3

Recombination

> "In mathematics you don't understand things. You get used to them."
> John von Neumann

In this chapter, we will begin to investigate the effects of recombination on the patterns of genetic variability.

3.1 Two loci

Going forward in time, the dynamics of a Wright-Fisher model with two loci may be described as follows. To generate an individual in the next generation, with probability $1-r$ we copy both loci from one randomly chosen individual, while with probability r a recombination occurs and we copy the two loci from two randomly chosen individuals. Reversing our perspective, suppose we sample m individuals from the population. Each locus will have its own genetic history leading to a coalescent, but the two coalescent processes will be correlated, due to the fact that in the absence of recombination the two loci will be copied from the same parent.

3.1.1 Sample of size 2

We call the two loci the a locus and the b locus. Ignoring the numbering of the copies, using parentheses to indicate the loci at which the individual has genetic material ancestral to the sample, the possible states of the system are given in the first column of the next table. The second column describes the state by giving the number of (a)'s, (b)'s, and (ab)'s. The third gives the number of a's, n_a, and the number of b's, n_b,

	$(a),(b),(ab)$	n_a, n_b	
$(a\,b)\,(a\,b)$	$(0,0,2)$	$2,2$	
$(a\,b)\,(a)\,(b)$	$(1,1,1)$	$2,2$	
$(a)\,(a)\,(b)\,(b)$	$(2,2,0)$	$2,2$	
$(a)\,(b)\,(b)$	$(1,2,0)$	$1,2$	Δ_1
$(b)\,(a)\,(a)$	$(2,1,0)$	$2,1$	Δ_1
$(a\,b)\,(b)$	$(0,1,1)$	$1,2$	Δ_1
$(a\,b)\,(a)$	$(1,0,1)$	$2,1$	Δ_1
$(a\,b)$	$(0,0,1)$	$1,1$	Δ_2
$(a)\,(b)$	$(1,1,0)$	$1,1$	Δ_2

To explain the notation, the initial state is $(ab)(ab)$, i.e., we have sampled two individuals and examined the states of their two loci. If the first event going backwards in time is a coalescence, the chain enters (ab) and the process stops before it gets interesting. If the first event is a recombination, the two copies in one individual become separated and the state is $(ab)(a)(b)$. At this point, a second recombination might produce $(a)(b)(a)(b)$, or a coalescence might produce $(ab)(b)$, $(ab)(a)$ or return us to the original $(ab)(ab)$.

The next figure shows a possible realization. The a locus is denoted by a solid dot and the b locus by an open circle. At the bottom, the first individual is on the left and the second on the right. On the right edge of the figure, we have indicated the sequence of states. The parts are sometimes listed in a different order from the table of states to make it easier to connect the state with the picture.

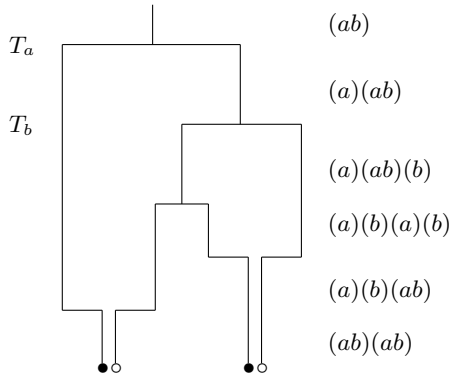

Fig. 3.1. Realization of the two-locus coalescent with recombination.

Let T_a be the coalescence time for the a locus and T_b be the coalescence time for the b locus. Our next task is to compute $E_x(T_aT_b)$ for the possible

initial states x. We will use the notation in the table above. If $n_a = 1$ or $n_b = 1$, then $T_a T_b = 0$, so we will use Δ_1 to denote the set of configurations where one coalescence has occurred, and Δ_2 the configurations where both have. Considering the Wright-Fisher model, speeding up time by a factor of $2N$, and following the standard practice of letting $\rho = 4Nr$, the rates become

from/to	$(0,0,2)$	$(1,1,1)$	$(2,2,0)$	Δ_1	Δ_2	total
$(0,0,2)$	.	ρ	.	.	1	$\rho+1$
$(1,1,1)$	1	.	$\rho/2$	2	.	$(\rho/2)+3$
$(2,2,0)$	.	4	.	2	.	6

In the first row, $(0,0,2) \to (1,1,1)$, i.e., $(ab),(ab) \to (a),(b),(ab)$ by two possible recombinations, or $(0,0,2) \to \Delta_2$ by coalescence. In the second row, if the (a) and (b) coalesce, $(1,1,1) \to (0,0,2)$, while the other two possible coalescence events lead to Δ_1, and if recombination happens to (ab), $(1,1,1) \to (2,2,0)$. When the state is $(2,0,0) = (a),(a),(b),(b)$, four possible coalescence events lead back to $(1,1,1)$, while two lead to Δ_1.

Theorem 3.1. *Let $v(x)$ be $cov(T_a, T_b)$ when the initial state is x.*

$$v(0,0,2) = \frac{\rho+18}{\rho^2+13\rho+18} \quad v(1,1,1) = \frac{6}{\rho^2+13\rho+18}$$
$$v(2,2,0) = \frac{4}{\rho^2+13\rho+18} \tag{3.1}$$

These are sometimes referred to as $C_{ij,ij}$, $C_{ij,ik}$, $C_{ij,k\ell}$ with i,j,k,ℓ distinct numbers that indicate the individuals sampled at the two locations.

Proof. Since $ET_a = ET_b = 1$, we have

$$\operatorname{cov}(T_a, T_b) = ET_aT_b - ET_aET_b = ET_aT_b - 1$$

Let $u(x) = E_x(T_aT_b)$. To get an equation for $u(x)$, let J be the time of the first jump, and let X_J be the state at time J:

$$\begin{aligned} E(T_aT_b|J,X_J) &= E((T_a - J + J)(T_b - J + J)|J,X_J) \\ &= E((T_a - J)(T_b - J)|J,X_J) + JE(T_a - J|J,X_J) \\ &\quad + JE(T_b - J|J,X_J) + J^2 \end{aligned}$$

Let $v_a(X_J)$ and $v_b(X_J)$ be 1 if coalescence has not occurred at the indicated locus and 0 if it has. Noting that J is independent of X_J and taking the expected value gives

$$u(x) = E_x u(X_J) + E_x J \cdot (E_x v_a(X_J) + E_x v_b(X_J)) + E_x J^2$$

If J is exponential with rate λ, then $EJ = 1/\lambda$ and $EJ^2 = 2/\lambda^2$. Using this with our table of rates gives

$$u(0,0,2) = \frac{\rho}{\rho+1}\left(u(1,1,1) + 2\cdot\frac{1}{\rho+1}\right) + \frac{1}{\rho+1}\left(u(\Delta_2) + 0\cdot\frac{1}{\rho+1}\right) + \frac{2}{(\rho+1)^2}$$

since the probabilities of the two transitions are their rate over the total. Here and in the next two equations $E_x J^2$ is last and does not depend on the state X_J the chain jumps to. Similar reasoning gives

$$\begin{aligned} u(1,1,1) &= \frac{(\rho/2)}{(\rho/2)+3}\left(u(2,2,0) + 2\cdot\frac{1}{(\rho/2)+3}\right) \\ &\quad + \frac{1}{(\rho/2)+3}\left(u(0,0,2) + 2\cdot\frac{1}{(\rho/2)+3}\right) \\ &\quad + \frac{2}{(\rho/2)+3}\left(u(\Delta_1) + 1\cdot\frac{1}{(\rho/2)+3}\right) + \frac{2}{((\rho/2)+3)^2} \\ u(2,2,0) &= \frac{4}{6}\left(u(1,1,1) + 2\cdot\frac{1}{6}\right) + \frac{2}{6}\left(u(\Delta_1) + 1\cdot\frac{1}{6}\right) + \frac{2}{6^2} \end{aligned}$$

Simplifying and using $u(\Delta_i) = 0$ gives

$$\begin{aligned} u(0,0,2) &= \frac{\rho}{\rho+1}u(1,1,1) + \frac{2}{\rho+1} \\ u(1,1,1) &= \frac{1}{(\rho/2)+3}u(0,0,2) + \frac{(\rho/2)}{(\rho/2)+3}u(2,2,0) + \frac{2}{(\rho/2)+3} \\ u(2,2,0) &= \frac{2}{3}u(1,1,1) + \frac{1}{3} \end{aligned}$$

$u(x) = E_x(T_a T_b)$. What we want to compute is the covariance $v(x) = u(x) - 1$, which satisfies the following equations (substitute $u(x) = v(x) + 1$):

$$\begin{aligned} v(0,0,2) &= \frac{\rho}{\rho+1}v(1,1,1) + \frac{1}{\rho+1} \\ v(1,1,1) &= \frac{1}{(\rho/2)+3}v(0,0,2) + \frac{(\rho/2)}{(\rho/2)+3}v(2,2,0) \\ v(2,2,0) &= 2v(1,1,1)/3 \end{aligned} \tag{3.2}$$

Rearranging the second equation, and then using the third equation,

$$\begin{aligned} v(0,0,2) &= \left(\frac{\rho}{2}+3\right)v(1,1,1) - \frac{\rho}{2}v(2,2,0) \\ &= \left(\frac{\rho}{2}+3-\frac{\rho}{3}\right)v(1,1,1) = \frac{\rho+18}{6}v(1,1,1) \end{aligned}$$

Using this in the first equation, we have

$$v(0,0,2) = \frac{\rho}{\rho+1}\cdot\frac{6}{\rho+18}v(0,0,2) + \frac{1}{\rho+1}$$

Multiplying each side by $\rho+1$, we have

$$1=\left(\rho+1-\frac{6\rho}{\rho+18}\right)v(0,0,2)=\frac{\rho^2+13\rho+18}{\rho+18}v(0,0,2)$$

which gives the first formula. The first equation in (3.2) implies

$$\begin{aligned}v(1,1,1)&=\frac{(\rho+1)v(0,0,2)-1}{\rho}\\&=\frac{1}{\rho}\left(\frac{\rho^2+19\rho+18}{\rho^2+13\rho+18}-1\right)=\frac{6}{\rho^2+13\rho+18}\end{aligned}$$

The final result follows from $v(2,2,0)=2v(1,1,1)/3$. □

3.1.2 Sample of size n

To generalize the previous argument to larger sample sizes, we give up on trying to solve the equations to get formulas and content ourselves with a recursion that can be solved numerically. Let τ_a and τ_b be the total size of the genealogical trees for the two loci. Kaplan and Hudson (1985, see pages 386–387), wrote equations for $E(\tau_a\tau_b)$. However, Pluzhnikov and Donnelly (1996, see page 1260) noticed that the recursion for the covariance is much nicer. Suppose that $n_a=i+k$ chromosomes are sampled at locus a and $n_b=j+k$ chromosomes are sampled at locus b, in such a way that k chromosomes are common to the two samples, and let $F(i,j,k)=\operatorname{cov}(\tau_a,\tau_b)$ when the initial configuration is (i,j,k). Let $\ell=i+j+k$. The configuration $x=(i,j,k)$ changes at rate

$$\beta_x=\frac{\ell(\ell-1)+\rho k}{2}$$

due to coalescence of two of the ℓ chromosomes or recombination separating the a and b loci on one of the k chromosomes with both.

Theorem 3.2. *If $x=(i,j,k)$ and X is the state after the first jump, then*

$$F(x)=E_xF(X)+\frac{2k(k-1)}{\beta_x(n_a-1)(n_b-1)}$$

Proof. This clever proof is from Tavaré's St. Flour notes. We start by deriving the formula for conditional covariances. Since $E\{E(Y|X)\}=EY$, we have

$$\begin{aligned}\operatorname{cov}(\tau_a,\tau_b)&=E(\tau_a\tau_b)-E(\tau_a)E(\tau_b)\\&=E\{E(\tau_a\tau_b|X)\}-E\{E(\tau_a|X)E(\tau_b|X)\}\\&\qquad+E\{E(\tau_a|X)E(\tau_b|X)\}-E(\tau_a)E(\tau_b)\\&=E\{\operatorname{cov}(\tau_a,\tau_b|X)\}+\operatorname{cov}(E(\tau_a|X),E(\tau_b|X))\end{aligned}\tag{3.3}$$

Letting J be the time of the first jump, we can write $\tau_a = n_a J + \tau'_a$ and $\tau_b = n_b J + \tau'_b$, where the waiting times after J, τ'_a and τ'_b, and the state X are independent of J. From this we get

$$\begin{aligned} E\{\operatorname{cov}(\tau_a, \tau_b|X)\} &= n_a n_b \operatorname{var}(J) + E\{\operatorname{cov}(\tau'_a, \tau'_b|X)\} \\ &= (n_a n_b)/\beta^2 + E_x F(X) \end{aligned} \tag{3.4}$$

where for simplicity we have dropped the subscript x on β. To evaluate the second term in (3.3), we recall that $E(E(\tau_a|X)) = E\tau_a$, so

$$\operatorname{cov}(E(\tau_a|X), E(\tau_b|X)) = E\{(E(\tau_a|X) - E\tau_a)) \cdot (E(\tau_b|X) - E\tau_b))\}$$

Let N_a and N_b be the number of a and b lineages after the jump. If $h(m) = 2\sum_{j=2}^{m} 1/j$, then

$$E(\tau_c|X) - E\tau_c = \frac{n_c}{\beta} + h(N_c) - h(n_c) \tag{3.5}$$

To compute the expected value of the product, we have to look in detail at the transition rates:

(i,j,k) to	at rate	N_a	N_b
$(i+1, j+1, k-1)$	$r_1 = \rho k/2$	n_a	n_b
$(i-1, j-1, k+1)$	$r_2 = ij$	n_a	n_b
$(i-1, j, k)$	$r_3 = ik + i(i-1)/2$	$n_a - 1$	n_b
$(i, j-1, k)$	$r_4 = jk + j(j-1)/2$	n_a	$n_b - 1$
$(i, j, k-1)$	$r_5 = k(k-1)/2$	$n_a - 1$	$n_b - 1$

From this it follows that

$$Eh(N_a) - h(n_a) = -\frac{2}{n_a - 1} \cdot \frac{r_3 + r_5}{\beta} \qquad Eh(N_b) - h(n_b) = -\frac{2}{n_b - 1} \cdot \frac{r_4 + r_5}{\beta}$$

Hence, using (3.5),

$$\begin{aligned} &E\{(E(\tau_a|X) - E\tau_a)) \cdot (E(\tau_b|X) - E\tau_b))\} \\ &\quad = \frac{n_a n_b}{\beta^2} - \frac{n_b}{\beta} \cdot \frac{2}{n_a - 1} \frac{r_3 + r_5}{\beta} - \frac{n_a}{\beta} \cdot \frac{2}{n_b - 1} \frac{r_4 + r_5}{\beta} \\ &\qquad + \frac{4}{(n_a - 1)(n_b - 1)} \cdot \frac{r_5}{\beta} \end{aligned}$$

To simplify, notice that $E(E(\tau_c|X) - E\tau_c) = 0$ and (3.5) imply

$$\frac{n_a}{\beta} = \frac{2}{n_a - 1} \frac{r_3 + r_5}{\beta} \qquad \frac{n_b}{\beta} = \frac{2}{n_b - 1} \frac{r_4 + r_5}{\beta}$$

so the second and third terms from the previous formula are equal to -1 times the first and we have

$$\operatorname{cov}(E(\tau_a|X), E(\tau_b|X)) = -\frac{n_a n_b}{\beta^2} + \frac{2k(k-1)}{\beta(n_a - 1)(n_b - 1)}$$

Adding this to (3.4) gives the desired result. □

Samples of size 2

To make the connection between the new formula and the old one, recall that the jump rate in state $x = (i, j, k)$ is $\beta_x = [\ell(\ell - 1) + \rho k]/2$, where $\ell = i + j + k$, so we have

state	ℓ	β_x
$(0, 0, 2)$	2	$1 + \rho$
$(1, 1, 1)$	3	$3 + (\rho/2)$
$(2, 2, 0)$	4	6

Since $F(i, j, k) = 0$ for all the other states, and $2k(k - 1) = 0$ unless $k = 2$, consulting the table of rates before Theorem 3.1, and using Theorem 3.2 gives

$$
\begin{aligned}
F(0,0,2) &= \frac{\rho}{\rho+1}F(1,1,1) + \frac{4}{\rho+1} \\
F(1,1,1) &= \frac{1}{(\rho/2)+3}F(0,0,2) + \frac{(\rho/2)}{(\rho/2)+3}F(2,2,0) \\
F(2,2,0) &= 2F(1,1,1)/3
\end{aligned}
$$

and we have $F(i, j, k) = 4v(i, j, k)$. The factor of 4 comes from the fact that we are considering the total size of the tree, which for a sample of size 2 is 2 times the coalescence time, so the covariance is four times as large.

Samples of size 3

To explain how the equations in Theorem 3.2 can be solved, consider the problem of computing $F(0, 0, 3)$. To do this we have to consider the other configurations with $n_a = 3$ and $n_b = 3$: (1,1,2), (2,2,1), and (3,3,0); the configurations with $n_a = 2$ and $n_b = 3$: (0,1,2), (0,2,1), and (2,3,0); and of course the configurations with $n_a = 3$ and $n_b = 2$. We have already computed the values for $n_a = 2$, $n_b = 2$. Those with $n_a = 1$ or $n_b = 1$ are 0.

To compute the values for the states with $n_a = 2$ and $n_b = 3$, we begin by identifying the transition rates:

n_a, n_b	to/from	(0,1,2)	(1,2,1)	(2,3,0)
2,3	(0,1,2)		2	
2,3	(1,2,1)	ρ		6
2,3	(2,3,0)		$\rho/2$	
1,3	(0,2,1)		1+0	
1,3	(1,3,0)			0+1
2,2	(0,0,2)	2+0		
2,2	(1,1,1)		2+1	
2,2	(2,2,0)			0+3
1,2	(0,1,1)	1		
	total $= \beta_x$	$\rho + 3$	$(\rho/2) + 6$	10
	$g(x)$	$2/\beta_x$	0	0

The final row gives the value of $g(x) = 2k(k-1)/[\beta_x(n_a-1)(n_a-2)]$. It is comforting to note that the total rate in each case $\beta_x = [\ell(\ell-1)+\rho k]/2$, where $\ell = i+j+k$. Using the notation from the proof of Theorem 3.2, the entries in the table with ρ's come from r_1. The numbers on the first two rows come from r_2. The entries with plus signs come from r_3 or r_4. The remaining rate $(0,1,2) \to (0,1,1)$ in the lower left is an r_5.

Using the values we have computed for $n_a = 2$, $n_b = 2$, and recalling that when $n_a = 1$, $F = 0$, we get three equations in three unknowns. To make the equations easier to write, we let $b = \rho+3$, $c = (\rho/2)+6$, and $d = \rho^2+13\rho+18$.

$$
\begin{aligned}
F(0,1,2) &= \frac{\rho}{b}F(1,2,1) + \frac{2}{b}\cdot\frac{\rho+18}{d} + \frac{2}{b}\\
F(1,2,1) &= \frac{2}{c}F(0,1,2) + \frac{\rho/2}{c}F(2,3,0) + \frac{3}{c}\cdot\frac{6}{d}\\
F(2,3,0) &= \frac{6}{10}F(1,2,1) + \frac{3}{10}\cdot\frac{4}{d}
\end{aligned}
$$

In matrix form, they become

$$
\begin{pmatrix} 1 & -\frac{\rho}{b} & 0 \\ -\frac{2}{c} & 1 & -\frac{\rho}{2c} \\ 0 & -\frac{6}{10} & 1 \end{pmatrix}
\begin{pmatrix} F(0,1,2) \\ F(1,2,1) \\ F(2,3,0) \end{pmatrix}
=
\begin{pmatrix} 2(\rho+18+d)/bd \\ 18/cd \\ 12/10d \end{pmatrix}
$$

The equations can be solved by row reducing the matrix to upper triangular form, but the answer is not very pretty.

For a sample of size n there are $(n-1)^2$ values of $2 \le n_a, n_b \le n$, which one must tackle by considering all the $n_a + n_b = m$ for $m = 4, 5, \ldots 2n$. For a given value of n_a, n_b we have a system of $\min\{n_a, n_b\} + 1$ equations to solve. Each system can be solved in at most $O(n^2)$ operations, so of order n^4 computations are needed, which is feasible for samples of size 100. One could, as Ethier and Griffiths (1990) suggested, write $O(n^3)$ equations for $F(i,j,k)$ with $2 \le n_a, n_b \le n$, but then solving the matrix equations would take n^6 steps.

3.2 m loci

Our task in this section is to generalize the two-locus results to m linearly arranged loci, each of which follows the infinite sites model, and to let $m \to \infty$ to get a model of a segment of DNA where recombination can occur between any two adjacent nucleotides. The number of mutations per generation per locus is assumed to have a Poisson distribution with mean u/m. Recombination does not occur within subloci, but occurs between adjacent subloci at rate $r/(m-1)$ per generation. With this assumption, the recombination rate between the most distant subloci is r.

3.2.1 Samples of size 2

Let S_2 be the number of segregating sites in a sample of size n and let $f_2(x) = (x+18)/(x^2+13x+18)$ be the covariance of the coalescence time for samples of size 2 at two loci with scaled recombination rate x between them.

Theorem 3.3. *For the infinite sites model with recombination*

$$var\,(S_2) = \theta + \theta^2 \int_0^1 2(1-y) f_2(y\rho)\, dy \tag{3.6}$$

If $\rho = 0$, $f_2(0) = 1$ and this reduces to (1.22): $\text{var}\,(S_2) = \theta + \theta^2$.

Proof. Let S_2^j be the number of segregating sites in the jth locus in a sample of two alleles.

$$\text{var}\,(S_2) = \sum_{i=1}^{m} \text{var}\,(S_2^j) + \sum_{1\le i\ne j\le m} \text{cov}\,(S_2^i, S_2^j)$$

If we let $\theta = 4Nu$, then it follows from (1.22) that

$$\text{var}\,(S_2^j) = \frac{\theta}{m} + \left(\frac{\theta}{m}\right)^2$$

Let T_2^i be the coalescence time of the two copies of locus i. This distribution of S_2^i given T_2^i is Poisson with mean $(\theta/m)T_2^i$, so $E(S_2^i|T_2^i) = (\theta/m)T_2^i$. The numbers of segregating sites S_2^i, S_2^j are conditionally independent given T_2^i and T_2^j, so

$$E(S_2^i S_2^j | T_2^i, T_2^j) = \left(\frac{\theta}{m}\right)^2 T_2^i\, T_2^j$$

and $\text{cov}\,(S_2^i, S_2^j) = (\theta/m)^2\,\text{cov}\,(T_2^i, T_2^j)$. Using (3.1), we see that the variance of the total number of segregating sites is

$$\text{var}\,(S_2) = \theta + \frac{\theta^2}{m} + \frac{\theta^2}{m^2}\sum_{k=1}^{m-1} 2(m-k) f_2\left(\frac{k\rho}{m-1}\right)$$

where $\rho = 4Nr$, since there are $2(m-k)$ pairs $1 \le i, j \le n$ with $|i-j| = k$. Letting $m \to \infty$, setting $y = k/m$, and noting that the sum approximates an integral gives the indicated result. □

3.2.2 Samples of size n

Let S_n be the number of segregating sites in a sample of size n, recall $h_n = \sum_{i=1}^{n-1} 1/i$, and let $f_n(x)$ be the covariance between the total time in the genealogical trees for two loci with scaled recombination rate x between them. This can be computed numerically using Theorem 3.2.

Theorem 3.4. *For the infinite sites model with recombination,*

$$var\,(S_n) = \theta h_n + \frac{\theta^2}{4}\int_0^1 2(1-y)f_n(y\rho)\,dy \tag{3.7}$$

Here, in contrast to (3.6), θ^2 is divided by 4. This is due to the fact that we consider the total size of the tree, which for a sample of size 2 is two times the coalescence time, and hence has a variance four times as large. Note that the mutational variance θh_n is the same as the case of no recombination, but the genealogical variance is reduced by recombination. Comparing with (1.23)

$$\text{var}\,(S_n) = \frac{\theta}{2}E(T_{tot}) + \left(\frac{\theta}{2}\right)^2 \text{var}\,(T_{tot})$$

we see that the integral gives $\text{var}\,(T_{tot})$, a fact that can be seen from the derivation.

Proof. To compute the variance of the number of segregating sites for a sample of size $n > 2$, we again begin with

$$\text{var}\,(S_n) = \sum_{i=1}^{m} \text{var}\,(S_n^j) + \sum_{1\le i\ne j\le m} \text{cov}\,(S_n^i, S_n^j)$$

(1.22) implies that

$$\text{var}\,(S_n^i) = \frac{\theta}{m}\sum_{j=1}^{n-1}\frac{1}{j} + \left(\frac{\theta}{m}\right)^2 \sum_{j=1}^{n-1}\frac{1}{j^2}$$

Let τ_n^i be the total time in the tree for the ith locus. This distribution of S_n^i given τ_n^i is Poisson with mean $(\theta/2m)\tau_n^i$, so $E(S_n^i|\tau_n^i) = (\theta/2m)\tau_n^i$,

$$E(S_n^i S_n^j|\tau_n^i,\tau_n^j) = \left(\frac{\theta}{2m}\right)^2 \tau_n^i\tau_n^j$$

and $\text{cov}\,(S_n^i, S_n^j) = (\theta/2m)^2\,\text{cov}\,(\tau_n^i,\tau_n^j)$. The scaled recombination rate between i and j is $4Nr(j-i)/(m-1)$. Combining our results gives

$$\begin{aligned}\text{var}\,(S_n) = \theta\sum_{j=1}^{n-1}\frac{1}{j} + \frac{\theta^2}{m}\sum_{j=1}^{n-1}\frac{1}{j^2}\\ + \frac{\theta^2}{4m^2}\sum_{k=1}^{m-1} 2(m-k)f_n\left(\frac{k\rho}{(m-1)}\right)\end{aligned}$$

Letting $m\to\infty$, setting $y = j/m$, and noting that the sum approximates an integral gives the indicated result. □

3.2.3 Pairwise differences

The computations above can also be used to study the variance of the number of pairwise differences, Δ_n, but the formula is nicer since it is explicit rather than in terms of $f_n(x)$, which must be computed numerically for $n > 2$.

Theorem 3.5. *For the infinite sites model with recombination,*

$$var(\Delta_n) = \frac{\theta(n+1)}{3(n-1)} + \frac{2\theta^2}{n(n-1)} \int_0^1 2(1-x) \frac{\rho x + (2n^2+2n+6)}{(\rho x)^2 + 13(\rho x) + 18}\, dx \tag{3.8}$$

When $\rho = 0$, this reduces to Tajima's result in (1.30).

$$\text{var}\,(\Delta_n) = \theta \frac{n+1}{3(n-1)} + \theta^2 \frac{2(n^2+n+3)}{9n(n-1)}$$

Proof. Following the appendix of Pluzhnikov and Donnelly (1996), we suppose the m loci are nucleotides and write

$$\Delta_n = \sum_{a=1}^{m} \binom{n}{2}^{-1} \sum_{i<j} \delta_{i,j}^a$$

where $\delta_{i,j}^a = 1$ if nucleotide a is different in sequences i and j.

$$\begin{aligned}\text{var}\,(\Delta_n) = \sum_{a=1}^{m} \text{var}\, &\left(\binom{n}{2}^{-1} \sum_{i<j} \delta_{i,j}^a \right) \\ &+ \binom{n}{2}^{-2} \sum_{a \neq b} \sum_{i<j} \sum_{k<\ell} \text{cov}\,(\delta_{i,j}^a, \delta_{k,\ell}^b)\end{aligned}$$

The summand in the first term is the variance of the heterozygosity, which was computed in (1.15). Plugging in the per locus mutation rate θ/m and noting that when m is large $2 + \theta/m \approx 2$, (1.15) simplifies to

$$\begin{aligned}&\frac{\theta}{m} \cdot \frac{2}{n(n-1)} \left[1 + \frac{2(n-2)}{2} + \frac{(n-2)(n-3)}{6} \right] \\ &= \frac{\theta}{m} \cdot \frac{6 + (6n-12) + (n^2-5n+6)}{3n(n-1)} = \frac{\theta}{m} \cdot \frac{n+1}{3(n-1)}\end{aligned}$$

As in the previous calculation,

$$\text{cov}\,(\delta_{i,j}^a, \delta_{k,\ell}^b) = \left(\frac{\theta}{2m} \right)^2 \text{cov}\,(\tau_{i,j}^a, \tau_{k,\ell}^b)$$

where the τ's are the tree lengths for the indicated samples of size 2. If we let $z = (b-a)\rho/(m-1)$ be the scaled recombination rate between loci a and b

then by the calculation for (1.16) and the covariance for coalescence times in Theorem 3.1 (multiplied by 4),

$$\binom{n}{2}^{-2}\sum_{i<j}\sum_{k<\ell}\operatorname{cov}\left(\tau^a_{i,j},\tau^b_{k,\ell}\right)$$
$$=\binom{n}{2}^{-1}\frac{4}{z^2+13z+18}\left[(z+18)\cdot 1+6\cdot 2(n-2)+4\cdot\binom{n-2}{2}\right]$$

since for each of the $\binom{n}{2}$ values of $i<j$ there is one $k<\ell$ with $i=k$ and $j=\ell$, $2(n-2)$ values with $|\{i,j\}\cap\{k,\ell\}|=1$, and $\binom{n-2}{2}$ with $|\{i,j\}\cap\{k,\ell\}|=0$. A little algebra now gives

$$=\binom{n}{2}^{-1}\frac{4[z+(2n^2+2n+6)]}{z^2+13z+18}$$

Using the fact that there are $2(m-k)$ pairs a,b with $|b-a|=k$, we have

$$\binom{n}{2}^{-2}\sum_{a\neq b}\sum_{i<j}\sum_{k<\ell}\operatorname{cov}\left(\delta^a_{i,j},\delta^b_{k,\ell}\right)$$
$$=\left(\frac{\theta}{2}\right)^2\frac{2}{n(n-1)}\frac{1}{m}\sum_{k=1}^{m}\frac{2(m-k)}{m}\frac{4[\frac{\rho k}{m-1}+(2n^2+2n+6)]}{\left(\frac{\rho k}{m-1}\right)^2+13\frac{\rho k}{m-1}+18}$$

Writing $x=k/(m-1)$ and letting $m\to\infty$ gives the indicated result. □

Although the calculus is somewhat unpleasant, one can evaluate the integral in (3.8) to get a formula first derived by Wakeley (1997).

Theorem 3.6. *For the infinite sites model with recombination,*

$$var(\Delta_n)=\theta\frac{(n+1)}{3(n-1)}+\theta^2 f(\rho,n) \tag{3.9}$$

where $a_n=\rho-2n(n+1)+7$, $b_n=2n(n+1)(13+2\rho)-\rho-55$, *and*

$$f(\rho,n)=\frac{2}{n(n-1)\rho^2}[-2\rho+a_nL_1+b_nL_2]$$

with $L_1=\log\left(\frac{\rho^2+13\rho+18}{18}\right)$ *and* $L_2=\log\left(\frac{(2\rho+13-\sqrt{97})(13+\sqrt{97})}{(2\rho+13+\sqrt{97})(13-\sqrt{97})}\right)$.

Proof. Changing variables $y=\rho x$, and letting $u_n=2n^2+2n+6$, the integral in (3.8) becomes

$$\frac{2}{\rho^2}\int_0^\rho(\rho-y)\frac{y+u_n}{y^2+13y+18}\,dy$$

The quadratic in the denominator has roots

$$r_1 > r_2 \quad \text{where} \quad r_i = \frac{-13 \pm \sqrt{97}}{2}$$

so we write the integrand as

$$\frac{-y^2 + (\rho - u_n)y + u_n\rho}{y^2 + 13y + 18} = -1 + \frac{(\rho - u_n + 13)y + (u_n\rho + 18)}{(y - r_1)(y - r_2)}$$

To evaluate the integral, we note that

$$\begin{aligned}
\int_0^\rho \frac{1}{y^2 + 13y + 18}\, dy &= \frac{1}{r_1 - r_2} \int_0^\rho \frac{1}{y - r_1} - \frac{1}{y - r_2}\, dy \\
&= \frac{1}{r_1 - r_2} \log\left(\frac{\rho - r_1}{-r_1} \cdot \frac{-r_2}{\rho - r_2}\right) = L_2 \\
\int_0^\rho \frac{y}{y^2 + 13y + 18}\, dy &= \frac{1}{r_1 - r_2} \int_0^\rho \frac{r_1}{y - r_1} - \frac{r_2}{y - r_2}\, dy
\end{aligned}$$

Recalling $r_i = (-13 \pm \sqrt{97})/2$ and $r_1 - r_2 = \sqrt{97}$, we have

$$\begin{aligned}
&= -\frac{13}{2} L_2 + \frac{1}{2} \int_0^\rho \frac{1}{y - r_1} + \frac{1}{y - r_2}\, dy \\
&= -\frac{13}{2} L_2 + \frac{1}{2} L_1
\end{aligned}$$

since $L_1 = \log\left(\frac{\rho - r_1}{-r_1} \cdot \frac{\rho - r_2}{-r_2}\right)$. Using these formulas and recalling that $u_n = 2n(n+1) + 6$, the integral becomes

$$\begin{aligned}
J_n &= -\rho + (\rho - 2n(n+1) + 7)\left(-\frac{13}{2} L_2 + \frac{1}{2} L_1\right) \\
&\qquad + ([2n(n+1) + 6]\rho + 18) \cdot \frac{2}{2} L_2 \\
&= -\rho + \frac{\rho - 2n(n+1) + 7}{2} L_1 + \frac{2n(n+1)(13 + 2\rho) - \rho - 55}{2} L_2 \\
&= -\rho + \frac{a_n}{2} L_1 + \frac{b_n}{2} L_2
\end{aligned}$$

Remembering the factors of $2/n(n-1)$ and $2/\rho^2$ we have left behind, the desired result follows. □

The reason for interest in (3.9) is that it allows us to construct an estimator for ρ. Hudson (1987) was the first to do this, but we will follow Wakeley's (1997) improvement.

Theorem 3.7. *Let $k_{\ell,m}$ be the number of differences between sequences ℓ and m, and let*

$$S_\pi^2 = \frac{2}{n(n-1)} \sum_{\ell < m} (k_{\ell m} - \Delta_n)^2$$

be the variance of the $\binom{n}{2}$ *pairwise differences.*

$$E(S_\pi^2) = \theta\frac{2(n-2)}{3(n-1)} + \theta^2 g_\pi(\rho, n) \tag{3.10}$$

where $\alpha_n = (n+1)\rho - (n-7)$, $\beta_n = (15n-1)\rho + (49n-55)$, *and*

$$g_\pi(\rho, n) = \frac{(n-2)}{n(n-1)\rho^2}\{-2\rho(n+1) + \alpha_n L_1 + \beta_n L_2\}$$

and L_1 *and* L_2 *are as in (3.9).*

Proof. S_π^2 can be rewritten as the second moment of the sample minus the square of the mean of the sample:

$$S_\pi^2 = \left[\frac{2}{n(n-1)}\sum_{\ell<m} k_{\ell m}^2\right] - \Delta_n^2$$

Since $Ek_{\ell m} = \theta = E\Delta_n$, we have

$$E(S_\pi^2) = E(k_{\ell m}^2) - E(\Delta_n^2) = \text{ var}\,(\Delta_2) - \text{ var}\,(\Delta_n)$$

since $k_{\ell m}$ is the number of pairwise differences between two sequences.

From this and (3.9) it follows that

$$E(S_\pi^2) = \theta - \theta\frac{n+1}{3(n-1)} + \theta^2[f(\rho, 2) - f(\rho, n)]$$

This gives the θ term in (3.10). To compute $n(n-1)\rho^2[f(\rho,2) - f(\rho,n)]$, we write

$$\begin{aligned} -n(n-1)\rho^2 f(\rho, n) = -2\{&-2\rho + (\rho - 2n(n+1) + 7)L_1 \\ &+([4n^2 + 4n - 1]\rho + [26n(n+1) - 55])L_2\} \\ n(n-1)\rho^2 f(\rho, 2) = (n^2 - n)\{&-2\rho + (\rho - 5)L_1 + (23\rho + 101)L_2\} \end{aligned}$$

in order to prepare for the miracle that every term in the difference has $(n-2)$ as a factor

$$\begin{aligned} -2\rho(n^2 - n - 2) &= -2\rho(n-2)(n+1) \\ \rho L_1(n^2 - n - 2) &= \rho L_1(n-2)(n+1) \\ L_1(-5n^2 + 5n + 4n^2 + 4n - 14) &= -L_1(n-2)(n-7) \\ \rho L_2(23n^2 - 23n - 8n^2 - 8n + 2) &= \rho L_2(n-2)(15n-1) \\ L_2(101n^2 - 101n - 52n^2 - 52n + 110) &= L_2(n-2)(49n-55) \end{aligned}$$

Combining our calculations, we see that the coefficient of θ^2 in $E(S_\pi^2)$ is given by the indicated formula. □

To introduce the estimator now, we follow Wakeley (1997) and write π for Δ_n. Since $E\pi^2 = \text{var}(\pi) + (E\pi)^2$, (3.9) implies

$$E\pi^2 = \theta\frac{(n+1)}{3(n-1)} + \theta^2[f(\rho,n)+1]$$

so the following is an unbiased estimator of θ^2:

$$\frac{\pi^2 - [(n+1)/3(n-1)]\pi}{f(\rho,n)+1}$$

and one can estimate ρ by solving

$$S_\pi^2 = \pi\frac{2(n-2)}{3(n-1)} + g_\pi(\rho,n)\frac{\pi^2 - [(n+1)/3(n-1)]\pi}{f(\rho,n)+1}$$

Example 3.1. Wakeley (1997) applied his estimator to a data set of Schaeffer and Miller (1993), who sequenced $n = 99$ individuals in a 3.5 kb region containing the alcohol dehydrogenase genes *Adh* and *Adh-dup* of *Drosophila psuedoobscura*. The data set had 359 polymorphic sites with 27 having 3 nucleotides segregating, for a total of 386 mutations. Since

$$\frac{386\cdot 385}{2}\cdot 3500 = 21.23$$

the large number of double hits is only a few more than we expect. Wakeley discarded the sites that were hit twice, computing $\pi = 31.7$ and a moment estimator of $\rho = 282$. Simulations suggested a 95% confidence interval of $[172, 453]$. Note that the estimate of ρ is about nine times the estimate of θ, i.e., in this region of the genome the recombination rate is about nine times the mutation rate.

3.3 Linkage disequilibirum

Linkage disequilibrium (LD) refers to the nonindependence of alleles at different loci. For example, suppose that allele A at locus 1 and allele B at locus 2 are at frequencies π_A and π_B, respectively. If the two loci were independent, then the AB haplotype would have frequency $\pi_{AB} = \pi_A\pi_B$. If this is not the case, then the two loci are in linkage disequilibrium and we let

$$D_{AB} = \pi_{AB} - \pi_A\pi_B$$

If r is the recombination probability between the two loci, then adding a superscript to indicate the generation number

$$\pi_{AB}^t = (1-r)\pi_{AB}^{t-1} + r\pi_A^{t-1}\pi_B^{t-1}$$

so if we ignore fluctuations in gene frequencies,

$$D_{AB}^t = (1-r)D_{AB}^{t-1} \cdots = (1-r)^t D_{AB}^0 \tag{3.11}$$

A wide variety of statistics have been proposed to measure LD. For probabilists and statisticians, the most natural is the square of the correlation coefficient

$$r^2 = \frac{D_{AB}^2}{\pi_A \pi_a \pi_B \pi_b} \tag{3.12}$$

A second commonly used measure, introduced by Lewontin (1964), is

$$D' = \begin{cases} \frac{D_{AB}}{\min(\pi_A\pi_b, \pi_B\pi_a)} & \text{if } D_{AB} > 0 \\ \frac{D_{AB}}{\min(\pi_A\pi_B, \pi_a\pi_b)} & \text{if } D_{AB} < 0 \end{cases}$$

To explain this formula, we will prove

Theorem 3.8. $D' \in [-1,1]$ *with the extremes achieved when one of the four combinations is absent from the population.*

In contrast, $r^2 = 1$ when there are only two combinations in the population: AB and ab, or Ab and aB.

Proof. Clearly, $\pi_{AB} \le \min\{\pi_A, \pi_B\}$. This implies

$$\pi_{AB} - \pi_A\pi_B \le \min\{\pi_A(1-\pi_B), \pi_B(1-\pi_A)\} = \min(\pi_A\pi_b, \pi_B\pi_a)$$

and $D' \le 1$. To get the bound $D' \ge -1$, note that $D_{AB} = -D_{Ab}$ and use the first result with the roles of B and b interchanged to conclude

$$D_{Ab} \le \min(\pi_A\pi_B, \pi_b\pi_a)$$

At this point, we have shown $D' \in [-1,1]$. If $D' = 1$, then $\pi_{AB} = \min\{\pi_A, \pi_B\}$. If $\pi_{AB} = \pi_A$, then $\pi_{Ab} = 0$. If $\pi_{AB} = \pi_B$, then $\pi_{aB} = 0$. If $D' = -1$, then we note $D_{AB} = -D_{Ab}$ and use the previous argument. □

The computation of Er^2 is made difficult by the correlation between the numerator and denominator. Many people believe, see e.g., Hartl and Clark (2007, page 532), the following:

Mythical result. If $\rho = 4Nr$ then $Er^2 = 1/(1+\rho)$.

An Internet article on LD I found attributes this result to Hill and Robertson (1968). However, all they have to say about this is that the limiting value of Er^2 appears to approach $1/\rho$ as Nr increases, see on page 229. Song and Song (2007) have developed numerical methods to compute Er^2 which allow them to prove that $Er^2 \sim 1/\rho$ as $\rho \to \infty$.

Ohta and Kimura (1971) argued that unless one or both of the allele frequencies took values near 0 or 1,

$$Er^2 \approx \frac{ED_{AB}^2}{E(\pi_A\pi_a\pi_B\pi_b)} \equiv \sigma_d^2 \tag{3.13}$$

(where the $\equiv$ indicates the second equality is the definition of σ_d^2), and they used diffusion theory (see Theorem 8.13 at the end of Section 8.2) to show

Theorem 3.9.

$$\sigma_d^2 = \frac{10+\rho}{22+13\rho+\rho^2}$$

Note that when ρ is large, $\sigma_d^2 \approx 1/\rho$. Here we will give McVean's (2002) derivation based on coalescent theory. Note that here the genotype frequencies are those of the population. One can compute the expected value for a sample, but the result is quite messy; see (10) in McVean (2002).

Proof. To simplify the first calculation, it is useful to consider the general case where there may be more than two alleles. In this case,

$$D_{\alpha,\beta} = f_{\alpha,\beta} - p_\alpha q_\beta$$

where p_α and q_β are the frequencies of α and β at the two loci, $f_{\alpha,\beta}$ is the frequency of the α,β haplotype, and one defines the square of the disequilibrium by

$$D^2 = \sum_{\alpha,\beta}(f_{\alpha,\beta} - p_\alpha q_\beta)^2$$

In the case of two alleles, $D_{11} = -D_{12} = -D_{21} = D_{22}$, so $D^2 = 4D_{\alpha,\beta}^2$. Note that $\sum_{\alpha,\beta} D_{\alpha,\beta} = 1 - 1 = 0$, so for two alleles we have

$$ED_{\alpha,\beta} = 0 \tag{3.14}$$

without any assumption other than a symmetric role for the two alleles.

To compute the second moment, we note that

$$\begin{aligned} ED^2 &= \sum_{\alpha,\beta}(f_{\alpha,\beta} - p_\alpha q_\beta)^2 \\ &= \sum_{\alpha,\beta} f_{\alpha,\beta}^2 - 2f_{\alpha,\beta}p_\alpha q_\beta + p_\alpha^2 q_\beta^2 \\ &= F_{ij,ij} - 2F_{ij,ik} + F_{ij,k\ell} \end{aligned} \tag{3.15}$$

where the $F_{ij,k\ell}$ is the probability that sequences i and j are equal at the first locus and sequences k and ℓ are equal at the second locus, and the sequences i, j, k, and ℓ are chosen at random. This is (A6) from Hudson (1985), who attributed the result to Strobeck and Morgan (1978). Since we are doing our calculation for the entire population, we can suppose that the indices i, j, k, and ℓ are distinct.

When there are only two alleles at each locus, the square of the disequilibrium coefficient is independent of how the alleles are defined, so we can

consider $F^*_{ij,k\ell}$ the probability that the derived mutation occurs in sequences i and j at the first locus and in sequences k and ℓ at the second locus. It follows from (3.15) that we have

$$ED^2_{AB} = F^*_{ij,ij} - 2F^*_{ij,ik} + F^*_{ij,k\ell}$$

Let I^h_{ij} be the branch length leading from the most recent common ancestor of i and j to the most recent common ancestor of the sample at locus $h = 1, 2$, and τ^h the total size of the tree for the population at locus h. We assume that the units for these times are $2N$ generations. Assuming the mutation rate μ is the same at the two sites and letting $u = 2N\mu$, we have

$$F^*_{ij,k\ell} = \frac{E(uI^1_{ij}e^{-u\tau^1}uI^2_{k\ell}e^{-u\tau^2})}{E(u\tau^1 e^{-u\tau^1}u\tau^2 e^{-u\tau^2})}$$

Taking the limit as $u \to 0$ to eliminate the mutation rate gives

$$F^*_{ij,k\ell} = \frac{E(I^1_{ij}I^2_{k\ell})}{E(\tau^1\tau^2)}$$

Let t^h_{ij} be the time of the most recent common ancestor of i and j at locus h, and let T^h be the time of the most recent common ancestor of the population at locus h. Writing $I^h_{ij} = T^h - t^h_{i,j}$ and using symmetry to conclude that $E(T^1 t^2_{ij})$ does not depend on i, j (assuming they are distinct), we have

$$\begin{aligned}
&E(I^1_{ij}I^2_{ij}) - 2E(I^1_{ij}I^2_{ik}) + E(I^1_{ij}I^2_{k\ell}) \\
&\quad = (1-2+1)E(T^1T^2) - (1-2+1)E(T^1t^2_{ij} + T^2t^1_{ij}) \\
&\qquad + E(t^1_{ij}, t^2_{ij}) - 2E(t^1_{ij}, t^2_{ik}) + E(t^1_{ij}, t^2_{k\ell})
\end{aligned}$$

The means $Et^h_{ij} = 1$, so we have

$$ED^2_{AB} = \frac{\operatorname{cov}(t^1_{ij}, t^2_{ij}) - 2\operatorname{cov}(t^1_{ij}, t^2_{ik}) + \operatorname{cov}(t^1_{ij}, t^2_{k\ell})}{E(\tau^1\tau^2)}$$

A similar approach can be used on the denominator of σ^2_d, which represents the probability that when two alleles i and j are drawn with replacement at the first locus and another two k and ℓ are drawn with replacement at the second locus, then i and k have the mutant alleles, and j and ℓ do not.

$$\begin{aligned}
E(\pi_A\pi_a\pi_B\pi_b) &= \lim_{\mu\to 0} \frac{E(ut^1_{ij}e^{-u\tau^1}ut^2_{k\ell}e^{-u\tau^2})}{E(u\tau^1e^{-u\tau^1}u\tau^2e^{-u\tau^2})} \\
&= \frac{E(t^1_{ij}t^2_{k\ell})}{E(\tau^1\tau^2)} = \frac{\operatorname{cov}(t^1_{ij}, t^2_{k\ell}) + 1}{E(\tau^1\tau^2)}
\end{aligned}$$

Combining this with the formula for ED^2_{AB}, we have

$$\sigma_d^2 = \frac{\operatorname{cov}(t_{ij}^1, t_{ij}^2) - 2\operatorname{cov}(t_{ij}^1, t_{ik}^2) + \operatorname{cov}(t_{ij}^1, t_{k\ell}^2)}{\operatorname{cov}(t_{ij}^1, t_{k\ell}^2) + 1}$$

Using (3.1) now, we have

$$\sigma_d^2 = \frac{18 + \rho - 2(6) + 4}{4 + 18 + 13\rho + \rho^2} = \frac{10 + \rho}{22 + 13\rho + \rho^2}$$

which proves the desired result. □

LD in humans

Understanding the extent and distribution of linkage disequilibrium in humans is an important question because LD plays a fundamental role in the fine scale mapping of human disease loci, see e.g., Risch and Merikangas (1996) and Cardon and Bell (2001). The technique known as association mapping differs from traditional pedigree studies in that marker-disease associations are sought in populations of unrelated individuals. To explain the idea behind this approach, imagine that a disease causing mutation has just occurred in a population. The chromosome on which this mutation occurred contains specific alleles in neighboring polymorphic loci. At first, the mutation will only be observed in conjuction with these alleles, so the association (or LD) with these alleles will be high. Through time these associations will dissipate because of recombination, but the closest loci will experience the fewest recombinations and hence retain the highest levels of LD. Thus, by looking for significant correlations between disease state and alleles, we can hope to identify the region in which the disease causing genetic mutation lies.

One of the surprising patterns revealed by the construction of a dense genome-wide map of single nucleotide polymorphisms (SNPs) was the slow decay of LD. Kruglyak's (1999) simulation study suggested that useful levels of LD were unlikely to extend beyond an average distance of about 3 kilobases (kb) in the general population. In contrast, Reich et al. (2001) found in a study of 19 randomly selected regions in the human genome that LD in a United States population of northern European descent typically extends 60 kb from common alleles. These findings were confirmed and further quantified by Dawson et al. (2002), who measured LD along the complete sequence of human chromosome 22, and Ke et al. (2004), who studied a contiguous 10 megabase segment of chromosome 20. In both cases, see Figure 1 of Dawson et al. (2002) and Figure 3 of Ke et al. (2004), LD as measured by the square of the correlation coefficient r^2 is about 0.1 at 100 kb. If, as Ardlie, Kruglyak, and Seielstad (2002) argued, $r^2 > 1/3$ is the limit of useful LD, this occurs at about 30 kb.

3.4 Ancestral recombination graph

As we have seen in the first two sections of this chapter, analytical results for genealogies with recombination are difficult and messy. In this section we will

show that it is reasonably easy to simulate the process. Hudson (1983) was the first to do this. Here we follow the formulation of Griffiths and Marjoram (1997). Taking a continuous perspective, we consider a segment of DNA and rescale it to be the unit interval $[0, 1]$. If we suppose that the segment is small enough so that two recombinations in one generation can be ignored, then the dynamics of a Wright-Fisher model (going forward in time) may be described as follows. To generate a chromosome in the next generation, with probability $1 - r$ we copy all of the contents from one randomly chosen individual. With probability r a recombination occurs. We pick a point uniformly along the chromosome and two individuals at random from the population. We copy the genetic material to the left of that point from the first individual and copy the material to the right from the second.

Reversing our perspective leads to a genealogical process, which in the case of a sample of size 4 can be drawn as in the next figure. Lines merging indicate coalescence events and are marked with letters. Splits indicate a recombination at the indicated location, with the convention that at recombination events the left half comes from the individual on the left and the right half comes from the individual on the right.

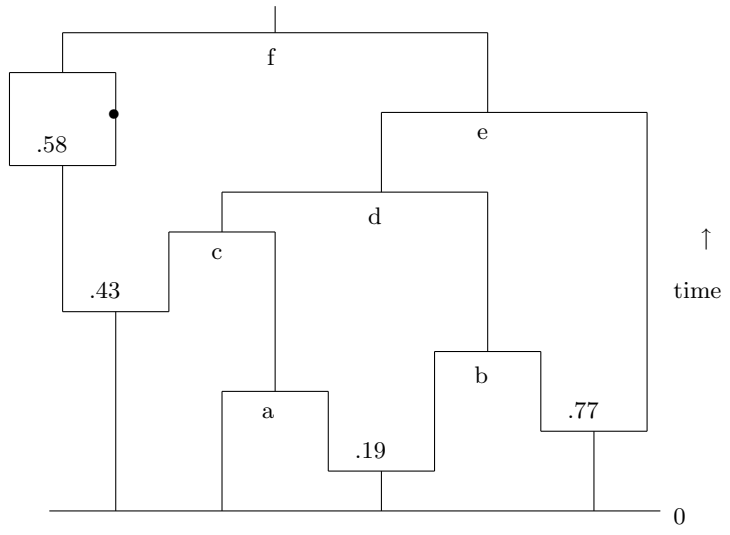

Fig. 3.2. Realization of the ancestral recombination graph for a sample of size 4.

The number of ancestors in the sample $2Nt$ units back in time Y_t is a birth and death process. When there are k ancestors,

the death rate due to coalescence of lineages is $\mu_k = k(k-1)/2$,

the birth rate due to recombinations is $\lambda_k = k\rho/2$, where $\rho = 4Nr$.

The first rate is by now familiar. To check the second, note that the probability of no recombination in one generation is

$$(1-r)^k \approx 1 - kr = 1 - \frac{1}{2N} \cdot \frac{k\rho}{2}$$

Because of the quadratic birth rate compared to the linear birth rate, if we start with $m > 1$ sequences then with probability 1, there is a time τ_m where the number of ancestors is 1. Griffiths (1991) has calculated

Theorem 3.10.

$$E\tau_m = \frac{2}{\rho} \int_0^1 \left(\frac{1 - v^{m-1}}{1-v} \right) (e^{\rho(1-v)} - 1)\, dv \tag{3.16}$$

When $m = 2$, this becomes

$$\begin{aligned} E\tau_2 &= \frac{2}{\rho} \int_0^1 e^{\rho(1-v)} - 1\, dv = \frac{2}{\rho} \left(-\frac{1}{\rho} e^{\rho(1-v)} - v \right) \Big|_0^1 \\ &= \frac{2}{\rho} \left(-\frac{1}{\rho} - 1 + \frac{1}{\rho} e^{\rho} \right) = \frac{2}{\rho^2} \cdot (e^{\rho} - 1 - \rho) \end{aligned}$$

Note that as $\rho \to 0$, $E\tau_2 \to 1$, the result for the ordinary coalescent, but $E\tau_2$ grows exponentially fast as ρ increases, which means that it will be very slow to simulate the ancestral recombination graph for large regions.

Proof. We use the methods of Section 1.5. We begin by computing hitting probabilities. The number of lineages in the genealogy, Y_t, changes

$$\begin{aligned} k &\to k+1 \ \text{ at rate } \rho k/2 \\ k &\to k-1 \ \text{ at rate } k(k-1)/2 \end{aligned}$$

so the embedded jump chain, X_n,

$$\begin{aligned} k &\to k+1 \ \text{ with prob. } p_k = \rho/(\rho + k - 1) \\ k &\to k-1 \ \text{ with prob. } q_k = (k-1)/(\rho + k - 1) \end{aligned}$$

We want to find a function ϕ so that $\phi(X_n)$ is a martingale. For this we need

$$\phi(k) = \phi(k+1)p_k + \phi(k-1)q_k$$

or, rearranging,

$$\phi(k+1) - \phi(k) = \frac{q_k}{p_k}(\phi(k) - \phi(k-1))$$

X_n has state space $\{1, 2, \ldots\}$. Setting $\phi(2) - \phi(1) = 1$ and iterating gives

$$\phi(k) - \phi(k-1) = \prod_{j=2}^{k-1} \frac{q_j}{p_j} = \frac{(k-2)!}{\rho^{k-2}}$$

(By convention, $\prod_{j=2}^{1} q_j/p_j = 1 = 0!/\rho^0$.) Taking $\phi(1) = 0$, we have

$$\phi(m) = \sum_{k=2}^{m} \frac{(k-2)!}{\rho^{k-2}}$$

Let $T_k = \min\{n \geq 0 : X_n = k\}$ be the time of the first visit to k. Since $\phi(X_n)$ is a martingale and the absorbing state 1 has $\phi(1) = 0$,

$$P_m(T_k < \infty) = \begin{cases} 1 & k \leq m \\ \frac{\phi(m)}{\phi(k)} & k > m \end{cases}$$

Let $T_k^+ = \min\{n \geq 1 : X_n = k\}$ be the time of the first return to k. If we start at k, the only way to avoid returning to k is to go from $k \to k-1$ on the first jump and then not come back, so

$$P_k(T_k^+ = \infty) = \frac{k-1}{\rho + k - 1} \left(1 - \frac{\phi(k-1)}{\phi(k)}\right)$$

If the process reaches k, then the number of visits to k, N_k, will have a geometric distribution with mean $1/P_k(T_k^+ = \infty)$, so

$$E_m N_k = \frac{P_m(T_k < \infty)}{P_k(T_k^+ = \infty)}$$

and it follows that

$$E_m N_k = \begin{cases} \frac{\rho+k-1}{k-1} \left(\frac{\phi(k)}{\phi(k)-\phi(k-1)}\right) & k \leq m \\ \frac{\phi(m)}{\phi(k)} \cdot \frac{\rho+k-1}{k-1} \left(\frac{\phi(k)}{\phi(k)-\phi(k-1)}\right) & k > m \end{cases}$$

Returning to Y_t, the continuous-time process that gives the size of the genealogy, we let S_k be the amount of time Y_t spends at k. Since the rate of jumps out of k is $k(\rho + k - 1)/2$, and $\phi(k) - \phi(k-1) = (k-2)!/\rho^{k-2}$,

$$E_m S_k = \frac{2}{k(\rho + k - 1)} E_m N_k = \begin{cases} \frac{2}{k(k-1)} \cdot \phi(k) \cdot \frac{\rho^{k-2}}{(k-2)!} & k \leq m \\ \phi(m) \cdot \frac{2}{k(k-1)} \cdot \frac{\rho^{k-2}}{(k-2)!} & k > m \end{cases}$$

Using $x \wedge y = \min\{x, y\}$, we can combine the two formulas into one:

$$E_m S_k = 2\phi(k \wedge m) \frac{\rho^{k-2}}{k!} = 2\frac{\rho^k}{k!} \sum_{j=2}^{k \wedge m} \frac{(j-2)!}{\rho^j}$$

where in the last expression we have multiplied the numerator and denominator by ρ^2.

Since $T_1 = \sum_{k=2}^{\infty} S_k$, we have

$$E_m T_1 = 2 \sum_{k=2}^{\infty} \frac{\rho^k}{k!} \sum_{j=2}^{k \wedge m} \frac{(j-2)!}{\rho^j}$$

Interchanging the order of the summation and letting $i = k - j$, the above

$$= 2 \sum_{j=2}^{m} \sum_{k=j}^{\infty} (j-2)! \frac{\rho^{k-j}}{k!} = 2 \sum_{j=2}^{m} \sum_{i=0}^{\infty} (j-2)! \frac{\rho^i}{(i+j)!}$$

To see the first step note that the sum is over $k \geq 2$, $2 \leq j \leq m$, $j \leq k$.

To relate the last formula to the answer given above, we note that

$$\begin{aligned} &\frac{2}{\rho} \int_0^1 \frac{1 - v^{m-1}}{1 - v} \left(e^{\rho(1-v)} - 1 \right) dv \\ &\quad = \frac{2}{\rho} \int_0^1 \sum_{k=0}^{m-2} v^k \sum_{\ell=1}^{\infty} \frac{\rho^\ell (1-v)^\ell}{\ell!} \, dv \end{aligned}$$

Letting $\ell = i + 1$ and $k = j - 2$, this becomes

$$2 \sum_{j=2}^{m} \sum_{i=0}^{\infty} (j-2)! \rho^i \int_0^1 \frac{v^{j-2}}{(j-2)!} \frac{(1-v)^{i+1}}{(i+1)!} \, dv$$

Integrating by parts $j - 2$ times gives

$$\int_0^1 \frac{v^{j-2}}{(j-2)!} \frac{(1-v)^{i+1}}{(i+1)!} \, dv = \int_0^1 \frac{(1-v)^{i+j-1}}{(i+j-1)!} \, dv = \frac{1}{(i+j)!}$$

and we have proved the desired result. □

The time τ_m is only an upper bound on the MRCAs for all nucleotides because (i) some of the ancestors in the genealogy, such as the one marked by the black dot in the example in Figure 3.2 above, may have no genetic material that is ancestral to that in the sample, and (ii) different segments of the chromosome will have MRCAs at different times $\leq \tau_m$. In the example, the four chromosome segments have the following genealogies. Here the letters correspond to the coalescence events in the ancestral recombination graph, while $\times$ indicates where the recombination occurred to change the tree to the next one.

3.4.1 Simulation

In simulating the ancestral recombination graph, we do not want to create ancestors that have no genetic material in common with the sample. To explain

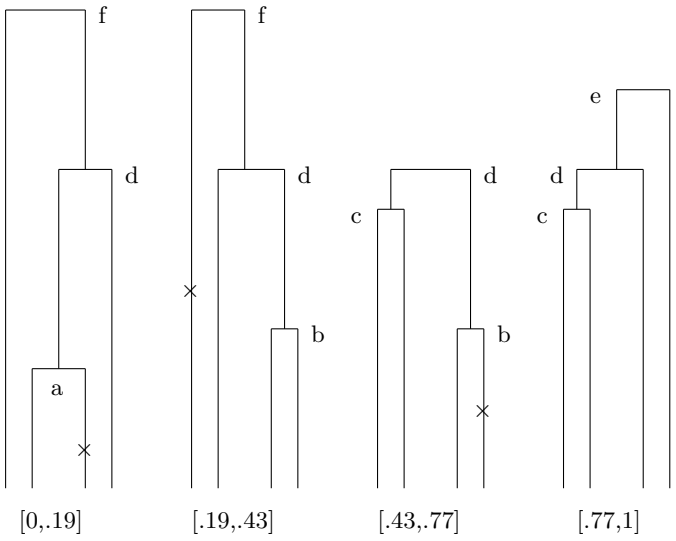

Fig. 3.3. Coalescent trees for the four segments in our example.

how to avoid this, we will, for simplicity, consider the two-locus case. Generation t signifies the population t generations before the present. We consider a sample of size n from the current population (generation 0). Let $g(t)$ be the number of chromosomes of generation t that contain genetic material at either locus that is directly ancestral to genetic material of the sample. Let $d(t)$ be the number of ancestral chromosomes that contain copies of both loci.

Let t_k be the time of the kth event (i.e., recombination or coalescence). If time is written in units of $2N$ generations, then $t_{k+1} - t_k$ has an exponential distribution with rate

$$\lambda(t_k) = d(t_k)\rho/2 + g(t_k)(g(t_k) - 1)/2$$

The next event is a recombination with probability

$$\frac{d(t_k)\rho/2}{\lambda(t_k)}$$

In this case, we pick one of the $d(t_k)$ chromosomes that contain copies of both loci to split. The event is a coalescence with probability

$$\frac{g(t_k)(g(t_k) - 1)/2}{\lambda(t_k)}$$

In this case, we pick one pair of ancestral chromosomes to coalesce. Of course, after either of these events, we must compute the new values of $g(t_{k+1})$ and $d(t_{k+1})$. When we finally reach the point at which there is only one chromosome, then we can move forward in time assigning mutations to the branches as in the ordinary coalescent. See Hudson (1991) for more details. The extension to a chromosomal segment is fairly straightforward. However, we are all fortunate that Hudson has created and made publicly available a program *ms* that will do these simulations.

`http://home.uchicago.edu/~rhudson1/source.html`

Coalescence trees along the chromosome

Wiuf and Hein (1999) developed a different way of constructing the ancestral recombination graph for a sequence. To start, one generates a genealogy for the endpoint $x_0 = 0$ in the usual way. Let r be the recombination rate per unit distance and $\rho = 4Nr$. If τ^0 is the total size of the tree (measured in units of $2N$ generations), then the distance x_1 until we encounter the next recombination is exponential with rate $\tau^0\rho/2$. To generate the genealogy at x_1, pick a point U^1 uniformly from the tree. At this point, the portion of the chromosome to the right of the recombination has a different path, so we simulate to see where it should rejoin the tree.

The simplest, but unfortunately incorrect, way to do this is to erase the part of the branch from that point to the next coalescence event, creating a floating lineage, and then simulate to determine where the lineage should reattach to the rest of the tree. In the example drawn, from the time U^1 marked by the $\times$ to T_3^0 the coalescence rate is 3, from time T_3^0 to T_1^0 the rate is 2, and after T_1^0 at rate 1. Thus, the new time of the most recent common ancestor may be larger than the old one. The coalescence times in the new tree are labeled T_4^1, T_3^1, T_2^1, and T_1^1, i.e., the superscript indicates tree numbers along the chromosome, and the subscript is the number of lineages just before the coalescence event.

This recipe is simple, but, unfortunately, it is wrong. If the new lineage coalesces with the erased one, then until there is another recombination the new lineage must follow the choices made by the erased lineage until it reattaches to the tree. This event will happen with significant probability, since from the $\times$ until time T_3^0 the erased lineage is one of four possible targets for coalescence, and from T_3^0 to T_2^0 it is one of three. The unfortunate consequence of this is that one cannot erase the lineage but must keep adding paths to the diagram, increasing the complexity and the computer storage requirements of the computation. We refer the reader to Wiuf and Hein (1999) for details about how to correctly implement the algorithm.

3.4.2 Two approximate algorithms

Given the problems that arise from the ghosts of previous branches, it is natural to forget them to produce a Markovian process that approximates

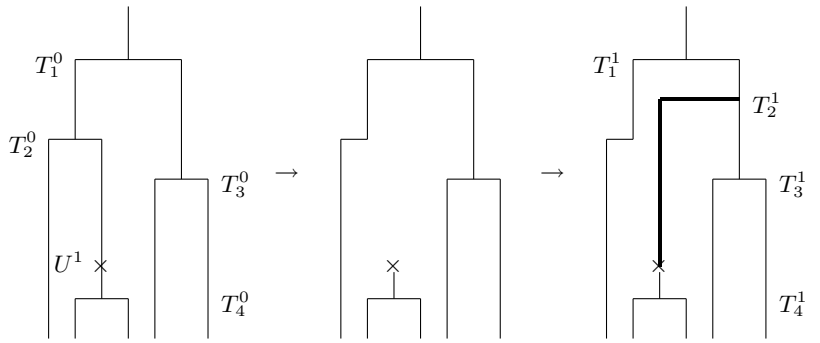

Fig. 3.4. The wrong way of moving from one tree to the next.

how the genealogical tree of a sample changes as we move along a chromosome. McVean and Cardin (2005) implemented the approximation as it was described above, and named it the spatial Markovian coalescent (SMC). Marjoram and Wall (2006) modified the approximation so the old lineage is not erased until after the coalescence point of the new one has been determined, which allows for the possibility that the new lineage coalesces with the one that was to be erased and no change occurs in the genealogy. In both papers, simulations show that the improved algorithm produces simulated data that for many statistics is close to that resulting from simulation of the full coalescent. Marjoram and Wall developed software (*FastCoal*) for implementing these approximations, which is available from the authors. The most important reason for being interested in the approximation is that the amount of computation needed for the SMC and the MW algorithms increases linearly with the length of chromosome simulated, while the ancestral recombination graph requires an exponentially increasing amount of work.

To obtain some insight into the workings of these approximations, we will consider $n = 2$ and for simplicity only the SMC, which is less accurate than the MW chain, but nicer to compute with. When $n = 2$, there is only one coalescence time, so we let H_k be the height of the kth tree. The place at which the recombination occurs, $U^1 = \xi_1 H_0$, where ξ_1 is uniform on $(0, 1)$. The new height $H_1 = U^1 + \eta_1$ and η_1 is exponential with mean 1. Since $H_1 = \xi_1 H_0 + \eta_1$, it is easy to see that if

$$H_0 =_d \sum_{n=1}^{\infty} \eta_n \prod_{m=1}^{n-1} \xi_m$$

then H_1 also has this distribution, so this is the stationary distribution π for the discrete-time chain. Since trees of height y stay around for a mean time

$1/2y$, the continuous-time stationary distribution has

$$\mu(dy) = \frac{1}{c} \cdot \frac{1}{2y}\pi(dy) \quad \text{where} \quad c = \int \frac{1}{2y}\pi(dy)$$

The first thing to prove is that

Theorem 3.11. *The stationary distribution for the height of the SMC tree for a sample of size 2, μ, is exponential with mean 1, which is the correct distribution for the coalescent.*

Proof 1. To prove this, we begin by taking the expected value

$$E(H_1|H_0) = \frac{H_0}{2} + 1$$

to conclude that in equilibrium $EH = EH/2 + 1$ and $EH = 2$. We will now prove by induction that $EH^k = (k+1)!$. To do this we begin with the observations that (i) the uniform distribution on (0,1), ξ, has

$$E\xi^m = \int_0^1 x^m \, dx = \frac{1}{m+1}$$

and (ii) integration by parts and induction shows that the mean 1 exponential distribution, η, has

$$E\eta^m = \int_0^\infty x^m e^{-x} \, dx = m \int_0^\infty x^{m-1} e^{-x} \, dx = m!$$

Now, since H and $H\xi + \eta$ have the same distribution, if the formula $EH^m = (m+1)!$ is correct for all powers $m < k$, then

$$\begin{aligned} EH^k &= \sum_{m=0}^{k} \binom{k}{m} E(H^m)E(\xi^m)E(\eta^{k-m}) \\ &= EH^k \frac{1}{k+1} + \sum_{m=0}^{k-1} \frac{k!}{m!(k-m)!}(m+1)!\frac{1}{m+1}(k-m)! \end{aligned}$$

Each term in the sum is $k!$, so we have

$$\frac{k}{k+1}EH^k = k \cdot k! \quad \text{and hence} \quad EH_k = (k+1)!$$

From the moments we see that $\pi(dx) = xe^{-x} \, dx$, so $c = 1/2$ and $\mu(dy) = e^{-y} \, dy$. □

Proof 2. We begin by computing the transition kernel for the discrete-time Markov chain $H_0, H_1, \ldots$. Suppose $H_n = x$. Breaking things down according to the value of $z = H_n\xi$, we have

$$\text{if } y > x \quad K(x,y) = \int_0^x \frac{dz}{x} e^{-(y-z)} = \frac{e^{x-y} - e^{-y}}{x}$$
$$\text{if } y < x \quad K(x,y) = \int_0^y \frac{dz}{x} e^{-(y-z)} = \frac{1 - e^{-y}}{x}$$

We will now show that $\pi(x) = xe^{-x}$ satisfies the *detailed balance condition*: $\pi(x)p(x,y) = \pi(y)p(y,x)$, which implies

$$\int \pi(x)p(x,y)\,dx = \pi(y) \int p(y,x)\,dx = \pi(y)$$

To check the detailed balance condition, we can assume that $y > x$. In this case, our formulas imply

$$\begin{aligned} \pi(x)K(x,y) &= xe^{-x} \cdot (e^{x-y} - e^{-y})/x \\ &= e^{-y} - e^{-(x+y)} \\ &= ye^{-y} \cdot (1 - e^{-x})/y = \pi(y)K(y,x) \end{aligned}$$

The detailed balance condition implies not only that the π is a stationary distribution, but also that the chain is *reversible*; in equilibrium the chain looks the same going forwards or backwards. This is natural since a chromosome consists of two complementary strands of DNA and has no inherent orientation. □

As McVean and Cardin (2005) explained, a nice way of thinking about the SMC is that it is a modification of the ancestral recombination graph in which coalescence is forbidden if the ancestors do not have genetic material in common. In the two-locus case, this means that (a)'s may coalesce with (ab)'s and (b)'s with (ab)'s but (a)'s and (b)'s are forbidden from coalescing. From this we see that, in a large population, the covariance of the coalescence times T_a and T_b is ≈ 0 if the initial sampling configuration is $(a)(ab)(b)$ or $(a)(a)(b)(b)$. Thus, the covariance for the state $(ab)(ab)$ can only be nonzero if coalescence occurs before recombination, and in this case it is 1, so for the SMC

$$E_\pi(H_0 H_t - 1) = \frac{1}{1+t}$$

since $\rho/2 = t/2$ is the scaled recombination rate for the interval $[0,t]$, and there are two individuals in the sample subject to recombination. In comparison to the exact answer given in (3.1),

$$\frac{1}{1+t} < \frac{t+18}{t^2+13t+18} < 1.2797 \frac{1}{1+t}$$

the maximum relative error of 28% occurring when $t = 4.24$. This is a pretty large error. However, if we consider the variance of the number of segregating sites, $\text{var}\,(S_2)$, computed exactly in (3.6), the error will be less since we will integrate $E_\pi(H_0 H_t - 1)$ over a range of t values.

Using the forbidden coalescence reasoning, one can compute the joint distribution $P_\pi(H_0 = x, H_t = y)$. If coalescence occurs before recombination, an event of probability $1/(1+t)$, then the joint distribution (H_0, H_t) is $(t+1)e^{-(1+t)x}1_{\{x=y\}}$. When recombination occurs before coalescence, an event of probability $t/(1+t)$, the additional time for the two coalescences are independent mean 1 exponentials, so if $x < y$, the joint density is

$$\begin{aligned}&\frac{t}{t+1}\int_0^x (1+t)e^{-(1+t)z}e^{-(x-z)}e^{-(y-z)}\,dz\\ &= t\int_0^x e^{(1-t)z}e^{-x}e^{-y}\,dz = t\frac{e^{(1-t)x}-1}{1-t}e^{-(x+y)}\end{aligned}$$

By symmetry, the formula for $y < x$ is

$$t\frac{e^{(1-t)y}-1}{1-t}e^{-(x+y)}$$

Dividing by $P_\pi(H_0 = x) = e^{-x}$, we have the transition probability

$$P(H_t = y|H_0 = x) = \begin{cases} e^{-tx}\delta_x & x = y\\ t\frac{e^{(1-t)x}-1}{1-t}e^{-y} & x < y\\ t\frac{e^{(1-t)y}-1}{1-t}e^{-y} & x > y\end{cases}$$

3.5 Counting recombinations

Suppose that there are two alleles, A and a, at site i and two alleles, B and b, at site j. As we observed in the analysis of the data of Ward et al. (1991) in Section 1.4, if there is no recombination and each site has been hit only once by mutation, then at most three of the four gametic types AB, Ab, aB, and ab can be present. Thus, if we assume that each site has been hit only once by mutation, then there must have been a recombination in between i and j. To get a lower bound, R_M, on the number of recombinations that have occurred, Hudson and Kaplan (1985) set $d(i,j) = 1$ when all four gametes are present and $d(i,j) = 0$ otherwise. To compute R_M, we represent the (i,j) with $d(i,j) = 1$ as an open interval and apply the following algorithm:

- Delete all (m,n) that contain another interval (i,j).
- Let (i_1, j_1) be the first interval not disjoint from all the others. If (m,n) has $i_1 < m < j_1$, then delete (m,n). Repeat until done.

Following Hudson and Kaplan (1985), we will analyze Kreitman's (1983) data on the alcohol dehydrogenase locus of *Drosophila melanogaster*. In the data set, F and S indicate the fast and slow alleles that are caused by the substitution at position 32. The first sequence is a reference sequence, the other 11 are the data. Here we have ignored the six sites in Kreitman's data at which there have been insertions and/or deletions.

```
                 1111111111222222222233333333334444
    12345678901234567890123456789012345678901234567890123
ref CCGCAATATGGGCGCTACCCCCGGAATCTCCACTAGACAGCCT
 1S ........AT........TT.ACA.TAAC..............
 2S ..C..............TT.ACA.TAAC..............
 3S ...................................A....T.A
 4S ................GT.................A..TA...
 5S ...AG...A.TC..AGGT..................C......
 6S ..C............G..............T.T.CAC....T.
 1F ..C............G.............GTCTCC.C......
 2F TGCAG...A.TCG..G.............GTCTCC.CG.....
 3F TGCAG...A.TCG..G.............GTCTCC.CG.....
 4F TGCAG...A.TCG..G.............GTCTCC.CG.....
 5F TGCAGGGGA....T.G....A...G....GTCTCC.C......
```

It should be clear from the first step in the algorithm that for each i we only have to locate $j_i = \min\{j > i : d(i,j) = 1\}$, because the other intervals with left endpoint i will contain (i, j_i). An example should help clarify the procedure. The first stage is to find (i, j_i) for $1 \le i \le 43$. In doing this, we can ignore sites 6, 7, 8, 10, 14, 15, 21, 25, and 39–43, which have only one mutation since these cannot be part of a pair with $d(i,j) = 1$. Note that 13, 19–30, 32, 34, 37, and 38 are consistent with all of the rows that followed, so they do not produce intervals. The remaining (i, j_i) pairs are as follows:

		(1,11)	(2,11)	**(3,4)**
		(4,17)	(5,17)	**(9,16)**
		(11,17)	(12,17)	**(16,17)**
(17,36)	(18,36)	(31,36)	(33,36)	**(35,36)**
				(36,37)

On each row, the interval at the end in bold-faced type is a subset of the previous intervals. The five intervals in boldface are disjoint, so $R_M = 5$. This conflicts with the value $R_M = 4$ reported by Hudson and Kaplan (1985); however, one can verify the five comparisons in the table above.

By looking at the data more closely, one can infer a larger number of recombinations. To make this easier to see, we delete sequences 3F and 4F, which are identical to 2F, and then delete the columns with singletons:

```
       |   11|1|11122222222333333|3|3
   123|45912|6|78902346789012345|6|7
1:1S ...|..A..|.|..TTACATAAC......|.|.
2:2S ..C|.....|.|..TTACATAAC......|.|.
3:3S ...|.....|.|.................|A|.
4:4S ...|.....|.|GT..............|A|.
5:5S ...|AGATC|G|GT..............|.|C
6:6S ..C|.....|G|...........T.T.C|A|C
7:1F ..C|.....|G|...........GTCTCC|.|C
8:2F TGC|AGATC|G|...........GTCTCC|.|C
9:5F TGC|AGA..|G|...........GTCTCC|.|C
```

Myers and Griffiths (2003) improved Hudson's bound by observing that if in a set of S columns there are K haplotypes, then there must be at least $K - S - 1$ recombinations. ($S = 2$ and $K = 4$ gives the four-gamete test.) In the example above, if we look at columns 3, 9, and 12, we see six haplotypes:

.A. C.. ... .AT CAT CA.

so there must have been at least $6 - 3 - 1 = 2$ recombinations in the interval $(3, 12)$.

Given some method of computing lower bounds $b_{i,j}$ on the number of recombinations in the interval (i, j), we can compute a better bound by

$$B_{i,j} = \max_{i<k<j} b_{i,k} + B_{k,j}$$

and a lower bound on the number in whole region by B_{1,S_n}, where S_n is the number of segregating sites. If $b_{i,k} = 1$ when all four gametes are present, this reduces to Hudson and Kaplan's R_M. Myers and Griffiths (2003) developed an estimator R_h where $b_{i,k}$ is computed from the haplotype bound applied to various subsets of $[i, j]$. For this data set it turns out that $R_h = 6$. Bafna and Bansal (2006) further improved the method and showed that there must be at least seven recombinations in this data set.

Song and Hein (2005) have a different approach, which in this case shows that the lower bound of 7 is optimal. The next figure gives trees that are consistent with the mutation patterns in the six regions of the sequence separated by vertical lines. Black dots indicate where recombinations can be introduced to move to the next tree in the seqeunce. Circles indicate the locations of mutations needed in the regions.

While the lower bounds are ingenious, they do not get close to the number of recombinations. Bafna and Bansal (2006) investigated the mean values of Hudsons' R_M, Myers and Griffiths' (2003) program Recmin, and their own bound R_g in comparison with the actual number of recombinations R for various values of the scaled recombination rate $\rho = 4Nr$ in a sample of size 100, when the scaled mutation rate $\theta = 10$.

ρ	1	5	20	50
R_M	1.02	3.03	6.29	9.39
Recmin	1.23	4.88	13.58	24.86
R_g	1.23	5.09	15.80	31.54
R	5.21	27.19	126.45	388.76

An assessment of R_M was done much earlier by Hudson and Kaplan (1985), who used simulations to obtain an estimate from their lower bound. In Kreitman's Adh data, the number of segregating sites is $S = 43$, so $\theta = 4Nu$ can be estimated by

$$S \Big/ \sum_{i=1}^{10} 1/i = \frac{43}{2.928968} = 14.68$$

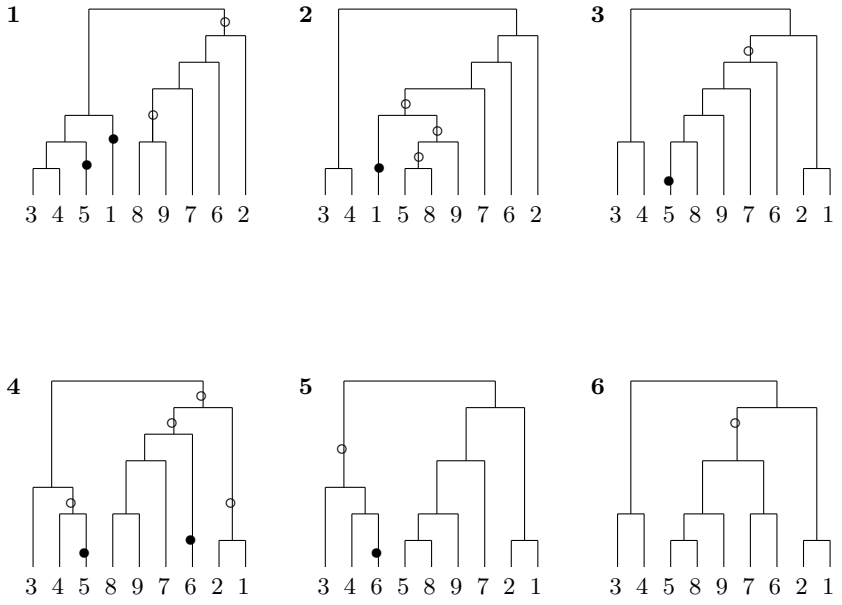

Fig. 3.5. A seven recombination scenario for the Adh data.

If $\theta = 15$, the average values of R_M were 3.7 for $R = 10$, 5.5 for $R = 20$, and 8.8 for $r = 50$. Interpolating suggests $R \approx 18$. To get a confidence interval for R, Hudson and Kaplan (1985) did further simulations with a fixed number of segregating sites to argue that R should be between 5 and 150, which is not very informative.

3.6 Estimating recombination rates

Earlier we saw two not very satisfactory methods of estimating the scaled recombination rate $\rho = 4Nr$. Hudson and Kaplan's (1985) R_M based on the four-gamete test, and its more sophisticated relatives, typically detect only a small fraction of the recombination events that have occurred. Wakeley's (1997) estimator based on the variance of the pairwise differences is easy to calculate but is biased, and not very effective even for samples of size 99.

One of the problems with Wakeley's estimator is that it describes the data with a single summary statistic, ignoring most of the available information. At the other extreme, likelihood approaches estimate the probability of

observing a given data set under an assumed population genetic model. Griffiths and Marjoram (1996), Kuhner, Yamata, and Felsenstein (2000), Nielsen (2000), and Fearnhead and Donnelly (2001) developed estimators of ρ based on likelihood methods. Each of these methods uses computationally intensive statistical methods to approximate the likelihood curve, and even the most efficient of these can accurately estimate this curve only for small data sets.

Given the difficulties of computing the full likelihood, various compromises have been introduced. Fearnhead and Donnelly (2002) split the region of interest into several subregions. They computed the full likelihood for the data in each of the subregions and then multiplied to obtain a composite likelihood. Wall (2000) used simulation to find the parameter values ρ_H and ρ_{HRM} that maximized the likelihood of $(\rho|H)$ and $(\rho|H, R_M)$, where H is the number of haplotypes, and R_M is Hudson and Kaplan's (1985) estimator. Hudson (2001) developed a composite likelihood method by examining pairs of segregating sites. In this section, we will explain his method, but first we consider

3.6.1 Equations for the two-locus sampling distribution

Suppose we sample n chromosomes and obtain information about two loci with alleles A_0 and A_1 and B_0 and B_1. Let n_{ij} be the number of sampled chromosomes that carry allele A_i at locus a and B_j at locus b. Let $q(\mathbf{n};\theta,\rho)$ be the probability of observing $\mathbf{n} = (n_{00}, n_{10}, n_{01}, n_{11})$ for the indicated parameters. The first sign of trouble is the large number of probabilities we have to compute. The number of vectors $\mathbf{n}$ with $n_{00}+n_{10}+n_{01}+n_{11} = n$ and $n_{ij} \geq 0$ is the same as the number of ways of writing $n+4 = m_{00}+m_{10}+m_{01}+m_{11}$ with $m_{ij} = n_{ij}+1 \geq 1$, which, by an earlier calculation, is $\binom{n+3}{3}$. For $n = 10$ this is

$$\frac{13 \cdot 12 \cdot 11}{3!} = 286$$

For $n = 40$ it is 12,341.

Golding (1984) was the first to develop a recursive equation for these probabilities. Here, we will follow the approach of Ethier and Griffiths (1990). As in the case of the recursion of Pluzhnikov and Donnelly (1996) given in Theorem 3.2, in order to obtain a closed set of equations we must consider a larger class of probabilities that allow some members of the sample to be specified only at one location. Let c_{ij} be the number of (A_i, B_j) chromosomes (sampled at both loci with allele A_i at locus a and B_j at locus b), let a_i be the number of $(A_i, \cdot)$ chromosomes (sampled only at the a locus with allele A_i), and let b_j be the number of $(\cdot, B_j)$ chromosomes (sampled only at the b locus with allele B_j). In what follows, we will assume that it is not known which of the alleles is ancestral.

Despite its enormous size, the recursion is straightforward to write. We use a subscript $\cdot$ to indicate a variable that has been summed over. Let $a_\cdot = \sum_i a_i$, $b_\cdot = \sum_j b_j$, $c_{\cdot j} = \sum_i c_{ij}$, $c_{i\cdot} = \sum_j c_{ij}$ and $c_{\cdot\cdot} = \sum_{ij} c_{ij}$. $n_a = a_\cdot + c_{\cdot\cdot}$ is the sample size at the a locus, $n_b = b_\cdot + c_{\cdot\cdot}$ is the sample size at the b locus, and

$n = a. + b. + c..$ is the total sample size. If $n_a = 1$ or $n_b = 1$, there is only one locus and probabilities can be computed from the folded site frequency spectrum.

Jumps happen at total rate

$$\lambda_{a,b,c} = [n(n-1) + \rho c.. + \theta_A(a. + c..) + \theta_B(b. + c..)]/2$$

due to coalescence, recombination, and mutation. Letting $e_i \in R^2$ and $e_{ij} \in R^4$ be vectors with 1's in the indicated position and 0 otherwise, the jumps due to coalescence and recombination occur as follows:

$$\begin{array}{cc} (a - e_i, b, c) & a_i(a_i - 1 + 2c_{i\cdot})/2 \\ (a, b - e_j, c) & b_j(b_j - 1 + 2c_{\cdot j})/2 \\ (a, b, c - e_{ij}) & c_{ij}(c_{ij} - 1)/2 \\ (a - e_i, b - e_j, c + e_{ij}) & a_i b_j \\ (a + e_i, b + e_j, c - e_{ij}) & \rho c_{ij}/2 \end{array}$$

To check these, note that $(A_i, \cdot)$ chromosomes can only coalesce with $(A_i, \cdot)$ or (A_i, B_j) chromosomes. Finally, as in the dual of the Wright-Fisher model with mutation, mutations can kill lineages. However, since we assume that all mutants were created by one mutation, jumps can only occur when there is one of the type left:

$$\begin{array}{cc} (a, b + e_j, c - e_{ij}) & \theta_A/2 \text{ if } a_i = 0,\ c_{ij} = 1,\ c_{i\cdot} = 1 \\ (a - e_i, b, c) & \theta_A/2 \text{ if } a_i = 1,\ c_{i\cdot} = 0 \\ (a + e_i, b, c - e_{ij}) & \theta_B/2 \text{ if } a_j = 0,\ c_{ij} = 1,\ c_{\cdot j} = 1 \\ (a, b - e_j, c) & \theta_B/2 \text{ if } a_i = 1,\ c_{i\cdot} = 0 \end{array}$$

Other mutations kill the chain because they create a configuration not consistent with the pattern we are looking for.

To illustrate the use of these equations, we will consider the simplest possible situation, $n = 2$.

Theorem 3.12. *Let x_1, x_2 and y_1, y_2 be the alleles present in our sample at the two loci. If we let $\theta = \theta_A + \theta_B$,*

$$q(2) = \frac{2(3+\theta)(6+\theta) + \rho[(2+\theta) + 2(6+\theta)\phi] + \rho^2\phi}{2(1+\theta)(3+\theta)(6+\theta) + \rho(2+\theta)(13+3\theta) + \rho^2(2+\theta)}$$

and $\phi = 1/(1+\theta_A) + 1/(1+\theta_B)$, then

$$\begin{aligned} P(x_1 = x_2, y_1 = y_2) &= q(2) \\ P(x_1 = x_2, y_1 \neq y_2) &= (1+\theta_A)^{-1} - q(2) \\ P(x_1 \neq x_2, y_1 = y_2) &= (1+\theta_B)^{-1} - q(2) \\ P(x_1 \neq x_2, y_1 \neq y_2) &= 1 - (1+\theta_A)^{-1} - (1+\theta_B)^{-1} - q(2) \end{aligned}$$

This agrees with (2.9) in Ethier and Griffiths (1990). As a check, we note that when $\rho = 0$ this is $1/(1+\theta)$, while if we let $\rho \to \infty$, the limit is

$$\frac{\phi}{2+\theta} = \frac{1}{1+\theta_A+1+\theta_B} \cdot \left(\frac{1}{1+\theta_A} + \frac{1}{1+\theta_B} \right) = \frac{1}{1+\theta_A} \cdot \frac{1}{1+\theta_B}$$

the answer for independent loci.

Proof. One locus results imply $P(x_1 = x_2) = (1+\theta_A)^{-1}$ and $P(y_1 = y_2) = (1+\theta_B)^{-1}$, so the second and third equations follow from the first one. The fourth one then follows from the fact that the four probabilities add to 1. Thus, it is enough to show that $P(x_1 = x_2, y_1 = y_2) = q(2)$.

The event $x_1 = x_2$, $y_1 = y_2$ occurs if and only if there is no mutation before coalescence, so we use the three states from calculations with Section 3.1: (0,0,2), (1,1,1), and (2,2,0), which we will abbreviate by giving their third coordinates. When we do this, the total transition rates are

$$\lambda_2 = 1+\theta+\rho \qquad \lambda_1 = 3+\theta+\frac{\rho}{2} \qquad \lambda_0 = 6+\theta$$

For states with $(n_a, n_b) = (2,1)$, $(1,2)$, or $(1,1)$, the two-locus homozygosities are $1/(1+\theta_A)$, $1/(1+\theta_B)$, and 1, respectively. Using this with the transition rates, and letting $\phi = 1/(1+\theta_A) + 1/(1+\theta_B)$, we have

$$\begin{aligned} q(2) &= \frac{\rho}{\lambda_2} q(1) + \frac{1}{\lambda_2} \\ q(1) &= \frac{1}{\lambda_1} q(2) + \frac{\rho/2}{\lambda_1} q(0) + \frac{1}{\lambda_1}\phi \\ q(0) &= \frac{4}{\lambda_0} q(1) + \frac{1}{\lambda_0}\phi \end{aligned} \tag{3.17}$$

Inserting the third equation into the second, we have

$$q(1) = \frac{1}{\lambda_1} q(2) + \frac{2\rho}{\lambda_0\lambda_1} q(1) + \frac{\rho+2\lambda_0}{2\lambda_0\lambda_1}\phi$$

Rearranging, we have

$$q(2) = \frac{\lambda_0\lambda_1 - 2\rho}{\lambda_0} q(1) - \frac{\rho+2\lambda_0}{2\lambda_0}\phi$$

The first equation in (3.17) implies $q(1) = (\lambda_2/\rho)q(2) - 1/\rho$. Inserting this in the previous equation, we have

$$q(2) = \frac{\lambda_0\lambda_1\lambda_2 - 2\rho\lambda_2}{\rho\lambda_0} q(2) - \frac{\lambda_0\lambda_1 - 2\rho}{\rho\lambda_0} - \frac{\rho+2\lambda_0}{2\lambda_0}\phi$$

Rearranging gives

$$\frac{\lambda_0\lambda_1\lambda_2 - \rho\lambda_0 - 2\rho\lambda_2}{\rho\lambda_0} q(2) = \frac{\lambda_0\lambda_1 - 2\rho}{\rho\lambda_0} + \frac{\rho+2\lambda_0}{2\lambda_0}\phi$$

and hence we have

$$q(2) = \frac{2\lambda_0\lambda_1 - 4\rho + (\rho^2 + 2\lambda_0\rho)\phi}{2[\lambda_0\lambda_1\lambda_2 - \rho\lambda_0 - 2\rho\lambda_2]} \tag{3.18}$$

Recalling the definitions of the λ_k, we see that the denominator is

$$\begin{aligned} &= 2(1+\theta)(3+\theta)(6+\theta) + \rho(1+\theta)(6+\theta) + 2\rho(3+\theta)(6+\theta) + \rho^2(6+\theta) \\ &\quad - 2\rho(6+\theta) - 4\rho(1+\theta+\rho) \end{aligned}$$

The coefficient of ρ^2 in the denominator is $(6+\theta) - 4 = 2+\theta$, while that of ρ is

$$\begin{aligned} (1+\theta+6+2\theta-2)(6+\theta) - 4(1+\theta) &= (3\theta+5)(\theta+6) - 4 - 4\theta \\ &= 3\theta^2 + 19\theta + 26 = (3\theta+13)(\theta+2) \end{aligned}$$

The numerator of the formula for $q(2)$ in (3.18) is

$$2(3+\theta)(6+\theta) + \rho(6+\theta) - 4\rho + \rho^2\phi + 2(6+\theta)\rho\phi$$

Combining our computations gives

$$q(2) = \frac{2(3+\theta)(6+\theta) + \rho[(2+\theta) + 2(6+\theta)\phi] + \rho^2\phi}{2(1+\theta)(3+\theta)(6+\theta) + \rho(2+\theta)(13+3\theta) + \rho^2(2+\theta)}$$

which proves the desired result. □

3.6.2 Simulation methods

Given the analytical complications of Golding's (1984) equation, Hudson (1985) used a clever simulation technique to estimate the probabilities. He generated a two-locus genealogy of a sample of n chromosomes, but then instead of generating a single sample from a pair of trees, he calculated the distribution of $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$ conditional on the pair of trees. To explain this we need some notation. Let $\mathcal{E}$ be the sequence of coalescence and recombination events that define the tree. Given the sequence of events E_i, the time T_i between E_{i-1} and E_i has an exponential distribution with mean determined by the configuration of ancestral lineages during the interval.

Let τ_A and τ_B be the total length of the trees, measured in units of $4N$ generations so that the scaled mutation rate is θ. Let $I(\mathcal{E}, \mathbf{n}, j, k) = 1$ if mutations on the jth branch of the a locus tree and on the kth branch of the b locus tree would produce the sample configuration $\mathbf{n}$, and 0 otherwise. As the notation indicates, this depends on the event sequence $\mathcal{E}$ and not on the time sequence $\mathcal{T}$. Letting a_j and b_k be the length of the two branches, the probability of the sample configuration $\mathbf{n}$ being produced is

$$I(\mathcal{E}, \mathbf{n}, j, k)(1 - e^{-\theta a_j})(1 - e^{-\theta b_k})e^{-\theta(\tau_A - a_j)}e^{-\theta(\tau_B - b_k)}$$

since we need at least one mutation on each of the selected branches and none on the rest of the trees. Note that the identification of the branches that

produce the desired mutation pattern depends on whether or not we consider A_0 and B_0 to be the ancestral alleles and A_1 and B_1 to be derived, which Hudson calls $a - d$ specified samples, but this is otherwise irrelevant to the details of the procedure.

We will apply this formula when the a and b loci are single nucleotides, so θ is small and the above is

$$\approx \theta^2 I(\mathcal{E}, \mathbf{n}, j, k) a_j b_k$$

Summing over j and k and then taking expected value with respect to the joint distribution of the $\mathcal{E}$ and $\mathcal{T}$ sequences, we have

$$q_u(\mathbf{n}; \theta, \rho) \approx \theta^2 h_u(\mathbf{n}; \rho)$$

where $h_u(\mathbf{n}; \rho) = E\left(\sum_{j,k} I(\mathcal{E}, \mathbf{n}, j, k) a_j b_k\right)$. The subscript u indicates that this is the unconditioned version of the probability, i.e., we are not conditioning on the event that the two sites are variable.

To estimate $h_u(\mathbf{n}; \rho)$, we generate a sequence of genealogies with events $\mathcal{E}_i$, $1 \le i \le m$, and let

$$\hat{h}_u(\mathbf{n}; \rho) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j,k} I(\mathcal{E}_i, \mathbf{n}, j, k) a_j(i) b_k(i)$$

In the case of a constant-size population model, this can be done more efficiently by using

$$\tilde{h}_u(\mathbf{n}; \rho) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j,k} I(\mathcal{E}_i, \mathbf{n}, j, k) E(a_j b_k | \mathcal{E}_i)$$

In words, we replace the observed times by their conditional expectation given the event sequence $\mathcal{E}_i$. This reduces the variance of the estimator and eliminates the need to simulate the time sequence.

Most applications of the two-locus sampling distribution will focus on pairs of sites in which both are polymorphic in the sample. That is, we must consider the probability

$$q(\mathbf{n}, \theta, \rho | \text{two alleles at each locus}) = \frac{q_u(\mathbf{n}; \theta, \rho)}{\sum_{\mathbf{m}} q_u(\mathbf{m}; \theta, \rho)}$$

Again, we are interested in

$$q_c(\mathbf{n}, \rho) = \lim_{\theta \to 0} q(\mathbf{n}, \theta, \rho | \text{two alleles at each locus}) = \frac{h_u(\mathbf{n}; \rho)}{\sum_{\mathbf{m}} h_u(\mathbf{m}; \rho)}$$

which we can estimate without specifying θ. Hudson has done this for sample sizes 20, 30, 40, 50, and 100 and a range of ρ values between 0 and 100, and these are available on his web page.

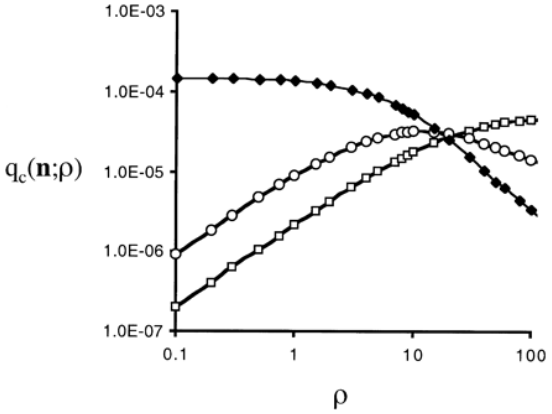

Fig. 3.6. Likelihood curves for three sample configurations.

3.6.3 Composite likelihood estimation of ρ

The graph above, which is Figure 2 in Hudson (2001), shows that if we try to use one pair of sites to estimate ρ, then the likelihood curves are often monotonic, resulting in estimates of 0 or ∞. Consider, for the moment, the mythical situation that we have k independent pairs of loci, each with scaled recombination rate ρ. The overall likelihood is

$$L(\mathbf{n}_1, \mathbf{n}_2, \ldots \mathbf{n}_k; \rho) = \prod_{i=1}^{k} q_c(\mathbf{n}_i; \rho)$$

and this is maximized to produce the estimate $\hat{\rho}$. To characterize statistical properties of $\hat{\rho}$, we consider

$$E_{\rho_0}(\log q_c(\mathbf{n}; z)) = \sum_{\mathbf{n}} q_c(\mathbf{n}; \rho_0) \log q_c(\mathbf{n}; z)$$

which is the expected log-likelihood given that the true parameter is ρ_0. The second derivative of this function with respect to z, evaluated at ρ_0 is inversely proportional to the asymptotic variance of the maximum likelihood estimate. More precisely, for a sample of size k

$$\mathrm{var}_{\rho_0,k}(\hat{\rho}) \approx \frac{-1}{k(\partial^2/\partial z^2) E_{\rho_0}(\log q_c(\mathbf{n}; z))|_{z=\rho_0}}$$

The next graph, which is Figure 5 in Hudson (2001), shows $E_{\rho_0}(\log q_c(\mathbf{n}; z))$ and a quadratic function fitted to several points near $\rho_0 = 5$, which suggests that this approximation may be accurate when k is not large.

Since it is trivial that smaller values of ρ can be estimated with less absolute error than larger values, it is interesting to look instead at the coefficient of variation:

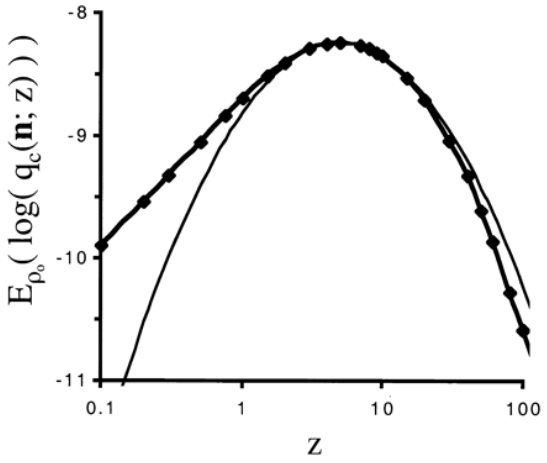

Fig. 3.7. Expected log-likelihood curve and its quadratic approximation.

$$\frac{\mathrm{var}_{\rho_0,k}(\hat{\rho})}{\rho_0^2} \approx \mathrm{var}_{\rho_0,k}(\log(\hat{\rho})) \approx \frac{-1}{k(\partial^2/\partial(\log z)^2)E_{\rho_0}(\log q_c(\mathbf{n}; z))|_{z=\rho_0}}$$

Figure 6 of Hudson (2001), which is given here as Figure 3.8, shows this quantity as a function of ρ_0. It achieves a minimum at $\rho_0 = 5$ and to quote Hudson, "shows that pairs separated by ρ in the range $2 - 15$ are best for estimating ρ."

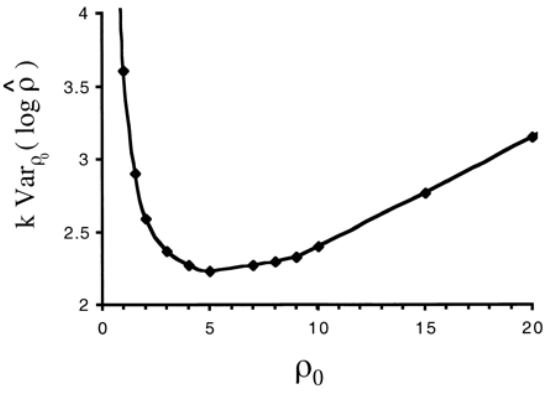

Fig. 3.8. Asymptotic variance of the log of the MLE

In practice, different pairs of polymorphic sites will be different distances apart and hence will have different recombination rates. Letting ρ_b be the recombination probability per base pair and d_i the distance between the ith pair, we can write the likelihood as

$$L(\mathbf{n}_1, \mathbf{n}_2, \ldots \mathbf{n}_k; \rho_b) = \prod_{i=1}^{k} q_c(\mathbf{n}_i; \rho_b d_i)$$

To define his composite likelihood estimator ρ_{CL}, Hudson uses the log-likelihood:

$$\sum_{i<j} \log q_c(\mathbf{n}_{ij}; \rho_b d_{ij})$$

where $\mathbf{n}_{ij}$ is the observed sampling distribution at sites i and j and d_{ij} is the distance between the two sites. Hudson investigated ρ_{CL} by simulation and showed that it and Wall's estimator ρ_{HRM} performed much better than Wakeley's moment estimator ρ_{wak} and Hey and Wakeley's (1997) γ.

Example 3.2. Hudson (2001) applied his method to estimating ρ from a survey of human variation on the X chromosome by Taillon-Miller et al. (2000). In this study, 39 SNPs were surveyed in three population samples, but Hudson only considered the sample of 92 CEPH males. The parameter ρ_b was estimated by maximizing the composite likelihood for (i) all 39 SNPs, (ii) the 14 SNPs in Xq25, and (iii) the 10 SNPs in or near Xq28. Here the q refers to the q arm of the X chromosome, and 25 and 28 refer to chromosome bands, which can be observed under microscopes, and in the era before whole genome sequencing was used to describe the locations of genes. For loci on the X chromosome, the methods described above will result in an estimate of $2Nr$ since these loci do not experience recombination in males. Hudson multiplied his computer output by 2 to give an estimate of $4Nr$ and reported estimates of (i) 9×10^{-5}, (ii) 8.8×10^{-5}, and (iii) 9×10^{-5}. These conclusions contrast with those of Taillon-Miller et al. (2000), who observed that linkage disequilibrium was high in Xq25 but almost nonexistent in Xq28.

Fearnhead (2003) investigated the consistency of the composite likelihood estimators of Fearnhead and Donnelly (2002), and the pairwise likelihood of Hudson (2001), when one examines an increasing number of segregating sites for a fixed sample size. He proved that the composite likelihood is consistent and proved the consistency of a truncated pairwise likelihood, which is based on the product of the likelihoods for all pairs of sites that are less than some distance R apart. Smith and Fearnhead (2005) compared the accuracy of these two methods and the pseudolikelihood method of Li and Stephens (2003) using simulated sequence data. They found that the performance was similar but that the pairwise likelihood method could be improved by including contributions to the log-likelihood only for pairs of sites that are separated by some prespecified distance.

3.7 Haplotypes and hot spots

Sequencing of the human genome revealed that there are sizable regions over which there is little evidence of recombination and a small number of SNPs

is sufficient to describe most of the genetic variation. For example, Patil et al. (2001) found that 80% of human chromosome 21 variation can be described by only three SNPs per block. The average observed block length was 7.8 kb, but the longest block stretched more than 115 kb and contained 114 SNPs. Results of Daly et al. (2001) on a 500 kb region on chromosome 5q31, implicated as containing a genetic risk factor for Crohn's disease, showed large haplotype blocks with limited diversity punctuated by apparent sites of recombination. Gabriel et al. (2002) studied haplotype patterns across 51 autosomal regions and showed that the human genome could be parsed objectively into haplotype blocks.

Reich et al. (2002) argued that the simplest explanation for LD is that the population under study experienced an extreme founder effect or bottleneck, and that a severe bottleneck occurring 800-1600 generations ago could have generated the LD they observed. Frisee et al. (2001) argued that the higher levels of LD in European populations, when compared to African populations, support a contribution from the bottleneck in the founding of non-African populations. However, Ardlie, Kruglyak, and Seielstad (2002) carried out simulations of several models of demographic history, including bottlenecks and expansion, and found that any model with nucleotide diversity in the empirically observed range had useful levels of LD limited to approximately 10 kb. In the other direction, Anderson and Slatkin (2004) argued that rapid population growth could account for the patterns observed in the data from 5q31. For more on the patterns of LD found in various population genetics models, see Pritchard and Przeworski (2001).

One possible explanation for the observed haplotype blocks is that recombinations are not uniformly spread along a chromosome but preferentially occur in hot spots. Analysis of recombination breakpoints and crossover events in sperm typing experiments have demonstrated the presence of recombination hot spots in several genomic locations. The first work was done by Jeffreys, Ritchie, Neumann (2000) and Jeffreys, Kauppi, Nuemann (2001). For a survey, see Kauppi, Jeffreys, and Keeney (2004).

Since experimental confirmation of hot spots is technically challenging and time consuming, Li and Stephens (2003) introduced statistical procedures based on a heuristic formula for haplotype probabilities to detect fine scale variation in recombination rates from population data. Using a method based on Hudson's (2001) two-sample distributions, McVean et al. (2004) found evidence of rate variation spanning four orders of magnitude and suggested that 50% of all recombinations take place in less than 10% of the sequence.

Myers et al. (2005) followed up on this work applying the method to 1.6 million SNPs genotype in samples from three samples: 24 European Americans, 23 African Americans, and 24 Han Chinese from Los Angeles by Perlegen, see Hinds et al. (2005). They estimated that there is a hotspot every 50 kb or so in the human genome, with approximately 80% of recombinations occuring in 10 to 20% of the sequence. With more than 25,000 hotspots at their disposal, Myers et al. (2005) also found the first set of sequence features

substantially overrepresented in hot spots relative to cold spots. Although the motifs are neither necessary nor sufficient for hot spot activity, their top scoring candidate ($CCCTCCCT$) played a role in 11% of them.

While recombination hot spots exist and will produce haplotype blocks, it remains unclear how much rate variation, if any, is needed to account for observed haplotype blocks. Phillips et al. (2003) found that only about one third of chromosome 19 was covered in haplotype blocks and that there was no reason to invoke recombination hot spots in order to explain the observed blocks. Wang et al. (2002) and Zhang et al. (2003) found through extensive coalescent simulations that haplotype blocks were observed in models where recombination crossovers were randomly and uniformly distributed.

A second mystery is that despite 99% identity between human and chimpanzee DNA sequences, ther is virtually no overlap between these two species in the locations of their hot spots. Ptak et al. (2005) showed that the well studied TAP2 hotspot in humans was absent in chimpanzees. Winckler et al. (2005) found that 18 recombination hot spots covering 1.5 megabases in humans were absent in chimpanzees and observed no correlation between estimates of fine scale recombination rates.