THE TWO-LOCUS ANCESTRAL GRAPH

R. C. Griffiths
Mathematics Department
Monash University

Abstract

In a population genetics two-locus model with recombination an offspring has either a single parent gene, or is a recombinant from two parent genes. The number of ancestors, backward in time, of a sample of genes can thus decrease or increase and is found to be a birth and death process. Instead of a one-locus ancestral tree the ancestral paths of a sample of gene pairs are described by a graph with leaves as the sample genes and an eventual common ancestor where all paths from the leaves lead. In this paper properties of the two-locus ancestral graph and the two marginal ancestral trees are studied.

1. Introduction. In population genetics the ancestral tree of a sample of genes plays an important role in a probabilistic description of the sample. Kingman (1982) studies the *coalescent process* which describes the relationship between a sample of genes and their ancestors. The ancestral tree of a sample in a large population can be described by a death process where ancestors coalesce at a rate of $r(r-1)/2$. A tree is constructed by beginning with leaves at the n edges and joining edges where ancestors coalesce, keeping edge lengths proportional to the times between coalescence. The root is the first common ancestor of the sample. Mutations occur (to ancestors) along the edges of the tree according to a Poisson process of rate $\theta/2$ and determine the allelic configuration in the sample. The infinitely-many-alleles model is characterized by every mutation producing a type entirely new to the population. A nice review article on ancestral trees is Tavare (1984).

In a multi-locus model with recombination the relationship between a sample of genes and their ancestors is complicated because of recombination splitting up genes.

A gene in the two-locus model is described by an ordered pair $(x, y) \in [0,1] \times [0,1]$. The allele types at locus A and locus B are x and y. Stochastic evolution of the population is described by a discrete-time Wright-Fisher model. There are a fixed number of M genes in each generation. Genes in generation $\tau+1$ are produced from those in generation τ in the following way. Choose a single parent at random from the preceeding generation with

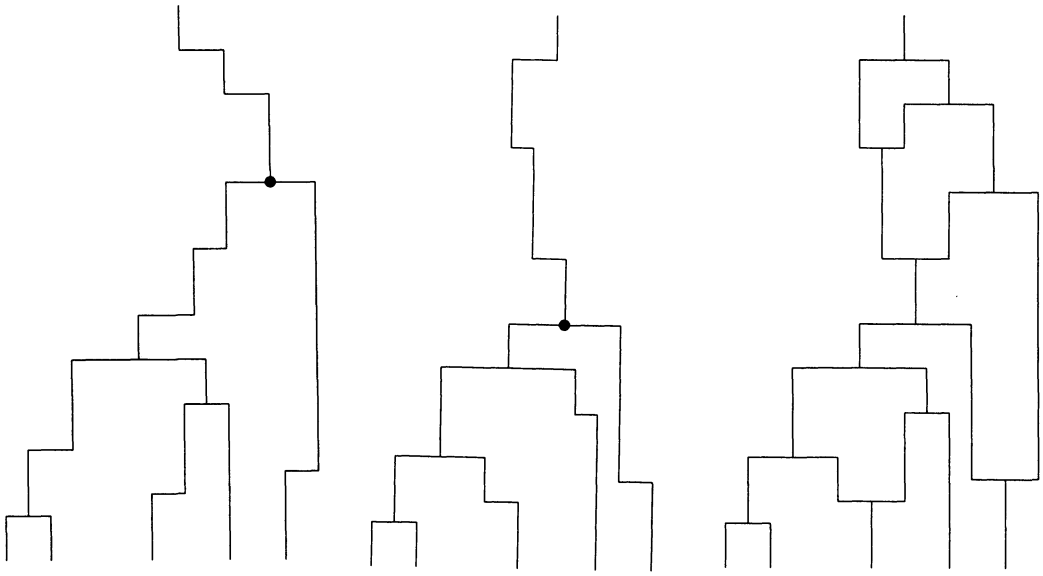
probability 1 - r_M. With probability r_M, the parent is formed by recombining two genes chosen at random from the preceeding generation; if the first chosen is (α, β) and the second chosen is (γ, δ), then the parent is of type (α, δ). If the parents type is (x, y) then the offspring's type is (x', y'), where $x' = x$ with probability 1 - u_A or $x' = Z$, a uniform random variable on [0, 1], with probability u_A, and similarly for y'. (Z is considered a member of an independent, identically distributed sequence of random variables where different mutations are represented by distinct random variables.) The labeling by uniform random variables is just a convenient way of achieving unique labels for mutations in the infinitely-many-alleles model.

A diffusion time scaling is to measure time in units of M generations and let $M \to \infty$ while $\theta_A = 2Mu_A$, $\theta_B = 2Mu_B$, $\rho = 2Mr_M$ are held constant. In Ethier and Griffiths (1990a) the Wright-Fisher model is set up as a measure-valued Markov process which converges to a measure-valued diffusion. An atom of the measure at (x, y) represents the relative frequency of that type in the population. The convergence is robust under a number of different models (eg. Moran model). Our interest here is in describing the process of a sample of n genes' ancestors backwards in time, rather than in convergence of the process. Events occurring between the sample and the sample's parents in the preceeding generation (to order M^{-1} in probability) are: mutation to a gene in the sample at the A locus, with probability nu_A; mutation to a gene in the sample at the B locus, with probability nu_B; coalescence, when the sample has n - 1 parents, with probability $n(n - 1)/2M$; and recombination, when one gene in the sample is constructed from two in the preceeding generation, with probability nr_M.

Let $n(t)$ be the number of ancestors at time t backwards of a sample of $n(0) = n$ genes in the limit process. This includes recombinant ancestors. Then $\{n(t); t \geq 0\}$ is a birth and death process with rates $\lambda_k = k\rho/2$, $\mu_k = k(k - 1)/2$. Let W_n be the waiting time to first visit state 1. W_n has a proper distribution because of the quadratic death and linear birth rates. It is convenient to stop the process at this time, since the genetic composition of the sample is determined by then. If the process was continued it would have a modified Poisson stationary distribution $\{(e^{\rho} - 1)^{-1} \rho^j/j!, j = 1,2,...\}$.

2. Two-locus Ancestral Graph. A *two-locus ancestral graph* is drawn proportional to a real-time scale starting from n end edges, then joining edges at a vertex when two ancestors of the sample coalesce, or appending two edges at a

Figure 1
Two-locus ancestal graph



Ancestal graph

Ancestral tree
Locus A

Ancestral tree
Locus B

vertex to represent ancestors when recombination occurs. An example is shown in Figure 1. By convention at recombination the genes represented by the left upper edge and the lower edge have the same type at the A locus, and similarly for the right edge and the B locus. Probabilistically the graph is described by choosing two edges at random to be joined when a death occurs in the process $\{n(t);\ t \geq 0\}$, and choosing an edge at random to append two edges to when a birth occurs. Mutations occur along the edges of the graph (given the edge lengths) according to a Poisson process of rate $\theta/2 = (\theta_A + \theta_B)/2$, with probabilities θ_A/θ and θ_B/θ of being on locus A or locus B. Knowing the graph and where mutations have occurred along the edges completely determines the configuration of allele pairs in the sample of n genes. Locus A's marginal coalescent tree is found by tracing a path from the n genes upward in the graph, always taking the left path at recombination events, similarly for locus B, taking the right path. The marginal common ancestors may occur in the graph before the (first) common ancestor of the pairs. Figure 1 illustrates the marginal trees in a graph. First marginal ancestors are shown as dots. If $\rho = 0$ there is no recombination and the common ancestors are the same. As $\rho \to \infty$ the waiting time to the common ancestor of the pairs tends to infinity, but the marginal times to common ancestors remain finite, their distribution not depending on ρ.

Hudson (1983), Kaplin and Hudson (1985) use genealogical methods in studying an m-loci model with recombination, where they consider the collection of m correlated family trees.

In a single-locus model the waiting time T_n until there is a common ancestor of n genes is a sum of mutually independent exponential random variables with rates

$n(n-1)/2, \ldots, 1$. The density of T_n is known (see Tavare (1984)) and

$$E(T_n) = \sum_{r=2}^{n} 2/(r(r-1)) = 2(1-n^{-1}).$$

As $n \to \infty$ the distribution of T_n converges to a proper distribution.

Theorem 1. Let W_n be the waiting time until there is a common ancestor of a sample of n genes in the two-locus ancestral graph, then

$$E(W_n) = 2\rho^{-1}\int_0^1 \frac{1-v^{n-1}}{1-v}(e^{\rho(1-v)}-1)\,dv. \tag{2.1}$$

Proof. By considering the waiting time until the first event in the birth and death process with rates $\lambda_r = r\rho/2$, $\mu_r = r(r-1)/2$, $r = 1,2,\ldots$, $E(W_1) = 0$, and

$$E(W_n) = \frac{2}{n(n+\rho-1)} + \frac{n-1}{n+\rho-1}E(W_{n-1}) + \frac{\rho}{n+\rho-1}E(W_{n+1}) \tag{2.2}$$

$n = 2,3,\ldots$.

Couple the birth and death process with one where there is a reflecting barrier at $b > 1$ by deleting all excursions $\geq b$. Denote W_n^b as the waiting time to 1 in the modified process. $E(W_n^b)$ satisfies a similar recursion to above, but for $n = 2,3,\ldots, b-1$, and $E(W_b^b) = E(W_{b-1}^b)$. It is straightforward to solve the modified system of equations and obtain

$$E(W_n^b) = 2\sum_{k=2}^{n}(k-2)!\sum_{j=0}^{b-1-k}\rho^j/(j+k)!$$

$n = 2,\ldots, b-1$.

The limit as $b \to \infty$ is

$$E(W_n) = 2\sum_{k=2}^{n}\sum_{j=0}^{\infty}\rho^j\frac{\Gamma(k-1)\Gamma(j+2)}{\Gamma(j+k+1)(j+1)!}$$

$$= 2\rho^{-1}\int_0^1\frac{1-v^{n-1}}{1-v}(e^{\rho(1-v)}-1)\,dv. \quad \square$$

Extreme cases of (2.1) are $E(W_n) = 2(1 - n^{-1})$, when $\rho \to 0$ and $E(W_n) \to \infty$ when $\rho \to \infty$. A limit formula is

$$E(W_\infty) = 2\rho^{-1}\int_0^1\frac{e^{\rho u}-1}{u}du.$$

The waiting time W_n can be decomposed into the waiting times spent in the states 2,3, ..., before the process is absorbed at 1.

Theorem 2. Let W_{nj} be the waiting time spent while there are j ancestors of a sample of n genes in the two-locus ancestral graph, then $W_n = \Sigma_{j=2}^{\infty}W_{n,j}$, and

$$E(W_{n,j}) = 2\sum_{k=2}^{\min(j,n)}(k-2)!\,\rho^{j-k}/j!, \tag{2.3}$$

$j = 2,3,\ldots$

Proof. This is similar to the proof of Theorem 1. $E(W_{n,j})$ satisfies a similar system of equations to (2.2), but with the first term on the right multiplied by the Kronecker delta $\delta_{n,j}$. □

If $\rho = 0$, then $E(W_{n,j}) = 2/(j(j-1))$, agreeing with known results, since then $W_{n,j}$ has an exponential distribution. $E(W_{n,2}) = 1$, not depending on ρ. As $\rho \to \infty$, $E(W_{n,j}) \approx 2\rho^{j-2}/(j(j-1))$. If $n \to \infty$, the upper summation limit in (2.3) is replaced by j.

It is of interest to consider the maximum number of ancestors of a sample that there can be at any time instant before a common ancestor in the two-locus ancestral graph.

Theorem 3. Let M_n be the maximum number of edges in a cross section of a two-locus ancestral graph of a sample of n genes, then

$$P(M_n \le k) = \frac{\Sigma_{j=n-1}^{k-1} j!/\rho^j}{\Sigma_{j=0}^{k-1} j!/\rho^j}, \tag{2.4}$$

$k = n, n+1, \ldots$.

Proof. Denote $p_n(k) = P(M_n \le k)$. Clearly for $k = 1,2 \ldots$,

$$p_n(k) = \frac{n-1}{n+\rho-1} p_{n-1}(k) + \frac{\rho}{n+\rho-1} p_{n+1}(k),$$

$n = 2, \ldots, k$, and $p_1(k) = 1$, $p_n(k) = 0$ if $n > k$. Denote $u_n(k) = p_n(k) - p_{n-1}(k)$, then

$$\rho u_{n+1}(k) = (n-1)\, u_n(k),$$

$$u_{j+1}(k) = \frac{(j-1)!}{\rho^{j-1}} u_2(k), \quad j = 1, \ldots, k.$$

Summing this equation and using $p_{k+1}(k) = 0$,

$$p_{n+1}(k) - 1 = (p_2(k) - 1) \sum_{j=0}^{n-1} j!/\rho^j, \quad p_2(k) = \frac{\Sigma_{j=1}^{k-1} j!/\rho^j}{\Sigma_{j=0}^{k-1} j!/\rho^j} \tag{2.5}$$

Simplifying (2.5) gives (2.4). □

As $n \to \infty$, $n^{-1} M_n \to 1$, in probability, and from (2.4), $P(M_n > n + d) \approx (n^{-1}\rho)^{d+1}$, $d = 0, 1, \ldots$.

The two-locus ancestral graph will now be examined in greater detail. Given a two-locus graph G construct two marginal coalescent trees T_A and T_B. It is convenient to consider the marginal trees up to the common ancestor vertex in G, although the common ancestor vertices in T_A and T_B may occur earlier. Denote the edge set of a graph by $E(\cdot)$. Partition the edges of G into $A = E(G) \cap E(T_A) \cap E(T_B)^c$, $B = E(G)E(T_A)^C \cap E(T_B)$,

$C = E(G) \cap E(T_A) \cap E(T_B)$, $D = E(G) \cap E(T_A)^C \cap E(T_B)^C$.

. As an example all four types of edge occur in Figure 1. Let $\mathrm{E_t}$ be the edges of a cross section of G taken at time t backwards. Denote $n_A(t) = |E_t \cap A|$, $n_B(t) = |E_t \cap B|$, $n_C(t) = |E_t \cap C|$, $n_D(t) = |E_t \cap D|$ and $n(t) = |E_t|$.

Ancestors represented by edges in A do not contain genetic material which has any influence on locus B, similarly for B and locus A, and those represented by D do not contain material influencing either locus A or B.

Our interest is in the Markov process $(n_A(t), n_B(t), n_C(t), n_D(t))$. Think of edges in G as being particles of possible types (1,0), (0,1), (1,1) or (0,0), according as to whether they belong to A, B, C or D. That is, the type of an edge $e \in G$ is $(I_{E(T_A)}(e), I_{E(T_B)}(e))$, where $I(\cdot)$ denotes an indicator function. Two edges of types (α, β), (γ, δ) $((\alpha, \beta, \gamma, \delta) \in \{0, 1\})$, coalesce to $(\alpha \vee \beta, \gamma \vee \delta)$ ($\vee$ denotes logical or). If recombination occurs to an edge of type (α, β) the edge splits into two edges of types $(\alpha, 0)$ and $(0, \beta)$. Once convinced that this particle evolution is consistent with the type definitions in the partition of G it is clear that $(n_A(t), n_B(t), n_C(t), n_D(t))$ is Markovian.

Transition rates are :

$$(a, b, c, d) \rightarrow \begin{cases} (a+1, b+1, c-1, d), & c\rho/2 \\ (a-1, b-1, c+1, d), & ab \\ (a-1, b, c, d), & ac + a(a-1)/2 \\ (a, b-1, c, d), & bc + b(b-1)/2 \\ (a, b, c-1, d), & c(c-1)/2 \\ (a, b, c, d+1), & (a+b+d)\rho/2 \\ (a, b, c, d-1), & d(a+b+c) + d(d-1)/2 \end{cases}$$

with $a + b + c + d > 1$.

The marginal transition rates of $(n_A(t), n_B(t), n_C(t))$ do not depend on $n_D(t)$,

so it too is Markovian.

Transition rates are :

$$(a,b,c) \to \begin{cases} (a+1, b+1, c-1), & c\rho/2 \\ (a-1, b-1, c+1), & ab \\ (a-1, b, c), & ac + a(a-1)/2 \\ (a, b-1, c), & bc + b(b-1)/2 \\ (a, b, c-1), & c(c-1)/2, \end{cases}$$

with $a + b + c > 1$.

This process is used in Ethier and Griffiths (1990a), with a different interpretation, in showing that the two-locus measure-valued diffusion is ergodic. Although $n(t)$ is not a monotonic process, both $n_A(t) + n_C(t)$ and $n_B(t) + n_C(t)$ are non-increasing processes, which are eventually absorbed into 1,1.

It is clear how to extend the two-locus graph to multiple loci, although the extension is not studied here. Consider m loci with a total recombination rate ρ. Given a recombination event, it splits a gene between locus j and $j + 1$ with probability ρ_j/ρ, $j = 1, ..., m - 1, \rho = \rho_1 + ... + \rho_{m-1}$. The ancestral graph is still similar to the illustration in Figure 1. Edges are labeled by elements of $\{0, 1\}^m$. Coalescence is similar to the two-locus case, and if recombination occurs between locus j and $j + 1$ to an edge of type $(\alpha_1, ..., \alpha_m)$ the upper joining left and right edges have respective types $(\alpha_1, ..., \alpha_j, 0, ..., 0)$, $(0, ..., 0, \alpha_{j+1}, ..., \alpha_m)$. Results of Theorems 1,2,3 apply also to multiple loci.

Hudson(1983), Kaplin and Hudson (1985) use the process (2.6) and its extension to m-loci in their study of the homozygosity and number of segregating sites in a sample of genes. They were not concerned with the total graph, but just the correlated trees at the m loci.

A system of equations which is useful in several theorems is the following. Let υ and β be functions on $\mathbf{N} \times \mathbf{N} \times \mathbf{N} \times \mathbf{R}$ satisfying

$$\begin{aligned} \nu(a,b,c;\rho) &= \beta(a,b,c;\rho) + \frac{c\rho}{n(n-1)+c\rho}\upsilon(a+1,b+1,c-1;\rho) \\ &+ \frac{2ab}{n(n-1)+c\rho}\nu(a-1\ \ b-1, c+1;\rho) \ + \frac{a(a+2c-1)}{n(n-1)+c\rho}\nu(a-1,b,c;\rho) \qquad (2.7) \\ &+ \frac{b(b+2c-1)}{n(n-1)+c\rho}\nu(a,b-1,c;\rho) + \frac{c(c-1)}{n(n-1)+c\rho}\nu(a,b,c-1;\rho)\,. \end{aligned}$$

where $n = a + b + c$, $a + c > 1$, $b + c > 1$. The function β and boundary conditions $\nu(1, 1, 0;\rho)$, $\nu(0, 0, 1; \rho)$, $\nu(1, 0, 0; \rho)$, $\nu(0, 1, 0; \rho)$ are assumed to be known.

The system arises by considering the first transition in the process $(n_A(t), n_B(t), n_C(t))$ beginning at (a,b,c). The overall transition rate away from (a,b,c) is $(n(n - 1) + c\rho)/2$. Proofs of Theorems 4,6,7 are straightforward using systems of equations of the form (2.7), so are omitted. Ethier and Griffiths (1990b) show how to solve systems of equations such as (2.7) numerically. Defining the degree of $\nu(a, b, c;\rho)$ as $a + b + 2c$, terms on the right side of (2.7) are of degree less than or equal to $a + b + 2c$. Knowing terms of lesser degree the others can be found by solving a tridiagonal system of equations.

Let $W_{A,n}$ and $W_{B,n}$ be the respective waiting times until a sample of n genes has common ancestors at locus A and locus B.

Theorem 4. Let $\xi(a, b, c;\rho)$ be the expected waiting time until $n_A(t) + n_C(t) = 1$, $n_B(t) + n_C(t) = 1$, beginning with $n_A(0) = a$, $n_B(0) = b$, $n_C(0) = c$.

$E(\max(W_{A,n}, W_{B,b})) = \xi(0, 0, n;\rho)$ is the expected waiting time until a sample of n genes has common ancestors at locus A and locus B. ξ satisfies the system (2.7) with $\beta(a, b, c; \rho) = 2/(n(n - 1) + c\rho)$ and $\xi(1, 1, 0; \rho) = 0$, $\xi(0, 0, 1; \rho) = 0$, $\xi(1, 0, 0; \rho) = 0$, $\xi(0, 1, 0; \rho) = 0$. □

Illustrative values of the waiting time are shown in Table 1 for a sample of size 50.

Table 1

Expected waiting time to marginal common ancestors, n = 50.

ρ	0	0.1	0.5	1.0	2.0	5.0	10.0	20.0	∞
μ	1.96	2.01	2.14	2.25	2.36	2.46	2.50	2.51	2.52

The waiting time until a sample has marginal common ancestors is finite as $\rho \to \infty$, being bounded below by $2(1 - n^{-1})$ and above by $4(1 - n^{-1})$, however the waiting time until a common ancestor in G tends to infinity. $\xi(a, b, c; \rho)$ is monotonic in ρ and therefore bounded by $\xi(a + c, b + c, 0; \infty)$. One argues that if $\rho = \infty$, edge types (1,1) are instantaneously split into a pair (1,0) and (0,1), so $\xi(a, b, c; \infty) = \xi(a + c, b + c, 0; \infty)$. In this case there is an easy recursion,

$$\xi(a,b,0;\infty) = \frac{2}{n(n-1)} + \frac{2ab}{n(n-1)}\xi(a,b,0;\infty)$$

$$+\frac{a(a-1)}{n(n-1)}\xi(a-1,b,0;\infty) + \frac{b(b-1)}{n(n-1)}\xi(a,b-1,0;\infty),$$

$a, b = 1,2, \ldots, n = a + b > 1$ and $\xi(1,1,0;\infty) = 0$.

Simplifying,

$$\xi(a,b,0;\infty) = \frac{2}{a(a-1)+b(b-1)} + \frac{a(a-1)}{a(a-1)+b(b-1)}\xi(a-1,b,0;\infty)$$

$$+\frac{b(b-1)}{a(a-1)+b(b-1)}\xi(a,b-1,0;\infty), \tag{2.8}$$

$a, b = 2,3, \ldots$ and

$$\xi(a,1,0;\infty) = 2(1-a^{-1}),\ a = 1,2,\ldots,\ \xi(1,b,0;\infty) = 2(1-b^{-1}),\ b = 1,2,\ldots .$$

If $\rho = \infty$, $W_{A,n}$ and $W_{B,n}$ are independent and their distribution known. It is therefore possible to work out a (series) formula for $\xi(a,b,0;\infty)$, but (2.8) is convenient for calculation. Illustrative values of the waiting time when $\rho = \infty$ are shown in Table 2.

Table 2

Expected waiting time to marginal common ancestors, $\rho = \infty$.

n	2	10	20	50	100	500	1000
μ	1.50	2.36	2.46	2.52	2.54	2.56	2.56

If $\rho = 0$ the process essentially behaves as a single-locus one and $\xi(0,0,n;0) = 2(1-n^{-1})$.

$E(\min(W_{A,n}, W_{B,n}))$ can be found from the relationship

$$E(\min(W_{A,n}, W_{B,n}) + \max(W_{A,n}, W_{B,n})) = E(W_{A,n} + W_{B,n}) = 4(1-n^{-1}).$$

It is possible to find the distribution of $\max(W_{A,2}, W_{B,2})$ and $\min(W_{A,2}, W_{B,2})$ explicitly.

Theorem 5. The joint Laplace transform

$$\psi(\lambda_1,\lambda_2) = E(\exp\{-\lambda_1\max(W_{A,2},W_{B,2}) - \lambda_2\min(W_{A,2},W_{B,2})\}) =$$

$$\left(1+\frac{\lambda}{1+\rho}\right)^{-1}\left(\frac{1}{1+\rho}+\frac{\rho}{1+\rho}\phi(\lambda_1,\lambda_2)\right),$$

where

$$\phi(\lambda_1,\lambda_2)=(1-\beta(\phi_1(\lambda_1,\lambda_2)-1))^{-1}\,\phi_2(\lambda_1,\lambda_2),\ \lambda=\lambda_1+\lambda_2,\ \beta=\frac{2\rho(4+\rho)}{\rho^2+13\rho+18},$$

and

$$\phi_1(\lambda_1,\lambda_2)=\left(1+\frac{2\lambda}{6+\rho}\right)^{-1}\left(\gamma\left(1+\frac{\lambda}{6}\right)^{-1}+(1-\gamma)\left(1+\frac{\lambda}{1+\rho}\right)^{-1}\right),\ \gamma=\frac{1+\rho}{4+\rho},$$

$$\phi_2(\lambda_1,\lambda_2)=\left(1+\frac{2\lambda}{6+\rho}\right)^{-1}\left(\delta_1(1+\lambda_1)^{-1}\left(1+\frac{\lambda}{6}\right)^{-1}+\delta_2\left(1+\frac{\lambda}{1+\rho}\right)^{-1}+\delta_3(1+\lambda_1)^{-1}\right),$$

$$\delta_1=\frac{\rho(1+\rho)}{\rho^2+13\rho+18},\delta_2=\frac{6}{\rho^2+13\rho+18},\delta_3=\frac{12(1+\rho)}{\rho^2+13\rho+18}.$$

An alternative form for the Laplace transform is

$$\psi(\lambda_1,\lambda_2)=\frac{A}{(1+\lambda_1)B} \tag{2.9}$$

where

$$A=\lambda_1(2\lambda^2+\lambda(\rho+18)+2\rho+36)+2\lambda^2+\lambda(5\rho+18)+2\rho^2+26\rho+36$$

$$B=2\lambda^3+\lambda^2(3\rho+20)+\lambda(\rho^2+19\rho+54)+2\rho^2+26\rho+36.$$

The mean values are

$$E(\max(W_{A,2},W_{B,2}))=\frac{3\rho^2+38\rho+36}{2(\rho^2+13\rho+18)},$$

$$E(\min(W_{A,2},W_{B,2}))=\frac{\rho^2+14\rho+36}{2(\rho^2+13\rho+18)}. \tag{2.10}$$

Proof. Let $f_{abc}(\lambda_1,\lambda_2)$ be the Laplace transform of the waiting times, given $n_A(0)=a$, $n_B(0)=b$, $n_C(0)=c$. The waiting times in state (a,b,c) have a Laplace transform

$$g_{abc}(\lambda_1,\lambda_2)=\left(1+\frac{2\zeta}{n(n-1)+c\rho}\right)^{-1},$$

where

$$\zeta = \begin{cases} \lambda \text{ if } a+b>1, b+c>1 \\ \lambda_1 \text{ if } a+c=1, b+c>1 \text{ or } a+c>1, b+c=1 \end{cases}$$

Let $f_{abc} \equiv f_{abc}(\lambda_1, \lambda_2)$ and $g_{abc} \equiv g_{abc}(\lambda_1, \lambda_2)$. By considering the first transition from particular values of (a,b,c),

$$f_{002} = g_{002}\left(\frac{\rho}{1+\rho} f_{111} + \frac{1}{1+\rho}\right),$$

$$f_{111} = g_{111}\left(\frac{\rho}{6+\rho} f_{220} + \frac{2}{6+\rho} f_{002} + \frac{4}{6+\rho} f_{101}\right),$$

$$f_{101} = g_{101}\left(\frac{\rho}{2+\rho} f_{210} + \frac{2}{2+\rho}\right),$$

$$f_{210} = g_{210}\left(\frac{2}{3} f_{101} + \frac{1}{3}\right),$$

$$f_{220} = g_{220}\left(\frac{2}{3} f_{111} + \frac{1}{3} f_{210}\right).$$

The required Laplace transform is f_{002}. Solving the equations,

$$f_{101} = f_{210} = (1+\lambda_1)^{-1} \text{ and } f_{111} = \phi(\lambda_1, \lambda_2),$$

Substituting in the first equation produces the Laplace transform in the statement of the theorem. It is straightforward to find the means (2.10) from the Laplace transform (2.9). □

Corollary.

$$E(exp\{-\lambda_1 W_{A,2} - \lambda_2 W_{B,2}\}) = \frac{1}{2}(\psi(\lambda_1, \lambda_2) + \psi(\lambda_2, \lambda_1)).$$

Proof. This is clear by symmetry. □

The first form of $\psi(\lambda_1, \lambda_2)$ in Theorem 5 is an expression in terms of a mixture. Another way to invert $\psi(\lambda_1, \lambda_2)$ is to use a partial fraction expansion of (2.9). The distribution has a singular component where $W_{A,2} = W_{B,2}$. Express (2.9) as

$$\psi(\lambda_1, \lambda_2) = \psi_0(\lambda) + \psi_1(\lambda_1, \lambda_2)$$

where

$$\psi_0(\lambda) = (2\lambda^2 + (\rho+18)\lambda + 2\rho + 36)/B$$

$$\psi_1(\lambda_1, \lambda_2) = (1+\lambda_1)^{-1} 2\rho (2\lambda + \rho + 12)/B. \tag{2.11}$$

The singular part of the distribution has Laplace transform $\psi_0(\lambda)$ and the continuous part $\psi_1(\lambda_1, \lambda_2)$. In particular

$$P(W_{A,2} = W_{B,2}) = \psi_0(0) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}. \tag{2.12}$$

It is possible to show that $B(\lambda, \rho)$, the cubic in λ appearing in the denominator of $\psi(\lambda_1, \lambda_2)$, has three distinct negative roots for all $\rho \geq 0$. The Laplace transforms $\psi_0(\lambda)$, $\psi_1(\lambda_1, \lambda_2)$, $\psi(\lambda, 0)$, $\psi(0, \lambda)$, can be inverted by partial fraction techniques after finding the roots of $B(\lambda, \rho)$ for a given numerical value of ρ. An expression of the form

$$(1+\lambda_1)^{-1}(1+\lambda/\beta)^{-1}, \beta > 0$$

occurring in the expansion of $\psi(\lambda_1, \lambda_2)$ corresponds to a density

$$\beta e^{-u-(\beta-1)v}, 0 < v < u < \infty.$$

If $\rho = 0$ both the waiting times are identical, and have an exponential distribution with mean 1. If $\rho \to \infty$ the waiting times are the maximum and minimum of two independent exponential random variables, with mean 1, and so

$$P(\max(W_{A,2}, W_{B,2}) \leq w) = (1-e^{-w})^2, \quad P(\min(W_{A,2}, W_{B,2}) \leq w) = 1-e^{-2w}.$$

Hudson and Kaplan (1985) introduced the idea of the number of recombination events in a sample of genes in an m-loci model with recombination. With two loci their definition of the number of recombination events occurring in a sample of n genes is the number of events occurring to ancestors of the sample which have both loci belonging to the marginal trees. Denote their number of recombination events by R_n°. Ethier and Griffiths (1990a) define their number of recombination events R_n as the number of recombination events in a two-locus graph before the common ancestor. Of course $R_n \geq R_n^\circ$. They show that R_n is distributed as the number of steps to the right in a random walk starting at n with absorbing state 1 and transition probabilities for $m \geq 2$,

$$m \to \begin{cases} m-1, & \text{probability } (m-1)\ (m-1+\rho), \\ m+1, & \text{probability } \rho\ (m-1+\rho) \end{cases}.$$

An explicit formula for the probability generating function of R_n is given in Ethier and Griffiths (1990b), where it is also shown that

$$E(R_n) = \rho \int_0^1 \frac{1-(1-v)^{n-1}}{v} e^{\rho v} dv.$$

They also show that there is a stochastic representation related to the random walk which gives the two-locus sampling distribution. This is a particularly attractive representation because of its simplicity, and is a good reason for having such an extended definition of ancestors. Here is a description of the process. Run a random walk starting at n until it hits 1.

Transition probabilities of the random walk are :

$$r \to \begin{cases} r-1, & (r-1)\ (r-1+\theta+\rho), \\ r, & \theta\ (r-1+\theta+\rho), \\ r+1, & \rho\ (r-1+\theta+\rho), \end{cases}$$

where $\theta = \theta_A + \theta_B$.

Let τ be the hitting time of 1. Construct samples of genes (pairs of loci) at $\tau - 1, \tau - 2, \ldots, 0$ in the following way. Start with a sample of size 1 at τ. Let $\{U_t, t = 0, 1, 2, \ldots\}$ be the random walk and $\{Z_t = U_{t+1} - U_t, t = 0, 1, \ldots\}$. If $Z_t = -1$, duplicate a sample member at random. If $Z_t = 0$, choose a sample member to mutate, at either locus A or B with respective probabilities θ_A/θ or θ_B/θ. If $Z_t = +1$, choose two members of the sample without replacement to recombine. The sample of n at time 0 has the two-locus sampling distribution.

Theorem 6. Let $\eta(a, b, c; \rho)$ be the expected number of transitions $(n_A(t), n_B(t), n_C(t)) \to (n_A(t^+) + 1, n_B(t^+) + 1, n_c(t^+) - 1)$ before $n_A(t) + n_C(t) = 1$, $n_B(t) + n_C(t) = 1$ beginning with $n_A(0) = a$, $n_B(0) = b$, $n_C(0) = c$.

Then $E(R_n^\circ) = \eta(0, 0, n; \rho)$. η satisfies the system (2.7) with $\beta(a, b, c; \rho) = c\rho/(n(n-1) + c\rho)$ and $\eta(1, 1, 0; \rho) = 0$, $\eta(0, 0, 1; \rho) = 0$, $\eta(1, 0, 0; \rho) = 0$, $\eta(0, 1, 0; \rho) = 0$. A particular case is $E(R_2^\circ) = 6\rho/(6 + \rho)$. □

Table 3 illustrates results from Theorem 6 for a sample of size 50.

Table 3

Recombination events in a samples' ancestors.

ρ	$E(R_{50})$	$E(R^{\circ}_{50})$
0	0	0
0.1	4.58×10^{-1}	4.43×10^{-1}
0.5	2.52	2.13
1.0	5.78	4.07
2.0	1.62×10	7.51
5.0	2.12×10^{2}	1.57×10
10.0	2.49×10^{4}	2.56×10
20.0	5.12×10^{8}	3.94×10

As $\rho \to \infty$, $R_n \to \infty$, but $R^{\circ}_n < \infty$ and has a proper limit distribution. Using an argument similar to that for $\xi(a, b, c; \infty)$, $\eta(a, b, c; \infty) = c + \eta(a + c, b + c, 0; \infty)$. $\eta(a, b, c; \infty)$ satisfies a similar system of equations to (2.8), but with the term $2/(a(a - 1) + b(b - 1))$ replaced by $2ab/(a(a - 1) + b(b - 1))$, $\eta(a, 1, 0; \infty) = 2\Sigma_{j=1}^{a-1} j^{-1}$, $a = 2,3, \ldots$, $\eta(1, b, 0; \infty) = 2\Sigma_{j=1}^{b-1} j^{-1}$, $b = 2,3, \ldots$, , and $\eta(1, 1, 0; \infty) = 0$. Thus if $\rho = \infty$, $E(R^{\circ}_n) = n + \eta(n, n, 0; \infty)$.

If the first time that $n_A(t) + n_C(t) = 1$, $n_B(t) + n_C(t) = 1$ is when $n_A(t) = 0$, $n_B(t) = 0$, $n_C(t) = 1$ then the first common ancestor in the two-locus graph and the two first marginal common ancestors coincide. This produces the following theorem.

Theorem 7. Let

$$T = \inf\{t \geq 0;\ n_A(t) + n_C(t) = 1,\ n_B(t) + n_C(t) = 1\}.$$

and

$$\zeta(a, b, c; \rho) = P(n_C(T) = 1 \mid n_A(0) = a, n_B(0) = b, n_C(0) = c).$$

Then ζ satisfies the system (2.7) with $\beta(a, b, c; \rho) = 0$ and $\zeta(1, 1, 0; \rho) = 0$, $\zeta(0, 0, 1; \rho) = 1$, $\zeta(1,0, 0; \rho) = 0$, $\zeta(0, 1, 0; \rho) = 0$. A particular case is

$$\zeta(0,0,2;\rho) = \frac{(2+\rho)(9+13\rho)}{(1+\rho)(\rho^2+17\rho+18)}. \quad \square$$

Table 4 illustrates probabilities from Theorem 7 for $n = 50$.

Table 4

Probability that marginal and two-locus ancestors are identical, $n = 50$.

ρ	0	0.1	0.5	1.0	2.0	5.0	10.0	20.0
Probability	1.0	0.998	0.969	0.909	0.785	0.525	0.326	0.182

It is possible to modify coalescent trees so that ancestral lines are lost by coalescence or by mutation (Griffiths (1980), Tavaré (1984)). With a single locus the number of ancestral lines at time t back is a death process with rate $n(n + \theta - 1)/2$. Lines are lost by coalescence, at rate $n(n - 1)/2$ and by mutation at rate $n\theta/2$. In the two-locus case the process $(n_A(t), n_B(t), n_C(t))$ can be modified to include loss by mutation.

Transitions rates are:

	$(a+1, b+1, c-1)$,	$c\rho/2$
	$(a-1, b-1, c+1)$,	ab
	$(a-1, b, c)$,	$ac + a(a-1)/2 + a\theta_A/2$
$(a, b, c) \to$	$(a, b-1, c)$,	$bc + b(b-1)/2 + b\theta_B/2$
	$(a, b, c-1)$,	$c(c-1)/2$
	$(a+1, b, c-1)$,	$c\theta_B/2$
	$(a, b+1, c-1)$,	$c\theta_A/2$,

with $a + b + c > 1$.

The expected waiting time until $n_A(t) + n_C(t) = 1$ and $n_B(t) + n_C(t) = 1$ in Theorem 4 can be modified by considering the transitions (2.13).

One way to derive the probability generating function, $P_n(z)$, for the number of allele types from a sample of n from the one-locus model is to consider back in time whether lines are lost by mutation or coalescence. Each line lost by mutation corresponds to a distinct allele type in the sample. By considering the first event back in time, for $n \geq 2$,

$$(n+\theta-1)P_n(z) = (n-1)P_{n-1}(z) + \theta z P_n(z). \qquad (2.13)$$

Solving (2.13) with the boundary condition $P_1(z) = z$,

$$P_n(z) = \prod_{j=1}^{n} \frac{\theta z + j - 1}{\theta + j - 1}.$$

The proof of the next theorem, in the two-locus case, is clear using this approach. The theorem is also proved in Ethier and Griffiths (1990b) by applying the generator of the measure-valued diffusion. It is possible to compute the joint distribution of the number of allele types at the two loci, but there is no simple formula for it.

Theorem 8. Let (K, L) be the number of allele types at locus A and locus B in a sample of n genes. The *pgf* of (K, L) is $P(0, 0, n; z_A, z_B)$ in the system of equations

$$\begin{aligned}
&(n(n-1) + c\rho + (a+c)\theta_A + (b+c)\theta_B) P(a, b, c; z_A, z_B) = \\
&\quad c\rho P(a+1, b+1, c-1; z_A, z_B) + 2abP(a-1, b-1, c+1; z_A, z_B) \\
&\quad + a(a+2c-1+\theta_A z_A) P(a-1, b, c; z_A, z_B) \\
&\quad + b(b+2c-1+\theta_B z_B) P(a, b-1, c; z_A, z_B) \\
&\quad + c(c-1) P(a, b, c-1; z_A, z_B) \\
&\quad + c\theta_B z_B P(a+1, b, c-1; z_A, z_B) + c\theta_A z_A P(a, b+1, c-1; z_A, z_B).
\end{aligned}$$

Boundary conditions are $P(1, 1, 0; z_A, z_B) = P(0, 0, 1; z_A, z_B) = z_A z_B$. □

References

[1] Ethier, S. N. and Griffiths, R. C. (1990a). The neutral two-locus diffusion model as a measure-valued diffusion. *Adv. Appl. Prob.* **22**, 773-786.

[2] Ethier, S. N. and Griffiths, R. C. (1990b). On the two-locus sampling distribution. *J. Math. Biol.* To appear.

[3] Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Pop. Biol.* **17** 37-50.

[4] Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* **23** 183-201.

[5] Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109** 611-631.

[6] Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111** 147-164.

[7] Kaplin, N. L. and Hudson, R. R. (1985). The use of sample genealogies for studying a selectively neutral m--loci model with recombination. *Theor. Pop. Biol.* **28** 382-396.

[8] Kingman, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235-248.

[9] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* **26** 119-164.