5 The coalescent

Until now, in our outline of the Wright-Fisher model, we have shown how to predict the state of the population in the next generation $(t + 1)$ given that we know the state in the current generation (t). This is the classical approach in population genetics and follows the evolutionary process forward in time. This view is most useful if we want to predict the evolutionary outcome under various scenarios of mutation, selection, population size and structure, etc. that enter as parameters into the model. However, these model parameters are not easily available in natural populations. Usually, we rather start out with data from a present-day population. In molecular population genetics, this will be mostly sequence polymorphism data from a population sample. The key question then becomes: What are the evolutionary forces that have shaped the observed patterns in our data? Since these forces must have acted in the history of the population, this naturally leads to a genealogical view of evolution backward in time. This view in captured by the so-called coalescent process (or simply *the coalescent*), which has caused a small revolution in molecular population genetics since its introduction in the 1980's. There are three main reasons for this:

- The coalescent is a valuable mathematical tool to derive analytical results that can be directly linked to observable data.
- The coalescent leads to very efficient simulation procedures.
- Most importantly, the coalescent allows for an intuitive understanding of patterns in DNA polymorphism data and of how these patterns result from evolutionary processes.

For all these reasons, we will introduce this modern backward view of evolution in parallel to the classical forward picture.

The coalescent process describes the genealogy of a population sample. The key event of this process is therefore that, going backward in time, two or more individuals share a common ancestor. We can ask, for example: what is the probability that two individuals from the population today (t) have the same ancestor in the previous generation $(t - 1)$? For the neutral Wright-Fisher model, this can easily be calculated because all individuals pick a parent at random. If the population size is $2N$ the probability that two individuals choose the same parent is

$$p_{c,1} = \Pr[\text{common parent one generation ago}] = \frac{1}{2N}. \tag{5.1}$$

Given the first individual picks its parent, the probability that the second one picks the same one by chance is 1 out of $2N$ possible ones. This can be iterated into the past. Given that the two individuals did not find a common ancestor one generation ago maybe they found one two generations ago and so on. We say that the lines of descent from the two

individuals *coalescence* in the generation where they find a common ancestor for the first time. The probability for coalescence of two lineages exactly t generations ago is therefore

$$p_{c,t} = \Pr\left[\begin{array}{l}\text{two lineages coalesce}\\ t \text{ generations ago}\end{array}\right] = \frac{1}{2N}\Big(1-\frac{1}{2N}\Big)^{t-1}.$$

Mathematically, we can describe the *coalescence time* as a random variable that is geometrically distributed with success probability $\frac{1}{2N}$. Figure 5.1 shows an example for the common ancestry like it can be generated by a simulation animator, such as the Wright-Fisher animator on www.coalescent.dk. In this case the history of just two individuals is highlighted. Going back in time there is always a chance that they choose the same parent. In this case they do so after 11 generations. In all the generations further back in time they will automatically also have the same ancestor. The common ancestor in the 11th generation in the past is therefore called the *most recent common ancestor* (MRCA).

The coalescence perspective is not restricted to a sample of size two but can be applied to any number of individuals. For a sample of size n from the Wright-Fisher model of size $2N$, the probability of coalescence in a single generation is

$$\begin{aligned} p_{c,1}^{(n)} &= 1-\Big(1-\frac{1}{2N}\Big)\cdot\Big(1-\frac{2}{2N}\Big)\cdots\Big(1-\frac{n-1}{2N}\Big) = 1-\prod_{i=1}^{n-1}\Big(1-\frac{i}{2N}\Big)\\ &= \frac{1}{2N}\sum_{i=1}^{n-1} i + \mathcal{O}\Big[\Big(\frac{n}{N}\Big)^2\Big] = \frac{1}{2N}\binom{n}{2} + \mathcal{O}\Big[\Big(\frac{n}{N}\Big)^2\Big]. \end{aligned} \tag{5.2}$$

We can interprete this result as follows. In a sample of size n, there are $\binom{n}{2}$ possible coalescence events between pairs of individuals. If we assume that $n \ll N$, multiple coalescence events in a single generation can be ignored and the leading order term in $p_{c,1}^n$ just accounts for the probability of a single pairwise coalescence event in the sample in the previous generation. Multiple coalescence events and coalescence events of more than two lineages simultaneously (so-called "multiple mergers") only contribute to the error term $\sim \mathcal{O}[N^-2]$, which can be ignored for small samples in a large population. In this approximation, the coalescence probability after t generation in a sample of size n becomes

$$p_{c,t}^{(n)} \approx \frac{1}{2N}\binom{n}{2}\cdot\left(1-\frac{1}{2N}\binom{n}{2}\right)^{t-1}. \tag{5.3}$$

We can then construct the genealogical history of the sample in a two-step procedure:

1. First, fix the topology of the coalescent tree. I.e., decide (at random), which pairs of genealogical lineages from individuals in a sample coalesce first, second, etc., until the MRCA of the entire sample is found.

2. Second, specify the times in the past when these coalescence events have happened. I.e., draw a so-called coalescent time for each coalescent event. This is independent of the topology.

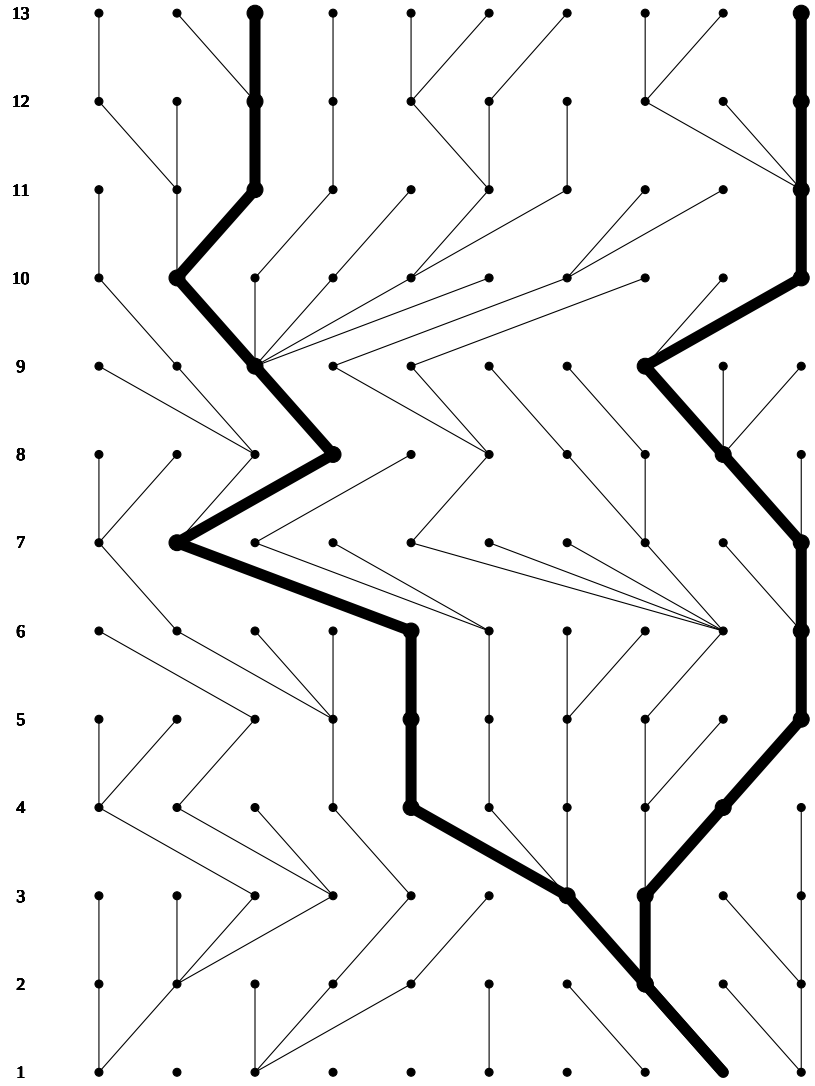

Figure 5.1: The coalescent of two lines in the Wright-Fisher Model

5.1 Topologies

With only pairwise coalescence events, the topology of a coalescence tree is easy to model. Consider a sample of size n and represent the ancestry of this sample as coalescing lineages back in time. Since each coalescence event reduces the number of ancestral lines by one, it takes $n-1$ such events to reach the MRCA as the root of the tree. We say that the tree is *in state* k at some time t in the past if there are k ancestral lines at this time. Looking further back in time, all $k(k-1)/2$ pairs of lines that can be chosen from these k lines are equally likely to be involved in the next coalescence event. If we start the coalescent process with labeled individuals (representing the n tips of the tree in our sample), we thus have

$$\prod_{k=2}^{n} \binom{k}{2} = \frac{n!(n-1)!}{2^{n-1}} \tag{5.4}$$

different *labeled and time-ordered histories*, where we do not only distinguish who coalesces with whom, but also different time orders in the coalescence events. In many cases, however, we are not interested in the genealogy of a specific sample, but in the statistical properties of (e.g. neutral) coalescent trees, such as the number of subtrees of a certain size – irrespectively of any labels at the tips of the tree. In this case, it is sometimes easier to construct the coalescent tree forward in time: For a tree currently in state k, we simply pick one of the lines at random and split it to obtain state $k+1$. For example, we can prove the following

Theorem 1 *Take a random coalescent tree of size n and consider the k branches that exist at state k of the tree. Let λ_i, $i = 1, \ldots, k$ be the number of offspring of the ith branch. Such a branch is also called a branch of* size λ_i. *Then, the offspring $(\lambda_1, \ldots, \lambda_k)$ of all k branches is uniformly distributed over all k-dim vectors with entries $\lambda_i \in \mathbb{N}_+$ and $\sum_i \lambda_i = n$.*

- Note that there are

$$\binom{n-1}{k-1} \tag{5.5}$$

  such vectors. To see this, imagine that we distribute n (identical) balls over k (labeled) groups. We can put all n balls next to each other in a single line and then place vertical lines between the balls to delimit the groups. Then, $k-1$ demarcation lines are needed, which can go in any of the $n-1$ spaces between the balls.

Proof To prove the theorem, consider first a specific history, forward in time, starting at state k: Imagine that the first $\lambda_1 - 1$ split events all occur in descendants of the first branch, followed by $\lambda_2 - 1$ split events in offspring of the second branch, and so on until the k-th branch, which needs $\lambda_k - 1$ split events in its offspring. The probability of this

particular history is

$$\left(\frac{1}{k}\cdot\frac{2}{k+1}\cdots\frac{\lambda_1-1}{k+\lambda_1-2}\right)\cdot\left(\frac{1}{k+\lambda_1-1}\cdots\frac{\lambda_2-1}{k+\lambda_1+\lambda_2-3}\right)\cdots$$
$$\cdots\left(\frac{1}{k+\sum_{i=1}^{k-1}\lambda_i-k+1}\cdots\frac{\lambda_k-1}{k+\sum_{i=1}^{k}\lambda_i-k-1}\right)=\frac{(k-1)!\prod_{i=1}^{k}(\lambda_i-1)!}{(n-1)!}\,. \qquad (5.6)$$

As long as there are λ_i-1 splitting events in the descendants of the ith branch ($i=1,\ldots,k$), we will always obtain the same distribution $(\lambda_1,\ldots,\lambda_k)$, irrespective of the order of these splitting events. If we can calculate the probability of each of these alternative histories in a stepwise procedure like in (5.6), it is easy to see that the only difference to (5.6) is a permutation of the numbers in the numerator. We conclude that the probability of all alternative histories to obtain a specific offspring distribution $(\lambda_1,\ldots,\lambda_k)$ is identical. The number of alternative histories for a given distribution is given by the multinomial coefficient $\binom{n-k}{\lambda_1-1,\ldots,\lambda_k-1}$, and thus

$$\Pr[(\lambda_1,\ldots,\lambda_k)]=\frac{(k-1)!(n-k)!}{(n-1)!}=\binom{n-1}{k-1}^{-1}\,. \qquad (5.7)$$

- The splitting scheme is also known as the *Polya urn scheme* in the mathematical literature. This scheme starts with an urn containing k balls with k different colors. Then, each round, take out one ball, put it back in and add another ball of the same color.

- For $k=2$, the result says that if we pick one of the branches after the first split, the size of this branch will be uniformly distributed on $1,2,\ldots,n-1$. In the limit $n\to\infty$, we obtain a coalescent tree of the "whole population". Then, the proportion X of lines that derive from the left branch after the first split is uniformly distributed on the interval $(0,1)$. Consider now the coalescent tree of a random sample of size m. The MRCA of the sample tree will be the same one as for the population tree unless either all m lines or no lines at all trace back to the left branch after the first split of the population tree. This occurs with probability

$$\int_0^1\Big(x^m+(1-x)^m\Big)dx=\frac{2}{m+1}\,.$$

  The probability that the population MRCA coincides with the sample MRCA is thus

$$1-\frac{2}{m+1}=\frac{m-1}{m+1}\,. \qquad (5.8)$$

  In more general, if we pick a subsample of size m of a sample of size n, the probability that both samples go back to the same MRCA is

$$\frac{(m-1)(n+1)}{(m+1)(n-1)}\,. \qquad (5.9)$$

For a proof, consider the n-tree after the first split and calculate the probability that all m lines of the subsample go back to the left branch,

$$p_l = \frac{1}{n-1}\sum_{k=m}^{n-1} \frac{k}{n}\frac{k-1}{n-1}\cdots\frac{k-m+1}{n-m+1} = \frac{m!(n-m)!}{(n-1)n!}\sum_{k=m}^{n-1}\binom{k}{m} = \frac{n-m}{(n-1)(m+1)}$$

using the summation formula Eq. (**??**). The result (5.9) is then obtained as $1-2p_l$, since the m lines can either go back to the left or the right branch with equal probability.

- In general, the uniform distribution over the branch sizes leads to a much higher variance in branch size than expected under a binomial or multinomial distribution: neutral coalescent trees can be both balanced or unbalanced.

Number of possible rooted and unrooted trees

In the examples above, we did not distinguish trees according to their branch length, but we have still accounted for the order of coalescence events. However, we can also count coalescence trees without any reference to time order.

For a sample of size n, we have $n-1$ coalescence events until we reach the MRCA (the *root*). This creates $2n-1$ so-called *vertices* in the tree: n are external (the *leaves*) and $n-1$ are internal. Every vertex has a branch directly leading to the next coalescence event. Only the root, which is also a vertex in the tree, does not have a branch. This makes $2n-2$ branches in a rooted tree with n leaves. As two branches lead to the root, the number of branches in an unrooted tree with n leaves is $2n-3$.

Let $\mathcal{B}_n$ be the number of topologies of unrooted trees with n leaves. We can derive this number recursively. Assume we have a tree with $n-1$ leaves, representing the first $n-1$ sampled sequences. We can ask in how many ways the nth sequence can be added to this tree. There are $2n-5$ branches in a tree with $n-1$ leaves. Since any branch can have the split leading to the nth leave, we obtain

$$\mathcal{B}_n = (2n-5)\mathcal{B}_{n-1}.$$

It is easy to see that there is only a single unrooted tree with three leaves. Thus

$$\mathcal{B}_n = 1\cdot 3\cdot 5\cdots(2n-7)\cdot(2n-5) = (2n-5)!!\,. \tag{5.10}$$

5.2 Coalescence times

For the branch lengths of the coalescent tree, we need to know the coalescence times. For a sample of size n, we need $n-1$ times until we reach the MRCA. As stated above, these times are independent of the topology. Mathematically, we obtain these times most conveniently by an approximation of the geometrical distribution by the exponential distribution for large N:

- If X is geometrically distributed with small success probability p and t is large then

$$\Pr[X \geq t] = (1-p)^t \approx e^{-pt}.$$

This is the distribution function of an exponential distribution with parameter p.

Let t_n be the time until the first coalescence occurs in a smaple of size n. This time is geometrically distributed according to

$$\Pr[t_n > t] = \left[1 - \frac{\binom{n}{2}}{2N}\right]^t = \left[1 - \frac{n(n-1)}{4N}\right]^t. \tag{5.11}$$

The mean waiting time until the first coalescence event is $\mathrm{E}[t_n] = 4N/n(n-1)$ and thus proportional to the population size. It is standard to integrate this dependence into a "coalescent time scale"

$$\tau := \frac{t}{2N}.$$

We can then take the limit $N \to \infty$ to obtain a stochastic process with a continuous time parameter τ. Coalescence times $T_n := t_n/2N$ in this limiting process are distributed like

$$\Pr[T_n > \tau] = \lim_{N\to\infty} \left[1 - \frac{\binom{n}{2}}{2N}\right]^{2N\tau} = \exp\left[-\tau\binom{n}{2}\right]. \tag{5.12}$$

In a sample of size n, the time to the first coalescence is thus exponentially distributed with parameter $\lambda = n(n-1)/2$. The fact that in the coalescent the times are exponentially distributed enables us to derive several important quantities.

- The time to the MRCA,

$$T_{\mathrm{MRCA}}(n) = \sum_{k=2}^{n} T_k,$$

is the sum of $n-1$ mutually independent exponentially distributed random variables. Its expectation and variance derive to

$$\mathrm{E}[T_{\mathrm{MRCA}}(n)] = \sum_{k=2}^{n} \mathrm{E}[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right) \tag{5.13}$$

and

$$\mathrm{Var}[T_{\mathrm{MRCA}}(n)] = \sum_{k=2}^{n} \mathrm{Var}[T_k] = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = 8\sum_{k=2}^{n}\frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)^2. \tag{5.14}$$

We have $\mathrm{E}[T_{\mathrm{MRCA}}(n)] \to 2$ for large sample sizes $n \to \infty$. Note that $\mathrm{E}[T_{\mathrm{MRCA}}(2)] = 1$, so that in expectation more than half of the total time to the MRCA is needed for the last two ancestral lines to coalesce. Similarly, $\mathrm{Var}[T_{\mathrm{MRCA}}(n)] \to 4\pi^2/3 - 12 \approx 1.16$ for $n \to \infty$ is dominated by $\mathrm{Var}[T_2] = 1$.

- Due to the independence of the coalescence times, the full distribution of $T_{\mathrm{MRCA}}(n)$ can be derived as an $(n-2)$-fold convolution,

$$f_{T_{\mathrm{MRCA}}(n)}(\tau) = \sum_{k=2}^{n} \binom{k}{2} \exp\left[-\binom{k}{2}\tau\right] \prod_{j=2, j\neq k}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{k}{2}}. \tag{5.15}$$

- For the total tree length,

$$L(n) = \sum_{k=2}^{n} kT_k,$$

we obtain the expected value

$$\mathrm{E}[L(n)] = \sum_{k=2}^{n} k\,\mathrm{E}[T_k] = 2\sum_{k=2}^{n} \frac{1}{k-1} = 2\sum_{k=1}^{n-1} \frac{1}{k}. \tag{5.16}$$

and the variance

$$\mathrm{Var}[L(n)] = \sum_{k=2}^{n} k^2\,\mathrm{Var}[T_k] = 4\sum_{k=1}^{n-1} \frac{1}{k^2}. \tag{5.17}$$

Increasing the sample size will mostly add short twigs to a coalescent tree. As a consequence, also the total branch length

$$\mathrm{E}[L(n)] \approx 2(\log(n-1) + \gamma) \quad ; \quad \gamma = 0.577216\ldots.$$

increases only very slowly with the sample size (γ is the Euler constant). The variance even approaches a finite limit $2\pi^2/3 \approx 6.58$ for $n \to \infty$.

- Again, also the entire distribution can be derived and takes a relatively easy form,

$$f_{L(n)}(\tau) = \frac{n-1}{2} \exp[-\tau/2]\Big(1 - \exp[-\tau/2]\Big)^{n-2} \tag{5.18}$$

- As we have seen above, the probability that the coalescent of a sample of size n contains the MRCA of the whole population is $(n-1)/(n+1)$ (for large, finite N). An important practical consequence of these findings is that, under neutrality, relatively small sample sizes (typically 10-20) will usually be enough to gain all statistical power that is available from a single locus.

5.3 Polymorphism patterns

In order to generate DNA diversity patterns using the coalescent, we need to add mutations to the process. This can be done according to any of the mutation schemes introduced in section (**??**). Most frequently used are the infinite sites and the infinite alleles model, which we will discuss in the following.

The key insight for the description of neutral DNA diversity using the coalescent is that neutral mutations do not interfere with the genealogy: *state* (the genotype) and *descent* (the genealogical relationships) are decoupled for neutral evolution. This is easy to see from the time-forward dynamics, since parents carrying different variants of a neutral allele are still equivalent concerning the distribution of their offspring in all future generations. If we want to create a random neutral polymorphism pattern using the coalescent process, we can therefore pick a genealogy first (as described in the previous section) and decide on the state later on. This is done by so-called *mutation dropping*, where mutations are added to all branches of the tree.

Let us first discuss the infinite sites mutation scheme. I.e. each mutation hits a new site (and thus leads to a new allele) and all mutations on a genealogy remain visible. If a mutation occurs on a branch of size i in the genealogy of n individuals, it will give rise to a polymorphism with frequency i/n of the derived allele. This means: the mutant allele is seen in i out of n sequences in the sample. Note that we do not need to know the precise time for the origin of the mutations in the genealogy, all that is needed is the total number of mutations that fall on each branch. On genealogical time scales (as opposed to phylogenetic time scales), we can usually assume that the mutation rate u (per haploid individual and generation) is constant.

For a branch of length l, we therefore directly get the number of neutral mutations on this branch by drawing from a Poisson distributed with parameter $2Nlu$. The factor $2N$ accounts for the fact that branch length l is measured on the coalescent time scale (in units of $1/2N$). In particular, the total number of mutations in an entire coalescent tree of length L is Poisson distributed with parameter $2NLu$. Let S be the number of segregating (polymorphic) sites in a sample. Since each polymorphic site corresponds to exactly one mutation on the tree under the infinite sites model, we have

$$\Pr[S=k] = \int_0^\infty \Pr[S=k|\ell] \cdot f_{L(n)}(\ell) d\ell = \int_0^\infty e^{-2N\ell u} \frac{(2N\ell u)^k}{k!} \cdot f_{L(n)}(\ell) d\ell \,.$$

For the expectation that means

$$\begin{aligned} \mathrm{E}[S] &= \sum_{k=0}^{\infty} k \Pr[S=k] = \int_0^\infty \frac{\ell\theta}{2} e^{-\ell\theta/2} \Big(\sum_{k=1}^{\infty} \frac{(\ell\theta/2)^{k-1}}{(k-1)!} \Big) \cdot f_{L(n)}(\ell) d\ell \\ &= \frac{\theta}{2} \int_0^\infty \ell \; f_{L(n)}(\ell) d\ell = \frac{\theta}{2} \mathrm{E}[L(n)] = \theta \sum_{i=1}^{n-1} \frac{1}{i} = a_n \theta \end{aligned} \tag{5.19}$$

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \,, \tag{5.20}$$

and where

$$\theta = 4Nu$$

is the standard population mutation parameter. Note that the distribution of S does not depend on the coalescent topologies, but only on the distribution of the coalescence times.

The mismatch distribution

For a Poisson distributed random variable, the time interval between consecutive events is exponentially distributed. There is therefore an alternative way to derive the equilibrium heterozygosity H (or the number of polymorphic sites in a sample of size 2) using the coalescent. If we follow the genealogy of two copies of a homologous site back in time, two things can happen first: (1) either one of the two mutates or (2) they coalesce. If they coalesce first they are identical by descent, if one of the two mutates, they are not identical. For both processes, the time back to the first event is exponentially distributed. Since mutation (in either lineage) occurs at rate $2u$ and coalescence occurs at rate $1/2N$, we directly obtain using Eq. (**??**),

$$H(u,N) = \frac{2u}{2u + (1/2N)} = \frac{\theta}{\theta + 1}\,. \tag{5.21}$$

We can easily extend this result and ask for the probability that we find precisely k differences among the two sequences. Under the assumptions of the infinite-sites model, and using that we can re-start the Poisson process after every event,

$$\Pr[\pi = k] = \Big(\frac{\theta}{\theta+1}\Big)^k \frac{1}{\theta+1}\,, \tag{5.22}$$

which is a modified geometrical distribution. Note that this is not the distribution of pairwise mismatches in a larger sample, which will be correlated due to a shared history. However, under the standard neutral model, we should see this distribution if we sequence from independent loci along the genome (e.g. counting mismatches among the two copies carried by a diploid individual).

The site frequency spectrum

The total number S of polymorphic sites is the simplest so-called *summary statistic* of polymorphism data. There are many more. As a next step, we can ask for the number S_i of mutations of a given size i (mutations that are observed in i out of n sequences in the sample). To derive the expected value $\mathrm{E}[S_i]$, we proceed in two steps. First, we ask for the probability that a branch at state k of the coalescent process is of size i,

$$P[i|k] := \Pr[\text{Probability for branch at state } k \text{ to be of size } i]\,.$$

From Theorem 1 we directly obtain $P[i|k]$ as

$$P[i|k] = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{5.23}$$

Here, the numerator counts all different possibilities to distribute $n-i$ descendants over $k-1$ branches – i.e. all remaining descendants after we have assigned i descendants to the focal branch. Note that $P[i|k]$ does not depend at all on the coalescent times, but only

on aspects of the topology. In the second step, we ask for the expected number $\mathrm{E}[S^{(k)}]$ of mutations on a branch at state k. Noting that the length of such a branch is T_k, this is easily derived (analogous to Eq. 5.19),

$$\mathrm{E}[S^{(k)}] = \frac{\theta}{2}\,\mathrm{E}[T_k] = \frac{\theta}{k(k-1)}\,. \tag{5.24}$$

In contrast to $P[i|k]$, this expression does not depend on the topologies, but only on the coalescent times. Using the independence of coalescent times and topologies, we now obtain the expected number of mutations of size i as

$$\begin{aligned}
\mathrm{E}[S_i] &= \sum_{k=2}^{n} kP[i|k] \cdot \mathrm{E}[S^{(k)}] \\
&= \sum_{k=2}^{n} \frac{\theta}{k-1} \frac{(n-i-1)!(k-1)!(n-k)!}{(k-2)!(n-i-k+1)!(n-1)!} \\
&= \frac{\theta}{i\binom{n-1}{i}} \sum_{k=2}^{n} \binom{n-k}{i-1} \\
&= \frac{\theta}{i\binom{n-1}{i}} \sum_{k=2}^{n} \left(\binom{n-k+1}{i} - \binom{n-k}{i} \right) \\
&= \frac{\theta}{i\binom{n-1}{i}} \cdot \binom{n-1}{i} = \frac{\theta}{i}\,,
\end{aligned} \tag{5.25}$$

where we have used Eq. (**??**). The expected number of mutations of size i is thus θ/i. Together, these numbers define the (expected) *site frequency spectrum* of sample taken from a standard neutral population.

- The frequencies of the expected normalized site frequency spectrum are $p_i = 1/(a_n i)$. They are independent of θ. The characteristic $(1/i)$-shape is a prime indicator of "neutrality".

- We can easily obtain an empirical site frequency spectrum from any polymorphism data. This empirical spectrum can then be compared to the spectrum predicted under neutrality. Note that we need data from many independent (unlinked) loci to observe the *expected* spectrum. For any single locus, the spectrum can differ considerably, because we only have a single coalescent history.

- To determine the size of a given polymorphism in the sample, we need to know the ancestral state at the locus. In practice, this is inferred from a so-called outgroup (usually a single consensus sequence from a closely related sister species). If the ancestral state cannot be determined, we can work with the so-called *folded site frequency spectrum*, with mutation classes $\tilde{S}_i = S_i + S_{n-i}$ for $i < n/2$ and $\tilde{S}_i = S_i$ for $i = n/2$.

Infinite alleles and haplotype statistics

So far, we have considered polymorphism patterns under the assumption of the infinite sites model, where all mutations that occur during the genealogy of a sample remain visible as a polymorphic site. Depending on the type of the mutation, however, this may not always be true. For example, the infinite sites model does not easily generalize to insertion/deletion mutations. Alternatively, we may focus on the entire haplotype in a chromosomal window and just ask for the distribution of different types (ignoring any information about the mutational distances between these types). Questions like these can be addressed within the framework of the infinite alleles model.

Just like in the case of the infinite sites model, we can construct the genealogical tree first and add mutations later on. However, for the infinite alleles model, only the latest mutations (the ones closest to the leaves of the tree) will be observed. As a consequence, major parts of the genealogy do not influence the pattern. We can account for this by adding mutations already as we build the genealogy. Once we encounter the first mutation in the ancestry of an individual, we know the state of this this ancestor and of all its descendants. So, before we construct the genealogy further back in time, we can stop (or *kill*) this branch. This leads to the so-called *coalescent with killings*, where we have two kinds of events:

1. As before, coalescence events occur at rate $k(k-1)/2$ on the coalescence time scale for a tree in state k (i.e. with k ancestral lines).

2. In addition, we directly account for mutation events, which occur at rate $k\theta/2$ in state k. Each mutation "kills" the corresponding branch.

Let K_n be the number of different haplotypes that we observe in a sample of size n. We are interested in the probability that K_n takes a certain value k. By following the coalescent with killings back in time to the first event (either coalescence or mutation), we can relate the values for the distribution of K_n to the corresponding values for K_{n-1},

$$P[K_n = k] = \frac{\theta}{\theta + n - 1} \cdot P[K_{n-1} = k-1] + \frac{n-1}{\theta + n - 1} \cdot P[K_{n-1} = k]\,. \tag{5.26}$$

As initial condition, we have $P[K_1 = 1] = 1$. To solve this recursion, observe that the denominator in both terms in (5.26) is the same. Note also, that for $K_n = k$, we need to choose "mutation" $k-1$ times before we reach a sample of size 1. Each time, we pick up a factor of θ, like in the first term of (5.26). We thus can write

$$P[K_n = k] = \frac{\theta^{k-1}}{(\theta+1)(\theta+2)\cdots(\theta+n-1)} \cdot S(n,k) = \frac{\theta^k}{\theta_{(n)}} \cdot S(n,k) \tag{5.27}$$

where

$$\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$$

and the $S(n,k)$ are the so-called *Stirling numbers of the first kind*, which follow the recursion relation

$$S(n,k) = S(n-1,k-1) + (n-1)\cdot S(n-1,k)\,. \tag{5.28}$$

In analogy to the allele frequency spectrum, we can also ask for the frequency distribution of haplotypes. Let A_j be the number of haploytpes that appear j times in a sample of size n. For $K_n = k$, we thus have

$$\sum_{j=1}^{n} A_j = k \quad \text{and} \quad \sum_{j=1}^{n} jA_j = n\,,$$

and let $\boldsymbol{a} = (a_1,\ldots,a_n)$ be a realization of $(A_1,\ldots,A_n)$. We can prove the following

Theorem 2 *The combined distribution of the number and frequencies of haplotypes under the standard neutral model is given by the so-called Ewens' sampling formula,*

$$P_n[\boldsymbol{a}] = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}\,. \tag{5.29}$$

- One interpretation of this result is to view the A_j as independently Poisson distributed random variables with parameter (= expected value) θ/j, and then consider the marginal distribution under the condition $\sum_{j=1}^{n} jA_j = n$. Note that the distribution is strongly influenced by θ. For large $\theta > 1$, the distribution is dominated by singleton haplotypes $\sim \theta^{a_1}$, for small θ, a large number of singletons (large a_1) is unlikely.

Proof To prove the theorem, we extend the recursion method that we have used in our proof of the distribution of K_n. Note first that for $n = 1$, we have $a_1 = 1$ with probability 1, in accordance with (5.29). Define $\boldsymbol{e}_i = (0,\ldots,1,0,\ldots)$ as the ith unit vector (with entry 1 in the ith position). Now, start with a sample of size n and go back to the first event. With probability $\theta/(\theta+n-1)$, this is a mutation, which creates a new haplotype. This relates the partition $\boldsymbol{a}$ of the n haplotypes in the sample to the partition $\boldsymbol{a} - \boldsymbol{e}_1$ of the remaining $n-1$ types. If the first event is coalescence (which it will be with probability $(n-1)/(\theta+n-1)$), this decreases the frequency of one of the haplotypes (with at least two copies) by one. In terms of the allelic partitions, this turns $\boldsymbol{a}$ into $\boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}$ for some $j \in \{1,\ldots,n-1\}$. Conditioned on this latter partition for the tree at state $n-1$, the probability that (forward in time) the next split event will be in one of the haplotype classes with j copies is thus $(a_j+1)j/(n-1)$. This results in the recursion

$$P_n[\boldsymbol{a}] \;=\; \frac{\theta}{\theta+n-1} P_{n-1}[\boldsymbol{a}-\boldsymbol{e}_1] + \frac{n-1}{\theta+n-1} \sum_{j=1}^{n-1} \frac{(a_j+1)j}{n-1} P_{n-1}[\boldsymbol{a}+\boldsymbol{e}_j-\boldsymbol{e}_{j+1}]. \tag{5.30}$$

It remains to be shown that (5.29) fulfills this recursion. For this, note that (5.29) implies that

$$P_{n-1}[\boldsymbol{a} - \boldsymbol{e}_1] = \frac{(\theta + n - 1)a_1}{n\theta} P_n[\boldsymbol{a}] \tag{5.31}$$

$$P_{n-1}[\boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}] = \frac{(\theta + n - 1)a_{j+1}(j+1)}{(a_j + 1)nj} P_n[\boldsymbol{a}] \tag{5.32}$$

Inserting this into (5.30) yields

$$1 = \frac{a_1}{n} + \sum_{j=2}^{n-1} \frac{a_{j+1}(j+1)}{n} = \frac{1}{n} \sum_{j=1}^{n} a_j \,,$$

which holds true, since $\sum_j a_j = n$.

- The underlying combinatorial problem is also known as the *Hoppe urn scheme* in the mathematical literature. This scheme starts with k colored balls like the *Polya urn*, but adds a special back ball with weight θ. Each time a colored ball is drawn, the ball is returned with another ball of the same color. Each time the black ball is drawn, it is put back with another ball of a new color.

- Note that the marginal distribution for the allelic partition given the number of haplotypes can be written as

$$P_n[\boldsymbol{a}|K_n = k] = \frac{P_n[\boldsymbol{a}]}{P[K_n = k]} = \frac{n!}{S(n,k)} \prod_{j=1}^{n} \frac{1}{a_j! j^{a_j}} \,, \tag{5.33}$$

  which is, in particular, independent of θ. In statistical terms this means that all information about θ is already contained in the number of haplotypes found in a sample: K_n is a *sufficient statistic*. Knowledge about their distribution does not add any further information.

5.4 Coalescent and statistics

Coalescent trees show the genealogical relationships between two or more sequences that are drawn from a population. This should not be confounded with a phylogenetic tree that shows the relation of two or more species. Indeed, both "trees" have entirely different roles for the theory of evolution. In phylogenetics, one is usually interested in the one "true tree" and the parameters of this tree (such as split times) are estimated from data. In contrast, there is no single "true tree" for a set of individuals from a population. Indeed, the genealogy will usually be different for different loci. For example, at a mitochondrial locus your ancestor is certainly your mother and her mother. However, if you are a male, the ancestor for the loci on your Y-chromosome is your father and his father. So the genealogical tree will look different for a mitochondrial locus than for a Y-chromosomal

locus. But even for a single locus, we are usually not able to reconstruct a single "true coalescence tree" and this is not the goal in coalescent studies. Instead, coalescent histories are used as a statistical tool for inferences about an underlying model.

The general idea is as follow. We define an evolutionary model that depends on a number of biological parameters (such as mutation rates, population sizes, selection coefficients). Under this model, we obtain a distribution of coalescent histories and (consequently) a distribution of polymorphism patterns that is predicted under this model. We can then compare measured data with the predicted distribution to make statistical inferences. Usually, there is a twofold goal:

1. to reject (or not) the underlying model. This is true, in particular, for the neutral model as the standard null model of population genetics.

2. to estimate model parameters. Note that the parameters of the coalescent tree (coalescent times, topology) are generally not model parameters. They are "integrated out" in the statistical treatment.

In some easy cases (notably the neutral model), key aspects of the distribution of polymorphism patterns can be obtained analytically using coalescent theory. In many other cases, this is no longer possible. However, even in these cases, the coalescent offers a highly efficient simulation framework that is routinely used in statistical simulation packages.

Estimators for the mutation parameter θ

All population genetic models, whether forward or backward in time, depend on a set of biological parameters that must be estimated from data. In the standard neutral model, there are two such parameters: the mutation rate u and the population size N. However, since both parameters only occur in the combination $\theta = 4Nu$, the population mutation parameter is effectively the only parameter of the model. From our derivation of the expected site frequency spectrum, we easily obtain several estimators for θ. In principle, we can use the total number of mutations of any size class to define an unbiased estimator $\hat{\theta}_i$,

$$\mathrm{E}[S_i] = \frac{\theta}{i} \quad \longrightarrow \quad \hat{\theta}_i := i \cdot S_i \,. \tag{5.34}$$

In practice, widely used estimators are linear combinations across mutations of different size classes. They can be distinguished according to the relative weight that is put on a certain class. The most important ones are the following:

1. *Watterson's estimator*,

$$\hat{\theta}_\mathrm{W} := \frac{S}{a_n} = \frac{1}{a_n} \sum_{i=1}^{n-1} S_i = \frac{1}{a_n} \sum_{1 \le i \le n/2} \tilde{S}_i \,, \tag{5.35}$$

uses the total number of segregating sites and puts an equal weight on each mutation class. The last equation expresses $\hat{\theta}_\mathrm{W}$ in terms of frequencies of the folded spectrum.

Remember that the distribution of S – and thus of $\hat{\theta}_{\mathrm{W}}$ – is independent of the coalescent topologies, but only depends on the coalescent times.

2. Let π_{ij} be the number of differences among two sequences i and j from our sample. We have $\mathrm{E}[\pi_{ij}] = \mathrm{E}[S(n=2)] = \theta$. If the sample size is just two, this corresponds to Watterson's estimator. In a larger sample, we can still take the pairwise difference as our basis and average over all $n(n-1)/2$ pairs. This leads to the *diversity-based estimator* (sometimes also called *Tajima's estimator*),

$$\hat{\theta}_{\pi} := \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij} \,. \tag{5.36}$$

We can also express $\hat{\theta}_{\pi}$ in terms of the (folded) frequency spectrum as follows,

$$\hat{\theta}_{\pi} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i) S_i = \binom{n}{2}^{-1} \sum_{1 \le i \le n/2} i(n-i) \tilde{S}_i \,. \tag{5.37}$$

Whereas Watterson's estimator weights all frequency classes equally, $\hat{\theta}_{\pi}$ puts the highest weight on classes with an intermediate frequency. In contrast to $\hat{\theta}_{\mathrm{W}}$, it also depends on the distribution of tree topologies. The estimator is often also just written as $\hat{\pi}$.

3. *Fay and Wu's estimator*,

$$\hat{\theta}_{\mathrm{H}} := \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i^2 S_i \,, \tag{5.38}$$

puts a hight weight on mutation classes of the unfolded spectrum with a high frequency of the derived allele. In contrast to the other estimators, it is not a summary statistic of the folded spectrum and thus requires knowledge of the ancestral state.

4. Finally, the *singleton estimator* $\hat{\theta}_{\mathrm{s}}$ uses the singletons of the folded spectrum,

$$\hat{\theta}_{\mathrm{s}} := \frac{n-1}{n} \Big(S_1 + S_{n-1} \Big) = \frac{n-1}{n} \tilde{S}_1 \,. \tag{5.39}$$

It has all its weight at both ends of the unfolded spectrum.

Test statistics for neutrality tests

Estimators of any model parameter, such as θ, will only produce meaningful results if the assumptions of the underlying model hold. In our case, we have assumed standard neutral evolution. In addition to the absence of selection, this includes the assumptions of a constant population size and no population structure. But how can we know whether these assumptions do hold (at least approximately) for a given data set? This question asks

for a test of the model assumptions. As it turns out, the availability of various different estimators of the same quantity θ is helpful for the construction of such a test.

The key idea is to consider the difference among two different estimators, such as $\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}$. Under standard neutrality, this quantity should be close to zero, whereas significant deviations indicate that the model should be rejected. The most widely used test statistic that is constructed in such a way is *Tajima's D*,

$$D_\mathrm{T} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}]}}\,. \tag{5.40}$$

The denominator of D_T is used for normalization and makes the distribution of the statistic (almost) independent of θ and of the sample size. Tajima has shown that D_T is approximately β-distributed. Today, however, the exact distribution under the standard neutral null model is usually obtained (resp. approximated to arbitrary precision) by computer simulations. For a given significance level α, one can then specify the critical upper and lower bounds for D_T, beyond which the null model should be rejected. Test statistics that are constructed in a similar way are *Fu and Li's D*,

$$D_\mathrm{FL} := \frac{\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}}{\sqrt{\mathrm{Var}[\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}]}} \tag{5.41}$$

and *Fay and Wu's H*,

$$H_\mathrm{FW} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}]}}\,. \tag{5.42}$$

To understand, which kind of deviations from the standard neutral model are picked up by the three summary statistics, it is instructive to consider the contribution of the site frequency classes S_i to the numerator of each statistic. For example, D_T will be negative if we have an excess of very low or very high frequency alleles, whereas it will be positive if many sites segregate at intermediate frequencies.