2.5 Natural selection: I. Background and models

At its most fundamental level, evolution proceeds through changes in allele frequencies over time. In the next three chapters we will discuss the role of natural selection in shaping genetic variation. This chapter describes basic models of population genetics with selection.

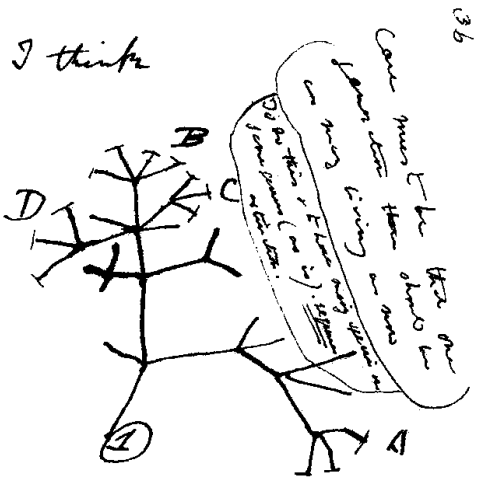

Figure 2.77: **Charles Darwin sketched this evolutionary tree** *in his notebook in 1837, to describe his monumental insight that species evolve from common ancestors.* [Link] *Public Domain.*

Evolution, adaptation, and the modern synthesis. **Charles Darwin**'s 1859 book *On the Origin of Species by Means of Natural Selection* launched a major revolution in the history of science. Darwin articulated two important principles:

(1) that different species evolve from common ancestors, a process that Darwin referred to as "descent with modification"; and

(2) that natural selection and the "struggle for life" provides a driving force for how species change and adapt over time.

These ideas are the fundamental organizing principles of biology: we can understand the similarities and differences among species in terms of the fact that species are descended from common ancestors, while at the same time, their traits evolve over time according to the principles of natural selection [a].

In Darwin's formulation of natural selection (also developed independently by his contemporary Alfred Russell Wallace), populations can adapt over time provided that three conditions are met:

(1) **Variation.** Individuals vary in their phenotypes;

(2) **Inheritance.** The phenotypes are at least partially inherited: i.e., children tend to resemble their parents;

(3) **Competition.** Not all individuals survive or reproduce equally; survival and/or reproductive success depend in part on phenotype;

Under these conditions, the traits that increase the probability of survival or reproduction tend to increase in frequency in the population.

In modern terms, we would say that if there is selection on certain phenotypes, and these phenotypes are (at least partly) controlled by genetic variation, then the genetic variants associated with the preferred phenotypes tend to increase in frequency [b].

Darwin amassed a wealth of evidence for his theory, drawing on natural history, paleontology, biogeography, and other fields. However a crucial gap was that the mechanism of inheritance – i.e., genetics – was not understood at all. At that time, the prevailing model of inheritance was known as "blending inheritance", namely that children represent

[a] *In 1973 the evolutionary biologist Theodosius Dobzhansky famously wrote that "Nothing in Biology Makes Sense Except in the Light of Evolution".*

[b] *Although we usually think of natural selection acting on phenotypes and genotypes, these same principles can act in other domains. In his 1976 book "The Selfish Gene", Richard Dawkins talked about the idea that the principles of natural selection can help to evolve, and spread, ideas in social networks. This idea has become increasingly relevant; you are surely familiar with the term he coined to describe this: "meme".*

some kind of blending, or averaging, of the characteristics of their parents. Blending inheritance would imply a steady loss of phenotypic variation over time, which would be seriously problematic for Darwin's theory since the theory requires the presence of heritable variation. Darwin recognized this as an important gap in his argument, and even endorsed an incorrect alternative model of inheritance called "pangenesis", in which parts of the body emitted particles called gemules that collected in the gonads.

The irony is that, unbeknownst to Charles Darwin, at the same exact time Gregor Mendel was working in Brno (now in the Czech Republic) on the experiments that would lead to Mendelian genetics. His experiments on peas, conducted between 1856 and 1863, showed that genetic information is inherited as discrete packets (i.e., alleles) rather than being blended. In contrast to blending inheritance, Mendelian inheritance means that allelic variation–and hence phenotypic variation–is transmitted from one generation to the next. This insight immediately rescues the Darwinian model. However, Mendel's findings were published in 1866 in an obscure natural history journal published in Brno (*Verhandlungen des naturforschenden Vereines in Brünn*), and were not widely known until the paper was rediscovered in 1900–long after both Darwin and Mendel were dead.

After the rediscovery of Mendel's work, there was a blossoming of genetics in the first half of the 20th Century including, for the first time, a clear understanding of alleles and transmission, a chromosomal theory of inheritance, and some understanding of the connections from genotype to phenotype. Most of the fundamental models of population genetics, including Hardy-Weinberg, the Wright-Fisher model, the basic models of natural selection that we will cover in this chapter, and quantitative genetic models of inheritance that we cover later, all date to this period [c]. This work joining together population genetics with Darwinian evolution in the early-to-mid 20th Century is referred to as the **Modern Synthesis**, and nowadays population genetics and molecular genetics are central pillars of evolutionary biology.

One key insight of the Modern Synthesis is that **evolution results from population genetic processes, played out over long timescales**. In population genetics, we study the forces that change allele frequencies or haplotype frequencies from one generation to the next; accumulated over hundreds, thousands or millions of years this results in adaptive changes, speciation, and everything else in evolutionary biology.

In these next three chapters, we will cover a modern understanding of how natural selection plays out in population genetics, using both theory and examples.

Fitness. In past chapters, our models have assumed that survival and reproduction is independent of genotype. But of course some genotypes do affect the ability of an individual to survive to adulthood, or to reproduce successfully.

Figure 2.78: **Fossil mosquito infected with the malaria plasmodium**, *preserved in amber. Malaria has been a major selective pressure in human history.* *Credit: George Poinar, Jr., [Link] CC BY-SA 2.0.*

[c] *It's striking that most fundamental principles of population genetics can be traced to this period when there was only a rudimentary understanding of genetics, and the molecular details were unknown. In contrast, coalescent theory came rather later (early 1980s), partly stimulated by the emergence of molecular data. The 21st Century has seen huge advances in statistical and computational techniques and the interpretation of modern data.*

To model this, we introduce the concept of **fitness**. Consider an individual at a certain point in the life-cycle (e.g., a newly fertilized egg), with genotype x at a certain variant or set of variants. We define the fitness of genotype x as the expected number of offspring, precisely one generation later, that descend from this individual. In other words, fitness measures the ability of genotype x to survive to reproductive age, to attract mates, and to reproduce successfully through one full turn of the life-cycle.

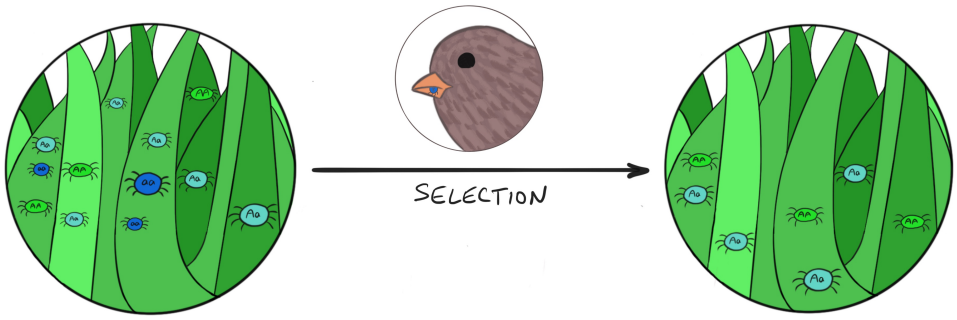

Figure 2.79: **Natural selection and fitness.** *Here, spiders with the aa genotype are blue and stand out from their background; as such they are more likely to be eaten by birds. Hence aa individuals have low fitness, lowering the frequency of the a allele among individuals at reproductive age.* Credit: Lucy Pritchard.

Notice that fitness is defined as an *expected* outcome – importantly, you can *think of fitness as the expected reproductive output for an individual with this genotype, averaging over the possible environments they may experience*, averaging over possible genotypes elsewhere in the genome, and averaging over the good or bad luck experienced by individuals of this genotype throughout their lives: what Hamlet called the "slings and arrows of outrageous fortune".

A basic fitness model. We're now ready to introduce a basic model of selection. We consider a single nucleotide position, with an ancestral allele, A, and a derived allele a.

We model the **relative fitness** of each genotype as follows. AA acts a reference group, defined to have fitness 1, and we measure the fitness of the other genotypes *relative* to that reference [235] [236]:

$$\begin{aligned} \text{Fitness of AA} &= 1 \\ \text{Fitness of Aa} &= 1 + hs \\ \text{Fitness of aa} &= 1 + s \end{aligned} \tag{2.61}$$

Here, s is referred to as the **selection coefficient**, and h is the **dominance coefficient**:
- If s is positive $(s > 0)$ then the derived allele is **advantageous**
- If s is zero then the derived allele is **neutral**
- If s is negative $(s < 0)$ then the derived allele is **deleterious**

Reflecting the sign of s, selection in favor of an advantageous allele is also referred to as **positive selection**; selection against a deleterious allele is **negative selection.**

To give you a sense of scale, the most strongly advantageous derived alleles in humans may have s of up to $\sim$3%. But there are many more ways to break genomes than to improve them: the effects of deleterious

variants can range from just very slightly negative, all the way down to $s = -1$ (which would imply that the derived allele is incompatible with life or reproduction).

Meanwhile, h measures the relative fitness of heterozygotes, and is known as the **dominance coefficient**. If the derived allele a is fully recessive then $h = 0$; and if a is fully dominant then $h = 1$. In rare cases h can be outside the range $[0, 1]$ leading to a special form of selection called balancing selection. Except where stated the figures below assume what is known as an **additive model (h = 0.5)**.

Frequency changes over time. How does selection change allele frequencies and genotype frequencies over time? We'll set p to be the current derived allele frequency, and $q = 1 - p$ as the ancestral frequency.

Genotype frequencies. Before selection the genotype frequencies are given by Hardy Weinberg proportions. The effect of selection is to change the genotype frequencies in proportion to their fitnesses.

For example, the frequency of the *aa* homozygote is p^2 before selection, and proportional to $p^2(1 + s)$ after selection:

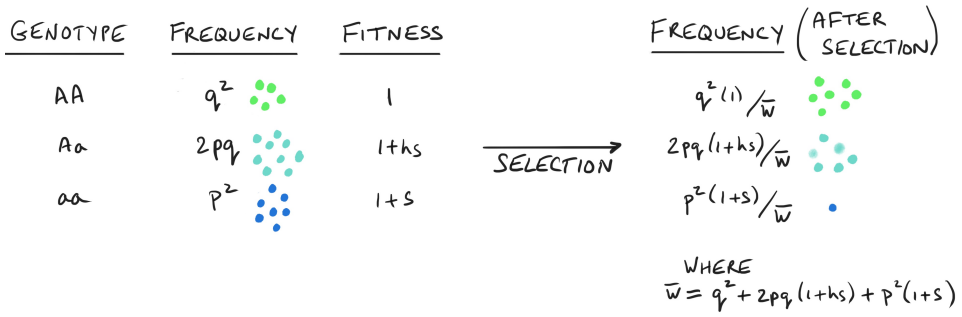

Figure 2.80: **Changes in genotype frequencies due to selection.** *Before selection the genotype frequencies are given by Hardy Weinberg proportions. After selection, the frequencies are multiplied by the genotype fitnesses – illustrated here for $s < 0$. The factor of $\overline{w}$ is used so that the genotype frequencies add to 1.*

By definition, frequencies have to add up to 1, so each of the terms above is divided by the total, a quantity known as the mean fitness:

$$\overline{w} = q^2 \cdot 1 + 2pq \cdot (1 + hs) + p^2 \cdot (1 + s). \tag{2.62}$$

Dividing by $\overline{w}$ simply rescales the frequencies to sum to 1.

Allele frequencies. And what is the expected frequency of a in the next generation? (We'll call this p'.) To get this we add together half the frequency of heterozygotes plus the frequency of *aa* homozygotes:

$$\mathrm{E}[p'] = \frac{pq(1 + hs) + p^2(1 + s)}{\overline{w}} \tag{2.63}$$

This expression isn't particularly illuminating, but we get something more useful if we look at the *change* in allele frequency, Δ_p from one generation to the next:

$$\Delta_p = \mathrm{E}[p'] - p. \tag{2.64}$$

Δ_p tells us whether p is increasing or decreasing over time (depending on whether Δ_p is positive or negative). After a small flurry of algebra [237], we find that

$$\Delta_p = \frac{pqs[p(1-h)+qh]}{\overline{w}}. \tag{2.65}$$

This expression is easier to interpret:

- When $p = 0$ or $q = 0$ there is no allele frequency change. That makes sense, because there's no variation for selection to act on.
- If $s = 0$ there's no selection, and no expected change in allele frequency.
- Third, and most important, if p lies between 0 and 1 we have the intuitive result that *if s is positive, then Δ_p is positive, meaning that the derived allele is favored, and tends to increase in frequency; if s is negative, the derived allele is disfavored and tends to decrease* [238].

What happens over multiple generations? We can iterate Equation 2.65 over multiple generations to predict the trajectory of a selected allele over time. This is known as a **deterministic** model, meaning that it assumes the trajectory of an allele is completely determined by the expectation. As you can see, selection drives favored alleles up towards fixation, and deleterious alleles to loss. The process in which favored alleles are pushed up to fixation is called a **selective sweep**.

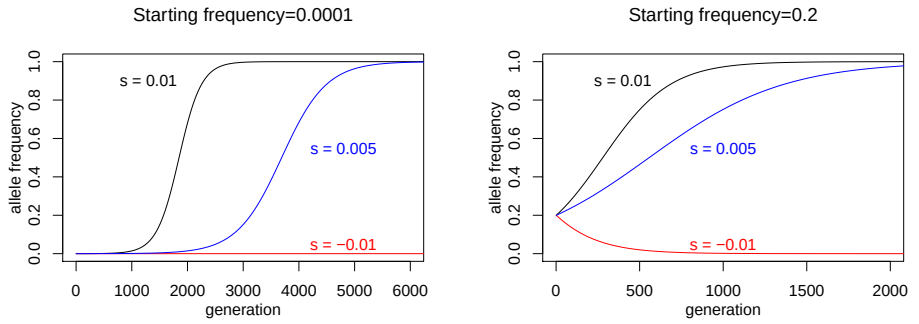

Figure 2.81: **Allele frequency trajectories of selected alleles, over time**. *The blue and black lines show frequency increases of advantageous alleles. The right-hand plot assumes an unreasonably high starting frequency of 0.2 to illustrate that selection drives deleterious alleles (in red) to low frequencies.*

The deterministic model is helpful for understanding the overall process, but it's also important to consider the random effects introduced by drift.

Frequency changes with selection and drift. In Chapter 2.1 I suggested that genetic drift of a new mutation is like a player's winnings over time in a casino. Even for advantageous alleles, the effects of random sampling are extremely important.

Suppose you walk into a casino to play Blackjack, and you play until you either go bust, or beat the house.

For Blackjack, assuming optimal play, players usually have an inherent disadvantage of 0.5–1.0% relative to the casino (the precise value depends

on the house rules). However, there are card-counting strategies that can potentially tilt the odds by 1–2% back towards the player, turning a small player disadvantage into a small player advantage (although these are frowned upon by the casinos [239]). You can think of the default Blackjack game as like selection on a mildly deleterious mutation ($s < 0$), and the game with card counting as like a mildly advantageous mutation ($s > 0$).

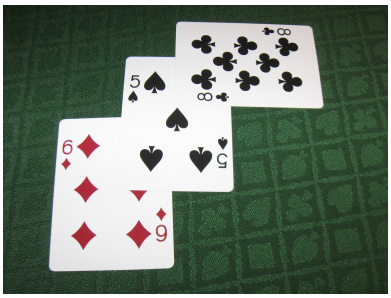

Figure 2.82: **Blackjack cards**. *Credit: Scott5114 [Link] Public Domain*

There are two key points here: First, with a small starting purse, you're likely to go bust quickly, regardless of whether you count cards or not. This is simply because, with small numbers, you're likely to have at least some bad luck that bankrupts you. This reflects the great importance of chance when you're working with small numbers.

But if you get lucky early on, then the power of large numbers starts to take over, and you can start to use a deterministic model to predict how your purse will grow – at least until the casino tosses you out!

Selection in the WF model. The same fundamental processes affect new mutations. So far we have considered the Wright Fisher model for neutral alleles, but it's easy to extend it to allow biased sampling due to selection.

Under the neutral model, if the current allele frequency is p in a population of size $2N$, then the allele frequency in the next generation would be

$$p' \sim \text{Binomial}(p, 2N) \quad [\text{neutral model}] \tag{2.66}$$

With selection, the allele frequency in the next generation is similar but centered on the *expected allele frequency with selection*, $\text{E}(p')$, as given by Equation 2.63:

$$p' \sim \text{Binomial}(\text{E}(p'), 2N) \quad [\text{with selection}] \tag{2.67}$$

Here's what this looks like, for one generation of sampling with relatively strong selection: $s = 0.01$. (A 1% selective advantage may not seem like much, but as we'll discuss shortly there are very few individual changes to the genome that can improve fitness by this much.)

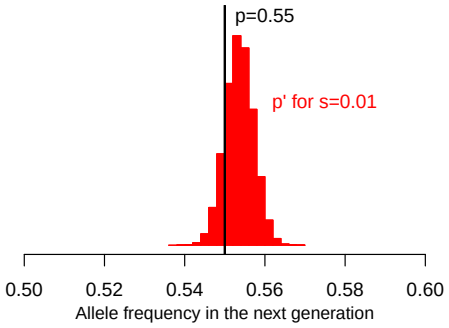

Figure 2.83: **Histogram of binomial sampling outcomes** *(p') after one generation of selection and drift with $s = 0.01$ and $2N = 20,000$. The starting allele frequency $p = 0.55$. (Compare to the neutral case, Figure 2.5.)*

As you can see, the overall distribution of outcomes (in red) is shifted toward higher frequencies than the initial frequency p. However, due to the random sampling process, there is variation in the resulting allele frequency, and even a chance that the frequency actually decreases, despite the upward selection pressure.

Now, let's look at **selection and random drift** together, over multiple generations. The next figure shows simulated trajectories from a starting frequency of 0.5, for a range of different selection coefficients. In the left panel, selection is strong enough that drift has little impact on the selected alleles (blue and red curves). In contrast, in the right panel, selection is just 1/10th as strong, and while the blue curves tend to be higher than the red curves on average, the randomness of drift means that some favored alleles (blue) fare worse than some deleterious alleles (red):

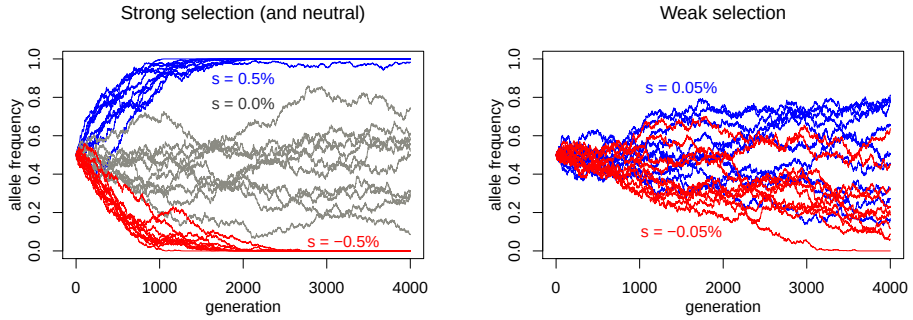

Figure 2.84: **Selection and drift of alleles from starting frequencies of 0.5** *in a population of* $N = 10^4$. *The left panel shows simulated trajectories for relatively strong selection* ($2Ns = 100$ *in blue, and* $2Ns = -100$ *in red); and neutral in gray for comparison. The right panel shows weaker selection* ($2Ns = 10$ *in blue, and* $2Ns = -10$ *in red).*

In the plots above, the directional effect of selection has to fight against the randomness imposed by genetic drift in a finite population. *When selection is strong enough, it overwhelms drift, and allele frequency curves are close to the deterministic trajectory. But when selection is weak, or the population is small, drift can effectively overwhelm selection.*

To quantify this, *a widely-used rule of thumb is that when* $2Ns$ *is in the range of about* -1 *to* 1, *selection is so weak that it is nearly overwhelmed by drift*. Such alleles are referred to as **nearly-neutral**. Alleles with $|2Ns|$ in the range of about 1 to 10 do feel the effects of selection, but are also heavily influenced by drift as you see above.

What sets this scaling for the nearly-neutral range? One way to think about this is that if $2Ns = 1$ then selection effectively adds (or removes) one copy of the alternate allele per generation somewhere in the population [240]. Below this threshold selection is almost entirely ineffective [241].

Most new mutations are lost, even if they are favorable. The last important point is that even strongly favored alleles are vulnerable to the vagaries of random sampling when they are rare [d]. To illustrate this, the simulations shown below started with 1000 new mutations with a 1% selective advantage. Despite the selective advantage, only about 11/1000 of the simulated alleles spread to fixation; the rest were rapidly lost from the population. As you can see, most trajectories stayed below 1% and were lost by drift; in contrast, nearly all of the trajectories that got above 1% went into deterministic growth and reached fixation:

[d] *This is analogous to the card-counter who walks into a casino with a small initial purse. Even if she has a long-term advantage, she is likely to go bust early on.*

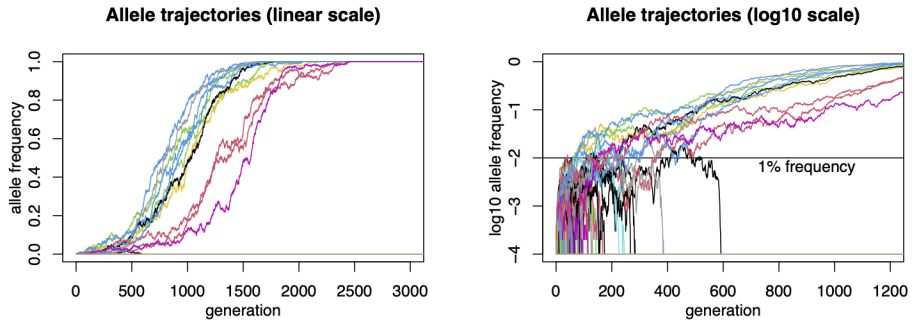

Figure 2.85: **Selection and drift of new favored mutations**. 1000 *simulated allele frequency trajectories with* $s = 0.01$*, starting from a single copy in a population of* $2N = 10^4$*. Around 99% of alleles were lost quickly, and are hard to see as they are effectively on top of each other along the* $y = 0$ *line. The right-hand panel shows the same data but with different axes: note the log*10 *scale on the y-axis to show rare variants more clearly.*

Fixation probabilities with selection. For strongly favored mutations, the probability that a new favored mutation fixes is only about $2hs$: this was 1% in the simulation above, close to the observed rate of $11/1000$. This fixation rate is much higher than the rate for neutral variants (i.e., $1/2N$) but still means that nearly all advantageous mutations are lost. For this reason, adaptation by new mutations can be highly inefficient [242].

A general formula for fixation probabilities with selection and drift was developed by the Japanese population geneticist Motoo Kimura in the 1950s (we'll hear from Kimura again soon, when we get to the Neutral Theory) [243]. For a new mutation with $h = 0.5$ the Kimura formula simplifies to

$$\text{Probability of fixation} = \frac{1 - e^{-s}}{1 - e^{-2Ns}}. \tag{2.68}$$

You can see this plotted here:

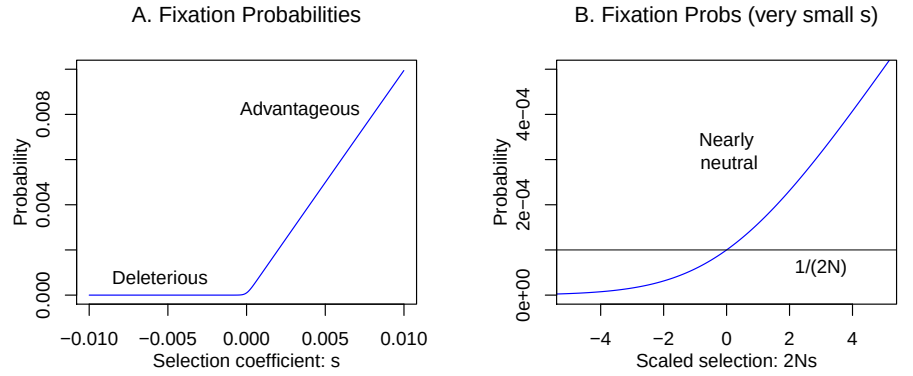

Figure 2.86: **Fixation probabilities of new mutations** *(here* $2N = 10^4$ *and* $h = 0.5$*).* **A.** *Fixation probabilities across a wide range of s.* **B.** *Nearly-neutral range: Same plot highly magnified near* $s = 0$*, with x-axis in units of* $2Ns$ *instead of s. The horizontal line at* $1/2N$ *shows the fixation probability for neutral mutations.*

The left-hand plot above illustrates that when selection is strong, the model does not depend on population size: strongly deleterious alleles have essentially no chance of fixing, and strongly advantageous mutations fix with probability $\sim 2hs$. Chance *does* matter for favored mutations, but only because it determines whether they start to spread when they are extremely rare [244].

But we see something quite different in the right-hand plot. This shows what happens in the **nearly-neutral range**, where selection is weak compared to drift (roughly $|2Ns| < 1$). These alleles drift very much like neutral alleles, and selection only modestly increases or decreases their chances of fixation [245].

We close this chapter with a deeper discussion of negative selection; we'll return to positive selection and balancing selection in the next chapter.

Purifying selection: protecting the genome against mutation. As we discussed in Chapter 1.5, our genomes suffer a barrage of mutations in every generation – around 70 per child. The vast majority of these are close to neutral, but among those with functional effects, the vast majority have deleterious effects. I think this is intuitive: if you introduce random typos into a written document, you're far more likely to reduce the quality of the writing than to improve it!

For this reason, the most common form of selection is against **deleterious** variants. The term "deleterious" refers to variants with fitness $s < 0$, and includes everything from severe disease-causing mutations to millions of variants across the genome with tiny effects on phenotypes and mildly negative effects on fitness. Selection against deleterious variants is referred to as **negative selection**; or sometimes **purifying selection** because it cleanses deleterious mutations from the genome.

Under a strictly deterministic model, a new deleterious mutation would not increase in frequency at all. But in practice, natural selection is competing against the randomness of genetic drift, and some deleterious alleles do manage to drift up to higher frequencies [e]. For this reason, at any given time, some variants segregating in a population are actually deleterious, but they tend to be at lower frequencies than neutral variants.

[e] *Going back to the gambling metaphor, even a completely rubbish player might win some money by luck early in a game, but they are extremely unlikely to keep winning indefinitely.*

Here you can see simulations comparing genetic drift of 1000 neutral variants (panels A and B) and 1000 deleterious variants with a fitness disadvantage of 5% (panels C and D). In each plot, all the trajectories were started from a single copy at time 0.

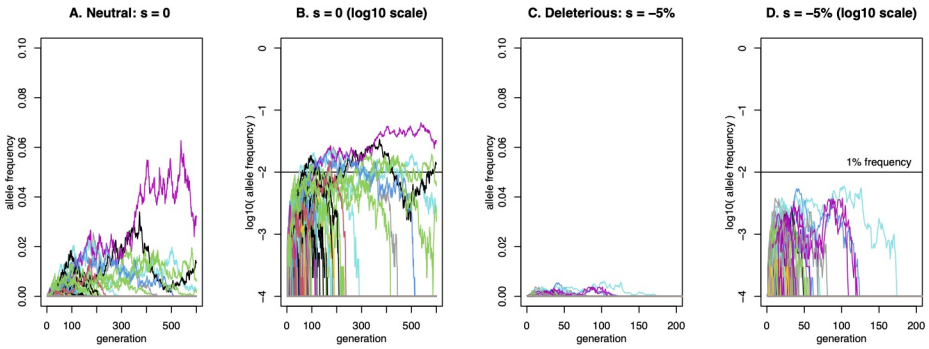

Figure 2.87: **Selection and drift of new mutations: neutral (panels A and B) and deleterious (panels C and D).** *Here panels A and B show the same data, but with the y-axis of B plotted on a log-scale to show more detail about rare variants. The same is true for C and D.*
Parameters: 1000 *simulated allele frequency trajectories for each panel, starting from a single copy at time* 0 *in a population of* $2N = 10^4$.

As you can see, the deleterious variants (C and D) are held at lower frequencies and are removed from the population much faster than the neutral variants.

Here, a useful approximation is that deleterious variants can drift up to a maximum frequency on the order of $\sim 1/(2N \cdot hs)$*, corresponding to selective removal of about one copy of the derived allele per generation.* This corresponds to 0.4% in Panels C and D above, which you can verify is close to the highest frequencies across the 1000 replicates.

The SFS for deleterious alleles. The plots above show trajectories of mutations over time, but in practice it's much easier to measure the distribution of allele frequencies across many different variants at a single point in time.

This is shown in the next plot, with theoretical distributions for neutral sites (red), nearly-neutral (blue), mildly deleterious (black) [246]. You can see that at low frequencies all three curves are similar, but at higher frequencies selection greatly reduces the numbers of deleterious alleles:

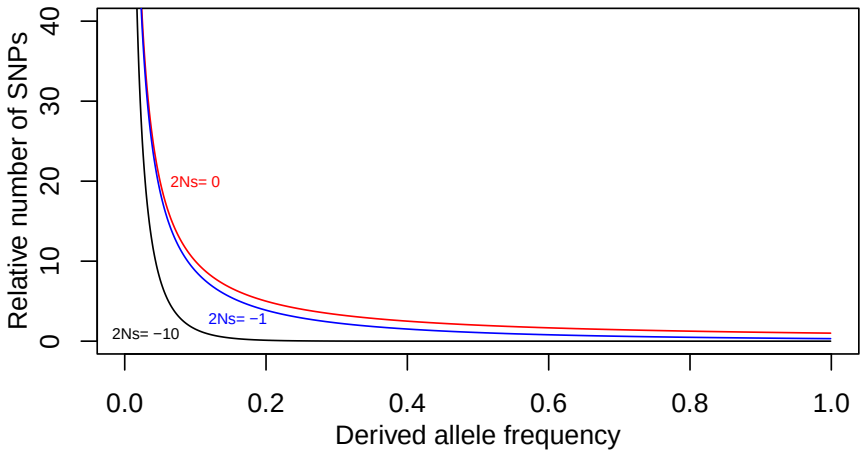

Figure 2.88: **Theoretical distributions for numbers of variants as a function of allele frequency, with weak purifying selection**. *The expected number of variants between frequencies p_1 and p_2 in a region of L basepairs is $4N\mu L$ times the integral from p_1 and p_2. Curves computed from theory in Sawyer and Hartl (1992) [Link].*

Hence, for a given number of base pairs, we see fewer total SNPs at deleterious sites, and the SNPs we do see tend to be at low frequencies.

We can also see similar effects in real data. The plot below, from 2005 [247], was one of the first to show the **site frequency spectrum (SFS)** [f] for sites with different levels of constraint [248].

This analysis tests the hypothesis that missense (nonsynonymous) variants are under purifying selection, and uses synonymous and noncoding variants as controls that are less often constrained [249]. Under this hypothesis, we would expect more of the missense variants to be at low frequencies.

[f] *Recall from Chapter 2.2 that the SFS shows the fraction of SNPs at each allele frequency.*

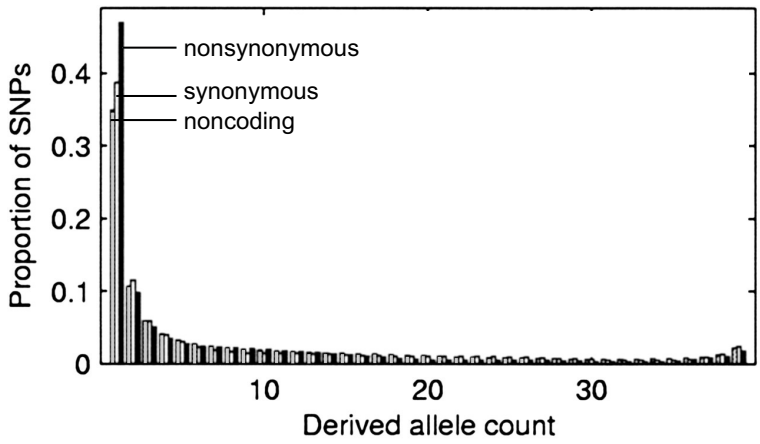

Figure 2.89: **SFS for different types of SNPs** *in a sample of size 40. Note that the plot is drawn differently than the theoretical plot, as here the histograms add up to 1 within each category. This plotting style emphasizes the relative shift toward rare variants for nonsynonymous SNPs. Credit: Modified Figure 1 from Scott Williamson et al (2005) [Link].*

Indeed, as you can see, around 48% of missense sites are singletons in

this data set, compared to around 35–38% of noncoding and synonymous sites [250]. *This reflects the fact that a large fraction of missense variants are under purifying selection.*

Sequence conservation between species is an important indicator of function. The SFS analysis is useful for showing that a *type* of variant (such as missense mutations) is under purifying selection, but it's not very useful for testing at individual sites [251].

However, recall that selection is extremely effective at preventing deleterious variants from fixing. So an alternative is to use **sequence conservation** between distant species to identify regions or sites that are functionally important. If we compare distantly related species, then a large fraction of neutral sites will show differences, but sites that are **functionally constrained** are much more likely to be shared.

This concept has been used to identify functional regions of the genome: for example important regulatory enhancers, or protein domains that are particularly crucial for protein function [252].

For example, the plot below shows sequence conservation between mouse and four distantly related vertebrates in the region around the TBX2 and TBX4 genes (highly conserved master regulators of limb development). Regions marked in blue are exons, are regions in red are putative noncoding elements. The boxed regions were shown to have regulatory activity in transgenic mouse experiments [253].

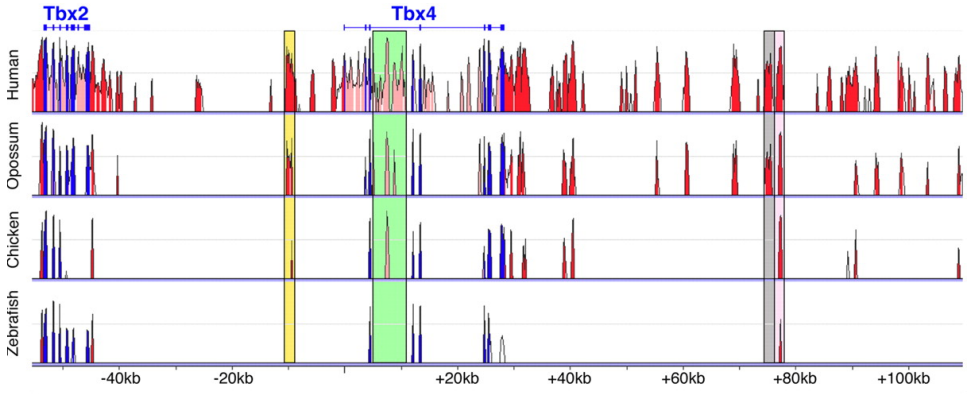

Figure 2.90: **Genome sequence conservation identifies functional elements.** *Each track shows regions with high sequence identity between mouse and the indicated species (the y-axis of each track ranges between 70 – 100% sequence identity). Coding exons are shown in blue and noncoding conserved regions in red.*
Credit: Figure 2 from Douglas Menke et al (2008) [Link]; CC BY.

Nearly-neutral mutations and the limits of natural selection. We've been talking about how natural selection tends to purge deleterious mutations. But as I discussed above, for variants with very weak selection, the vagaries of drift become more important than selection; we refer to these weakly selected variants as **nearly-neutral**.

There's no hard cutoff for a variant to be "nearly-neutral", but as I noted above, a common definition is $|2Ns| \leq 1$.

Here it's worth pausing to reflect on the fact that selection is an extraordinarily efficient process. To put this into numbers, if the human effective population size N_e is $\sim 15,000$, this implies that selection is efficient

down to around 3×10^{-5}, or three extra individuals surviving or reproducing per 100,000. Variants with s smaller than this are nearly-neutral.

And yet, while a single variant with a fitness cost of 10^{-5} is almost inconsequential, the combined impact of many nearly-neutral mutations can have meaningful effects. In particular, **the existence of the nearly-neutral zone places important limits on the extent to which natural selection can optimize genomes.**

This is especially true for species with small effective population size, including humans, since the size of the nearly-neutral zone depends on N. These species are much worse at safe-guarding their genomes from weakly deleterious mutations compared to species with larger populations including fruit flies, yeast, or E. coli.

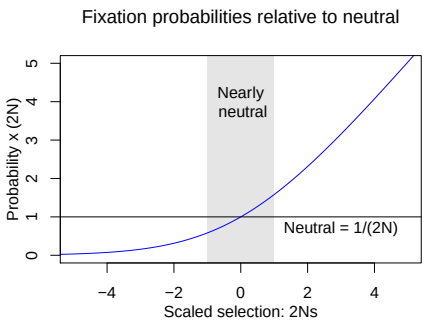

Figure 2.91: **The nearly-neutral zone**: *fixation probabilities of new mutations. A slightly deleterious mutation with* $2Ns = -1$ *is nearly as likely* ($0.58\times$) *as a new neutral variant to fix. A slightly advantageous mutation* ($2Ns = 1$) *is about* $1.6\times$ *more likely to fix than neutral.*

One setting where nearly-neutral mutations are relevant is for something called **codon bias**. As you know, different DNA triplets can code for the same amino acid (e.g., GGA, GGC, GGG, and GGT all encode glycine). Mutations that switch between alternative triplets encoding the same amino acid are referred to as synonymous. However, it turns out that some synonymous codons are slightly preferred over others, likely because they enable greater translation accuracy or speed. Preferences are species-specific and correlate with the abundances of the corresponding tRNAs [254].

These codon preferences can result in a *very slight selective benefit* to using one synonymous codon instead of another. But you can imagine that the fitness consequence of switching, for example, a single GGA to GGG in a single gene, is very very small. In consequence, the ability for a species to maintain codon usage bias depends on its effective population size – for species with sufficiently large N_e, codon switches can lie outside the nearly-neutral zone. As a result, many species with large N_e, such as in Drosophila, can maintain strong codon bias across the genome, while species with small N_e including humans cannot [255].

A second example comes from the difficulty that genomes have in controlling the spread of **transposable elements** (TEs). TEs are DNA elements that can copy themselves and reinsert the copies elsewhere in the genome, usually via an RNA intermediate [256]. While TE insertions do occasionally have salubrious effects [257], on the whole they are considered **selfish DNA**: they replicate because they can, but they do not benefit the host genome. Quite remarkably, it's estimated that more that 2/3 of the human genome was originally derived from transposable element insertions [258]. Moreover, 10% of your genome is made up by copies of just a single 300 bp element called **Alu**, which is present about in about 1 million copies [259]! Although a few Alu copies play functional roles in gene regulation, Alus are primarily parasitic elements.

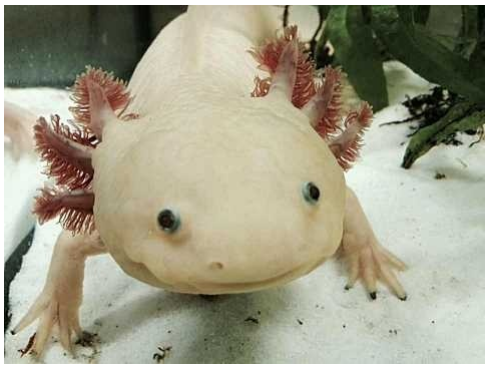

Figure 2.92: **Axolotl genomes are 10-fold larger than ours**. *The axolotl, a model organism for limb regeneration, has a huge 32GB genome chock-full of millions of transposable elements. This fact also provides the opportunity for a gratuitous axolotl photo.* Credit: th1098 [Link], CC BY-SA 3

The key problem is that the selective costs of most new TE insertions are very small. When an Alu is copied into a new location, there is a slight chance that it inserts into a functional region such as an exon, in which case it will probably be deleterious [260], and be removed by selection. But

if the new copy inserts into nonfunctional sequence, the added cost of the new copy is almost negligible – mainly the tiny cost of replicating a few hundred basepairs of additional DNA at every cell division [261] .

In fact, the marginal cost of each new Alu is so small that selection cannot effectively prevent individual Alus from fixing. At the same time, however, there are substantial genome-wide costs to carrying and replicating millions of TEs. In consequence, genomes have evolved trans-acting mechanisms for epigenetic silencing of TEs to try to reduce their rates of spreading [262].

A third example is **evolution of the mutation rate** [263]. The mutation rate depends on a number of factors: the rate of spontaneous damage and copying errors, as well as the ability of cells to fix these errors. These factors – in particular the complex machinery that cells use to prevent and repair errors – are of course evolved properties of organisms. What factors determine the evolution of the mutation rate?

Given that mutation is an essential component of evolution, you might think that some amount of mutation is helpful. That may be true for the long-term survival of a species, but from the viewpoint of an individual – which is what matters for natural selection – the overall effect of mutation is negative. The mutations your kids inherit may have no impact on their fitness, but if they do impact fitness, then it's *much* more likely that they have a negative effect than a positive effect.

Consider a new variant that makes the DNA repair machinery very slightly worse – such a variant is known as a **mutator** allele. Let's suppose this mutator increases the average genome-wide number of mutations by a single mutation. We can estimate that this mutator variant would decrease fitness by around 10^{-5}, which puts it in the nearly-neutral range for humans, and selection wouldn't be very good at removing it [264]. In contrast, a mutation that adds 10 new mutations per generation would have a ten-fold higher fitness cost and would be much more visible to selection [265].

This process creates what Michael Lynch has termed a **drift barrier**: natural selection cannot reduce the mutation rate indefinitely because below a certain point, any improvement to the mutation rate is nearly-neutral, and hence mainly governed by drift. **The mutation rate at which the drift barrier kicks in depends on population size.** Indeed, data on mutation rates of different organisms suggest that mutation rates are determined by the drift barrier model, as species with larger population sizes tend to have lower mutation rates:

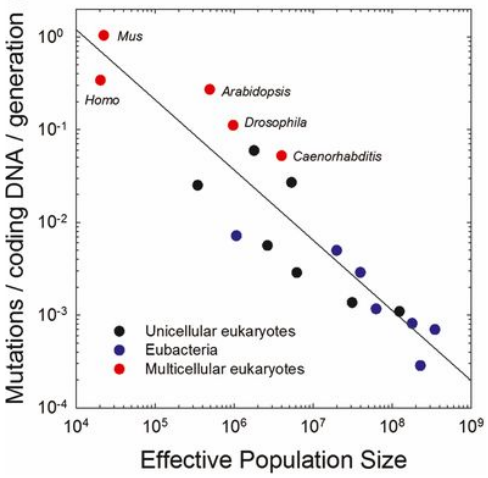

Figure 2.93: **Relationship between mutation rate and effective population size.** *The mutation rate here refers to the total number of mutations at protein-coding positions, per generation.* *Credit: Figure 1c from Way Sung et al 2012. [Link]*

Genetic load. Given these limits of natural selection, each of our genomes contains many deleterious variants. These come in two main categories:

- Each of us has *a unique personal collection of deleterious variants that are currently drifting at low to moderate frequencies* (and will eventually be removed from the population by natural selection).
- Like any other species, theory argues that *humans must also carry many fixed variants that are weakly deleterious*, but within the nearly-neutral range where selection is ineffective.

Together, these deleterious variants are referred to as genetic load.

It's been argued that the second of these categories – fixed nearly-neutral variants – leads to an evolutionary paradox. If we make the plausible assumption that many more mutations are slightly bad than slightly good, then we should predict an inexorable increase in genetic load over evolutionary time. One famous paper had the colorful title *"Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over?"* [266]. But clearly we're still around after 4 billion years of evolution so this argument cannot be fully correct. While the details are still not entirely clear, this argument likely under-states the ability of weakly advantageous mutations to counteract the accumulation of load [267]

Meanwhile, our burden of segregating deleterious variants is responsible for the genetic contributions to phenotypic variation and disease – and is something we'll come back to in much greater detail in Section 4 of the book.

In this chapter we have covered basic models of selection, with and without drift, and an overview of negative selection. In the next chapter we turn to a deeper consideration of positive selection.

Notes and References.

[235]In these models, the alleles compete against each other, but we assume that the population size is fixed by exogenous factors–perhaps food or other resources–and that selection at the variant in question does not directly drive population growth. This is referred to as "soft selection", and the genotype fitnesses are measured relative to one another. In contrast, in *hard selection* models, the genotypes have absolute fitness values, and this means that the population can grow, or grow faster, as fitter alleles increase in frequency. Soft selection models are theoretically more tractable, and usually a good approximation in humans where fitness gains from any single variant tend to be very small. Hard selection may be relevant in other situations–for example in modeling growth of E. coli on antibiotics, where an antibiotic resistance allele can allow a dramatic increase in growth rate.

[236]You'll often see this model parameterized slightly differently, denoting the fitness of each genotype by w with a subscript: i.e., w_{AA}, w_{Aa}, w_{aa}. But in the soft selection case what matters is the fitness of each genotype relative to the others, so we set the ancestral homozygote to be a *reference group*, and divide all three fitnesses by w_{AA}. Now the fitnesses are 1, w_{Aa}/w_{AA}, w_{aa}/w_{AA}, which we rewrite as $1, 1+hs, 1+s$. (We can do this provided that we don't have the special case of symmetric balancing selection $w_{AA} = w_{aa} \neq w_{Aa}$).

[237]First, recall that we want to compute $\Delta_p = \mathrm{E}[p'] - p$ where

$$\mathrm{E}[p'] = \frac{pq(1+sh) + p^2(1+s)}{q^2 + 2pq(1+sh) + p^2(1+s)} \tag{2.69}$$

We simplify the notation by using $\overline{w}$ in place of the denominator (pronounced w-bar, and referred to as "mean fitness"), and simplifying:

$$\begin{aligned} \overline{w} &= q^2 + 2pq(1+sh) + p^2(1+s) && (2.70)\\ &= q^2 + 2pq + 2pqsh + p^2 + p^2 s && (2.71) \end{aligned}$$

Noting that $p + q = 1$ and $q^2 + 2pq + p^2 = 1$ we simplify this to

$$\overline{w} = 1 + 2pqsh + p^2 s \tag{2.72}$$

Now we're ready to start calculating Δ_p as follows:

$$\begin{aligned} \Delta_p &= \frac{pq(1+sh) + p^2(1+s)}{\overline{w}} - p \times \frac{\overline{w}}{\overline{w}} && (2.73)\\ &= [pq(1+sh) + p^2(1+s) - p[1 + 2pqsh + p^2 s]/\overline{w} && (2.74)\\ &= p[q(1+sh) + p(1+s) - 1 - 2pqsh - p^2 s]/\overline{w} && (2.75)\\ &= p[q + qsh + p + ps - 1 - 2pqsh - p^2 s]/\overline{w} && (2.76)\\ &= p[qsh + ps - 2pqsh - p^2 s]/\overline{w} && (2.77)\\ &= ps[qh + p - 2pqh - p^2]/\overline{w} && (2.78)\\ &= ps[qh + pq - 2pqh]/\overline{w} && (2.79)\\ &= pqs[h + p - 2ph]/\overline{w} && (2.80)\\ &= pqs[h(1-2p) + p]/\overline{w} && (2.81)\\ &= pqs[h(q-p) + p]/\overline{w} && (2.82)\\ &= pqs[p(1-h) + qh]/\overline{w} && (2.83) \end{aligned}$$

which gives us the desired result.

[238]We assume that h is in the range of $[0, 1]$; in the next chapter we'll discuss balancing selection, which can happen when h is outside the range $[0, 1]$. Also note that $\overline{w}$ is positive under reasonable conditions.

[239]Overview of card counting: [Link], and an example of a card-counting technique: [Link]. And a classic movie scene about counting cards from Rain Man: [Link].

[240]To be more precise, if the allele is at frequency p, selection would add or remove $2Nsp$ copies in expectation. So for a common allele this is of order 1.

[241]A second intuition for why $2Ns = 1$ represents the lower bound for selection is that the expected change in allele frequency ($\mathrm{E}(\Delta_p)$ due to selection is on the order of $sp(1-p)$, while the variance in allele frequency due to drift ($\mathrm{Var}(\Delta_p)$ is $p(1-p)/2N$. So the expected change due to selection trumps the change in variance when $2Ns >> 1$.

[242]A nice description of the math for the haploid case is given by Otto and Whitlock (1997). Otto and Whitlock also point out that the fixation rate of new mutations is much higher in growing populations, and this is probably important in some ecological settings. See also Pritchard et al (2010) for further discussion of these issues:

Otto SP, Whitlock MC. The probability of fixation in populations of changing size. Genetics. 1997;146(2):723-33

Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010;20(4):R208-15

[243]Kimura M. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics. 1957:882-901

Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(6):713

[244]For strong positive selection, if the alleles are lucky enough to reach more than a handful of copies then the deterministic dynamics take over, and this randomness at very low numbers is independent of N. In fact the dynamics at very low sample numbers are often modeled as branching processes, ignoring the total population size. When $s > 0$, the branching process either goes extinct quickly or goes to infinity (i.e., fixation).

[245]You may be wondering what happened to the distinction between census population size N and effective population size N_e. I've been focusing on the ideal Wright-Fisher model where they are the same. For more general models both can matter: the initial frequency of a mutation depends on N (i.e., it is $1/2N$), but the rate of the drift depends on N_e. It's worth noting that N_e is a useful hack that gives us insight into complicated models, while not always being a perfect approximation. For example, fixation probabilities of advantageous alleles can be dramatically different with population size changes in a way that is not modeled by the neutral N_e. You can see this by noting that exponential growth (which is not well-modeled by a single N_e) gives new mutations a big boost; the same will be true to a smaller extent even with fluctuating population sizes (where N_e is traditionally computed as the harmonic mean of N); see Otto and Whitlock (1997). Meanwhile, Simons et al explored the interactions between selection, drift and population size changes, and found complicated effects on genetic load:

Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nature Genetics. 2014;46(3):220-4.

[246]The theoretical prediction for the number of sites at frequency p given mutational input $4N\mu$ is

$$4N\mu \frac{1 - e^{-2\gamma(1-x)}}{(1 - e^{-2\gamma})p(1-p)} \tag{2.84}$$

where $\gamma = 2Ns$. You can find derivations for this leading up to Equation 11 of Sawyer and Hartl (1992), and Equations 33 and 35 in the review by Senupathy and Hannenhalli (2008):

Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132(4):1161-76

Sethupathy P, Hannenhalli S. A tutorial of the poisson random field model in population genetics. Advances in bioinformatics. 2008;2008

[247]Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of Sciences. 2005;102(22):7882-7

[248]Recall from Chapter 2.2 that the SFS can be used to estimate population histories. Since the SFS is also influenced by selection, the demographic analysis would usually be restricted to putatively neutral sites, such as synonymous or noncoding sites.

[249]For real data we don't (yet) know the actual selection coefficients for most types of sites, but it's common to use synonymous and noncoding sites as proxies for a more-neutral baseline. While these sites may occasionally have functional effects such as altering splicing or transcription factor binding, they usually have little selection compared to coding sites.

[250]Note: It's not entirely clear why the noncoding sites have fewer singletons than synonymous in this analysis. I suspect it may reflect differences in sequence composition and mutation rates between exons and noncoding regions rather than major differences in functional constraint

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).

[251]If we see a common variant at a site then we can be confident this site is not under selective constraint. But even neutral sites generally don't have common variants so this test lacks sensitivity. However, there are new approaches that can detect strong selection in very large samples:

Agarwal I, Przeworski M. Mutation saturation for fitness effects at human CpG sites. Elife. 2021;10:e71513

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv. 2022:2022-03

[252]These methods are no longer as widely used for predicting gene regulation as recent improvements in functional genomics are far more interpretable, including providing cell-type specific information. Nonetheless the general principles are still important.

[253]Menke DB, Guenther C, Kingsley DM. Dual hindlimb control elements in the Tbx4 gene and region-specific control of bone size in vertebrate limbs. Development. 2008

[254]e.g., Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134(2):341-52.

[255]Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics. 2006;7(2):98-108

Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular biology and evolution. 2008;25(3):568-79

Hershberg R, Petrov DA. Selection on codon bias. Annual review of genetics. 2008;42:287-99

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Molecular biology and evolution. 2018;35(5):1092-103

[256]Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosome Research. 2018;26:25-43

[257]Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philosophical Transactions of the Royal Society B. 2020;375(1795):20190347

[258]de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 2011;7(12):e1002384

[259]Deininger P. Alu elements: know the SINEs. Genome biology. 2011;12(12):1-12

[260]Deininger PL, Batzer MA. Alu repeats and human disease. Molecular genetics and metabolism. 1999;67(3):183-93

[261]There is some tiny cost from the fact that it has to be copied every time the cell divides: the nucleotides, the energetic cost, and the copying time. If the Alu inserts inside an intron, it is must also be transcribed every time the gene is transcribed. Pairs of nearby Alu elements also occasionally trigger incorrect chromosome pairing and recombination

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. The American Journal of Human Genetics. 2006;79(1):41-53

Kim S, Cho CS, Han K, Lee J. Structural variation of Alu element and human disease. Genomics & informatics. 2016;14(3):70. Another potential issue arises from inverted Alu repeats in mRNA can form double stranded RNA (dsRNA). Since dsRNA is a hallmark of some viruses (and not ordinarily present in human mRNA), this can trigger an inappropriate (auto)immune response. There is an entire machinery evolved to edit dsRNA to reduce double-strand pairing

Chung H, Calis JJ, Wu X, Sun T, Yu Y, Sarbanes SL, et al. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. Cell. 2018;172(4):811-24

[262]e.g., Yang F, Wang PJ. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. In: Seminars in cell & developmental biology. vol. 59. Elsevier; 2016. p. 118-25.

[263]Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences. 2012;109(45):18488-92

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics. 2016;17(11):704-14

[264]To get a ballpark estimate, let's suppose that mutations in 1% of the genome would have an average deleterious effect on fitness of 10^{-3}. Assuming these numbers, each new mutation in the genome produces an average fitness cost of 10^{-5}, per generation (usually zero, and occasionally much higher, depending on where the mutation lands). There's an additional complication which is that the precise selective effect that a mutator allele experiences as the result of the mutations it produces is slightly more complicated because it can experience those effects over multiple generations. However in a recombining organism, it recombines away from the damage it produces at a rate of $1/2$ per generation. Lynch et al (2016) give the fitness effect of a mutator allele as being $\approx 2s\Delta(U_D)$, where s is average fitness effect of a new mutation, $\Delta(U_D)$ is the change in genome-wide mutation number caused by the mutator, and the factor of 2 reflects the average number of generations that the mutator is in the same genome as the mutations it causes. (Lynch 2016)

[265]For examples of mutator evolution in action see e.g.,

Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. Elife. 2017;6:e24284

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, et al. A natural mutator allele shapes mutation spectrum variation in mice. Nature. 2022;605(7910):497-502

[266]Kondrashov AS. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? Journal of theoretical biology. 1995;175(4):583-94

[267]One hypothesis is that protein evolution involves a lot of weakly deleterious substitutions that are repaired by very slightly advantageous compensatory mutations that maintain overall function.