 probability (i.e., 40%) of rolling a 20 within the first ten rolls:

$$1-(1-\frac{1}{20})^{10}. \tag{2.17}$$

Can you be confident that you will eventually roll at '20' if you are patient enough? Yes. Using this formula, we find that there is a 64% chance of getting a '20' within 20 rolls, 99.4% probability of getting a '20' within 100 rolls, and 99.996% within 200 rolls. The probability of eventually getting a '20' converges to 1 as you roll infinitely long.

Lastly, an important property of the geometric distribution is that **the average waiting time to the first success is simply $1/p$**: so in this example, 20 rolls.

Understanding waiting-time distributions: the exponential distribution. The geometric distribution measures time in terms of a discrete number of events or trials. But for our purposes we can approximate the geometric with a continuous distribution called the **exponential distribution** [Link]. For our setting, the two distributions are virtually equivalent [158], but the exponential distribution is much easier to work with.

Like the geometric, the exponential distribution is also used to model waiting times, but in settings where time is measured in continuous units. For example, I might ask: "How long will it be until the next earthquake on the Stanford campus?". Let λ be the rate of earthquakes per day [159]. Then, according to the definition of the exponential distribution, the probability that the next earthquake will occur exactly t days from now is

$$\lambda e^{-\lambda t} \tag{2.18}$$

and the total probability of having an earthquake any time within the next t days is

$$1-e^{-\lambda t}. \tag{2.19}$$

An example of this function is plotted below. Finally, the average waiting time [160] to the next earthquake is $1/\lambda$. Notice this has the same form as the average waiting time in the geometric distribution, $1/p$.

In our models, we are interested in waiting times until coalescent events. We measure time in generations, and set λ to be the rate of coalescence per generation, namely $(1/2N)$; hence the average coalescence time will be $2N$ generations.

The time distribution for two samples. We're now ready to model the distribution of coalescent times for two copies of a locus.

Remember that each generation there is a probability $1/2N$ that the two copies will coalesce. As described above, we'll model the waiting time to coalescence using the exponential distribution, which an excellent ap-

proximation to the geometric (and easier to work with, mathematically).

The next plot shows what is called a "cumulative distribution" of coalescence times under this model (Equation 2.19 with $\lambda = 2N$). As we showed in the last chapter, for human populations, the longterm (effective) population size N is around $20,000$ [d].

The way to interpret this plot is that the y-axis shows the probability that two samples coalesce within the most recent t generations (plotted on the x-axis):

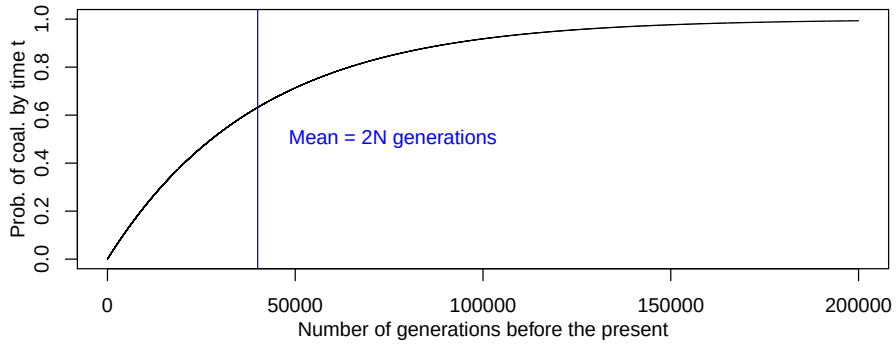

Figure 2.17: **Cumulative distribution for coalescence times.** *This shows the probability that two samples coalesce within the past t generations, assuming $N = 20,000$.*

As this plots shows, there is a 50% chance that coalescence occurs within the last $1.4N = 28,000$ generations (slightly less than the mean of $2N$ generations). And it's almost certain that coalescence occurs with $10N = 200,000$ generations. Note that if we assumed a different population size, this would change the numerical scale on the x-axis, but not the shape of the plot, which is simply proportional to N_e.

The last important point here is that *these timescales are really long* in terms of human evolution. Let's assume that the average generation time is about 25 years [161]... then the average coalescence time of $2N$ generations is 1 million years ago, before the appearance of anatomically modern humans.

The coalescent for larger samples. So far we have been talking about the coalescent for a pair of samples. Suppose instead that we sequence a particular locus in $m/2$ individuals, giving us a sample of m copies of the locus. Remember that the genealogy is embedded within the WF process.

How can we model the genealogy without having to bother with the WF process?

Imagine that we trace the ancestry of these m copies back in time. Going backward in time, *we will pick two of these lineages random* to coalesce into a common ancestor. Now (always looking backward in time) there are $m - 1$ copies. This process repeats until we get down to 2, and then finally to one common ancestor.

The process looks like this:

[d] *Remember that when we need to allow for the complexities of real world populations, the rate of coalescence depends on the* ***effective population size*** N_e*, rather than true population size N. For simplicity we'll discuss the models in terms of N, but you can think of subbing in N_e for real-life situations.*

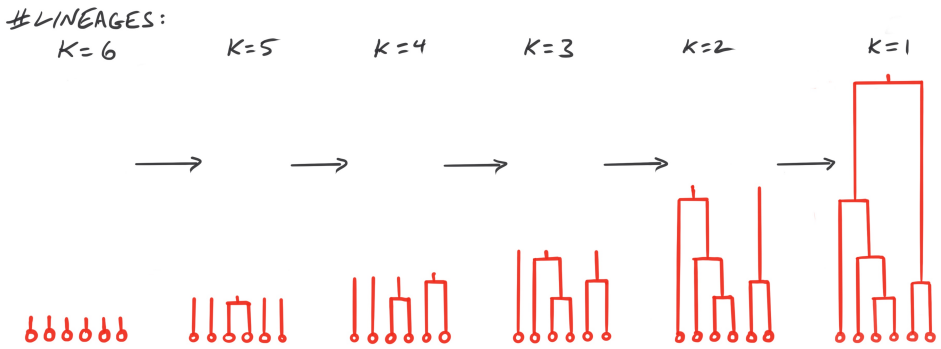

Figure 2.18: **Stepwise construction of a genealogy.** *At each step we randomly join two lineages. This results in a random* ***topology*** *– i.e., branching structure – that relates the m samples.*

Next we need to model the waiting times between coalescent events. We'll use T_k to be the number of generations when there are k lineages on the tree. (Here I use m as the number of samples in the present and k, ranging from 1 to m, as the number of distinct lineages at times in the past.) We showed above that T_2 has an exponential distribution, with a mean of $2N$ generations. What about larger values of k?

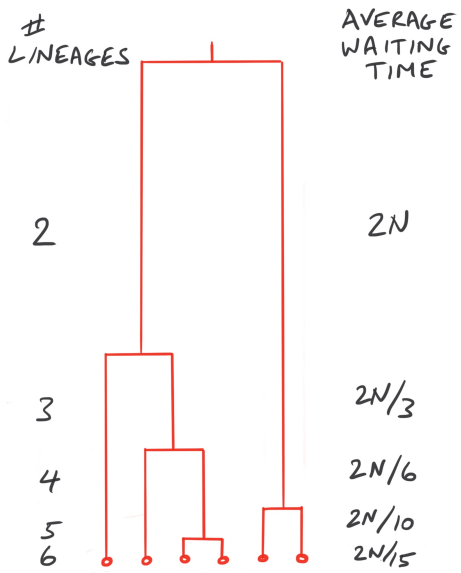

Figure 2.19: **Expected times in the genealogy.** *Here T_k labels the time during which there are k lineages. T_k is a random draw from an exponential distribution with mean $4N/k(k-1)$.*

How long does it take to go from k lineages to $k-1$?

To get this, we need to compute how long it takes for *any two* of the k lineages to merge into a single ancestor. The key thing here is that there are a lot of possible pairs that we could make out of k lineages. Specifically, there are [e]

$$\frac{k(k-1)}{2} \text{ possible pairs.} \tag{2.20}$$

Since there are $k(k-1)/2$ ways to get a possible coalescent event, this means that the waiting time to the first coalescence is reduced by a factor $2/k(k-1)$ compared to the waiting time when there are only two samples. Specifically, the waiting time when there are k lineages is exponen-

[e] *You can compute the number of possible pairs as follows. List the k lineages in some arbitrary order. The first lineage can pair with $k-1$ other lineages; the second can form $k-2$ pairs not counting the pair with the first lineage... and so on. The sum $(k-1)+(k-2)+(k-3)+...+2+1$ equals $k(k-1)/2$.*

tially distributed with mean:

$$\mathrm{E}[T_k] = \frac{4N}{k(k-1)} \tag{2.21}$$

For example when $k = 2$, the average waiting time is $2N$ generations. When $k = 10$, the average waiting time is 45-fold shorter: $2N/45$ generations [f]. When $k = 100$, the average waiting time is nearly 5000 times shorter: $2N/4950$.

In other words, **the most recent coalescent events – when there are many lineages – occur within a few generations, while the oldest coalescent events can easily take a million years.**

One question of particular interest is: *How long ago was the MRCA of a sample (or even of the entire population)?*

[f] *To get some intuition for this, imagine k cars with blindfolded drivers, driving erratically around a large parking lot. The time until the first crash is much shorter when there are many cars, and hence many possible pairs that could crash. E.g., with two cars the time until the first crash would be 45 times longer than for ten cars.*

Optional math on time to the MRCA. To compute the time to the MRCA, we add together the waiting times between each node. Here $T_{\mathrm{MRCA(m)}}$ is the random time to the MRCA for a sample of size m.

$$T_{\mathrm{MRCA(m)}} = T_2 + T_3 + T_4 ... + T_{m-1} + T_m, \tag{2.22}$$

where T_k represents the random waiting time during which there are k lineages, and is an exponential random variable with mean $4N/[k(k-1)]$. So the average time to the MRCA is:

$$\mathrm{E}[T_{\mathrm{MRCA(m)}}] = \sum_{k=2}^{m} E(T_k) \tag{2.23}$$

$$= \sum_{k=2}^{m} \frac{4N}{k(k-1)} \tag{2.24}$$

As the sample size gets large, this sum converges to a fixed value (the derivation requires techniques on infinite series):

$$\lim_{m\to\infty} \mathrm{E}[T_{\mathrm{MRCA(m)}}] = 4N. \tag{2.25}$$

In other words, as the sample size goes to infinity (or in practical terms, the entire population), what we see is that on average the most recent common ancestor for the entire population is $4N$ generations in the past.

The key result here is that for an average location in the genome, **the common ancestor for the entire population is 4N generations ago** ($\sim$2 million years, for humans). On average, **half of the total time back to the common ancestor is spent waiting for the last two lineages to coalesce.**

The genealogy has both random topology and random times. Before moving on, I want to emphasize one last important point about the coalescent: although we have been focusing on average properties, genealogies are inherently random, and vary in two important ways: *both the*

topology *(i.e., branching patterns) and coalescent times are random draws from the coalescent process*. This is illustrated below for genealogies with $m = 4$:

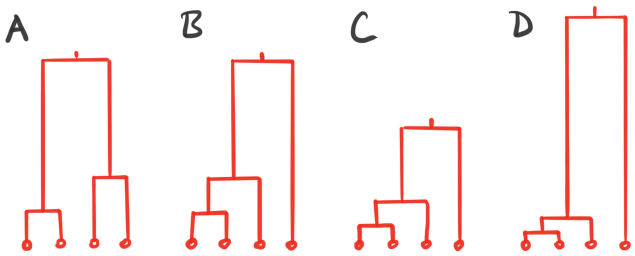

Figure 2.20: **Random outcomes of the coalescent process. A** *and* **B** *differ in their branching patterns (topologies), while* **B**, **C**, *and* **D** *have different coalescent times.*

In practice, the genealogies in different regions of the genome vary widely, and as we shall see next, this influences the allele frequencies and numbers of SNPs at any given locus.

Coalescent with mutation. So far, we have been talking about the genealogy. The genealogy reflects the inheritance of a DNA segment through time. Conceptually you can think of this as tracking the copying of DNA molecules through thousands of meioses, and reflecting the fact that a particular stretch of DNA in different people is a copy of the same ancestral DNA sequence in some distant ancestor.

Now we need to add mutations into the model. Patterns of genetic variation in modern samples reflect the combination of coalescence and mutation. As you get used to the structure of the coalescent, it provides a powerful tool for understanding patterns of genetic variation. We'll come back to this theme repeatedly in the upcoming chapters.

To make this concrete, let's suppose that we sequence a stretch of L base pairs ($L = 5$ kb, for example) in m samples (without recombination). We assume that new mutations arise at a rate μ per base pair per generation. It's going to be helpful now to label the lengths of branches on the tree (in generations); we'll do this using b_i for branch i:

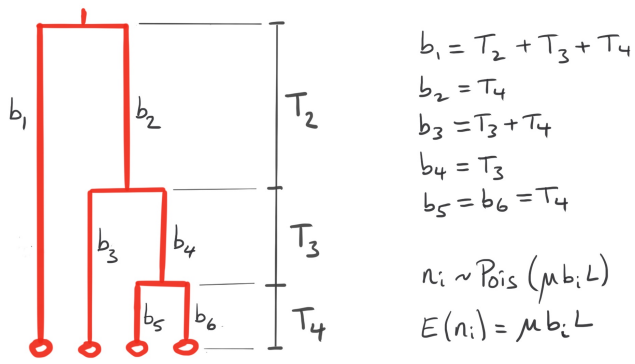

Figure 2.21: **Example of branch lengths in a genealogy.** *The branch lengths b_i show the length (in generations) of each branch. The numbering is arbitrary. Note that the specific branching patterns depend on the random tree topology.*

Notice above that we can write the branch lengths b_i in terms of the times between coalescent events (remember that T_k denotes the time when there are k lineages), although the specific branches and their lengths depend on the random topology.

Now, let n_i be the number of mutations on branch i. What is the **expected number of mutations $\mathbf{n_i}$**? This is the product of the mutation rate μ, branch length b_i, and sequence length L:

$$\mathrm{E}[n_i] = \mu b_i L \tag{2.26}$$

That is the *expected* number, but the actual number of mutations on any particular branch is random; this is modeled using the **Poisson distribution**. For more about the Poisson see [162].

Here's an example of what this might look like for a sample tree. Mutations are shown on each branch in blue; the tips of the tree (A-F) show six samples collected in the present day. *Mutations occur in ancestors along each branch, and are inherited by all the samples that lie below them.* So for example on the tree below, mutation 1 is inherited by sample A only, while mutation 2 is inherited by samples A-D. This means that we can go from the tree on the left, to the haplotypes on the right:

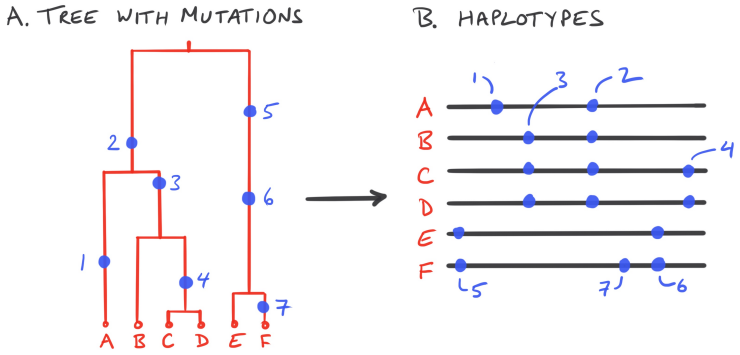

Figure 2.22: **Coalescent tree, mutations, and haplotypes. (A)** *Example genealogy with 7 mutations.* **(B)** *The corresponding haplotypes, with blue circles indicating derived alleles (at arbitrary locations).* *The labeling A-F corresponds to the sample labels in the tree. The assignment of alleles to haplotypes is entirely determined from panel (A). However the sequence positions of the mutations were assigned randomly while drawing panel (B).*

Trees, branches, and derived allele frequencies. The picture above hints at a key connection between the tree topology and the allele frequencies in a sample. If a branch is above j samples, then any mutation on that branch will occur exactly j times within the sample. This is shown below for two example topologies:

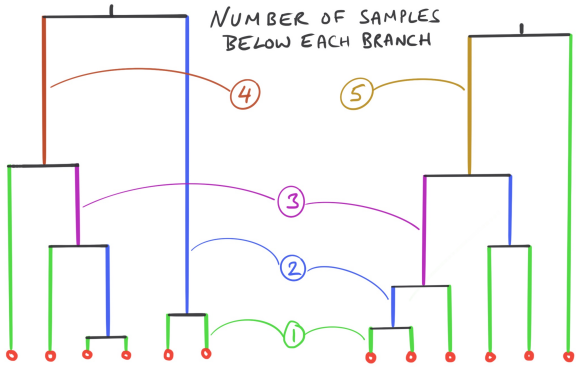

Figure 2.23: **Tree topologies and allele counts.** *Branches are colored according to the number of samples (tips of the tree) below each branch. For example, mutations that occur on blue branches will be present exactly twice in the sample. Notice that the branch lengths and possible allele counts differ between the random tree topologies.*

The branches labeled in green, above, are of particular interest as they lead to just a single sample. These are often referred to as **terminal branches**, and mutations that occur on them are referred to as **singletons** as they are found in only a single sample.

Quantitative aspects of variation in the coalescent. Thus far, I have described the coalescent at a conceptual level, as a way of understanding the structure of genetic variation. But we can also use it as a tool for making quantitative predictions about variation.

To start: *How many sequence differences can I expect between two samples, in a region of L basepairs?*

Recall that the coalescent time for two samples, T_2, is exponentially distributed with mean $2N$ generations, with an average μL mutations per generation along each branch. It follows that the expected number of mutations between each sample and the common ancestor is $T_2 \mu L$, and twice that for the total number of differences between the two modern day samples [163]:

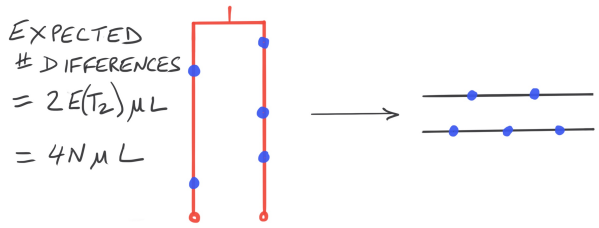

Figure 2.24: **Number of differences between two samples.** *The expected number of differences between two samples (equivalent to heterozygosity) is the product of their average coalescent time ($2N$) times the mutation rate along both branches, $2\mu L$.*

It's convenient to divide this by L, which gives us the expected number of differences per base pair. That is equivalent to **heterozygosity per site**, H, which we computed in the last chapter using the WF forward model:

$$H = 4N\mu. \tag{2.27}$$

Happily, the forward and backward approaches gives us the same result.

Most heterozygous SNPs are very old. I mentioned before that you have about 3 million heterozygous SNPs in your genome. How old are the mutations that produced these heterozygous alleles?

Using this model we know that for any random part of your genome, the average time to the common ancestor of two homologous copies is $2N$ generations (or about 1 million years). On average, a mutation occurs halfway along the branch to the common ancestor... this tells us that **the average variant in your genome is due to a mutation that happened 500,000 years ago(!)** and many are much older [164].

To put this into perspective, modern humans evolved in sub-Saharan Africa. About 70,000 years ago, some populations started spreading out of Africa into the Middle East, and then went on to colonize nearly all of the world's landmasses [g].

The number of SNPs found in a sample. The calculation above tells us the expected number of mutations in a sample of 2. How many mutations should we expect in a sample of size m?

[g] *I like to think there is some beauty in the fact that most of the heterozygous sites in your genome, or in mine, arose as mutations in distant ancestors in Africa half a million years ago.*

Optional: number of SNPs in a sample (the math). Suppose that you sequence a region of L basepairs in m samples. What is the expected number of variable sites (i.e., SNPs) that you detect? In addition to the result itself, this box illustrates the kinds of calculations that are (relatively) easy to do using the coalescent.

To get this, notice that we can break the problem down into two parts: (1) What is the **total branch length** – i.e., the sum of all the branch lengths; and (2) How many mutations do we expect to have occurred given the tree length?

To get the tree length, you might want to start by thinking about computing the length of every branch, and then adding all those together. But this is complicated because it depends on the branching structure of the tree, which is random. Instead, we can make the calculation easier by adding together a contribution from the time between each coalescent event. Specifically, what is the total branch length during the time when there are k lineages? Well the expected time is $4N/k(k-1)$, and there are k branches; multiplying these together gives $4N/(k-1)$ total branch length in this epoch. Next, adding together all the epochs, the expected total tree length is

$$\sum_{k=2}^{m} \frac{4N}{k-1}. \tag{2.28}$$

Then we can multiply by the mutation rate to get the expected number of variable sites (denoted S) in a sample of size m, in a region of L base pairs. After minor rearrangement and a shift in the sum index we get:

$$S = 4N\mu L \sum_{k=1}^{m-1} \frac{1}{k}. \tag{2.29}$$

When $m = 2$ this agrees with the result we got before for heterozygosity.

Equation 2.29 in the box provides an important result: the expected number of SNPs in a sample of size m.

One key point is that **as the sample size grows the MRCA time converges to 4N, while the number of segregating sites grows indefinitely at a rate proportional to the log of the sample size, $\ln(\mathbf{m})$** [h]. This is because as you increase the sample size, new samples usually add additional short branches near the bottom of the tree – slightly increasing the total branch length but not changing the MRCA time.

[h] *In human populations the number of rare variants actually grows a bit faster than ln(m), for reasons we'll explain shortly.*

The site frequency spectrum (SFS). Suppose that we collect genome sequence data from m samples. Let s_i be the number of SNPs at which the derived allele is present exactly i times. For example, s_1 gives us the number of singletons, s_2 the number of doubletons, and so on. The total number of SNPs, S, is related to s_i simply by summing over all the possible allele frequencies from 1 to $m-1$:

$$S = \sum_{i=1}^{m-1} s_i. \tag{2.30}$$

The vector of allele frequencies s_1, s_2, s_3, ..., is referred to as the **site frequency spectrum (SFS)**, and is a simple but important description of genetic variation [i].

[i] *As we'll discuss later, some types of natural selection, as well as other departures from the basic model such as recent population growth, can be detected because they distort the SFS away from this baseline model.*

What determines the SFS? Take a look back at Figure 2.23. The expected value of s_i depends on the amount of branch length that sits above exactly i samples: for example, s_2 depends on the amount of branch length that sits above pairs of samples. If we focus on genome-wide data, this has the effect that we will sample many different trees (in different parts of the genome) so that we can average over the randomness of the coalescent process.

The derivation of the SFS is beyond the scope of this book, but you can read about it here: [165]. Although the math is a little tricky, it produces the pleasingly simple result [j] that the expected number of variants with a derived allele frequency i is proportional to $1/i$:

$$\mathrm{E}[s_i] = \frac{1}{i} \times 4N\mu L. \tag{2.31}$$

[j] *This result also implies that the expected tree length above exactly i samples is $4N/i$.*

This distribution is plotted here:

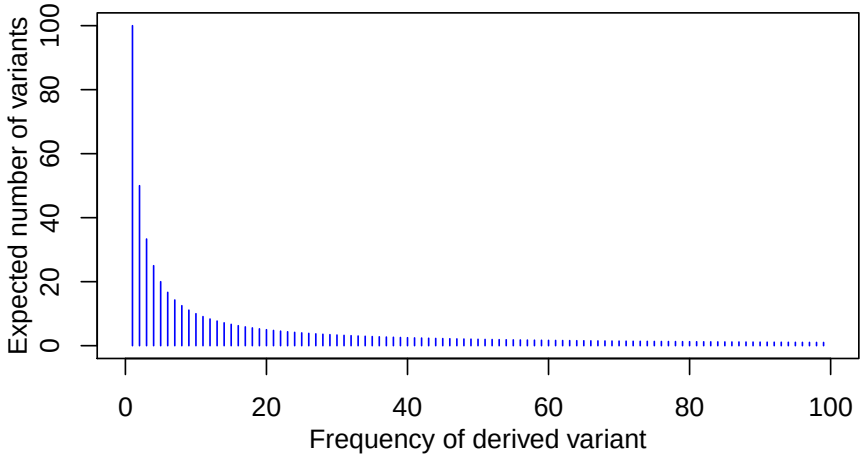

Figure 2.25: **The Site Frequency Spectrum (SFS).** *Here the expected SFS is plotted for $m = 100$ and $4N\mu L = 100$. Notice that most variants are rare. Here, 55% of the variants are below 10% frequency. This pattern is even more dramatic in large samples: in a sample of $m = 10,000$, 76% of the variants are at $< 10\%$ frequency.*

One key thing to notice is that most variants are rare. A useful rule of thumb is that allele frequencies are uniform on a log scale: in very large samples there are as many variants with derived frequencies between 0.1 and 1 as there are between 0.01 and 0.1, or between 10^{-5} and 10^{-4}.

Lastly, we can also get intuition for this from the WF model. In the last chapter I pointed out that most derived alleles are very rare, and only a small fraction are common: every new mutation starts out rare (i.e., at $1/2N$ frequency). Most are lost quickly, while only a few are lucky enough to drift up to become common. Thus, the WF model gives us a different conceptual tool to reach a similar conclusion.

The coalescent with population size changes. I have been describing the coalescent under the simplest possible population model: constant size and no population structure. This basic model is referred to as the **vanilla coalescent**.

But real populations often differ from this simple model, and it's important to think how this might affect the coalescent. In this section I'll de-

scribe how to think about two types of changing population size that are important for humans: bottlenecks and population growth.

Population bottlenecks. In population genetics, a bottleneck refers to a reduction in population size, often but not always followed by a return to the original population size. Bottlenecks are important because they greatly increase the rate of genetic drift.

Bottlenecks have been important features of human evolution, including during the spread of populations as they left Africa and colonized the globe during the past $\sim 80,000$ years [166]. This is why non-African populations have less genetic variation than Africans.

Bottlenecks have also reshaped patterns of variation in some populations within much more recent timescales – for example the ancestors of modern Jews went through a tight population bottleneck $\sim$1000 years ago [167].

In the WF model, we can think of the bottleneck as increasing the variance in allele frequencies: some alleles increase dramatically, while others decrease:

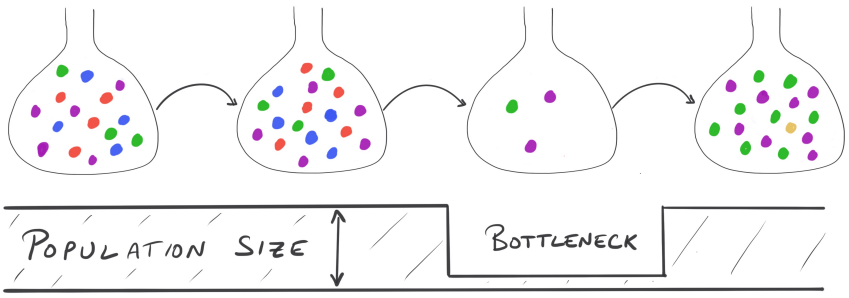

Figure 2.26: **WF drift through a bottleneck.** *Bottlenecks greatly increase the rate of drift due to low N_e.*

Of course, we can also think of this in terms of the coalescent. Remember that the rate of coalescence is $k(k-1)/4N$ per generation. If our model allows N to vary with time then, when N decreases, the rate of coalescence will increase at an inverse rate.

This means that we will get an increased rate of coalescence within the bottleneck, and fewer ancient lineages. The few lineages that predate the bottleneck are likely to have many descendants:

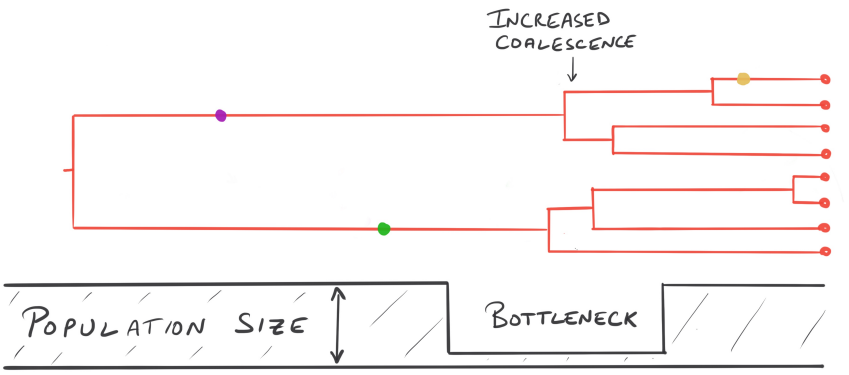

Figure 2.27: **Coalescent through a bottleneck.** *The rate of coalescence during the bottleneck is greatly increased due to low N_e. The purple and green mutations occurred on lineages that survived the bottleneck and are at high frequency in the final sample (at right). The yellow mutation postdates the bottleneck and is at low frequency. The tree here is tipped on its side to emphasize similarity to the WF picture above.*

This example also helps to **illustrate the intimate connection between coalescence and drift**: in a sense, drift in the WF model occurs *because* lineages are coalescing.

Population growth. Another key feature of real human populations is dramatic population growth, from ∼1 million in 10,000 BCE to ∼8 billion today. How did this affect the coalescent process, and genetic variation?

Here, the logic is opposite the bottleneck situation: a very large population size slows down the rate of coalescence at very recent times. As a result, **recent growth hugely increased the number of very rare variants**.

To understand this, it would be most natural to model population growth as following an exponential increase over time [168]. But the math for coalescence with exponential growth is a bit clunky and obscures the main points, so we'll consider a simpler model:

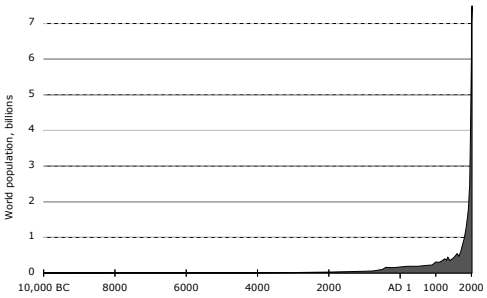

Figure 2.28: **Exponential human population growth.** *Estimates of total world population during the past 12,000 years.* *Credit: El T [Link], Public Domain. Data: [Link].*

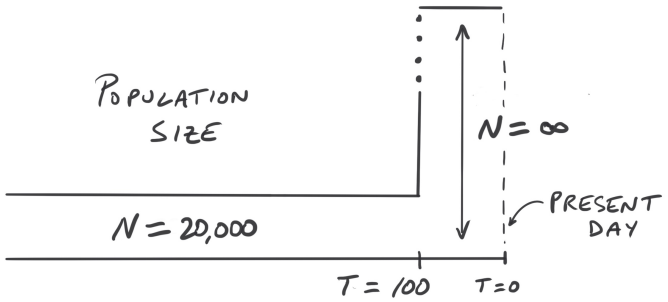

Figure 2.29: **Instantaneous growth.** *In this simplified model, the ancestral population size is* 20,000*, followed by instantaneous growth to infinite size* 100 *generations ago.*

In the model above, we consider a population that grew instantaneously to infinite size, 100 generations ago. How would this extreme model change the properties of trees, compared to a model of constant $N = 20,000$?

Recall that in the vanilla (constant size) model, for large samples the first coalescent events occur very quickly. But in the infinite growth model, there is no coalescence in the most recent time period, thus greatly extending the terminal branches:

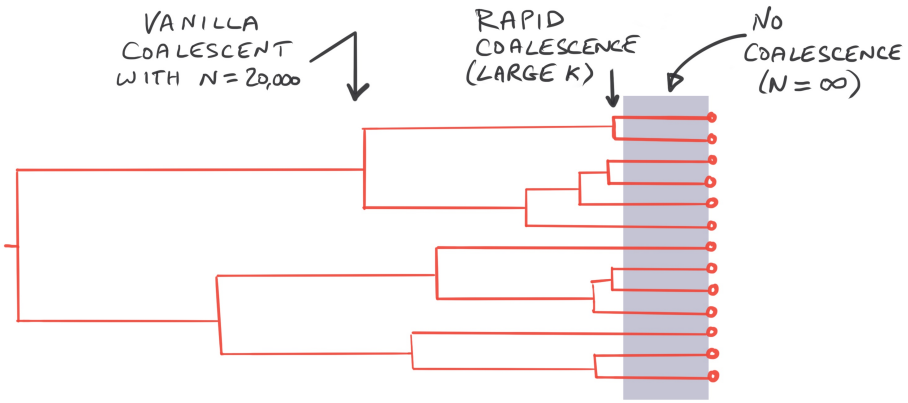

Figure 2.30: **Coalescent tree in the instantaneous growth model.** *There is no coalescence for the first* 100 *generations (grey region) due to infinite population size. Note: the picture is not drawn to scale.*

The longer terminal branches produce many more singleton mutations. Recall for the vanilla model that the expected number of singletons is

$4N \times \mu L$ (Equation 2.31).

But in the infinite growth model, every tip is extended by 100 generations. Since there are m tips, the expected number of singletons is now $(4N + 100m) \times \mu L$. So for example, in a sample of $m = 1000$, the number of singletons is more than doubled! Meanwhile, the deeper structure of the tree is unaffected, aside from slightly pushing back all the expected times.

While the infinite growth model is unrealistic it still provides valuable insight. Under a more realistic model of continuous exponential growth there is a strong reduction in the rate of recent coalescence relative to the vanilla model, thereby increasing the lengths of recent branches. In summary, **recent exponential growth leads to a dramatic increase in low frequency variants.**

Footprints of population history in real data. In a 2012 paper, Tennessen et al described genome-wide (exome) sequencing data from about 1100 individuals of African-American, and 1300 individuals of European-American ancestry [169] [170]. They found over 500,000 SNPs, of which 86% were at less than 0.5% frequency and 57% were singletons.

The plot below illustrates the SFS for these two samples: Each line plots the proportion of alleles (Y-axis) in bins of allele frequencies (X-axis). Both axes are on log-scales; on these axes the theoretical null (constant N) is approximately a straight line [171].

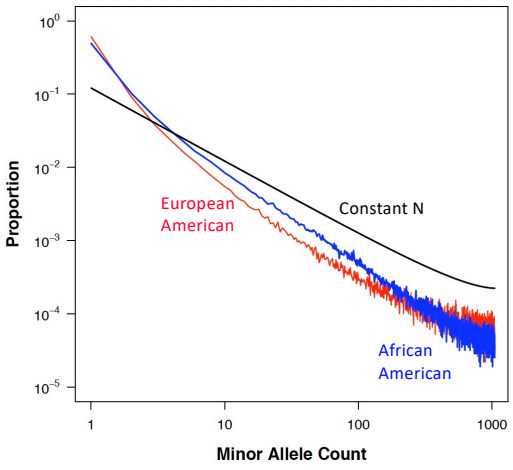

Figure 2.31: **The SFS in human populations: huge excess of rare variants.** *Notice that the real data (colored lines) are well above the theoretical prediction (black line) in the upper-right hand part of the plot.* *Credit: Modified Figure S9D from Jacob Tennessen et al 2012 [Link] Used with permission.*

As you can see, the real data from both populations show a much higher fraction of rare variants (higher in the upper right) compared to the null. This is direct evidence for rapid recent population growth.

The authors then fit a model of historical population sizes (often called a **demographic model**) that can fit the full SFS data. The model is shown below, including a tight European bottleneck, and extreme recent population growth to reflect the huge excess of rare variants relative to the vanilla coalescent model:

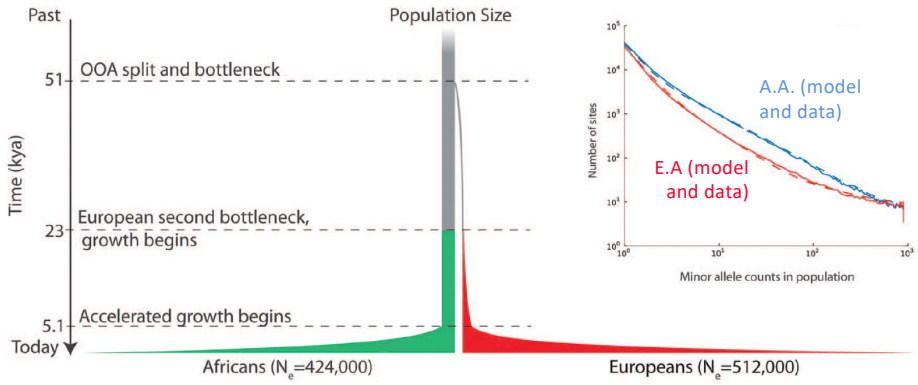

Figure 2.32: **Fitted demographic model.** *This model is designed to fit the SFS data for African- and European Americans (see upper-right panel).* ***The inferred model illustrates extreme recent growth in both populations, and a strong European bottleneck.*** *Note that more recent estimates include times and population sizes that are roughly doubled due to updates in mutation rate estimates since 2012. The image was simplified by not showing European mixing into African Americans in the last 400 years. Credit: Modified Figure 2B from Jacob Tennessen et al 2012 [Link] Used with permission.*

As you can see in the inset in the upper right of the plot above, the proposed model provides a good fit to the observed SFS. While the precise parameter estimates vary among papers in this area, all of them agree on the presence of a tight bottleneck for non-African populations, and extreme recent population growth.

The coalescent and the fixation process. Thus far we've been using the coalescent to understand genetic variation. But it also provides a useful intuition for understanding how alleles fix. *Crucially: A variant is fixed in the present day if and only if it was present in the population MRCA.*

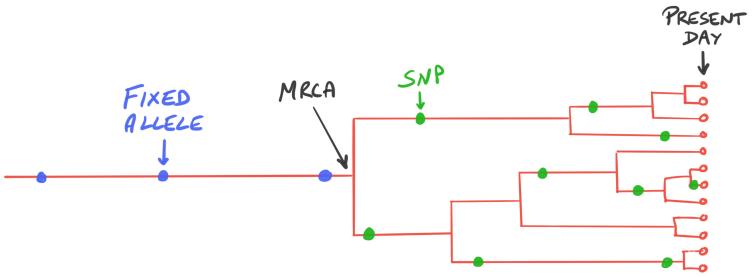

Figure 2.33: **Fixed variants are present in the population MRCA.** *The blue variants are fixed in the present day population because they were carried by the MRCA; the green variants are SNPs.* *Assume that the MRCA shown for this sample is in fact the MRCA of the entire population.*

This now provides intuition for two important results that I stated in the last chapter:

The probability of fixation for an allele now at frequency p, is simply p. You now know that any present-day sample has a common ancestor sometime in the past. Flipping this around, if we imagine going far enough forward in time (on the order of $4N$ generations), we know that exactly one copy of this locus will eventually be a common ancestor of the entire current population. So for a SNP now segregating at frequency p, there is a probability p that in the future a lineage carrying it will become the ancestor of everyone.

The average time to fixation for a new mutation is $4N$ generations. The logic here is similar: If a new mutation eventually fixes, this means that it

is destined to become the common ancestor of everyone in a future population. We know that the expected time back to a common ancestor is $4N$ generations. Forward in time, it takes approximately $4N$ generations until the first time that the mutation is the common ancestor of everyone [172]:

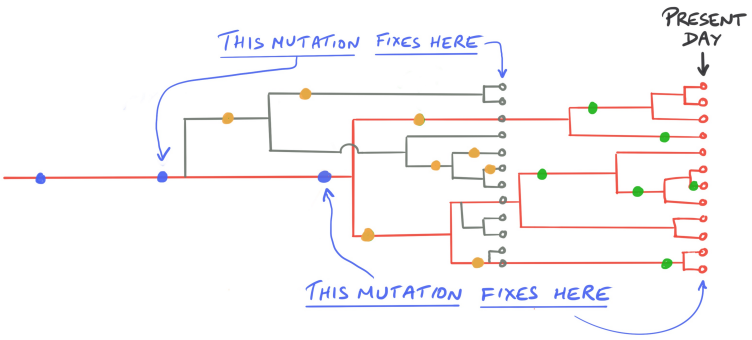

Figure 2.34: **Mutations fix at different times** *depending on when their lineages first become the population MRCA. All three blue mutations are fixed in the final population at the far right, but the first time at which they fix depends on the structure of the coalescent.* *Also of interest: the yellow variants are SNPs in the gray sample, but many of them are lost by the time of the final (red) sample.*

Coalescent simulation of haplotype variation [Optional]. As in the previous chapter we end with a basic outline for how to simulate haplotypes, this time using the vanilla coalescent model. If you're good at programming you may wish to try this [173].

Data storage: It's useful to create a *data structure that represents nodes of the tree*. There are m of these to represent each of the present day samples, and $m-1$ for the ancestral tree nodes. Each node stores the time, as well as a pointer to the parent node, and to each child. (The child nodes are null for the present day samples, and the parent pointer is null for the MRCA.) It also stores a list of the derived variants present at this node.

You'll also want:
- a list of the *locations of mutations* within the sequenced region;
- the *current time* before present, for use while constructing the tree;
- a *list of current active lineages (nodes)*, for use while constructing the tree.

Construct tree:
Initialize the current time at 0.
The initial active lineages are the m present-day samples. Set the times for these to 0 and all their children to null.

for (k starting at $k = m$, down to $k = 2$) do:
{

- *Coalesce lineages:* Pick two of the active lineages at random to coalesce. Create a new node, with these two lineages as children. Drop those from the active list and replace with the new node;
- *Generate node time:* Update the current-time by adding a random time $\sim$ $Exponential(4N/k/(k-1))$. Set the time at the new node equal to the new current-time;
- *Update lineage counter:* $k = k - 1$

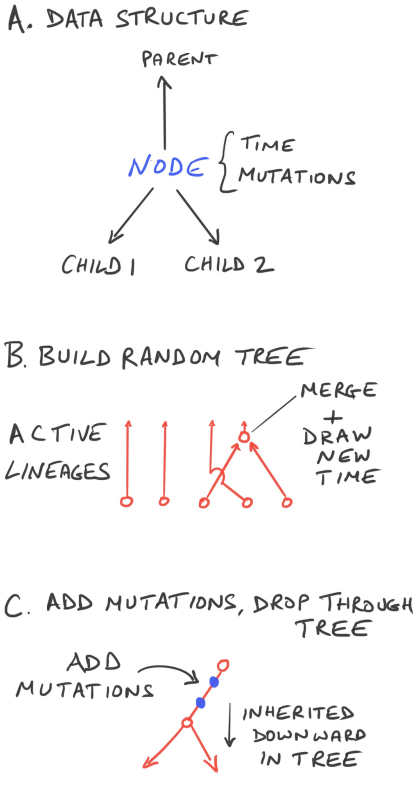

Figure 2.35: **Coalescent simulations.**

}

Add mutations: Starting from the top of the tree, visit each branch i in turn and do:

{

- *Calculate the branch length* b_i as the elapsed time between the parent node and the child node;
- *Simulate the number* of mutations as $\sim Poisson(b_i \mu L)$;
- *Simulate the position* of each mutation as $\sim Uniform(0, L)$;
- *Drop the mutations* down through every node below the mutated branch to the present-day samples.

}

Comments. This is conceptually a bit more complicated than the Wright-Fisher pseudocode, but it's far more computationally efficient, as we don't need to track a huge number of ancestors that are not relevant to variation in the present-day sample.

This type of algorithm provides an extremely efficient tool for simulating genetic data. As a rule, coalescent simulations are much faster than WF simulations, but they can be less versatile, and more difficult to modify to new situations. There are numerous **free software packages** for coalescent simulations, including **msprime** [Link].

Well done! In the last two chapters you have learned the two most fundamental tools for understanding patterns of genetic variation. In the next two chapters we'll discuss how to fold recombination and population structure into these basic models.

Notes and References.

[154]Credit for finding this quote goes to the late Paul Joyce: [Link].

[155]We'll talk more about these early data in Chapter 2.7, along with the other major conceptual development of the 1970s and 80s, the Neutral Theory.

[156]Inspiration for the coalescent was motivated in part by developments in population genetics during the 1970s. John Kingman (later Sir John Kingman) was a mathematician at the University of Oxford with particular interest in stochastic processes. He came to this problem after conversations with a group of Australian population geneticists: Pat Moran, Warren Ewens, and Geoff Watterston. In a trio of papers published in 1982, Kingman framed the process in highly mathematical terms and published in mathematical journals; in one of these he coined the term "coalescent" (hence the occasional name "Kingman Coalescent" for this model). Kingman only worked in population genetics for a couple of years. Despite the huge impact of the coalescent work, Kingman commented to me many years later (2022) that "Coalescent theory is very far from the thing I am most proud of", preferring instead his contributions in queuing theory (which later became important in the development of the internet [Link]), and perhaps his role as a university administrator, including as head of the University of Bristol (England) starting in 1985.

Meanwhile, Richard (Dick) Hudson was a PhD student at the University of Pennsylvania and at UC Davis. He published a pair of papers a year after Kingman (but unaware of Kingman's work) that describe–almost as an afterthought– the nuts and bolts of the basic coalescent model, as well as important extensions to handle the coalescent with recombination, all for the purpose of performing highly efficient simulations. He later went on to develop extensive tools for coalescent simulation.

The third key person, Fumio Tajima, a Japanese scientist then at the University of Texas Houston, published a 1983 paper that outlines the structure of genealogies and the coalescent and showed how this can be used to derive important sample statistics in population genetics. Published in the same year as Hudson's work, in some ways Tajima's presentation is the most modern in flavor (and is the paper in which I first encountered the coalescent as a graduate student, some ten years later).

Kingman JFC. The coalescent. Stochastic processes and their applications. 1982;13(3):235-48,
Kingman JF. Origins of the coalescent: 1974-1982. Genetics. 2000;156(4):1461-3,
Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 1983:203-17,
Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201,
Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437-60

[157]Early, highly readable reviews of the coalescent were written by Dick Hudson and Magnus Nordborg. (You can find online versions of the book chapters via Google Scholar: for Hudson 1990 see [Link]; for Nordborg 2000 see [Link])

Hudson RR. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. 1990;7(1):44
Hudson R. The how and why of generating gene genealogies. Mechanisms of molecular evolution. 1993:23-36
Nordborg M. Coalescent theory. Handbook of Statistical Genomics: Two Volume Set. 2019:145-30 .

[158]Differences between the geometric and exponential only arise in very special settings: for example when the sample size is large compared to the total population, and also in problems looking at coalescence within relatives.

[159]At the time of writing there have been two major earthquakes at Stanford (in 1906 and 1989) since its founding in 1885. So a simple-minded estimate of λ for major earthquakes would be $\sim 4 \times 10^{-5}$ per day. For an entirely gratuitous picture of a smashed car outside Stanford's Old Chem Building in 1989 see [Link]. USGS data: [Link].

[160]The mean of the exponential distribution with rate parameter λ is given by

$$\int_{t=0}^{\infty} t \cdot \lambda e^{-\lambda t}\, dt = \lambda^{-1}. \tag{2.32}$$

[161]Estimates for long-term average generation times are in the 25-30 year range. I chose 25 here to make round numbers, and that's roughly balanced by using a population size on the high end for human populations.

[162]The **Poisson Distribution** is a widely used model for the (random) number of rare events that occur in a specified time – for example the random number of earthquakes in a 100-year period. It depends on a single parameter, which gives the expected number of events. To read more see [Link].

$$\text{number of mutations} \sim Poisson(\mu L b_i) \tag{2.33}$$

[163] We want to compute the expected number of pairwise differences, m, between two samples under a constant population size model. Note that m is distributed as Poisson$(2\mu LT)$, where μ is the mutation rate per base pair per generation, L is the length of the region in base pairs, and T is the realized coalescent time of the two samples. We use $\Pr[T]$

to denote the probability density function for T (i.e., the exponential distribution with mean $2N$). Then we have:

$$\begin{aligned} \mathrm{E}[m] &= \int_0^\infty \mathrm{E}[m|T]\ \Pr[T]\mathrm{d}t && (2.34)\\ &= \int_0^\infty (2\mu LT)\ \Pr[T]\mathrm{d}t && (2.35)\\ &= 2\mu L \int_0^\infty T\ \Pr[t]\mathrm{d}t && (2.36)\\ &= 2\mu L\ \mathrm{E}[T] && (2.37)\\ &= 2\mu L 2N = 4N\mu L && (2.38) \end{aligned}$$

or simply $4N\mu$ per base pair.

[164]The mean is actually a bit older than this even, because there's an additional ascertainment effect in which the distribution of coalescent times at sites with variation is older than the unconditional mean.

[165]For a proof of the θ/i result, by Richard Hudson, see
Hudson RR. A new proof of the expected frequency spectrum under the standard neutral model. Plos One. 2015;10(7):e0118087

[166]Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences. 2005;102(44):15942-7

[167]Waldman S, Backenroth D, Harney É, Flohr S, Neff NC, Buckley GM, et al. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. Cell. 2022;185(25):4703-16

[168]The classic paper on exponential growth is
Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129(2):555-62

[169]Tennessen JA, Bigham AW, O'connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9

[170]I'm highlighting this work because it illustrates our major points. There is a long history of papers in this area, with sample sizes and genome coverage generally increasing over time.

[171]The slight uptick at the right occurs because the data are plotted in terms of the minor allele frequency instead of derived allele frequency.

[172]This argument is not entirely rigorous, and the classic results on this use forward-in-time diffusion theory.

[173]Here is a link to some similar sample code by Goncalo Abecasis [Link]. When I get time I expect to post a file that follows this code more closely.