Lecture 1: mutation and drift

Population geneticists study evolution and genetic inheritance using mathematical models. This lecture introduces the Wright-Fisher model, which is a simple model for two fundamental "forces" of population genetics, *mutation* and *drift*.

Key terms:

- A *site* is a position in the genome that may vary
- An *allele* is a possible DNA sequence at a site
- A *mutation* is the event that changes one allele into another
- A *variant* is a set of multiple alleles at the same site
- A *genotype* is the allele(s) of an individual or haplotype
- The *frequency* of an allele is the proportion of haplotypes that carry it

Mutation rates. The *mutation rate* refers to an expected number of mutation events, sometimes measured per site per generation, other times per genome, or per gene, etc. In humans, the combined mutation rate of single-base substitutions per genome per generation is around 100. In bacteria, this number is just ~1/1000, due to their smaller genome and smaller mutation rate per generation per site. However, bacteria have a very short generation time, so their mutation rate per *human* generation time is large. Human mutation rates also vary widely across sites. At methylated cytosine-guanine dinucleotides (CpGs), the rate of substitution (from C to T or G to A) is over 10 times higher than at non-CpGs (>10^-7 vs 10^-8).

We expect that most of these mutations do not have any functional effect, and that they are not under any type of selection. Some fraction will have deleterious effects, and a very small fraction can possibly be beneficial. Lecture II introduces models of evolution involving selection. Models not involving selection are called *neutral* models.

Equilibrium frequency with mutation alone. Suppose that a site has two alleles, *A* and *a*, and let p be the frequency of allele a in a very large population. Suppose that every generation, p changes only due to mutation: with probability μ, each A becomes an a; with probability ν, each a becomes an A. We can solve analytically for the equilibrium allele frequency, meaning the frequency such that an equal number of $A \to a$ and $a \to A$ mutations occur. The former is $(1-p)\mu$, the latter $p\nu$. Equating these, we find that the equilibrium frequency p^* satisfies:

$$\frac{p^*}{1-p^*} = \frac{\mu}{\nu}$$

This is a simple example of a kind of approach that will reoccur, where we find the change in allele frequency and set it equal to zero in order to solve for an 'equilibrium' value.

Infinite sites. Under the *infinite sites assumption*, the mutation rate per site, μ, goes to zero while the number of sites, S, goes to infinity. An infinite-sites model is parameterized by its combined mutation rate across sites, m. Under infinite sites, all carriers of an allele inherit it from a common ancestor; there is no back-mutation and no recurrent mutation.

Under infinite sites, we can study drift separately from mutation. Due to drift, every allele is eventually *lost* from the population (reaching frequency 0 and remaining at 0 forever), or *fixed* (reaching frequency 1).

Key terms:
- *Fixation* and *loss*: the event that allele frequency reaches 1 or 0, respectively
- *Fixation probability*: probability an allele eventually fixes
- *Time until fixation*: number of generations to fixation, if an allele does eventually fix
- *Time until loss*: number of generations to loss

Wright-Fisher model. There are multiple specific models that can be used to study these phenomena, but we'll focus on the Wright-Fisher model. Under Wright-Fisher, we have a population of size $2N$ haplotypes (the factor of two is a convention that we apply both to diploid and to haploid species). Time is indexed by an integer, t, which counts generations. Each generation, a new population is formed by drawing $2N$ independent samples *with replacement* from the old one. If the population at time t has an allele with frequency p_t, then the number of carriers at time $t+1$ is:

$$x_{t+1} = 2Np_{t+1} \sim Binomial(2N, p_t).$$

This distribution has mean $E(p_{t+1}|p_t) = p_t$ and variance $Var(p_{t+1}|p_t) = \frac{1}{2N}p_t(1-p_t)$.

Genetic drift is the change in allele frequencies over time due to the randomness of reproduction. Wright-Fisher is a model of drift. The rate of drift per generation under Wright-Fisher is the variance of p_{t+1}, which is inversely proportional to N.

Hardy-Weinberg equilibrium. In a diploid population, it makes sense to discuss the number of homozygotes and heterozygotes. If we randomly pair haplotypes to form diploids, then we obtain the following expected genotype frequencies, as a function of the allele frequency p:

$$p_{aa} = p^2$$
$$p_{Aa} = 2p(1-p) = H$$
$$p_{AA} = (1-p)^2$$

A site (or set of sites) with these proportions of diploid genotypes is said to be at Hardy-Weinberg equilibrium (HWE). We will return to this subject when we discuss population structure, which causes deviations from HWE.

In general, however, we will ignore the distinction between diploid and haploid populations. In the current setting, with no population structure and no selection, it is equivalent to (1)

sample a diploid to reproduce, then pick one of its two haplotypes as the one that is passed on, or (2) to just sample a haplotype.

Fixation probability. Under Wright-Fisher, an allele with starting frequency p_0 has fixation probability $u = p_0$, which can be shown using a martingale argument. (A *martingale* is a stochastic process with constant mean). We know that $E(p_1|p_0) = p_0$, and likewise,

$$E(p_2|p_0) = E(E(p_2|p_1)|p_0) = E(p_1|p_0) = p_0,$$

and so on. Assume that after T generations, the allele has either been fixed or lost: $p_T \in \{0,1\}$. (T is called a *stopping time*). Then:

$$E(p_T|p_0) = p_0 = u \cdot 1 + (1-u) \cdot 0 = u.$$

Fixation time. In a population of size $2N \gg 1$, an allele that starts with a low initial allele frequency $p_0 \ll 1$ has an expected fixation time of

$$\bar{t}_{fix} = 4N.$$

This is the expected time to fixation *conditional upon fixation*. It implies that common alleles are older in a large population. It is traditionally derived using diffusion, which is a continuous-time approximation. We can gain intuition by analyzing the expected change in p per generation conditional upon fixation.

For an allele with starting frequency p_0, conditional upon eventual fixation, the expectation of p_1 is no longer p_0. Instead:

$$\begin{aligned} E(p_1|p_0, \text{eventual fixation}) &= \frac{E(p_1 P(\text{eventual fixation}|p_1)|p_0)}{E(P(\text{eventual fixation}|p_1)|p_0)} \\ &= \frac{E(p_1^2|p_0)}{E(p_1|p_0)} \\ &= \frac{E(p_1|p_0)^2 + Var(p_1|p_0)}{E(p_1|p_0)} \\ &= p_0 + \frac{1-p_0}{2N}. \end{aligned}$$

We find that the expected change in allele frequency, $p_1 - p_0$, is inversely proportional to N, consistent with the expected time to fixation being proportional to N. When the allele is rare ($p_0 \ll 1$), the rate of increase is one allele per generation.

Heterozygosity. The standard measure of genetic diversity in a population – how different are individuals from each other? – is *heterozygosity*. The heterozygosity of a site is the probability that two random haplotypes, sampled with replacement, have a different allele at that site:

$$H = P(x \neq x').$$

x and x' are i.i.d., and under our biallelic model, they follow a Bernoulli distribution. At time t, the heterozygosity of a site with alleles at frequency p_t and $1 - p_t$ is:

$$\begin{aligned} H_t &= p_t(1-p_t) + (1-p_t)p_t \\ &= 2p_t(1-p_t). \end{aligned}$$

What is the expected heterozygosity in generation $t+1$, conditional on H_t? Pick two haplotypes from generation $t+1$, with replacement. They are *the same haplotype* with

probability $1/2N$, in which case they must have the same allele. They are different haplotypes with probability $1 - 1/2N$, in which case the probability they have a different allele is exactly the heterozygosity of generation t. Thus,

$$E(H_{t+1}|H_t) = \frac{1}{2N} \cdot 0 + \left(1 - \frac{1}{2N}\right) \cdot H_t = \frac{2N-1}{2N} H_t.$$

Heterozygosity decays to zero over time, at a geometric rate of $\frac{2N-1}{2N}$ per generation, if not restored by mutation.

Mutation-drift balance. Mutation continually restores genetic diversity. Suppose that a population of size $2N$ haplotypes, with initial heterozygosity H, undergoes mutation, which flips a fraction μ of haplotypes from one allele to another. The expected heterozygosity afterwards is:

$$H' = H(1 - 2\mu + \mu^2) + (1 - H)(2\mu - \mu^2)$$

where $H(1 - 2\mu + \mu^2)$ is the probability that you sample two haplotypes which previously had different alleles, and they still have different alleles after mutation, and $(1 - H)(2\mu - \mu^2)$ is the probability that you sample two haplotypes which previously had the same alleles, and exactly one of them mutated. The factor of 2 in 2μ comes from the fact that there are two haplotypes, either of which can mutate; the μ^2 terms account for the possibility of double-mutation. Here, it matters that we assume a biallelic model; the classical setting for mutation-drift balance assumes "infinite alleles".

Neglecting μ^2 terms:

$$\begin{aligned} H' &\approx H(1 - 2\mu) + (1 - H)(2\mu) \\ &= H(1 - 4\mu) + 2\mu. \end{aligned}$$

The expected change in heterozygosity due to mutation is $(2 - 4H)\mu$. We combine the forces of mutation and drift find that:

$$E(H_{t+1}|H_t) \approx H_t - \frac{H_t}{2N} + (2 - 4H_t)\mu.$$

.A quantity is at *balance* when the expected change in that quantity is zero. *Mutation-drift balance* refers to the phenomenon that for some value $H_t = H^*$, the expected change $H_{t+1} - H_t$ is zero. We can solve for H^*:

$$0 \approx -\frac{H^*}{2N} + (2 - 4H^*)\mu = -H^* + 4N\mu - 8N\mu H^*.$$

This gives the result that $H^* \approx \frac{4N\mu}{1+8N\mu}$. The classical result, under infinite alleles, is $H^* \approx \frac{4N\mu}{1+4N\mu}$, which differs because an infinite-alleles model excludes back-mutation. These approximations are accurate when μ is small but $4N\mu$ might be large. They both agree in the that when $4N\mu \ll 1$,

$$H^* \approx 4N\mu.$$

This approximation can be easily derived using the *coalescent*, which will be discussed in a later lecture.

Note for students familiar with stochastic processes. In general, solving an equation of the form $E(\Delta X|X = X^*) = 0$ for an equilibrium value X^* does not give the *expected* value of that

quantity, or the average over time, because random deviations from the equilibrium value may have nonzero mean. In this particular case, however, the function $E(\Delta H|H)$ was an affine linear function of *H*, having the form $f(X) = a + bX$. Such a function commutes with integration: $E\big(f(X)\big) = f\big(E(X)\big)$. If we apply this fact to the stationary distribution of H, we find:

$$E(\Delta H) = 0 = E\big(E(\Delta H|H)\big) = E\big(\Delta H\big|H = E(H)\big).$$

Thus, H^* is both the equilibrium value and the mean of the stationary distribution.

The site frequency spectrum (SFS). The SFS is the distribution of derived allele frequencies across sites. Let s_k be the number of derived alleles having allele count exactly equal to $k \in \{1, \dots, 2N-1\}$ in a population at mutation-drift balance, under infinite sites. The SFS in this setting is:

$$E(s_k) = \frac{4N\mu}{k}.$$

The number of sites with allele count at most k is a harmonic series, which grows logarithmically:

$$\sum_{k=1}^{n} \frac{1}{k} \approx 0.58 + \log n$$

This implies, for example, that the number of sites with allele frequency in the range $(0.01,0.02)$ is approximately the same as the number of sites in the range $(0.02,0.04)$ or $(0.04,0.08)$.

The SFS is usually derived using *diffusion*, a continuous-time model, which is outside of the scope of this unit. Diffusion approaches involve partial differential equations for probability density functions parameterized by the time t, the allele frequency p_t, and the starting frequency p_0. These equations are called the *Kolmogorov forward equation*, which relates partial derivatives with respect to t and p_t, and the *Kolmogorov backward equation*, which relates partial derivatives with respect to t and p_0.

Lecture 2: selection part I

Selection is the most salient topic in evolutionary biology. Evidence for selection is everywhere, to such an extent that one might think it the *only* topic. In fact, the effects of selection can only be understood in the context of other forces, particularly mutation and drift. In this lecture we cover the ways that selection, mutation, and drift interact. In Lecture 5 we will cover the ways that selection also interacts with migration and linkage.

Luria-Delbruck experiment. In 1943, Salvador Luria and Max Delbruck performed a famous experiment resolving an open question about the interaction of selection and mutation. *Does selection act upon standing variation in a population* (Darwinian selection)*, or does it spur the production of advantageous mutations within a single generation* (Lamarckian selection)*?* They expanded a small initial colony of bacteria, giving rise to a large population containing many new alleles at various frequencies. Then, at generation 0, they introduced a phage (virus) that killed most of the colony, except for resistant strains. They counted the number of survivors and repeated the experiment many times.

Under the Lamarckian hypothesis, resistance mutations arise after each individual bacterium encounters the phage, independently in each individual. Because these are independent events, the number of survivors follows a Poisson distribution with some rate.

Under the Darwinian hypothesis, pre-existing mutations confer protection. Occasionally, such a mutation would have occurred early in the pre-challenge expansion, such that a large fraction of the colony is resistant and there are many survivors. Other times, such a mutation never occurs, and there are no survivors. More generally, the number of survivors is *overdispersed*, with variance much greater than its mean (i.e., greater than expected under the Poisson). Luria and Delbruck worked out this distribution – a complex bit of theoretical population genetics – and showed that it fit their experimental data.

This clarified the relationship between mutation and selection. Selection affects the frequency of existing alleles; it does not guide mutation to produce them.

Wright-Fisher with selection. Let p_0 be the frequency of an allele a of a biallelic variant a/A at time 0 in a population of size $2N$ haplotypes. Suppose that carriers and non-carriers of a have a different *relative fitness*, which is proportional to the expected number of offspring. By convention, we define the relative fitness of non-carriers to be $W_A = 1$. The *average fitness* is:

$$\bar{W} = pW_a + (1-p)W_A.$$

The expected number offspring for each haplotype j is $W_j/\bar{W}$. As a result, the expected allele frequency at generation $t+1$ is:

$$E(p_{t+1}) = \frac{p_t W_a}{\bar{W}}.$$

The ratio $W_a/\bar{W}$ is called the *absolute fitness* and is denoted w_a. Unlike W_a, w_a is a function of the allele frequency (which affects $\bar{W}$).

One way to understand W_a is that it is an *odds ratio*. The "odds" corresponding to a probability p are $o(p) = \frac{p}{1-p}$. We have that:

$$W_a = \frac{o\big(E(p_{t+1})\big)}{o(p_t)}.$$

It is often more convenient to work with relative rather than absolute fitness because the latter changes as a function of allele frequency. When a beneficial allele *a* is rare in the population, carriers can have a much larger number of offspring on average than the population mean. When it is already common, this becomes mathematically impossible. For example, for an allele with frequency 0.5, the maximum possible absolute fitness is 2.

The *selection coefficient* s_a is a conventional parameterization of relative fitness:

$$s = W_a - W_A = W_a - 1.$$

It is also sometimes defined as $s = \log W_a$ (which is equivalent for $W_a \approx 1$), in which case it is interpreted as a log-odds ratio, similar to a coefficient in logistic regression.

The expected change in allele frequency from one generation to the next is:

$$E(\Delta p) = p(w-1) = p\left(\frac{1+s}{\bar{W}} - 1\right) = p\left(\frac{1+s-(1+ps)}{\bar{W}}\right)$$
$$= \frac{sp(1-p)}{\bar{W}}.$$

Key terms:

- The *relative fitness* of a haplotype or individual is proportional to their expected number of offspring in the next generation
- The *average fitness* of a population is the mean relative fitness
- The *absolute fitness* is the relative fitness divided by the average fitness
- The *selection coefficient* of an allele is the effect of the allele on relative fitness

Positive and negative selection. *Positive selection* or *adaptation* refers to selection in favor of a derived allele. *Negative selection* or *purifying selection* refers to selection in favor of an ancestral allele. Negative selection is more common than positive selection. In this course, we use a positive selection coefficient for positive selection, a negative selection coefficient for negative selection. Unfortunately, both conventions are used; you will often encounter positive selection coefficients used to indicate negative selection.

Simultaneous effects. When analyzing the change in allele frequency from one generation to the next, you should normally model selection, mutation and drift as though they occur

simultaneously. For example, suppose that the average change in frequency due to mutation is:

$$\Delta_{\text{mutation}}(p) = \mu(1-p)$$

and the average change due to selection is:

$$\Delta_{\text{selection}}(p) = sp$$

You may model the total change in a generation as:

$$\Delta_{\text{total}}(p) = \Delta_{\text{mutation}}(p) + \Delta_{\text{selection}}(p) = \mu(1-p) + sp$$

instead of, for example,

$$\begin{aligned}&\Delta_{\text{mutation}}(p) + \Delta_{\text{selection}}\big(p + \Delta_{\text{mutation}}(p)\big)\\ &= \mu(1-p) + s\big(p + \mu(1-p)\big)\\ &= \mu(1-p) + sp + s\mu(1-p).\end{aligned}$$

The interpretation of the latter formula is that selection acts after mutation, on the allele frequency including the change due to mutation. If $s\mu$ is small, then the extra term that is introduced by non-simultaneous changes is negligible.

Selection on diploids. Diploid organisms can have fitness that is different from the sum of their two haplotypes. This is commonly parameterized using a *dominance coefficient*, h, with relative fitnesses:

$$\begin{aligned}W_{AA} &= 1\\ W_{Aa} &= 1 + hs\\ W_{aa} &= 1 + s\end{aligned}$$

When $h = 1/2$, selection is *additive*, and it can be modeled equivalently by analyzing haploids, which we do for the remainder of the lecture. However, the following equation allows you to analyze non-additive (*overdominant* or *underdominant*) selection, in addition to other exotic phenomena:

$$E(\Delta p) = \frac{p(1-p)}{2}\frac{d}{dp}\log \bar{W}.$$

This equation is *Wright's Formula*. Under Hardy-Weinberg equilibrium, $\bar{W}$ is:

$$\bar{W} = p^2 W_{aa} + q^2 W_{AA} + 2pqW_{Aa}.$$

Fixation probability of a beneficial allele. We can use a *branching process* to derive the approximate fixation probability of a rare, beneficial allele a. The intuition is that there are two stages:

- When p is small, the different copies of s do not compete with each other: $\bar{W} \approx 1$
- When p is large, such that $\bar{W} > 1$, fixation is almost certain.

In the stochastic stage, we may assume that the number of offspring for each haplotype having a is independent. This gives rise to a branching process where:

$$X = \text{number of offspring for each carrier} \sim Poisson(1+s).$$

Let v be the *extinction probability* of this branching process for a starting value of 1. v follows the following recurrence relation:

$$v = P(\text{extinction}) = E\big(P(\text{extinction}|X)\big) = E(v^X)$$

where X is the number of offspring in generation 1, and v^X is the probability that every one of the X offspring goes extinct. The function $G(z) = E(z^X)$ is the *probability generating function* of X, which we can look up. For the Poisson distribution with $\lambda = 1 + s$:

$$G(z) = e^{(1+s)(z-1)}.$$

We must solve:

$$v = e^{(1+s)(v-1)}.$$

One solution is $v = 1$, which corresponds to the initial state that $p = 0$ already. Let $u = 1 - v$. We take a Taylor expansion of the right-hand side as follows:

$$e^{-u(1+s)} = 1 - u(1+s) + \frac{1}{2}[u(1+s)]^2 + \cdots$$
$$\approx 1 - u - us + \frac{1}{2}u^2$$

where we have kept terms up to order 2 in s and u (for example, we dropped the 4th-order term $\frac{1}{2}(su)^2$ and the 3rd-order term $-\frac{1}{3!}u^3$). We find that:

$$1 - u \approx 1 - u - us + \frac{1}{2}u^2$$

which has solutions $u = 0$ (as before) and $u = 2s$.

Fixation probability of a deleterious allele. We cannot analyze deleterious alleles using a similar argument, which would imply that the fixation probability is always zero. An exact formula, applicable to advantageous or deleterious variants with starting frequency $1/2N$, is that:

$$\text{Fixation probability} = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}.$$

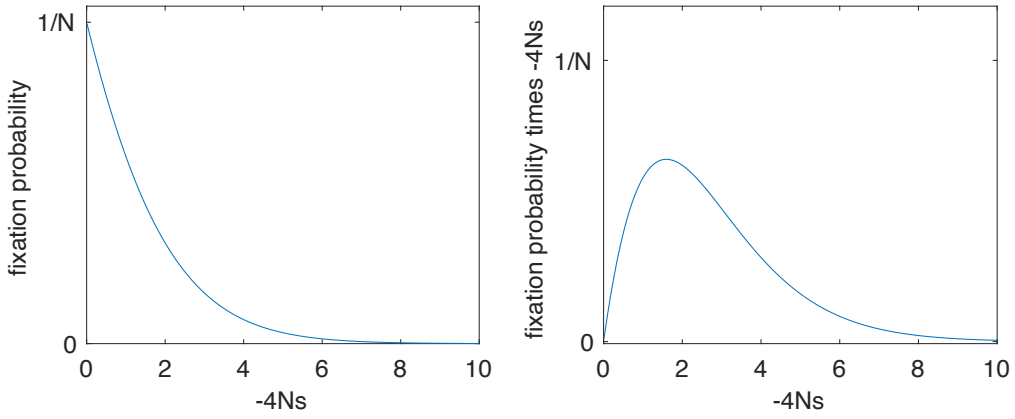
This formula is undefined when $s = 0$, but the limit as $s \to 0$ agrees with the Lecture 1 result $u = 1/2N$. When $s > 0$ is large enough that $e^{-4Ns} \approx 0$, but small enough that $e^{-2s} \approx 1 - 2s$, it agrees with approximation $u \approx 2s$.

Inefficiency of selection. Due to genetic drift, there exists a large regime where *selection is inefficient*:

- An advantageous mutation is unlikely to fix, because it is often lost due to drift
- A deleterious mutation can sometimes fix in a finite population

Deleterious alleles that may sometimes fix are often called "nearly neutral." There is no precise definition of this regime, but its size depends upon the population size. The existence of a nearly neutral regime gives rise to a phenomenon called the *drift barrier*, which we will discuss in Lecture 5.

The drift barrier. The drift barrier is a fundamental limitation upon the ability of a finite population to reach a fitness optimum. Because weakly deleterious mutations often fix in a finite population, the average fitness of a population depends upon its effective population size. The drift barrier can be illustrated by the second panel of this plot.



There is a "sweet spot," roughly $-4Ns \in [1,4]$, where deleterious alleles are able to fix despite the effect of selection. The size of the sweet spot depends on N (note the x-axis scale). For $-4Ns \ll 1$, selection is weak enough that fixation is inconsequential. For $-4Ns \gg 4$, the fixation probability becomes vanishingly small. In between, the fixation of deleterious alleles produces genetic load, and a population may suffer illness as a result. This phenomenon is population size dependent; smaller populations suffer greater load.

Mutation-selection balance. On the other hand, negative selection can be strong enough to overwhelm the effects of drift. Consider a highly deleterious allele ($s \ll 0$) and further suppose that the mutation rate times the population size is large:

$$2N\mu \gg 0$$

Every single generation, many new copies of this allele arise in the population, such that a fraction $p_0 = \mu$ of haplotypes carry the mutation *de novo*. This fraction is noisy in a finite population, but it becomes increasingly stable as $2N\mu \to \infty$. Moreover, some additional fraction p_1 of haplotypes inherit the mutation from an ancestor who had the mutation *de novo*. This fraction is smaller than p_0 because of selection:

$$p_1 = p_0(1+s)$$

Therefore, the total frequency in the population is:

$$\begin{aligned} p &= p_0 + p_1 + p_2 + \cdots \\ &= p_0 + (1+s)p_0 + (1+s)^2 p_0 + \cdots \end{aligned}$$

Recall that $\frac{1}{1-x} = 1 + x + x^2 + \cdots$. Letting $x = 1 + s$,

$$p = \frac{p_0}{1-(1+s)} = -\frac{\mu}{s}.$$

This derivation disregards the possibility of back-mutation, which is acceptable when $p \ll 1$. When $s \geq 0$, the series does not converge.

A different way to arrive at the same conclusion is to reason that at mutation-selection balance, the number of new mutations per generation should equal the number of deleterious alleles lost to selection. The former is μ, and the latter is $-ps$ because a

fraction p of haplotypes have the mutation, and of those, a fraction $-s$ are lost. Solving $\mu + ps = 0$ gives $p = -\mu/s$.

Estimating genic constraint. Usually, one cannot produce a useful estimate for the selection coefficient of a variant because the only information available is its frequency. This is because:

- Most sites lack a common variant, such that large s cannot be ruled out.
- Essentially all sites, even CpGs, have small enough μ that they are usually rare even when $s = 0$.

An exception to this rule occurs when μ and s are both large, and this is the case for protein truncating variants (PTVs) in many genes. PTVs ablate gene function via a variety of mechanisms: they might introduce a premature stop codon, delete a start codon, produce a frameshift indel, or cause the inclusion of an intron. Approximately 5% of all coding mutations are PTVs. If we assume that all PTVs in a gene are functionally equivalent, and that they have the same selection coefficient, then we can pool PTVs into a single 'variant' with some combined allele frequency and some combined mutation rate. The selection coefficient of PTVs in a gene is called s_{het}; genes with large s_{het} are said to be *constrained*.

s_{het} can be estimated using following procedure:

1. Identify all substitutions in a gene which would, if observed, truncate the protein
2. Estimate the mutation rate at each of these sites
3. Estimate the actual allele frequency at each site in a large sequencing dataset
4. Compute

$$\hat{s}_{het} = -\frac{\sum_{\text{sites } i} \mu_i}{\sum_{\text{sites } i} p_i}$$

A limitation of this approach is that it is very noisy for short genes; the denominator can be zero or close to zero. There exist various strategies to address this. The widely used gnomAD resource provides observed vs. expected variant counts and allele frequencies, which could be used to estimate s_{het}, but they choose instead to report a measure called LOEFF. This is a lower-bound estimate for the fraction of observed over expected loss of function variants.

A majority of human genes have s_{het} between -0.001 and -0.1, and a handful have s_{het} values of nearly -1.

Lecture 3: population structure and demography

Population structure refers to genetic differences between individuals that results from nonrandom mating over time. It is often due to geography. *Demography* refers to the genetic history of a population, including for example migration events that might break down population structure. The human population is structured at multiple levels, and our demographic history strongly shapes the patterns of genetic variation that are observed around the world today.

Key terms:

- *Population*: a set of individuals being modeled together
- *Demography*: events (like migration, population size change) that occurred in the history of a population and affect its genetics
- *Ancestry*: the demographic history of an individual or group of individuals, not to be confused with race and ethnicity
- *Divergence*: when a randomly mating population splits into two populations with less-than-random mating between them
- *Migration*: the movement of individuals between randomly mating populations, leading to *admixture*

Changes in population size could be analyzed separately from divergence and migration, but the latter are often accompanied by the former in human history, so it makes sense to consider them in tandem.

Divergence. In the absence of migration, the divergence of two populations from a shared source population can be studied by considering the divergence of each derived population from the source population separately. This is because allele frequency trajectories in each population are independent of each other. If a neutral allele begins at time 0 with a frequency $p^{(0)}$ and undergoes drift for T generations in a population of constant size $N \gg t$, then the mean and variance of the frequency at time T, $p^{(T)}$, are:

$$E\left(p^{(T)}\right) = p^{(0)}$$

$$Var\left(p^{(T)}\right) \approx \frac{Tp^{(0)}\left(1-p^{(0)}\right)}{2N},$$

where the approximation is equivalent to:

$$H_T = H_0\left(1-\frac{1}{2N}\right)^T \approx H_0\left(1-\frac{T}{2N}\right).$$

If two populations diverse from the same ancestral population, and they have population sizes N_1 and N_2, then the difference between their frequencies is:

$$Var\left(p_1^{(T)} - p_2^{(T)}\right) = Var\left(p_1^{(T)}\right) + Var\left(p_2^{(T)}\right) - Cov\left(p_1^{(T)}, p_2^{(T)}\right)$$

$$\approx Tp^{(0)}\left(1-p^{(0)}\right)\left(\frac{1}{2N_1}+\frac{1}{2N_2}\right).$$

Fixation index. One way to understand population structure is to consider a study that ascertains $n_1, \ldots, n_K$ haplotypes from populations $1, \ldots, K$. We can then study properties of the "study population," which can be viewed as weighted union of the source populations. For a sample i, let x_i denote its genotype for some allele, and let p_i be the frequency of that allele in the population from which sample i is drawn. The allele frequency in the study population is:

$$p = E(p_i).$$

The heterozygosity is:

$$H = P(x_i \neq x_j) = Var(x_i - x_j) = 2p(1-p).$$

A measure of the total difference among source populations is the *fixation index*, F_{ST}. F_{ST} is a ratio of between-population vs. between-haplotype variation in a weighted union of populations. The between-population variation is:

$$H_{\text{pop}} \coloneqq Var(p_i - p_j) = 2Var(p_i).$$

The ratio is:

$$F_{ST} \coloneqq \frac{H_{\text{pop}}}{H} = \frac{Var(p_i)}{Var(x_i)}.$$

When the allele has been either fixed or lost in every population, then $p_i = x_i$ for all i, and $F_{ST} = 1$, hence its name. If $p_i = p$ for all i, then $F_{ST} = 0$.

Calculating F_{ST}. F_{ST} is commonly calculated for a pair of populations, implicitly giving them equal weight. In this case, we can write p_1 and p_2 for the allele frequency in each population and $p = (p_1 + p_2)/2$. The variance of $p_i \in \{p_1, p_2\}$ is

$$Var(p_i) = E((p_i - p)^2) = \frac{1}{4}(p_1 - p_2)^2.$$

Let $q_1 = 1 - p_1, q_2 = 1 - p_2$. The denominator is:

$$\begin{aligned} Var(x_i) = p(1-p) &= \frac{p_1 + p_2}{2}\left(1 - \frac{p_1 + p_2}{2}\right) \\ &= \frac{1}{4}(p_1 + p_2)(q_1 + q_2). \end{aligned}$$

Thus, F_{ST} for this variant is:

$$F_{ST} = \frac{Var(p_i)}{Var(x_i)} = -\frac{(p_1 - p_2)(q_1 - q_2)}{(p_1 + p_2)(q_1 + q_2)}$$

where we used that $p_1 - p_2 = -(q_1 - q_2)$.

Like the heterozygosity, F_{ST} is usually taken across many sites. When performing a meta-analysis of ratios such as this, it is best to average the numerator and denominator first, taking a "ratio of averages" instead of an unstable "average of ratios:"

$$\widehat{F_{ST}} = \frac{\text{average of numerators}}{\text{average of denominators}} = -\frac{\sum_{\text{variants}}(p_1 - p_2)(q_1 - q_2)}{\sum_{\text{variants}}(p_1 + p_2)(q_1 + q_2)}.$$

Common values of F_{ST} in humans between continental ancestry groups are around 0.1. A typical within-continent value might be 0.02. Looking back to the definition of F_{ST}, these

numbers can be interpreted to mean that between-individual variation is much greater than between-group variation.

F_{ST} over time. Over a short period of time T, for a small starting value of F_{ST}, the expected change in F_{ST} can be calculated as follows. For a sampled individual i, let N_i be the population size of the source population from which they are drawn. The change is approximately:

$$\Delta F_{ST} \approx \frac{\Delta H_{\text{pop}}}{H}$$

and the numerator is:

$$\begin{aligned} 2\Delta Var(p_i) &= 2E\big(Var(\Delta p_i|i)\big) \\ &= 2E\left(\frac{T}{2N_i}p_i(1-p_i)\right) \\ &\approx TE\left(\frac{1}{N_i}\right)H \\ &= \frac{T}{N_e}H \end{aligned}$$

where we used that Δp is uncorrelated across populations in the first line, and we assumed that $F_{ST} \ll 1$ initially. This gives us:

$$\Delta F_{ST} \approx \frac{T}{N_e}.$$

Migration and admixture. Ancient DNA evidence has demonstrated that human evolutionary history is characterized by a constant churn of migration, displacement, and genetic admixture. *Migration* is the movement of individuals among populations. *Admixture* is the resulting pattern of genetic mixture, involving recombination between haplotypes of varying local ancestry. Migration from one population to another is modeled using a rate parameter, m, which might often vary with time. (A short-lived migration event is called a *pulse*.) In each generation, a fraction m of haplotypes in a group are sampled from some other group, and $1-m$ from within the group. Over time, the population will reach mutation-migration-drift balance, with some equilibrium value of F_{ST} that depends upon the rate of migration and the rate of drift (which depends upon population size). It does not really depend on mutation, except that mutation is needed to maintain some variation. The effect of drift on F_{ST} is to cause it to increase at a constant rate. The effect of migration is to cause it to decrease at a rate that depends on F_{ST} (in particular, the rate is 0 if F_{ST} is 0).

In sexually reproducing populations, migration gives rise to observable *admixture*. Individuals in the admixed population have long tracts of local ancestry from one source population or another. It is possible to make inferences about past migration events by identifying these tracts. For example, the age of the migration event can be estimated by looking at the length of these tracts, as they tend to be broken up by recombination over time.

Geometric view. We may imagine the vector of allele frequencies in each population, $p_k^{(t)}$, as a random walk in high dimensional space, each beginning at the same point $p^{(0)}$. In the absence of migration, these vectors are always mutually orthogonal in relation to the ancestral population. (This is a generally useful fact to know: in high-dimensional space, two random vectors are approximately orthogonal). In the presence of migration, they are nonorthogonal, because their changes are correlated. F_{ST} measures the average squared distance of the populations from each other.

Principal components analysis (PCA). PCA is a widely used approach to estimate population structure. Given some points in a vector space V, it finds an (affine) subspace of V that can be viewed in two equivalent ways:

1. It minimizes the distance between the subspace and the data, i.e., the sum of squared residuals
2. It maximizes the explained variance, i.e., the variance of the data projected onto U

If V is the space of possible genotypes, i.e., R^L where L is the number of sites, then any random subspace will explain $1/L$ of the variance on average. On the other hand, if our data come from two different populations with allele frequency vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, then the subspace spanned by $\boldsymbol{p}_1 - \boldsymbol{p}_2$ will explain a fraction of the variance equal to F_{ST}.

To perform PCA, you follow these steps:

1. Mean-center genotype matrix *X*
2. Standardize columns of *X*
3. Top eigenvector(s) of *XX'*
4. Make a scatterplot
5. Avoid overinterpreting it!

Indeed, PCA comes with some major caveats. One is that specific axes are generally uninterpretable: if there is some *d*-dimensional structure in your data, PCA may succeed in finding the corresponding subspace, but its *d* axes will not have any interpretation individually. In particular, they won't pick out specific populations or tell you about admixture proportions; you'll need to use a different method for that. Second, if structure is weak or your sample size is too small, your PCs will be totally uncorrelated with the true structure. There exists a phase change phenomenon whereby structure only becomes detectable at a threshold; below the threshold, it is totally undiscoverable (by PCA or any other method).

Lecture 4: linkage disequilibrium and recombination

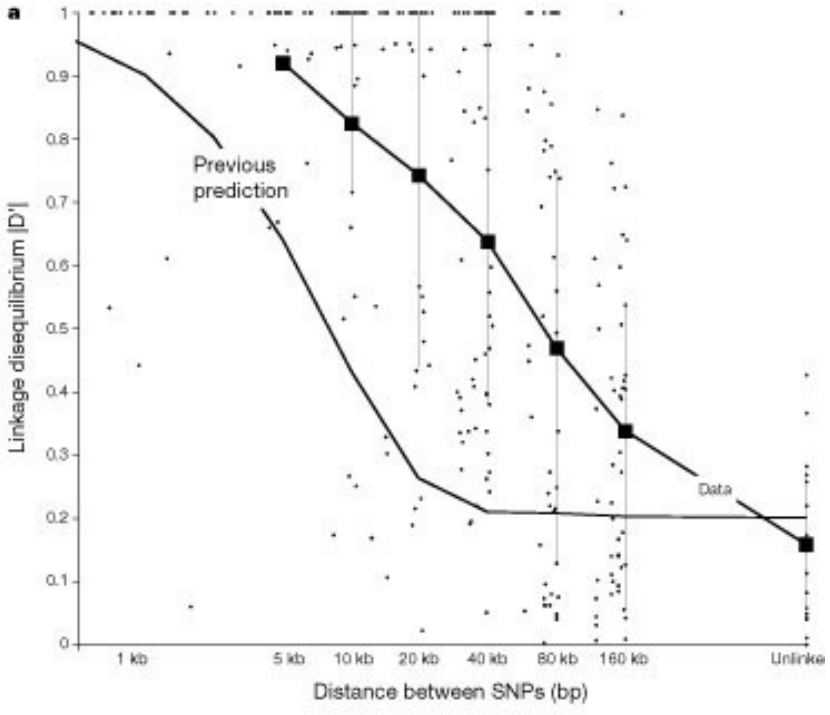
So far, we have considered models of just a single site, or of multiple sites segregating independently. In fact, different sites co-segregate, as they are physically located upon the same DNA molecule. This phenomenon – *linkage* – produces correlations between alleles. These correlations play a key role in modern statistical genetics, as when modeling the correlation between traits and alleles, it is relevant what correlations those alleles have with each other.

Key terms:

- *Linkage* is the physical connection between different sites
- *Linkage disequilibrium* (LD) is the correlation between alleles at different sites due to haplotype structure
- *Recombination* breaks apart haplotypes and causes LD to decay with distance
- *Coalescence* is when two lineages, looking backward in time, converge upon a common ancestor

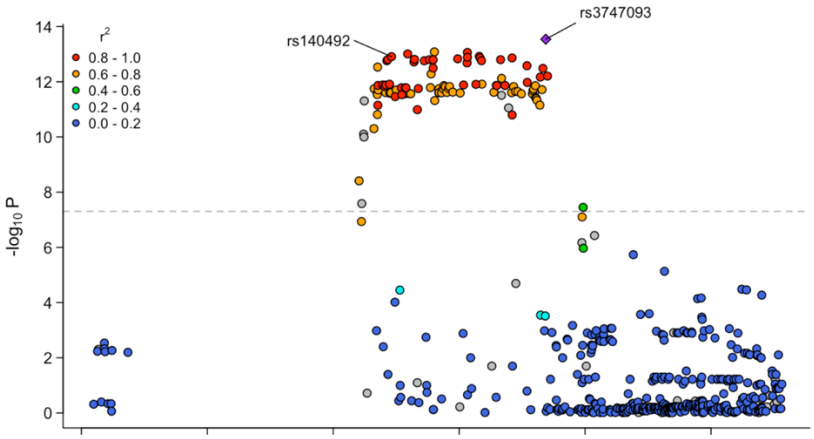
LD in humans. Around the time of the Human Genome Project, LD in humans was a topic of intense interest. A major motivation for the project was to map genetic associations with common diseases. Previously, *linkage studies* were used to map rare disease-causing loci in human families, by assaying a sparse array of markers (tandem repeats). If one of these was in linkage with the disease-causing mutation, it could be used to trace the causal locus within families, even if the causal variant was different. For common disease, it would be necessary to do something similar among non-relatives, relying on much shorter shared haplotypes. Thus, population geneticists produced estimates of the amount of LD present in humans.

In 2001, the HapMap project produced some of the first estimates of LD from human sequencing data. They found that there was around 10 times more LD than predicted previously, due to incorrect assumptions about human demography. This was good news for the feasibility of genome-wide association studies, and indeed, modern methods for imputation, phasing, and genetic association mapping take advantage of LD in numerous ways.



Reich et al. 2001; also see Abecasis et al. 2001

Part of the reason that the typing + imputation works so well, especially for common variants, is that common variants will often have LD partners in perfect or near-perfect LD. This can be visualized using a "LocusZoom plot" for a genome-wide association study:



Notice that the "lead SNP", in purple, has dozens of LD partners with $r^2 > 0.8$. This is typical of common SNPs in general.

Measuring linkage. The *map distance* between two sites is defined as the expected number of recombination events between those sites, per generation. It is quantified in centimorgans (cM). If two sites are 1cM apart, it means that there is one recombination

event between them on average per 100 generations. The human genome is around 3000cM in length, meaning that we have 1-2 recombination events per chromosome per generation. On average, 1cM is approximately 1Mb in humans, but the recombination rate – the number of morgans per base – varies widely. Per genome, the rate of recombination and the rate of mutation are roughly equal.

Measuring linkage disequilibrium. LD is canonically measured by three metrics, called D, D', and r^2. We'll cover D and r^2, which are the covariance and the (squared) correlation.

For a pair of alleles a, b at different sites, let p_a, p_b be their allele frequencies, and let p_{ab} be the frequency of the ab haplotype. Let x_a be the random variable which is 1 if a haplotype has a, and 0 otherwise, and likewise for x_b. D is:

$$D \coloneqq p_{ab} - p_a p_b = cov(x_a, x_b).$$

If the two alleles are in linkage equilibrium (LE), segregating independently, then $p_{ab} = p_a p_b$ and $D = 0$. However, a drawback of D is that when two alleles are in perfect LD ($p_{ab} = p_a = p_b$), its value can still be small. Both D' and r^2 address this problem. r^2 is:

$$r^2 = \frac{D^2}{p_a(1-p_a)p_b(1-p_b)} = corr(x_a, x_b)^2.$$

Under Hardy-Weinberg equilibrium, r^2 is the same whether x_a and x_b are defined as haploid genotypes (0/1-valued) or diploid genotypes (0/1/2-valued). Suppose that a diploid individual has haplotypes x_a, x_b and x'_a, x'_b. Then:

$$\begin{aligned} cov(x_a + x'_a, x_b + x'_b) &= cov(x_a, x_b) + cov(x_a, x'_b) + cov(x'_a, x_b) + cov(x'_a, x'_b) \\ &= cov(x_a, x_b) + cov(x'_a, x'_b) = 2D \end{aligned}$$

and the same occurs in the denominator of the correlation, such that:

$$corr(x_a + x'_a, x_b + x'_b)^2 = r^2.$$

This relies on HWE because without it, non-independence of x and x' causes cross-terms not to drop out.

With several sites, their pairwise LD can be described using a single correlation matrix, usually denoted R. This matrix plays a central role in statistical genetics.

Accumulation of LD due to drift. If two random variables have correlation $r = 0$, and you draw n samples from their joint distribution, the sample correlation has mean and variance:

$$E(\hat{r}) = 0, \qquad var(\hat{r}) \approx \frac{1}{n}.$$

(A more accurate formula is that $var(\hat{r}) = (1 - r^2)^2/(n-2)$.) Thus, suppose that the r^2 between two alleles in a population of size N at generation t is $r_t^2 \ll 1$. In the next generation:

$$E(r_{t+1}^2) \approx r_t^2 + \frac{1}{2N}.$$

Decay of LD due to recombination. In recombining organisms, there is a competing effect: recombination breaks down haplotypes over time, at a rate that depends upon map distance. The net effect of the accumulation of new alleles and the decay of LD with time is that old haplotypes are *shorter* (in map distance) but have *greater nucleotide diversity* (in the number of segregating sites, and thus the number of LD partners, per unit of map distance).

The rate of decay in LD due to recombination is approximately exponential. Suppose that two alleles a, b start out with some covariance $D(0) = p_{ab}(0) - p_a(0)p_b(0)$ at generation 0, and they have a map distance c morgans. We will assume that $\frac{1}{N} \ll c \ll 1$: in this regime, we can ignore drift in allele frequencies (because $\frac{1}{N}$ is small) and we can also ignore the possibility of double recombination (because c is small). Thus, the change in D equals the change in p_{ab}. The change in p_{ab} is $cp_ap_b - cp_{ab}$: with rate p_ap_b, two different haplotypes unite to create an ab haplotype, and with rate cp_{ab}, an ab haplotype is split broken into two different haplotypes. (It could be broken in a way that still creates an ab haplotype, which is taken care of in the cp_ap_b term). Thus, we find that

$$E(D_1) \approx D_0(1-c)$$

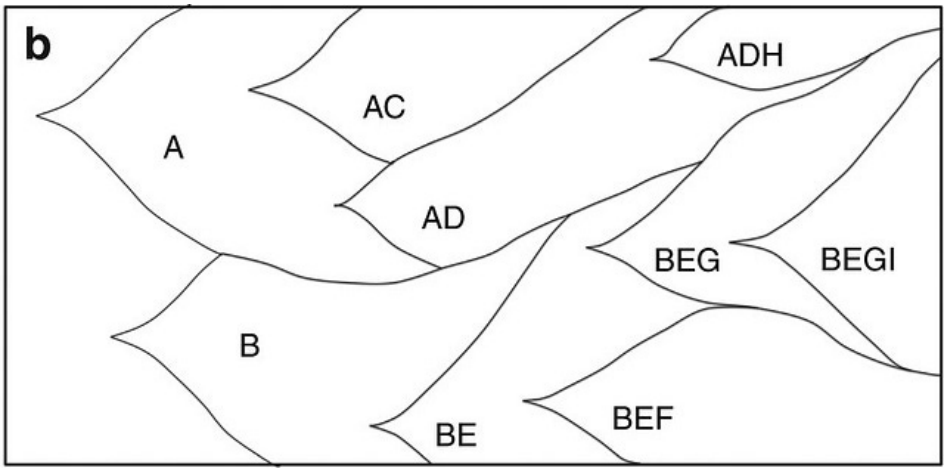
and more generally,

$$E(D_t) \approx D_0(1-c)^t.$$

Similarly:

$$E(r_t^2) \approx r_0^2(1-c)^{2t}.$$

Fixation and loss upon a specific background. Suppose an allele *b* arises as a mutation on a haplotype with allele *a*, at a perfectly linked site. Eventually, if *b* is not lost, then it will fix upon the *a* carrying haplotype. When this occurs, *a* and *b* will be in perfect LD. We can view the accumulation of LD over time, and in particular the accumulation of perfect LD partners, as the loss of *haplotype* diversity, akin to the loss of *allelic* diversity that we have seen previously. Just as allelic diversity is restored by mutation, haplotype diversity is restored by recombination.

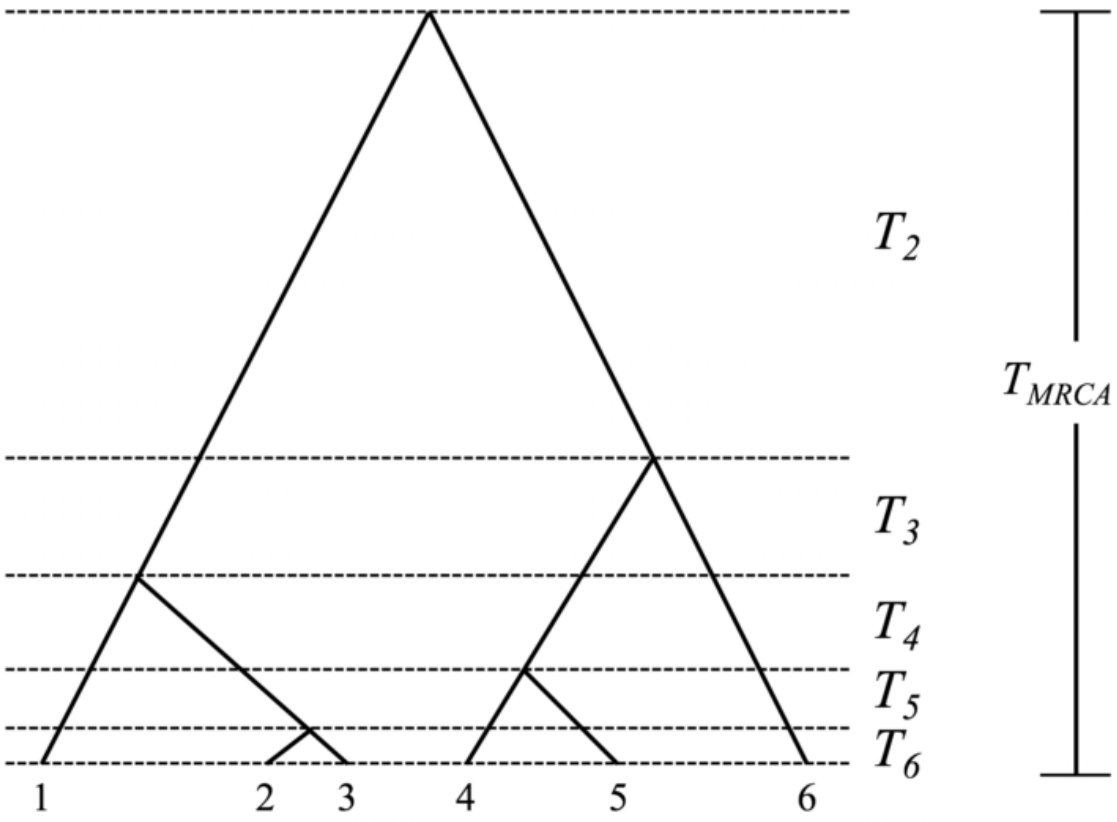


Peabody et al. 2017

LD between alleles of different frequency. In units of r^2, rare and common alleles cannot be in high LD with each other. Moreover, the vast majority of rare alleles arise on different lineages, and these too have very low LD. In summary:

- A rare and a common allele are always in low LD.
- Two rare alleles are usually in low LD, occasionally in high LD.
- Two common alleles are usually in high LD.

The coalescent. Useful perspective is often gained by analyzing the history of a contemporary population looking backward in time, and in particular, by drawing a *coalescent tree*. Suppose we sample from a contemporary population comprised of two ancestry groups, which share a recent common origin T generations ago. At a site, we say that two samples *coalesce* at time t if their most recent common ancestor at that site was t generations ago.



If there are $1 < k \leq 2N$ lineages remaining, how many generations will we have to wait until some pair of lineages coalesce? For any given pair of lineages, we will have a waiting time that follows a geometric distribution with rate $\frac{1}{2N}$. The number of pairs that might coalesce is k choose 2, which equals $k(k-1)/2$. Thus, the number of generations T_k until *some* pair coalesces is approximately:

$$T_k \sim \text{Exponential}\left(\frac{k(k-1)}{4N}\right)$$

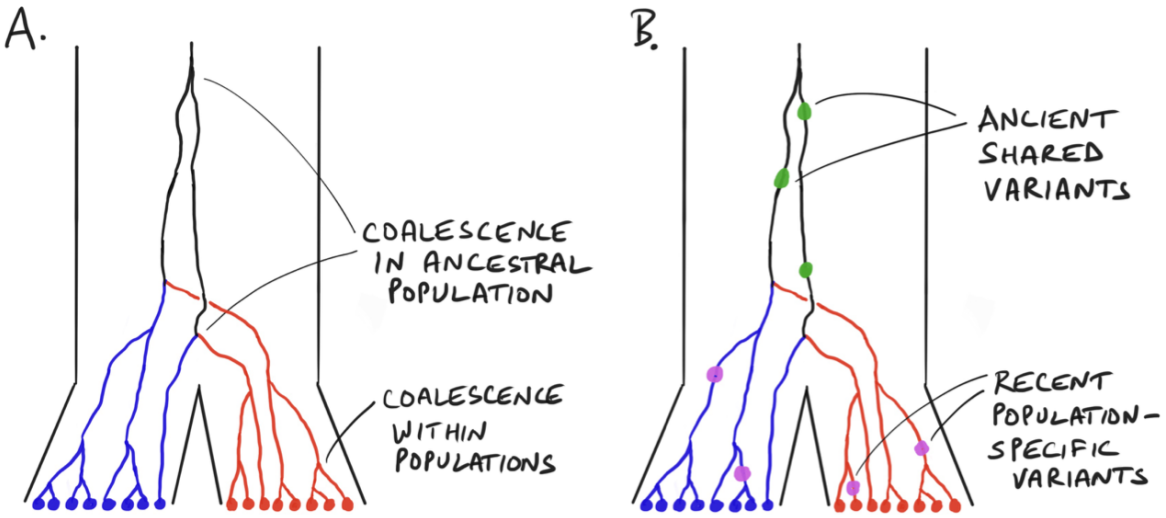
which has mean $E(T_k) = \frac{4N}{k(k-1)/2}$. We use an exponential instead of a geometric distribution here to allow $k(k-1) > 4N$. This means that we are now using a continuous-time model. It would be possible to continue using a discrete-time model, but we would have to allow multiple coalescence events to occur in the same generation.

The quadratic dependence upon k in the denominator is significant. It means that coalescence is much slower when there are fewer lineages remaining, and in fact, half of the total coalescence time is spent waiting for the last 2 lineages to coalesce.

During that time, mutations occur. Without recombination, these mutations will be in perfect LD – positive or negative – at the present time. Therefore, at any given locus, all of the most common alleles will tend to be in perfect LD with each other. Of course, they were not in perfect LD when they first arose; but over time, all of the combinations except for two of them (e.g., AB and ab) were lost, leaving them with $r^2 = 1$.

Coalescent perspective on heterozygosity. Recall that heterozygosity is defined as the probability that two haplotypes have a different allele at some site. We may calculate this using the coalescent as follows. In each generation, two haplotypes have a probability $1/2N$ of coalescing, such that their expected coalescence time is $2N$ generations. The total branch length separating them is $4N$ generations on average. Thus, if $\mu \ll \frac{1}{4N}$, the probability that a mutation occurs in between them is $4N\mu$. We previously obtained the same argument by equating the increase in heterozygosity due to mutation with the decrease due to drift; you might be able to see how this argument would generalize more easily, for example, to the case of a non-constant population size.

Coalescent perspective on F_{ST}. Consider the scenario that an ancestral source population split into two derived populations, with no migration between them, D generations ago. (D is the divergence time). For two samples from different groups, their coalescence time t will always be prior to (greater than) the divergence time. For two samples from the same group, their coalescence time might be either before or after D.

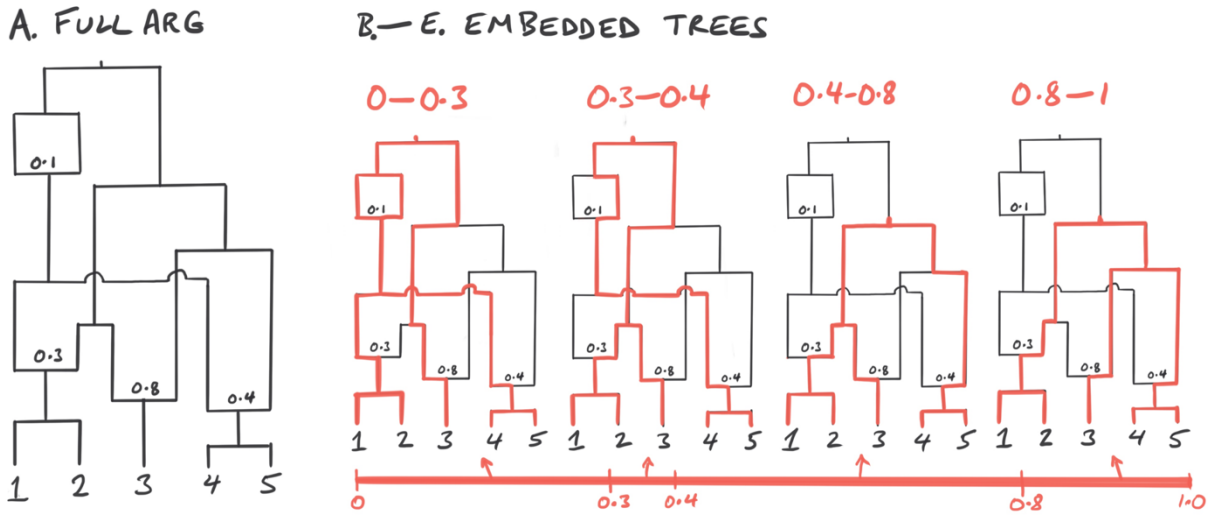


Pritchard, 2024.

More generally, let i, j be two haplotypes, let t_{ij} be their coalescence time, and let d_{ij} be their ancestral divergence time. We may view t_{ij} as having two components: $t = d + t_{\text{within}}$, where t_{within} is the coalescence time for the two samples before their populations diverged. This leads to the following interpretation of F_{ST}:

$$F_{ST} = \frac{\bar{t} - \bar{t}_{\text{within}}}{\bar{t}} = \frac{\bar{d}}{\bar{t}}.$$

The ancestral recombination graph. Recombination causes the coalescent tree to differ from one site to the next. We may view the effect of recombination as causing lineages to split apart: the same haplotype has one parent at one position, another at the next. The resulting graph is called the *ancestral recombination graph* (ARG). At any position in the genome, the subgraph of the ARG comprising nodes and edges which span that position form a coalescent tree.



Pritchard, 2024

The internal nodes of the ARG represent ancestral haplotypes. They are contiguous chunks of ancestral material that have certain descendants (possibly different descendants in different trees) and that span some set of positions.

A major topic of research in modern population genetics is the inference of ARGs and the development of applications for inferred ARGs. Inferred ARGs are useful for population genetic inference because as compared with sequence data, they are closer to the evolutionary process that we wish to infer. They can be used to scan for signals of selection and to infer demographic events. They are also potentially useful for computational reasons in statistical genetic applications.