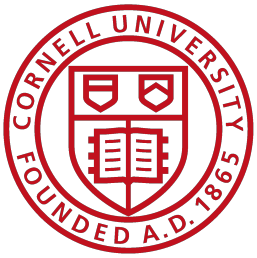
Population Genetics

BIOMG 4810 / BTRY 4810

Philipp W. Messer

November 22, 2016



Department of Biological Statistics
& Computational Biology
Cornell University
102J Weill Hall
Ithaca, NY 14853
phone: 607-255-3984
messer@cornell.edu

Contents

Chapter 1

Probability theory

1.1 Sample space and probability

Consider a random experiment that can produce a number of discrete outcomes (e.g. rolling a dice). We call the set Ω of all possible outcomes the sample space (e.g. $\Omega = \{1, 2, 3, 4, 5, 6\}$ in the dice example). We further assume that there is a function that attaches a value $f(x)$ to each element $x \in \Omega$ and satisfies:

$$\text{(i)} \quad 0 \leq f(x) \leq 1 \quad \text{and} \quad \text{(ii)} \quad \sum_{x \in \Omega} f(x) = 1. \tag{1.1}$$

The function $f(x)$ is called the probability mass function (PMF). Subsets $A \subset \Omega$ are called events (e.g. the specific number $A = \{3\}$ in the dice example, or the subset $A = \{2, 4, 6\}$ of all even numbers of a dice). Events that contain more than one element from the sample space are interpreted in the sense that a single outcome of the random experiment results in any of the elements from this set (e.g. throwing a dice yields any even number in the dice example). The probability of any given event A is then defined as:

$$\Pr(A) = \sum_{x \in A} f(x). \tag{1.2}$$

For two events, A and B, the probability of either A or B occurring is:

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B), \tag{1.3}$$

where $\Pr(A \text{ and } B)$ denotes the joint probability that both A and B occur (defined by the probability of the intersection of sets A and B). A and B are said to be mutually exclusive if their joint probability is zero,

$$\Pr(A \text{ and } B) = 0. \tag{1.4}$$

Two events are said to be independent if their joint probability is the product of their individual probabilities,

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B). \tag{1.5}$$

The conditional probability $\Pr(A \,|\, B)$ is the probability of A, given the occurrence of B. It holds that:

$$\Pr(A \text{ and } B) = \Pr(B \,|\, A)\Pr(A) = \Pr(A \,|\, B)\Pr(B). \tag{1.6}$$

Rearrangement of Equation (1.6) yields Bayes' theorem:

$$\Pr(A \,|\, B) = \frac{\Pr(B \,|\, A)\Pr(A)}{\Pr(B)}. \tag{1.7}$$

Example: rolling a dice

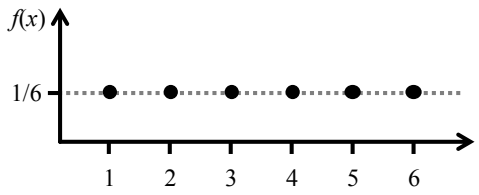


Figure 1.1: $\Omega = \{1, 2, 3, 4, 5, 6\}$. We define $f(x) = 1/6$ for all $x \in \Omega$ (fair dice). Let event $A = \{4\}$ denote rolling the number 4, then $\Pr(A) = 1/6$. Define $B = \{1, 3, 5\}$, i.e., rolling an uneven number. Then $\Pr(B) = f(1)+f(3)+f(5) = 1/2$. Events A and B are mutually exclusive, $\Pr(A \text{ and } B) = 0$ (the intersection of sets A and B is empty). Further, $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) = 1/6 + 1/2 = 2/3$. Let's define event $C = \{1, 2\}$, then $\Pr(C) = 1/3$ and $\Pr(B \,|\, C) = 1/2$.

1.2 Random variables

A random variable (RV) is a numeric variable (X) that describes the outcome of a random experiment, where each value can be associated with a probability that the experiment yields the specific value as its outcome. Discrete RVs are defined by the PMF specifying these probabilities, $\Pr(X = x) = f(x)$.

Expectation value

The expectation value of a discrete RV is defined by:

$$E[X] = \sum_{x \in \Omega} x \Pr(X = x). \tag{1.8}$$

This expectation value can be interpreted as the mean of X, estimated over an infinite number of trials. In our dice example, $E[X] = 1 \times 1/6 + 2 \times /1/6 + \cdots + 6 \times 1/6 = 3.5$.

1. For any two RVs, X and Y, it holds that:

$$E[X + Y] = E[X] + E[Y]. \tag{1.9}$$

2. Any function $g(X)$ of X defines again a random variable, and it holds that:

$$E[g(X)] = \sum_{x \in \Omega} g(x) \Pr(x = X). \tag{1.10}$$

Variance

The variance of a discrete RV is the expectation value of the squared deviation from its mean:

$$V[X] = E[(X - E[X])^2]. \tag{1.11}$$

1. The variance can be related to the mean of the square of the RV:

$$V[X] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - E[X]^2. \tag{1.12}$$

2. For two independent RVs, X and Y, the variance of their sum equals:

$$V[X+Y] = V[X] + V[Y]. \tag{1.13}$$

3. The variance of a linear function of a random variable is given by:

$$V[a+bX] = b^2 V[X]. \tag{1.14}$$

The Bernoulli RV

The Bernoulli RV describes an experiment with only two possible outcomes, $\Omega = \{0, 1\}$: 'success' $(X = 1)$ occurs with probability p, whereas 'failure' $(X = 0)$ occurs with probability $1 - p$. The classical example is the toss of a coin with $p = 1/2$ for a fair coin. The PMF of a Bernoulli RV is given by:

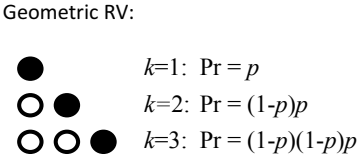
$$\Pr(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}, \quad 0 \leq p \leq 1. \tag{1.15}$$

Expectation value: $E[X] = \sum_x xP(X = x) = 1 \times p + 0 \times (1-p) = p$. Variance: $E[X^2] = 1^2 \times p + 0^2 \times (1-p) = p \Rightarrow V[X] = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$.

The Geometric RV

The geometric RV describes the number of trials until the first success is observed in a series of independent Bernoulli trails with individual success probability p. The sample space of a geometric RV is $\Omega = \{1, 2, \dots\}$ and its PMF is given by:

$$\Pr(X = k) = p(1-p)^{k-1}, \quad 0 \leq p \leq 1. \tag{1.16}$$



Expectation value: $E[X] = 1/p$. Variance: $V[X] = 1/p^2$.

The Binomial RV

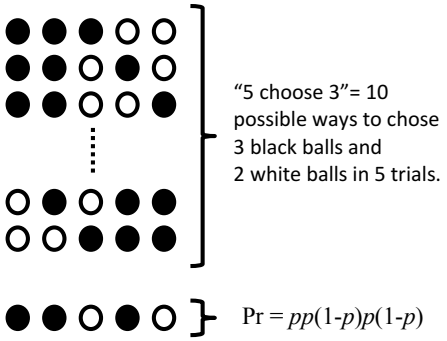
A binomial RV describes the sum of n independent Bernoulli RVs (i.e., the number of successes, since each success adds one to the sum). The sample space is $\Omega = \{0, 1, \cdots, n\}$ and the PMF is given by:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq p \leq 1. \tag{1.17}$$

The binomial coefficient "n choose k" specifies the number of different configurations in which k successes can occur in a total of n trials. Formally:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad 0 \leq k \leq n. \tag{1.18}$$

The notation $n!$ refers to the "factorial number", defined as $n! = n \times (n-1) \times \cdots \times 1$. For example, "3 choose 2" yields $(3 \times 2 \times 1)/(2 \times 1 \times 1) = 3$. Indeed, there are three possible configurations in which two successes can occur in three trials $\{1,1,0\}, \{1,0,1\}, \{0,1,1\}$.



The expectation value of a Binomial RV can be directly obtained from the fact that it is the sum of n Bernoulli RVs, $X = Y_1 + Y_2 + \cdots + Y_n$. Using Equation (1.9), we obtain:

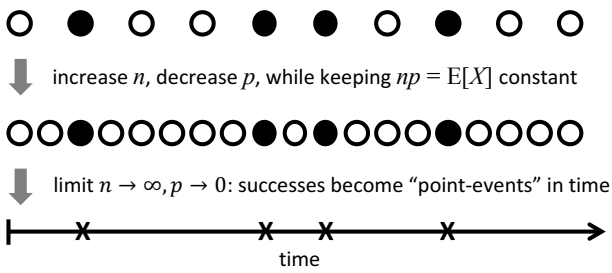
$$E[X] = E[Y_1 + Y_2 \cdots + Y_n] = nE[Y_i] = np. \tag{1.19}$$

Similarly, since the individual Bernoulli RVs are independent of each other, we can use Equation (1.13) to calculate the variance of a Binomial RV, yielding:

$$V[X] = nV[Y_i] = np(1-p). \tag{1.20}$$

The Poisson point process

The Poisson process describes events that occur in time and can be interpreted as a continuous version of the Bernoulli process. Consider a sequence of n independent Bernoulli trials, each with success probability p. The expected overall number of successes across all n trials is the expectation value of a Binomial RV: $E[X] = np$. Let's consider a scenario in which we perform, over the same time interval, twice as many Bernoulli trials: $n \to 2n$. At the same time, however, we half the success probability of each individual trial: $p \to p/2$. The expected overall number of successes remains the same in this case: $E[X] = 2np/2 = np$. We can push this approach further and let the number of individual Bernoulli trials go to infinity ($n \to \infty$), while the success probability for each individual trial goes to zero ($p \to 0$), such that the product $np = E[X]$ remains constant. This defines a Poisson process, in which successes become "events" that occur at specific points in time.



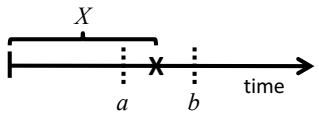
Poisson processes are specified by a single number, the rate parameter $\lambda = E[X]/\text{time} > 0$, which defines how many events are expected to occur, on average, per unit of time (e.g. seconds, years, generations). A key property of the Poisson process is that it is "memoryless". This means that the number of events that occur

in non-overlapping intervals are independent RVs. In particular, the probability that an event occurs during a given interval does not depend on whether an event has just occurred in the preceding interval. The Poisson process is widely used for describing natural phenomena that occur as independent events over time, such as impacts of large meteorites on earth, radioactive decay, or mutations occurring in a genome.

The Exponential RV

The exponential RV describes the waiting time until the first occurrence in a Poisson process with rate λ. Its sample space can therefore be any positive real number: $\Omega = \mathbb{R}^+$. This situation is different from the Bernoulli, geometric, and binomial RVs we discussed so far, which were all defined over a discrete sample space. In contrast to these discrete RVs, the exponential RV is a continuous variable. Unfortunately our standard approach of assigning probabilities to each particular value of the RV using PMFs does no longer work for such continuous RVs, because the probability of any given exact value would always be infinitesimally small. Instead, we will define such continuous RVs in terms of a so-called probability density function (PDF). For a continuous RV with PDF $f(x)$, the probability of observing a value somewhere in the interval between a and b is then given by the integral:

$$\Pr(a \leq X < b) = \int_a^b f(x)dx. \tag{1.21}$$



The exponential RV for a Poisson process with rate λ is specified by the PDF:

$$f(x) = \lambda e^{-\lambda x}, \tag{1.22}$$

which can be derived mathematically from our above definition of the Poisson process. According to Equation (1.21), the probabilities that the first event occurs prior to time t, or after time t, respectively, are then:

$$\Pr(X < t) = \int_0^t f(x)dx = 1 - e^{-\lambda t} \quad \text{and} \quad \Pr(X > t) = \int_t^\infty f(x)dx = e^{-\lambda t}. \tag{1.23}$$

Note that due to the "memoryless" property of the Poisson process, the exponential RV also describes the waiting time between two events in a Poisson process (we can just set the time to zero when the first event has occurred, and then ask how long we have to wait until the next event occurs). The mean of an exponential RV with rate parameter λ is given by $E[X] = \int_0^\infty x f(x)dx = 1/\lambda$. This makes sense: the average waiting time between two events is the inverse of the rate at which events occur. The variance is $V[X] = 1/\lambda^2$.

Competing Poisson processes

Consider two independent Poisson processes. Process A occurs at rate α and process B occurs at rate β. Assume that both processes start at the same time. What is the probability that the first event that occurs is an event from the A process, rather than from the B process?

$$\Pr[\,\text{A event occurs first}\,] = \int_0^\infty \Pr[X_\beta > t]\, f_\alpha(t)\, dt = \int_0^\infty e^{-\beta t} \alpha\, e^{-\alpha t} dt = \frac{\alpha}{\alpha + \beta}. \tag{1.24}$$

The probability that the first event is from the A process is thus simply the relative rate of that process over the combined rate of both processes. This result makes intuitive sense: if you consider two processes that occur at equal rates, $\alpha = \beta$, both processes will have equal probability $1/2$ to produce the first event.

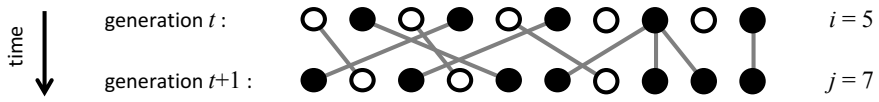
Chapter 2

Genetic drift

2.1 The Wright-Fisher model

In a finite population, the population-frequency of an allele may change over time merely due to stochastic variation in the reproductive success among individuals. We call this effect random genetic drift (RGD). For studying drift in statistically meaningful ways, we rely on idealized models that capture key biological aspects of the process while remaining mathematically tractable. One of the most commonly used model for this purpose is named after two of the founders of population genetics, Sewall Wright and Ronald Fisher.

The WF model describes a population with discrete, non-overlapping generations. In each generation the entire population is replaced by the offspring from the previous generation. Parents are chosen via random sampling with replacement. Consider a locus with two alleles (e.g. "black" and "white") in a haploid population of size N evolving under this model. Let i denote the number of individuals that carry the black allele in generation t.



When randomly picking a parent from generation t for an individual in generation $t+1$, the probability that this parent has a black allele is $p = i/N$. We pick a parent this way for every individual in generation $t+1$. The resulting number of individuals that carry a black allele in generation $t+1$ is therefore a binomial random variable. Its PMF is given by:

$$\Pr(j) = \binom{N}{j} p^j (1-p)^{N-j} \qquad 0 \leq i, j \leq N. \tag{2.1}$$

Properties of the WF model

- The WF model is a so-called Markov process (a stochastic process for which the transition probabilities to the next state are determined solely by the present state).
- Expected allele counts remain constant across generations: $E[j] = np = N(i/N) = i$.
- The variance between two generations is $V[j] = np(1-p) = N(i/N)(1-i/N) = i(1-i/N)$.
- It is straightforward to incorporate non-constant population sizes $N(t)$ into the WF model by taking larger or smaller samples in each generation.

- Rather than absolute allele counts, we are often more interested in the population frequency of an allele, $x_t = i/N$. In that case:

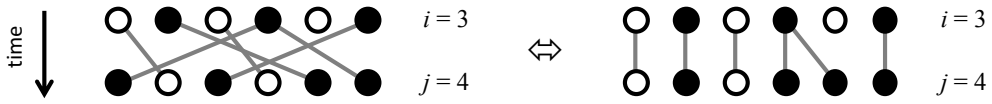
$$E[x_{t+1}] = E\left[\frac{j}{N}\right] = \frac{1}{N}E[j] = \frac{i}{N} = x_t, \tag{2.2}$$

$$V[x_{t+1}] = V\left[\frac{j}{N}\right] = \frac{1}{N^2}V[j] = \frac{i}{N^2}\left(1 - \frac{i}{N}\right) = \frac{x_t(1-x_t)}{N}. \tag{2.3}$$

  RGD does not actually change the expected frequency of an allele. To understand this intuitively, consider a black allele that is initially present in exactly half of the population, $x_t = 0.5$. If drift were to systematically increase the frequency of the black allele over time, it would also systematically decrease the frequency of the white allele. Yet both alleles are equivalent in terms of reproductive success. The expectation value of each allele's frequency thus has to remain constant.

  However, allele-frequencies are still expected to fluctuate between generations, there is just no systematic bias towards either direction. According to Equation (2.3), these fluctuations will be largest when the allele is at frequency $x_t = 0.5$. Furthermore, fluctuations will be larger in small populations than in large populations. This inverse relationship between population size N and magnitude of allele-frequency fluctuations between generations is a fundamental aspect of RGD.

- The actual order of individuals in each generation of a WF model does not matter and we can entangle the parent-offspring relationships by reshuffling individuals in generation $t+1$ to avoid line crossings. The following two outcomes are equivalent with respect to their genealogical relationships:

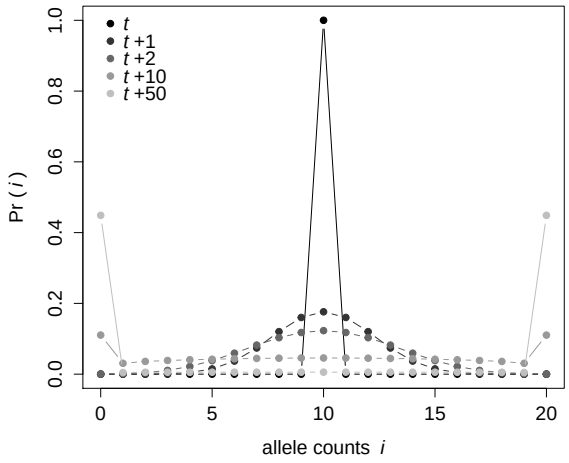


- Given the outcome for generation $t+1$, we can progress the WF model to generation $t+2$. Let i, j, k denote the numbers of black alleles in generations $t, t+1, t+2$, respectively. The transition probability from generation t to generation $t+2$ is then given by:

$$\Pr(k \mid i) = \sum_{j \in \{0,\cdots,N\}} \binom{N}{k} p^k (1-p)^{N-k} \Pr(j \mid i), \quad p = j/N. \tag{2.4}$$

- In principle, we can calculate transition probabilities over longer time intervals through iterative summation over all possible intermediate states, each iteration adding another summation to Equation (2.4). However, these calculations do not yield analytical expressions that are easy to work with. Fortunately there is a much more elegant approach that will allow us to obtain approximative transition probabilities over longer time intervals using the so-called diffusion approximation. We will touch this topic later in the class when we will incorporate selection into the model.

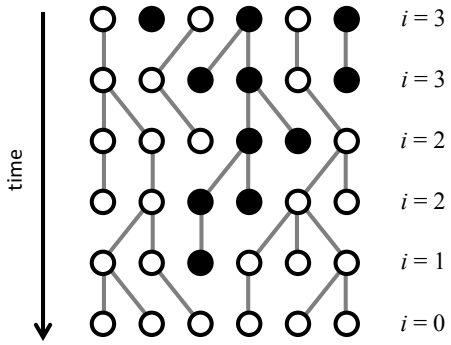
  For now, we content ourselves with the fact that we can easily go through a number of these iterations with the help of a computer. For example, the plot below shows the resulting probabilities to observe i counts of the black allele after $1, 2, 10, 50$ generations in a WF model with $N = 20$ individuals, which started with the black allele being present in 10 individuals initially:



Note that these transition probabilities describe the probabilities of observing a particular outcome of the experiments if we were to run the experiment many times. We could measure such distributions by counting how often we observe a particular outcome among a very large number of experiments. Each particular experiment, of course, has a specific number of black alleles in each generation, for example 13 black alleles in generation $t + 2$. In that case, going to the next generation $(t + 3)$ then involves drawing a random number from the binomial distribution specified by Equation (2.1), which yields again a specific number, e.g. 9 black alleles, and so on.

2.2 Fixation and loss of alleles

In each generation, there is a certain probability that the frequency of the black allele will be different from that in the previous generation. Over time, these changes will accumulate, such that the allele will either become fixed or lost. We call fixation $(x = 1)$ and loss $(x = 0)$ the two absorbing states of the model, because once the allele has reached one of these states, it remains there perpetually (we are not considering mutations yet).



In the WF model, the probability that an allele eventually becomes fixed is simply its initial frequency,

$$\Pr(\text{fixation} \,|\, x_0) = x_0 \quad \text{and thus} \quad \Pr(\text{loss} \,|\, x_0) = 1 - x_0. \tag{2.5}$$

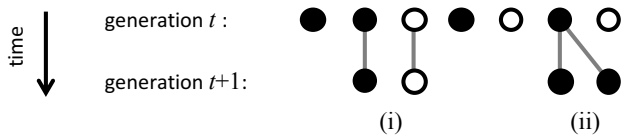
In particular, the fixation probability of an allele initially present in a single copy will be $1/N$. We can prove Equation (2.5) as follows: We already showed that the expectation value of the allele-frequency does not change between generations, $E[x_1] = x_0$. By recursion, this has to hold true for all times $t > 0$, so it

will still be true in the limit of very long times: $\lim_{t\to\infty} E[x_t] = x_0$. However, we know that after a very long time the allele must have either become fixed or lost. The state space of x_t in the limit of very large times therefore has only two values, $\Omega = \{0, 1\}$. The probabilities corresponding to these two state are the probabilities of loss and fixation, respectively. Using the definition of the expectation value, we obtain: $x_0 = \lim_{t\to\infty} E[x_t] = \sum_{x\in\Omega} x\Pr(x) = 0 \times \Pr(\text{loss} \,|\, x_0) + 1 \times \Pr(\text{fixation} \,|\, x_0) = \Pr(\text{fixation} \,|\, x_0)$.

2.3 Decay of heterozygosity

As a consequence of random genetic drift, every allele in the WF model will ultimately become either fixed or lost in the population. Polymorphic loci will thus tend to become monomorphic over time, thereby systematically reducing levels of diversity in the population. As a quantitative measure for the level of diversity at a given locus, we will define heterozygosity (H) to be the probability that two randomly drawn alleles from the population are "different by state" (e.g. one is black while the other is white). In the following, we will calculate how heterozygosity decays over time in the WF model.

Consider a locus with two different alleles (e.g. black and white) in a haploid population of size N. Assume that in generation t we have heterozygosity H_t. In order to calculate the expected heterozygosity in the next generation, H_{t+1}, we will distinguish two possibilities: (i) We could pick two alleles that happen to have different parents in generation t, in which case the probability that the two alleles are different by state, by definition, is H_t; (ii) Alternatively, we could pick two individuals that happen to have the same parent in generation t. In this case, the two alleles will always be identical by state.



The probability of picking two individuals that originate from the same parent is $1/N$ in our WF population of size N, while the probability that they have different parents is $1 - 1/N$. This yields the recursion:

$$H_{t+1} = \frac{1}{N} \times 0 + \left(1 - \frac{1}{N}\right) \times H_t. \tag{2.6}$$

We can apply this recursion iteratively over many generations. If we start with heterozygosity H_0 in generation zero, after t generations we will have:

$$H_t = \left(1 - \frac{1}{N}\right)^t H_0 \approx e^{-t/N} H_0. \tag{2.7}$$

This result shows that genetic diversity is lost in the WF model at an exponential rate. The key parameter determining the rate of loss is again the inverse of the population size N. In a small population, diversity will be lost much faster than in a large population.

2.4 Diploid populations

Up to this point we have only discussed asexual populations of haploid individuals, which had some rather convenient implications for the WF model: each individual carries only a single allele at any given locus, inherited from a single parent in the previous generation. However, many species, including most animals, are diploid species that reproduce sexually. In this case, each individual carries a set of two alleles at each locus, one allele inherited from each of its two parents. In the following we will study how sexual reproduction and diploidy affect allele frequencies and how this can be incorporated into the WF model of random genetic drift.

Punnett squares

Consider a single locus with two alleles (A and a) in a sexually reproducing, diploid population. An individual can be one of three different genotypes at this locus: AA, Aa, and aa (we assume that there is no order among the two alleles in individuals, so that Aa and aA can be considered the same genotype). Individuals with genotypes AA and aa are called homozygotes, Aa individuals are called heterozygotes. We can summarize all possible outcomes of matings between two parents in such a population using a so-called Punnett square. The rows in the square indicate the three possible genotypes of the first parent, the columns those of the second parent (we do not yet worry about whether individuals have distinct sexes or not). The cells in the Punnet square then specify the possible child genotypes that can result from a mating between parents with the respective row and column genotypes, together with their relative probabilities.

Let us assume that the alleles that are transmitted to the child are chosen randomly from each parent (random transmission). For example, a parent with genotype Aa should have equal probability $1/2$ of transmitting the A allele or the a allele to its child. This scenario is described by the following Punnet square:

	AA	Aa	aa
AA	AA 1	$AA : Aa$ 1/2 : 1/2	Aa 1
Aa	$AA : Aa$ 1/2 : 1/2	$AA : Aa : aa$ 1/4 : 1/2 : 1/4	$Aa : aa$ 1/2 : 1/2
aa	Aa 1	$Aa : aa$ 1/2 : 1/2	aa 1

Note that by adjusting the probabilities in the cells we could easily incorporate non-random segregation of alleles into this framework, for instance if the A allele were a transmission distorter, such that Aa heterozygotes would transmit the A allele to their offspring with a probability of, say, 0.9, while the a would be transmitted only with probability 0.2.

2.5 Hardy-Weinberg equilibrium

Punnet squares allow us to calculate the expectation values of genotype frequencies, x_{AA}, x_{Aa}, and x_{aa} in a sexually-reproducing diploid population across generations. We will make several additional assumptions:

- Reproduction occurs via random mating.
- If there are distinct sexes, genotype frequencies are equal in all sexes.
- The population is very large, such that drift can be effectively neglected.
- There is no mutation, selection, or population structure.

Under these assumption, the probability that a randomly chosen individual in generation $t+1$ is the result of a mating between parents of two given genotypes is simply the product of the population frequencies of these two genotypes in generation t. For example, the probability that a randomly chosen individual in generation $t+1$ stems from two parents that both have genotype AA, is simply $x_{AA}(t)^2$, and so on. In order to calculate the expectation values of the different genotype frequencies in generation $t+1$, given their frequencies in generation t, we then need to sum over all possible matings that can generate the given genotype, multiplied

by the respective probabilities that such matings occurs, and the probabilities that they will actually produce an offspring of the given genotype. For AA homozygotes, we obtain:

$$\begin{aligned} E[x_{AA}(t+1)] &= x_{AA}^2(t) + \frac{x_{AA}(t)x_{Aa}(t)}{2} + \frac{x_{Aa}(t)x_{AA}(t)}{2} + \frac{x_{Aa}^2(t)}{4} \\ &= \left[x_{AA}(t) + \frac{x_{Aa}(t)}{2}\right]^2 \\ &= x_A^2(t). \end{aligned} \tag{2.8}$$

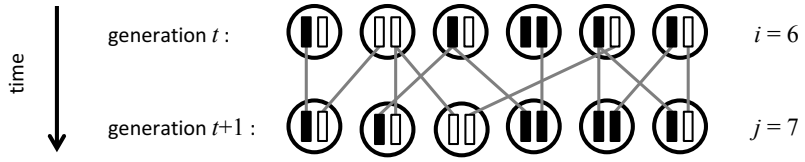
Here $x_A(t)$ denotes the overall frequency of the A allele in the population in generation t. Due to symmetry, $E[x_{aa}(t+1)]$ can be obtained by simply exchanging $A \leftrightarrow a$. The frequency of heterozygotes then follows from the condition that all genotype frequencies have to add up to 1. Combining these equations, we obtain $E[x_A(t+1)] = x_A(t)$, consistent with a haploid WF model. Genotype frequencies, on the other hand, can change between generations, depending on their initial values. For example, consider a population that initially consists only of homozygotes at equal frequency: $x_{AA}(t) = x_{aa}(t) = 1/2$. In this case: $E[x_{AA}(t+1)] = x_{AA}^2(t) = 1/4$. Due to symmetry, we also have $E[x_{aa}(t+1)] = 1/4$, and therefore $E[x_{Aa}(t+1) = 1 - E[x_{AA}(t+1)] - E[x_{aa}(t+1)] = 1/2$. Thus, after a single round of random mating, half of the population will be made up of heterozygotes. Note that in the next generation, $t+2$, the expectation values of genotype frequencies will then no longer change in the model, which can be shown by explicitly calculating the expectation values. This is an example of so-called Hardy-Weinberg equilibrium (HWE), which connects allele frequencies with genotype frequencies in a diploid population via the following relations:

$$\begin{aligned} x_{AA} &= x_A^2 \\ x_{Aa} &= 2x_A x_a \\ x_{aa} &= x_a^2. \end{aligned} \tag{2.9}$$

As we have shown above, HWE will be attained after only a single generation of random mating under the above assumptions. Breaking any of these assumptions can potentially lead to deviations from HWE. Possible scenarios that could do this include assortative mating, population structure, transmission ratio distortion, gene conversion, mutation, selection, small population size, and various others.

2.6 Diploid WF model

In order to model drift in a diploid population of N individuals, we can modify the haploid model such that, for each individual in generation $t+1$, we chose a pair of random parents in generation t. We then randomly pick one allele from each parent to assign to the child. As the following figure illustrates, such a model with random mating and random transmission is practically equivalent to a haploid WF model of size $2N$:



Discrepancies between the two models only arise in very small populations, because in our diploid model it is not allowed that both alleles of an individual trace back to the same individual (which would no longer be considered mating). However, these discrepancies become quickly irrelevant as N increases.

2.7 Effective population size

The WF model makes several idealizing assumptions, such as discrete generations and random mating. In real populations, these assumptions will often not hold. Nevertheless, the WF model can still prove very useful for helping us understand the basic features of RGD in a real population. One commonly used approach for this is to try to map a real population onto a corresponding WF model with a similar amount of drift. In particular, we want to describe the real population by an effective population size (N_e) specifying the population size of a WF model with the same amount of RGD as the real population. These effective population sizes are typically much smaller than the actual number of individuals in the real population. In order to define an effective population size, we need to decide how we want to quantify the amount of RGD. This can be done in several ways, each yielding its own version of an effective population size. Here we will focus on two such definitions: the *variance* effective population size and the *inbreeding* effective population size. In chapter 3, we will introduce yet another definition: the *coalescent* effective population size.

Variance effective size

One of the key features of genetic drift is that it causes allele frequencies to fluctuate stochastically between generations. The magnitude of these fluctuations is described by their variance. For a haploid WF model, we calculated: $V[x_{t+1}] = x_t(1-x_t)/N$. In our diploid model, we simply had to exchange $N \to 2N$. We can turn this relation around and use the empirically measured variance $V'[x]$ in a real population to define the variance effective population size for the corresponding WF model:

$$N_e = \begin{cases} \dfrac{x(1-x)}{V'[x]} & \text{(haploid)} \\ \dfrac{x(1-x)}{2V'[x]} & \text{(diploid).} \end{cases} \tag{2.10}$$

Note that the variance effective population size can in principle be a function of allele frequency if $V'[x]$ is not actually proportional to $x(1-x)$. Variance N_e can be a useful concept for describing real populations when it is possible to measure $V'[x]$ empirically from observed frequency changes over time.

Inbreeding effective size

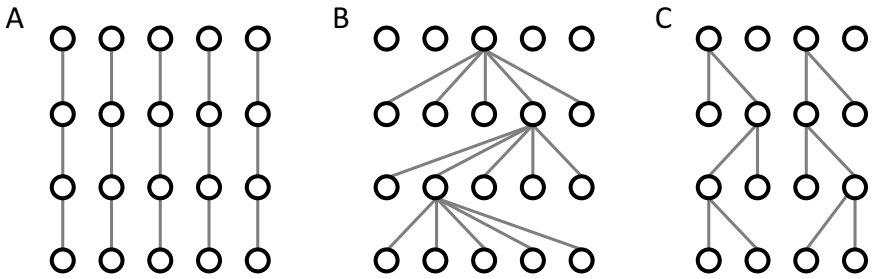
Another key feature of genetic drift is that two randomly chosen alleles in a population can come from the same parent allele in the previous generation. In this case, we say the two alleles are identical-by-decent (IBD). We have already seen that this feature of the WF model is a key determinant for the rate at which heterozygosity decays in the model. In a haploid WF model of size N, the probability that two alleles are IBD in the parent generation is simply $\Pr(\text{IBD}) = 1/N$. Note that IBD is defined on the level of alleles, rather than individuals. In a diploid population, we therefore calculate the probability of IBD in the parent generation by first picking two random children, and then picking one random allele from each. In a diploid WF population of N individuals without separate sexes, this yields $\Pr(\text{IBD}) = 1/(2N)$. We can use these relations to define the so-called inbreeding effective population size by:

$$N_e = \begin{cases} \dfrac{1}{\Pr(\text{IBD})} & \text{(haploid)} \\ \dfrac{1}{2\Pr(\text{IBD})} & \text{(diploid),} \end{cases} \tag{2.11}$$

where $\Pr(\text{IBD})$ is the probability of IBD in the parent generation in our real population, which we could try to obtain empirically through measurement, or infer from other knowledge about how reproduction, mating, etc. proceeds in this population.

Example 1: modified WF models

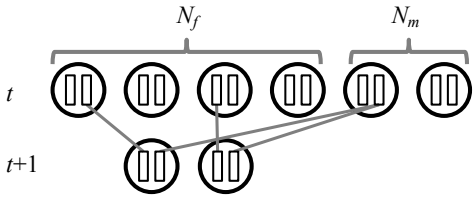
Let us study effective population sizes in three simple modifications of the WF model where parents are no longer chosen randomly from the previous generation but according to certain modified rules:



Scenario (A) describes a population in which each parent always has exactly one child. In this case, $\Pr(\text{IBD}) = 0$. The inbreeding effective population size is therefore infinite. Scenario (B) describes the opposite extreme, where all children always stem from the same (randomly chosen) parent in the previous generations. In this case, $\Pr(\text{IBD}) = 1$, and thus inbreeding $N_e = 1$. Scenario (C) describes a population of $N = 4$ individuals, in which exactly two random individuals reproduce each generation. Each of them always has two children. This yields $\Pr(\text{IBD}) = 1/3$ (after choosing the first individual, there will be only three individuals left to chose from and only one of them will be IBD with the first one). Therefore, inbreeding $N_e = 3$ in this case. All of these scenarios are obviously rather artificial, but they illustrate nicely how different aspects of the reproductive process will affect effective population sizes.

Example 2: dioecious population with biased sex-ratio

Most sexual species are dioecious, meaning that they have two sexes. Matings can only occur between a male and a female. Consider a dioecious, diploid, randomly mating population, that evolves over discrete generations and consists of N_m males and N_f females. Each child in generation $t+1$ will inherit one allele from a male and the other allele from a female in generation t.



If we pick two random individuals in generation $t+1$ and then randomly pick one allele from each, these two alleles can only be IBD if both alleles are either inherited from a male, or both alleles are inherited from a female. These scenarios occur with probability $1/4$ each. If the two alleles were both inherited from a female, the probability that they actually trace back to the same allele in a single female is $1/(2N_f)$ (there are $2N_f$ alleles in the pool of N_m females). If both alleles were inherited from a male, the probability that they trace back to the same allele in a single male is $1/(2N_m)$, accordingly. Taken together, the probability that the two alleles are IBD is thus given by:

$$\Pr(\text{IBD}) = \frac{1}{4} \times \frac{1}{2N_f} + \frac{1}{4} \times \frac{1}{2N_m} = \frac{1}{8}\left(\frac{1}{N_f} + \frac{1}{N_m}\right). \tag{2.12}$$

The inbreeding effective population of this population is therefore:

$$N_e = \frac{1}{2\mathrm{Pr}(\mathsf{IBD})} = \frac{4N_f N_m}{N_f + N_m}. \tag{2.13}$$

As an example, imagine a zoo population of 40 primates with 20 males and 20 females. Due to dominance hierarchy all females, yet only one of the males, can actually mate. Thus, $N_m = 1$ and $N_f = 20$. In this case: $N_e = 4N_f N_m/(N_f + N_m) = 4 \cdot 20 \cdot 1/21 \approx 3.8$. Obviously there will be a problem in defining a WF model for which the population size is not an integer. Nevertheless, the result still makes sense in the context of other aspects of RGD, such as describing the rate at which genetic diversity is lost over time in such a population according to Equation (2.7).

Chapter 3

Mutation

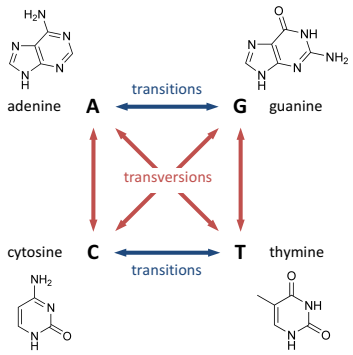
Over time, genetic drift will lead to fixation or loss of genetic variants, thereby systematically eliminating genetic variation from a population. In a real population, this trend is counterbalanced by mutational processes generating new genetic variants upon which other evolutionary processes can then act. Understanding the interplay between mutation and random genetic drift is of fundamental importance for understanding patterns of genetic diversity within populations, and the accumulation of genetic differences between species over time.

3.1 Mutation models

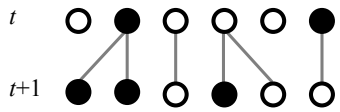
Mutations can come in many forms, such as nucleotide mutations, DNA insertions and deletions, duplications, inversions, translocations, etc. In multicellular organisms, mutations occur throughout their lifetime in most of their cells. Some of these mutations can have a profound impact on the individual, for instance when they initiate uncontrolled cell-growth that leads to cancer. However, for many question in population genetics those mutations that occur in somatic cells are largely irrelevant (as long as we do not consider their potential effect on phenotype and fitness). Only mutations that occur in the germline are typically inherited to future generations. We often do not bother about the complexities on the cell level, nor the specifics of each individual mutation. Instead, we want to describe mutations by idealized models that retain key aspects of the process while remaining simple enough so we can study them analytically. Three such models are widely used:

k-alleles model

The k-alleles model assumes that there exist only a finite number of possible allele states $\{A_1, \ldots, A_k\}$ at a locus and mutations can convert alleles between these states. The classic example of a k-alleles model are the four possible nucleotides $\{A, C, G, T\}$ at a single nucleotide position in the genome.



To incorporated the k-alleles model into our WF framework, we need to implement the possibility that allele states can change in individuals, which we will model by a Bernoulli process (success = mutation occurs, failure = mutation does not occur). This is also referred to as mutation "dropping". The following scenario illustrates such an extended WF model for a population of $N = 6$ individuals, assuming a 2-alleles model (black and white) in which two mutations occur in generation $t + 1$:



Generally there can be distinct probabilities for mutations occurring between each two states, $A_i \to A_j$, which can be described by a matrix of mutation probabilities. For example, in the $\{A, C, G, T\}$ model we can have the following 12 individual mutations probabilities:

$$\begin{pmatrix} \bullet & \Pr(A|C) & \Pr(A|G) & \Pr(A|T) \\ \Pr(C|A) & \bullet & \Pr(C|G) & \Pr(C|T) \\ \Pr(G|A) & \Pr(G|C) & \bullet & \Pr(G|T) \\ \Pr(T|A) & \Pr(T|C) & \Pr(T|G) & \bullet \end{pmatrix} \tag{3.1}$$

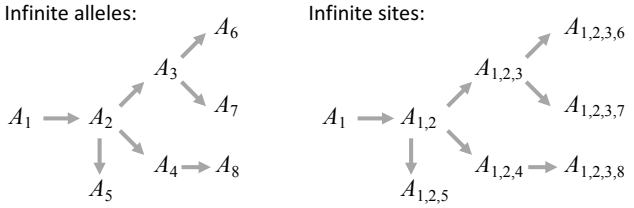
Here $\Pr(C|T)$, for instance, specifies the probability that a T allele mutates into a C allele, given that the individual carries a T allele at the nucleotide locus. The WF model has discrete generations, and mutation probabilities are therefore typically defined as the probability that a given mutation occurs per individual per generation. Note also that all these probabilities are conditional probabilities, i.e., they depend on the current allele state of the individual. For example, if the current allele state of an individual is T, the combined probability that any mutation occurs in this individual at this locus is then $\Pr(A|T) + \Pr(C|T) + \Pr(G|T)$.

Infinite alleles model

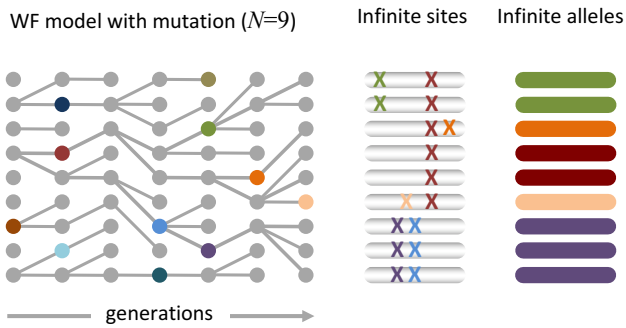
In contrast to the k-alleles model, the infinite alleles model assumes that there is no limit to the number of possible allele states at a locus and that each mutation produces a new state that has not previously existed in the population: $A_1 \to A_2 \to A_3 \to \cdots$. The infinite alleles model is commonly used for describing the evolution of haplotypes in a sufficiently large genomic region, where one can assume that every mutational event creates a novel haplotype that has not previously existed in the population.

Infinite sites model

The infinite sites model is an extension of the the infinite alleles models, which preserves some information on the mutational history of each allele (yet not their particular chronological sequence). This allows for the definition of an evolutionary distance between two alleles in terms of the number of mutational steps by which they differ:



The infinite sites model is often used for describing nucleotide mutations across a larger genomic region, where each mutation event can be assumed to produce a point mutation at a novel site where no previous mutation has yet occurred in the population. We can construct an explicit instantiation of the infinite sites model as follows: Consider a diploid WF population of N individuals. As our locus, we define a non-recombining genomic region of length L base pairs. Nucleotide mutations at individual sites of the locus occur at rate μ per generation per allele, which we assume to be constant across all L sites in our locus. We are not interested here in the particular nucleotide states at these sites, and simply refer to any nucleotide change as a mutation. On average, at any given site, we expect $2N\mu$ mutations to occur per generation in the population overall, and $2N\mu L$ mutations across the whole locus. The infinite sites model can be obtained as the limit of this model for $L \to \infty$ and $N\mu \to 0$, while $N\mu L$ remains finite. In other words, we consider a locus with infinitely many sites, where the probability that a mutation occurs at any particular site is negligible. In this case, a new mutation will always occur at a novel site where no preexisting mutation is already segregating in the population. Consequently there will never be more than two different alleles at a polymorphic site and we can also neglect "back mutations". We will designate the two different alleles at each polymorphic locus as "wildtype" and "mutant" allele. Obviously the infinite sites model is an idealization that can never apply exactly for any finite locus, but it will be a very useful model for describing a locus with $L \gg 1$ in a population with $N\mu \ll 1$. The following figure illustrates the connection between infinite sites and infinite alleles models, interpreted in the context of a haploid WF model:



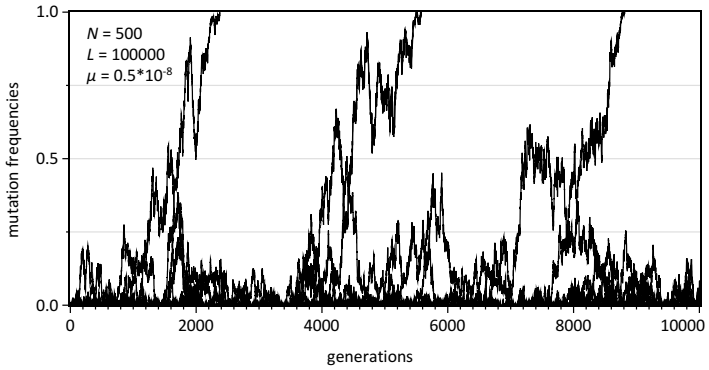
3.2 The molecular clock

Consider a genomic locus in diploid WF population with N individuals, evolving under an infinite sites model as specified above. If we focus exclusively on a particular nucleotide site inside this locus where a mutation has just occurred, evolution at this site will be equivalent to that of a WF model with just two alleles (wildtype and mutant). In particular, drift will ultimately lead to either fixation or loss of the mutant, with fixation occurring with probability $\Pr(\text{fixation} \,|\, x_0) = x_0 = 1/(2N)$. The overall rate at which new mutations are expected to arise in the population across our locus is $2N\mu L$, each of them having probability $1/(2N)$ of ultimately becoming fixed in the population. The product of the two specifies the overall rate at which mutations become fixed in the population across the locus. We will write this overall rate of fixation as the product of the fixation rate per nucleotide site, d, which is also called the nucleotide substitution rate, and the length L of the locus. Together, we obtain:

$$dL = 2N\mu L \times \frac{1}{2N} = \mu L \quad \Rightarrow \quad d = \mu \tag{3.2}$$

Hence, the nucleotide substitution rate simply equals the nucleotide mutation rate in this model. Maybe surprisingly, it is independent of the population size, or in fact any other population-level characteristics. This is the consequence of the following symmetry: while in a larger population more new mutations will occur in the

population overall, each individual such mutation will have a lower fixation probability, and both these effects cancel out. Below is an illustration showing how neutral mutations in a simulated WF model evolving under an infinite sites model can occasionally drift to fixation in the population:



The product $dL = \mu L$ describes the rate at which substitutions are expected to accumulate across the locus of an infinite sites model if drift and mutation are the only evolutionary processes acting. Assuming that μ is constant over time, we then expect $\mu L \times t$ mutations to become fixed in the population over an evolutionary time span of t generations. This relationship has important applications, as it relates mutation rate, evolutionary time, and expected number of fixed mutation at a neutrally evolving locus with each other. Given estimates of two of these quantities, we can then use this relationship to infer the third. This influential idea is known as the *molecular clock* hypothesis and was originally conceived by Emile Zuckerkandl and Linus Pauling in the 1960's.

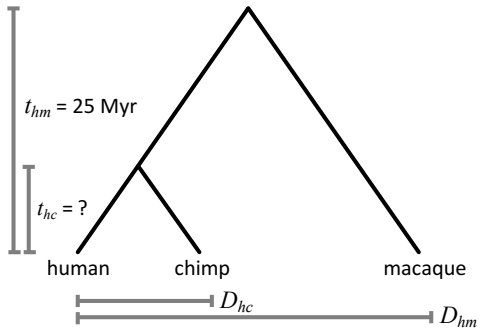
One important application of the molecular clock is the possibility to date speciation times. Consider a neutrally-evolving genomic locus of length L base pairs in two species that split from a common ancestor t_s generations in the past. In this case we expect $\mu L \times t_s$ mutations to have fixed along each of the two lineages. As long as we can neglect the possibility that the same mutations may have occurred in both species, or that more than one substitution has occurred at the same site in any lineage, we can estimate the overall number of substitutions that have occurred along both lineages by simply counting the number of observed nucleotide difference (D) at the locus between the two species. We expect:

$$E[D] = \mu L \times 2t_s. \tag{3.3}$$

The factor 2 here arises because substitutions can have occurred in either of the two lineages. An important caveat of the molecular clock hypothesis lies in the speicifc assumptions it is based on, namely that mutation and drift are the only processes acting and that mutation rates remain approximately constant over time. Both assumptions are violated in most realistic systems and much research has been devoted towards study how such violations can be corrected for in our evolutionary models.

Example: dating the human-chimpanzee divergence time

Paleontological evidence suggests that humans and rhesus macaque monkeys shared a common ancestor until around $t_{hm} = 25$ Myr ago. The present-day human genome differs from the present-day macaque genome, on average, at about 7 out of 100 nucleotide sites ($D_{hm}/L = 0.07$). We can use these two numbers to obtain an estimate of the single-nucleotide mutation rate: $\mu = D_{hm}/(2t_{hm}L) = 1.4 \times 10^{-9}$ per year.

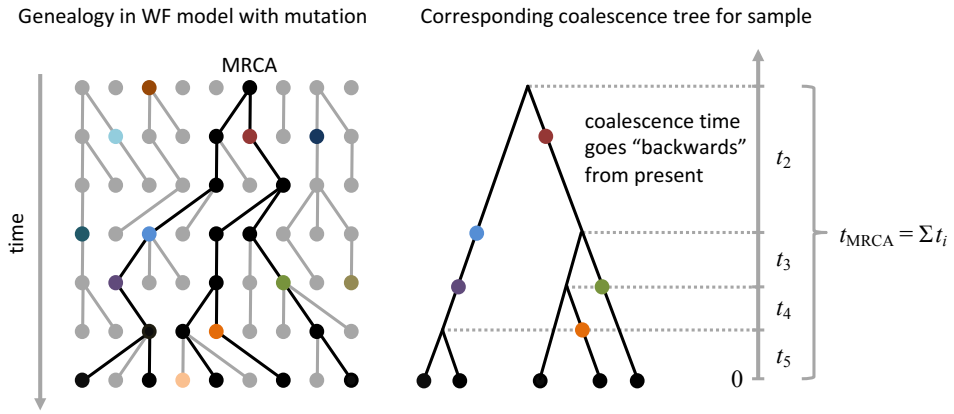


Sequencing of the chimpanzee genome revealed that humans and chimps differ at about 1.2 out of 100 sites ($D_{hc}/L = 0.012$). Given our previous estimate of the mutation rate, we can use this divergence estimate to date the human-chimp split to $t_{hc} = D_{hc}/(2\mu L) = 4.3$ Myr ago. This estimate is somewhat smaller than estimates based on paleontological evidence, which date the split event to 5-6 Myr ago. One possible explanation for this discrepancy could be a change in generation time: macaque monkeys have a shorter generations, so the mutation rate per year might be higher in macaques compared with chimps and humans. Other possible explanations include changes in the actual mutation rate, natural selection, error in the timing of the human-macaque split, or sequencing errors.

Chapter 4

Coalescence theory

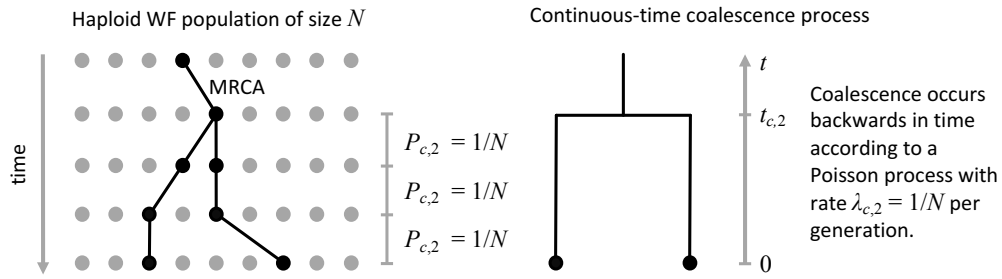
In real populations it is typically unrealistic to obtain the genetic information for every individual. Instead, observations will often rely on smaller samples of individuals. One of the most fruitful insights in modern population genetics was the realization that for understanding the evolutionary patterns in a population it often suffices to focus on the mutations that occurred along the genealogy of such population samples, rather than the population as a whole. The mathematical framework that captures this approach, coalescence theory, was developed in the 1980's by J. F. C. Kingman, R. C. Griffiths, R. R. Hudson, F. Tajima and others. It has since come to be the workhorse of modern population genetics, especially for scenarios in which selective forces can be ignored. Due to its focus on population samples, results from coalescence theory are often directly relatable to empirical observations and provide a very intuitive understanding for the interplay between patterns and processes. In addition, the coalescent process offers a very efficient framework for numerical simulations. The following figure illustrates the connection between genealogies on the population level in a WF model with mutation and the corresponding coalescence tree of a sample of five alleles from this population, taken at present. The patterns of genetic diversity observed in the sample will be determined by the interplay of coalescence events (where two lineages meet in a common ancestor) and mutation events occurring along the lineages of this genealogical tree:



4.1 Coalescence in a sample of size two

The coalescence process is a continuous-time Markov process for modeling the genealogy of a population sample. The idea is similar to the WF model, a discrete-time Markov process we introduced as model for describing the effects of random genetic drift in a population. However, in contrast to the coalescence process,

which modeled the genealogical history of every individual in the population, the coalescence process only focuses on those lineages that are relevant for our specific population sample. The key event of the coalscence process is that two lineages in the sample can meet (coalesce) in their most recent common ancestor (MRCA) when going backward in time:



Consider a haploid WF population of size N. In the WF model, two randomly chosen alleles coalesce with probability $P_{c,2} = 1/N$ in the previous generation, which is simply the probability that the two alleles are identical-by-decent (IBD) in the previous generation. In the continuous-time coalescent framework, we will describe such coalescence events by a Poisson process with pairwise coalescence rate $\lambda_{c,2}$, which we model going "backwards" in time from the point where the alleles were sampled. The waiting time until coalescence in a WF model is then an exponential RV with rate-parameter $\lambda_{c,2} = 1/N_c$. According to Equation (1.23), the probability that two lineages have not yet coalesced after t generations is given by:

$$\Pr(t_{c,2} > t) = \int_t^\infty \lambda_{c,2} \cdot e^{-\lambda_{c,2} x} dx = e^{-t/N}. \tag{4.1}$$

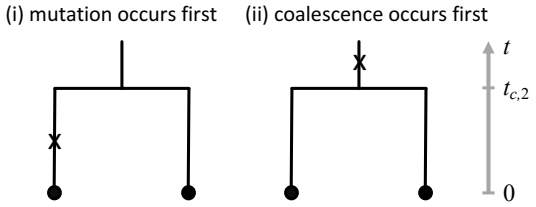
Given that $t_{c,2}$ is an exponential RV, the expected time until two alleles coalesce will be $E[t_{c,2}] = 1/\lambda_{c,2} = N$ generations. Extension to a diploid WF population is straightforward: In this case, $P_{c,2} = \lambda_{c,2} = 1/(2N)$, and thus $\Pr(t_{c,2} > t) = e^{-t/(2N)}$ and $E[t_{c,2}] = 1/\lambda_{c,2} = 2N$ generations.

4.2 Heterozygosity under mutation and drift

Neutral mutations do not interfere with the genealogy of a population sample because they have no effect on the reproductive success of the individuals that carry them. In a neutral scenario, we can thus treat coalescence and mutation by two independent Poisson processes. In particular, we can generate the coalescence tree first, and add mutations to the tree afterwards ("mutation-dropping"). Mutations that occur in individuals that do not lie along the genealogy of our population sample can be ignored, as those mutations will not be observed in the sample. Along the lineages of a coalescence tree mutations are dropped at the same rate at which they occur in individuals, which can be done according to any of the mutation models specified above.

Infinite sites model

Consider a locus evolving under an infinite sites model with per-nucleotide mutation rate μ in a population with pairwise coalescence rate $\lambda_{c,2}$. We are interested in the expected heterozygosity per nucleotide site (H), which we define as the probability that two randomly drawn alleles from the population differ at at a randomly drawn nucleotide site. In order to differ, a mutation must have occurred in one of the lineages at that site prior to the two alleles coalescing in the past (in our infinite sites model we neglect the probability that mutations could have occurred in both lineages at the same site, as well as back-mutations). The probability that the two alleles differ at the site is therefore simply the probability that a mutation occurs in either of the lineages, prior to the two lineages coalescing:



Mutation and coalescence are independent Poisson processes in our model, with coalescence of the two lineages occuring at rate $\lambda_{c,2}$ and mutations occurring at rate 2μ (the factor 2 arises here because mutations can occur in either of the two lineages). According to Equation (1.24), the probability that a mutation occurs prior to coalescence is then simply

$$H = \frac{2\mu}{2\mu + \lambda_{c,2}} = \frac{\Theta}{1+\Theta} \approx \Theta \quad \text{with} \quad \Theta = \frac{2\mu}{\lambda_{c,2}} = 2\mu E[t_{c,2}]. \tag{4.2}$$

The approximation $\Theta/(1+\Theta) \approx \Theta$ is appropriate when $\Theta \ll 1$. Note that we haven't made any assumption here about the population, other than that the lineages of the two randomly drawn alleles coalesce at rate $\lambda_{c,2}$ per generation, which we assume to be constant over time. Given some specific model, we can then estimate $E[t_{c,2}]$ for this model and calculate Θ explicitly. For example, in a haploid WF population with N individuals, we have $\lambda_{c,2} = 1/N$ and thus $\Theta = 2N\mu$. In a diploid WF model, we have $\lambda_{c,2} = 1/(2N)$ and thus $E[t_{c,2}] = 2N$ generations and $\Theta = 4N\mu$. In a realistic diploid population with inbreeding effective population size N_e, we have $\lambda_{c,2} = 1/(2N_e)$ and thus $\Theta = 4N_e\mu$.

Equation (4.2) illustrates an important consequence of the interplay between mutation and drift: While in the WF model without mutation heterozygosity decays exponentially, the incorporation of a mutation process leads to the maintenance of an equilibrium level of heterozygosity, specified by $\Theta/(1+\Theta)$. The parameter Θ is a key parameter in population genetics that we will encounter frequently over the course of the class.

Infinite alleles model

The extension of the above theory to an infinite alleles model is trivial. In an infinite alleles model there are no individual nucleotide sites and H is then simply defined as the probability that two alleles at the locus are different by state. If mutations occur at rate μ at the locus, all of the above arguments still apply without any adjustment needed. In particular, we again obtain $H = \Theta/(1+\Theta)$ with $\Theta = 2\mu E[t_{c,2}]$.

4.3 Estimating nucleotide diversity in a population sample

One of the key advantages of the coalescence process is that it provides a framework in which we can directly relate statistical predictions of an evolutionary model with empirically observable data, such as the patterns of nucleotide diversity in a genomic region we sequenced in a population sample. In the following we will define several statistical quantities that can be measured empirically in such data and are commonly used for describing patterns of nucleotide diversity. Consider a genomic region of length L nucleotides that we sequenced in a population sample of size n (obtained by randomly sampling a set of n alleles from the population). All sequences in our sample were then "aligned" against each other. We assume that the assumptions of the infinite sites model can be applied to our locus, so that we can effectively ignore multiple mutations occurring at the same nucleotide position and back-mutations. Let us define the following statistical quantities/terms:

- segregating site: nucleotide site where at least two sequences have a different nucleotide in our sample. We also call a segregating site a single-nucleotide polymorphism (SNP).
- S: overall number of segregating sites in the sample.
- major/minor SNP allele: more/less frequent nucleotide at a segregating site.
- d_{ij}: number of nucleotide sites at which sequences i and j differ from each other ($d_{ij} = d_{ji}$, $d_{ii} = 0$).
- $\overline{d_{ij}}$: average nucleotide distance between two sequences:

$$\overline{d_{ij}} = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2} \tag{4.3}$$

  Note that this average does not include distances of sequence with themselves.

- h_k: heterozygosity at nucleotide site k in our sample (without replacement). If k is not a segregating site, we have $h_k = 0$. At a segregating site where the minor allele is present in i copies, we have:

$$h_k = \frac{i}{n}\left(1 - \frac{i-1}{n-1}\right) + \left(1 - \frac{i}{n}\right)\frac{i}{n-1} = \frac{2i(n-i)}{n(n-1)} \tag{4.4}$$

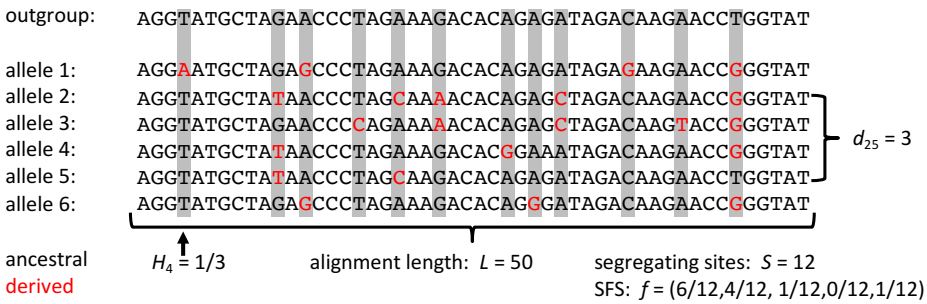
- π: average heterozygosity per nucleotide site:

$$\pi = \overline{d_{ij}}/L = \overline{h_k} = \frac{1}{L}\sum_{k=1}^{L} h_k \tag{4.5}$$

- ancestral/derived SNP allele: If we know which of the two alleles at a segregating site is the wildtype and which is the mutant (e.g. through comparison with an outgroup species) we call the wildtype allele the ancestral allele and the mutant allele the derived allele.
- site frequency spectrum (SFS): distribution of derived allele frequencies across all segregating sites. If we denote with s_i the number of segregating sites where the derived allele is present in i copies, then:

$$f = (s_1/S, s_2/S, \cdots, s_{n-1}/S) \tag{4.6}$$

- folded SFS: the site frequency spectrum estimated using only minor allele frequencies (typically used when we do not know which allele is ancestral/derived).
- singleton: segregating site at which the minor allele is present in only a single copy in the sample.
- summary statistic: any statistical quantity estimated from the population sample, such as S, π, and f.



4.4 Tajima's estimator

We derived in Equation (4.2) that for the infinite sites model under neutral evolution, the expected level of nucleotide heterozygosity is $H = \Theta/(1+\Theta) \approx \Theta$, where the approximation is appropriate when $\Theta \ll 1$. For an actual population sample, we can obtain an empirical estimate of the average level of heterozygosity per nucleotide by measuring $\pi = \overline{h_k}$. This allows us to obtain an educated guess for the value of Θ in the population through measurement of π in a population sample, which is called Tajima's estimator:

$$\Theta \approx \widehat{\Theta}_\pi = \pi. \tag{4.7}$$

Estimators are typically used for inferring a population parameter from the data that we are unable to observe directly (here $\Theta = 2\mu E[t_{c,2}]$) and they are indicated by a hat symbol above the parameter to be inferred. The subscript indicates that this estimator of Θ is based on an empirical measurement of π. If the sequenced locus is long enough and/or we have a large enough sample size, our estimator $\widehat{\Theta}_\pi$ from the population sample should converge to the true Θ of the population.

4.5 The coalescence effective population size

For a WF model, we calculated that $E[t_{c,2}] = N$ generations in a haploid model and $E[t_{c,2}] = 2N$ generations in a diploid model. Similar to how we defined variance and inbreeding effective population sizes, we can use these relations to define yet another version of N_e, which we call the *coalescence effective population size*:

$$N_e = \begin{cases} E[t_{c,2}] & \text{(haploids)} \\ \\ E[t_{c,2}]/2 & \text{(diploids)} \end{cases} \tag{4.8}$$

Coalescence, inbreeding, and variance effective population sizes are often equal, but not always. Coalescence and inbreeding N_e, for instance, are closely related due to the fact that, in each generation, the inverse of the inbreeding effective population size determines the pairwise coalescence rate in that generation. However, differences between the two definitions can arise when the inbreeding effective population size is not constant over time. We will talk more about this when we discuss demography.

In a real population, we can use Tajima's estimator and an independent estimate of the nucleotide mutation rate to obtain an empirical estimate of the coalescence N_e via the relation $\Theta = 2\mu E[t_{c,2}] \approx \pi$, which connects the level of nucleotide heterozygosity at a neutral locus, the nucleotide mutation rate, and the expected pairwise coalescence time. For example, using a sample of a neutral locus from a diploid population, we can infer $N_e = E[t_{c,2}]/2 \approx \widehat{\Theta}_\pi/(4\mu) = \pi/4\mu$.

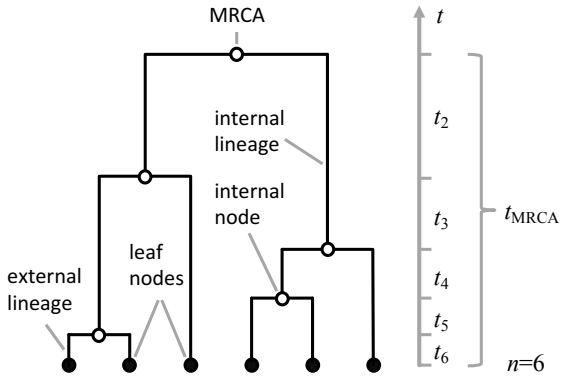
Example: pairwise coalescence time and effective population sizes in human populations

In human autosomes, we observe approximately 12 nucleotide differences per 10000 nucleotides between two randomly sampled genomes from individuals of African descent (i.e., $\pi \approx 0.0012$). In genomes sampled from individuals of European descent, we observe only around 8 differences per 10000 nucleotides (i.e., $\pi \approx 0.0008$). Assuming a single nucleotide mutation rate of $\mu \approx 2 \times 10^{-8}$ per generation, this yields: $E[t_{c,2}] = \pi/(2\mu) = 0.0012/(4 \times 10^{-8}) = 30000$ generations for individuals of African decent and $E[t_{c,2}] = 20000$ generations for individuals of European decent, corresponding to coalescence N_e of 15000 and 10000 individuals, respectively.

4.6 The coalescence process for larger samples

In order to generalize the coalescent process to larger samples, we have to account for the fact that coalescence events can occur between any pair of lineages in the sample. In a sample of size n, there are initially 'n choose

$2' = n(n-1)/2$ possible pairs of lineages that can coalesce. After the first coalescence event, $n-1$ lineages remain, and so on, until all lineages have reached their MRCA. We can describe this process through a *coalescence tree*, where the alleles in our sample constitute the external (or leaf) nodes and lineages connect nodes with each other:



Let us define t_k as the time interval during which there are exactly k lineages present in the tree ($2 \leq k \leq n$). The coalescence rate during this time interval is then:

$$\lambda_{c,k} = \binom{k}{2} \times \lambda_{c,2} = \frac{k(k-1)\lambda_{c,2}}{2}, \tag{4.9}$$

where $\lambda_{c,2}$ is the pairwise coalescence rate (e.g. $\lambda_{c,2} = 1/N$ in a haploid WF population of size N). Given the coalescence rates $\lambda_{c,k}$, we can calculate the expected waiting time until the next coalescence event:

$$E[t_k] = \frac{1}{\lambda_{c,k}} = \frac{2}{k(k-1)\lambda_{c,2}}. \tag{4.10}$$

The time until all alleles have reached their MRCA is then the sum over all intervals t_k:

$$E[t_{\mathrm{MRCA}}] = \sum_{k=2}^{n} E[t_k] = \frac{2}{\lambda_{c,2}} \sum_{k=2}^{n} \frac{1}{k(k-1)}. \tag{4.11}$$

Using this result, we can also calculate the total length of all lineages in the tree. By definition, for each k in the interval $2 \leq k \leq n$ there will be k lineages of length t_k present in the tree. The expected total length of all lineages combined – the so-called total tree length (t_{total}) – is therefore given by:

$$E[t_{\mathrm{total}}] = \sum_{k=2}^{n} kE[t_k] = \frac{2}{\lambda_{c,2}} \sum_{k=2}^{n} \frac{1}{k-1} = \frac{2}{\lambda_{c,2}} \sum_{k=1}^{n-1} \frac{1}{k}. \tag{4.12}$$

Mutations can occur along all lineages in the tree, and we will again model such mutations by a Poisson process with constant rate μ. For a locus of length L nucleotides evolving under infinite sites assumptions, the total number of mutations expected to occur along the tree, which will determine the number of segregating sites (S) in our sample, is then:

$$E[S] = \mu L E[t_{\mathrm{total}}]. \tag{4.13}$$

Note that we again made use of the fact that we can treat coalescence and mutation as two independent processes, so that we can build the coalescence tree first and then add mutations to the tree afterwards.

Example: coalescence in a WF population

Consider a sample of size n, drawn randomly from a diploid WF populations of N individuals. In this case, $\lambda_{c,2} = 1/(2N)$, and therefore:

$$\lambda_{c,k} = \frac{k(k-1)}{4N}, \tag{4.14}$$

$$E[t_k] = \frac{4N}{k(k-1)}, \tag{4.15}$$

$$E[t_{\mathrm{MRCA}}] = 4N \sum_{k=2}^{n} \frac{1}{k(k-1)} = 4N \left(\frac{1}{2} + \frac{1}{3 \cdot 2} + \cdots + \frac{1}{n(n-1)} \right), \tag{4.16}$$

$$E[t_{\mathrm{total}}] = 4N \sum_{k=1}^{n-1} \frac{1}{k} = 4N \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1} \right), \tag{4.17}$$

$$E[S] = \mu L E[t_{\mathrm{total}}] = \Theta L \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1} \right) \quad \text{with} \quad \Theta = 4N\mu. \tag{4.18}$$

In a haploid WF population, $\lambda_{c,2} = 1/N$. We therefore simply have to replace $2N \to N$ in the above equations.

4.7 Watterson's estimator

Equation (4.18) relates a summary statistic (S) we can measure from a population sample to an unknown evolutionary parameter (Θ). We previously used a similar relation based on the average nucleotide heterozygosity in our sample (π) to construct Tajima's estimator $\widehat{\Theta}_\pi$. Accordingly, we can can construct another estimator for Θ based on Equation (4.18), which is called Watterson's estimator:

$$\widehat{\Theta}_w = \frac{S}{L \sum_{k=1}^{n-1} 1/k}. \tag{4.19}$$

Both Tajima's estimator and Watterson's estimator will converge to the true value of Θ if we cann estimate them over many samples in an ideal WF population. Each single estimate, of course, can be somewhat different from the true value because of noise due to a finite sample size. The two estimators will no longer converge if certain assumptions of the WF are violated in the real population. Two examples for scenarios that will cause such deviations are demographic events and selection. Comparing both estimators in a real population and testing whether their estimates show larger differences than expect by noise alone can therefore be used to infer violations of the WF model assumptions, such as selection or demography.

4.8 Site frequency spectrum in the infinite sites model

In addition to the number of segregating sites in the sample (S) and the average nucleotide heterozygosity (π), in section 4.3 we had introduced a third summary statistic, the site frequency spectrum (SFS). We defined the SFS as the distribution of the numbers of segregating sites (s_i) at which the derived allele is present in $0 < i < n$ copies: $f = (s_1/S, s_2/S, \cdots, s_{n-1}/S)$. Consider a genomic locus of length L nucleotides, evolving in a WF population under infinite sites model assumptions with nucleotide mutation rate μ. For a population sample sample of size n, it then holds that:

$$E[s_i] = \frac{\Theta L}{i}, \quad 0 < i < n, \tag{4.20}$$

where $\Theta = 2N\mu$ in a haploid WF population of size N and $\Theta = 4N\mu$ in a diploid WF population. We will not prove this result here, since all currently know derivations are rather cumbersome. For an instructive proof, see e.g. RR Hudson, *PLoS One*, e0118087 (2015).

4.9 Allele frequency spectrum in the infinite alleles model

Under the infinite alleles model, every mutation event creates a new allele that maintains no information on its mutational history. In a sense, this makes the coalescent process simpler, because large parts of the genealogy will have no influence on the allele-patterns observed in a sample. Each lineage needs to be followed only until the most recent mutation, whereas mutations and coalescence events further back in time can be ignored. The mathematical process for describing this scenario is the so-called coalescent with 'killings', in which coalescence and mutation are modeled simultaneously and each lineage is terminated once a mutation is encountered. The process stops when only a single lineage remains.

Since there is no information in the comparison of two alleles other than whether they are different or the same, the primary quantities of interest in this model are how many different alleles there are in the sample, and how many of those alleles are present in $1, 2, \dots$ copies. We will denote these counts by a vector $(a_1, a_2, \dots)$. The probability of observing a given set of counts $(a_1, a_2, \dots, a_n)$ in a random sample of size n alleles from a locus evolving under an infinite alleles model with mutation rate μ in a WF population is then:

$$P[(a_1, a_2, \dots, a_n)] = \frac{n!}{\Theta_{(n)}} \prod_{j=1}^{n} \frac{(\Theta/j)^{a_j}}{a_j!}, \tag{4.21}$$

where $\Theta_{(n)} = \Theta(\Theta + 1) \cdots (\Theta + n - 1)$ with $\Theta = 2N\mu$ in a haploid WF population and $\Theta = 4N\mu$ in a diploid WF population of size N. Note that the number of distinct alleles in the sample is simply $k = \sum_{i=1}^{n} a_i$. Equation 4.21 is known as Ewen's sampling formula. It can be used, for example, for describing the expected distribution of haplotype frequencies at a neutrally evolving locus in a WF population. For a proof of Ewan's sampling formula, we refer the reader to John Wakeley's book "Coalescent Theory: An Introduction".

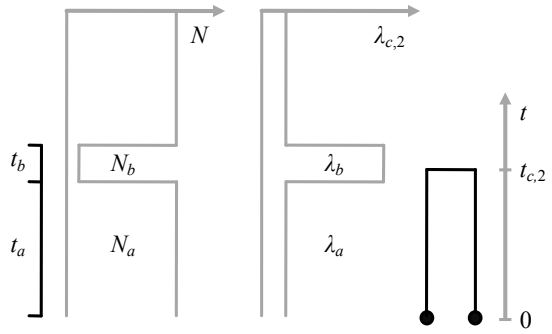
Chapter 5

Demography

We have assumed so far that population size remains constant over time. Real populations, however, can fluctuate in size quite dramatically over time. We call the sequence of population size changes over time the demographic history of a population. The events that cause these changes, such as population bottlenecks and population expansions, are called demographic events. In this chapter, we will study how demographic events affect random genetic drift and the expected patterns of genetic diversity in a population.

5.1 Population bottlenecks

Consider a population that consisted of N_a individuals over the past t_a generations with constant pairwise coalescence rate λ_a during this time. Prior to this, it experienced a population bottleneck that lasted for t_b generations, during which its size was only $N_b < N_a$ individuals and the pairwise coalescence rate was $\lambda_b > \lambda_a$. For simplicity, we assume that population size changes occurred instantaneously.



We can calculate the probability that the lineages in a sample of size two coalesce before reaching the bottleneck, using an exponential distribution with constant pairwise coalescence rate λ_a:

$$\Pr(t_{c,2} < t_a) = 1 - e^{-\lambda_a t_a}. \tag{5.1}$$

The probability that coalescence does not occur prior to, but during the bottleneck, is then given by:

$$\Pr(t_a < t_{c,2} < t_a + t_b) = e^{-\lambda_a t_a}(1 - e^{-\lambda_b t_b}). \tag{5.2}$$

Example: large population with severe ancient bottleneck

Consider a haploid WF population undergoing the above bottleneck scenario, with current population size $N_a = 10^6$ and bottleneck size $N_b = 100$. In this case, $\lambda_a = 10^{-6}$ and $\lambda_b = 10^{-2}$. We further assume that the

bottleneck occurred $t_a = 10^4$ generations ago and lasted for $t_b = 200$ generations. The probability that the two alleles coalesce more recently than the bottleneck is $\Pr(t_{c,2} < t_a) = 1 - e^{-\lambda_a t_a} \approx 0.01$. The probability of coalescence during the bottleneck is $\Pr(t_a < t_{c,2} < t_a + t_b) \approx (1 - 0.01) \times (1 - e^{-2}) \approx 0.86$.

Thus, in almost 90% of cases the two alleles coalesce during the bottleneck, whereas more recent coalescence is very unlikely, even though the bottleneck occurs 10^4 generations ago and lasts for only 200 generations. Note that since most coalescence will occur during the bottleneck, we have $E[t_{c,2}] \approx t_a$, and thus the coalescence effective population size will be $N_e \approx t_a$. For a locus evolving under an infinite sites model with mutation rate μ, we obtain $E[\pi] = 2\mu E[t_{c,2}] \approx 2\mu t_a$. This means that the expected level of diversity in this scenario is mostly determined by when the bottleneck occurred, rather than the actual size of the population.

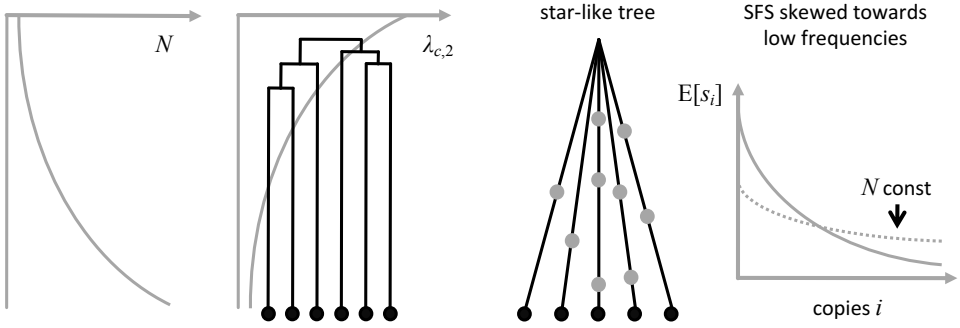
5.2 Population expansions

Another important class of demographic events are population expansions. In a sense, every population bottleneck is associated with an expansion when the population recovers from the bottleneck. However, in contrast to the above scenario, where we assumed that the population expanded instantaneously after a bottleneck, here we want to study a continuously growing population.

In principle, we can calculate the probability that pairwise coalescence occurs in this model prior to time t using Equation (1.23) with a time-dependent coalescence rate, $\lambda_{c,2}(t)$, which we integrate over from 0 to t:

$$\Pr(t_{c,2} < t) = \int_0^t \lambda_{c,2}(x) e^{-\lambda_{c,2}(x)} dx. \tag{5.3}$$

While this integral can be solved for several standard cases, such as exponential growth, we want to discuss only the qualitative aspects of such scenarios here. The key insight for understanding the expected patterns of diversity in a rapidly expanding populations is that the rate of coalescence is initially low, but grows rapidly as one goes further back in time. As a result, coalescence trees look quite different from those in a constant-size scenario. Specifically, the external branches will be relatively long compared with internal branches, and most of the coalescence events should occur in a narrow window of time. We call such trees "star-like".

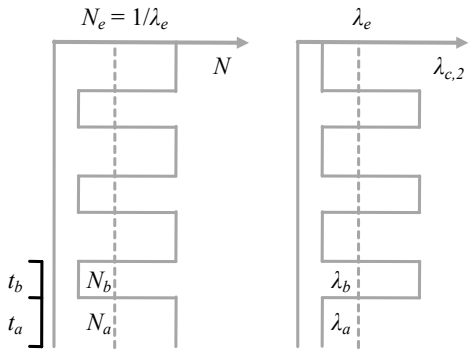


The patterns of diversity expected for star-like coalescence trees differ from the trees in a constant-sized population in that we expect the SFS to be skewed towards low-frequency polymorphisms and that there will be an excess of singletons due to mutations on the long external branches.

5.3 Fluctuating population size

The third scenario we want to study is a population that goes though rapid population cycles, such as those observed in predator-prey systems or in insect populations that fluctuate over yearly seasons. To model such

demographies, we will initially use a highly simplified two-phase scenario, in which we assume that there are only two different population phases that periodically alternate (large: N_a and λ_a; small: N_b and λ_b). Between each two phases the population size changes instantaneously. We assume that the N_a phase always lasts exactly t_a generations, while the t_b phase always lasts t_b generations. Let us further assume that pairwise coalescence is slow compared with the time-scale over which population size changes occur, i.e., $t_a + t_b \ll E[t_{c,2}]$, meaning that on average we have to wait many cycles for a pairwise coalescence event to happen.



Can we map this scenario onto a WF model with constant pairwise coalescence rate λ_e, such that the fluctuating model and the constant model yield the same $E[t_{c,2}]$? In such a constant model, the probability that a coalescence event occurs over the course of one full cycle would then simply be:

$$\Pr(t_{c,2} < t_a + t_b) = 1 - e^{-\lambda_e(t_a+t_b)} \approx \lambda_e(t_a + t_b), \tag{5.4}$$

where the last approximation, $e^x \approx 1 + x$, is a so-called Taylor-approximation, which holds for $x \ll 1$. This condition is met because of our assumption that coalescence is unlikely to occur within one given cycle. In the fluctuating model, coalescence can occur in the first *or* the second phase of each cycle, so we have to add the probabilities for each case, which we already derived in Equations (5.1) and (5.2), yielding:

$$\begin{aligned}
\Pr(t_c < t_a + t_b) &= \Pr(t_{c,2} < t_a) + \Pr(t_a < t_{c,2} < t_a + t_b) && (5.5)\\
&= \left(1 - e^{-\lambda_a t_a}\right) + e^{-\lambda_a t_a}\left(1 - e^{-\lambda_b t_b}\right) && (5.6)\\
&\approx \lambda_a t_a + \lambda_b t_b. && (5.7)
\end{aligned}$$

Here we again used $e^x \approx 1 + x$. If both the fluctuating and constant size model should yield the same $E[t_{c,2}]$, both need to have the same (small) probability that pairwise coalescence occurs over the course of any given cycle. We can therefore equate both probabilities, obtaining:

$$\lambda_e = \frac{\lambda_a t_a + \lambda_b t_b}{t_a + t_b}. \tag{5.8}$$

If we assume a diploid WF model during each stage, we can relate pairwise coalescence rates to population sizes, $\lambda_{c_2} = 1/(2N)$, which yields:

$$N_e = \frac{1}{\lambda_e} = \frac{t_a + t_b}{t_a/N_a + t_b/N_b}. \tag{5.9}$$

This N_e specifies the coalescence effective population size for a scenario with constant population size that corresponds to our original scenario with fluctuating population size, such that both scenarios have the same expected time until pairwise coalescence: $N_e = E[t_{c,2}]$.

Example: seasonal population cycles in fruit flies

Consider a population of fruit flies (modeled under WF assumption) that alternates between a census population size of $N_a = 10^8$ individuals in the summer and $N_b = 10^2$ in the winter. We assume that $t_a = 10$ generations occur in the summer and $t_b = 2$ generations in the winter (nobody really knows what fruit flies do in the winter in colder areas). The coalescence effective population size in this scenario will then be given by:

$$N_e = \frac{10 + 2}{10/10^8 + 2/10^2} \approx 600. \tag{5.10}$$

Note that this effective population size is much closer to the small winter population size than the enormous summer population size, even though winter lasts only 2 generations while summer last 10 generations in our model. This is an important general feature of random genetic drift, which tends to be dominated by the phases of small population sizes when pairwise coalescence rates are high.

The harmonic mean

We can extend the above example of a population alternating between two phases to scenarios that cycle over an arbitrary number of phases, $N_a, N_b, N_c, \ldots$, lasting over times $t_a, t_b, t_c, \ldots$, respectively. The crucial assumption is again that the probability of coalescence within a single cycle remains still small, i.e., population size changes much faster than the time scale over which pairwise coalescence occurs. In this case, the coalescence effective population size will be given by:

$$N_e = \frac{\sum_i t_i}{\sum_i t_i / N_i}. \tag{5.11}$$

Equation (5.11) is known as the harmonic mean of $N(t)$ over the time interval of one cycle. This harmonic mean is quite different from a regular (arithmetic) mean in that it is dominated by the phases where N is small. The regular mean, in contrast, weights large and small phases more equally. The harmonic mean is frequently encountered in cases where averages are taken over different rates, which in our case are the coalescence rates during each phase:

5.4 Maximum likelihood estimation

We have seen above how demographic events, such as bottlenecks and population expansions, can produce characteristic signatures in the patterns of genetic diversity in a population sample. Conversely, by analyzing patterns of genetic diversity, we may be able to draw inferences about the demographic history a population has experienced. A number of powerful statistical frameworks have been developed for this purpose. One of the most widely-used approaches is maximum likelihood (ML) estimation, a very general statistical approach that can also be applied in many other contexts beyond demographic inference.

The maximum likelihood approach is based on the concept of a likelihood function, which specifies the probability of observing some given empirical data under a given evolutionary model, conditional on an unknown parameter p of the model:

$$L(p) = \Pr(\text{data} \,|\, p). \tag{5.12}$$

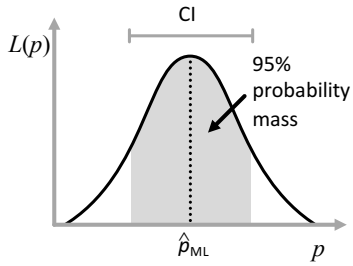
For example, in the case of demographic inference, the data could be the measurement of a summary statistics in a population sample, such as π, S, or the SFS, the model could be a WF population that has experienced a population bottleneck at some point in the past, and the unknown parameter could be the severity of this bottleneck. Note that both the data and the parameter could also be vectors. For example, the data could be measurements of both π and S, and p could be a vector including both the duration of the bottleneck as well as the population size during the bottleneck.

The likelihood function $L(p)$ specifies for each p the probability of observing the given data outcome under the given model. We can then ask which of all possible values yields the highest probability of producing the data we actually observed. This particular value of p is called the maximum likelihood estimator of p:

$$\widehat{p}_{\mathrm{ML}} = \underset{p}{\operatorname{argmax}}\, L(p). \tag{5.13}$$

Intuitively it makes sense to use $\widehat{p}_{\mathrm{ML}}$ as our "best guess" for the actual value of p. Any other choice would have a lower probability of generating the data we actually observed. For discrete parameters p, it can in fact be shown that $\widehat{p}_{\mathrm{ML}}$ always converges to the true parameter value as one increases the sample size over which the data is estimated (and assuming the given model is actually the correct model).

The likelihood function also provides a straightforward way for estimating confidence intervals (CI) for the parameter p. For example, we can find the boundaries $p_{\max}$ and $p_{\min}$ for which the area under the likelihood curve encompasses a given overall probability mass, say 0.95, which would then define the 95% confidence intervals for our parameter p.



Maximum likelihood is an incredibly powerful method for parameter estimation, but also comes with an important caveat: While this approach allows us to infer our best guess for a parameter p under a given model, it does not directly inform us about whether we have actually assumed the correct model. In fact, we will always obtain some maximum likelihood estimate, regardless of how unrealistic our assumed model may be. For instance, the real population in our bottleneck example may have experienced several successive bottlenecks, not just a single one. By wrongly assuming a model with only a single bottleneck, our maximum likelihood estimate of the bottleneck size would be rather meaningless. Statisticians have developed various approaches to address this issue. Goodness of fit tests, for example, can be used to estimate how well the assumed model does actually explain the observed data, and so-called likelihood ratio tests can be used to compare the goodness of fit between different models.

Example 1: estimating allele frequency from a population sample

As a simple example of the maximum likelihood approach, let us consider the problem of estimating the true population frequency (p) of an allele, given that we have observed the allele in k copies in a population sample of size n (our data). Our model is simple binomial sampling with replacement. In this case, we can directly calculate the probability of observing the allele in k copies in our sample, given that the true population frequency is p, which is the binomial probability:

$$L(p) = \Pr(k, n \,|\, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{5.14}$$

The maximum of this likelihood function, and thus our maximum likelihood estimator for p, lies at $\widehat{p}_{\mathrm{ML}} = k/n$, which makes intuitive sense.

Example 2: estimating likelihoods using simulations

The above example was straightforward because we could directly calculate the likelihood function for our model of binomial sampling. This is not always the case. As models become more complex, we can often no longer find an analytic solution for the likelihood function. Only slightly more complex demographic models, such as two bottlenecks or a bottleneck followed by an exponential expansion, typically push us to a point already where we can no longer analytically calculate the likelihood for site-frequency-spectra. However, it is still possible to perform maximum likelihood estimation in such scenarios even in the absence of an analytical likelihood function, as long as we are able to numerically simulate the model.

For example, consider some complex demographic model of which we know all evolutionary parameters, except the duration of one specific bottleneck, which is our unknown parameter ($p = t_t$). We also obtained a population sample from this population, from which which we measured the number of segregating sites ($\text{data} = S_{\text{obs}}$). Further assume we have a numerical simulation that allows us to produce random population samples under the given evolutionary model. In this case, we can estimate the likelihood function numerically by running a large number of these simulations with varying values of p. We can then obtain the ML estimate from this empirical likelihood function. The procedure will be as follows:

1. Define a grid of possible values of t_b that you consider worth testing.

2. Estimate the empirical likelihood function for each of these values as the ratio between the number of all simulation runs (n_{all}) for the given value of t_b, and the number of runs that actually yielded S_{obs} segregating sites (n_{match}):
$$\Pr(S_{\text{obs}} \,|\, t_b) \approx \frac{n_{\text{match}}}{n_{\text{all}}}. \tag{5.15}$$
Obviosuly the more simulations you run, the closer the ratio will converge to the true probability.

3. Find the particular value of t_b that maximizes the empirical likelihood function.

Note that this approach can be easily extended to the simultaneous estimation of several parameters. In this case, we need to run the simulations over a multidimensional grid of parameter values, each dimension corresponding to a separate parameter of the model. Whether such inferences can be performed in reasonable time hinges on the availability of efficient simulation approaches.

5.5 Bayesian inference

In the maximum likelihood approach, all parameter values p are initially assumed to be equally likely, and we then try to find the particular value of p that yields the highest probability of generating the observed data under our given model. Bayesian inference extends this approach, by allowing the parameter to have a non-uniform probability distribution at the outset. This probability distribution can be used to incorporate prior expectations for p, in the sense that we might consider some values to be more likely than others.

For example, imagine we want to infer the true population frequency of a neutral mutation in a WF model. On average, such mutations are more likely to be present at low frequency in the population than at high frequency, according to Equation (4.20). Assume we observe a randomly chosen mutation in $k = 1$ copies in a sample of size $n = 2$. In that case, it would be reasonable to assume that this mutation is actually segregating in the population at a lower frequency than the maximum likelihood estimate $\widehat{p}_{\text{ML}} = k/n = 0.5$, we just happened to catch this particular mutation in one of the two alleles in our sample. We can incorporate such *a priori* expectations into our inference framework using Bayes' theorem

$$\Pr(p \,|\, \text{data}) = \frac{\Pr(\text{data} \,|\, p)\Pr(p)}{\Pr(\text{data})}, \tag{5.16}$$

which we derived in Equation (1.1) in a general probability framework, but here we define events A and B to refer to our parameter p and our data, respectively. The interpretation of the four individual terms in this equation is as follows:

- $\Pr(p \mid \text{data})$: We call this quantity the *posterior probability*. It assigns a probability to each value of p that takes into account the data, as well as our *a priori* expectations for p.
- $\Pr(\text{data} \mid p)$: This is just the regular *likelihood function* we already introduced above.
- $\Pr(p)$: This term is called the *prior probability* for the parameter p, which allows us to incorporate our expectations about the probabilities of observing different parameter values.
- $\Pr(\text{data})$: This so-called *marginal likelihood* specifies the probability of observing the data after the parameter has been integrated out: $\Pr(\text{data}) = \int \Pr(\text{data} \mid p)\Pr(p)dp$. It is also sometimes referred to as the *model evidence*. In practice, this quantity is often ignored, as it is not actually a function of p and therefore does not affect which value of p maximizes the posterior distribution.

Given that the marginal probability is constant with respect to p, the posterior probability is proportional to the product of the likelihood and the prior probability:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{5.17}$$

The posterior probability can thus be interpreted as an "updated" likelihood, which also takes our prior expectations for the probabilities of different values of p into account. The analog to the maximum likelihood estimator in Bayesian inference is the so-called maximum a posteriori estimator, specifying the value of p that maximizes the posterior probability:

$$\widehat{p}_{\text{MAP}} = \underset{p}{\operatorname{argmax}}\, P(p \mid \text{data}). \tag{5.18}$$

Example: estimating allele frequency from a population sample (Bayesian style)

Let us come back to the example of measuring the true population frequency p of a mutation, given that we observed it in k copies in a sample of size n. This time, we want to incorporate *a priori* knowledge in a Bayesian framework. Specifically, we want to assume that the mutation is a neutral mutation in a WF population. In this case, our *a priori* expectation for p will be the expected frequency distribution of such mutations in the population, which according to Equation (4.20) is inversely proportional to the mutation's population frequency: $\Pr(p) \propto 1/p$. The posterior probability distribution is then given by:

$$\Pr(p \mid \text{data}) \propto L(p) \times \frac{1}{p} = \binom{n}{k} p^{k-1}(1-p)^{n-k}. \tag{5.19}$$

The maximum of this function (i.e., our MAP estimate $\widehat{p}_{\text{MAP}}$) will be somewhat smaller than k/n, the original maximum likelihood estimate, consistent with the fact that we attach larger *a priori* probabilities to smaller values of p. The MAP and ML estimates converge towards each other only as the sample size goes to infinity.

Chapter 6

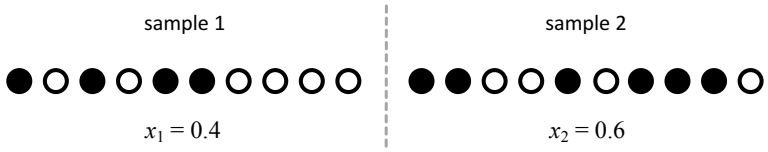
Population structure

The evolutionary models we considered so far all made the assumption that populations are well-mixed, with no subdivisions of any kind. In the haploid WF model, this assumption is constituted by the rule that all individuals have equal probability of being a parent of any given individual in the next generation. In a sexually reproducing species, this additionally implies that there is random mating between individuals. We call populations for which this holds *panmictic*.

Real populations often behave quite differently from a panmictic population. For example, populations could be geographically subdivided, such that individuals from a given region would be more likely to have parents from that same region. Populations could also be subdivided in the sense that ancestry could be influenced by environmental, hereditary, or social factors, in which case we could split a population into different compartments, with individuals from within a compartment being more likely to share ancestry than individuals from two different compartments. The individual populations in such subdivided populations are sometimes called subpopulations or, in a geographical context, demes.

6.1 Quantifying population subdivision

Consider a bi-allelic locus (black, white) in a subdivided population. Assume we observe the black allele at frequency x_1 in a sample from population 1 and at frequency x_2 in a sample from population 2:



We define the heterozygosity in sample i to be $H_i = 2x_i(1 - x_i)$. The average heterozygosity of the two samples is then:

$$H_S = \overline{H_i} = x_1(1 - x_1) + x_2(1 - x_2). \tag{6.1}$$

The total heterozygosity when both samples are pooled together is:

$$H_T = 2\overline{x}(1 - \overline{x}) \quad \text{with} \quad \overline{x} = \frac{x_1 + x_2}{2}. \tag{6.2}$$

The relative difference between H_S and H_T provides a measure for how much the individual frequencies in the two samples differ from each other, which is called Wright's fixation index:

$$F_{ST} = \frac{H_T - H_S}{H_T}. \tag{6.3}$$

In the above example, we have $H_S = 0.4 \cdot 0.6 + 0.6 \cdot 0.4 = 0.48$. Pooling both samples together, we have $\overline{x} = 0.5$, and thus $H_T = 2 \cdot 0.5 \cdot 0.5 = 0.5$. Wright's fixation index F_{ST} for the two samples is therefore:

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{0.5 - 0.48}{0.5} = 0.04. \tag{6.4}$$

6.2 The Wahlund effect

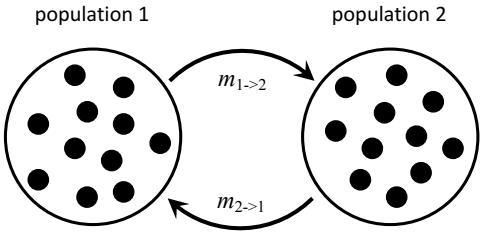
Using the definitions of H_S and H_T it can be easily proven that always $0 \leq H_S \leq H_T \leq 1$, and that equality, $H_S = H_T$, holds if and only if all $x_1 = x_2 = \overline{x}$. Furthermore, F_{ST} is maximal if heterozygosity is zero in each individual sample, i.e. each is fixed for one of the two alleles, while the pooled sample is not fixed. In this case, $F_{ST} = 1$. In summary, we have:

$$F_{ST} = \begin{cases} 0 & x_1 = x_2 \\ > 0 & x_1 \neq x_2 \\ 1 & \text{each sample fixed for different allele} \end{cases} \tag{6.5}$$

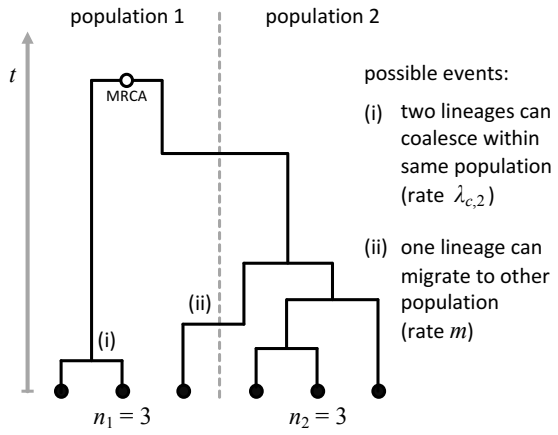
Any form of population subdivision will typically lead to differences in allele frequencies between individual populations. As a result, the average heterozygosity levels within individual populations will be lower compared with the heterozygosity in the population overall. Hence, subdivision should always lead to $F_{ST} > 0$, which is called the *Wahlund effect*. F_{ST} is a commonly used measure for quantifying levels of population subdivision

6.3 The structured coalescent

Population subdivision might not always be strict, and individuals could still occasionally migrate from one population into another. We can account for this process by introducing *gene flow* between populations, which we will model by a Poisson process with migration rates $m_{i \to j}$, specifying the rates at which a randomly chosen allele from population i will migrate into population j per generation. Note that we define migration rates per allele, not individual. When modeling migration events in a diploid population, each migrant individual will bring with it two alleles to the new population. Thus, the migration rate of alleles will be twice as high as the migration rate of individuals in this case.



The coalescence process in a subdivided population is more complicated than in a panmictic population, because we have to account for the possibility that alleles can migrate between populations. We can model this process by allowing lineages to "switch" populations at their given migration rates. We further assume that pairwise coalescence can only occur between individuals that happen to be in the same population. This assumption will always be appropriate in our continuous-time coalescence model, where the probability that two events (migration and coalescence) occur at exactly the same time is zero. We call this extension to subdivided populations with migration the *structured coalescence* process.

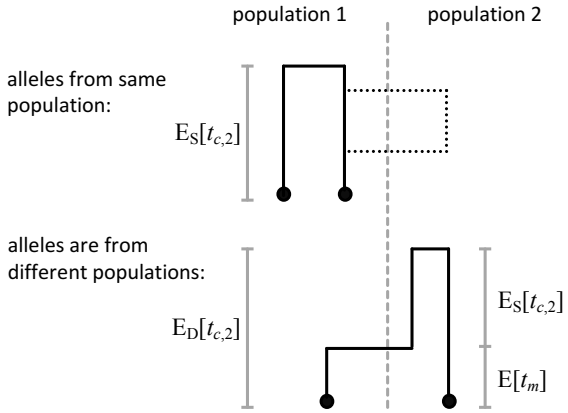


6.4 Coalescence times in subdivided populations

How do population subdivision and migration affect the expected time to pairwise coalescence, $E[t_{c,2}]$, in a sample of two randomly drawn alleles, and thus the expected number of mutations that should have occurred between them? To calculate this, we will consider a simple scenario of two populations, linked by symmetric migration rate m. Within each population, the pairwise coalescence rates is $\lambda_{c,2}$ (same for both populations). We have to distinguish two cases:

1. Both alleles are sampled from the same population (expected coalescence time $E_S[t_{c,2}]$).
2. The alleles are sampled from different populations (expected coalescence time $E_D[t_{c,2}]$).

Note that even if the two alleles are sampled from the same population, it could still be that a lineage has migrated before reaching the MRCA (e.g. dashed lines in illustration below). In fact, there could have been many such migration events prior to coalescence if migration rates are sufficiently large. If the two alleles are sampled from different populations, we know that there must have been at least one such migration event before they can reach their MRCA.



Rather than studying all possible configurations for these trees, we will apply a mathematical trick that will allow us to relate the expected coalescence times $E_S[t_{c,2}]$ and $E_D[t_{c,2}]$ to each other. In the D scenario, where both alleles are sampled from different populations, we know that the first event must be a migration event, as the two lineages cannot coalesce while they are in different populations. The expected waiting time for

such a migration event to happen is $E[t_m] = 1/(2m)$ (the factor two arises because either of the two lineages can migrate). Once this first migration event has occurred, both lineages are in the same population. Thus, we will be in the S scenario, where by definition we will have to wait an additional $E_S[t_{c,2}]$ generations until coalescence. Therefore:

$$E_D[t_{c,2}] = \frac{1}{2m} + E_S[t_{c,2}]. \tag{6.6}$$

Let's consider the scenario where both alleles were sampled from the same population. In this case, both coalescence and migration can occur, and the overall rate at which any of the two events will occur is simply the sum of their individual rates, $2m + \lambda_{c,2}$. Thus, the expected waiting time until the first event occurs (which can be either coalescence or migration), will be $1/(2m + \lambda_{c,2})$. If this is a coalescent event, we are done. But if this a migration event, we will be in the D scenario and thus, by definition, will have to wait another $E_D[t_{c,2},]$ generations until eventual coalescence. Since both coalescence and migration are independent Poisson processes, the probability that the first event is a migration event is $2m/(2n + \lambda_{c,2})$, therefore:

$$E_S[t_{c,2}] = \frac{1}{2m + \lambda_c} + \frac{2m}{2m + \lambda_c} E_D[t_{c,2}]. \tag{6.7}$$

Equations (6.7) and (6.6) build a system of two equations with two unknowns, $E_S[t_{c,2}]$ and $E_D[t_{c,2}]$. We can solve this system by plugging one equation into the other, yielding:

$$E_S[t_{c,2}] = \frac{2}{\lambda_c} \quad \text{and} \quad E_D[t_{c,2}] = \frac{1}{2m} + \frac{2}{\lambda_c}. \tag{6.8}$$

Example: two diploid WF populations

Consider two diploid WF populations with $N_1 = N_2 = N$ and symmetric migration rate m between them. In this case, $\lambda_{c,2} = 1/(2N)$. Therefore, $E_S[t_{c,2}] = 4N$ and $E_D[t_{c,2}] = 1/(2m) + 4N$. Note that $E_S[t_{c,2}]$ is exactly twice the expected pairwise coalescence time of a single diploid WF population of size N.

Interpretation

Our calculation of the expected pairwise coalescence times $E_S[t_{c,2}]$ and $E_D[t_{c,2}]$ in a subdivided population with migration revealed some important results:

1. Surprisingly, $E_S[t_{c,2}]$ does not actually depend on the migration rate. The reason for this lies in a peculiar symmetry of the system. If migration rate is low, the first event will typically be coalescence. However, in those unlikely case where the first event happens to be migration, it will then take a very long until another migration event occurs that brings both lineages back into the same population, which is prerequisite for their coalescence. On the other hand, if migration rate is high, the first event will be migration more often, but it will also take less time for back-migration. Both effects effectively cancel out, such that the expected pairwise coalescence time for two alleles sampled form the same population is actually independent of m.

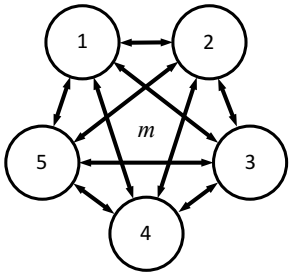
2. In our model with two subpopulations, $E_S[t_{c,2}]$ is simply twice the value of the expectation for a single panmictic population with the same pairwise coalescence rate $\lambda_{c,2}$. In can be shown that this result holds in fact more generally: In a scenario with d populations with symmetric migration rate m between each pair, and equal pairwise coalescence rate $\lambda_{c,2}$ within them, one obtains:

$$E_S[t_{c,2}] = \frac{d}{\lambda_c} \quad \text{and} \quad E_D[t_{c,2}] = \frac{1}{2m} + \frac{d}{\lambda_{c,2}}. \tag{6.9}$$

3. As expected, $E_D[t_{c,2}]$ is larger than $E_S[t_{c,2}]$, with the difference between them getting smaller as migration rate increases. The transition between a subdivided population and an effectively well-mixed population occurs when $md \approx \lambda_{c,2}$.

6.5 Genetic differentiation in an island model

Real populations could be subdivided into many more populations with migration between them. We refer to such models as island models. Consider an idealized island model with $d = 5$ populations. For simplicity, let us further assume that each population has the same pairwise coalescence rate $\lambda_{c,2}$ for alleles within the same population, and that migration occurs at the same rate m between any pair of populations.



Consider a locus evolving under an infinite sites model with nucleotide mutation rate μ in such a population. If we pick two random alleles from the population, these two alleles could have either been picked from the same island (scenario S), or from different islands (scenario D). The expected number of pairwise differences between the alleles in the two scenarios will be given by:

$$E_S[\pi] = 2\mu E_S[t_c] = \frac{2\mu d}{\lambda_{c,2}} \tag{6.10}$$

$$E_D[\pi] = 2\mu E_D[t_c] = 2\mu \left(\frac{1}{2m} + \frac{d}{\lambda_{c,2}} \right). \tag{6.11}$$

The average heterozygosity per site in a sample taken from a single island will therefore be:

$$H_S \approx E_S[\pi] = \frac{2\mu d}{\lambda_{c,2}}. \tag{6.12}$$

If we pool all islands together and pick a random sample from this total population, the probability that two randomly picked alleles will actually be from the same island is $\Pr[S] = 1/d$, while $\Pr[D] = 1 - \Pr[S]$. With these probabilities we can estimate the expected heterozygosity in the total population:

$$\begin{aligned} H_T &= \Pr[S] \times E_S[\pi] + \Pr[D] \times E_D[\pi] \\ &= 2\mu \left[\frac{d}{\lambda_{c,2}} + \frac{1}{2m} \left(1 - \frac{1}{d} \right) \right]. \end{aligned} \tag{6.13}$$

This allows us to calculate the expected level of genetic differentiation between islands, estimated by F_{ST}:

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{(d-1)/d}{(d-1)/d + 2md/\lambda_{c,2}}. \tag{6.14}$$

Example 1: two diploid WF populations with symmetric migration

Consider the above example of two islands ($d = 2$), where each island is a diploid WF population of size N (then $\lambda_{c,2} = 1/(2N)$ within each population). Migration is assumed to occur at symmetric migration rate m between both islands. The expected value of F_{ST} between the two populations is then:

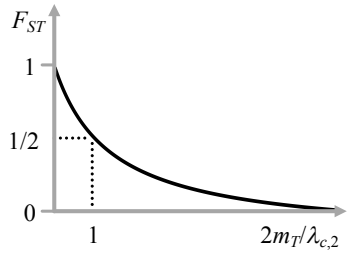
$$E[F_{ST}] = \frac{1/2}{1/2 + 4m/\lambda_{c,2}} = \frac{1}{1 + 8Nm}. \tag{6.15}$$

Example 2: infinitely many islands

As the number of islands in a population increases, the number of possible pairs of islands between which migration could occur grows rapidly (quadratically in d). In this case, it can make more sense to define migration rates in terms of the overall influx and outflow of migrants in each island. Assuming a symmetric migration rate m between each pair, we can define $m_T = (d-1)m$ to be the total rate of migration into and out of any given island per generation. In the limit of a infinitely many islands ($d \to \infty$), while assuming that the total migration rate per island remains constant, we obtain:

$$\lim_{d\to\infty} E[F_{ST}] = \frac{1}{1 + 2m_T/\lambda_{c,2}}. \tag{6.16}$$

This equation shows that genetic differentiation in the island model decays quickly as the total migration rate reaches the same magnitude as the pairwise coalescence rate within each island. Once $m_T > \lambda_{c,2}$, distinguishing the island model from a single well-mixed population will become more and more difficult.

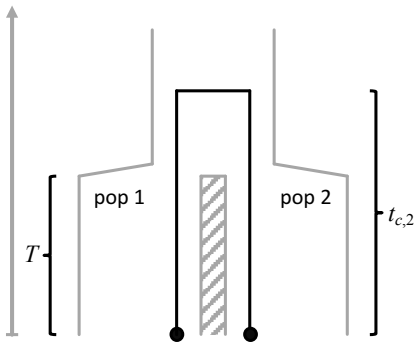


For WF populations, this transition occurs at $Nm \approx 1$, which corresponds to a single allele migrating per island per generation. This shows how tremendously effective migration can be at leveling out genetic differentiation between populations.

6.6 Divergence after a population split

Migration rates between demes in a subdivided population do not necessarily have to remain constant over time. One important counter-example is an originally well-mixed population that becomes subdivided into separate demes with no further migration between the demes after their split. Models of this type are frequently used to describe the process of speciation and we call them *divergence models*.

Consider the following scenario of a population that split into two subpopulation T generations ago with no further migration between the subpopulations after their split. What is the level of genetic differentiation we expect to observe between two alleles sampled from these populations?



For simplicity, let us assume that the pairwise coalescence rate $\lambda_{c,2}$ is the same in each subpopulation and that this was also the pairwise coalescence rate in the original population prior to the split. In this case, two alleles sampled from the same deme will simply coalesce at a constant rate $\lambda_{c,2}$. Two alleles sampled from different demes, however, cannot coalesce more recently than the populations split, since there was no migration after the split. Prior to the split, they also coalesce at rate $\lambda_{c,2}$. The expected time to coalescence for two alleles sampled from the same deme (S) or different demes (D) will thus be:

$$E_S[t_c] = \frac{1}{\lambda_{c,2}} \quad \text{and} \quad E_D[t_c] = T + \frac{1}{\lambda_{c,2}}. \tag{6.17}$$

Using the same assumptions we made above when calculating the expected F_{ST} in the island model, we obtain:

$$H_S \approx E_S[\pi] = \frac{2\mu}{\lambda_{c,2}} \tag{6.18}$$

$$\begin{aligned} H_T &\approx \Pr[S] \times E_S[\pi] + \Pr[D] \times E_D[\pi] \\ &= \frac{1}{2} \times \frac{2\mu}{\lambda_{c,2}} + \frac{1}{2} \times \frac{2\mu}{L}\left(T + \frac{1}{\lambda_{c,2}}\right) \\ &= \frac{2\mu}{\lambda_{c,2}} + \mu T \end{aligned} \tag{6.19}$$

The expected level of genetic differentiation between the two subpopulations will therefore be:

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{T}{T + 2/\lambda_{c,2}}. \tag{6.20}$$

We can see that F_{ST} becomes on the order of one once the split time T becomes on the order of $1/\lambda_{c,2}$ generations. In a diploid WF population of size N, this corresponds to $2N$ generations. We can interpret this result in the following way: two populations that split from each other more than $2N$ generations ago will be almost complete genetically differentiated, i.e., there will no longer remain a lot of shared genetic polymorphism between them.

Chapter 7

Genetic linkage

Chapter 8

Selection

8.1 Evolution by natural selection

In 1859, Charles Darwin published the theory of evolution by natural selection in his famous book "On the origin of species", which is considered by many as the foundation of evolutionary biology. Darwin's theory can be summarized by four basic observations:

1. Phenotypes vary within a population
2. Phenotypic differences can affect survival and reproductive success
3. Phenotypes are (to some extent) heritable
4. Successful phenotypes will become more prevalent in the population over time

Modern evolutionary synthesis

The modern evolutionary synthesis introduces genetics and the laws of heritability into Darwin's theory. In particular, it recognizes that heritable phenotypic variation is due to underlying genetic variation. The modern synthesis also includes random genetic drift as an additional evolutionary process to natural selection, which Darwin had not recognized in his original theory. Major figures in the development of the modern synthesis included Sewall Wright, J. B. S. Haldane, Ronald Fisher, Ernst Mayr, and Theodosius Dobzhansky.

8.2 Fitness in haploids

Consider a locus with two different alleles (A, a) in a haploid population. Let $N_A(t)$ denote the number of individuals that carry allele A in generation t. We define the *absolute fitness* (ω_A) of allele A as the expected growth rate per generation of the number of individuals that carry A:

$$E[N_A(t+1)] = \omega_A \times N_A(t). \tag{8.1}$$

Note that this definition only specifies an expectation value. In a single realization, the actual number $N_A(t+1)$ could be different from the expectation as a consequence of random genetic drift.

This general definition of fitness in terms of expected growth rates combines the various aspects by which natural selection can affect survival and reproductive success into a single quantity. For example, it does not matter whether selection affects the rate at which individuals with allele A produce offspring, or whether it affects the viability of this offspring. Both aspects are integrated in the overall growth rate. Note that the

absolute fitness of an allele does not actually inform us about whether individuals carrying one allele do better than individuals carrying the other allele. The whole population could be growing. In order to compare the fitness between the two alleles at our locus, we define their relative fitness as:

$$\frac{\omega_A}{\omega_a} = 1 + s_{Aa}, \tag{8.2}$$

where s_{Aa} is called the *selection coefficient* of allele A over allele a. The selection coefficient tells us whether carrying allele A provides an advantage, disadvantage, or no difference over allele a with regard to growth rate. Specifically, we denote the following cases:

$$\begin{array}{rcl} s_{Aa} > 0 & : & A \text{ is } \textit{advantageous} \text{ over } a \\ s_{Aa} = 0 & : & A \text{ is } \textit{neutral} \text{ with regard to } a \\ s_{Aa} < 0 & : & A \text{ is } \textit{deleterious} \text{ over } a \end{array}$$

For example, if $s_{Aa} = 0.1$, allele A will on average grow at a ten percent faster rate than allele a. For a given pair of growth rates, ω_A, ω_a, we have $s_{aA} = -s_{Aa}/(1 + s_{Aa}) \approx -s_{Aa}$, where the last approximation holds if $s_{Aa} \ll 1$, which we will typically assume. Note that while we can also define growth rates and selection coefficients for every given point in time, it may not necessarily hold that their values actually remain constant. In fact, these quantities could depend on all sorts of factors, such as the actual frequencies of the alleles in the population (frequency-dependent selection) or the specific point in time (time-dependent selection), etc. However, for now we will assume that at least the selection coefficient between the two alleles (and thus the ratio of their growth rates) is in fact constant. In this case, let us denote the frequency trajectory of A in the population by $x(t)$. The frequency trajectory of a is then $1 - x(t)$. We have:

$$x(t) = \frac{N_A(t)}{N_A(t) + N_a(t)}. \tag{8.3}$$

Given frequency x in generation t, the expected frequency of A in generation $t + 1$ will be:

$$E[x(t+1)] = \frac{\omega_A N_A(t)}{\omega_A N_A(t) + \omega_a N_a(t)} = \frac{(1+s)x}{1+sx}. \tag{8.4}$$

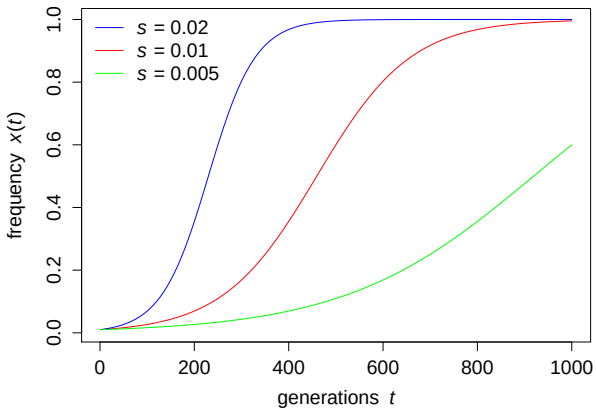
Here we wrote $s_{Aa} = s$, for simplicity. From this result, we can calculate the expected change in the frequency of the A allele over the course of one generation:

$$E[\Delta x] = E[x(t+1)] - x(t) = \frac{sx(1-x)}{1+sx} \approx sx(1-x), \tag{8.5}$$

where the last approximation holds if $s \ll 1$, which we will generally assume. Thus, the expected change in frequency due to selection is a function of both the selection coefficient and the current frequency of the allele. The expected change is maximal when the allele is at intermediate frequency ($x = 1/2$), while it becomes very small when the allele is either at very low or at very high frequency. Equation (8.5) describes a process of so-called "logistic growth". If there were no drift and the frequency of the A allele would follow its expectation value exactly, the solution to this equation would be given by:

$$x(t) = \frac{x_0}{x_0 + (1 - x_0)e^{-st}}. \tag{8.6}$$

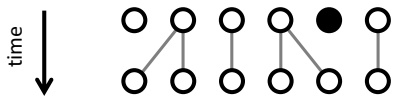
Here x_0 specifies the initial frequency of the allele in generation $t = 0$. The figure below shows the expected trajectories $x(t)$ according to Equation (8.6) for three different selection coefficients, $s = 0.005, 0.01, 0.02$, using a starting frequency of $x_0 = 0.01$:



8.3 Fixation probabilities under selection and drift

Consider a locus with two different alleles (A, a) in a haploid population and let $x(t)$ denote the frequency trajectory of the A allele in the population. We assume the A allele has a selection coefficient $0 < s \ll 1$ over allele a and is initially present in only a single individual in the population.

If $x(t)$ would follow its expectation value from Equation (8.4) exactly, the number of individuals that carry allele A would grow continuously in the population according to Equation (8.5). However, random genetic drift will lead to stochastic fluctuations in $x(t)$ over time. As a result of these fluctuations, the A allele can be lost despite being advantageous. For example, consider the following model in which the black circle denotes the individual with the A allele and white circles denote individuals with the a allele:



Assume we run the experiment 100 times and observe 30 runs in which the A individual does not actually contribute any children to the next generation, 45 runs in which it has one child, and 35 runs in which it has two children. In this case, $\overline{N_A(t+1)} = (40 + 2 \times 40)/100 = 1.2$, consistent with A having a selection coefficient of $s \approx 0.2$. Nevertheless, the allele is still lost quickly in 30% of our runs.

We can estimate the probability that a beneficial allele A with selection coefficient s, which is initially present in a single individual, will be lost in the next generation in a WF model. In order to do this, we first have to modify the WF model such that it takes into account the expected allele frequency change due to selection, which we can do by increasing the probability that the A allele is chosen as a parent. For a neutral allele in a haploid WF population of size N, this probability was $1/N$. For our selected allele, we will simply adjust this probability to $(1 + s)/N$. This ensures that the A allele, on average, has a growth rate of $(1 + s)$ when at low frequency. In this case, the probability that none of the children in the next generation stems from the parent with the A allele will be:

$$\Pr(\text{loss in next generation}) = \left(1 - \frac{1+s}{N}\right)^N \approx e^{-(1+s)}, \tag{8.7}$$

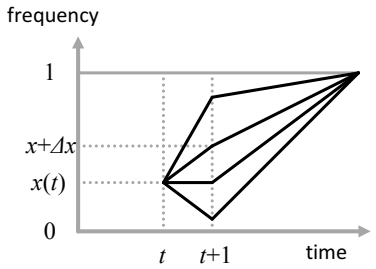
where the last approximation assumes $N \gg 1$. For example, if individuals with the A allele, on average, have 10% more offspring per generation than those with the a allele, the probability that A is lost in the next generation is still $e^{-1.1} \approx 0.33$ – only slightly less than the corresponding probability ($e^{-1} \approx 0.37$) of a neutral

allele. Furthermore, even if A is not lost right away, it still has to survive the second generation, the third generation, and so on. Thus, the overall probability that the allele will be lost is actually much higher.

In the long run, the A allele will either be fixed or lost. To calculate the overall fixation probability, $u(x)$, given an initial frequency x, we first decompose the fixation probability into a sum of transition probabilities over all intermediate frequencies in the next generation:

$$u(x) = \sum_{\Delta x} \Pr(\Delta x) u(x + \Delta x). \tag{8.8}$$

Here $\Pr(\Delta x)$ denotes the probability that A changes in frequency by a given amount Δx over one generation, in which case it will end up at frequency $x + \Delta x$. The fixation probability at this new frequency will be $u(x + \Delta x)$. In our decomposition, we then sum over all possible frequency changes Δx:



As long as frequency changes tend to be small, we can approximate $u(x + \Delta x)$ by a Taylor series expansion:

$$u(x + \Delta x) \approx u(x) + u'(x)\Delta x + \frac{1}{2}u''(x)\Delta x^2. \tag{8.9}$$

Plugging this expansion into the above decomposition of $\pi(x)$ and using the fact that the expectation of a sum is the sum of the expectations, we obtain:

$$\begin{aligned} u(x) &\approx \sum_{\Delta x} \Pr(\Delta x)[u(x) + u'(x)\Delta x + \frac{1}{2}u''(x)\Delta x^2] \\ &= u(x)\sum_{\Delta x} \Pr(\Delta x) + u'(x)\sum_{\Delta x}[\Pr(\Delta x)\Delta x] + \frac{1}{2}u''(x)\sum_{\Delta x}\left[\Pr(\Delta x)\Delta x^2\right] \end{aligned} \tag{8.10}$$

We can identify $\sum_{\Delta x} \Pr(\Delta x) = 1$ and $\sum_{\Delta x}[\Pr(\Delta x)\Delta x] = E[\Delta x]$. Furthermore, $\sum_{\Delta x}\left[\Pr(\Delta x)\Delta x^2\right] = E[\Delta x^2] = V[\Delta x] + E[\Delta x]^2 \approx V[\Delta x]$, where the last approximation holds if the mean and variance of Δx are both small and similar in magnitude. We can then subtract $\pi(x)$ from both sides to obtain:

$$u'(x)E[\Delta x] + \frac{1}{2}u''(x)V[\Delta x] = 0. \tag{8.11}$$

This is a linear differential equation of second order. Since a fixed allele will always remain fixed in our model and a lost allele will always remain lost, we have two boundary conditions:

$$u(0) = 0 \quad \text{and} \quad u(1) = 1. \tag{8.12}$$

The solution to Equation (8.11) with the given boundary conditions is covered in elementary books on differential equations and yields:

$$u(x) = \frac{\int_0^x e^{-B(y)}dy}{\int_0^1 e^{-B(y)}dy} \quad \text{with} \quad B(y) = 2\int_0^y \frac{E[\Delta x]}{V[\Delta x]}dy. \tag{8.13}$$

In Equation (8.5), we already calculate that $E[\Delta x] \approx sx(1-x)$. The variance in the change in allele frequency is determined by drift, $V[\Delta x] = x(1-x)/N_e$, where N_e specifies the variance effective size of the population. Together we obtain:

$$B(y) = 2\int_0^y \frac{sx(1-x)N_e}{x(1-x)} = 2N_e sy \tag{8.14}$$

This finally yields the fixation probability:

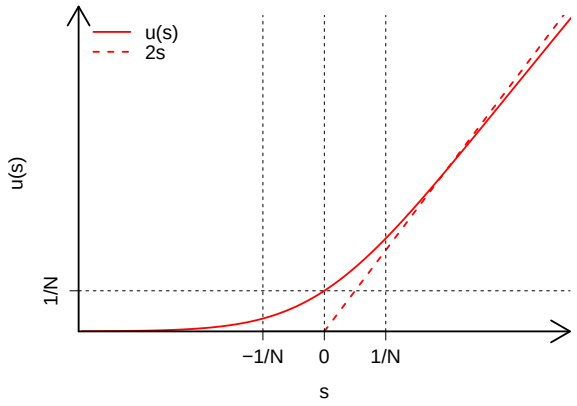
$$u = \frac{1-e^{-2N_e sx}}{1-e^{-2N_e s}}. \tag{8.15}$$

Special case: new mutations in WF model:

In order to achieve a better understanding of this very general result, let us consider the special case of a new mutation with selection coefficient s over the wild type allele, initially present in a single individual ($x = 1/N$) in a haploid WF population of size N. In this case $N_e = N$,and Equation (8.15) thus simplifies to

$$u = \frac{1-e^{-2s}}{1-e^{-2Ns}}. \tag{8.16}$$

We see that the fixation probability of a new mutation depends on both its selection coefficient and the product Ns. Three qualitative regimes with respect to the relation between s and N are typically distinguished:



1. **Strongly advantageous regime ($s \gg 1/N$):**

   In this regime, the denominator of Equation (8.16) approaches one. Therefore, $u \approx 1 - e^{-2s} \approx 2s$, where the last approximation holds under our assumption of $s \ll 1$. The fixation probability of a new mutation is thus simply twice its selection coefficient in this regime. For example, a new mutation with $s = 0.1$ has a fixation probability of $u \approx 2s = 0.2$ (meaning that it still gets lost in 4 out of 5 cases).

2. **Nearly neutral regime ($-1/N < s < 1/N$):**

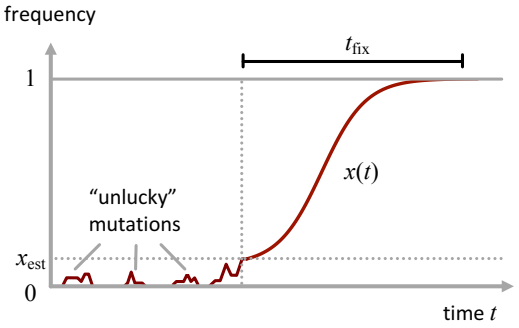
   In this regime, selection is weak compared with the fluctuations caused by random genetic drift and the fixation probability of a new mutation is close to that of a neutral mutation ($u = 1/N$). Note that Equation (8.16) is technically undefined for $s = 0$. However, this discontinuity is easily removables since u converges to $1/N$ from both sides. Population size N impacts the formula for the fixation probability primarily in that it determines which selection coefficients can still be considered "effectively neutral".

3. **Strongly deleterious regime ($s \ll -1/N$):**

   In this regime, the fixation probability decays rapidly as selection coefficients become more deleterious. However, even deleterious mutations still have a non-zero – even though very small – probability of actually fixing in the population.

8.4 The fixation process

In order to illustrate the process by which mutations can become fixed in a population, consider a single locus in a haploid WF population of size N. Every individual initially carries the wild type allele. Mutations can create a new advantageous allele (A) with selection coefficient $s > 0$. When such a new mutation occurs in an individual, the A allele will initially be present at frequency $x = 1/N$ in the population. Most such mutations will be lost quickly due to random genetic drift if $s \ll 1$, which we will assume. The probability that a given mutation escapes loss and eventually reaches fixation is $2s$, twice its selection coefficient.



The reason why most mutations are lost quickly is that when $x(t)$ is still small the fluctuations due to drift can outweigh the tendency of natural selection to increase the frequency of the allele, such that the allele might be lost when drift pushes its frequency to the "absorbing boundary" at frequency zero. However, some mutations may reach higher frequencies in the population, and the higher this frequency, the more likely it is that the mutation will actually go to fixation. For example, consider a mutation that managed to reach a frequency $x = 1/(N_e s)$ in the population. According to Equation (8.15), such a mutation will then fix with probability $\pi(x) \approx 1 - e^{-2} \approx 0.87$, meaning it will fix in almost nine out of ten cases. We call this particular frequency the establishment frequency:

$$x_{\text{est}} = \frac{1}{N_e s}. \tag{8.17}$$

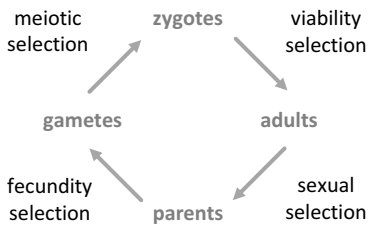
A mutation that has reached establishment frequency is very likely to go all the way to fixation and it will do so, on average, following the logistic growth trajectory described in Equation (8.6) with initial value $x_0 = x_{\text{est}}$. We can thus calculate how long it will take until the allele is expected to be present in half of the population, $x(t_{1/2}) = 0.5$, which yields $t_{1/2} = \log(N_e s)/s$. The expected time until the allele will have gone all the way to fixation will be around twice this value:

$$E[t_{\text{fix}}] \approx \frac{2\log(N_e s)}{s}. \tag{8.18}$$

For example, given a selection coefficient of $s = 0.1$ and $N_e = 10^4$, we have $\log(N_e s) \approx 7$, and thus $t_{\text{fix}} \approx 140$ generations. If selection were one hundred times weaker than this, $s = 0.001$, we would obtain $\log(N_e s) \approx 2$, and thus $t_{\text{fix}} \approx 4000$ generations. Note that the logarithm does not matter much here, as it grows very slowly.

8.5 Selection in diploids

Fitness and selection in diploids is somewhat more complicated than selection in haploids because individuals do not just carry one allele per locus, but two alleles. Each possible genotype could be associated with a different fitness. It therefore becomes important during which stage of the life cycle natural selection operates. We classify the different life stages of a sexually reproducing diploid organism into four phases, and selection can operate along each transition:



1. Viability selection: affects which zygotes survive into adulthood ("struggle for existence").
2. Sexual selection: affects the chances of mating and number of mates per individual.
3. Fecundity selection: affects the number of gametes produced by each mating pair.
4. Meiotic selection: affects the probability that gametes can successfully fuse to form zygotes.

Viability selection

Viability selection most closely resembles Darwin's original idea of selection resulting from differences among individuals in their abilities to compete over limited resources, leading to the survival of the fittest. Darwin attributed the differences in survival to differences in phenotypes. In the modern synthesis, we associate at least some of these differences with genetic differences.

At a bi-allelic locus (A, a) there are three different possible genotypes in diploids: aa, aA, and AA. Each genotype could have a different chance of surviving from the zygote stage to adulthood. We specify the respective survival probabilities by the so-called *viability coefficients*:

$$\begin{aligned} \nu_{AA} &= \Pr(AA \text{ individual survives to adulthood}) \\ \nu_{Aa} &= \Pr(Aa \text{ individual survives to adulthood}) \\ \nu_{aa} &= \Pr(aa \text{ individual survives to adulthood}) \end{aligned} \tag{8.19}$$

Given that viabilities are probabilities, they have to be smaller than or equal to one. In contrast to the absolute fitness values we used in haploids, viabilities thus cannot inform us about absolute growth rates, which will also depend on how many zygotes are actually produced in the population per generation.

Under the assumption of random mating, we can calculate the expected genotype frequencies in zygotes in generation $t+1$, given the frequency x of the A allele in adults in generation t:

$$\begin{aligned} E[x_{AA}(t+1)] &= \frac{\nu_{AA}}{\overline{\nu}} x^2 \\ E[x_{Aa}(t+1)] &= \frac{\nu_{Aa}}{\overline{\nu}} 2x(1-x) \\ E[x_{aa}(t+1)] &= \frac{\nu_{aa}}{\overline{\nu}} (1-x)^2. \end{aligned} \tag{8.20}$$

Here $\overline{\nu} = \nu_{AA}x^2 + \nu_{Aa}2x(1-x) + \nu_{aa}(1-x)^2$ specifies the mean viability across all individuals in the population in generation t. Each AA heterozygote in generation $t+1$ will contribute two A alleles to the population, whereas each heterozygote will contribute only one. The expected frequency of the A allele in generation $t+1$ is therefore:

$$\begin{aligned} E[x(t+1)] &= E[x_{AA}(t+1)] + \frac{1}{2}E[x_{Aa}(t+1)] \\ &= \frac{\nu_{AA}x^2 + \nu_{Aa}x(1-x)}{\overline{\nu}}. \end{aligned} \tag{8.21}$$

This expectation value is the diploid analog under viability selection to Equation (8.4) in haploids. We can use this result to calculate the frequency trajectory $x(t)$ under the assumption that there is no drift, such that the population frequency of the A allele would follow its expectation value exactly. In contrast to the simple haploid scenario, where beneficial alleles would continuously increase in frequency until they become fixed in the population, the diploid case is more complicated and can result in qualitatively different outcomes, such as balancing selection, where the A allele is maintained at an intermediate frequency. We will discuss some of these scenarios in more detail below.

Equation (8.21) shows that only the ratios of viabilities matter for determining the expected change in allele frequencies over time. We can again express these ratios using the concept of a selection coefficient. However, because there are three possible genotypes in diploids, it is not enough to simply specify one selection coefficient. Instead, we define:

$$\frac{\nu_{AA}}{\nu_{aa}} = 1 + s \quad \text{and} \quad \frac{\nu_{aA}}{\nu_{aa}} = 1 + hs, \tag{8.22}$$

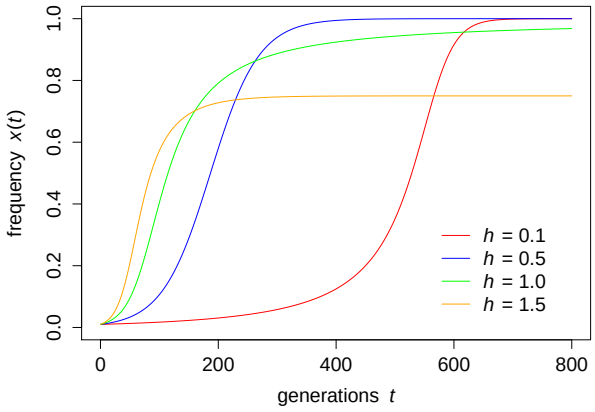
where h is called the dominance coefficient of allele A over allele a. Thus, the selection coefficient specifies the relative viability of the AA homozygote over the aa homozygote, and h specifies how much of this difference between the two homozygotes is already observed in the comparison of aA heterozygotes over aa homozygotes. The value of h for a given pair of alleles has important ramifications on the expected frequency trajectory of the alleles in the population. We generally distinguish the following cases:

$h = 0$	:	A is *recessive* to a	directional selection
$0 < h < 1$	:	A and a are *codominant*	
$h = 1/2$	:	A and a are *additive*	
$h = 1$	:	A is *dominant* to a	
$h > 1$	:	A is *overdominant* to a	balancing selection
$h < 0$	:	A is *underdominant* to a	

If A is recessive ($h = 0$) to a, there is no difference in viability between the aA and aa genotypes. Hence, even if A is beneficial ($s > 0$), this advantage will only be felt once there are actually AA homozygotes in the population. Analogously, recessive deleterious alleles only carry a cost in homozygotes but not yet in heterozygotes. In contrast, dominant alleles ($h = 1$) show all their effects on viability already in heterozygotes. Codominant alleles fall somewhere between these two extremes. A special case of codominance is additivity ($h = 1/2$), where the viability of the heterozygote lies exactly in the middle between the two homozygotes.

Whenever $0 \leq h \leq 1$ an A allele with $s > 0$ will still tend to grow continuously in frequency until fixation (conditional on that it successfully overcomes random genetic drift). This is qualitatively different for overdominance ($h > 1$) and underdominance ($h < 0$). An overdominant beneficial allele actually provides the largest benefit when present in heterozygotes. Under such so-called *heterozygote advantage*, the allele will not actually be driven to fixation, since homozygotes have lower viability than heterozygotes. Instead, selection will try to maintain the allele at an intermediate frequency. We call such a scenario *balancing selection*, in contrast to the standard *directional selection*, where selection drives the allele either to fixation or to loss, depending on its selection coefficient. For an underdominant A allele, the viability of the aa homozygote is actually

between that of the heterozygote and the AA homozygote. In that case the allele could be beneficial ($s > 0$) but the heterozygote would actually do worse than the aa homozygote, effectively preventing this allele from invading the population. The figure below shows the expected trajectories $x(t)$ according to Equation (8.21) for a selection coefficients $s = 0.05$ and starting frequency $x_0 = 0.01$ under four different dominance scenarios:



Dominance coefficients also have important consequences for the establishment probabilities of new mutations in a population, i.e., the probability that a new mutation survives early stochastic loss to random genetic drift. Note that in diploids, due to the possibility of balancing selection, not all mutations that successfully establish will ultimately go to fixation. Thus, in contrast to haploids, the establishment probability is no longer the same as the fixation probability.

To the extent that we can assume Hardy-Weinberg-Equilibrium to hold in our population, a new mutation will initially be present exclusively in heterozygotes. This is because when x is still really small, the expected frequency of heterozygotes, $2x(1-x)$, is much larger than x^2, the expected frequency of homozygotes. As a result, the establishment probability of a new mutation will be determined by the viability of the heterozygote. Assuming the mutation initially arises in a single copy, $x_0 = 1/(2N)$, in a WF model with $N_e = N$, we obtain:

$$\Pr(\text{new mutation successfully establishes}) = \frac{1 - e^{-2hs}}{1 - e^{-4Ns}} \approx 2hs. \tag{8.23}$$

A beneficial mutation therefore has a higher probability of establishing in the population if it is dominant. A recessive mutation, on the other hand, will have similar probability of being lost to drift as a neutral mutation.

Chapter 9

Molecular evolution

9.1 The neutral theory of molecular evolution

There has been a long-standing debate about whether the interplay between mutational processes and random genetic drift suffices to describe most of the patterns of molecular genetic variation we can observe in real biological populations. The neutral theory of molecular evolution posits that the vast majority of genetic variation is in fact selectively neutral, and that DNA substitutions between species are therefore primarily the result of drift. In contrast, selectionists argue that natural selection plays an important role in maintaining genetic diversity within populations and driving substitutions between species. The mastermind behind the development of neutral theory was the famous Japanese population geneticist Motoo Kimura, who introduced this theory in 1968. His ideas were revolutionary at the time and sparked a fierce debate between neutralist and selectionists over the following years.

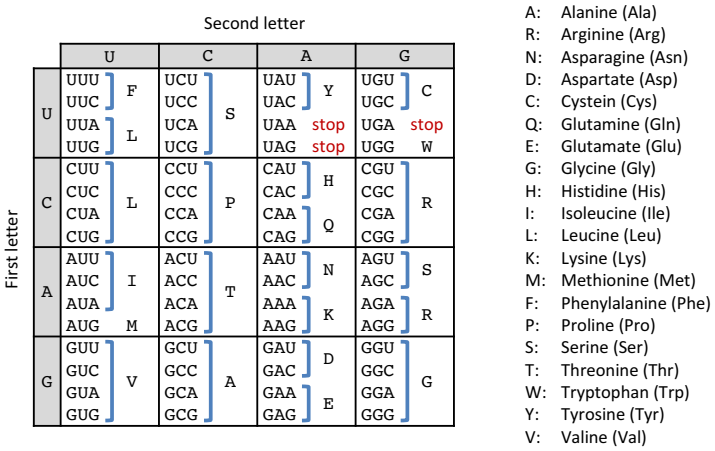
The neutral theory has been tremendously influential in helping us understand how stochastic processes can contribute to evolutionary dynamics. However, with the advent of population genomic data, it became increasingly clear that natural selection does often play an important role in molecular evolution. Moreover, there is growing evidence that the pillar of neutral theory – random genetic drift – may not even always be the dominant stochastic process acting on neutral genetic variation, as it could frequently be outrivaled by other stochastic processes such as genetic draft, which describes allele frequency changes due to linkage with nearby selected mutations. Nevertheless, the neutral theory has firmly established itself as our standard null model in population genetics and will likely continue to play an important role as the baseline scenario for the development of improved evolutionary theories.

9.2 Inferring selection with dN/dS-type tests

One key prediction of neutral theory is that the substitution rate (d) per nucleotide site in the population should equal the mutation rate (μ) at which new mutations arise in individuals, as we derived in Equation (3.2). In a genomic region where all new mutations are strictly neutral, we thus expect to observe $d \approx \mu$. If we observe a significantly lower rate of substitution in a given genomic region, this would be indicative of the presence of at least some *negative selection* (sometimes also called *purifying selection*) because for some mutations the probability of becoming fixed in the populations must have been lower compared with that of neutral mutations. On the other hand, if we observe a significantly higher rate of substitution in a given region, this would be indicative of at least some *positive selection* having acted in that region, such that certain mutations fixed with a higher probability than expected under neutrality. Thus, by comparing estimates of mutation rates with estimates of substitution rates for a given genomic region, we can in principle infer the presence of positive and negative selection in such a region.

Rather than comparing substitution rates with mutation rates, which can be difficult to measure, we can make use of the fact that we expect $d \approx \mu$ in a region we know is in fact evolving neutrally. Thus, if we can find such a region and assume that mutation rates are constant along the genome, we can simply use the substitution rate in this neutral reference region as a proxy for the genome-wide mutation rate. We can then compare the substitution rate we observed in a test region with the substitution rate we observed in our neutral reference region to infer the presence of selection in the test region. This approach has the additional advantage that any systematic errors in the measurement of substitution rates, such as uncertainty of the exact divergence time between two species when such rates are inferred from a sequence alignment, are expected to affect both d and d_0 in similar ways, and thus effectively cancel out in the ratio d/d_0. Of course this approach hinges on our ability to find a suitable reference region that is indeed evolving neutrally.

One commonly used strategy for this is to take synonymous sites in a protein-coding region (i.e., sites where mutations will not actually change the encoded amino acid) as our neutral reference region. Because such synonymous mutations do not change the amino acid sequence of the protein, one often assumes that they will have no measurable effect on phenotype, and therefore should be selectively neutral (this assumption is obviously debatable). The *genetic code* informs us about which specific nucleotide mutations in a protein-coding DNA sequence will be synonymous and which will change the amino acid:



A: Alanine (Ala)
R: Arginine (Arg)
N: Asparagine (Asn)
D: Aspartate (Asp)
C: Cystein (Cys)
Q: Glutamine (Gln)
E: Glutamate (Glu)
G: Glycine (Gly)
H: Histidine (His)
I: Isoleucine (Ile)
L: Leucine (Leu)
K: Lysine (Lys)
M: Methionine (Met)
F: Phenylalanine (Phe)
P: Proline (Pro)
S: Serine (Ser)
T: Threonine (Thr)
W: Tryptophan (Trp)
Y: Tyrosine (Tyr)
V: Valine (Val)

Note that not all sites in a protein-coding region are strictly synonymous or non-synonymous. At some sites, a subset of all possible mutations will result in a change of the amino acid, while the other mutations will not. For example, consider the codon UUA, which encodes for Leucine. A mutation A→G at the third position of this codon will not change the amino acid, but a mutation A→C will. However, for some amino acids, such as Alanine, no mutations at the third codon position will ever change the amino acid. We call such sites in a protein-coding region the *4D synonymous sites*.

4D synonymous sites thus provide a set of genomic positions we can use as our neutral reference against which to compare our test region we want to detect selection in. In the classical dN/dS test, the test region is constituted by the non-synonymous sites themselves. That is, we ask whether mutations that change an amino acid have experienced positive or negative selection over the course of evolution, or whether such mutations were selective neutral. To do this, we measure the rate dS of synonymous substitutions at 4D synonymous sites and compare it with the rate dN of non-synonymous substitutions at non-synonymous sites (for example by studying the pairwise sequence alignment of the two othologous copies of a gene of interest in two species).

Note that the genetic code is structured in a way that any mutation at the second position in a codon does in fact always change the amino acid, so second codon positions are always strictly non-synonymous sites. There are also only a few amino acids where a mutation at the first codon position does not necessarily change

the amino acid. For the sake of simplicity, we will simply neglect these cases in our examples and just assume that, regardless of the actual codon, the first and second codon positions are always strictly non-synonymous sites. The ratio dN/dS then provides a test for selection at non-synonymous sites that can inform us about whether selection has favored or disfavored amino acid changes in the particular region.

$dN/dS \approx 1$: compatible with neutrality
$dN/dS < 1$: some negative selection has been acting on non-synonynmous mutations
$dN/dS > 1$: some positive selection has been acting on non-synonynmous mutations.

For example, consider the following alignment of two orthologous mRNA sequences taken from two species that shared a common ancestor in the past. We assume that all of the differences we see between these two sequences are actually fixed in their populations (i.e., each such difference reflects a mutation that has occurred in one of the species and then became fixed over the course of evolution). We also assume that there were no additional fixations occurring we no longer see because of subsequent fixations:

```
mRNA1:      ACUCCGAACGGGGCGCCA
AA seq:     -T--P--N--G--A--P-

mRNA2:      ACGCCGAUCGGCGCGACA
AA seq:     -T--P--I--G--A--T-

S sites (4D): 001001000001001001
N sites:      110110110110110110
```

In this example we observe two synonymous substitutions and two non-synonynmous substitutions. Overall, there are five 4D synonymous sites in the sequence and 12 non-synonymous sites, yielding:

$$dN = 2/12 \quad \text{and} \quad dS = 2/5 \quad \Rightarrow \quad dN/dS = 5/12 \approx 0.41 < 1. \tag{9.1}$$

Thus, we conclude that at least some purifying was likely acting on non-synonymous mutations on this locus (here we also assume that our result is actually significantly different from $dN/dS = 1$, which is probably not the case given the small numbers). Some general remarks on the dN/dS test:

1. A ratio $dN/dS < 1$ does not exclude that there was also some positive selection. The result simply means that we observed fewer non-synonymous mutations than we would expected under neutrality. Some negative selection has therefore been acting. However, unless $dN = 0$, those non-synonymous mutations that did occur could still have been driven by positive selection. Analogously, a dN/dS-ratio larger than one does not exclude that there was also some negative selection.

2. The dN/dS ratio is sometimes referred to as the K_a/K_s ratio, where the letter a stands for amino acid-changing mutations and the letter s stands for synonymous mutations. Both of these notations specify esentially the same thing.

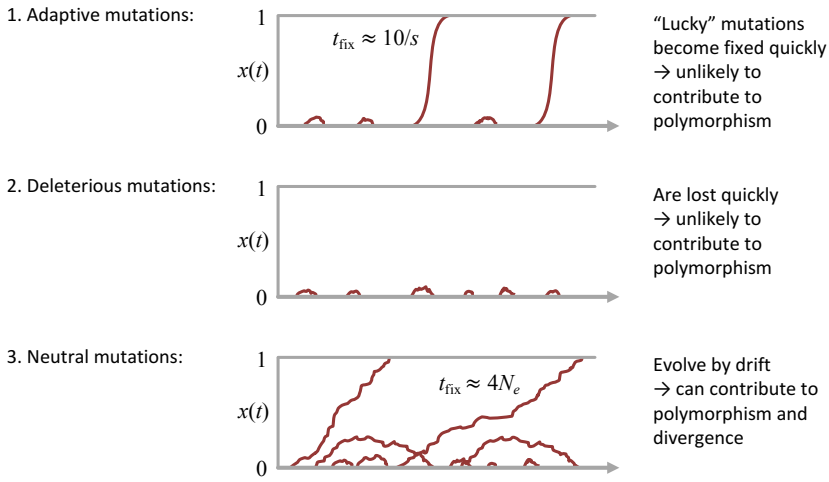
3. We can estimate dN/dS for a genomic region using a pairwise alignment from two species. For instance, we could measure dN/dS from the mRNA alignment of the orthologous copies of a gene in human and mouse to infer selection in this gene from the human-mouse ancestor until the present. Note that the test will not allow us to distinguish whether selection was acting in one or the other lineage, or in both. This would require that we were able to measure the substitution rates along each lineage separately.

4. If we have a multiple sequence alignment between many species that is deep enough such that under neutrality we expect several substitutions to have occurred at each nucleotide site along their phylogeny, we can then estimate dN/dS ratios for individual sites in the alignment. This has become an important approach for identifying functionally relevant sites in a protein, and sophisticated statistical methods have been developed for this purpose (e.g. phylogenetic analysis by maximum likelihood, PAML).

9.3 The McDonald-Kreitman test

The dN/dS test can only provide us with information of the kind "some negative/positive selection must have been acting in the test region". Unfortunately it does not tell us whether maybe both were acting, or what fraction of the substitutions that did occur were actually driven by positive selection. In order to answer such questions, we need to disentangle the individual contributions of positive selection, negative selection, and neutral evolution to the overall rate of substitution in the test region. This can be achieved by the McDonald-Kreitman (MK) test, named after its two inventors John McDonald and Martin Kreitman. The MK tests utilizes information on both substitution rates and levels of polymorphism.

Consider a test region and a neutral reference region that both have the same mutation rate μ per site. In the test region, some unknown fraction ρ of new mutations will be neutral, some will be adaptive, and the rest will be deleterious. In the neutral reference region, all mutations are assumed to be neutral. There are a number of additional assumptions we need to make about the mutations in the test region: First, we assume that the adaptive mutations that successfully establish in the population go to fixation quickly, such that we are unlikely to catch them as polymorphisms in a population sample. Second, we assume that the deleterious mutations are strongly deleterious and are therefore lost quickly. In this case, we are also unlikely to catch deleterious mutations in a population sample. Finally, the neutral mutations in the test region will be subject to drift similar to the mutations in the neutral reference region. The neutral mutations in the test region will thus contribute to both polymorphism and divergence. However, we expect to observe fewer such neutral mutations in the test region compared with the reference region, since only a fraction ρ of the new mutations in the test region are actually neutral. The following picture illustrates the trajectories of mutations in the test region under these assumptions:



To measure the level of polymorphism in the test region, we can simply count the number of segregating sites in a population sample. Let p and p_0 denote the observed numbers of segregating sites (per sequenced nucleotide site) in a population sample in the test region and the neutral reference region, respectively. Under our assumptions, we expect that $p/p_0 = \rho$, since only a fraction ρ of the mutations in the test region are actually neutral, and these mutations are the only ones that we expect to contribute to polymorphism. Thus, we can use the ratio p/p_0 as a proxy for the fraction of new mutations in the test region that are neutral.

Substitution rates in the test and reference region can be measured from sequence alignments in the same way as we did for the dN/dS test. In the neutral reference region, we expect the substitution rate per site to equal the mutation rate: $d_0 = \mu$. The substitution rate in the test region can be written as the sum of the contributions from adaptive substitutions d^+, and neutral substitutions, which should occur at rate $\rho\mu$ in the

test region:

$$d = d^+ + \rho\mu. \tag{9.2}$$

Using our proxies $\rho = p/p_0$ and $\mu = d_0$, the fraction of substitutions in the test region that were adaptive (α) is then given by:

$$\alpha = \frac{d^+}{d} = 1 - \frac{p}{p_0}\frac{d_0}{d}. \tag{9.3}$$

This allows us to estimate how much adaptation has occurred in the test region by combining information about polymorphism levels and divergence.

Example: MK-test for protein evolution

We can apply the MK-test to protein evolution by using non-synonymous sites as our test region and synonymous sites as out neutral reference region. In this case, $d = dN$ and $d_0 = dS$, both can be estimated in the region of interest from a pairwise mRNA alignment. We can then estimate pN and pS in the same region as the numbers of non-synonymous and synonymous SNPs observed in a population sample, yielding:

$$\alpha \approx 1 - \frac{pN}{pS}\frac{dS}{dN}. \tag{9.4}$$

Note that the overall number of N and S sites in the region cancels out in the ratio. Hence, we do not actually have to normalize dN, dS, pN, and pS by the overall numbers of N and S sites in the region. Instead, we can simply estimate dN and dS as the actual numbers of N and S differences observed in the alignment, and estimate pN and pS as the actual numbers of N and S segregating sites observed in the sample.

Example: Rejecting neutrality

The original introduction of the MK test did not actually focus on calculating α, but was simply a test whether we can reject the null-hypothesis that $\alpha = 0$. This is equivalent to asking whether dN/dS is significantly different from pN/pS. Given a set of four observed numbers dN, dS, pN, and pS, we can answer this question using a chi-square test on the two-by-two contingency table:

	synonymous	non-synonymous
divergence	dS	dN
polymorphism	pS	pN

For example, in the original MK paper the values $dS = 17$, $pS = 42$, $dN = 7$, and $pN = 2$ were measured for the *Adh* locus (which encodes alcohol dehydrogenase) in *Drosophila*. The chi-squared test rejects the null hypothesis that $E[dN]/E[dS] = E[pN]/E[pS]$ with $p = 0.0128$.

The asymptotic MK test:

The key problem of the MK test lies in its assumption that there are only strongly deleterious mutations, which do not contribute to polymorphism. If slightly or moderately deleterious mutations are common, these mutations could contribute noticeably to levels of polymorphism even though they would still be unlikely to fix in the population, thereby biasing estimates of α downwards. One strategy to address this problem is to simply exclude polymorphisms where the derived allele is below a certain cut-off frequency. The higher one choses this cut-off frequency, the lower the proportion of slightly deleterious polymorphisms in the sample. However, higher cutoffs will also result in fewer polymorphism that remain for inference, thereby increasing noise.

The asymptotic MK test provides a more elegant solution to this problem by studying how the inferred values of α change when focusing on polymorphisms in specific derived-allele frequency classes separately.

Let $pN(x)$ and $pS(x)$ denote the measured levels of non-synonymous and synonymous polymorphism when considering only those polymorphisms where the derived allele is present at frequency x. In practice, one will typically obtain these estimates over discrete frequency bins. For example, when using ten frequency bins, x_1 could specify all SNPs with derived allele frequency in the range $0 < x \leq 0.1$, etc. By summing over all frequency classes, we regain the overall polymorphism levels: $pN = \sum_i pN(x_i)$ and $pS = \sum_i pS(x_i)$.

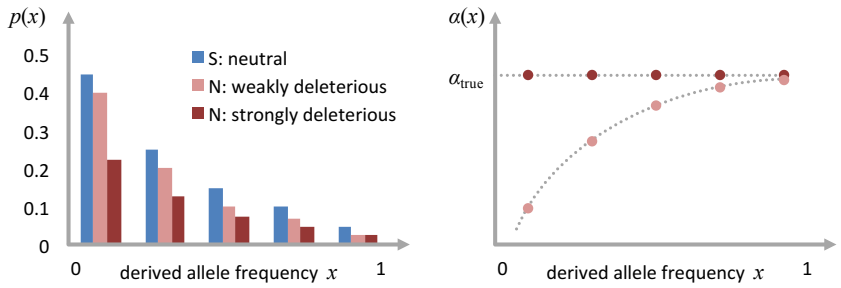
Importantly, slightly or moderately deleterious mutations can inflate $pN(x)$ for low frequencies x, but this effect should decrease as x increases, because selection will tend to prevent these mutations from reaching higher frequencies. Consequently, we expect that the ratio $pN(x)/pS(x)$ should overestimate the fraction of non-synonymous mutations that are neutral (ρ) only for small x, but should approach the correct value of ρ as x becomes larger. To make use of this insight, let us define

$$\alpha(x) = 1 - \frac{dN}{dS}\frac{pN(x)}{pS(x)} \tag{9.5}$$

as the value of α we obtain when considering only the SNPs with derived allele within frequency-class x. The value of $\alpha(x)$ should then converge to the true value as x approaches one:

$$\alpha_{\text{true}} = \lim_{x \to 1} \alpha(x). \tag{9.6}$$

In practice, estimation of this limit will be complicated by the fact that one will typically find only very few polymorphisms at high derived allele frequencies. It therefore makes sense to estimate $\alpha(x)$ for all frequency classes and then infer the value at $x \to 1$ by extrapolating a curve that was fitted using data from all classes:



Note that the shape of this curves also bears information about the extent to which deleterious mutations actually contribute to $pN(x)$. If all polymorphism were in fact neutral, $\alpha(x)$ should be approximately constant over all values x, whereas the presence of deleterious mutations will result in lower values of $\alpha(x)$ for smaller x. To obtain the asymptotic value of $\alpha(x)$ in the limit $x \to 1$, one can try to fit an exponential function of the form $\alpha(x) \approx a + b \exp(-cx)$ to the data. This makes intuitive sense for the case where deleterious mutations all have the same selection coefficient, and the contribution of deleterious mutations to levels of non-synonymous polymorphisms should thus decay approximately exponentially with increasing frequency. However, it is not exactly clear which functional form should be fitted in scenarios where selection coefficients are drawn from a broader distribution, unless the actual distribution of fitness effects were known.