Recombination

March 23, 2015
1 Two loci

Our previous models focused on a single locus which does not experience recombination. Such is the case with the mitochondrial genome in many animal species, the chloroplast genome in some plant species, and the majority of the Y chromosome in male mammals and the W chromosome in female birds. However, in most species, at least among the eukaryotes, the majority of the nuclear genome does recombine. This has important implications for the population genetics and evolution of such loci, including the fact that recombining loci no longer necessarily share the same genealogy. On the one hand, this makes the analyses much harder; now, in order to understand patterns of variation in a sample of chromosomes, we must cope with multiple genealogies. On the other hand, we can also take advantage of the variation of the genealogy along a chromosome to design better statistical tools for inferring demographic history and selection from sequence data. Furthermore, the interaction between recombination and selection is a particularly important theme in evolutionary biology. Recombination is itself a source of variation that can be acted on by natural selection and recombination between loci targeted by selection can increase the rate of adaptation of a population by reducing competition between beneficial alleles segregating at linked loci.

In this section we will analyze the simplest scenario involving just two loci that recombine at some rate \( r \) per generation. To this end, we consider a generalization of the Wright-Fisher model, which makes the following assumptions:

- Constant population size: \( 2N \) diploid individuals;
- Non-overlapping generations;
- Neutrality;
- Each chromosome in generation \( t + 1 \) is descended from either one or two chromosomes present in the previous generation as follows. With probability \( 1 - r \), both loci are inherited from a single randomly chosen chromosome present in generation \( t \). Otherwise, each locus is inherited from two different chromosomes, which are chosen at random.

Let us label the two loci \( a \) and \( b \). If we restrict attention to a single locus, then this model reduces to the usual Wright-Fisher model without recombination and so all of our previous results apply. In particular, the genealogy at each locus can be described using Kingman’s coalescent. On the other hand, because the two loci can recombine, there is no longer any guarantee that the genealogy at the \( a \) locus is necessarily the same as the genealogy at the \( b \) locus. Suppose that we have sampled two chromosomes at random from the population. To characterize the genealogical history of this sample at both loci, we introduce the following notation. Let us write \( (ab) \) for a chromosome which contains material ancestral to the sample at both loci. Likewise, let us write \( (a-) \) for a chromosome that contains material ancestral to the sample only at the \( a \) locus and \( (-b) \) for a chromosome that contains material ancestral to the sample only at the \( b \) locus. (We could also omit the dashes and just write \( (a) \) and \( (b) \), as Durrett (2008) does, but the notation used here will be somewhat more transparent when we turn to larger numbers of loci.)

If we measure time in units of \( 2N \) generations and we write \( \rho = 4Nr \) for the population recombination rate, then for sufficiently large \( N \) the genealogy of this sample can be described by a continuous-time Markov chain that takes values in the following state space:
Let $J$ be the system of linear equations for the quantities this jump. By conditioning on $x$, $J$, we have sampled two chromosomes, it will be convenient to calculate the covariance for any possible initial state. This will be denoted $v(x)$, where $x$ is the initial state, which will be represented by one of the configurations shown in the table above. If the sample contains only a single ancestral copy of either locus (i.e., $x \in \Delta_1$ or $x \in \Delta_2$), then at least one of the two coalescent times $T_a$ or $T_b$ is equal to 0 and so $v(x) = 0$ as well. On the other hand, if $x$ is equal to $(0,0,2)$, $(1,1,1)$ or $(2,2,0)$, then the sample contains two ancestral copies of both loci. In this case $\text{Var}(T_a) = \text{Var}(T_b) = 1$ and so $v(x)$ is also equal to the correlation between $T_a$ and $T_b$.

**Theorem 1.** Under the two-locus Wright-Fisher model, the covariances between the coalescent times at each locus are

\[
\begin{align*}
v(0,0,2) &= \frac{\rho + 18}{\rho^2 + 13\rho + 18} \quad & v(1,1,1) &= \frac{6}{\rho^2 + 13\rho + 18} \\
v(2,2,0) &= \frac{4}{\rho^2 + 13\rho + 18}.
\end{align*}
\]

**Proof.** Let $u(x) = \mathbb{E}_x[T_aT_b]$ be the expected value of the product of the coalescent times when the initial state is $x$ and observe that for states $x$ with two ancestral copies of each locus we have $v(x) = \text{Cov}(T_a, T_b) = \mathbb{E}_x[T_aT_b] - \mathbb{E}[T_a] \cdot \mathbb{E}[T_b] = u(x) - 1$.

Let $J$ be the time of the first jump and let $X_J$ be the state of the sample immediately following this jump. By conditioning on $J$ and $X_J$ and exploiting the Markov property, we can derive a system of linear equations for the quantities $u(x)$. First observe that

\[
\begin{align*}
\mathbb{E}[T_aT_b|J, X_J] &= \mathbb{E}[(T_a - J + J)(T_b - J + J)|J, X_J] \\
\end{align*}
\]
Write \( n_a(x) \) and \( n_b(x) \) for the number of ancestral copies at the \( a \) or \( b \) locus, respectively, in configuration \( x \). Then, because the genealogy of the sample is governed by a Markov process, we have

\[
\begin{align*}
\mathbb{E}_x[\mathbb{E}((T_a - J)(T_b - J)|J,X_J)] &= \mathbb{E}_x[u(X_J)] \\
\mathbb{E}_x[J \cdot \mathbb{E}[(T_a - J)|J,X_J]] &= \mathbb{E}_x[J] \cdot \mathbb{P}_x(n_a(X_J) = 2) \\
\mathbb{E}_x[J \cdot \mathbb{E}[(T_b - J)|J,X_J]] &= \mathbb{E}_x[J] \cdot \mathbb{P}_x(n_b(X_J) = 2).
\end{align*}
\]

To understand why the second and third identities are true, first recall that for any continuous-time Markov chain, the jump time \( J \) is independent of the new state \( X_J \) that the chain occupies after the jump. Furthermore, since the genealogy at each locus is governed by Kingman’s coalescent, the jump time \( J \) times at each locus when our sample contains two chromosomes. As expected, \( v(0, 0, 2) = 1 \) when \( \rho = 0 \) since \( T_a = T_b \) in that case, while \( v(0, 0, 2) \sim 1/\rho \to 0 \) as \( \rho \to \infty \) since the two genealogies are independent in the infinite recombination limit. Indeed, the correlation is a decreasing function of \( \rho \), reflecting the fact that as the recombination rate increases, the two loci

Similarly, the expectations \( \mathbb{E}_x[u(X_J)] \) can be expressed as linear combinations of terms of the form \( u(y) \) by accounting for all of the possible transitions out of state \( x \) and their transition probabilities:

\[
\begin{align*}
\mathbb{E}_{(0,0,2)}[u(X_J)] &= \frac{\rho}{\rho + 1} u(1,1,1) + \frac{1}{\rho + 1} u(\Delta_2) \\
\mathbb{E}_{(1,1,1)}[u(X_J)] &= \frac{\rho/2}{3 + \rho/2} u(2,2,0) + \frac{1}{3 + \rho/2} u(0,0,2) + \frac{2}{3 + \rho/2} u(\Delta_1) \\
\mathbb{E}_{(2,2,0)}[u(X_J)] &= \frac{4}{6} u(1,1,1) + \frac{2}{3} u(\Delta_1).
\end{align*}
\]

As explained above, \( u(\Delta_1) = u(\Delta_2) = 0 \). Combining these results leads to the following system of three equations in three unknowns:

\[
\begin{align*}
u(0,0,2) &= \frac{\rho}{\rho + 1} u(1,1,1) + \frac{2}{\rho + 1} \\
u(1,1,1) &= \frac{2}{\rho + 6} u(0,0,2) + \frac{\rho}{\rho + 6} u(0,0,2) + \frac{4}{\rho + 6} \\
u(2,2,0) &= \frac{2}{3} u(1,1,1) + \frac{1}{3}.
\end{align*}
\]

Solving these and then using the identity \( v(x) = u(x) - 1 \) leads to the expressions that appear in the statement of the theorem.

The quantity that is of interest to us is \( v(0,0,2) \), which is the correlation between the coalescent times at each locus when our sample contains two chromosomes. As expected, \( v(0,0,2) = 1 \) when \( \rho = 0 \) since \( T_a = T_b \) in that case, while \( v(0,0,2) \sim 1/\rho \to 0 \) as \( \rho \to \infty \) since the two genealogies are independent in the infinite recombination limit. Indeed, the correlation is a decreasing function of \( \rho \), reflecting the fact that as the recombination rate increases, the two loci
are increasingly likely to have distinct genealogies. Furthermore, the actual correlation depends only on the composite parameter $\rho = 4Nr$, which quantifies the relative rates of recombination ($r$) and coalescence ($1/2N$). Thus, when $r$ is much larger than $1/2N$, recombination will dominate coalescence and the correlation between the two loci will be weak.

The preceding result can be generalized to larger samples. Suppose that $n_a = i + k$ chromosomes are sampled at locus $a$ and $n_b = j + k$ chromosomes are sampled at locus $b$ and that exactly $k$ of these are common to both samples. Let $\tau_a$ and $\tau_b$ be the total branch lengths of the genealogies at locus $a$ and $b$, respectively. If we define $F(i, j, k) = \text{cov}(\tau_a, \tau_b)$ to be the covariance of the coalescent times at each locus when the initial configuration of the sample is $x = (i, j, k)$, then using arguments similar to those given in the proof of Theorem 1, we can show that the quantities $F(x)$ satisfy a recursive system of equations.

**Theorem 2.** Suppose that $l$ chromosomes with initial configuration $x = (i, j, k)$ are sampled from a population governed by the two-locus Wright-Fisher model and let $X$ be the state of the sample after the first event in the genealogy. Then

$$F(x) = \mathbb{E}_x[F(X)] + \frac{2k(k-1)}{\beta_x(n_a-1)(n_b-1)},$$

where

$$\beta_x = \left(\frac{1}{2}\right) + \frac{kp}{2}$$

is the total rate of change to a sample with configuration $x$.

If we let $p(x, y)$ denote the transition probability from configuration $x$ to $y$ under the genealogical process, then we can rewrite the above identity as

$$F(x) = \sum_y p(x, y)F(y) + c(x),$$

where the sum is over all states $y$ that can be reached from $x$ in a single event and $c(x)$ is the constant on the right-hand side of the identity that appears in the theorem. Since neither recombination nor coalescence will increase the number of chromosomes ancestral to a locus, while coalescent events can decrease this number by one at one or both loci, we can solve this system recursively by using the results for configurations $y$ with $n_a(y) + n_b(y) < n$ to solve for $F(x)$ for configurations $x$ with $n_a(x) + n_b(x) = n$. In general, this must be done numerically except when $n$ is very small.

This last result can also be used to calculate the variance of the number of segregating sites, $S_n$, in a sample of $n$ chromosomes from a population governed by the infinite sites model with recombination between loci. More formally, we consider a sequence of Wright-Fisher models with $m$ loci arranged linearly along a chromosome, where recombination occurs between adjacent loci at rate $r/(m-1)$ per generation and each locus mutates at rate $\mu/m$ per generation. By identifying each chromosome with the unit interval $[0, 1]$, one can show that in the limit as $m \to \infty$ these processes converge to a limit in which mutations occur at rate $\mu$ per generation according to the infinite sites model, while recombinations occur at rate $r$ per generation at uniformly-distributed positions along $[0, 1]$. Let us call this limiting model the infinite sites model with recombination. Because recombination does not affect the marginal distribution of the genealogy at an individual site, the mean number of segregating sites does not depend on $\rho$ and so $\mathbb{E}[S_n] = h_n\theta$, where $h_n = \sum_{i=1}^{n-1} 1/i$, as previously shown. In contrast, because the variance of $S_n$ depends on the covariance of the genealogies at different sites, it does depend on the recombination rate.
Theorem 3. Under the infinite sites model with recombination,

\[ \text{Var}(S_n) = \theta h_n + \frac{\theta^2}{4} \int_0^1 2(1 - y)f_n(\rho y)dy \]

where \( f_n(x) \) is the covariance between the total time in the genealogies of two loci that recombine at (scaled) rate \( x \).

Proof. We prove the result by first considering the approximating model with \( m \) loci and then taking the limit as \( m \to \infty \). If \( S'_n \) to be the number of segregating sites at the \( j \)'th locus in this model, then since \( S_n = S'_1 + \cdots + S'_m \), we have

\[ \text{Var}(S_n) = \sum_{i=1}^m \text{Var}(S'_i) + \sum_{1 \leq i \neq j \leq m} \text{cov}(S'_i, S'_j). \]

Using our results for the infinite sites model without recombination, we know that

\[ \text{Var}(S'_i) = \frac{\theta}{m} \sum_{j=1}^{n-1} \frac{1}{j} + \left( \frac{\theta}{m} \right)^2 \sum_{j=1}^{n-1} \frac{1}{j^2}. \]

If \( \tau_i \) is the total branch length in the genealogy at the \( i \)'th locus, then conditional on \( \tau_i \), \( S'_i \) is Poisson-distributed with mean \((\theta/2m)\tau_i\). Furthermore, \( S'_i \) and \( S'_j \) are conditionally independent given \( \tau_i \) and \( \tau_j \) and so

\[ \text{cov}(S'_i, S'_j) = \left( \frac{\theta}{2m} \right)^2 \text{cov}(\tau_i, \tau_j). \]

Notice that the scaled recombination rate between locus \( i \) and locus \( j \) is \( 4Nr(j - i)/(m - 1) \) and that there are \( m - k \) pairs of loci between 1 and \( m \) at this distance from each other. It follows that

\[ \text{Var}(S_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \frac{\theta^2}{m} \sum_{j=1}^{n-1} \frac{1}{j^2} + \frac{\theta^2}{4m^2} \sum_{k=1}^{m-1} 2(m - k) f_n \left( \frac{k \rho}{m - 1} \right). \]

The main result follows by letting \( m \to \infty \) and observing that the second term on the right-hand side tends to 0, while the third term is a Riemann sum for the desired integral. \( \square \)

If \( \rho = 0 \), then \( f_n(0) = \text{Var}(\tau_n) = 4g_n \) is constant and the result reduces to the one previously obtained when studying the infinite sites model without recombination: \( \text{Var}(S_n) = \theta h_n + \theta^2 g_n \). On the other hand, since \( f_n(\rho y) \to 0 \) for every \( y > 0 \) as \( \rho \to \infty \), it follows that in this limit we have \( \text{Var}(S_n) = \theta h_n \). Indeed, in this limit, every site effectively has its own, independently-determined genealogy and so the contribution to the variance of \( S_n \) from the randomness in the genealogy is averaged out. Furthermore, since \( f_n(\rho y) \) is a decreasing function of \( \rho \) for every \( y > 0 \), we also see that the variance of \( S_n \) is decreasing in \( \rho \).

2 Linkage Disequilibrium

Suppose that a population is variable at two loci and that alleles \( A \) and \( a \) are segregating at the first locus while alleles \( B \) and \( b \) are segregating at the second locus. Let \( p_{AB} \) denote the frequency of the two-locus haplotype \( AB \) in the population and let \( p_A = 1 - p_a \) and \( p_B = 1 - p_b \) denote the marginal frequencies of the alleles at each locus. If the loci are unlinked and the population is governed by the Wright-Fisher model, then the two loci will evolve independently and so we expect, on average, for the frequency of each two-locus haplotype to be equal to
the product of the marginal frequencies of the alleles that it contains, e.g., \( p_{AB} = p_{APB} \). The \textbf{linkage disequilibrium} of \( A \) and \( B \) is a statistic that measures how far the population is from this expectation:

\[
D_{AB} = p_{AB} - p_{APB}.
\]

In fact, even under the idealized conditions just described, we would not expect \( D_{AB} \) to be exactly equal to 0 because genetic drift causes the allele and haplotype frequencies to fluctuate at random, so that \( D_{AB} \) itself is a random variable with a distribution that depends on the population size.

Now suppose that the two loci are linked, with recombination rate \( r \) per generation. If the population is infinite, then we can ignore genetic drift and derive a recursive formula showing how the linkage disequilibrium changes from generation to generation. In the absence of genetic drift, the allele frequencies are constant and so we have \( p_A^t = p_A \) and \( p_B^t = p_B \) for all \( t \geq 0 \). In contrast, recombination causes the haplotype frequencies to change over time, e.g.,

\[
p_{AB}^{t+1} = (1 - r)p_{AB}^t + rp_{APB}.
\]

Subtracting \( p_{APB} \) from both sides gives

\[
D_{AB}^{t+1} = p_{AB}^{t+1} - p_{APB} = (1 - r)p_{AB}^t - (1 - r)p_{APB} = (1 - r)D_{AB}^t,
\]

which shows that the linkage disequilibrium in an infinite population decays geometrically at rate \((1 - r)\), i.e., \( D_{AB}^{t+1} = (1 - r)^t D_{AB}^0 \). In particular, as long as the recombination rate \( r \) is positive, \( D_{AB}^t \) will tend to 0 as \( t \to \infty \). In reality, populations are finite and so the actual distribution of linkage disequilibrium between two linked loci depends on both the recombination rate and the population size, with smaller populations typically exhibiting larger values of \(|D_{AB}|\) for any given recombination rate.

Although we could also calculate the linkage disequilibria \( D_{ab}, D_{A B} \) and \( D_{a B} \), all of these quantities are related in a simple way. Indeed, noting that \( p_A = p_{AB} + p_{Ab} \) and \( p_B = p_{AB} + p_{aB} \), and that \( p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1 \), we can rewrite \( D_{AB} \) as follows

\[
D_{AB} = p_{AB} - p_{APB} = p_{AB}(p_{AB} + p_{Ab} + p_{aB} + p_{ab}) - (p_{AB} + p_{Ab})(p_{AB} + p_{aB}) = p_{AB}p_{ab} - p_{Ab}p_{aB}.
\]

This shows that \( D_{AB} = D_{ab} = -D_{Ab} = -D_{aB} \), which in turn implies that the sign of \( D_{AB} \) is determined by our usually arbitrary choice of the alleles \( A \) and \( B \). In addition, the range of possible values of \( D_{AB} \) is determined by the marginal allele frequencies. Indeed, given \( p_A \leq p_B \), the largest possible value of \( D_{AB} \) will be realized when all of the \( A \) alleles are contained within \( AB \) haplotypes, in which case \( p_{AB} = p_A \) so that

\[
D_{AB} = p_A - p_{APB} = p_A(1 - p_B) = p_{APB}.
\]

Similarly, the smallest possible value of \( D_{AB} \) will be realized when the frequency of \( AB \) is equal to zero, in which case

\[
D_{AB} = -p_{APB}.
\]

This implies that the linkage disequilibrium between even two tightly linked loci may be small if the marginal allele frequencies are close to 0 or 1. For this reason, several alternative measures of linkage disequilibrium have been proposed that control for the marginal allele frequencies. For example, if we define

\[
D' = \begin{cases} 
\frac{D_{AB}}{\min\{p_{APB}, p_{AB}\}} & \text{if } D_{AB} > 0 \\
\frac{D_{AB}}{\min\{p_{APB}, p_{aB}\}} & \text{if } D_{AB} < 0 
\end{cases}
\]
then it follows from the bounds given above that \(-1 \leq D' \leq 1\) for all marginal allele frequencies. Another measure that is widely used is

\[
r^2 = \frac{D_{AB}^2}{p_{AB}p_Bp_B},
\]

which can be interpreted as the square of the correlation coefficient of the two indicator variables \(1_A\) and \(1_B\). Indeed, \(1_A\) is a Bernoulli random variable with mean \(p_A\) and variance \(p_A(1 - p_A) = p_{AB}\), while the covariance of \(1_A\) and \(1_B\) is

\[
\text{cov}(1_A, 1_B) = \mathbb{E}[1_A1_B] - \mathbb{E}[1_A]\mathbb{E}[1_B] = p_{AB} - p_{AP}p_B = D_{AB},
\]

since \(1_A1_B = 1\) if and only if a randomly chosen chromosome carries the haplotype \(AB\). Recalling that the correlation coefficient between any pair of random variables takes values between \(-1\) and 1, it follows that \(r^2\) takes values between 0 and 1 and is equal to 1 when the population only contains the genotypes \(AB\) and \(ab\) (so-called coupling gametes) or the genotypes \(Ab\) and \(aB\) (the so-called repulsion gametes).

It is sometimes of interest to compare the observed linkage disequilibrium between two loci with the expected disequilibrium assuming a null model such as the infinite sites model with recombination. Excess linkage disequilibrium can arise because of population bottlenecks, population structure or selection. Nowadays this would be done through coalescent simulations, but some analytical results are available that illustrate the competing effects of recombination and genetic drift. Because the expected value of a ratio such as \(r^2\) is very difficult to calculate, we will instead calculate the ratio of the expected values of the numerator and denominator in \(r^2\). The following result is due to Ohta and Kimura (1971).

**Theorem 4.** Under the infinite sites model, the linkage disequilibrium between a pair of segregating sites separated by a genetic distance of \(\rho = 4Nr\) is approximately

\[
\sigma_d^2 = \frac{\mathbb{E}[D_{AB}^2]}{\mathbb{E}[p_{AP}p_Bp_B]} = \frac{10 + \rho}{22 + 13\rho + \rho^2}.
\]

**Proof.** Since \(D_{AB} = D_{ab} = -D_{Ab} = -D_aB\), it follows that \(D_{AB}^2 = D_{ab}^2 = D_{Ab}^2 = D_{aB}^2\) and so \(\sigma_d^2\) does not depend on which alleles are designated \(A\) and \(B\). In particular, we can assume that \(A\) and \(B\) are the derived alleles at each locus, in which case

\[
\mathbb{E}[D_{AB}^2] = \mathbb{E}[(p_{AB} - p_{APB})^2] \\
= \mathbb{E}[p_{AB}^2] - 2\mathbb{E}[p_{AB}p_{APB}] + \mathbb{E}[p_{APB}^2] \\
= F_{ij,ij}^* - 2F_{ij,ik}^* + F_{ij,kl}^*.
\]

where \(F_{ij,kl}^*\) is the conditional probability that sequences \(i\) and \(j\) both carry the derived allele at the first locus and that sequences \(k\) and \(l\) both carry the derived allele at the second locus when \(i, j, k\) and \(l\) are sampled at random (with replacement) and we are given that both sites are segregating.

To calculate the probabilities \(F_{ij,kl}^*\), let \(l^h_i\) denote the branch length at locus \(h\) from the most recent common ancestor of sequences \(i\) and \(j\) to the most recent common ancestor of the entire population and let \(u^h\) denote the total branch length of the genealogy of the entire population at this locus, both measured in units of \(2N\) generations. If the mutation rate at each locus is \(v\) mutations per \(2N\) generations, then because the number of mutations along a portion of the tree of branch length \(L\) is Poisson-distributed with mean \(uL\), we have

\[
F_{ij,kl}^* = \frac{\mathbb{E}[u^1_i e^{-u^1_{ij}} \cdot u^2_j e^{-u^2_{ij}}]}{\mathbb{E}[u^1 e^{-u^1} \cdot u^2 e^{-u^2}]}.
\]
Notice that the denominator is the probability that the two sites are segregating, while the numerator is the probability that the mutations were transmitted to the sampled sequences. Under the infinite site model, the mutation rate at any particular site is 0 and so we take the limit as $u \to 0$ to obtain
\[
F^*_{ij,kl} = \frac{\mathbb{E}[I_{ij}^t \cdot I_{kl}^2]}{\mathbb{E} \left[ \tau^1 \cdot \tau^2 \right]}
\]

Let $t^h_{ij}$ be the time to the most recent common ancestor of sequences $i$ and $j$ at locus $h$ and let $T^h$ be the time to the most recent common ancestor of the entire population at the same locus. Since $I_{ij}^t = T^h - t^h_{ij}$ and the expectations $\mathbb{E}[T^1h_{ij}^2]$ do not depend on $i$ and $j$, it follows that
\[
\mathbb{E} \left[ I_{ij}^1 I_{ij}^2 \right] - 2 \mathbb{E} \left[ I_{ij}^1 I_{ik}^2 \right] + \mathbb{E} \left[ I_{ij}^2 I_{kl}^2 \right]
= (1 - 2 + 1) \mathbb{E} \left[ T^1 T^2 \right] - (1 - 2 + 1) \mathbb{E} \left[ T^1 t_{ij}^1 + T^2 t_{ij}^2 \right] + \mathbb{E} \left[ t_{ij}^1 t_{ij}^2 \right] - 2 \mathbb{E} \left[ t_{ij}^1 t_{ik}^2 \right] + \mathbb{E} \left[ t_{ij}^2 t_{kl}^2 \right]
= (\mathbb{E} \left[ t_{ij}^1 t_{ij}^2 \right] - 1) - 2 (\mathbb{E} \left[ t_{ij}^1 t_{ik}^2 \right] - 1) + (\mathbb{E} \left[ t_{ij}^1 t_{kl}^2 \right] - 1)
= \text{cov}(t_{ij}^1, t_{ij}^2) - 2 \text{cov}(t_{ij}^1, t_{ik}^2) + \text{cov}(t_{ij}^1, t_{kl}^2).
\]

Here we have used the fact that $\mathbb{E}[t_{ij}^1] = 1$ when time is measured in units of $2N$ generations. It follows that the second moment (and variance) of $D_{AB}$ is equal to
\[
\mathbb{E}[D_{AB}^2] = \frac{\text{cov}(t_{ij}^1, t_{ij}^2) - 2 \text{cov}(t_{ij}^1, t_{ik}^2) + \text{cov}(t_{ij}^1, t_{kl}^2)}{\mathbb{E}[\tau^1 \cdot \tau^2]}.
\]

To calculate the denominator of $\sigma^2_d$, observe that $\mathbb{E}[p_{ApaBPB}]$ is the probability that in a sample of four sequences $i$, $j$, $k$ and $l$, $i$ and $k$ carry the derived alleles at locus 1 and 2, respectively, while $j$ and $l$ carry the ancestral alleles at these loci. Arguing as before, this can be shown to be equal to
\[
\mathbb{E}[p_{ApaBPB}] = \lim_{u \to 0} \frac{\mathbb{E}[ut_{ij}^1 e^{-ut_{ij}^1} \cdot ut_{kl}^2 e^{-ut_{kl}^2}]}{\mathbb{E}[ut^1 e^{-u \tau^1} \cdot ut^2 e^{-u \tau^2}]} = \frac{\mathbb{E}[t_{ij}^1 \cdot t_{kl}^2]}{\mathbb{E}[\tau^1 \cdot \tau^2]} - \frac{\text{cov}(t_{ij}^1, t_{kl}^2) + 1}{\mathbb{E}[\tau^1 \cdot \tau^2]}.
\]

Taking the ratio of these two terms, the denominators cancel and we are left with
\[
\sigma^2_d = \frac{\text{cov}(t_{ij}^1, t_{ij}^2) - 2 \text{cov}(t_{ij}^1, t_{ik}^2) + \text{cov}(t_{ij}^1, t_{kl}^2)}{\text{cov}(t_{ij}^1, t_{kl}^2) + 1}.
\]

The covariances in this expression were evaluated in Theorem 1:
\[
\text{cov}(t_{ij}^1, t_{ij}^2) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}
\]
\[
\text{cov}(t_{ij}^1, t_{ik}^2) = \frac{6}{\rho^2 + 13\rho + 18}
\]
\[
\text{cov}(t_{ij}^1, t_{kl}^2) = \frac{4}{\rho^2 + 13\rho + 18}.
\]

Substituting these identities into the previous expression then leads to the formula given in the statement of this theorem.

Taking $\rho = 0$ in the above expression shows that the expected value of $r^2$ between a pair of completely linked sites is approximately $\sigma^2_d = 5/11 \approx 0.455$. That this is less than 1 reflects the operation of genetic drift. Furthermore, $\sigma^2_d$ decreases like $\rho^{-1}$ as $\rho \to \infty$. 
\[\square\]
3 The Ancestral Recombination Graph

3.1 Basics

In section 1 we formulated a Markov process which describes the genealogical relationships of a sample of chromosomes at two recombining loci. Our goal in this section is to extend this formulation to entire chromosomes. To this end, we will represent the chromosome by the interval \([0, 1]\) and we will assume that the recombination rate between sites is constant across the chromosome, so that the probability that a recombination event occurs within an interval \([a, b] \subset [0, 1]\) is proportional to the length of the interval. Suppose that the population evolves according to the Wright-Fisher model with recombination, i.e., each chromosome is either copied from a single chromosome chosen uniformly at random from the previous generation (probability \(1 - r\)) or it is the product of recombination between two chromosomes chosen uniformly at random from the previous generation (probability \(r\)). In the latter case, the breakpoint is uniformly distributed on \([0, 1]\) and all of the sites to the left of the breakpoint are inherited from one of the two recombining chromosomes and all of the sites to the right are inherited from the other. We will define \(\rho = 4Nr\) to be the scaled recombination rate between the two ends of the chromosome.

Now suppose that we have sampled \(n\) chromosomes at random from a population governed by this model. If the recombination rate \(r\) is positive, then the genealogy of the sample will typically change as we move from one end of the chromosome to the other. However, the collection of genealogies indexed by position can be described by a Markov process known as the ancestral recombination graph or ARG for short. Here we will measure time in units of \(2N\) generations and we will assume that the population size \(N\) is large enough that we can approximate the correct discrete-time process by a continuous-time process. Each branch in the ARG will correspond to a series of chromosomes that form a single lineage over the entire length of the chromosome. Looking backwards in time, the ARG is shaped by two kinds of events. First, each pair of branches coalesces at rate \(\frac{1}{2}\) so that the total rate of coalescence when there are \(k\) branches in the graph is \(\frac{k(k - 1)}{2}\), as in Kingman’s coalescent. In this case, coalescence will occur along the entire length of the chromosome and the number of branches in the graph will decrease from \(k\) to \(k - 1\).

Secondly, at rate \(\frac{\rho}{2}\), each branch can experience a recombination event, in which case it splits into two new branches. A breakpoint is chosen uniformly at random from \([0, 1]\) and sites to the left of the breakpoint are inherited from one of the two newly created branches, while sites to the right of the breakpoint are inherited from the other branch. The total rate of recombination when there are \(k\) branches in the graph is \(kp\) and the number of branches increases by 1 following each recombination event.

If we let \(Y_t\) denote the number of branches in the ARG at time \(t\), then \(Y = (Y_t : t \geq 0)\) is itself a continuous-time Markov chain. In fact, because each event affecting the ARG causes the number of branches to either increase or decrease by 1, \(Y\) is a continuous-time birth-death process with birth rate \(\lambda_k = \frac{k}{2}\) and death rate \(\mu_k = \frac{kp}{2}\) when \(Y_t = k\). Although there is no upper bound to the number of branches that can be contained in the ARG at any given time, the fact that the death rate is quadratic in \(k\) while the birth rate is linear in \(k\) can be used to show that this number is almost surely finite. In particular, we can use the properties of birth-death processes to show that with probability 1 there will be a time \(\tau_n < \infty\) when the ARG contains only one branch. Necessarily the sample of \(n\) chromosomes will have reached the most recent common ancestor at all sites by this time. The next theorem shows how the expected value of this time depends on the population recombination rate \(\rho\).

**Theorem 5.** Let \(Y = (Y_t : t \geq 0)\) be the branch counting process corresponding to the ARG and let \(\tau_n = \inf\{t : Y_t = 1\}\) be the first time when the ARG for a sample of \(n\) chromosomes contains
a single branch. Then
\[
\mathbb{E}[	au_n] = \frac{2}{\rho} \int_0^1 \left( \frac{1-v^n-1}{1-v} \right) \left( e^{\rho(1-v)} - 1 \right) dv.
\]

**Proof.** By conditioning on the first event in the birth-death process and invoking the Markov property, the expectations \( T_n = \mathbb{E}[\tau_n] \) satisfy the following two-term recursion

\[
T_n = \frac{2}{n(\rho + n - 1)} + \frac{n-1}{\rho + n - 1} T_{n-1} + \frac{\rho}{\rho + n - 1} T_{n+1}.
\]

Letting \( \Delta_n = T_n - T_{n-1} \), this can be rearranged into a one-term recursion

\[
\Delta_{n+1} = \frac{n-1}{\rho} \Delta_n - \frac{2}{n} \rho.
\]

To solve this recursion, we will make the following ansatz: assuming that each term \( \Delta_n \) is a smooth function of \( \rho \), we can expand \( \Delta_n \) in a power series in \( \rho \)

\[
\Delta_n = \sum_{k=0}^{\infty} c_{n,k} \rho^k.
\]

When \( \rho = 0 \), the ARG reduces to Kingman’s coalescent and so \( T_n = 2(1 - 1/n) \) is just the mean time to the most recent common ancestor of a sample of \( n \) chromosomes. This shows that

\[
c_{n,0} = \Delta_n(\rho = 0) = \frac{2}{n(n - 1)}.
\]

To solve for the higher-order coefficients, we substitute the ansatz into the one-term recursion, which gives the identity

\[
\sum_{k=0}^{\infty} c_{n+1,k} \rho^k = -\frac{2}{n\rho} + \sum_{k=0}^{\infty} (n-1)c_{n,k} \rho^{k-1} = \sum_{k=0}^{\infty} (n-1)c_{n,k+1} \rho^k,
\]

where we have used the formula for \( c_{n,0} \) to simplify the expression on the right-hand side. For these two power series to be pointwise equal, their coefficients must also be equal, which leads to the following recursion

\[
c_{n,k+1} = \frac{1}{n-1} c_{n+1,k}.
\]

If we apply this recursion repeatedly to the coefficient on the right-hand side, then we will eventually arrive at an identity involving \( c_{n+k+1,0} \), which we know from above. This shows that

\[
c_{n,k} = \frac{2(n-2)!}{(n+k)!}
\]

for \( n \geq 2 \) and \( k \geq 0 \) and so

\[
\Delta_n = 2 \sum_{k=0}^{\infty} \frac{(n-2)!}{(n+k)!} \rho^k.
\]

To solve for \( T_n \), we use the fact that \( \Delta_2 = T_2 - T_1 = T_2 \) (since \( T_1 = 0 \)), giving

\[
T_n = \sum_{m=2}^{n} \Delta_m = 2 \sum_{m=2}^{n} \sum_{k=0}^{\infty} \frac{(m-2)!}{(m+k)!} \rho^k.
\]
To show that this expression is equal to the integral that appears in the statement of the theorem observe that
\[
\frac{2}{\rho} \int_0^1 \left(1 - v^{n-1}\right) \left(e^{\rho(1-v)} - 1\right) dv = \frac{2}{\rho} \int_0^1 \left(\sum_{m=0}^{n-2} y^m\right) \left(\sum_{k=1}^\infty \frac{\rho^k (1-y)^k}{k!}\right) dy
\]
\[
= \frac{2}{\rho} \sum_{m=0}^{n-2} \sum_{k=1}^\infty \frac{\rho^k}{k!} \int_0^1 y^m (1-y)^k dy
\]
\[
= \frac{2}{\rho} \sum_{m=0}^{n-2} \sum_{k=1}^\infty \frac{\rho^k}{k!} \frac{m!k!}{(m+k+1)!}
\]
\[
= \frac{2}{\rho} \sum_{m=2}^n \sum_{k=0}^\infty \frac{(m-2)!}{(m+k)!} \rho^k.
\]
\[
\square
\]

Taking \( n = 2 \) in Theorem 5 shows that
\[
E[\tau_2] = \frac{2}{\rho} \int_0^1 \left(e^{\rho(1-v)} - 1\right) dv = \frac{2}{\rho^2} (e^\rho - 1 - \rho),
\]
which grows exponentially in \( \rho \). This shows that the waiting time until the ARG contains just a single branch is typically very large when the recombination rate is high. In fact, the situation is not quite as bad as this computation suggests, because we need only run the ARG until every site in the chromosome has reached its most recent common ancestor. This will certainly have happened by the time that the ARG contains a single branch, but often it will happen well before that time. Additional savings in computational time can be found by avoiding recombination events that create new branches that contain no sites that are ancestral to the sample. For example, if recombination causes one of the sampled lineages to split into two new branches that contain ancestral material in the intervals \([0, x]\) and \([x, 1]\), respectively, then we can ignore recombination events occurring at breakpoints greater than \( x \) in the first branch and less than \( x \) in the second branch since either of these would create a superfluous branch. Note, however, that we cannot disregard all recombination events taking place within non-ancestral material. For example, if a branch contains ancestral material in the disjoint intervals \([0, 0.2]\) and \([0.7, 1]\), as could happen if two recombinant branches coalesce, then we cannot disregard a recombination event that has a breakpoint in the interval \((0.2, 0.7)\), because this will create two new branches that both contain ancestral material.

3.2 The Spatial Markov Coalescent

As suggested by Theorem 5, the computational burden of simulating the ARG over long DNA sequences can be very high. For this reason, several authors have formulated processes that approximate the ARG but which are much less computationally demanding. One of these, the so-called spatial Markov coalescent or SMC, works by approximating the sequence of genealogies \( A = (T_x : 0 \leq x \leq 1) \) along the chromosome. To explain how the SMC arises and in what sense it is an approximation to the full ARG, we first describe an alternative approach to simulating the ARG due to Wiuf and Hein.

These authors proposed an algorithm which simulates the ARG by starting at one end of the chromosome, say at \( x = 0 \), and moving progressively to the other end. Recalling that the marginal distribution of the genealogy in the ARG at any particular site \( x \) is governed by Kingman’s coalescent, we can initiate the process by simulating the tree at \( x = 0 \) according to ordinary
coalescent. Let this tree be denoted \( T_0 \) and let \( \tau_0 \) be the total branch length in \( T_0 \). Since recombination events occur at constant rate \( \rho/2 \) per branch in the ARG, it can be shown that the left-most breakpoint will occur at a position \( x_1 \) which is exponentially distributed with rate \( \tau_0 \rho/2 \). In this case \( T_0 \) will be the genealogy at all positions \( x \in [0, x_1) \), whereas the genealogy at \( x_1 \) may change to a new tree \( T_1 \). To simulate \( T_1 \), let \( U_0 \) be uniformly distributed on \( T_0 \), i.e., first choose a branch in \( T_0 \) with probability proportional to its length and then choose a position uniformly at random along this branch. \( U_0 \) determines both the branch in \( T_0 \) affected by the first recombination event as well as the time when that event occurs. At that time, the recombination event creates a new branch, carrying material ancestral to the sample to the right of the breakpoint \( x_1 \), which can experience one of the following three fates: (1) it coalesces with the branch affected by the recombination event, in which case \( T_1 = T_0 \); (2) it coalesces with another branch in \( T_0 \), in which case \( T_1 \) coincides with \( T_0 \) everywhere except along the new lineage connecting \( U_0 \) back to \( T_0 \); or (3) it coalesces with the lineage ancestral to the most recent common ancestor of \( T_0 \) with similar consequences for \( T_1 \) as in case (2).

The next breakpoint \( x_2 \) and tree \( T_2 \) are determined in a similar fashion, with one new complication. Let \( \tau_1 \) be the total branch length of \( T_1 \) and define \( x_2 \) by specifying that the difference \( x_2 - x_1 \) is exponentially distributed with rate \( \tau_1 \rho/2 \). Then \( T_1 \) is the genealogy at all positions \( x \in [x_1, x_2) \), while the genealogy at position \( x_2 \) may change to a new tree \( T_2 \). To determine the new tree, we again choose a point \( U_1 \) uniformly at randomly along \( T_1 \) and create a new branch carrying material ancestral to the sample to the right of the breakpoint \( x_2 \). This branch can then either re-coalesce to \( T_1 \) or, if \( T_1 \) and \( T_0 \) are different, it can coalesce with any branches contained in \( T_0 \) that are not also contained in \( T_1 \).

In general, given trees \( T_0, \ldots, T_n \) corresponding to segments \( [0, x_1), \ldots, [x_{n-1}, x_n) \), the new breakpoint \( x_{n+1} \) and the new tree \( T_{n+1} \) are determined as above. This is done until the right endpoint of the interval \( [0, 1] \) is reached, in which case every site will have been assigned to a tree. What makes this process difficult to simulate and to analyze is the fact that the new tree \( T_{n+1} \) depends on all of the trees \( T_0, \ldots, T_n \) corresponding to sites to the left of \( x_{n+1} \). In particular, because of this long-range dependence, it follows that the process \( A = (T_x : 0 \leq x \leq 1) \) does not satisfy the Markov property, since conditioning on the tree at any site \( x \) does not remove the dependence between the trees to the right and to the left of \( x \).

The spatial Markov coalescent is obtained by suppressing the long-range dependence in the preceding construction. In the version formulated by McVean and Cardin, the tree \( T_{n+1} \) is generated by first choosing a point \( U_{n+1} \) uniformly at random from \( T_n \) and replacing the lineage below \( U_{n+1} \) by a new lineage which is allowed to re-coalesce with \( T_n \). This differs from the full ARG in two respects: (1) the branch experiencing the recombination event is erased below \( U_{n+1} \) so that it cannot coalesce with the newly-created branch, and (2) the newly-created branch can only re-coalesce with \( T_n \) and not with any of the trees to the left of \( T_n \). With these modifications, it is clear that the SMC is indeed a Markov process and it can be shown that it has the correct marginal distributions, i.e., the genealogy at any single position \( x \) is governed by Kingman’s coalescent. On the other hand, the joint distributions of the SMC differ from those of the ARG. To illustrate this discrepancy, suppose that \( n = 2 \) and consider the trees \( T_0 \) and \( T_1 \) at positions 0 and 1. In this case, the recombination rate between the two positions is \( \rho \) and the covariance between the coalescent times \( \tau_0 \) and \( \tau_1 \) at these positions is given by Theorem 1

\[
\text{cov} (\tau_0, \tau_1) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}.
\]

In contrast, if the SMC is restricted to two loci, the tree \( T'_0 \) at position \( x = 0 \) is simulated first using Kingman’s coalescent. Then, with probability \( 1/(1 + \rho) \), no recombination occurs, in which case the tree \( T'_1 \) at position \( x = 1 \) coincides with \( T'_0 \). Otherwise, a recombination does occur between the two positions and the tree \( T'_1 \) is generated by an independent simulation of
Kingman’s coalescent. In this case the covariance between the coalescent times $\tau'_0$ and $\tau'_1$ is
\[
\text{cov}(\tau'_0, \tau'_1) = \frac{1}{1+\rho},
\]
which is strictly less than $\text{cov}(\tau_0, \tau_1)$. In other words, the SMC underestimates the correlation between trees at different sites, which is not surprising given the nature of its construction. Furthermore, since the ratio of the two covariances is
\[
\frac{\text{cov}(\tau'_0, \tau'_1)}{\text{cov}(\tau_0, \tau_1)} = \frac{\rho^2 + 13\rho + 18}{\rho^2 + 18\rho + 18},
\]
the relative error is greatest for intermediate values of $\rho$ and decays to 0 both when $\rho = 0$ and when $\rho \to \infty$.

4 Detecting and Quantifying Recombination

A variety of methods have been proposed to detect evidence of recombination in sequence data. At one end of the spectrum is a class of methods which attempt to give a lower bound on the number of recombination events that are needed for the data to be consistent with the infinite sites model. Our starting point is the so-called four gametes test. Suppose that we have sequenced a sample of $n$ chromosomes at two biallelic loci and let 0 and 1 denote the ancestral and derived alleles at each locus, respectively. Initially the population will consist entirely of 00 haplotypes, but following the first mutation, which we will assume occurs at the left-hand locus, it will contain a mixture of 00 and 10 haplotypes. Since each locus can mutate at most once under the infinite sites model, only the right-hand locus can now mutate and this can happen either to a 00 chromosome or to a 10 mutation. In the first case, the population will consist of a mixture of 00, 10 and 01 haplotypes, while in the second case it will consist of a mixture of 00, 10 and 11 haplotypes. However, in no case will the fourth haplotype be generated unless there is either a recombination event or an additional mutation. To see how recombination can generate the fourth haplotype, observe that a recombination event between haplotypes 10 and 01 can generate both the 00 and 11 haplotypes, while a recombination event between haplotypes 00 and 11 can generate both the 10 and 01 haplotypes. It follows that if we can detect a pair of segregating sites in a sequence alignment where all four haplotypes are present, then either we must reject the infinite sites model or we can infer that a recombination event occurred at a position somewhere between the two sites.

By applying the four-gametes test repeatedly to a sequence alignment, we can find a lower bound on the number of recombinations that have affected the sample. The following algorithm is due to Hudson & Kaplan (1985):

1. For each pair of segregating sites, $i$ and $j$, set $d(i, j) = 1$ if all four haplotypes are present in the sample and $d(i, j) = 0$ otherwise. Let $\mathcal{I}$ be the set of all intervals $(i, j)$ with $d(i, j) = 1$.
2. Remove all intervals $(m, n) \in \mathcal{I}$ that contain another interval $(i, j) \in \mathcal{I}$.
3. Let $(i_1, j_1)$ be the left-most interval in $\mathcal{I}$. If $(m, n) \in \mathcal{I}$ overlaps $(i_1, j_1)$, then remove $(m, n)$. Repeat until there are no more overlapping intervals.

When completed, this algorithm will generate a collection of non-overlapping intervals that each must contain at least one recombination event by the four-gametes test. The number of such intervals is therefore the minimum number of recombination events that must be invoked if the infinite sites model is not rejected. In fact, we can improve upon this algorithm by using the
following observation. If \( b_{i,j}^{(0)} \) is a lower bound on the number of recombination events in the interval \((i,j)\), then a possibly better bound can be found by setting

\[
\hat{b}_{i,j}^{(1)} = \max_{i<k<j} \left( b_{i,k}^{(0)} + b_{k,j}^{(0)} \right).
\]

By recursively applying this rule, we will generate a non-decreasing sequence of lower bounds \( b_{i,j}^{(n)} \). Since \( b_{i,i+1}^{(n)} \leq 1 \) for all \( n \), this sequence will necessarily converge, giving a lower bound \( \hat{b}_{0,L} \) for the entire region \( 0,L \). See Myers & Griffiths (2003) and Bafna & Bansal (2006) for implementations of this and related algorithms.

Simulations of the ARG have shown that the various estimators of the minimum number of recombinations tend to greatly underestimate the actual number of recombinations, often by a factor of 5-10. For this reason, a variety of other methods have been investigated which attempt to estimate the recombination rate \( \rho \) itself. These methods fall into two classes, depending on whether the estimators are based on the likelihood function of the full data or are calculated from summary statistics. For example, several authors have published software packages that use MCMC to sample ARG’s with probabilities conditional on the full sequence data. These can then be used to estimate either the full likelihood function or the posterior density of \( \rho \) given an alignment. In principle, because these estimates use all of the information contained in the data, they should be superior to those based on summary statistics. Unfortunately, in practice, this is often not the case due to the astronomical size of the state space of the ARG when there are more than a few sequences containing more than a few segregating sites. Noting that there are \( (2n-3)!! = (2n-3) \cdot (2n-1) \cdot \ldots \cdot 1 \) binary rooted trees with \( n \) leaves, the number of possible ARGs with at most \( k \) breakpoints is equal to this number raised to the \( k \) power. This is problematic even for MCMC-based algorithms, which may fail to converge.

One alternative to the full likelihood-based approach is to calculate the composite likelihood of the data using pairs of segregating sites. Suppose that we have sampled \( n \) chromosomes, which contain \( S \) segregating sites. Assuming that each such site is segregating two alleles, say 0 and 1, then by restricting attention to a pair of such sites, \( i \) and \( j \), the data can be summarized by four numbers, \( n_{00} + n_{01} + n_{10} + n_{11} = n \), which we will represent by a vector \( \mathbf{X}_{ij} \). Let \( q_c(\mathbf{X}_{ij}, \rho, \theta) \) be the probability of observing the two-site data when the scaled recombination rate between the two sites is equal to \( \rho \) and the scaled mutation rate at each site is equal to \( \theta \). Assuming that we can calculate this probability, the composite likelihood of the full alignment \( \mathbf{X} \) is defined to be the product of all of the two-site likelihoods

\[
L_c(\mathbf{X}, \rho, \theta) = \prod_{i<j} q_c(\mathbf{X}_{ij}, d_{ij}\rho_b, \theta),
\]

where \( \rho_b \) is the scaled recombination rate between neighboring sites and \( d_{ij} \) is the physical distance along the chromosome between the \( i \)'th and the \( j \)'th segregating sites. (This assumes that the genetic distance between two sites scales linearly with the physical distance.) Point estimates of the population recombination rate and the population mutation rate can then be derived by finding the values of \( \rho_b \) and \( \theta \) that maximize the composite likelihood.

The composite likelihood differs from the full likelihood in two respects. First, although every pair of segregating sites contributes to the composite likelihood, \( L_c \) does not contain information about higher-order associations between segregating sites. Consequently, the composite likelihood ignores some of the information content of the data. Secondly, by multiplying the pairwise likelihoods, we are effectively assuming that the sampling distributions at distinct pairs of sites are independent. While this may be approximately true when the pairs are non-overlapping and separated by large genetic distances, it will certainly not be true of overlapping or nested pairs of sites or even of non-overlapping pairs of sites that are tightly linked. However, by assuming independence in these cases, we are inflating the apparent information content of the data,
which causes the composite likelihood function to be much more peaked about its mode than is the true likelihood function. For this reason, we cannot use the curvature of the composite likelihood function to estimate the precision of the composite likelihood estimates. On the other hand, Fearnhead (2003) proved that the composite likelihood estimates are consistent for any fixed sample size \( n \geq 2 \) in the sense that as the sequence length increases to infinity, the composite likelihood estimates converge in distribution to the true values.

Although the two-locus probabilities \( q_c \) can be shown to satisfy a recursive system of linear equations, the number of sample configurations is so large that it is not practical to solve this system except for small \( n \). Indeed, for fixed \( n \), the number of such configurations is equal to the number of integer vectors \((n_{00}, n_{10}, n_{01}, n_{11})\) with non-negative components that sum to \( n \). Since this is also equal to the number of integer vectors with positive components that sum to \( n + 4 \), this number is

\[
\binom{n+3}{3} = \frac{1}{6}n(n-1)(n-2).
\]

For example, for \( n = 20 \), there are 1140 such configurations, while for \( n = 50 \), there are 19600 configurations. For this reason, Monte Carlo simulations are used to estimate the \( q_c \). The most straightforward approach is to simulate the two-locus ARG. If we carry out \( K \) simulations, generating realizations \( x_1, \cdots, x_K \), then the empirical frequency of configuration \( x \) is an unbiased estimator of \( q_c(x, \rho, \theta) \):

\[
\hat{q}_c(x, \rho, \theta) \equiv \frac{1}{K} \sum_{i=1}^{K} 1_{x}(x_i).
\]

Notice that these simulations must be carried out for multiple values of \( \rho \) and \( \theta \). While this approach is easy to implement, it is also very computationally intensive even for modest values of \( n \) due to the large number of configurations that have small, but non-negligible probabilities. Indeed, if a configuration has probability \( p \), then on average we will have to perform \( 1/p \) simulations before we obtain a realization of the ARG that has that configuration.

An alternative approach, which has been implemented by McVean et al. (2004), uses importance sampling to generate a biased sample of realizations of the ARG with greater representation of the observed configurations. Suppose that \( X \) is a random variable with density \( p \) and let \( q \) be another density on the state space of \( X \) such that \( q(x) > 0 \) for any \( x \) with \( p(x) > 0 \). If this condition is satisfied, we say that \( q \) is absolutely continuous with respect to \( p \). Our goal is to estimate the expectation \( \mathbb{E}[f(x)] \), where \( f \) is a real-valued function. In ordinary Monte Carlo simulation, we would generate \( K \) independent realizations of \( X \), say \( x_1, \cdots, x_K \), and use these to calculate the empirical average

\[
\frac{1}{K} \sum_{i=1}^{K} f(x_i).
\]

In importance sampling, we instead generate \( K \) independent realizations of the distribution \( q \), say \( y_1, \cdots, y_n \), and use these to calculate the weighted empirical average

\[
\frac{1}{K} \sum_{i=1}^{K} \frac{p(y_i)}{q(y_i)} \cdot f(y_i).
\]

In this expression, the weights \( p(y)/q(y) \) compensate for the fact that the sampled values come from distribution \( q \) rather than \( p \). This also is an unbiased estimate of \( \mathbb{E}[f(x)] \) and, depending on the choice of \( q \), it may lead to a much more efficient algorithm in the sense that fewer simulations are required to achieve any given precision of the estimate.
It can be shown that the optimal proposal distribution for estimating the two-locus likelihood $q_c(x, \rho, \theta)$ is the conditional distribution of the ARG given that the sample configuration is $x$. This is optimal in the sense that, for a fixed number of simulations, this proposal distribution will minimize the variance of the IS estimate. Furthermore, by conditioning on the configuration, we ensure that every realization can be used in the estimate. Unfortunately, the conditional distribution is unknown. However, McVean and colleagues showed how to construct an approximation to this distribution which is easy to implement and which has good statistical properties. By applying this method to genome-wide sequence data from human HapMap project, they were able to estimate local recombination rates across the human genome. This work revealed the existence of numerous recombination hotspots in which the local recombination rate is elevated by up to two orders of magnitude above the genome-wide average.