3 Genetic Drift

In the first part of the lecture, we have described the evolutionary dynamics using a *deterministic* framework that does not allow for stochastic fluctuations of any kind. In a deterministic model, the dynamics of allele (or genotype) frequencies is governed by the expected values: mutation and recombination rates determine the expected number of mutants or recombinants, and fitness defines the expected number of surviving offspring individuals. In reality, however, the number of offspring of a given individual (and the number of mutants and recombinants) follows a distribution. Altogether, there are three possible reasons why an individual may have many or few offspring:

- *Good or bad genes*: the heritable genotype determines the distribution for the number of surviving offspring. Fitness, in particular, is the expected value of this distribution and determines the allele frequency change due to natural selection.
- *Good or bad environment*: the offspring distribution and the fitness value may also depend on non-heritable ecological factors, such as temperature or humidity. These factors can be included into a deterministic model with space- or time-dependent fitness values.
- *Good or bad luck*: the actual number of offspring, given the distribution, will depend on random factors that are not controlled by either the genes nor the external environment. This gives rise to the stochastic component in the change of allele frequencies: *random genetic drift*.

For a general evolutionary system, we can define classes of individuals according to genotypes and environmental parameters. Because of the law of large numbers, genetic drift can be ignored if and only if the number of individuals in each class tends to infinity (or if the variance of the offspring distribution is zero). Note that effects of genetic drift may be relevant even in infinite populations if the number of individuals in a focal allelic class is finite.

3.1 The Wright-Fisher model

The Wright-Fisher model (named after Sewall Wright and Ronald A. Fisher) is maybe the simplest population genetic model for genetic drift. We will introduce the model for a single locus in a haploid population of constant size N. Further assumptions are no mutation and no selection (neutral evolution) and discrete generations. The life cycle is as follows

1. Each individual in the parent generation produces an equal and very large number of gametes (or seeds). In the limit of seed number $\to \infty$, we obtain a so-called *infinite gamete pool*.
2. We sample N individuals from this gamete pool to form the offspring generation.

Sewall Wright, 1889–1988, was an American geneticist. Wright's earliest studies included investigation of the effects of inbreeding and crossbreeding among guinea pigs, animals that he later used in studying the effects of gene action on coat and eye color, among other inherited characters. His papers on inbreeding, mating systems, and genetic drift make him a principal founder of theoretical population genetics, along with R.A. Fisher and JBS Haldane. Wright's most eminent contribution to population genetics is his concept of *genetic drift* and his development of mathematical theory combining drift with the other evolutionary forces. He was also the inventor/discoverer of key concepts like the *fitness landscape* and the *inbreeding coefficient* and originated a theory to guide the use of inbreeding and crossbreeding in the improvement of livestock (adapted from Encyclopedia Britannica and Wikipedia).

Obviously, this just corresponds to *multinomial sampling with replacement* directly from the parent generation according to the rule:

- Each individual from the offspring generation picks a parent at random from the previous generation and inherits the genotype of the parent.

Remarks

- Mathematically, the probability for $k_1, \ldots, k_N$ offspring for individual number $1, \ldots, N$ in the parent generation is given by the multinomial distribution with
$$\Pr\Big[k_1, \ldots, k_N \Big| \textstyle\sum_i k_i = N\Big] = \dfrac{N!}{\prod_i k_i! N^N}. \tag{3.1}$$
- The number of offspring of a given parent individual is binomially distributed with parameters $n = N$ (number of trials) and $p = 1/N$ (success probability):
$$\Pr\big[k_1\big] = \binom{N}{k_1} \left(\frac{1}{N}\right)^{k_1} \left(1 - \frac{1}{N}\right)^{N-k_1}.$$
- Under the assumption of *random mating* (or *panmixia*), a diploid population of size N can be described by the haploid model with size $2N$, if we follow the lines of descent of all gene copies separately. Technically, we need to allow for selfing with probability $1/N$.
- The Wright-Fisher model can easily be extended to non-constant population size $N = N(t)$, simply by taking smaller or larger samples to generate the offspring generation.
- As long as the population is unstructured and evolution is neutral, the offspring distribution is invariant with respect to exchange of individuals in each generation. We can use this symmetry to disentangle the genealogies, as shown in Figure (3.3).
- Inclusion of mutation, selection, and migration (population structure) is straightforward, as shown in later sections.

Figure 3.1: The first generation in a Wright-Fisher Model of 5 diploid or 10 haploid individuals. Each of the haploids is represented by a circle.

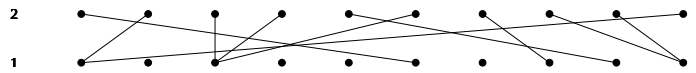

Figure 3.2: The second generation (first offspring generation) in a Wright-Fisher Model.

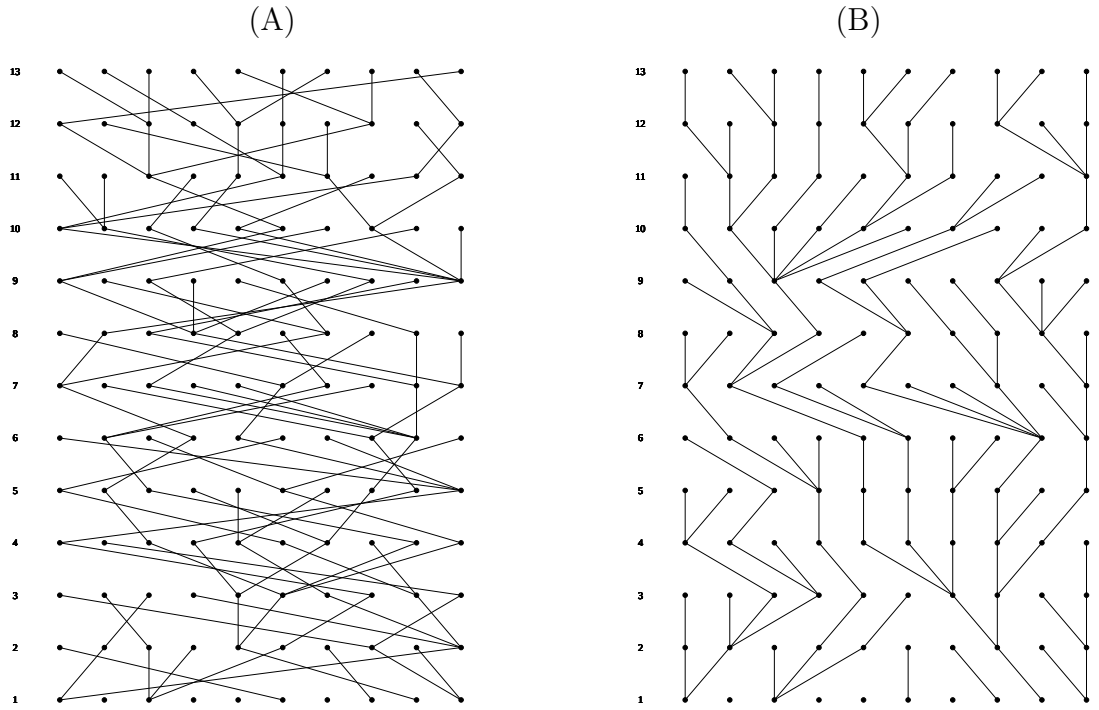

Figure 3.3: The *tangled* and *untangled version* of the Wright-Fisher Model after several generations. Both pictures show the same process, except that the individuals in the untangled version have been shuffled to avoid line crossings. The genealogical relationships are still the same, but the children of one parent are now put next to each other and close to the parent. (Web resource: www.coalescent.dk, Wright-Fisher simulator.)

3.2 Consequences of genetic drift

Genetic drift is the process of random changes in allele frequencies in populations. We will now study the effects of genetic drift quantitatively using the Wright-Fisher model. To

this end, consider a single locus with two neutral alleles a and A in a diploid population of size N. We thus have a haploid population size (= number of gene copies) of $2N$. We denote the number of A alleles in the population at generation t as n_t and its frequency as $p_t = n_t/2N$. The transition probability from state $n_t = i$ to state $n_{t+1} = j$, $0 \leq i, j, \leq 2N$ is given by

$$P_{ij} := \Pr[n_{t+1} = j | n_t = i] = \binom{2N}{j} \cdot \left(\frac{i}{2N}\right)^j \cdot \left(1 - \frac{i}{2N}\right)^{2N-j} . \tag{3.2}$$

This defines the transition matrix $\mathbf{P}$ with elements P_{ij}, $0 \leq i, j \leq 2N$, of a time-homogeneous Markov chain. If $\mathbf{x}_t$ is the probability vector (of length $2N + 1$) on the state space at generation t, we have $\mathbf{x}_{t+1} = \mathbf{x}_t\mathbf{P}$. Some elementary properties of this process are:

1. For the expected number of A alleles, we have $\mathrm{E}[n_1|n_0] = 2N \cdot \frac{n_0}{2N} = i = n_0$, and thus $\mathrm{E}[n_1] = \mathrm{E}[n_0]$ and
$$\mathrm{E}[p_t] = \mathrm{E}[p_0] .$$
The expected allele frequency is constant. The stochastic process defined by the neutral Wright-Fisher model is thus a *martingale*. This holds true, in more general, for any neutral model of *pure random drift* (no mutation and selection) in an unstructured population. We can also express this in terms of the expected change in allele frequencies as $\mathrm{E}[\delta p|p = p_0] = \mathrm{E}[p_1 - p_0] = 0$.

2. For the variance among replicate offspring populations from a founder population with frequency $p_0 = n_0/2N$ of the A allele, we obtain: $\mathrm{Var}[n_1|n_0] = 2Np_0(1 - p_0)$ and thus
$$V := \mathrm{Var}[p_1|p_0] = \frac{p_0(1 - p_0)}{2N} .$$
The variance is largest for $p_0 = 1/2$. In terms of allele frequency changes, we can also write $\mathrm{Var}[\delta p|p = p_0] = \mathrm{Var}[p_1 - p_0] = \mathrm{Var}[p_1|p_0] = V$.

3. There are two absorbing states of the process: Fixation of the A allele at $p_t = 1$, corresponding to a probability vector $\mathbf{x}^{(1)} = (0, 0, \ldots, 1)$, and loss of the allele at $p_t = 0$, corresponding to $\mathbf{x}^{(0)} = (1, 0, \ldots, 0)$. Both $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$ are left eigenvectors of the transition matrix with eigenvalue 1.

4. The absorption probabilities in $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$ for an initial frequency $p_0 = n_0/2N$ are given by the corresponding right eigenvectors, $\mathbf{y}^{(0)}$ and $\mathbf{y}^{(1)}$, with normalization $\mathbf{x}^{(i)} \cdot \mathbf{y}^{(j)} = \delta_{ij}$: If we define as π_i the fixation probability (absorption in $\mathbf{x}^{(1)}$) for a process that starts in state $p_0 = i/2N$, then $\mathbf{y}^{(1)} = (0, \pi_1, \pi_2, \ldots, \pi_{2N-1}, 1)$. Indeed, we have the single-step iteration
$$\pi_i = \sum_{j=0}^{2N} P_{ij}\pi_j$$
which is just the eigenvalue equation for $\mathbf{P}$ with eigenvalue $\lambda = 1$.

5. For a neutral process with two absorbing states, we can immediately determine the fixation probability from the martingale property of the process. Assume that we start in state $p_0 = i/2N$. Since any process will eventually be absorbed in either $\mathbf{x}^{(0)}$ or in $\mathbf{x}^{(1)}$, we have

$$\lim_{t\to\infty} \mathrm{E}[p_t] = \frac{i}{2N} = \pi_i \cdot 1 + (1-\pi_i)\cdot 0 \quad \Rightarrow \quad \pi_i = \frac{i}{2N}\,.$$

In particular, the fixation probability of a single new mutation in a population is $\pi_1 = 1/2N$.

Random genetic drift has consequences for the variance of allele frequencies among and within populations. For the variance among colonies that derive from the same ancestral founder population, we have already derived above that $V = p_0(1-p_0)/2N$ after a single generation. After a long time, we get

$$V_\infty = \lim_{t\to\infty} \Big(\mathrm{E}[(p_t)^2] - \big(\mathrm{E}[p_t]\big)^2 \Big) = p_0 - p_0^2 = p_0(1-p_0)\,.$$

The variance among populations thus increases with drift to a finite limit. To measure variance within a population, we define the homozygosity F_t and the heterozygosity H_t as follows

$$F_t = p_t^2 + (1-p_t)^2 \quad ; \quad H_t = 2p_t(1-p_t) = (1-F_t)\,.$$

The homozygosity (heterozygosity) is the probability that two randomly drawn individuals carry the same (a different) allelic state, where the same individual may be drawn twice (i.e. with replacement). We can generalize this definition for a model with k different alleles with frequencies $p_t^{(1)}, \ldots, p_t^{(k)}$ and $\sum_i p_t^{(i)} = 1$,

$$F_t = \sum_{i=1}^{k} \big(p_t^{(i)}\big)^2 = 1 - H_t\,.$$

We obtain the single-step iteration

$$F_t = \frac{1}{2N} + \Big(1 - \frac{1}{2N}\Big) F_{t-1}$$

Indeed, if we take two random alleles (with replacement) from the population in generation t, the probability that we have picked the same allele twice is $1/2N$. If this is not the case, we choose parents for both alleles in the previous generation $t-1$. By definition, the probability that these parents carry the same state is F_{t-1}. From this we get for the heterozygosity

$$H_t = \Big(1 - \frac{1}{2N}\Big) H_{t-1} = \Big(1 - \frac{1}{2N}\Big)^t H_0 \approx H_0 \exp[-t/2N]\,.$$

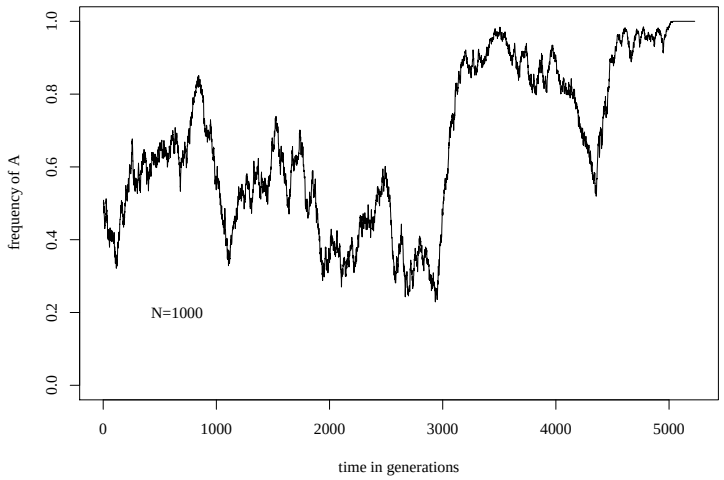

Figure 3.4: Frequency curve of one allele in a Wright-Fisher Model. Population size is $2N = 2000$ and time is given in generations. The initial frequency is 0.5.

We see that drift reduces variability within a population and $H_t \to 0$ as $t \to \infty$. The characteristic time for approaching a monomorphic state is given by the (haploid) population size. We can derive the half time for H_t as follows

$$\frac{H_t}{H_0} = \left(1 - \frac{1}{2N}\right)^{t_{1/2}} \approx \exp[-t_{1/2}/2N] := \frac{1}{2}$$

and thus

$$t_{1/2} = 2N \log[2] \approx 1.39N\,.$$

The half time scales with the population size. Note that the time scale to approach a monomorphic state does not depend on the number of alleles that are initially present.

Note finally that heterozygosity and homozygosity (as defined here) should not be confused with the frequency of heterozygotes and homozygotes in a population. Both quantities only coincide under the assumption of random mating. For this reason, some authors (e.g. Charlesworth and Charlesworth 2010) prefer the terms *gene diversity* for H_t and *identity by descent* for F_t.

Exercises

1. We have defined homozygosity and heterozygosity by drawing individuals with replacement. How do the formulas look like if we define these quantities without replacement (which is sometimes also done in the literature)?

2. Consider the neutral Wright-Fisher model with a variable population size. What is then the fixation probability of a new mutant that arises in generation 1?

4 Neutral theory

In a pure drift model, genetic variation within a population can only be eliminated, but never created. To obtain even the most basic model for evolution, we need to include mutation as the ultimate source for new variation. Just these two evolutionary forces, mutation and drift, are the only ingredients of the so-called *neutral theory*, developed by Motoo Kimura in the 50s and 60s. Kimura famously pointed out that models without selection already explain much of the observed patterns of polymorphism within species and divergence between species. Importantly, Kimura did not claim that selection is not important for evolution. It is obvious that purifying selection is responsible for the maintenance of functional important parts of the genome (e.g. in coding regions). However, Kimura claimed that most differences that we see within and among populations are not influenced by selection. Today, selection is thought to play an important role also for these questions. However, the neutral theory is the standard null-model of population genetics. This means, if we want to make the case for selection, we usually do so by rejecting the neutral hypothesis. This makes understanding of neutral evolution key to all of population genetics.

Motoo Kimura, 1924–1994, published several important, highly mathematical papers on random genetic drift that impressed the few population geneticists who were able to understand them (most notably, Wright). In one paper, he extended Fisher's theory of natural selection to take into account factors such as dominance, epistasis and fluctuations in the natural environment. He set out to develop ways to use the new data pouring in from molecular biology to solve problems of population genetics. Using data on the variation among hemoglobins and cytochromes-c in a wide range of species, he calculated the evolutionary rates of these proteins. Extrapolating these rates to the entire genome, he concluded that there could not be strong enough selection pressures to drive such rapid evolution. He therefore decided that most evolution at the molecular level was the result of neutral processes like mutation and drift. Kimura spent the rest of his life advancing this idea, which came to be known as the "neutral theory of molecular evolution" (adapted from `http://hrst.mit.edu/groups/evolution`.)

4.1 Mutation schemes

There are three widely used schemes to introduce (point) mutations to a model of molecular evolution:

1. With a finite number of alleles, we can define transition probabilities from any allelic state to any other state. For example, there may be k different alleles A_i, $i = 1, \ldots, k$ at a single locus and a mutation probability from A_i to A_j given by μ_{ij}. Then $\mu_i = \sum_{j \neq i} \mu_{ij}$ is the total mutation rate per generation in state A_i. Mutation according to this scheme is most easily included into the Wright-Fisher model as an additional

step on the level of the infinite gamete pool,

$$\mathbf{p}_t \to \mathbf{p}'_{t+1} = \mathbf{p}_t \cdot \mathbf{U}$$

where $\mathbf{p}_t$ is the (row) vector of allele frequencies and the mutation matrix $\mathbf{U}$ has elements μ_{ij} for $i \neq j$ and $\mu_{ii} \equiv 1 - \mu_i$. We then obtain the frequencies in the next generations $\mathbf{p}_{t+1}$ from $\mathbf{p}'_{t+1}$ by multinomial sampling as in the model without mutation.

2. If we take a whole gene as our locus, we get a very large number of possible alleles if we distinguish different amino acid sequences. In particular, back mutation to an ancestral allelic state becomes very unlikely. In this case, it makes sense to assume an effectively infinite number of alleles in an evolutionary model,

$$A_1 \to A_2 \to A_3 \to \ldots$$

Usually, a uniform mutation rate u from one allelic state to the next is assumed. Formally, the *infinite alleles model* corresponds to a Markov chain with an infinite state space.

3. In the infinite alleles model, we assume that the latest mutation erases all the memory of the previous state. Only the latest state is visible. However, for a stretch of DNA, point mutation rates at a single site (or nucleotide position) are very small. We can thus assume that subsequent point mutations will always happen at different sites and remain visible. This leads to the so-called *infinite sites model* for mutation that is widely applied in molecular evolution. In particular, under the assumptions of the infinite sites model (no "double hits"), we can count the number of mutations that have occurred in a sequenced region – given that we have information about the ancestral sequence.

4.2 Predictions from neutral theory

We can easily derive several elementary consequences of neutral theory, given one of the mutation schemes above.

- Under the infinite sites model, new mutations enter a population at a constant rate $2Nu$, where u is the mutation rate per generation and per individual for the locus (stretch of DNA sequence) under consideration. Since any new mutation has a fixation probability of $1/(2N)$, we obtain a neutral substitution rate of

$$k = 2Nu \cdot \frac{1}{2N} = u\,.$$

Importantly, the rate of neutral evolution is independent of the population size and also holds if $N = N(t)$ changes across generations. As long as the mutation rate u can be assumed to be constant, neutral substitutions occur constant in time. They define a so-called *molecular clock*, which can be used for molecular dating of phylogenetic events.

- For the evolution of the homozygosity F_t or heterozygosity H_t under mutation and drift, we obtain for the infinite alleles model or the infinite sites model

$$F_t = 1 - H_t = (1-u)^2\Big(1 - \Big(1 - \frac{1}{2N}\Big)H_{t-1}\Big)\,.$$

  In the long term, the population will approach a state where both forces, mutation and drift balance. We thus reach an equilibrium, $H_t = H_{t-1} = H$, with

$$H = \frac{1-(1-u)^2}{1-(1-u)^2(1-1/2N)} = \frac{\Theta(1-u/2)}{\Theta(1-u/2)+(1-u)^2} \approx \frac{\Theta}{\Theta+1}$$

  where $\Theta = 4Nu$ is the population mutation parameter. In the case with a finite number of alleles, we need to account for cases where one allelic state can be produced by multiple mutations (i.e., F_t measures the identity in state rather than just the identity by descent). For two alleles with symmetric mutation at rate u in both directions,

$$1 - H_t = (1-2u)\Big(1 - \Big(1 - \frac{1}{2N}\Big)H_{t-1}\Big) + 2u\Big(1 - \frac{1}{2N}\Big)H_{t-1}$$

  and thus

$$H = \frac{\Theta}{2\Theta + 1 - 4u} \approx \frac{\Theta}{2\Theta+1}\,.$$

- For the special case of the expected *nucleotide diversity*, denoted as $\mathrm{E}[\pi]$, where the focus is on a single nucleotide site, we usually have $\Theta \ll 1$. We can then further approximate

$$\mathrm{E}[\pi] = H_{\text{nucleotide}} \approx \Theta\,,$$

  independently of the mutational scheme that is used.