2.1 Genetic Drift: What happens to alleles over time?

The two copies of your genome (one inherited from your mum and one from your dad) differ at about 3 million SNPs. Each of these arose as a point mutation some time in the past: about 70 are new mutations from your parents, while most of them are inherited from very distant ancestors. In fact, most SNPs that you carry arose as mutations in distant ancestors, hundreds of thousands of years ago, living in sub-Saharan Africa. In this chapter, and the next one, I'll explain why.

Every generation, new mutations are introduced into the population (around 70 per child). You can imagine tracking what happens to these mutations over time. Most mutations are lost from the population within a few generations, but sometimes a mutation can increase in frequency by chance alone. The random changes in allele frequencies over time are known as **genetic drift** [a].

[a] *Until Chapter 2.5 we'll assume that all variation is **neutral**: i.e., that there is no advantage to having one allele or the other. This is a good assumption for the vast majority of point mutations.*

You can think about the spread of a new mutation as being like what would happen if you walked into a casino with a dollar. You decide that you are going to keep playing until you either go bust or you beat the house. Most likely, you go bust pretty quickly, but if you have some early luck, you might be able to build up your cash reserves and play for a while. Very very rarely (theoretically at least) you might be able to play long enough to bankrupt the casino [136].

This is how it is for a new mutation. Most mutations are **lost** from the population within just a few generations (that's like you going bust in the casino). But a tiny fraction spread by chance to be common. And a very few, eventually, spread throughout the entire population, so that the newer allele reaches frequency 1 (that's like you bankrupting the casino). We refer to this situation as **fixation**; or we say that the new variant has **fixed**.

Figure 2.1: **The casino analogy** *is not entirely accurate because real casinos have a built-in advantage which means that you have slightly less than fair odds of winning each bet; that's not to mention the big burly guys who come over when you start to beat the house and discuss loudly how much they enjoy busting kneecaps.*
Credit: Lucy Pritchard

Before we go on, I need to remind you of some jargon: At the position of a mutation, we'll refer to the *original* allele as the **ancestral allele**, and the *new* allele will be the **derived allele**.

A new derived allele (i.e., a new mutation) starts out with 1 copy in the population. If use **N to denote the number of individuals** in the population, then the starting **allele frequency p** of a derived allele is

$$p = \frac{1}{2N}.$$

The factor of 2 in the denominator is to account for the fact that chromosomes come in pairs so everyone has two copies of each locus (for the autosomes [137]).

Over time, due to genetic drift, the allele frequency p either drifts down to zero (the derived allele is lost) or eventually drifts up to 1 (the derived allele is fixed). As I will explain, a new allele is usually lost within a few generations, but fixation takes tens of thousands of generations.

Thought experiment: random changes in allele frequencies over time. The tiny and remote island of Pitcairn is situated in the south Pacific, roughly halfway between New Zealand and Peru. It is currently home to about 50 inhabitants. They are descended from a party of 29 founders who landed there in 1790: nine mutineers from the British ship Bounty, along with 20 Polynesians they had kidnapped. The population has never exceeded 250 inhabitants [138].

Figure 2.2: **Pitcairn Island.** *NOAA, Public Domain [Link]*

Imagine an A/G SNP that was present in the founding population of the island. Suppose that the derived allele G was at a frequency of 55% in the founding group in 1790. Assuming there's no advantage to having either A or G in terms of either survival or reproduction....should we expect that G would stay at a constant 55% frequency over time?

Answer: Probably not. In each generation, the kids get a random sample of the alleles from the previous generation. Since there are so few people in each generation, the number of G alleles will vary by chance from one generation to the next, as illustrated below. This random change is genetic drift.

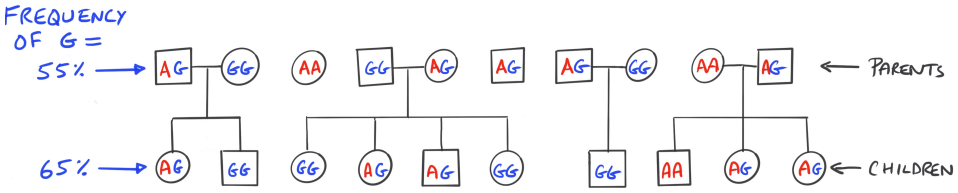

Figure 2.3: **Random sampling of alleles.** *Allele frequencies change from one generation to the next due to random processes: how many children each person has and which alleles they pass on.*

Although most human populations may seem very different from the population of Pitcairn Island, **genetic drift occurs in all populations, though usually much more slowly**.

The Wright-Fisher (WF) model of genetic drift. The Wright Fisher model provides a framework for modeling how allele frequencies change over time [b].

If you wanted to model genetic drift on Pitcairn Island, you might try to get a pedigree for the population over time, and then try to understand how allele frequencies might change through this pedigree. But in practice we don't have pedigrees like this in most populations and, in any event, the complexities of real-world pedigrees tend to obscure the general principles of how frequencies change over time [140].

Instead, the Wright-Fisher model proposes some simplifying assumptions that allow us to understand the fundamentals of population genetics within a very basic model for how allele frequencies change over time. As we will show in the next few chapters, this model is naturally extendable to cover all the other main processes in population genetics, including recombination, natural selection, and population size changes and population structure. While the structure of the model is relatively simple, it provides a powerful framework for understanding genetic variation in real populations.

[b] *The **WF model** [139] is named after two early-20th century founders of population genetics, Sewall Wright and Ronald Fisher.*

We start by assuming a population with N individuals ($2N$ copies of each locus). We assume that there are discrete generations, and that the N individuals mate at random to generate N individuals who form the next generation, ignoring constraints on the sexes of parents [c].

To make the model as simple as possible, you can think of all $2N$ alleles being thrown into a giant bag. Then, we generate the genotypes in the next generation as follows: reach into the bag, draw out an allele at random, write it down, and throw the allele back into the bag. (In a probability class, this process is referred to as *sampling with replacement*.) This is illustrated here:

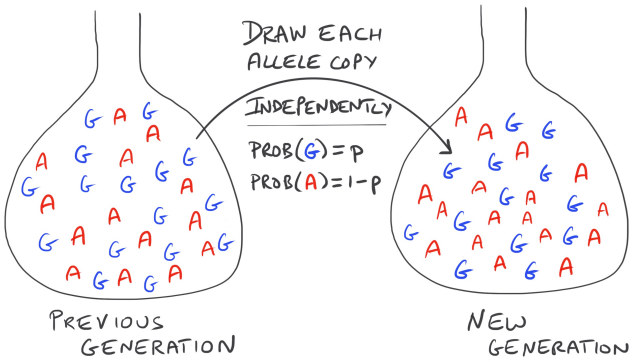

Figure 2.4: **WF sampling.** *Imagine that all* $2N$ *copies of a site are thrown into a big bag (left). We draw alleles out of this bag to make the new generation (right). After we draw out an allele and record it on the right, we drop the original back into the bag on the left. We do this* $2N$ *times to make the new generation.*

This process gives rise to a probability distribution called the **binomial distribution**, which we'll describe shortly.

Before we get to that, here is what this looks in practice. Here I'm assuming a starting allele frequency of $p = 0.55$ as before. Let's suppose that we do a single generation of Wright-Fisher sampling: what is the range of possible outcomes?

The histograms below show the distributions of possible outcomes from repeating this experiment many times in populations of two different sizes.

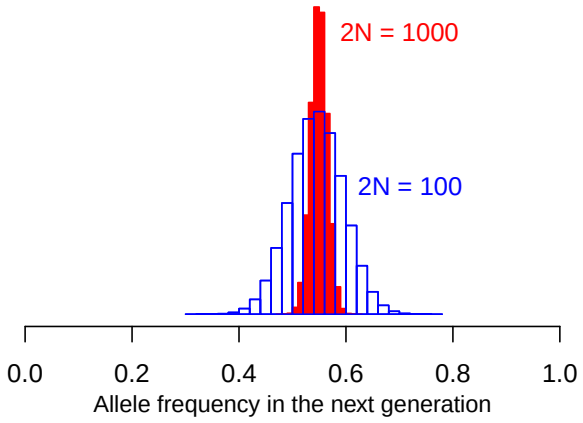

Figure 2.5: **Genetic drift in a single generation.** *Histograms of binomial sampling outcomes for* p_1 *given* $2N$=1000 *(red) vs.* $2N$=100 *(overlaid in blue).*

As you see, both populations are centered on the previous allele frequency (0.55) [d], but the range of outcomes is much wider in the smaller population. This is intuitive, because there is a greater amount of randomness from sampling in the smaller population.

[c] *Here we focus on the numbers of each allele, and ignore the pairing of alleles in diploid genotypes. When we need to think about genotypes in Chapter 2.5, we can predict the proportions using Hardy Weinberg.*

[d] *To be more precise, the* ***expected value*** *of the new distribution equals the frequency in the previous generation. In statistics,* ***an expected value indicates the average (mean) of a distribution****.*

Binomial sampling.

Binomial sampling comes up in many contexts where we make a series of random, independent draws, and each time there is a probability p of one outcome, and $q = 1 - p$ of the other outcome [141].

In WF sampling, we make $2N$ independent draws to create the next generation. Suppose that k is the number of times we draw the derived allele. In this case, the allele frequency after one generation of sampling is $k/2N$, which we will refer to as p_1. The expected value of p_1 (denoted $\mathrm{E}(p_1)$) is simply

$$\mathrm{E}(p_1) = \mathrm{E}(\frac{k}{2N}) = p, \tag{2.2}$$

meaning that *on average* the frequency in the next generation is centered on the current frequency. This doesn't tell the whole story, however, because as shown in the histograms above, the actual, observed p_1 can vary around p. We can measure how much p_1 varies using the **variance** of p_1. By definition, variance measures the average squared difference from the mean:

$$\mathrm{Var}(p_1) = \mathrm{E}[(p_1 - p)^2] \tag{2.3}$$

Using standard properties of the binomial distribution we can show that:

$$\mathrm{Var}(p_1) = \frac{p(1-p)}{2N} \tag{2.4}$$

Notice that the variance in p_1 is inversely proportional to the population size N. This makes sense: a larger population size means that you're getting a bigger sample of the allele frequency from the previous generation. Another important quantity is the **standard deviation** (SD), which is the square root of the variance, namely:

$$\mathrm{SD}(p_1) = \sqrt{\mathrm{Var}[p_1]} \tag{2.5}$$

$$= \sqrt{\frac{p(1-p)}{2N}} \tag{2.6}$$

A very useful rule of thumb is that 95% of the time p_1 will be within two standard deviations of p.

Election polling also follows a binomial distribution. We can get some intuition for binomial sampling by thinking about a completely different context where it comes up: election polling. Suppose that Dumbledore and Voldemort are running against each other for President.

In a particular state, 55% of voters plan to vote for Dumbledore, and 45% for Voldemort. To get a pre-election poll, we phone 100 people, chosen at random, to ask whom they plan to vote for. Assuming that we can get a representative sample of the voting population, the binomial distribution tells us that there is a 95% chance our estimate will be within two standard deviations of the true value: i.e., between 45.1% and 64.9% for Dumbledore [142] [143].

However, suppose that we phone 1,000 people instead of 100. Now we expect a much more accurate estimate: in the range 51.8–58.1%.

These examples illustrate two properties: first, each time we do the survey we get a random estimate centered around the true value. Second, the random error is reduced with a larger sample compared to a smaller sample. Both properties are relevant for allele frequency sampling.

Binomial sampling over successive generations produces genetic drift.

So far, we have talked about genetic drift for a single generation. Now let's think about what happens over the course of many generations.

The crucial thing now is that the result of binomial sampling in one generation gives you the starting point for binomial sampling in the next generation. This will produce a series of allele frequency changes over time called a Markov chain, or more colorfully, a **random walk** [e].

Let's go back to Pitcairn Island. In our hypothetical example, the derived allele started at a frequency p_0=0.55 in the founders (the subscript 0 on p_0 is to indicate that this is generation 0, before any kids have been born on the island). Let's suppose that in the next generation, due to random sampling, the frequency of A goes up to 60% (p_1=0.60). Now, we repeat the random sampling to create generation 2–but this time, the input frequency of A is 0.60, so the expected distribution is centered around 0.60. This process repeats, with the frequency of A drifting up or down by chance depending on the previous frequency. So for example, we might get a sequence of allele frequencies like this:

p_0=.55, p_1=.60, p_2=.57, p_3=.60, p_4=.52, ...

but another SNP with the same starting frequencies might go

p_0=.55, p_1=.45, p_2=.42, p_3=.30, p_4=.35, ...

This is illustrated here:

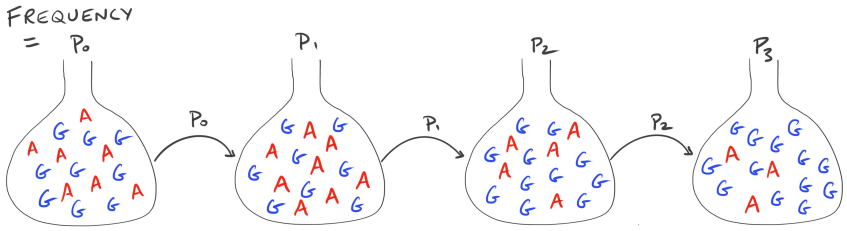

Figure 2.6: **WF sampling over multiple generations.** *The allele frequency in each generation is a binomial sample centered on the allele frequency in the previous generation. Over many generations, this randomness allows the frequencies to drift away from the initial starting point.*

Drift works the same way in larger populations, but the *rate* of drift is slower, simply because the binomial variation is smaller in each generation. The next plots show simulations of this process in populations of different sizes. Each line plots an independent random outcome:

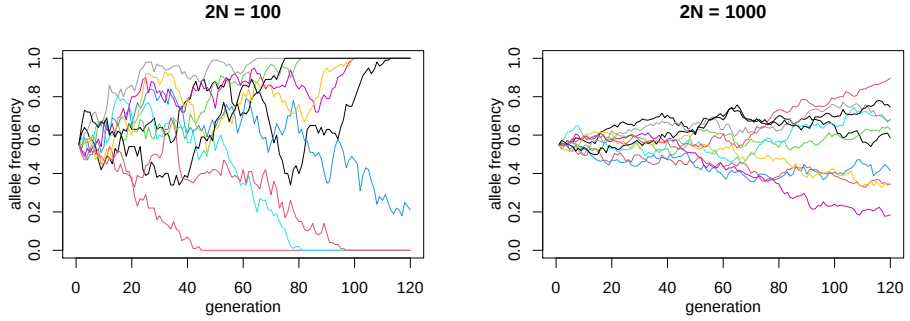

Figure 2.7: **Simulations of genetic drift** *from a starting allele frequency of 0.55. Each plot shows ten independent simulations. Notice that the range of possible outcomes diverges much faster in the smaller population size, with some simulations reaching fixation (frequency=1) or loss (frequency=0) within the timescale of the simulation.*

Eventually, the G allele will either reach 100% frequency, in which case we say that it has *fixed*, or 0% in which case we say it has been *lost*. In random walk theory, 0 and 1 are referred to as *absorbing states*: meaning that the random walk ends if it reaches those values.

[e] *Genetic drift is an example of a mathematical model known as a random walk. You can imagine a drunkard stumbling backwards and forwards along a number line with walls at 0 and 1 until he bumps into either wall and stops.*

It's outside the scope of this book, but infinite random walks in 2 and 3 dimensions have very interesting properties. I have always enjoyed the aphorism from mathematician Shizuo Kakutani that "A drunk man will find his way home, but a drunk bird may get lost forever." [Link].

Mutation and drift. So far we have been talking about drift of alleles that are already common. But in practice, each SNP starts life as a new mutation. Some mutations drift up to become common. How can we model this?

We'll assume that each mutation creates a new allele that did not exist previously in the population. This is known as an **infinite sites** assumption (this simplifies the math and is usually a good approximation [144]). Under this assumption, each new mutation has a starting allele frequency of one copy in the population: i.e., $p_0 = 1/2N$. The new allele now drifts until it either reaches loss or fixation:

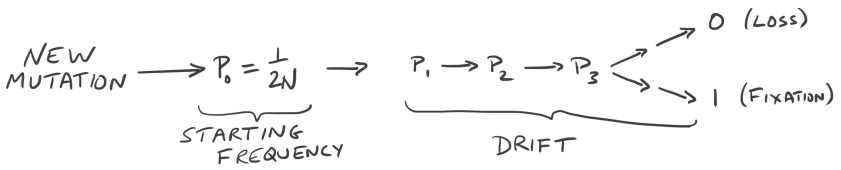

Figure 2.8: **The life-cycle of a SNP.** *A mutation generates a new variant. This is initially at frequency $1/2N$. Its frequency drifts until it eventually reaches loss or fixation.*

The next figure illustrates this process for 200 mutations introduced at different times in a population of 100 individuals. As you can see, most of the mutations stay rare and are quickly lost; however a few drift up to become more common and, in this example, one eventually reaches fixation.

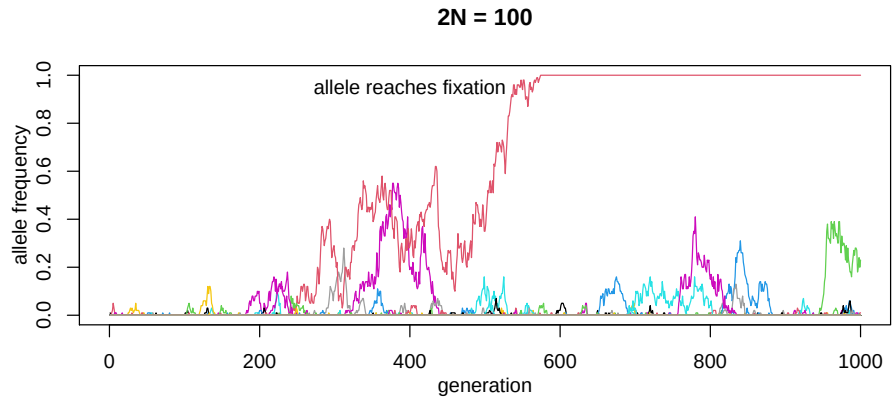

Figure 2.9: **Genetic drift of new mutations.** *Each line shows the simulated trajectory of a different mutation, starting at a random generation number, and drifting independently of the other mutations. This simulation included 200 mutations, most of which stayed rare and are hard to see on this plot.*

Most alleles in a population are very rare. Every new mutation starts out rare in the population (at a frequency of $1/2N$), and most are quickly lost, while a very small fraction drift up to become common. You can see this in the simulation plot above, where only a few of the 200 mutations drifted above 10% frequency [f].

What is the probability that a new mutation reaches fixation?

One very useful fact that we'll derive in the next chapter is that **the probability that a derived allele currently at frequency p will eventually fix is also p. For example, the probability that a new mutation will eventually fix is 1/2N.** Since human populations number in the tens-of-thousands to millions, the fraction of new mutations that eventually fix is very very small.

[f] *You can think about new mutations within our casino analogy. Think about what happens when a crowd of punters all walk into a casino with one dollar each and start gambling – most never get much money and go broke very quickly, but a very few lucky players build up a sizable purse and play for a while before going broke.*

Mutation, drift and the amount of genetic variation. If we put these concepts together, we're now ready to think about genetic variation in populations.

In Chapter 1.3 we discussed how to quantify the *amount* of genetic variation in different populations. One important measure of genetic diversity is **expected heterozygosity**. We define this as follows. Suppose that you sequence a genomic region on one homolog of one random individual, and on one homolog of a different individual. Expected heterozygosity is the average fraction of sites at which these two haploid sequences differ.

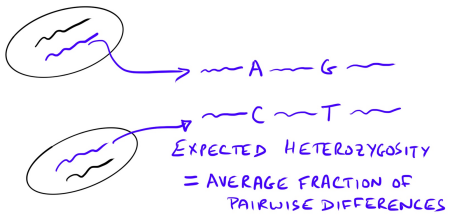

Figure 2.10: **Expected heterozygosity** *is the average fraction of sequence differences between a random pair of allele copies.*

In modern human populations, **expected heterozygosity is ∼0.5–1 heterozygous sites per kilobase**, depending on the population. (See Table 1.2, Chapter 1.3.)

What determines expected heterozygosity? First, mutation plays a critical role of creating new variation in the population. Secondly, the average effect of drift is to remove variation. (Of course, drift sometimes allows rare alleles to become common, but this is always transitory, and in the absence of new mutations, all variants eventually drift to fixation or loss.)

Thus, we can **understand expected heterozygosity as a balance between two forces: mutation, which inputs new variation, and drift, which tends to remove it.** The next box shows a derivation of expected heterozygosity under the WF model. You can skip this if you prefer.

Optional derivation: Computing expected heterozygosity.

Imagine picking two random copies of a locus from the population. We want to write down how the probability that a single nucleotide is heterozygous in the next generation depends on population size and mutation. We set H_1 to be the initial heterozygosity, and H_2 the heterozygosity in the next generation. We also define μ (pronounced mu) as the mutation rate per base pair per generation.

Imagine we pick two random alleles in generation 2. Under our model, there is a probability $1/2N$ that they are descended from the same parent allele in generation 1. This has an effect of reducing heterozygosity by a factor $1/2N$.

On the other hand, alleles that were identical in the parents (with probability $1 - H_1$) might have mutated in either parent: i.e., with probability $\sim 2\mu$. (For simplicity I'm ignoring some extra terms that relate to rare double events, such as two mutations or both mutation AND inbreeding; I'm also making a standard simplifying assumption that mutations always create new alleles.)

We can now write a simple recursion for the expected heterozygosity in generation 2, given what it is in generation 1:

$$H_2 = \underbrace{H_1 \times (1 - \frac{1}{2N})}_{\text{Het goes down by } 1/2N} + \underbrace{(1 - H_1) \times 2\mu}_{\text{Het goes up due to mutation}} . \quad (2.7)$$

This last equation tells us how heterozygosity changes from one generation to the next. Let's suppose we're at a steady state between loss of heterozygosity (from drift) and gain of heterozygosity (from mutation). In that case, we can consider an equilibrium value H that is the same on the left and right hand

sides of the equation, and solve for this:

$$H = H \times (1 - \frac{1}{2N}) \; + \; (1 - H) \times 2\mu \tag{2.8}$$

After some algebraic rearrangement we get

$$H = \frac{4N\mu}{1 + 4N\mu} \tag{2.9}$$

Since $4N\mu$ is usually very small ($\sim$0.1% in humans) it's customary to simplify this last expression to

$$H \; \approx \; 4N\mu \tag{2.10}$$

which matches the value we will derive in the next chapter using a very different technique called the coalescent.

To summarize the math, we just showed that the expected heterozygosity is $4N\mu$. In other words, heterozygosity is proportional to both population size N (because larger population size lowers the rate at which alleles are lost to drift) and mutation rate μ (because higher mutation rate increases the influx of new variation).

Moreover, it turns out that $4N\mu$ is a fundamental parameter in population genetics, that controls the amount of neutral genetic variation. We won't use this notation here, but it's so fundamental that it's sometimes given a special name, θ. We'll come back to interpreting $4N\mu$ on the next page, but we have to introduce *effective population size* first.

Effective population size. It's time for me to confess that the binomial sampling model requires several assumptions that you might think are silly: for example the population size is constant over time; mating occurs completely at random; we don't differentiate between males and females; generations don't overlap; and that there is no inherent tendency for some individuals to have more kids than others (in technical terms, we say that the number of kids is Poisson-distributed).

But it turns out that the simple model actually works well in place of more complicated scenarios, provided that we use a fudge factor called **effective population size** (N_e). Basically the idea is that we will use this number N_e in place of N, where N_e reflects the actual rate of drift under a more realistic scenario.

For example, if a few males have very many offspring, as happens in some species (and in historical examples like Genghis Khan and his sons who fathered huge numbers of offspring across central Asia [145]), this can greatly reduce N_e. Similarly, when the population size fluctuates over time, N_e is strongly shifted toward the times when the population size is smallest, because drift happens much faster in those generations [146] [g].

In future, when we're talking about theoretical models, it's most conve-

Figure 2.11: **Male elephant seals fighting for dominance.** *The winners can control harems of up to 50 females. High variation in reproductive success reduces N_e relative to the actual population size N.* *Credit: Hullwarren CC BY-SA 3.0 [Link]*

[g] *Most of the ways in which real populations differ from the idealized WF model tend to reduce N_e.*

nient to describe the models in terms of N, for idealized populations. But when we talk about data from real populations, we almost always need to think about the data in terms of N_e instead of N. As we'll see below, N_e is usually much smaller than a census estimate of the population size.

Estimating N_e from data. Remember from above that in the idealized model, the expected heterozygosity per site is given by $4N\mu$. When we look at real populations, we replace this with $4N_e\mu$ to reflect that the actual rate of drift is controlled by *effective* population size.

As we noted above, the effective heterozygosity in humans ranges from 5×10^{-4} to 1×10^{-3} per base pair, depending on the population. The mutation rate is about 1.3×10^{-8} per base pair, per generation. If we use N_e in place of N then for the higher end of this range we have

$$4N_e\mu = 1 \times 10^{-3} \tag{2.11}$$

$$N_e \approx 20,000 \tag{2.12}$$

and $N_e \approx 10,000$ for populations at the lower end of the range [147]. In summary, the long-term effective population sizes of humans are around 10,000–20,000 individuals.

These estimates may seem absurdly low, given that the current world population is almost 8 billion. In part N_e is so small because it's a type of average [148] over roughly the last million years and the human population was far smaller for most of that time than it is now; the effective size is also made smaller by the various other ways that real human populations differ from the ideal model. But it's difficult to fully interpret exactly why effective population sizes are what they are [149]. Nonetheless, N_e provides a powerful tool for modeling patterns of genetic variation, especially if we allow it to vary over time–as we will in the next chapter.

The WF model with haplotypes. So far we have been discussing mutation and drift for individual SNPs. But of course, each SNP is contained within a DNA sequence, which may contain multiple variant sites (also known as a *haplotype* [h]). How should we think about mutation and drift in the context of haplotypes?

[h] Genetic variation is inherited within **haplotypes** and we'll talk a lot about these in later chapters.

Here we're going to introduce a basic haplotype model. Importantly, this type of model can be extended in many ways – for example with recombination, selection, or population structure – and we'll use this basic model as a scaffold again in later chapters.

First, let's assume we want to model mutation and drift for a genomic region of L basepairs in length. Now, each generation will comprise two steps: (1) Mutations can arise anywhere in the sequence, at a rate μ per base pair, per haploid sequence. (2) In the sampling process, each haploid sequence in the next generation is drawn at random from the previous generation, sampling with replacement. (Similar to before, this is like putting all $2N$ haplotypes in a bag, and drawing out the next generation one at a time, always writing down the new haplotype, and throwing the old one back in the bag.) This is shown here:

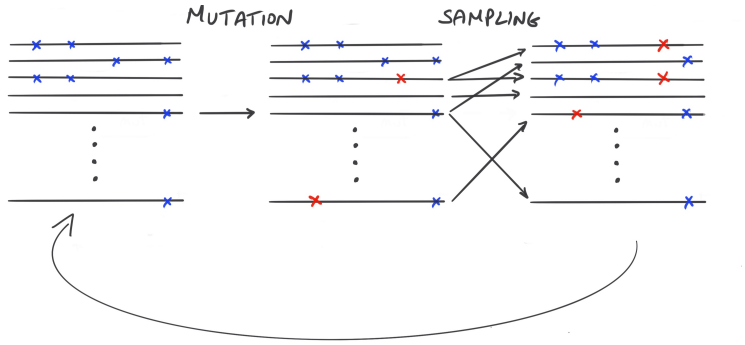

Figure 2.12: **A WF model for haplotypes.** *Each line is a haploid sequence (there are 2N of these to represent the entire population). Blue crosses indicate derived alleles at SNPs. New mutations (red crosses) are placed at random locations within the sequence. WF sampling acts on the haplotypes instead of on alleles.*

WF simulation of haplotype variation. One powerful feature of the WF model is that we can use it to simulate data under a wide variety of evolutionary models [i].

To close the chapter, we'll take the intuition suggested above, and turn this into pseudocode (a kind of recipe) that we could use to simulate haplotypes. If you have experience with programming you could try this out. Even if you don't have experience with programming, think about the steps here, and how they relate to the underlying model.

One key idea here is that we will start with an arbitrary genotype matrix, and then iterate through many generations of mutation and WF sampling until the simulation reaches equilibrium levels of genetic variation (and the starting point doesn't matter any more). For reasons we'll cover in the next chapter this takes at least about $4N$ generations [150]. The genotype matrix at the end of the simulation is a random draw from this mutation-drift equilibrium.

Here's some basic pseudocode for a Wright Fisher model with mutation:

- *Genotype matrix:* Create a genotype matrix G, that contains $2N$ rows (each row is a haplotype) and L columns (each column is a site in the sequence). We'll designate four possible nucleotides using the integers 0, 1, 2, 3 [151].
- *Initialization:* Set every entry in the genotype matrix G to 0.
- *for* `generation` *in* 1 *to* `Max-Generations` *do:*
  {
  - *Mutation:* For each site in each row of G, mutate the existing allele with probability μ.
  - *WF sampling:* Create a new temporary genotype matrix, named G'. For each row of G', pick a random integer u between 1 and $2N$, inclusive. Copy row u from G into G' (this simulates WF sampling with replacement). When all $2N$ entries of G' are filled, copy G' back into G before starting the next generation.

  }

[i] *This approach to simulation is called* ***forward simulation*** *to distinguish it from the backward-in-time approaches we will encounter in the next chapter.*

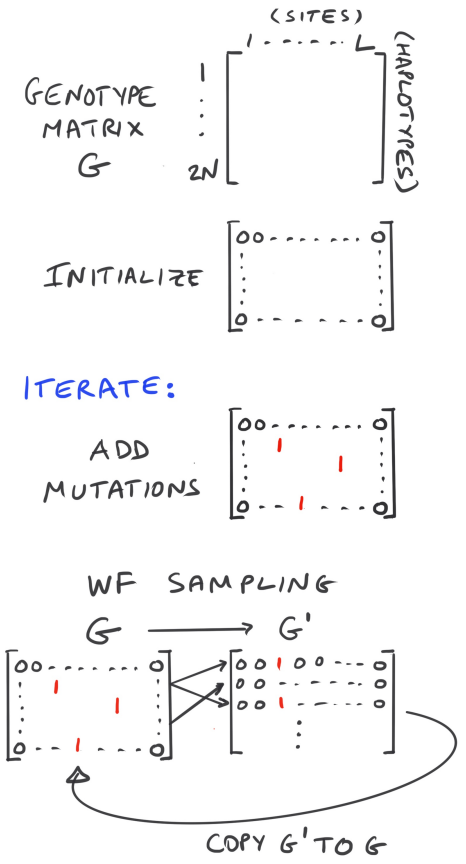

Figure 2.13: **Illustration of WF pseudocode.**

Possible improvements [Optional]. The pseudocode above is ok, but it's slower than necessary and it wastes memory because (i) many haplotypes are identical, and (ii) most sites are not variable. We could make this much more efficient if we just keep track of the distinct haplotypes, and how many there are of each haplotype. We also don't need to store all the nonvariable sites – instead we can just store the positions of derived variants. Lastly, when we add mutations, we can generate the total number of new mutations each generation using a single Poisson random variable with mean $2N\mu L$, and then modify the existing haplotypes at random positions [152].

Simulation software. Forward simulations provide a flexible approach for modeling population genetic data, and can be applied to a wide range of possible models. They are usually computationally slower than backward simulations (next chapter) but aside from very simple models they are easier to implement and far more flexible. A popular software package called **SLiM** provides a powerful toolkit for simulating a wide range of interesting models [Link] [153].

In this chapter we introduced the Wright Fisher model and the fundamental concept of genetic drift (and the interplay of mutation and drift). Nearly everything else in population genetics depends on the basic processes of mutation and drift. In the next chapter, we'll introduce the coalescent, which gives us a very different way to understand drift.

Notes and References.

[136]In practice the size of your cash holdings over time when gambling in a casino is more analogous to the drift of a deleterious variant, since casino betting is set up to favor the house. We'll describe drift of deleterious alleles in Chapter 2.5.

[137]The counts would be different for sex chromosomes: there are $N/2$ Y chromosomes, and $3N/2$ X chromosomes, assuming equal numbers of males and females.

[138]You can read more about Pitcairn Islands here: [Link] and specifically about the mutiny here [Link]. The peak population size was 250 inhabitants in 1936.

Another example of an extremely isolated population is Tristan da Cunha. This is a tiny island in the south Atlantic– at 1700 miles west of Cape Town in South Africa it is the most remote inhabited island in the world. Tristan da Cunha is currently home to about 270 people who descend mainly from 8 men and 7 women from Europe and the US who settled the island in 1816:

Soodyall H, Nebel A, Morar B, Jenkins T. Genealogy and genes: tracing the founding fathers of Tristan da Cunha. European Journal of Human Genetics. 2003;11(9):705-9

[139]Sewall Wright, RA Fisher, and a third scientist JBS Haldane, are often credited as developing many of the key ideas of modern population genetics, mainly in the first half of the 20th Century. This formed a key component of the so-called Modern Synthesis, which united Darwin's theory of evolution with the growing understanding of heredity started by Mendel.

[140]It's outside our scope here, but techniques for studying frequency changes in known pedigrees are referred to as *gene dropping*. For an excellent example see

Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, et al. Allele frequency dynamics in a pedigreed natural population. Proceedings of the National Academy of Sciences. 2019;116(6):2158-64

[141]Binomial sampling. The probability of getting k successes is

$$\frac{n!}{k!(n-k)!}p^k q^{n-k}, \tag{2.13}$$

where the function $n!$ is pronounced "n factorial" and calculated as $n \times (n-1) \times (n-2) \times \cdots 3 \times 2$. For more on the binomial see [Link].

[142]Here we approximate the sampling distribution as binomial, assuming that the size of the poll is much smaller than the number of voters. The standard deviation of the binomial proportion is $\sqrt{p(1-p)/n}$ where p is the true proportion and n is the number of voters that we phoned (instead of $2N$ for number of allele). The true estimate will lie within $+/-$ two standard deviations about 95% of the time.

[143]These example are meant as illustrations, but in practice, the biggest challenge in election polling is not binomial sampling error but getting a representative sample of the voting population. In particular, it may be more difficult to reach some types of likely voters than others. For this reason, analysis of polling data usually involves techniques to reweight the samples to better reflect the expected demographic and political composition of likely voters.

[144]Remember that only about 0.1% of sites are common SNPs so this is a very useful approximation for most applications within species. However the assumption breaks down in analyses of very large sample sizes, especially at hypermutable CpG sites. It also doesn't work well for phylogenetic models of distantly related species as over longer timescales a larger fraction of the sites have accumulated substitutions.

Harpak A, Bhaskar A, Pritchard JK. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genetics. 2016;12(12):e1006489.

[145]About 8% of the men in central Asia carry a single Y chromosome haplotype that is estimated to descend from a common ancestral haplotype 1000 years ago. The age and geographic distribution of the haplotype suggest that it was likely spread by Genghis Khan and his male relatives:

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The genetic legacy of the Mongols. The American Journal of Human Genetics. 2003;72(3):717-21

Balaresque P, Poulet N, Cussat-Blanc S, Gerard P, Quintana-Murci L, Heyer E, et al. Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. European Journal of Human Genetics. 2015;23(10):1413-22

[146]When population size fluctuates rapidly over generations, the effective population size is given by the harmonic mean. Long-term changes in N are less-well modeled by a simple change in N_e.

[147]I'm rounding here since all the other numbers are somewhat rounded (and in any event heterozygosity varies across the genome and across populations). Given these particular numbers, the precise value of N_e would be 19,230.

[148]The harmonic mean.

[149]It's difficult to fully interpret effective population size estimates. Humans have extremely low heterozygosity (and hence N_e) compared to a wide range of other species. Although chimpanzees and gorillas now have very small populations, they actually have higher long-term N_e than humans. Meanwhile, Neanderthals were even less diverse than modern humans, as are a few contemporary species with very small populations, such as lynx and wolverines. Although N_e can be difficult to interpret, it still provides a powerful tool for modeling patterns of genetic variation, especially if we allow N_e to vary over time as is typical in more advanced models.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLOS Biology. 2012;10(9):e1001388

[150]We want to run the simulation long enough to ensure that the simulation can reach a stationary distribution with respect to the amount of genetic variation (and so the starting point is no longer relevant). One way to think about this is that the population MRCA in the final generation (see the next chapter) should exist within the simulation. On average, the time to the MRCA is $4N$ generations, so we would want to run this for at least $4N$, and probably more like $10N$ generations to be safe.

[151]The way I'm writing this it's actually finite sites mutation, instead of the infinite sites model alluded to earlier. The finite sites model is a bit more intuitive here.

[152]We can also convert this into an infinite sites model by representing the mutated position using a real number on the interval [0,1]. Derived alleles will be represented by 1.

[153]Messer PW. SLiM: simulating evolution with selection and linkage. Genetics. 2013;194(4):1037-9

Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Molecular Biology and Evolution. 2019;36(3):632-7