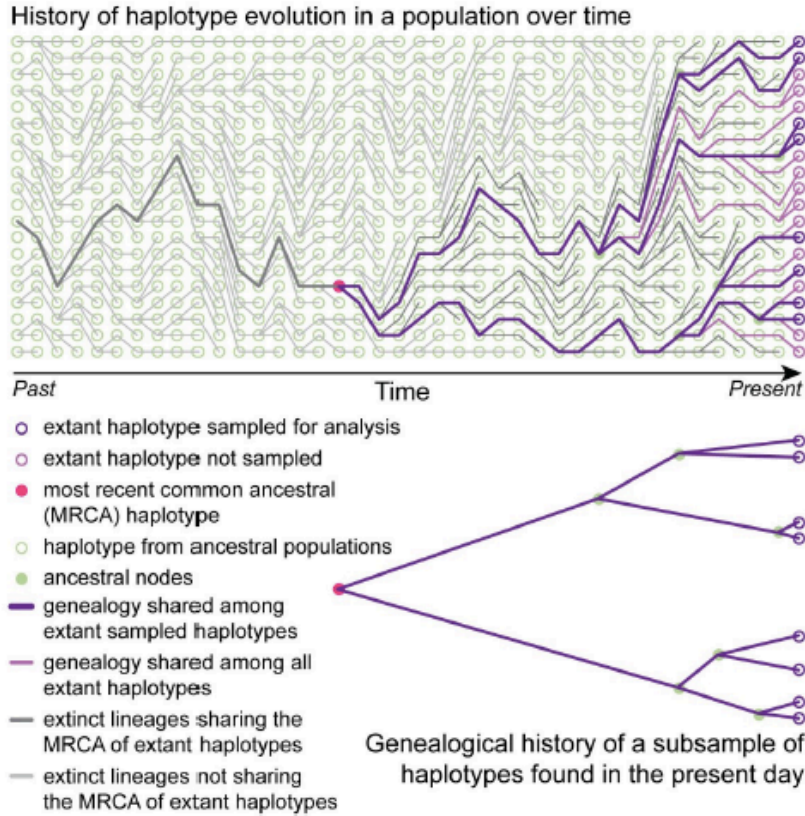


Introduction to Coalescence

In our discussion of identity by descent (or inbreeding), we looked forward in time. In the absence of mutation or migration, the probability of IBD increased over time approaching 1. This means, that eventually all the alleles in the population will be descended from a single common ancestor allele. In coalescent theory, we look backward in time to ask when two alleles last shared a common ancestor.



From book by A. Cutter (2019).

Time to coalescence for two alleles

In an ideal population of N diploids, the probability that two alleles have descended from the same allele in the previous generation (i.e., the probability of coalescence) is,

$$P_C = 1/(2N)$$

And the probability of not coalescing is

$$P_{NC} = 1 - P_C = 1 - 1/(2N)$$

The probability that two alleles coalesce $(t + 1)$ generations ago is

$$P_{C,t+1} = \frac{1}{2N} \left(1 - \frac{1}{2N} \right)^t$$

 Q4.1) Explain why this equation makes sense.

This result is an example of the well-studied *geometric distribution* that has the general form $P(x) = b(1 - b)^{x-1}$.

For this distribution, it is known that $E(x) = 1/b$ and $V(x) = (1 - b)/b^2$.

In our case, $b = 1/(2N)$, so that $E(T) = 2N$ and $V(T) = 4N^2(1 - 1/(2N)) \cong 4N^2$

Aside on the Exponential Distribution

If independent events are occurring at a rate of c events per instant of time, then the time until the *first* event follows an exponential distribution

$$P(t) = ce^{-ct}$$

which has the properties $E(t) = 1/c$ and $V(t) = 1/c^2$.

Because the expected number of coalescent events between two alleles per generation is equal to the probability the two alleles coalesce in 1 generation, we could use the exponential distribution to describe the time to coalescence with $c = 1/(2N)$. There are some subtle differences because the exponential distribution assumes time is continuous rather than parceled into discrete generations. Provided that N is not too small, the expected time to coalescence will be long so that time is effectively continuous.

Much of the work done in formal coalescence theory assumes that N is large and time is continuous. Although I do not use it here, most coalescent work measures time on the **coalescent time scale** $\tau = 2N$ generations. A bit more formally, this proceeds as follows.

The probability of going from i to $i - 1$ samples in a single generation is

$$P(i \rightarrow i - 1) = \binom{i}{2} \frac{1}{2N} \left(1 - \frac{1}{2N}\right) \left(1 - \frac{2}{2N}\right) \cdots \left(1 - \frac{i-1}{2N}\right)$$

Note this is

$$P(i \rightarrow i - 1) = \binom{i}{2} \frac{1}{2N} + O\left(\frac{1}{N^2}\right)$$

Instead, we multiply $P(i \rightarrow i - 1)$ by $2N$ (the coalescent time scale) and then take the limit as $N \rightarrow \infty$

$$\lim_{N \rightarrow \infty} (2N * P(i \rightarrow i - 1)) = \binom{i}{2}$$

This gives the *rate* of coalescence on the coalescent time scale. A rate is the expected number of events per unit time. So in the simplest case of $i = 2$, we expect coalescence to take 1 time unit ($2N$ generations).

Note that $P(i \rightarrow i - k) \sim O\left(\frac{1}{N^k}\right)$ so that for $k \geq 2$, $\lim_{N \rightarrow \infty} (2N * P(i \rightarrow i - k)) = 0$

Expected divergence between two alleles: If two alleles coalesced t generations ago, then each allele is will have accumulated μt mutations in that time, where μ is the mutation rate. Assuming that each mutation is different (infinite sites model of mutation), then the number of differences separating two sequences will be $2\mu t$. The expected value is $2\mu E(t) = 4N\mu$. The quantity $4N\mu$ arises so often in population genetics that it gets its own symbol $\theta = 4N\mu$.

Coalescence of multiple alleles.

Previously, we considered the time to coalescence of 2 alleles. Now let's consider 3 alleles.

With 3 alleles, the expected number of coalescent events in one generation is:

$$E(\text{\# of pairwise coalescent events}) = \frac{\binom{3}{2}}{2N} = \frac{3}{2N}$$

Q4.3) Explain why.

We can again employ the exponential distribution, to get the distribution of times to the first coalescence,

$$P(\text{1st coalescence in gen } t+1) = \left(\frac{3}{2N}\right) e^{-3t/2N}$$

which has the expected value of $E(t) = 2N/3$

The expectations and the probability given above apply only for so long as there are 3 separate alleles. After the first coalescence, only two alleles remain and we need to use the equations we previously derived for considering two alleles. The expected time for the coalescence of all 3 alleles would simply be the expected time to the first coalescence when there are 3 alleles ($2N/3$) plus the expected time to the first (and only remaining) coalescence when there are two 2 alleles ($2N$), giving a total time of $2(1 + 1/3)N$.

In general, if we are considering k alleles then

$$E(\text{\# of pairwise coalescent events}) = \binom{k}{2} \frac{1}{2N} = \frac{k(k-1)}{2} \frac{1}{2N} = \frac{k(k-1)}{4N}$$

Again using the exponential distribution,

$$P(\text{1st coalescence in gen } t+1) = \frac{k(k-1)}{4N} e^{-\frac{k(k-1)t}{4N}}$$

$$\text{and } E(t_k) = \frac{4N}{k(k-1)}$$

$E(t_k)$ is the expected time to go from k to $(k - 1)$ alleles.

Let us now consider all $2N$ alleles in the population and find the expected time for all of them to coalesce.

$$\begin{aligned} E(\text{time to coalescence of all alleles}) &= \sum_{k=2}^{2N} \frac{4N}{k(k-1)} \\ &= 4N \sum_{k=2}^{2N} \frac{1}{k(k-1)} \end{aligned}$$

Using the general relationship, $\sum_{x=2}^B \frac{1}{x(x-1)} = 1 - \frac{1}{B}$

$$E(\text{time to coalescence of all alleles}) = 4N(1 - 1/(2N)) \cong 4N.$$

Thinking forward in time, rather than backwards, this result tells us the expected time it takes a new allele to drift to fixation, given that it eventually becomes fixed.

Recall that if there are only two alleles, the expected time to coalescence between them will be $2N$. The result above indicates that half of the expected time to coalescence comes from the last two alleles.

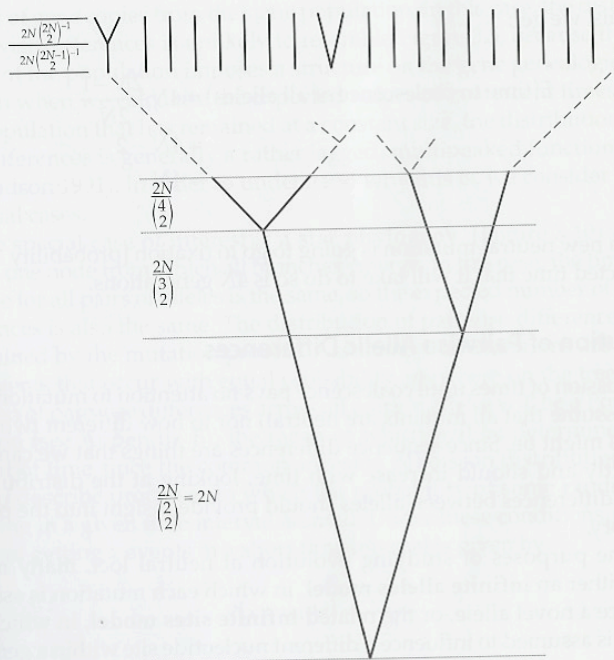


Figure 3.9 Coalescent diagram for all allele copies in a population.

Figure from 2004 book by S. Rice.

Expected Amount of Polymorphism

We want to know how many variable sites (X) we can expect to find in a sample of k alleles. Assuming each mutation occurs at a unique site (i.e., the infinite sites model of mutation), X will be equal to the number of mutations that have occurred during the total time over all the branches of these k alleles back to a common ancestor. If μ is the mutation rate per generation and Y_k is the total time over all the branches of these k alleles back to a common ancestor, then $X = \mu Y_k$. Thus, we just need to calculate Y_k . To do so, consider that there are k lineages that experience time to coalesce to $(k - 1)$ and there are $(k - 1)$ that experience time to coalesce to $(k - 2)$, and so on until there are only two lineages that experience the time to coalesce to 1 common ancestor. We previously calculated the time it takes to go from k to $(k - 1)$ lineages as $E(t_k) = 4N/(k(k - 1))$. We can now calculate the expected value of Y_k as,

$$E(Y_k) = \sum_{i=2}^k i E(t_i) = \sum_{i=2}^k i \frac{4N}{i(i-1)} = 4N \sum_{i=2}^k \frac{1}{(i-1)} = 4N \sum_{i=1}^{k-1} \frac{1}{i}$$

Thus, the expected number of variable sites is simply

$$E(X) = \mu E(Y_k) = 4N\mu \sum_{i=1}^{k-1} \frac{1}{i}$$

 Q4.4) Here we assumed an infinite sites model of mutation (each mutation affected a different site). In reality, there are a finite number of sites. The infinite sites assumption will be a problem when the mutation rate is high relative to the inverse of the coalescent time. Why is this a problem and explain whether it would cause us to over- or under-estimate X using the formula above.

Mutation-Drift Balance: Heterozygosity

We know that drift erodes variation and mutation introduces it. How much sequence variation do we expect if when both processes are acting? One way to address this issue is to ask about the average heterozygosity, \bar{H} , i.e., the probability that two randomly chosen alleles will have different sequences. Note here we aren't paying attention to how different two sequences are (infinite alleles rather than infinite sites).

Two alleles will only be different if there has been a mutation event since they descended from a common ancestor. We can follow the two alleles back in time until SOMETHING happens and ignore all the generations where nothing happens. The 'something' will either be that they coalesce or that there has been a mutation event. The probability that the something is a mutation gives the average heterozygosity:

$$\begin{aligned}
\bar{H} &= P(\text{mutation} \mid \text{something}) \\
&= \frac{P(\text{mutation})}{P(\text{something})} \\
&= \frac{P(\text{mutation})}{P(\text{mutation}) + P(\text{coalescence})} \\
&= \frac{2\mu}{2\mu + \frac{1}{2N}} \\
&= \frac{4N\mu}{4N\mu + 1} \\
&= \frac{\theta}{\theta + 1}
\end{aligned}$$

Q4.5) What were we assuming in denominator of the 3rd line?

Q4.6) Go back to our discussion of inbreeding (IBD) and mutation. Discuss the two results.

Structured coalescent – 3 approaches

In the coalescent model above, there was no “structure”. Any two alleles could be sampled and the expected time to coalescence would be the same; in this sense any two alleles were equivalent. However, there are a number of important biological situations where we do not expect any two alleles to have the same average coalescence time. Rather, there is some underlying structure that affects how likely two alleles are to coalesce in a given generation.

We will consider two examples. First, consider the island model of population subdivision. When a species is distributed across space in a series of demes where individuals tend to mate locally and rarely migrate, then two alleles within a deme will be more closely related, on average, than two alleles chosen from different demes. The expected time to coalescence is shorter for alleles from the same deme than for two alleles chosen from different demes. Intuitively, this is because two alleles in different demes cannot coalesce in the immediately preceding generation, there must first be a migration event that brings the two alleles into the same deme before coalescence can occur. The second case we will consider is selfing. In species with high rates of selfing, two alleles chosen from a single individual have a much higher rate of coalescing in the immediately preceding generation than two alleles sampled from different individuals. In both cases, at any given instant, the two alleles being considered can be in one of three states. Depending on the biological situation, there may be many more than three states but we will only consider the three state case but the same types of approaches can be applied to more complicated situations.

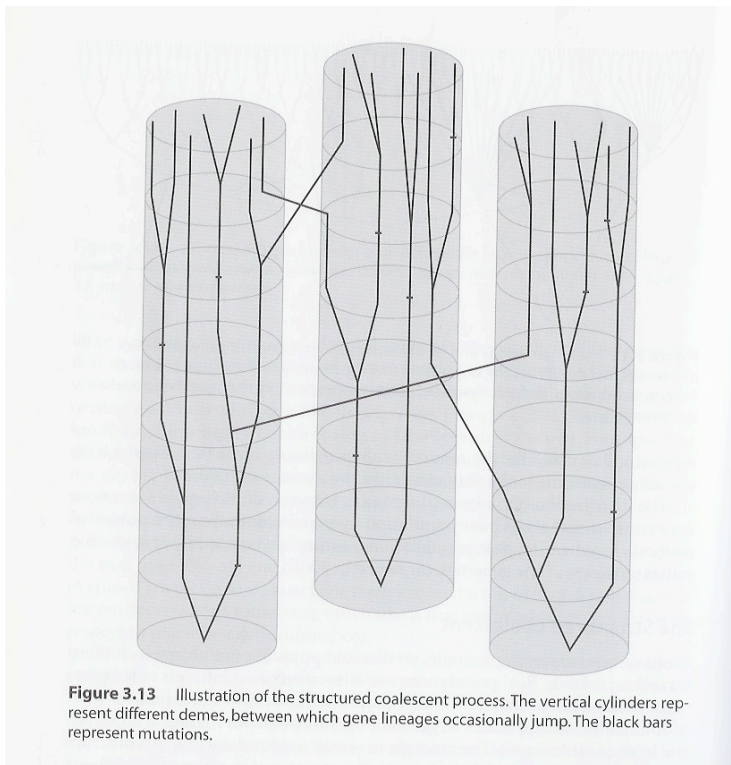
	Island Model	Selfing
State 1	Both alleles in same deme	?
State 2	Alleles in separate demes	?
State 3	Coalesced	Coalesced

Table 4.1. Possible states for two alleles in the models considered here.

There are a variety of ways to model the structured coalescent and we will look at three different approaches. It is worth noting that the distribution of coalescence times in the structured coalescence will not be an exponential distribution as it is in the simple unstructured coalescence model. In some structured models, the entire distribution can be calculated analytically but it can be tricky to do so. Here we will focus only the mean coalescence times.

Before going into the three approaches, a few notes and parameter definitions for the two models we will examine. Note we also make a few assumptions.

Island model: We assume that there are d demes, each consisting of n diploid individuals (total population size is $N = dn$). There is a low migration rate m among demes (m is the probability that an individual migrates out of a deme in a given generation). Migration is random among demes (i.e., some demes are *not* more strongly connected by migration than others). We will assume that $n \gg 1$ and $m \ll 1$. From this, we can ignore the possibility of more than one “event” occurring per generation (i.e., no more than one migration event or a coalescence).



From 2004 book by S. Rice.

Selfing model: We assume that individuals reproduce by (sporophytic) selfing with probability S and otherwise by random mating. The population is of size N .

Approach 1: “Time to leave current state”

This type of approach was used by Nordborg (1997, *Genetics*) and others. This is called “First Step Analysis” in Wakeley’s (2009) book.

In our model, we have three states (see table above). Let a_{ij} be the probability of moving from state i this generation to state j in the immediately preceding generation. We will assume the third state

represents coalescence, which is an absorbing state so $a_{31} = a_{32} = 0$ and $a_{33} = 1$. The other probabilities are given below.

Transition Probability	Island Model	Selfing Model
a_{11}	$1 - (a_{12} + a_{13})$?
a_{12}	$2m$?
a_{13}	$1/2n$?
a_{21}	$2m/(d-1)$?
a_{22}	$1 - (a_{21} + a_{23})$?
a_{23}	0	?

Table 4.2. Transition probabilities between states.

Q4.7) Go through the transition probabilities for the island model in Table 4.2 and explain why each one is what it is. (Remember the assumptions listed in the description of the island model.)

Q4.8) Go through the transition probabilities for the selfing model in Table 4.2 and explain why each one is what it is.

Let's consider the expected time to coalescence from State 1, $E[T_{C1}]$. As long as it remains in State 1, coalescence cannot occur. So, at a minimum, the coalescence time will be at least as long as the expected time to leave State 1. In any given generation, the probability of leaving State 1 is $(1 - a_{11})$. So the average waiting time to leave State 1 is $1/(1 - a_{11})$. [The waiting time to leave State 1 is geometrically distributed with parameter $b = (1 - a_{11})$ and the mean of the geometric distribution is $1/b$]. Conditional on leaving State 1, the two alleles could move into either State 2 or State 3 (coalescence). In the latter case, there is no additional time to coalescence. In the former case, we need to add in the expected coalescence time for two alleles in State 2, but weighted by the probability of moving into State 2 rather than State 3. This leaves us with:

$$E[T_{C1}] = \frac{1}{1 - a_{11}} + \frac{a_{12}}{a_{12} + a_{13}} E[T_{C2}] + \frac{a_{13}}{a_{12} + a_{13}} * 0$$

$$E[T_{C1}] = \frac{1}{1 - a_{11}} + \frac{a_{12}}{a_{12} + a_{13}} E[T_{C2}] \quad [4.1a]$$

Using the same logic, we obtain

$$E[T_{C2}] = \frac{1}{1 - a_{22}} + \frac{a_{21}}{a_{21} + a_{23}} E[T_{C1}] \quad [4.1b]$$

Simultaneously solving the two equations above for $E[T_{C1}]$ and $E[T_{C2}]$, we obtain

$$E[T_{C1}] = \frac{a_{12} + a_{21} + a_{23}}{a_{12}a_{23} + a_{13}(a_{21} + a_{23})}$$

and

$$E[T_{C2}] = \frac{a_{12} + a_{21} + a_{13}}{a_{12}a_{23} + a_{13}(a_{21} + a_{23})}$$

Making the appropriate substitutions into [4.1] for the island model, this means that the expected time to coalescence for two alleles sampled from the same deme is $E[T_{C1}] = 2dn$ and the expected time to coalescence for two alleles sampled from separate demes is $E[T_{C2}] = 2dn + (d-1)/2m$.

As expected from intuition, $E[T_{C2}] > E[T_{C1}]$ and reducing the migration rate increases $E[T_{C2}]$. Curiously, $E[T_{C1}]$ depends on the number of demes but not on the migration rate. An explanation for this result is that as m gets small the probability of migration to another deme decreases but the effect of a migration event on the time to coalescence gets bigger so in the end there is no net effect. Note the solution for $E[T_{C1}]$ cannot be correct if $m = 0$, because then each deme is a completely independent population and we can apply our result from analyzing an unstructured population, $E(T) = 2n \neq 2nd$. If we had we calculated the variance for our structured model, we would see that $V[T_{C1}] \rightarrow \infty$ when $m \rightarrow 0$, indicating that the associated average, $E[T_{C1}]$, is not meaningful. (Later we will see a method that lets us calculate the variance too.)

Under the infinite sites model of mutation, the number of base-pair differences between two alleles should be a linear function of their time since coalescence. By comparing multiple sequences from each of multiple demes, it is possible, in principle, to use the equations above to estimate deme size n and number d .

If we pick two alleles at random from anywhere in the metapopulation

$$E[T_{C.}] = \frac{1}{d}E[T_{C1}] + \left(1 - \frac{1}{d}\right)E[T_{C2}] = 2dn + \frac{(d-1)^2}{2dm}$$

Because variation between alleles is directly proportional to coalescence time, we can measure F_{ST} as

$$F_{ST} = \frac{E[T_{C.}] - E[T_{C1}]}{E[T_{C.}]} = \frac{(d-1)^2}{1 - 2d + d^2(1 + 4mn)}$$

This does not match the result we found in the “Identity By Descent” section because there we implicitly assumed an infinite number of demes. If we do the same here, we recover the classic result.

$$\lim_{d \rightarrow \infty} F_{ST} = \lim_{d \rightarrow \infty} \frac{(d-1)^2}{1 - 2d + d^2(1 + 4mn)} = \frac{1}{1 + 4mn}$$

<< I have removed the other two approaches as we will not be covering them. Consequently, there will be a gap in the numbering of equations. >>>

Coalescent times for neutral sites linked to selected sites

So far we have assumed complete neutrality. Now we will consider the coalescence of neutral sites that are linked to sites under selection. We will consider two forms of selection at a linked site: balancing selection and background selection. We will use the same idea as population structure to examine selection at a linked locus.

Balancing Selection at a Linked Locus

Imagine that there is balancing selection on locus S such that the S_1 allele is maintained at frequency p and the S_2 allele at frequency $q = (1 - p)$. In a population of N diploids, there will be $2N_1 = 2Np$ copies of the S_1 allele and $2N_2 = 2Nq$ copies of the S_2 allele, i.e., $N_1 = Np$ and $N_2 = Nq$. Imagine a closely linked neutral locus. From the perspective of this neutral locus, we can think of the S_1 alleles as representing one deme and the S_2 alleles as representing a second deme. An allele at the neutral locus that is currently in deme 1 (i.e., linked to an S_1 allele) can 'migrate' to deme 2 via recombination. Remember that m_1 is defined as the probability that an allele in deme 1 came from deme 2 in the previous generation. We want to know what the probability that an allele at the neutral locus that is linked to S_1 was linked to S_2 in the previous generation. Let G_{11} , G_{22} , and G_{12} be the number of S_1 homozygotes, S_2 homozygotes and the heterozygous individuals (here G_{12} represents all heterozygotes, $S_1S_2 + S_2S_1$). Because recombination is only relevant in heterozygotes, the number of 'migrant' alleles (not individuals) onto the S_1 background is rG_{12} and the same is true for the number of migrant alleles onto the S_2 background. We can then write,

$$m_1 = rG_{12}/2N_1 \text{ and } m_2 = rG_{12}/2N_2$$

(Note $m_1 = rG_{12}/2N_1$ is the number of alleles that migrated onto the S_1 haplotype over the total number of S_1 alleles.)

Assuming $G_{12} = 2pqN$, we get,

$$m_1 = rq \text{ and } m_2 = rp$$

Using this in eq 4.7 we get

$$E(T_{11}) = 2N + \frac{2Nq(p - q)}{4Nr pq + 1} \quad [\text{Eq.4.10a}]$$

$$E(T_{22}) = 2N + \frac{2Np(q - p)}{4Nr pq + 1} \quad [\text{Eq.4.10b}]$$

$$E(T_{12}) = 2N + \frac{1}{r} \quad [\text{Eq.4.10c}]$$

If two alleles at the neutral locus are picked at random (we don't know whether they are linked to S_1 or S_2), what is their expected time to coalescence?

$$E(T) = 2N + \frac{2pq(1 + Nr)}{r(1 + 4pqNr)}$$

Recall that in the absence of balancing selection, $E(T) = 2N$. The result above shows that balancing selection can increase the time to coalescence of closely linked neutral alleles. Sequence variation among alleles increases with their time to coalescence. Thus regions of the genome experience balancing selection should also have elevated levels of neutral variation. Note that per base pair recombination rates are $\sim 10^{-8}/\text{bp}$

 Q4.16) Show how to go from Eq. 4.7 to 4.10.

Q4.17) Why might it be bad to assume $G_{12} = 2pqN$?

You could also obtain the results in [4.10] using the “*time to leave current state*” approach by having four states:

State 1: both on the S_1 background

State 2: both on the S_2 background

State 3: one on the S_1 background and the other on the S_2 background

State 4: coalesced

Transition probability	Balancing Selection	Background Selection
a_{12}	0	0
a_{13}	$2qr$	$2qr$
a_{14}	$1/(2Np)$	$1/(2Np)$
a_{21}	0	0
a_{23}	$2pr$	$2p(hs + r)$
a_{24}	$1/(2Nq)$	$1/(2Nq)$
a_{31}	pr	$p(hs + r)$
a_{32}	qr	qr
a_{34}	0	0

Table 4.3. Transition probabilities for 4 state model (diploid with two different haplotype genetic backgrounds). Note that $a_{ii} = 1 - \sum_{j \neq i} a_{ij}$

Simultaneously solve the three equations:

$$E[T_{C1}] = \frac{1}{1-a_{11}} (1 + a_{12} E[T_{C2}] + a_{13} E[T_{C3}]) \quad [4.11a]$$

$$E[T_{C2}] = \frac{1}{1-a_{22}} (1 + a_{21} E[T_{C1}] + a_{23} E[T_{C3}]) \quad [4.11b]$$

$$E[T_{C3}] = \frac{1}{1-a_{33}} (1 + a_{31} E[T_{C1}] + a_{32} E[T_{C2}]) \quad [4.11c]$$

Substituting in the transition probabilities for the balancing selection model in Table 4.3 returns the same coalescent times given in [4.10].

Background Selection at a Linked Locus

We now consider the case where a linked locus experiences recurrent deleterious mutation.

Specifically, the wild-type S_1 allele mutates to the deleterious alternative S_2 at rate μ . Fitnesses of the three genotypes S_1S_1 , S_1S_2 , and S_2S_2 , are 1, $1 - hs$, and $1 - s$. We make the typical assumptions of the canonical mutation-selection balance model, $\mu \ll hs$, so the frequency of the deleterious S_1 allele is $q = \mu/hs$. Furthermore, for this coalescence analysis, we also need to assume that this frequency is stable, requiring $Nq \gg 1$.

As in the balancing selection model, we can use $m_1 = qr$. However, the other migration term is $m_2 = pr + pu/q$ because there are two ways for our focal neutral site to move onto an S_2 background from an S_1 background in the preceding generation. The focal neutral site can change backgrounds via recombination (which happens with probability pr). Alternatively, an S_1 background in the immediately preceding generation can mutate into an S_2 background. The chance that the focal S_2 was created by mutation in the immediately preceding generation is pu/q (because pu is the mutational input per generation of S_2 alleles and q is the total frequency of S_2 alleles, i.e., pu/q is the probability that a given S_2 allele was created by mutation in the immediately preceding generation). You can think of pu/q as the being the probability of being a **newly** created S_2 allele (“ pu ”) given that it is an S_2 allele (“ q ”). Because $q = \mu/hs$, we can write $m_2 = p(r + hs)$.

Now, we can use $N_1 = Np$, $N_2 = Nq$, $m_1 = qr$ and $m_2 = p(r + hs)$ in [4.7]. The resulting solution is a bit messy but can be greatly simplified if we assume that q , hs , $r \sim O(\xi)$ and that $1/N \sim O(\xi^3)$, i.e., doing a Taylor series on the result after using these assumptions. Then, we find that

$$E[T] = p^2 E[T_{11}] + 2pq E[T_{12}] + q^2 E[T_{22}] = 2N \left(1 - \frac{q}{(1 + \frac{r}{hs})^2} \right) \quad [4.12]$$

which is (almost) the same result derived by other quite different methods by Hudson & Kaplan (1995, *Genetics*) and by Nordborg, Charlesworth, and Charlesworth (1996, *Genetical Research*). In constructing our model, we said that the number of 'migrant' alleles (not individuals) onto the S_1 background is rG_{12} (because recombination is only relevant in heterozygotes) and the same is true for the number of migrant alleles onto the S_2 background. However, remember that the heterozygotes have reduced fitness so it should really be $r(1 - hs)G_{12}$ instead. If we use $r' = r(1 - hs)$ in place of r in [4.12] we recover the Nordborg *et al* (1996) result.

Though we obtained [4.12] using the *moment generating functions* approach, we could just as well have used the “*time to leave current state*”, following from [4.11] and using the appropriate values from Table 4.3. Nordborg (1997, *Genetics*) used a similar approach to obtain the same result. (Note, if we assume $1/N \sim O(\xi^2)$ the result does not exactly match Nordborg (1997). Nordborg argues that $N\mu \gg 1$ and his analysis is using an approximation that is based on the limit as $N \rightarrow \infty$.)

Let us consider [4.12]. The term in parenthesis shows the reduction in coalescent time due to background selection. Remembering that pairwise diversity π is proportional to coalescent time, the effect of background selection at site i on the focal neutral site can be thought of as

$$B_i = \frac{\pi}{\pi_0} = \left(1 - \frac{q_i}{(1 + \rho_i)^2} \right) \quad [4.13]$$

as where $\rho_i = r_i/h_i s_i$ and π_0 is the level of diversity in the absence of background selection.

Using $q = \mu/hs$, we can re-write [4.13] as

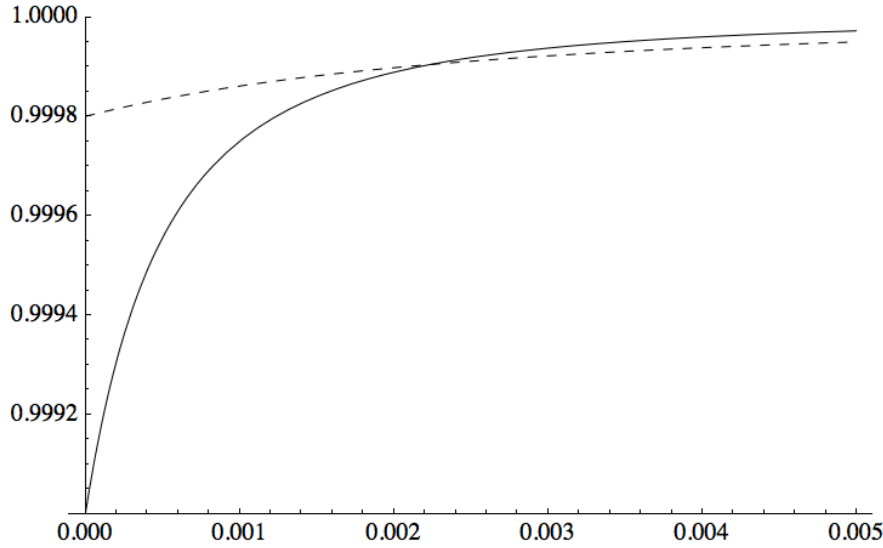
$$B_i = \frac{\pi}{\pi_0} = \left(1 - \frac{q_i}{(1 + \rho_i)^2} \right) = \left(1 - \frac{\mu_i}{h_i s_i (1 + r_i/h_i s_i)^2} \right) \quad [4.14]$$

Based on this result, Nordborg *et al.* 1996 (p. 162, with adjustments to match the notation used here) wrote:

As is intuitively expected, diversity decreases with lower r_i and higher μ_i . The effect of the selection coefficient is slightly more complicated. Differentiating [4.14] with respect to s_i , we see that the effect on diversity increases as s_i decreases, up to a maximum at $h_i s_i = r_i$, and decreases

thereafter. In other words, a weakly selected locus can cause strong background selection if it is tightly linked, but its importance declines rapidly with increasing r_i , whereas a strongly selected locus causes weaker background selection (i.e. greater π/π_0), but can do so from a greater distance.

Below is a plot of B_i as a function of r_i . The solid line is $h_s = 0.001$ and the dashed line is $h_s = 0.005$. (Other parameter: $\mu = 10^{-6}$)



Assuming multiplicative fitness effects and linkage equilibrium, the effects of background selection from multiple sites is be given by the product

$$B = \prod_i \left(1 - \frac{q_i}{(1 + \rho_i)^2}\right) \approx \exp \left[- \sum_i \frac{q_i}{(1 + \rho_i)^2} \right] = \exp \left[- \sum_i \frac{\mu_i}{h_i s_i (1 + \rho_i)^2} \right] \quad [4.15]$$

If we consider mutation only in the region where recombination distance is a linear function of physical distance (double crossovers can be ignored) and make the simplifying assumption that all mutations have the same effect (h_s) and occur at equal rate, then Nordborg *et al* (1996) showed

$$B = e^{-\frac{U}{M(1-hs)}}$$

where U is the total deleterious mutation rate in this region and M is the map length of the region (in Morgans). They pointed out that “the proportional effect of background selection is approximately the same as the density of new mutations per map unit, as pointed out previously by Hudson & Kaplan (1994, 1995) and Barton (1995) for the case of a neutral locus located in the center of a block of selected loci.”

The strength of background selection will vary across the chromosome (due to variation in mutation rate and gene density). Consequently, some parts of a chromosome will experience stronger reductions in diversity (i.e., a smaller N_e than other regions) and, in a metapopulation context, this can lead to elevated F_{ST} (which depends on $N_e m$) in some chromosomal regions compared to others. One

needs to consider this before assuming that regions of high F_{ST} must be due to divergent ecological selection.

Coalescent Simulations

It can be difficult or impossible to analytically calculate the distribution of coalescence times or metrics related to coalescence times (e.g., *Tajima's D*) in many cases. An alternative is using simulations. In principle, we could perform forward simulations to ask how likely a given set of parameters (e.g., N , μ , m , etc) would be to produce the observed results. This would be very computationally intensive because it would require tracking $2N$ alleles over a very long time and most of that information would not be used because the real data is represented by only a few samples ($n \ll N$). Coalescent simulations offer a fast and efficient alternative. In its simplest form, the process works as follows:

- A) Choose the model structure. (Allow for changes in population size? Allow for population subdivision?)
- B) Choose a set of parameter values you want to simulate.
- C) Build a genealogy:
 - a. Start with $k = n$ samples
 - b. Determine the rate of “something” that could affect the genealogy happens (e.g., a coalescence, a migration, etc.). Let's call that rate r_k .
 - c. Draw from an exponential distribution with rate r_k a waiting time for “something” to happen. Record this waiting time.
 - d. Determine which type of events by using probability of event type x relative to the total probability of “something” happening and pick at random which samples were involved. Adjust the system to reflect this event (e.g., one sample migrated, two samples coalesced). If the event was a coalescence, you now have 1 fewer samples (i.e., k goes from n to $n - 1$)
 - e. If $k = 1$, then stop, otherwise repeat b-e.
 - f. From the record of the waiting times and the corresponding events, a genealogy can be constructed
- D) Mutations can be added to the genealogy. For a branch of length t , randomly sample a random number from a Poisson distribution with mean μt (assuming for the sake of explanation that t is measured in generations rather than in units of $2N$ generations). Add that many mutations to that branch.
- E) Calculate your metric of interest (e.g., π , *Tajima's D*, etc.).
- F) Repeat steps C-E numerous times to generate a distribution of your metric of interest under the specified parameters.
- G) You can use this distribution to determine the probability of observing something close to the true value of your observed metric of interest under the specified model.
- H) If you repeat steps A-G for other parameter values (or other model structures), you can then answer whether one model structure or set of parameter values is more likely than others. (Or you can be Bayesian about the whole thing if you like.)

Coalescent Simulations with Recombination (*ancestral recombination graph*)

It is a little less obvious how to incorporate recombination into coalescent simulations. When there is recombination, different sites in a sequence may have different coalescent histories so the coalescent history of the sequence is represented by a complicated graph rather than a single tree, and this is very difficult to deal with analytically. However, if we consider a short enough region such that the rate of

recombination is not too high, then it is still relatively straightforward to perform coalescent simulations as outlined by Hudson (1983). Before thinking about the algorithm, let's think about recombination moving backwards in time. Starting with a single sample, a recombination event splits that sample into two samples, though only part of each new sample contains material from our original sample.

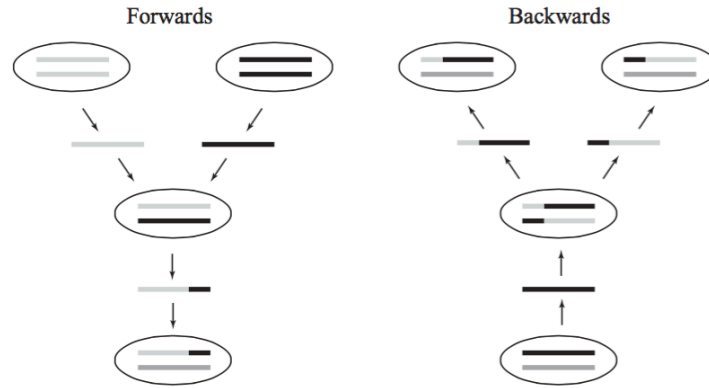


Figure 5.5 Hudson's recombination model on a continuous representation of a sequence. A sequence is made by recombination when an individual creates a haploid genome (sperm cell or egg). Looking forwards, two sequences are recombined into one recombinant sequence. Knowing the allelic states of the grandparent's chromosomes determines one of the child's two chromosomes; the other, the dark grey, originates from the second set of grandparents. Looking backwards, an individual chooses a chromosome from a parent. This chromosome is split onto two grandparental chromosomes. The child's dark grey chromosome is inherited through the other parent and the dark grey chromosomes in grandparents have unknown allelic states.

Copied from Hein, Schierup, and Wiuf. 2004. *Gene Genealogies, Variation and Evolution*.

Starting with a single sample of length L , a recombination event splits that sample into two samples, each of length L , though only part of each new sample contains material from our original sample. We call the parts that were in our original sample as “ancestral material”; that is the material we care about.

Without recombination, when we started with $k = n$ samples, the number of samples never increased; it stayed the same or decreased. With recombination, we can get an increase in the number of samples because every recombination event splits a sample, increasing by one the number of samples we need to track.

As long as recombination is low and N is large, we can safely assume the following: (a) no more than one of our samples is involved in recombination at a time; (b) recombination events and coalescence events do not occur at the same time. With these two assumptions (and other standard assumptions), there are only two possible events that can occur: (i) one of the k samples is split by recombination or (ii) two of the k samples coalesce. (For simplicity, we are ignoring other possible events like migration.) The chance per generation that any particular sample is split by sex is r . Measuring time on the coalescent time scale (in units of $2N$ generations), recombination occurs at rate $2Nr = \rho/2$ per sample (where $\rho = 4Nr$ is the population scaled recombination parameter). While we have k samples, the total rate of recombination for our sample is $k\rho/2$. On this time scale, the total rate of coalescence is $\binom{k}{2} = k(k-1)/2$. Remember, that recombination causes the number of samples (k) to increase whereas coalescence causes the number of samples to decrease. Because the rate of

coalescence is of order k^2 whereas the rate of recombination is of order k , we expect the number of samples to shrink in the long run, i.e., we will eventually find a grand common ancestor for the entire length of the sequence, though it might take a while.

In their book, *Gene Genealogies, Variation and Evolution*, Hein, Schierup, and Wiuf (2004), provide a basic outline for a coalescent simulation algorithm and an illustrated example, which I show below:

1. Start with $k = n$ genes.
2. For k sequences with ancestral material, draw a random number from the exponential distribution with parameter $k(k - 1)/2 + k\rho/2$. This is the time to the next event.
3. With probability $(k - 1)/(k - 1 + \rho)$ the event is a coalescence event, otherwise it is a recombination event.
4. If it is recombination, draw a random sequence and a random point on the sequence. Create an ancestor sequence with the ancestral material to the left of the chosen point and a second ancestor with the ancestral material to the right of the recombination point. Increase the number of ancestral sequences k by one and go to 1.
5. If it is a coalescence event choose two sequences among ancestral sequences at random and merge them into one sequence inheriting the ancestral material to both of the sequences. Decrease k by one. If $k = 1$ end the process, otherwise go to 1.

Q4.18) With respect to Step 2 above, explain why you should use that value as the rate for the exponential distribution.

Q4.19) With respect to Step 3 above, explain why you should use that value as the probability that the event is a coalescence.

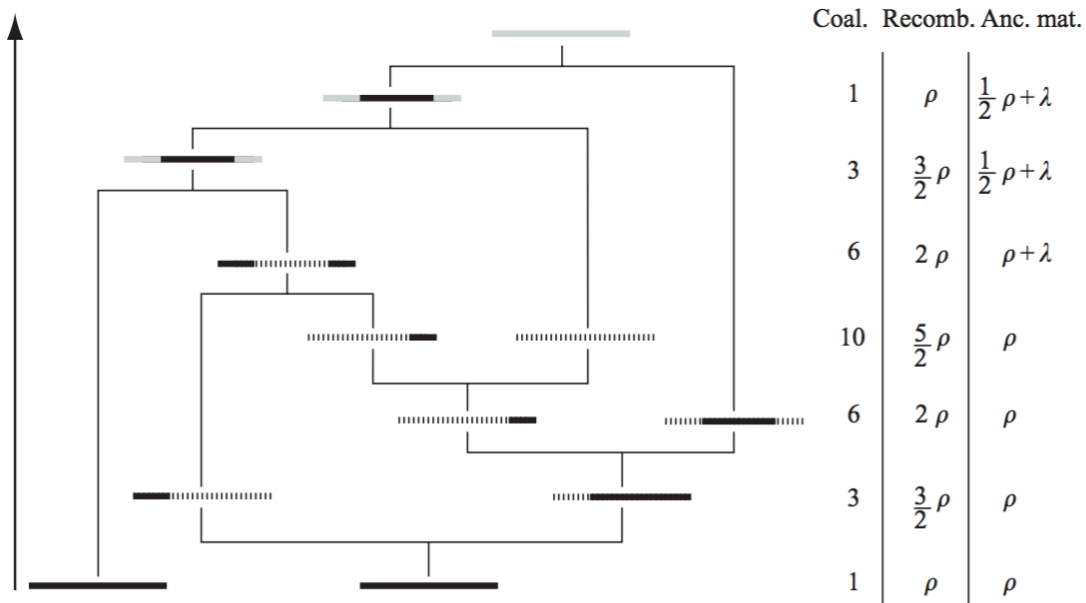


Figure 5.12 Black lines represent the sample sequences or ancestral sequence material to these. Dotted lines represent non-ancestral material. Light grey lines indicate that a MRCA has been found. The non-ancestral piece formed after the first coalescence event between two non-consecutive pieces of ancestral material is trapped material. Also shown is the rate of coalescence and recombination, and the amount of material spanned by ancestral material. λ is the length of the black bar in the sequence with dashed ends.

Hein *et al* (2004) point out that the simple algorithm given above has two features that reduce its efficiency. First, some parts of the sequence may completely coalesce long before the entire sequence does and it is unnecessary to track these parts after they coalesce. Second, recombination could generate a sample that carries no ancestral material (as in the 3rd event back in time of the illustrated example) and it is unnecessary to track such samples. Appropriate adjustments can be made to the simple algorithm to deal with these issues and improve simulation efficiency.

Q4.20) Explain what is happening at each step in the figure above.

Linkage disequilibrium among neutral alleles

McVean (2002 *Genetics*) showed how a coalescent approach could be used to predict the strength of linkage disequilibrium for neutral sites.

Because average disequilibrium will be zero (derived neutral SNPs are not expected to be found together more or less often than expected by chance), we instead focus on the square of disequilibrium. The square of the correlation coefficient is

$$r^2 \equiv \frac{D^2}{f_A(1-f_A)f_B(1-f_B)}$$

where $D = f_{AB} - f_A f_B$. There is no simple analytical expression for the expected value of r^2 . Instead, we consider the related quantity

$$\sigma^2 \equiv \frac{E[D^2]}{E[f_A(1-f_A)f_B(1-f_B)]} \quad [4.16]$$

He showed that this measured of disequilibrium could be expressed in terms of covariances of coalescence times for different sets of sequences

$$\sigma^2 = \frac{\text{Cov}[t_{x(i,j)}, t_{y(i,j)}] - 2\text{Cov}[t_{x(i,j)}, t_{y(i,k)}] + \text{Cov}[t_{x(i,j)}, t_{y(k,l)}]}{E[t_{x(i,j)}]E[t_{y(k,l)}] + \text{Cov}[t_{x(i,j)}, t_{y(k,l)}]}$$

where $t_{x(i,j)}$ is the time to coalescence of site x in sample i and sample j . Qualitatively, σ^2 is large when there is a covariance in coalescence time between two sites (x and y) when both sites come *from the same pair of samples* is strong relative to when both sites are not from the same pair of samples. Recombination puts two sites from the same sample into different samples, so that the coalescent history of that site can then be different from the other. This reduces linkage disequilibrium.

He showed that

$$\sigma^2 = \frac{10+\rho}{22+13\rho+\rho^2} \quad [4.26]$$

where $\rho = 4Nr$. The same result was obtained by Ohta and Kimura (1971) using a very different approach.

This result as presented here ignored two complications. First, as mentioned above, σ^2 approximates $E[r^2]$ only if allele frequencies aren't too extreme. Technically, then our calculation should be conditional on polymorphism being present in the sample at some minimum level. However, McVean (2002) notes that the error introduced by this problem is small. Second, we have done expectations assuming a very large (infinite) sample. McVean notes that "For finite sample size, a modification is required to include the possibility that i, j, k , and l are not all distinct..." (see his paper for the modification) but later says that the modification is negligible for large n (e.g., $n = 50$).

Finally, we note that as ρ becomes large, then [4.26] becomes

$$\sigma^2 \approx \frac{1}{\rho} \quad [4.27]$$

(This final approximation can be obtained by substituting $\rho = 1/z$ into [4.26], then doing a 1st order Taylor series around $z = 0$ and then replacing z with $1/\rho$.)

Remembering that $\rho = 4Nr = 2r/(1/2N)$ we can think of this linkage disequilibrium measure depending on the rate of coalescence relative to the rate recombination.