Demography

April 10, 2015
1 Effective Population Size

The Wright-Fisher model makes a number of strong assumptions which are clearly violated in many populations. For example, it is unlikely that any population has a constant size for more than a few generations. Nonetheless, this model has played a central role in the development of theoretical population genetics and many of the statistical tools used to analyze sequence data were derived using the Wright-Fisher model. As we saw earlier in the semester, the Moran model, which describes a population with overlapping generations, makes predictions about genetic variation that are nearly identical to those of the Wright-Fisher model apart from a change in time scale: namely, as far as genetic drift and coalescence are concerned, a population of size $2N$ governed by the Moran model behaves like a population of size $N$ governed by the Wright-Fisher model. For this reason, we say that the effective population size of the Moran model is half that of the Wright-Fisher model.

In fact, it can be shown that a much larger class of models behave, in some respect, like the Wright-Fisher model apart from a change in time scale and effective population sizes can be calculated for these models as well. Depending on which statistic is compared between models, different kinds of effective population sizes can be defined, although in each case the effective population size of a population governed by the Wright-Fisher model will be equal to the census population size. Some of the more important formulations include:

- The inbreeding effective population size
  \[ N_{e}^i = \frac{1}{2\pi_2} \]
  where $\pi_2$ is the probability that two genes sampled at random are copies of the same parental gene.

- The variance effective population size
  \[ N_{e}^v = \frac{p_t(1-p_t)}{\text{Var}(p_{t+1}|p_t)} \]
  where $p_t$ is the frequency of an allele $A$ at a variable locus.

- The coalescent effective population size $N_{e}^c$, which is defined if and only if it is the case that if time is run at rate $2N_{e}^c$, then the genealogy of a sample of $n$ genes is governed by Kingman’s coalescent (Sjodin et al., 2005).

In general, there is no guarantee that the inbreeding and variance effective population sizes will be equal and, in principle, these could vary from generation to generation or even depend on the genetic composition of the population. On the other hand, if the coalescent effective population size is well-defined, then it will usually be the case that all three quantities are equal. This reflects the fact that for $N_{e}^c$ to be well-defined for a model, that model must be comprehensively like the Wright-Fisher model apart from a change in time scale. In contrast, we can calculate the inbreeding and variance effective population sizes for a much larger class of models that have very different qualitative and quantitative behaviors even after the time scale is adjusted. In the remainder of this section, we will focus on the coalescent effective population size, which we will calculate for several models that generalize the Wright-Fisher model in some biologically important respect.

We start with a generalization of the Wright-Fisher model in which the population size changes from generation to generation in a cyclical fashion.
Theorem 1. Suppose that a diploid population is governed by a version of the Wright-Fisher model in which the population size $N_t$ cycles through the values $N_1, N_2, \ldots, N_T$, where $T$ is much smaller than the minimum population size. Then the coalescent effective population size is equal to the harmonic mean of the census population sizes,

$$N_e^c = \left( \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t} \right)^{-1}.$$ 

Proof. Provided that the minimum population size is sufficiently large and that $T$ is much smaller than this number, the probability that a pair of randomly sampled lineages does not coalesce during a single cycle is

$$\prod_{t=1}^{T} \left( 1 - \frac{1}{2N_t} \right) \approx 1 - \sum_{t=1}^{T} \frac{1}{2N_t} = 1 - \frac{T}{2N_e^c}.$$ 

It follows that the probability that the pair does not coalesce in $2N_e^c t$ generation is approximately equal to

$$\left( 1 - \frac{T}{2N_e^c} \right)^{2N_e^c t/T} \approx e^{-t},$$

which shows that pairs of lineages coalesce at rate 1 when time is measured in units of $2N_e^c$ generations. Furthermore, since each pair coalesce independently of the others and since multiple or simultaneous mergers are unlikely under the Wright-Fisher model in a sufficiently large population, this shows that the genealogy of a sample of $n$ lineages will be approximately governed by Kingman’s coalescent when time is measured in these units.

As stated, Theorem 1 is not completely rigorous, since we have not specified how much smaller $T$ must be than the $N_t$’s and the genealogy of a sample is only approximately governed by Kingman’s coalescent. To obtain a more formal result, we must let $N_t = N x_t$, where $x_1, \ldots, x_T > 0$ are fixed, and then take the limit as $N \to \infty$. In this case, it can be shown that the distribution of genealogies converges exactly to Kingman’s coalescent when time is measured in units of $N_e^c$ generations.

Theorem 1 has two important consequences. First, it shows that even if the population size fluctuates, the model still behaves like the ordinary Wright-Fisher model provided that the fluctuations occur sufficiently rapidly and the population size never gets too small. This is a reflection of the separation of time scales that exists between the slow process governing coalescence and the fast process governing the population size variation. In effect, the population size fluctuates so rapidly that lineages experience an ‘average’ population size that determines the rate of coalescence. Secondly, because the harmonic mean is disproportionately sensitive to small values, the effective population size will typically be closer to the smallest census population sizes that occur in the cycle than to the largest. For example, if $T = 2$ and $N_1 = 1000$ and $N_2 = 100,000$, then the effective population size will be $N_e^c \approx 1980$. Similarly, if $N_1 = 1000$ and $N_2 = \cdots = N_T = \infty$, then $N_e^c = 1000T$. This accords with the observation that estimated effective population sizes (using whichever definition) are usually much smaller than predicted census population sizes.

A result similar to Theorem 1 holds for Wright-Fisher models in which the population size fluctuates randomly from generation to generation. For example, if the population size is governed by a stationary ergodic Markov chain with stationary distribution $\pi_k = \mathbb{P}(N_t = k)$ on the set $\{N_{\text{min}}, \ldots, N_{\text{max}}\}$, then the coalescent effective population size is

$$N_e^c = \left( \sum_k \frac{\pi_k}{k} \right)^{-1}.$$
As in the periodic case, this result will hold if the minimum population size \( N_{\text{min}} \) is not too small and if the rate at which the chain converges to its stationary distribution is much greater than the rate at which coalescence occurs within the genealogy. Such a model could apply if, for example, the population size is governed by random fluctuations in the environment, e.g., drought or temperature.

We can also calculate the coalescent effective population size for a version of the Wright-Fisher model with two sexes. Suppose that in each generation there are \( N_f \) females and \( N_m \) males and that each individual chooses one chromosome at random from a female parent and one chromosome at random from a male parent.

**Theorem 2.** The coalescent effective population size of the two-sex Wright-Fisher model with \( N_f \) females and \( N_m \) males is

\[
N^e_c \approx \frac{4N_f N_m}{N_f + N_m}
\]

provided that \( N_f \) and \( N_m \) are both large.

**Proof.** Suppose that two chromosomes have been sampled at random from the current generation. For these two to coalesce in the previous generation, they must either be descended from the same female chromosome or the same male chromosome in the previous generation. Writing \( N = N_f + N_m \) for the total population size, we see that there are \( 2N(2N-1) \) pairs of chromosomes in the population, of which \( N(N-1) \) are both of maternal origin and \( N(N-1) \) are both of paternal origin. (Notice that while there may be different numbers of males and females in the population, both sexes contribute equal numbers of chromosomes to the next generation.) It follows that the probability that the two chromosomes share a common ancestor in the previous generation is

\[
\pi^2 = \frac{N(N-1)}{2N(2N-1)} \left( \frac{1}{2N_f} + \frac{1}{2N_m} \right) \approx \frac{1}{8} \frac{N_f N_m}{N_f + N_m} = \frac{1}{2N^e_c},
\]

where the approximation is accurate if \( N \) is large. This shows that pairs of lineages will coalesce at rate 1 if time is measured in units of \( N^e_c \) generations. The extension to Kingman’s coalescent for larger samples can then be deduced from the fact that gamete sampling is governed by the Wright-Fisher mechanism.

If the sex ratio is 1 : 1, then \( N_f = N_m = N/2 \) and \( N^e_c = N \), i.e., the effective population size of a two-sex Wright-Fisher model with equal numbers of males and females is the same as the census population size. On the other hand, if the sex ratio is extremely skewed, say \( N_f \gg N_m \), then \( N^e_c \approx 4N_m \) and then genetic drift and coalescence occur at a rate determined mainly by the rarer sex.

As a final example, we will investigate a class of models, called Cannings’ models, in which the numbers of chromosomes transmitted by each individual to the next generation can have different distributions than the multinomial distribution used in the Wright-Fisher model. Cannings’ models make the following assumptions:

1. Non-overlapping generations.
2. Constant population size with \( N \) diploid adults.
4. If we label the chromosomes present in generation \( t \) from 1 to \( 2N \), then the number of copies of chromosome \( i \) transmitted to generation \( t + 1 \) is a random variable \( \eta_i(t) \).
5. The variables \((\eta_1(t), \cdots, \eta_{2N}(t))\) are exchangeable, i.e., their joint distribution does not depend on the labeling.

6. Reproductive contributions will be independent across generations, e.g., \(\eta_1(t)\) and \(\eta_1(s)\) will be independent if \(t \neq s\).

In general, we will suppress the time parameter and simply write \(\eta_i\) for \(\eta_i(t)\). Suppose that two chromosomes are sampled at random from generation \(t + 1\). Conditional on the family sizes \(\eta_1, \cdots, \eta_{2N}\) produced in generation \(t\), these two chromosomes will share a common ancestor in this generation with probability

\[
\frac{2N}{2N(2N-1)} \sum_{i=1}^{2N} \frac{\eta_i(\eta_i - 1)}{2N(2N-1)}.
\]

Taking expectations and using exchangeability, it follows that the unconditional probability of coalescence one generation back is

\[
\pi_2 = \mathbb{E} \left[ \frac{2N}{2N(2N-1)} \right] = \frac{\mathbb{E}[\eta_1(\eta_1 - 1)]}{2N-1}.
\]

However, since \(2N = \eta_1 + \cdots + \eta_{2N}\), exchangeability also implies that \(\mathbb{E}[\eta_1] = 1\), which allows us to write

\[
\mathbb{E}[\eta_1(\eta_1 - 1)] = \mathbb{E}[\eta_1^2] - 1 = \text{Var}(\eta_1).
\]

This shows that the inbreeding effective population size for a Cannings’ model is

\[
N_e^i = \frac{2N - 1}{2\text{Var}(\eta_1)} \approx \frac{N}{\text{Var}(\eta_1)},
\]

where the approximation is accurate when \(N\) is large. In other words, the greater the within-generation fecundity variance between the members of a population, the smaller the effective population size will be, which in turn will typically lead to greater rates of genetic drift and reduced genetic variation. Indeed, in most the extreme case where all \(2N\) chromosomes are descended from a single chromosome in the previous generation, the effective population size will be \(1/2\) and genetic variation will collapse. At the other extreme, if every chromosome transmits exactly one copy of itself, then \(\eta_1 = \cdots = \eta_{2N} = 1\) and \(\text{Var}(\eta_1) = 0\), so that the effective population size will be infinite. In this case, there will be no coalescence and each lineage will be unrelated to all of the others.

Whether the coalescence effective population size is defined for a Cannings’ model depends on the relative rates of pairwise and three-way coalescent events. The following result is due to Möhle. Here we assume that we have a sequence of Cannings’ models, one for each value of \(N\), and we consider the corresponding sequence of genealogies for a sample of fixed size \(n\) as \(N \to \infty\). Since the family size distributions depend on \(N\), we will write \(\eta_i^{(N)}\) to make this dependence explicit.

**Theorem 3.** The sequence of genealogies converges to Kingman’s coalescent if and only if

\[
\lim_{N \to \infty} \frac{\mathbb{E}[\eta_1^{(N)}(\eta_1^{(N)} - 1)(\eta_1^{(N)} - 2)]/N^2}{\mathbb{E}[\eta_1^{(N)}(\eta_1^{(N)} - 1)]/N} = 0.
\]

The condition in the theorem implies that when \(N\) is large, the probability of a three way merger is negligible when compared with the probability of a two way merger. For example, this condition will be satisfied if the family sizes are uniformly bounded, i.e., if there is an integer \(M < \infty\) such that \(\mathbb{P}(\eta_i^{(N)} \leq M) = 1\) for all \(N \geq 1\). Furthermore, when this condition is satisfied, it is easy to show that the coalescent effective population size exists and is equal to the variance effective population size calculated above.
2 Population Dynamics

In the previous section, we saw that if the size of a population fluctuates rapidly and in either a stationary or cyclical fashion, then the genealogy of a sample of chromosomes from that population has the same statistical properties as the genealogy of a sample of chromosomes from a population of constant size equal to the coalescent effective population size. Key to this result is the separation of time scales between coalescence and the population dynamics. In this section we will explore an alternative scenario in which the population size changes on the same time scale as coalescence. In this case we will see that genealogies can be modeled not by the coalescent itself but rather by a time change of the coalescent. Furthermore, we will see that this relationship between genealogies and the demographic history of a population can be used to infer that history from genetic data.

Suppose that a population is governed by the Wright-Fisher model with variable population size and let $N_t$ denote the population size $t$ generations before the present. If two chromosomes are sampled at random from this population at time $t=0$ and we let $\tau_2$ denote the time to their most recent common ancestor, then the probability that they do not coalesce in the first $t$ generations is
\[
P(\tau_2 > t) = \prod_{s=1}^{t} \left( 1 - \frac{1}{2N_s} \right) \approx \exp \left( - \sum_{s=1}^{t} \frac{1}{2N_s} \right)
\]
where the approximation is accurate if $N_s \gg 1$ for all $s \in [0, t]$. In particular, if we define a new time scale $\tau = \tau(t)$ by setting
\[
\tau(t) = \sum_{s=1}^{t} \frac{1}{2N_s},
\]
then we can rewrite the previous result as
\[
P(\tau(\tau_2) > t) = P(\tau_2 > \tau^{-1}(t)) \approx \exp(-t),
\]
which shows that the pairwise coalescent time $\tau(\tau_2)$ is exponentially distributed on this new time scale.

In fact, a much stronger result is true. Suppose that $n$ chromosomes are sampled at time 0 and let $G = (G_t : 0 \leq t \leq \infty)$ be their genealogy, where $G_t$ describes the relationships between the $n$ lineages at time $t$. Provided that $n$ is much smaller than the minimal population size, the distribution of $G$ on the time scale $\tau$ is approximately the same as Kingman’s coalescent run up until the time $\tau_{max} = \sum_{s=1}^{\infty} \frac{1}{N_s}$:
\[
(G_t : 0 \leq t < \infty) \overset{d}{=} (G_{\tau} : 0 \leq \tau < \tau_{max}),
\]
where $\mathcal{G} = (\mathcal{G}_t : 0 \leq t < \infty)$ is the ordinary coalescent and the expression $X \overset{d}{=} Y$ means that the random variables $X$ and $Y$ have the same distribution. For this reason, we say that $G$ is (approximately) a time change of the coalescent, i.e., apart from a possibly non-linear change in time scale, $G$ and $\mathcal{G}$ have the same distribution. This correspondence makes it easy to simulate genealogies for populations with arbitrary population dynamics. First simulate a coalescent tree $G$ and then use the time change $\tau$ to construct a sample genealogy $G$ for the variable population model. Neutral mutations can then be assigned to $G$ using Poisson processes.

As an example, consider a two-stage model in which for the last $T_{exp}$ generations, the population has experienced exponential growth at rate $\gamma > 0$ per generation, while prior to that time the population size was equal to $N_A$:
\[
N_t = \begin{cases} 
N_0 e^{-\gamma t} & t \in [0, T_{exp}] \\
N_A & t > T_{exp}.
\end{cases}
\]
If we assume continuous variation in the population size, then we must also have $N_A = N_0 e^{-\gamma T_{\text{exp}}}$, which reduces the number of parameters needed to describe this model from four to three. In this case the time change $\tau$ is

$$
\tau(t) = \begin{cases} 
\frac{1}{2N_0 \gamma} (e^{\gamma t} - 1) & t \in [0, T_{\text{exp}}] \\
\frac{1}{2N_0 \gamma} (e^{\gamma T_{\text{exp}}} - 1) + \frac{t-T_{\text{exp}}}{2N_A} & t > T_{\text{exp}}.
\end{cases}
$$

Since $N_0 > N_A$, it follows that the pairwise coalescence rate increases from $1/2N_0$ to $1/2N_A$ as we go backwards in time, which will tend to increase the relative lengths of the external branches in the genealogy. Since this will also increase the proportion of segregating sites that are represented by singletons in the sample, Tajima’s D is expected to be negative under population expansion.

A demographic **bottleneck** is an event in which a population is transiently reduced to small size. This could happen because of transient changes in the environment, such as during a drought or an epidemic, as well as during founder events when a new population is established by a small number of individuals dispersing from a larger population. For example, transmission of infectious diseases often involve founder events in which each new infection develops from a small number of pathogens that successfully colonize the new host individual. Bottlenecks are often modeled by assuming that the population size is piecewise constant with

$$
N_t = \begin{cases} 
N_0 & t \in [0, T_0] \\
N_b & t \in (T_0, T_1] \\
N_A & t > T_1.
\end{cases}
$$

Here $N_b \ll N_0$, $N_A$ is the size of the population during the bottleneck, which lasts for $T_b = T_1 - T_0$ generations. In this case the time change $\tau$ is piecewise linear

$$
\tau(t) = \begin{cases} 
\frac{t}{2N_0} & t \in [0, T_0] \\
\frac{T_0}{2N_0} + \frac{t-T_0}{2N_A} & t \in (T_0, T_1] \\
\frac{T_0}{2N_0} + \frac{T_0}{2N_b} & t > T_1.
\end{cases}
$$

The genetic impact of a bottleneck depends on its severity, its duration and how recently it has occurred. For example, bottlenecks that are both very recent and very severe can give rise to star-shaped trees; this is expected if

$$
\frac{T_0}{2N_0} \ll 1 < \frac{T_b}{2N_b} \quad \text{and} \quad T_b \ll T_0.
$$

The first inequality implies that no coalescent events are expected to occur during the first $T_0$ generations, while the second inequality implies that the entire sample is expected to coalesce to its most recent common ancestor during the bottleneck. Since the third inequality implies that the bottleneck is effectively instantaneous when time is measured in units of $2N_0$ generations, the order in which coalescent events occur during this period will have a negligible impact on the site frequency spectrum of the sample and so the genealogy can be modeled as a star-shaped tree. In this case, every mutation will give rise to a singleton and so Tajima’s D will be negative.

In contrast, if the bottleneck is both older and either less severe and/or less prolonged, e.g.,

$$
\frac{T_0}{2N_0} \approx 1 \approx \frac{T_b}{2N_b},
$$

then some lineages may pass through the bottleneck without coalescing. In this case, the tree is expected to contain internal branches that are unusually long, resulting in an excess of mutations of intermediate frequency and a positive value of Tajima’s D.
2.1 Inference of Ancestral Population Dynamics

The fact that genetic variation and genealogical relationships in a population are affected by its demographic history means that we can use genetic data to make inferences about that history. One approach is to assume that the population dynamics can be described by a model with unknown parameters \( \theta_{\text{dem}} = (\theta_1, \cdots, \theta_d) \), which are then estimated from the sequence data. For example, in the two-stage exponential model discussed above the demographic parameters are \( \theta_{\text{dem}} = (N_0, \gamma, T_{\text{exp}}) \). In the past this was often done by first inferring the genealogy \( G \) of a sample of chromosomes from the sequence data \( D \) and then finding the values of the parameters that maximize the likelihood of \( G \) under the variable population size coalescent, i.e.,

\[
\hat{\theta}_{\text{dem}} = \arg \max_{\theta_{\text{dem}}} \mathbb{P}(G|\theta_{\text{dem}}).
\]

From a statistical point of view, this approach is unsatisfactory because it treats the genealogy \( G \) as if it were known with certainty when, in reality, there are usually many different genealogies with comparable likelihoods.

A better approach is to estimate both the genealogy and the demographic parameters jointly, which can be done using Bayesian methodology. Under neutral models of evolution, the substitution process along the genealogy is independent of the demographic history of the population and so the likelihood function can be written as

\[
\mathbb{P}(D|G, \theta_{\text{dem}}, \theta_{\text{sub}}) = \mathbb{P}(D|G, \theta_{\text{sub}}),
\]

where \( \theta_{\text{sub}} \) contains the parameters of the substitution model. The objective in this case is to sample from the joint posterior distribution of the genealogy and unknown parameters

\[
p(G, \theta_{\text{dem}}, \theta_{\text{sub}}|D) \propto p(G, \theta_{\text{dem}}, \theta_{\text{sub}}) \times \mathbb{P}(D|G, \theta_{\text{sub}}).
\]

The first term on the right-hand side is the prior distribution of the genealogy and unknown parameters and is usually represented as a product of three terms

\[
p(G, \theta_{\text{dem}}, \theta_{\text{sub}}) = p(\theta_{\text{sub}}) \times p(\theta_{\text{dem}}) \times \mathbb{P}(G|\theta_{\text{dem}}),
\]

where the conditional prior on the genealogy \( p(G|\theta_{\text{dem}}) \) is specified by the variable population size coalescent. If \( T_k, k = 1, \cdots, n - 1 \) is the time when the number of lineages in \( G \) is reduced from \( k + 1 \) to \( k \) and \( T_n = 0 \), then this is equal to

\[
p(G|\theta_{\text{dem}}) = \prod_{k=2}^{n} \frac{1}{2N_{T_k}} \exp \left( -\frac{1}{2N_s} \int_{T_k}^{T_{k+1}} \frac{1}{N_s} ds \right),
\]

where \( (N_t : t \geq 0) \) is the particular demographic history specified by \( \theta_{\text{dem}} \).

Because the posterior distribution is usually too complicated to be found analytically, MCMC methods are typically used to generate a sample of trees and parameter values

\[
\left( G^{(1)}, \theta_{\text{dem}}^{(1)}, \theta_{\text{sub}}^{(1)} \right), \cdots, \left( G^{(K)}, \theta_{\text{dem}}^{(K)}, \theta_{\text{sub}}^{(K)} \right),
\]

with this distribution. Point estimates of the demographic parameters can then be found by using this sample to estimate the posterior means

\[
\hat{\theta}_{\text{dem}} = \frac{1}{K} \sum_{i=1}^{K} \theta_{\text{dem}}^{(i)}.
\]

If a plausible parametric model for the population dynamics is not available, then the skyline plot offers a non-parametric alternative for estimating the ancestral population sizes. In its
original form, the skyline plot begins with a genealogy $G$ that has been inferred from the data and estimates the population size between times $T_{k+1}$ and $T_k$ using the method of moments estimator
\[
\hat{N}_k = \left(\frac{k}{2}\right)(T_k - T_{k+1}).
\]
The result is a population size process $(\hat{N}_t : 0 \leq t \leq T_{mrca})$ which is piecewise linear, i.e.,
\[
\hat{N}_t = \hat{N}_k \quad \text{if} \ t \in (T_{k+1}, T_k],
\]
where $k = 1, \cdots, n - 1$. Like the maximum likelihood method described above, this approach incorrectly treats the genealogy $G$ as a known rather than inferred object. In addition, because the times $T_{k+1} - T_k$ and population size estimates $\hat{N}_k$ are each based on a single coalescent event, the resulting estimates are very noisy.

Both of these problems can be addressed by implementing the skyline plot within a Bayesian context. In the simplest version, the population dynamics are assumed to be piecewise constant with a prescribed number of different population sizes
\[
N_t = N^{(i)} \quad \text{if} \ t \in (t_{i-1}, t_i]
\]
for $i = 1, \cdots, m$. Here $0 = t_0 < t_1 < t_2 < \cdots < t_m = T_{mrca}$ are the unknown times when the population size changes and these are assumed to coincide with coalescent events, as in the original skyline plot. However, since $m$ is usually chosen to be much smaller than $n$, each epoch $(t_{i-1}, t_i]$ will typically contain multiple coalescent events. Although non-parametric in spirit, the Bayesian skyline plot can be represented as a demographic model in which the demographic parameters $\theta_{dem}$ include the times $t_1, \cdots, t_m$ as well as the population sizes $N^{(1)}, \cdots, N^{(m)}$. An MCMC algorithm can then be used to repeatedly sample both a genealogy $G$ and a population history $(N_t : t \geq 0)$ from the posterior distribution. In more sophisticated implementations of the Bayesian skyline plot, the number of epochs $m$ is also treated as a random variable, allowing its posterior distribution to be inferred from the sequence data.

3 Subdivided Populations

3.1 Matrix Migration Models and the Structured Coalescent

One of the core assumptions of the Wright-Fisher model is that the population is panmictic, so that each pair of individuals is equally likely to be the parents of an individual alive in the next generation. However, when a population is subdivided into two or more subpopulations (or demes) that exist in different areas, individuals that belong to different subpopulations may be less likely to mate than those that belong to the same subpopulation. To investigate the population genetical consequences of population subdivision, we will consider a model which makes the following assumptions:

1. Generations are non-overlapping.
2. The population is subdivided into $D$ demes, labeled $1, \cdots, D$.
3. The adult population size of deme $i$ is $N_i = N x_i$, where $N = N_1 + \cdots + N_D$ is the total diploid population size.
4. Each chromosome present in deme $i$ in generation $t + 1$ is an independently-chosen copy of a chromosome that was present in the previous generation. Furthermore, with probability
$m_{ij}$, they are descended from a chromosome that was present in deme $j$, in which case the ancestral chromosome is chosen uniformly at random from the $2N_j$ chromosomes that were present in this deme.

The matrix $m = (m_{ij})$ is called the **backward migration matrix** because it specifies the probability that a lineage currently in deme $i$ originated in deme $j$. In particular, $m$ is a stochastic matrix: for each $i$, $m_{i1} + \cdots + m_{iD} = 1$. The backward migration matrix can also be derived from the forward migration rates as follows. Suppose that each chromosome initially gives rise to a large number of copies of itself, say $C$, and that a proportion $f_{ij}$ of the copies produced in deme $i$ move to deme $j$. In this case $f = (f_{ij})$ is called the **forward migration matrix** because it specifies the probability that a chromosome produced in deme $i$ will migrate to deme $j$. Following migration, deme $i$ will contain a total of $\sum_{j=1}^{D} N_j f_{ij} M$ of chromosomes, of which only $N_i$, chosen uniformly at random, will survive to adulthood. Then the backward migration rates are related to the forward migration rates by the formula

$$m_{ij} = \frac{N_j f_{ij}}{\sum_{k=1}^{D} N_k f_{ik}}.$$ 

The ancestral process describing the genealogy of a sample of chromosomes from a subdivided population governed by this model depends on the scaling relationship between the backward migration rates and the coalescent rate within each deme. We will first consider the case where migration and coalescence occur on the same time scale. Since the rate of coalescence within each deme is inversely proportional to the total population size $N$, this will be true if the backward migration rates $m_{ij}, j \neq i$ are also inversely proportional to $N$. In this case the backward migration matrix $m^{(N)}$ corresponding to the model with total population size $N$ can be written in the form

$$m^{(N)} = I + \frac{1}{4N} M,$$

where $M = (M_{ij})$ is a rate matrix which contains the scaled migration rates off the diagonal and which has the property that each row sum is equal to 0: for each $i$, $M_{i1} + \cdots + M_{iD} = 1$.

With these assumptions, it can be shown that if time is measured in units of $2N$ generations, then the sequence of ancestral processes indexed by the total population size $N$ converges to a continuous-time Markov chain known as a **structured coalescent**. This process was first described by Notohara (1990) and takes values in the set of non-negative integer vectors $k = (k_1, \cdots, k_D)$ where $k_i$ denotes the number of ancestral lineages contained in deme $i$. Two kinds of events can occur in the structured coalescent: either a pair of lineages contained in the same deme can coalesce or a single lineage can migrate from one deme to another. Letting $e_i$ denote the $D$-dimensional vector with 1 in the $i$'th position and 0’s in all other positions, these transitions have the following rates:

<table>
<thead>
<tr>
<th>event</th>
<th>transition</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>coalescence in deme $i$</td>
<td>$k \rightarrow k - e_i$</td>
<td>$\binom{k_i}{2}/x_i$</td>
</tr>
<tr>
<td>migration from deme $i$ to $j$</td>
<td>$k \rightarrow k - e_i + e_j$</td>
<td>$k_i M_{ij}/2$</td>
</tr>
</tbody>
</table>

This process is called the structured coalescent because in addition to keeping track of the total number of lineages ancestral to our sample, we also need to keep track of the locations of these lineages. In particular, because lineages can only coalesce when they occupy the same deme, chromosomes that are sampled from the same deme will typically be more closely related and therefore more genetically similar than are chromosomes that have been sampled from different demes. In other words, the demes will be sufficiently isolated from one another to become partially genetically differentiated.
If the backward migration matrix $m$ is held constant when $N$ tends to infinity, then migration will occur on a much faster time scale than coalescence. In this case, gene flow will be high enough to prevent the demes from becoming genetically differentiated and it will no longer be the case that chromosomes sampled from the same deme are more closely related than those sampled from different demes. Furthermore, provided that $m$ satisfies certain technical conditions, we can again exploit the separation of time scales between these two processes to show that the sequence of ancestral processes will converge to Kingman’s coalescent when time is rescaled by the coalescent effective population size. This result is due to Notohara (1993).

**Theorem 4.** Suppose that $m$ is the transition matrix of an irreducible aperiodic Markov chain with unique stationary distribution $\pi = (\pi_1, \ldots, \pi_D)$ and let $G^{(N)}(t) : t \geq 0$ be the ancestral process corresponding to the structured Wright-Fisher model with total population size $N$, deme sizes $(Nx_1, \ldots, Nx_D)$ and backward migration matrix $m$. Then the coalescent effective population size of this model is

$$N^c_e = \beta N$$

where

$$\beta = \left( \sum_{i=1}^{D} \pi_i^2 \frac{1}{x_i} \right)^{-1}$$

and the sequence of processes $(G^{(N)}(2N^c_et) : t \geq 0)$ converges in distribution to Kingman’s coalescent.

The effective population size can be understood by observing that two lineages can only coalesce if they belong to the same deme. Since the lineages move independently of each other, at stationarity, this will happen with probability $\pi_i^2$, in which case the two lineages will coalesce at rate inversely proportional to $N_i = Nx_i$.

### 3.2 The Symmetric Island Model

The best studied matrix migration model is the so-called **symmetric island model** in which all of the subpopulations are assumed to have the same size, say $N_i = N$ for $i = 1, \ldots, D$, and the migration rates are assumed to be independent of the locations of the source and destination demes, i.e., $m_{ii} = 1 - m$ and $m_{ij} = m/(D - 1)$ for all $j \neq i$. To study the consequences of population subdivision on genetic variation, let $u$ be the locus-wide mutation rate and define $\theta = 2Nu$ and $M = 4Nm$ to be the scaled mutation and migration rates, respectively. With these definitions, the evolutionary history of a pair of lineages can be affected by three kinds of events:

- the pair coalesces at rate 1 when they belong to the same deme;
- each lineage migrates to a new deme, chosen uniformly at random, at rate $M/2$;
- each lineage mutates at rate $\theta/2$.

Our first goal is to determine how the probability of identity by descent is affected by population subdivision. To this end, let $p_s = p_s(\theta)$ and $p_d = p_d(\theta)$ be the probabilities that a pair of lineages are identical by descent when they are picked from the same or from different demes, respectively.

**Theorem 5.** If $\gamma = M/(D - 1)$ and $\omega = \theta^2 + \theta(1 + D\gamma) + \gamma$, then under the symmetric island model

$$p_s = \frac{\theta + \gamma}{\omega}, \quad p_d = \frac{\gamma}{\omega}.$$
Proof. If two lineages are chosen at random, either from the same or from different demes, then by conditioning on the first event to occur in the ancestral process, we can derive the following recursive equations for $p_s$ and $p_d$:

$$
\begin{align*}
    p_s &= \frac{1}{1 + \theta + M} \cdot 1 + \frac{M}{1 + \theta + M} \cdot p_d \\
    p_d &= \frac{M}{\theta + M} \cdot \frac{1}{D - 1} \cdot p_s + \frac{M}{1 + \theta + M} \cdot \frac{D - 2}{D - 1} \cdot p_d.
\end{align*}
$$

To solve these, first observe that the second equation can be rearranged to give the identity

$$
p_d = \frac{\gamma}{\theta + \gamma} p_s.
$$

This can then be substituted into the first equation to obtain

$$
p_s = \frac{1}{1 + \theta + M} + \frac{M}{1 + \theta + M} \cdot \frac{\gamma}{\theta + \gamma} p_s,
$$

which can be solved for $p_s$, giving

$$
p_s = \frac{1}{1 + \theta + M} \cdot \frac{(1 + \theta + M)(\theta + \gamma)}{(1 + \theta + M)(\theta + \gamma) - M\gamma} = \frac{\theta + \gamma}{\omega},
$$

where

$$
\omega = (1 + \theta + M)(\theta + \gamma) - M\gamma = \theta^2 + (M + \gamma + 1)\theta + \gamma.
$$

The expression for $p_d$ can then be derived from the relation between $p_s$ and $p_d$ shown above. \[\square\]

As expected, since chromosomes sampled from the same deme are, on average, more closely related than chromosomes sampled from different demes, $p_s > p_d$ and the ratio $p_s/p_d = 1 + \theta/\gamma = 1 + (D - 1)\theta/M$ is an increasing function of the number of demes and a decreasing function of the scaled migration rate. On the other hand, if we take the limit as the scaled migration rate tends to infinity, i.e., $M \to \infty$, then both $p_s$ and $p_d$ converge to the same value

$$
p_s = p_d = \frac{1}{1 + D\theta} = \frac{1}{1 + 4DNu},
$$

which coincides with the probability of identity by descent in a panmictic population containing $DN$ diploid individuals.

Theorem 5 can be used to calculate the expected coalescent time for a pair of lineages sampled either from the same or from different demes.

**Theorem 6.** Let $t_s$ and $t_d$ denote the coalescent times of a pair of lineages sampled either from the same or from different demes in the symmetric island model. Then

$$
\mathbb{E}[t_s] = D \quad \mathbb{E}[t_d] = D + \frac{1}{\gamma}.
$$

Proof. Since the total migration rate when there are two lineages is $M$ and each migration event has probability $1/(D - 1)$ of bringing the two lineages together in the same deme, the average waiting time until two lineages initially occupying different demes enter the same deme is $(D - 1)/M = 1/\gamma$. From this it follows that $\mathbb{E}[t_d] = \mathbb{E}[t_s] + 1/\gamma$.

To calculate $\mathbb{E}[t_s]$, observe that the probability that two lineages are identical by descent is the probability that no mutations occur on their genealogy. Since mutations are governed by a
Poisson process operating at rate $\theta/2$ on each branch of the genealogy and since the total branch length is equal to $2t_s$, it follows that

$$p_s(\theta) = \mathbb{E}[e^{-\theta t_s}].$$

In other words, $p_s(\theta)$ is the Laplace transform of the density of the random variable $t_s$. Differentiating with respect to $\theta$ and then setting $\theta = 0$ gives

$$\mathbb{E}[t_s] = -p'_s(0).$$

From Theorem 5 we know that

$$p_s(\theta) = \frac{\theta + \gamma}{\theta^2 + (1 + D\gamma)\theta + \gamma}$$

and therefore

$$p'_s(\theta) = \frac{(\theta^2 + (1 + D\gamma)\theta + \gamma) \cdot 1 - (\theta + \gamma) \cdot (2\theta + (1 + D\gamma))}{(\theta^2 + (1 + D\gamma)\theta + \gamma)^2}$$

and

$$p'_s(0) = -D.$$ 

### 3.3 Infinitely-many Demes

If the total migration rate $m$ is assumed to be inversely proportional to the within-deme population size $N$, then the results of the previous section tell us that the ancestral process for the symmetric island model converges to a structured coalescent in the limit as $N \to \infty$. In this case, the number of demes $D$ is held constant. A very different limit is obtained if we allow $D$ to tend to infinity while holding $N$ and $m$ constant. In this case, Wakeley (1998) has shown that if chromosomes are sampled at random from different demes and if time is measured in units proportional to $D$ generations, then the genealogy instead tends to Kingman’s coalescent run at a certain rate. As with Theorem 4, this result is a consequence of a separation of time scales between coalescence, which occurs on a slow times scale of order $O(D)$, and both migration and within-deme coalescence, which occur on a fast time scale of order $O(1)$.

**Theorem 7.** Under the symmetric island model, the genealogy of a sample of chromosomes containing at most one individual from each deme will tend to Kingman’s coalescent as $D$ tends to infinity if time is measured in units of $2N_c^e$ generations, where

$$N_c^e = DN \left(1 + \frac{1}{M}\right)$$

and $M = 4Nm$.

**Proof.** Because of symmetry, the state of the ancestral process for a sample initially containing $n$ chromosomes can be represented by a vector $(x_1, \ldots, x_n)$, where $x_i$ is the number of demes that contain $i$ ancestral lineages and $x_1 + 2x_2 \cdots + nx_n$ is the total number of branches in the genealogy at that time. For convenience, we will say that a deme is occupied if it contains at least one lineage that is ancestral to the sample. If the present configuration is $(k, 0, \cdots, 0)$, i.e., each deme contains at most one ancestral lineage, then the configuration will change only if one of the lineages moves from its current deme to one of the remaining $k - 1$ demes that is occupied. Such events occur at rate $k \cdot m \cdot (k - 1)/(D - 1)$ and change the configuration to $(k - 2, 1, 0, \cdots, 0)$.

Suppose, then, that the configuration is $(k - 2, 1, 0, \cdots, 0)$. In this case, the configuration can change in several different ways, which are outlined below:
1. At rate $1/2N$, the two lineages occupying the same deme coalesce, resulting in the new configuration $(k - 1, 0, \cdots, 0)$.

2. At rate $2m \cdot (1 - \frac{k-2}{D-1})$, one of these two lineages moves to an unoccupied deme, restoring the configuration to $(k, 0, \cdots, 0)$.

3. At rate $2m \cdot (k-2)/(D-1)$, one of these two lineages moves to an occupied deme containing just one ancestral lineage, leaving the configuration unchanged.

4. At rate $(k-2)m \cdot 1/(D-1)$, one of the isolated lineages moves into the deme containing two ancestral lineages, resulting in the new configuration $(k - 3, 0, 1, 0, \cdots, 0)$.

5. At rate $(k-2)m \cdot (k-3)/(D-1)$, one of the isolated lineages moves into another deme containing an isolated lineage, resulting in the new configuration $(k - 4, 2, 0, \cdots, 0)$.

If $m$ and $N$ are held constant while $D \to \infty$, then because the last three events occur at rates that are order $O(D)$ times smaller than the first two events, we can ignore these possibilities in the limit. In particular, configurations with either $x_i > 0$ for $i \geq 3$ or $x_2 > 1$ will never be visited in this limit, while configurations with $x_2 = 1$ will only be transiently visited and will either result in a single pairwise coalescence event with probability

$$\frac{1}{2N} \frac{1}{2m + 1/2N} = \frac{1}{1 + M}$$

or restore the previous configuration with probability $M/(1 + M)$. It follows that pairwise coalescent events occur at rate

$$2 \binom{k}{2} m \left( \frac{1}{D-1} \right) \left( \frac{1}{1+M} \right) = \binom{k}{2} \left( \frac{D}{D-1} \right) \left( \frac{2m}{D(1+m)} \right)$$

when there are $k$ branches in the genealogy and setting this equal to $\binom{k}{2} \frac{1}{2N_c}$ gives

$$N_c \approx D \left( \frac{1+M}{4m} \right) \approx ND \left( 1 + \frac{1}{M} \right).$$

One consequence of Theorem 7 is that if $M = 4Nm \gg 1$, then $N_c \approx DN$, which is also the census population size. In other words, if the migration rate is high enough that lineages occupying the same deme are likely to disperse to different demes before they can coalesce, then the symmetric island model behaves like a single panmictic population with effective population size equal to its census population size. On the other hand, if $M \leq 1$, then the effective population size of the metapopulation can be much larger than its census population size, essentially because it takes a long time for lineages sampled from different demes to be gathered into the same deme where they can coalesce.

If multiple chromosomes are sampled from the same deme, then we have to account for the possibility that some of the lineages coalesce in that deme, a process that is effectively instantaneous on the slow time scale that governs coalescence between lineages sampled from different demes. To this end, suppose that $n$ chromosomes have been sampled from a single deme and let $L_n$ denote the number of lineages that escape from that deme into the metapopulation. In the infinitely many demes limit, the probability that a lineage returns to a deme that it has previously escaped is negligible and so we need only consider the random sequence of coalescence and dispersal events that affects the initial sample. This process can be represented by a continuous-time Markov chain in which each pair of lineages coalesces at rate $1/2N$ and each lineage disperses out of the deme and is killed at rate $m$. Equivalently, if time is measured in units
of $2N$ generations, then this is the coalescent with killing previously studied in the context of the infinite alleles model with killing rate $M = 4Nm$. In particular, if we equate $\theta$ with $M$, then the distribution of the number of lineages that escape the deme is the same as the distribution of the number of alleles in the infinite alleles model, i.e.,

$$P(L_n = k) = \frac{M^k}{M(k)}|S_n^k|,$$

where $M(k) = M(M+1)\cdots(M+k-1)$ and $|S_n^k|$ is the number of permutations of $n$ elements with exactly $k$ cycles, i.e., a Stirling number of the first kind.

By combining these two results, we can give a unified description of the genealogy of a stratified random sample of chromosomes from a metapopulation governed by the infinitely-many demes island model. Here stratified random sampling means that we first sample $d$ demes at random and then sample $n_1$ from the first deme, $n_2$ from the second deme, etc. Each deme then independently passes through the scattering phase in which a sequence of coalescence and migration events reduces the number of lineages in the $i$’th deme from $n_i$ to $L(i)$, which are scattered to different demes. At the end of the scattering phase, the collection phase begins with a total of $L = L^{(1)} + \cdots + L^{(d)}$ lineages in as many demes and then the results of Theorem 7 apply, i.e., the genealogy of these $L$ lineages is governed by Kingman’s coalescent when time is measured in units of $Nc$ generations.

### 3.4 Fixation Indices

The distribution of genetic variation in a subdivided population can be described in many ways, but among the most commonly used metrics are the fixation indices $F_{IS}$, $F_{ST}$ and $F_{IT}$ introduced by Wright (1951). To define these, consider a biallelic locus segregating two alleles $A$ and $a$ in a population subdivided into $D$ demes and let $p_i$ and $P_i$ be the frequencies of allele $A$ and genotype $AA$, respectively, in deme $i$. Assuming that the subpopulations are all of equal size, the global frequencies of $A$ and $AA$ are equal to

$$\bar{p} = \frac{1}{D} \sum_{i=1}^{D} p_i, \quad \bar{P} = \frac{1}{D} \sum_{i=1}^{D} P_i.$$  

The fixation index $F_{IT}$ is a measure of the deviation from Hardy-Weinberg equilibrium at the global level and is defined by the equation

$$\bar{P} = (1 - F_{IT})\bar{p}^2 + F_{IT}\bar{p}$$

giving the identity

$$F_{IT} = \frac{\bar{P} - \bar{p}^2}{\bar{p}(1 - \bar{p})}.$$  

Similarly, the fixation index $F_{IS}$ is a measure of the average deviation from Hardy-Weinberg equilibrium at the subpopulation level and is defined by the equation

$$\bar{P} = (1 - F_{IS})\bar{p}^2 + F_{IS}\bar{p},$$

where

$$\bar{p}^2 = \frac{1}{D} \sum_{i=1}^{D} p_i^2.$$  

This gives the identity

$$F_{IS} = \frac{\bar{P} - \bar{p}^2}{\bar{p} - \bar{p}^2}.$$
The third fixation index $F_{ST}$ is then defined by requiring that

$$1 - F_{IT} = (1 - F_{ST})(1 - F_{IS}),$$

giving

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}} = \frac{\bar{p}^2 - \overline{p^2}}{\overline{p}(1 - \overline{p})} = \text{Var}(p),$$

where $\text{Var}(p)$ is the variance of the allele frequency across subpopulations. Furthermore, since $\text{Var}(p) \leq \overline{p}(1 - \overline{p})$, this shows that $F_{ST}$ takes values in $[0, 1]$, with larger values corresponding to greater levels of genetic differentiation between subpopulations. In particular, if there is no differentiation, then $A$ will have the same frequency in every deme and $F_{ST} = 0$. In contrast, if each deme contains just one allele, then $\text{Var}(p) = \overline{p}(1 - \overline{p})$ and $F_{ST} = 1$.

The fixation indices can also be defined in terms of observed and expected heterozygosities at different levels. Let $H_T$ be the expected heterozygosity assuming that Hardy-Weinberg equilibrium holds at the global level, let $H_S$ be the expected heterozygosity assuming that Hardy-Weinberg equilibrium holds within each subpopulation, and let $H_I$ be the observed heterozygosity averaged over all of the subpopulations. Then

$$F_{IS} = \frac{H_S - H_I}{H_S} \quad \text{and} \quad 1 - F_{IS} = \frac{H_I}{H_S},$$
$$F_{IT} = \frac{H_T - H_I}{H_T} \quad \text{and} \quad 1 - F_{IT} = \frac{H_I}{H_T},$$
$$F_{ST} = \frac{H_T - H_S}{H_T} \quad \text{and} \quad 1 - F_{ST} = \frac{H_S}{H_T},$$

and, as before, the identity $1 - F_{IT} = (1 - F_{ST})(1 - F_{IS})$ holds. In practice, the heterozygosities and allele frequencies must be estimated from sample data, in which case the estimated fixation indices are typically denoted $G_{IS}$, $G_{ST}$ and $G_{IT}$.

The index $F_{ST}$ can also be given a genealogical interpretation. We start by rewriting the last equation as

$$F_{ST} = \frac{(1 - H_S) - (1 - H_T)}{1 - (1 - H_T)},$$

where $1 - H_S$ is the probability that two chromosomes sampled from the same subpopulation carry the same allele and $1 - H_T$ is the probability that two chromosomes sampled from the global population carry the same allele. Under the infinite alleles model, these quantities are equal to the probability that two alleles sampled either from the same subpopulation or from the global population are identical by descent. Denoting these probabilities $f_0$ and $\bar{f}$, respectively, we have

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}.$$

Thus, $F_{ST}$ can also be interpreted as a measure of how much more closely related two chromosomes sampled from the same deme are, on average, than two chromosomes sampled at random from the entire population.

Because the probability of identity by descent depends on the mutation rate at the locus under investigation, the same is true of $F_{ST}$. However, it can be shown that this dependence is weak when $\mu$ is small. To see this, let $T$ and $T_0$ be the pairwise coalescent times for the entire population and a single deme, respectively, and recall that the probability of identity by descent under the infinite sites model can be expressed as the Laplace transform of the pairwise coalescent time, e.g.,

$$f = \mathbb{E}[e^{-\mu T}] \approx 1 - \mu \bar{T},$$
where $\bar{T}$ is the mean pairwise coalescent time and the approximation is accurate if $\mu\bar{T} \ll 1$. Since a similar argument gives $f_0 \approx 1 - \mu\bar{T}_0$, the fixation index can be approximated by the quotient

$$F_{ST} \approx \frac{\bar{T} - \bar{T}_0}{\bar{T}},$$

which has a simple genealogical interpretation.

For the symmetric island model with $D$ demes, we know from Theorem 5 that $f_0 = p_s = (\theta + \gamma)/\omega$ and

$$\bar{f} = \frac{1}{D} p_s + \left(1 - \frac{1}{D}\right) p_d = \frac{1}{D} \frac{\theta}{\omega} + \frac{\gamma}{\omega},$$

since $p_d = \gamma/\omega$. This shows that $f_0 - \bar{f} = (1 - 1/D)(p_s - p_d) = (1 - 1/D)(\theta/\omega)$, which gives

$$F_{ST} = \frac{(1 - \frac{1}{D}) \theta}{\omega - \frac{\theta}{D} - \gamma} = \frac{(1 - \frac{1}{D}) \theta}{\theta^2 + \theta(1 + D\gamma - 1/D)} \approx \frac{D - 1}{D^2\gamma + D - 1},$$

where the approximation is accurate if $\theta \ll 1$. Recalling that $\gamma = M/(D - 1)$, the last expression can be rewritten as

$$F_{ST} \approx \frac{1}{1 + 4Nm\frac{D^2}{(D-1)^2}} \approx \frac{1}{1 + 4Nm},$$

where the second approximation is accurate if $D \gg 1$ and is exact in the infinitely-many demes limit. Thus, at least in the symmetric island model, we can expect the different demes to be significantly genetically differentiated if $M = 4Nm \ll 1$, i.e., if intrademic drift operates more rapidly than migration.