J. Appl. Prob. **50,** 256–271 (2013)
Printed in England

DUALITY BETWEEN THE TWO-LOCUS WRIGHT–FISHER DIFFUSION MODEL AND THE ANCESTRAL PROCESS WITH RECOMBINATION

SHUHEI MANO,* *The Institute of Statistical Mathematics*

Abstract

Known results on the moments of the distribution generated by the two-locus Wright–Fisher diffusion model, and the duality between the diffusion process and the ancestral process with recombination are briefly summarized. A numerical method for computing moments using a Markov chain Monte Carlo simulation and a method to compute closed-form expressions of the moments are presented. By applying the duality argument, the properties of the ancestral recombination graph are studied in terms of the moments.

Keywords: Duality; diffusion process; ancestral graph; recombination; population genetics

2010 Mathematics Subject Classification: Primary 60K35
Secondary 60J70; 92D15; 92D25

1. Introduction

In classical population genetics theory, the behavior of the frequency of a gene type (allele) has been a central issue (see, for example, [2]). The fate of the allele frequency is modeled by a diffusion process, where the population size is assumed to be sufficiently large. The diffusion limit, which is called the Wright–Fisher diffusion model, is expected to illustrate the actual evolution of the allele frequency in the population. Numerical methods for computing the likelihood of a sample taken from the equilibrium distribution have attracted much interest (see, for example, [22]). Explicit and closed-form expressions of the whole process are important in their own right. Unfortunately, their availability has been limited. For the one-locus, two-allele model without mutation and other evolutionary forces, closed-form expressions of the probability density of the allele frequency at a fixed time have been obtained in terms of orthogonal polynomials [10], [16]. In contrast, for two-locus models, known closed-form expressions have been limited to several moments of the distribution generated by the diffusion process [15], [20]. A comprehensive survey carried out in the early 1970s, which is still useful, is presented in [15]. Recently, closed-form expressions of a class of moments were obtained in terms of orthogonal polynomials [17].

The concept of duality has been a powerful tool in stochastic analysis of interacting particle systems [14]. In population genetics theory the moment dual of the Wright–Fisher diffusion model was introduced in [21]. The genealogical process of a sample taken from a population, which is known as the coalescent [12], has been useful for population genetics data analyses. The duality was applied to obtain branching-coalescent processes as models of natural selection [13] and conversion bias [19]. The number of ancestral lineages in a section of the ancestral graph is an ancestral process, analogously to coalescent genealogy. The dual of the one-locus,

Received 24 February 2012; revision received 20 July 2012.
* Postal address: The Institute of Statistical Mathematics and The Japan Science and Technology Agency, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. Email address: smano@ism.ac.jp

two-allele Wright–Fisher diffusion model with directional selection [11] is an ancestral process, which traces the ancestral lineages in a section of the ancestral selection graph. The dual is a birth-and-death process with linear birth and quadratic death rates. It was demonstrated that the properties of the birth-and-death process can be studied by referring to the classical results of the Wright–Fisher diffusion model [18]. For the multi-locus model, an analogue of the coalescent genealogy, called the ancestral recombination graph (ARG), was introduced in [7]. The two-locus ARG integrates marginal genealogies at the two loci. The ancestral process, which traces the ancestral lineages in a section of the ARG, is the dual of the two-locus, two-allele Wright–Fisher diffusion model [5].

In Section 2 we briefly summarize the known results for the moments of the distribution generated by the two-locus, two-allele Wright–Fisher diffusion model. In Section 3, the moment duality between the diffusion process and the ancestral process, which traces the ancestral lineages in a section of the ARG, is introduced. A numerical method for computing moments at a fixed time by a Markov chain Monte Carlo simulation is introduced. In Section 4, a method for computing closed-form expressions of the moments using ARG terminology is presented. In Section 5, applying the duality argument, the properties of the ARG are studied in terms of the moments.

2. Summary of known results for the moments

Consider a random mating monoecious diploid population consisting of N individuals. Two linked loci, A and B, segregate, where the recombination fraction between the two loci is r. Pairs of alleles A_1, A_2 and B_1, B_2 are in loci A and B, respectively. A diffusion limit measures the time in units of $2N$ generations and $2N \to \infty$, while $\rho = 4Nr$ is kept constant. Let the frequencies of gametes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 respectively be x_1, x_2, x_3, and $1-x_1-x_2-x_3$. The frequencies of alleles A_1 and A_2 are denoted by x and $1-x$, respectively, and those of the alleles B_1 and B_2 are denoted by y and $1-y$, respectively. Then $x = x_1 + x_2$ and $y = x_1 + x_3$. Set $z = x_1(1 - x_1 - x_2 - x_3) - x_2x_3$, which is a measure of the association between x and y, or the disequilibrium between the two loci. The limiting diffusion process $\{x_1(t), x_2(t), x_3(t); t \geq 0\}$ is defined in the simplex

$$K : 0 \leq x_1 \leq x_1 + x_2 \leq x_1 + x_2 + x_3 \leq 1.$$

Let $H = \Phi(K)$, where $\Phi(x_1, x_2, x_3) = (x, y, z)$ is a C^∞-diffeomorphism of K onto H. The generator of the diffusion process $\{x(t), y(t), z(t); t \geq 0\}$ in H is [20]

$$\begin{aligned}\mathcal{L} = {} & \frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2} + \frac{y(1-y)}{2}\frac{\partial^2}{\partial y^2} + z\frac{\partial^2}{\partial x\partial y} + z(1-2x)\frac{\partial^2}{\partial x\partial z} + z(1-2y)\frac{\partial^2}{\partial y\partial z} \\ & - z\left(1+\frac{\rho}{2}\right)\frac{\partial}{\partial z} + \frac{1}{2}\{xy(1-x)(1-y) + z(1-2x)(1-2y) - z^2\}\frac{\partial^2}{\partial z^2}. \end{aligned} \tag{2.1}$$

In the classical population genetics theory some problems of general interest concern events of fixation. The probability of the eventual fixation of an allele and the probability density of the time to fixation have been studied. Some of these properties can be studied in terms of the moments of the distribution generated by the model by using the moment inversion formula. The probability of the eventual fixation of a gamete in the two-locus, two-allele Wright–Fisher diffusion model governed by generator (2.1) is obtained immediately from the moments. In fact, since the stationary density is atomic, $\lim_{t\to\infty} \mathbb{E}[x(t)y(t)]$ gives the fixation probability of the gamete AB. For an allele, two types of fixation can be defined: one type of fixation

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

occurs when the first of the four alleles is lost and the second type of fixation occurs when an allele at the other locus is lost (a gamete fix). These fixation times are

$$T_1 = \inf\{t \geq 0; x(t)(1-x(t))y(t)(1-y(t)) = 0\},$$
$$T_0 = \inf\{t \geq 0; x(t)(1-x(t)) + y(t)(1-y(t)) = 0\},$$

respectively. The probability densities are

$$\mathbb{P}[T_1 < t] = \lim_{n\to\infty} \mathbb{E}[\{1 - x(t)(1-x(t))y(y)(1-y(t))\}^n],$$
$$\mathbb{P}[T_0 < t] = \lim_{n\to\infty} \mathbb{E}[\{1 - x(t)(1-x(t)) - y(y)(1-y(t))\}^n],$$

respectively. It seems impossible to obtain explicit and closed-form expressions for these limits (see Section 4). Nevertheless, some of the moments for which closed-form expressions are available are useful for obtaining upper bounds and approximate formulae for these probabilities [15]. For the same reason, a closed-form expression of the joint distribution of $(x(t), y(t), z(t))$ at a fixed time t is not available.

Let us introduce a classification of the moments of the distribution obtained from generator (2.1).

Definition 2.1. The *rank* and *class* of a moment, which is an expectation of a monomial $x^l u^m x_1^n$, $l, m, n \in \mathbb{Z}_+$, are $l + m + 2n$ and $n + \min\{l, m\}$, respectively. The rank is equal to or twice as large as the class.

Remark 2.1. The class-zero moments have closed-form expressions and they are the moments of the one-locus Wright–Fisher diffusion model [10]. The class-one moments have closed-form expressions [17] (see below).

Other moments whose closed-form expressions have been obtained are expectations of types of polynomials.

Lemma 2.1. (Lemma 3.6.1 of [15].) *The manifold of polynomials spanned by the set of polynomials* $\{x^l(1-x)^l y^m(1-y)^m z^n(1-2x)^a(1-2y)^a\}$, *where* $a = 0, 1$ *and* $l, m > 0$ *if* $n = 0$, *is closed under the operation of* $\mathcal{L}$.

Remark 2.2. The polynomials are zero on the boundary of the square $x(1-x)y(1-y) = 0$ and $z = 0$. Known closed-form expressions for the moments of polynomials of this type are $(l, m, n, a) = (1, 1, 0, 0)$, $(0, 0, 2, 0)$, and $(0, 0, 0, 1)$ (see [20]). Expressions for $(2, 1, 0, 0)$, $(1, 2, 0, 0)$, $(2, 0, 1, 1)$, $(0, 2, 1, 1)$, $(2, 0, 2, 0)$, $(0, 2, 2, 0)$, $(2, 2, 0, 0)$, $(1, 1, 1, 1)$, $(1, 1, 2, 0)$, $(0, 0, 3, 1)$, and $(0, 0, 4, 0)$ are given in [15], and involve eigenvalues whose closed-form expressions are not available.

In [17] a closed-form expression of $\mathbb{E}[x_1(t) \mid x(t) \in (0, 1)]$ was obtained using a limit of a closed-form expression of a class-one moment. This yielded an expression for the conditional covariance between x and y given that alleles A_1 and A_2 are segregated in locus A. This expression plays an important role in interpreting observable polymorphisms in population genetics data analysis. Here, we summarize some results given in [17] that will be used in the later sections. If the argument equals unity, the truncated hypergeometric series

$$y_n(a, b; c; z) = \sum_{i=0}^{n} \frac{(a)_i (b)_i}{(c)_i i!} z^i$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

is expressed by the generalized hypergeometric series [1]

$$y_n(a,b,c;1)=\frac{\Gamma(a+n+1)\Gamma(b+n+1)}{n!\,\Gamma(a+b+n+1)}{}_3F_2(a,b,c+n;c,a+b+n+1;1),$$

where ${}_3F_2(\cdot)$ denotes the generalized hypergeometric series. A trivial but useful identity is the following.

Lemma 2.2. ([17].) *For $m,n\in\mathbb{Z}_+$ and $a,b,c\in\mathbb{C}$,*

$$\begin{aligned}&\frac{n!\,\Gamma(a+b+n+1)}{\Gamma(a+n+1)\Gamma(b+n+1)}y_n(a,b;a+b+m+1;1)\\&\quad=\frac{m!\,\Gamma(a+b+m+1)}{\Gamma(a+m+1)\Gamma(b+m+1)}y_m(a,b;a+b+n+1;1).\end{aligned}\tag{2.2}$$

Remark 2.3. If $m=0$, (2.2) gives the identity

$$\begin{aligned}{}_2F_1(a,b;a+b+1;1)&={}_3F_2(a,b,a+b+n+1;a+b+1,a+b+n+1;1)\\&=\frac{n!\,\Gamma(a+b+n+1)}{\Gamma(a+n+1)\Gamma(b+n+1)}y_n(a,b,a+b+1;1)\\&=\frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b+1)},\end{aligned}$$

which is a special case of the Gauss hypergeometric theorem [1].

An expression of a power of p in terms of the Gegenbauer polynomial follows by the orthogonal and complete property [17]:

$$p^n=\sum_{m=2}^{n+2}2(2m-1)\frac{[n+1]_{m-1}}{(n+1)_{m+1}}(-1)^mT_{m-2}^1(1-2p),\qquad n\in\mathbb{Z}_+.\tag{2.3}$$

Here $T_m^1(\cdot)$ is the Gegenbauer polynomial, which is also denoted by $C_1^{(3/2)}(\cdot)$, and $[n]_m$ and $(n)_m$ are the falling and rising factorials, respectively. Using this expression, closed-form solutions of systems of differential equations for class-one moments were obtained. Let $\mu_{l,m,n}(t)=\mathbb{E}_{pqd}[x(t)^ly(t)^mz(t)^n]$ for $l,m,n\in\mathbb{Z}_+$.

Proposition 2.1. ([17].) *For $n\in\mathbb{Z}_+$,*

$$\mu_{n,0,1}(t)=\sum_{m=2}^{n+2}2(2m-1)\frac{[n+1]_{m-1}}{(n+1)_{m+1}}(-1)^mT_{m-2}^1(1-2p)d\mathrm{e}^{-(m(m-1)+\rho)t/2}.$$

Proposition 2.2. ([17].) *For $n\in\mathbb{Z}_+$,*

$$\mu_{n,1,0}(t)=pq+\frac{2d}{2+\rho}+\sum_{m=1}^{n-1}E_n^{(m)}\mathrm{e}^{-m(m+1)t/2}+\sum_{m=1}^{n}F_n^{(m)}\mathrm{e}^{-(\rho+m(m+1))t/2}$$

except for $\rho=(k+m)(k-m-1)$, $k=m+2,m+3,\ldots,n$, $m=1,2,\ldots,n$, where

$$\begin{aligned}E_n^{(m)}=(-1)^m\frac{[n]_{m+1}}{(n)_{m+1}}\Bigg[&\frac{2(2m+1)}{m(m+1)}p(1-p)qT_{m-1}^1(1-2p)\\&+2\left\{\frac{T_m^1(1-2p)}{2(m+1)+\rho}+\frac{T_{m-2}^1(1-2p)}{2m-\rho}\right\}d\Bigg]\end{aligned}$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

and

$$F_n^{(m)} = 2(-1)^m \frac{[n]_m}{(n)_m} \left\{ \frac{1}{2m+\rho} + \frac{1}{2(m+1)-\rho} \frac{(n-m)(n-m-1)}{(n+m)(n+m+1)} \right\} T_{m-1}^1(1-2p)d, \tag{2.4}$$

with the conventions that the first sum is 0 *if* $n = 1$ *and* $T_{-1}^1(\cdot) = 0$.

Proof. A sketch of the proof was given in [17]. A system of differential equations for the moments gives

$$E_n^{(m)} = \frac{[n]_{m+1}!\,(2m+1)!}{(n)_{m+1}(m+1)!\,m!} E_{m+1}^{(m)}, \qquad n \in \mathbb{Z}_+,\ m = 1, 2, \ldots, n-1,$$

and

$$\{(n+m)(n-m-1)-\rho\}F_n^{(m)} = 4n(2m+1)\frac{[n-1]_{m-1}}{(n+1)_{m+1}}(-1)^{m+1}T_{m-1}^1(1-2p)d + n(n-1)F_{n-1}^{(m)}, \qquad n \in \mathbb{Z}_+,\ m = 1, 2, \ldots, n, \tag{2.5}$$

with the initial condition

$$p^n q = pq + \frac{2d}{2+\rho} + \sum_{m=1}^{n-1} E_n^{(m)} + \sum_{m=1}^{n} F_n^{(m)}, \qquad n \in \mathbb{Z}_+.$$

Using (2.2) with $m = 1, 2$, it is straightforward to solve (2.5) for $F_n^{(m)}$, yielding the solution given in (2.4), except for $\rho = (k+m)(k-m-1)$, $k = m+2, m+3, \ldots, n$, $m = 1, 2, \ldots, n$ (the exceptional values given in [17] are incorrect). By setting $E_n^{(m)} = qE_{n,1}^{(m)}(p) + dE_{n,2}^{(m)}(p)$, the initial condition gives

$$\begin{aligned}
\sum_{m=1}^{n-1} E_{n,1}^{(m)}(p) &= p(p-1)\sum_{i=0}^{n-2} p^i \\
&= p(p-1)\sum_{i=0}^{n-2}\sum_{m=2}^{i+2} 2(2m-1)\frac{[i+1]_{m-1}}{(i+1)_{m+1}}(-1)^m T_{m-2}^1(1-2p) \\
&= p(p-1)\sum_{m=1}^{n-1} 2(-1)^{m+1}\frac{m!\,(m-1)!}{(2m)!} \\
&\qquad \times y_{n-m-1}(m+1, m, 2m+2; 1)T_{m-1}^1(1-2p) \\
&= p(1-p)\sum_{m=1}^{n-1}(-1)^m \frac{[n]_{m+1}}{(n)_{m+1}}\frac{2(2m+1)}{m(m+1)}T_{m-1}^1(1-2p)
\end{aligned}$$

for each $n = 2, 3, \ldots$, where the second equality holds by (2.3) and the last equality holds by (2.2) with $m = 0$. The expression for the summation of $E_{n,2}^{(m)}(p)$ follows by rearrangement of the terms.

3. Duality and a numerical method for computing moments

The process that traces the number of lineages, including nonancestral lineages (see below), in a section of a two-locus ARG is a birth-and-death process [7]. When there are i lineages, the birth rate is $i\rho/2$ and the death rate is $i(i-1)/2$. The birth-and-death process is identical to the number of ancestral lineages in a section of ancestral selection graph, and the moment

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

dual of the birth-and-death process is the Wright–Fisher diffusion of the one-locus, two-allele model with directional selection [18]. Note that an ARG involves gametes whose alleles are not ancestral to any allele in a sample. We denote alleles that are not ancestral to any allele in the sample with a minus sign (e.g. $A-$), since in principle the allelic state of the locus cannot be specified by the sample. For example, a gamete AB could be a recombinant descendant of $A-$, which in turn could be a recombinant descendant of $--$. The gamete $--$ is involved in the ARG, but its alleles are not ancestral to any allele in the sample. In this paper we discuss the ancestral process that generates the number of ancestral lineages in a section of the ARG. The ancestral lineages are a subset of lineages in the ARG. The stationary distribution of the process with the infinitely-many-allele mutation model was studied in [6]. A whole graph $\mathcal{G}$ includes marginal genealogies $\mathcal{T}_A$ and $\mathcal{T}_B$ at the loci A and B, respectively. We denote the edges of a graph by $E(\cdot)$. The edges of $\mathcal{G}$ are partitioned into $\mathcal{A} = E(\mathcal{G}) \cap E(\mathcal{T}_A) \cap E(\mathcal{T}_B)^{\mathrm{c}}$, $\mathcal{B} = E(\mathcal{G}) \cap E(\mathcal{T}_A)^{\mathrm{c}} \cap E(\mathcal{T}_B)$, $\mathcal{C} = E(\mathcal{G}) \cap E(\mathcal{T}_A) \cap E(\mathcal{T}_B)$, and $\mathcal{D} = E(\mathcal{G}) \cap E(\mathcal{T}_A)^{\mathrm{c}} \cap E(\mathcal{T}_B)^{\mathrm{c}}$. We call $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ ancestral lineages; $\mathcal{D}$ is not an ancestral lineage since it is not ancestral to any allele in the sample. Let $\mathcal{E}_t$ be the edges of a section of $\mathcal{G}$ taken at time t backwards. We denote the number of ancestral lineages by $a(t) = |\mathcal{E}_t \cap \mathcal{A}|$, $b(t) = |\mathcal{E}_t \cap \mathcal{B}|$, and $c(t) = |\mathcal{E}_t \cap \mathcal{C}|$. The marginal transition rates of $(a(t), b(t), c(t))$ do not depend on $|\mathcal{E}_t \cap \mathcal{D}|$ and the process is Markovian [5], [7]. The rates are

$$(a, b, c) \to \begin{cases} (a+1, b+1, c-1), & c\rho/2, \\ (a-1, b-1, c+1), & ab, \\ (a-1, b, c), & ac + a(a-1)/2, \\ (a, b-1, c), & bc + b(b-1)/2, \\ (a, b, c-1), & c(c-1)/2, \end{cases} \tag{3.1}$$

with $a + b + c > 1$. The backward equation for the joint probability generating function $\xi_{l,m,n}(t) = \mathbb{E}_{l,m,n}[p^{a(t)} q^{b(t)} g^{c(t)}]$ of the Markov chain $\{a(t), b(t), c(t); t \geq 0\}$ on the integer lattice $\mathbb{Z}_+^3 \setminus \mathbf{0}$ is

$$\begin{aligned} \frac{\mathrm{d}\xi_{l,m,n}}{\mathrm{d}t} = &-\frac{(l+m+n)(l+m+n-1)+n\rho}{2}\xi_{l,m,n} + \frac{n\rho}{2}\xi_{l+1,m+1,n-1} \\ &+ \frac{l(l-1+2n)}{2}\xi_{l-1,m,n} + \frac{m(m-1+2n)}{2}\xi_{l,m-1,n} + \frac{n(n-1)}{2}\xi_{l,m,n-1} \\ &+ lm\xi_{l-1,m-1,n+1} \end{aligned} \tag{3.2}$$

for $(l, m, n) \in \mathbb{Z}_+^3 \setminus \mathbf{0}$, where terms whose subscripts have negative integers are 0. It is straightforward to see that the moments $\nu_{l,m,n}(t) = \mathbb{E}_{p,q,g}[x(t)^l y(t)^m x_1(t)^n]$ also satisfy the system of differential equations (3.2). Therefore, a moment duality follows immediately [5]. We give a proof, since it is useful to introduce a numerical method for computing moments.

Lemma 3.1. *The diffusion process* $\{x(t), y(t), z(t); t \geq 0\}$ *in* H *with* $(x(0), y(0), x_1(0)) = (p, q, g)$ *and the Markov chain* $\{a(t), b(t), c(t); t \geq 0\}$ *in* $\mathbb{Z}_+^3 \setminus \mathbf{0}$ *whose transition rates are (3.1) with* $(a(0), b(0), c(0)) = (l, m, n)$ *are dual to each other:*

$$\mathbb{E}_{p,q,g}[x(t)^l y(t)^m x_1(t)^n] = \mathbb{E}_{l,m,n}[p^{a(t)} q^{b(t)} g^{c(t)}].$$

Proof. The system of differential equations (3.2) is equivalent to an integro-recurrence equation

$$\xi_{l,m,n}(t) = \int_0^t \mathcal{T}\xi_{l,m,n}(s) \mathrm{e}^{-\gamma_{l,m,n}(t-s)}\, \mathrm{d}s, \qquad l, m, n \in \mathbb{Z}_+,$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

where

$$\mathcal{T}\xi_{l,m,n} = \frac{n\rho}{2}\xi_{l+1,m+1,n-1} + \frac{l(l-1+2n)}{2}\xi_{l-1,m,n} + \frac{m(m-1+2n)}{2}\xi_{l,m-1,n} + \frac{n(n-1)}{2}\xi_{l,m,n-1} + lm\xi_{l-1,m-1,n+1} \tag{3.3}$$

and $\gamma_{l,m,n} = ((l+m+n)(l+m+n-1)+n\rho)/2$. The integro-recurrence equation is recast as

$$\xi_{l,m,n}(t) = \int_0^t \sum_{l',m',n'} \mathbb{P}[l'm'n' \mid l,m,n]\xi_{l',m',n'}(t-s)\gamma_{l,m,n}\mathrm{e}^{-\gamma_{l,m,n}s}\,\mathrm{d}s, \tag{3.4}$$

where the transition probability is given by dividing the rates in (3.1) by $\gamma_{a,b,c}$. Meanwhile,

$$\begin{aligned}\mathbb{E}_{l,m,n}[p^{a(t)}q^{b(t)}g^{c(t)}] &= \mathbb{E}_{l,m,n}[\mathbb{E}[p^{a(t)}q^{b(t)}g^{c(t)} \mid (a(s),b(s),c(s)) = (l',m',n')]] \\ &= \mathbb{E}_{l,m,n}[\mathbb{E}_{l',m',n'}[p^{a(t-s)}q^{b(t-s)}g^{c(t-s)}]] \\ &= \mathbb{E}_{l,m,n}[\xi_{l',m',n'}(t-s)],\end{aligned}$$

where the second equality follows by the strong Markov property. This expression is equivalent to (3.4). Therefore, $\xi_{l,m,n}(t) = \mathbb{E}_{l,m,n}[p^{a(t)}q^{b(t)}g^{c(t)}]$. On the other hand, it is straightforward to see by Itô's formula that the moments $\nu_{l,m,n}(t) = \mathbb{E}_{p,q,g}[x(t)^l y(t)^m x(t)^n]$ satisfy the system of differential equations (3.2).

The duality relation is useful for numerical computation of the moments $\nu_{l,m,n}(t)$ by simulating independent copies of $(a(t),b(t),c(t))$ by the Markov chain Monte Carlo simulation with transition probabilities (3.1). Consider the simulations stopped at time t. The average over $p^{a(t)}q^{b(t)}g^{c(t)}$ of the copies is then an unbiased estimator of the moment $\nu_{l,m,n}(t)$. A similar method was used for computing the likelihood of a sample in a varying environment [8]. The simulation can be stopped before t. Consider the hitting time

$$\tau = \inf\{s \geq 0;\, (a(s),b(s),c(s)) \in \mathcal{S}\},$$

where $\mathcal{S} = \{(0,0,1),(1,1,0)\}$ is the closed set of states from which a chain cannot exit. If $\tau < t$,

$$\begin{aligned}\mathbb{E}_{l,m,n}[p^{a(t)}q^{b(t)}g^{c(t)}] &= \mathbb{E}_{l,m,n}[\mathbb{E}[p^{a(t)}q^{b(t)}g^{c(t)} \mid (a(\tau),b(\tau),c(\tau)) = (l',m',n')]] \\ &= \mathbb{E}_{l,m,n}[\mathbb{E}_{l',m',n'}[p^{a(t-\tau)}q^{b(t-\tau)}g^{c(t-\tau)}]] \\ &= \mathbb{P}_{l,m,n}[(a(\tau),b(\tau),c(\tau)) = (0,0,1)]\nu_{0,0,1}(t-\tau) \\ &\quad + \mathbb{P}_{l,m,n}[(a(\tau),b(\tau),c(\tau)) = (1,1,0)]\nu_{1,1,0}(t-\tau),\end{aligned}$$

where the second equality follows by the strong Markov property, and

$$\nu_{0,0,1}(t-\tau) = g - \frac{\rho(g-pq)}{2+\rho}(1-\mathrm{e}^{-(2+\rho)(t-\tau)/2})$$

and

$$\nu_{1,1,0}(t-\tau) = pq + \frac{2(g-pq)}{2+\rho}(1-\mathrm{e}^{-(2+\rho)(t-\tau)/2}).$$

From these observations we have the following numerical method for computing the moments.

Proposition 3.1. *Set a Markov time $\sigma = t \wedge \tau$, where $x \wedge y = \min\{x,y\}$. An unbiased estimator of $\nu_{l,m,n}(t)$ is the average of the following values obtained by independent copies of*

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

$(a(\sigma), b(\sigma), c(\sigma))$ *simulated using the Markov chain Monte Carlo simulation with transition probabilities (3.1). If* $\sigma = t$, *the value is* $p^{a(\sigma)}q^{b(\sigma)}g^{c(\sigma)}$. *If* $\sigma = \tau$ *and* $(a(\sigma), b(\sigma), c(\sigma)) = (0, 0, 1)$, *the value is* $\nu_{0,0,1}(t-\tau)$. *If* $\sigma = \tau$ *and* $(a(\sigma), b(\sigma), c(\sigma)) = (1, 1, 0)$, *the value is* $\nu_{1,1,0}(t-\tau)$.

4. Closed-form expressions for the moments

Since the system of differential equations for the moments of the distribution generated by the two-locus, two-allele Wright–Fisher diffusion model (3.2) is the same as that for the joint probability generating function of the distribution of the number of ancestral lineages in a section of a two-locus ARG, the relationship among moments can be specified in terms of the events on the ARG. In (3.3), the first event is a recombination, the second and third events are marginal coalescences in $\mathcal{T}_A$ and $\mathcal{T}_B$, respectively, and the forth event is a joint coalescence in both $\mathcal{T}_A$ and $\mathcal{T}_B$. We call the fifth event a null coalescence, because coalescence events occur in neither $\mathcal{T}_A$ nor $\mathcal{T}_B$. The following lemma is obvious.

Lemma 4.1. *The manifold of moments spanned by the set of moments whose ranks and classes are equal to or smaller than specified values is closed under the operation of* $\mathcal{T}$. *Neither the class nor the rank of a moment change under recombination and null coalescence operations.*

It was shown in [15] that the moments $\mu_{l,m,n}(t)$ can be obtained recursively from the smaller rank moments. We present a method to compute the closed-form expressions of the moments $\nu_{l,m,n}(t)$ using ARG terminology, since it gives systematic insights into the computation. In our approach the concept of a class of a moment is essential. Since closed-form expressions of all moments whose classes are less than two are available (see Section 2), let us start with moments whose classes and ranks are two and $i(\geq 4)$, respectively: $\nu_{i-4,0,2}$, $\nu_{i-3,1,1}$, $\nu_{i-2,2,0}$, $\nu_{0,i-4,2}$, $\nu_{1,i-3,1}$, and $\nu_{2,i-2,0}$. The first three moments are closed by recombinations and the latter three moments are closed by null coalescences. Expressions for the latter three moments are obtained immediately by exchanging p and q in the expressions of the first three moments. The system of three differential equations for the first three moments whose ranks are j $(4 \leq j \leq i)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\nu_{j-4,0,2}\\ \nu_{j-3,1,1}\\ \nu_{j-2,2,0}\end{pmatrix} = \begin{pmatrix}-\frac{(j-2)(j-3)+2\rho}{2} & \rho & 0\\ j-3 & -\frac{(j-1)(j-2)+\rho}{2} & \frac{\rho}{2}\\ 0 & 2(j-2) & -\frac{j(j-1)}{2}\end{pmatrix}\begin{pmatrix}\nu_{j-4,0,2}\\ \nu_{j-3,1,1}\\ \nu_{j-2,2,0}\end{pmatrix} + \begin{pmatrix}\frac{(j-1)(j-4)}{2}\nu_{j-5,0,2}\\ \frac{(j-2)(j-3)}{2}\nu_{j-4,1,1}\\ \frac{(j-2)(j-3)}{2}\nu_{j-3,2,0}\end{pmatrix} + \begin{pmatrix}\nu_{j-4,0,1}\\ \nu_{j-3,0,1}\\ \nu_{j-2,1,0}\end{pmatrix}, \tag{4.1}$$

where $\nu_{-1} = 0$ by convention. The solution involves eigenvalues of the matrix in (4.1), which are roots of the cubic equation

$$\lambda^3 + \frac{3j^2-9j+8+3\rho}{2}\lambda^2 + \left\{\frac{3(j-1)^2(j-2)^2}{2} + (3j^2-11j+15)\rho + \rho^2\right\}\lambda + \frac{j(j-1)^2(j-2)^2(j-3)}{8} + \frac{(3j^4-22j^3+65j^2-86j+48)\rho + (j^2-5j+8)\rho^2}{4} = 0.$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

If $j = 4$, the system of differential equations (4.1) involves $\nu_{0,0,2}$, $\nu_{1,1,1}$, $\nu_{2,2,0}$, and class-one moments, and we can obtain closed-form expressions. In fact, the closed-form expressions of the moments whose classes and ranks are two and four, respectively, were obtained in [20]. Solving (4.1) iteratively in $j = 5, 6, \dots, i$, we obtain the closed-form expressions of the moments whose classes and ranks are two and i, respectively.

The moments whose classes and ranks are $k(\geq 3)$ and $i(\geq 2k)$, respectively, are computed in the same way. To obtain closed-form expressions of these moments, a system of $k + 1$ differential equations must be solved. The solution involves eigenvalues of a tridiagonal matrix A_k with

$$(A_k)_{s,s-1} = (s-1)(j-2k+s-1),$$
$$(A_k)_{s,s} = -\frac{(j-k+s-1)(j-k+s-2)+(k-s+1)\rho}{2},$$
$$(A_k)_{s,s+1} = \frac{k-s+1}{2}\rho,$$

for $1 \leq s \leq k+1$ and $2k \leq j \leq i$. The eigenvalues are the roots of the $(k+1)$th-degree algebraic equation. If $k \geq 4$, we cannot expect explicit closed-form expressions of the eigenvalues. The computation eventually involves moments whose classes and ranks are t and u, respectively, where $1 \leq t \leq k$ and $i - k + t \leq u \leq 2t$.

5. Properties of the ARG

We have discussed how to compute the moments of the distribution generated by the two-locus, two-allele Wright–Fisher diffusion model. These moments are useful for studying the two-locus ARG, since the expression for the moments of the diffusion $\nu_{l,m,n}(t)$ is the same as that for the joint probability generating function for the distribution of the number of ancestral lineages in a section of an ARG with the initial condition $(a(0), b(0), c(0)) = (l, m, n)$. We define the rank and class of the number of ancestral lineages in a section of an ARG, (l, m, n), by $l + m + 2n$ and $n + \min\{l, m\}$, respectively. An ARG of a class-zero sample is a marginal genealogy, whose properties are well known. In the following we consider a sample whose class is larger than 0. Since Lemma 3.1 gives

$$\lim_{t\to\infty} \nu_{l,m,n}(t) = \frac{2g}{2+\rho} + \frac{\rho pq}{2+\rho}, \qquad n + \min\{l, m\} \geq 1,$$

the stationary distribution of the number of ancestral lineages in a section of an ARG is

$$\frac{2}{2+\rho}\delta_{(0,0,1)} + \frac{\rho}{2+\rho}\delta_{(1,1,0)}.$$

The distribution of the number of ancestral lineages in a section of an ARG with the initial condition (0, 0, 2) can be obtained from the known closed-form expression of the moments of the distribution generated by the two-locus, two-allele Wright–Fisher diffusion model [20]. It seems that a general formula (applicable to all rank moments) for the distribution of a sample whose class is larger than 1 is not available. In contrast, we have a general formula for the distribution of class-one samples. The closed-form expression can be obtained by using closed-form expressions of the moments with a finite-series expansion of the Gegenbauer polynomial [3]:

$$T_m^1(1-2p) = \frac{1}{2}\sum_{i=0}^{m} \frac{(-m)_i (m+1)_{i+2}}{i!\,(i+1)!} p^i. \tag{5.1}$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

Let $\nu_{k,1,0}(t) = \sum_{l,m,n} f_{l,m,n}(t) p^l q^m g^n$, where

$$f_{l,m,n}(t) = \mathbb{P}_{k,1,0}[(a(t), b(t), c(t)) = (l, m, n)].$$

The distribution of sample $(1, 1, 0)$ is

$$f_{1,1,0}(t) = \frac{\rho}{2+\rho} + \frac{2}{2+\rho} e^{-(2+\rho)t/2}, \qquad f_{0,0,1}(t) = \frac{2}{2+\rho}(1 - e^{-(2+\rho)t/2}),$$

since

$$\nu_{1,1,0} = pq + \frac{2(g-pq)}{2+\rho}(1 - e^{-(2+\rho)t/2}).$$

For samples $(k, 1, 0)$, $k \geq 2$, from Proposition 2.2 and (5.1), the closed-form expressions have a general formula. For $i \geq 0$, we have

$$\begin{aligned} f_{i,0,1}(t) = {} & \frac{2}{2+\rho}\delta_{i,0} + g_{i,k}(t) \\ & + \sum_{m=i}^{k-1} \frac{(-1)^m [k]_{m+1}}{(k)_{m+1} i!\,(i+1)!} \\ & \qquad \times \left[\frac{(-m)_i (m+1)_{i+2}}{2(m+1)+\rho} + \frac{(2-m)_i (m-1)_{i+2}}{2m-\rho} \right] e^{-m(m+1)t/2}, \end{aligned}$$

and, for $i \geq 2$, we have

$$\begin{aligned} f_{i,1,0}(t) = {} & -g_{i-1,k}(t) \\ & - \sum_{m=i-1}^{k-1} \frac{(-1)^m [k]_{m+1}}{(k)_{m+1} (i-1)!\, i!} \\ & \qquad \times \left\{ (2m+1)(1-m)_{i-2}(m)_i \right. \\ & \qquad \left. + \left[\frac{(-m)_{i-1}(m+1)_{i+1}}{2(m+1)+\rho} + \frac{(2-m)_{i-1}(m-1)_{i+1}}{2m-\rho} \right] \right\} e^{-m(m+1)t/2} \end{aligned}$$

and

$$\begin{aligned} f_{1,1,0}(t) = {} & \frac{\rho}{2+\rho} - g_{0,k}(t) \\ & + \sum_{m=1}^{k-1} \frac{(-1)^m [k]_{m+1}}{(k)_{m+1}} \left[2m+1 - \frac{(m+1)(m+2)}{2(m+1)+\rho} - \frac{(m-1)m}{2m-\rho} \right] e^{-m(m+1)t/2}, \end{aligned}$$

where

$$\begin{aligned} g_{i,k}(t) = {} & \sum_{m=i+1}^{k-1} \frac{(-1)^m [k]_m}{(k)_m} \left[\frac{1}{2m+\rho} + \frac{1}{2(m+1)-\rho} \frac{(k-m)(k-m-1)}{(k+m)(k+m+1)} \right] \\ & \qquad \times \frac{(1-m)_i (m)_{i+2}}{i!\,(i+1)!} e^{-(m(m+1)+\rho)t/2}. \end{aligned}$$

We can obtain closed-form expressions for the distribution of samples $(k, 0, 1)$, $k \geq 1$, in a similar manner.

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

Let the waiting times until common ancestors of $\mathcal{T}_A$ and $\mathcal{T}_B$ respectively be

$$W_A = \inf\{s \geq 0; a(s) + c(s) = 1\}, \qquad W_B = \inf\{s \geq 0; b(s) + c(s) = 1\}.$$

Proposition 5.1. *The waiting time until a sample has a common ancestor at both of the two loci is given by*

$$\mathbb{P}_{l,m,n}[W_A \vee W_B \leq t] = \mathbb{P}_{l,m,n}[(a(t), b(t), c(t)) \in \mathcal{S}],$$

where $x \vee y = \max\{x, y\}$. *The waiting time until a sample has a common ancestor at one of the two loci is given by*

$$\mathbb{P}_{l,m,n}[W_A \wedge W_B \geq t] = \sum_{n'+\min\{l',m'\}\geq 2} \mathbb{P}_{l,m,n}[(a(t), b(t), c(t)) = (l', m', n')].$$

Remark 5.1. A recursion of the expectation of $W_A \vee W_B$ for the ARG with the initial condition $(0, 0, c)$ is given by Theorem 4 of [7]. Theorem 5 of [7] gives a closed-form expression for the joint Laplace transform of $W_A \vee W_B$ and $W_A \wedge W_B$ for the ARG with the initial condition $(0, 0, 2)$.

The idea of the number of recombination events in a sample was introduced in [9]. The number of recombination events on the two-locus ARG, including nonancestral lineages, was considered in [5] and [6], where a closed-form expression for the probability generating function of the number of recombination events was given. Here, we consider the number of recombination events on the ancestral lineages of an ARG. Let $s(t)$ be the number of recombination events occurring in $\mathcal{C}$ lineages of an ARG in a time interval $(0, t)$. The recombination events are a subset of the recombination events occurring in the whole lineages of the ARG.

Lemma 5.1. *The joint probability generating function of* $(a(t), b(t), c(t), s(t))$ *is*

$$\mathbb{E}_{l,m,n,0}[p^{a(t)}q^{b(t)}g^{c(t)}v^{s(t)}] = \mathbb{E}_{l,m,n,0}\left[p^{a_v(t)}q^{b_v(t)}g^{c_v(t)}\exp\left\{-\frac{\rho(1-v)}{2}\int_0^t c_v(u)\,\mathrm{d}u\right\}\right],$$

where $\{a_v(t), b_v(t), c_v(t); t \geq 0\}$ *is a modified process of* $\{a(t), b(t), c(t); t \geq 0\}$ *in which the recombination fraction is* rv, $0 \leq v \leq 1$.

Proof. Let $\zeta_{l,m,n}(t) = \mathbb{E}_{l,m,n,0}[p^{a(t)}q^{b(t)}g^{c(t)}v^{s(t)}]$. For $(l, m, n) \in \mathbb{Z}_+^3 \setminus \mathbf{0}$, we have

$$\begin{aligned}\frac{\mathrm{d}\zeta_{l,m,n}}{\mathrm{d}t} = &-\frac{(l+m+n)(l+m+n-1)+nv\rho}{2}\zeta_{l,m,n} + \frac{nv\rho}{2}\zeta_{l+1,m+1,n-1}\\ &+\frac{l(l-1+2n)}{2}\zeta_{l-1,m,n} + \frac{m(m-1+2n)}{2}\zeta_{l,m-1,n} + \frac{n(n-1)}{2}\zeta_{l,m,n-1}\\ &+lm\zeta_{l-1,m-1,n+1} - \frac{n(1-v)\rho}{2}\zeta_{l,m,n}\end{aligned}$$

with the initial condition $\xi_{l,m,n}(0) = p^l q^m g^n$. This is uniquely solved by means of the Feynman–Kac formula, giving the desired result.

Theorem 5.1. *The conditional probability generating function of* $\mathcal{A}$ *lineages in a section of an ARG with the initial condition* $(n, 0, 1)$ *given that no recombination events occur in a time interval* $(0, t)$ *is*

$$\mathbb{E}_{n,0,1,0}[p^{a(t)} \mid s(t) = 0] = \tilde{\nu}_{n,0,1}(t),$$

where $\tilde{\nu}_{n,0,1}(t)$ *is* $\nu_{n,0,1}(t)$, *setting* $q = g = 1$ *and* $\rho = 0$.

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

Proof. By (3.1) we see that $b_0(t) = 0$ and $c_0(t) = 1$ for all t, and the marginal process $\{a_0(t); t \geq 0\}$ is a death process with death rate $i(i+1)/2$ when $a_0(t) = i$. The joint probability generating function of $(a(t), b(t), c(t))$ given that no recombination events occur in a time interval $(0, t)$ is

$$\begin{aligned}\lim_{v\to 0} \mathbb{E}_{n,0,1,0}[p^{a(t)}q^{b(t)}g^{c(t)}v^{s(t)}] &= \mathbb{E}_{n,0,1,0}[p^{a(t)}q^{b(t)}g^{c(t)},\ s(t)=0] \\ &= \mathbb{E}_{n,0,1,0}\left[p^{a_0(t)}q^{b_0(t)}g^{c_0(t)}\exp\left\{-\frac{\rho}{2}\int_0^t c_0(u)\,\mathrm{d}u\right\}\right] \\ &= g\mathbb{E}_{n,0,1,0}[p^{a_0(t)}]\mathrm{e}^{-\rho t/2},\end{aligned}$$

where the first equality follows by Lebesgue's convergence theorem and the second equality follows by Lemma 5.1. Setting $p = q = g = 1$, we have

$$\mathbb{P}_{n,0,1,0}[s(t) = 0] = \mathrm{e}^{-\rho t/2},$$

while setting $q = g = 1$, we have

$$\mathbb{E}_{n,0,1,0}[p^{a(t)}, s(t) = 0] = \tilde{\nu}_{n,0,1}(t)\mathrm{e}^{-\rho t/2},$$

where $\tilde{\nu}_{n,0,1}(t) = \mathbb{E}_{n,0,1,0}[p^{a_0(t)}]$.

Remark 5.2. The explicit closed-form expression of $\tilde{\nu}_{n,0,1}(t)$ is given in Section 2, since $\nu_{n,0,1}(t) = \mu_{n+1,1,0}(t) + \mu_{n,0,1}(t)$. This expression follows immediately by considering the ARG. Recombination might occur on the single $\mathcal{C}$ lineage. According to the Poisson nature of recombination events, the probability that no recombination occurs on the single lineage is $\mathrm{e}^{-\rho t/2}$. The marginal process $\{a_0(t); t \geq 0\}$ follows the death process independently.

Lemma 5.2. *Let $\mathcal{S}$ be the absorbing states. The probability generating function of the number of recombination events on the ancestral lineages of an ARG until a sample has a common ancestor at both of the two loci is*

$$\mathbb{E}_{l,m,n,0}[v^{s(\tau)}] = \mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho(1-v)}{2}\int_0^{\tau_v} c_v(u)\,\mathrm{d}u\right\}\right],$$

where $\tau_v = \inf\{s \geq 0; (a_v(s), b_v(s), c_v(s)) = \mathcal{S}\}$.

Proof. Let $\zeta_{l,m,n} = \mathbb{E}_{l,m,n,0}[p^{a(\tau)}q^{b(\tau)}g^{c(\tau)}v^{s(\tau)}]$. For $(l, m, n) \in \mathbb{Z}_+^3 \setminus \mathbf{0}$, we have

$$\begin{aligned}0 = &-\frac{(l+m+n)(l+m+n-1)+nv\rho}{2}\zeta_{l,m,n} + \frac{nv\rho}{2}\zeta_{l+1,m+1,n-1} \\ &+ \frac{l(l-1+2n)}{2}\zeta_{l-1,m,n} + \frac{m(m-1+2n)}{2}\zeta_{l,m-1,n} + \frac{n(n-1)}{2}\zeta_{l,m,n-1} \\ &+ lm\zeta_{l-1,m-1,n+1} - \frac{n(1-v)\rho}{2}\zeta_{l,m,n}\end{aligned}$$

with the boundary conditions $\xi_{0,0,1} = g$ and $\xi_{1,1,0} = pq$. This boundary-value problem is uniquely solved by means of the Feynman–Kac formula. That is,

$$\begin{aligned}\zeta_{l,m,n} = &\, g\mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho(1-v)}{2}\int_0^{\tau_v} c_v(u)\,\mathrm{d}u\right\}, (a_v(\tau_v), b_v(\tau_v), c_v(\tau_v)) = (0,0,1)\right] \\ &+ pq\mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho(1-v)}{2}\int_0^{\tau_v} c_v(u)\,\mathrm{d}u\right\}, (a_v(\tau_v), b_v(\tau_v), c_v(\tau_v)) = (1,1,0)\right].\end{aligned}$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

On the other hand, we have

$$\mathbb{E}_{l,m,n,0}[p^{a(\tau)}q^{b(\tau)}g^{c(\tau)}v^{s(\tau)}] = g\mathbb{E}_{l,m,n,0}[v^{s(\tau)}, (a(\tau), b(\tau), c(\tau)) = (0, 0, 1)] + pq\mathbb{E}_{l,m,n,0}[v^{s(\tau)}, (a(\tau), b(\tau), c(\tau)) = (1, 1, 0)].$$

Thus, the probability generating functions of the number of recombination events on the ancestral lineages of an ARG until a sample has a common ancestor at both of the two loci, with the given state in which the sample path is absorbed, are

$$\mathbb{E}_{l,m,n,0}[v^{s(\tau)}, (a(\tau), b(\tau), c(\tau)) = (0, 0, 1)] = \mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho(1-v)}{2}\int_0^{\tau_v} c_v(u)\,\mathrm{d}u\right\}, (a_v(\tau_v), b_v(\tau_v), c_v(\tau_v)) = (0, 0, 1)\right]$$

and

$$\mathbb{E}_{l,m,n,0}[v^{s(\tau)}, (a(\tau), b(\tau), c(\tau)) = (1, 1, 0)] = \mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho(1-v)}{2}\int_0^{\tau_v} c_v(u)\,\mathrm{d}u\right\}, (a_v(\tau_v), b_v(\tau_v), c_v(\tau_v)) = (1, 1, 0)\right].$$

The proof is complete upon summation of these two probability generating functions.

Corollary 5.1. *The expected number of recombination events on ancestral lineages of an ARG until a sample has a common ancestor at both of the two loci is*

$$\mathbb{E}_{l,m,n,0}[s(\tau)] = \frac{\rho}{2}\mathbb{E}_{l,m,n,0}\left[\int_0^{\tau} c(u)\,\mathrm{d}u\right] = \frac{\rho}{2}\sum_{(l',m',n')\in\mathbb{Z}_+^3\setminus\{\mathbf{0},\mathcal{S}\}}\int_0^{\infty} k\mathbb{P}_{l,m,n,0}[(a(u), b(u), c(u)) = (l', m', n')]\,\mathrm{d}u.$$

Proof. Let $T_{l',m',n'}$ be the sojourn time of a sample path of the process of the number of ancestral lineages in a section of an ARG with the initial condition (l, m, n) that stays at state $(l', m', n') \notin \mathcal{S}$. Then

$$\begin{aligned}\mathbb{E}_{l,m,n,0}&\left[\int_0^{\tau} c(u)\,\mathrm{d}u\right] \\ &= \sum_{(l',m',n')\in\mathbb{Z}_+^3\setminus\{\mathbf{0},\mathcal{S}\}} n'\mathbb{E}_{l,m,n,0}[T_{l',m',n'}] \\ &= \sum_{(l',m',n')\in\mathbb{Z}_+^3\setminus\{\mathbf{0},\mathcal{S}\}} n'\mathbb{E}_{l,m,n,0}\left[\int_0^{\infty} I_{(l',m',n')}(a(t), b(t), c(t))\,\mathrm{d}t\right] \\ &= \sum_{(l',m',n')\in\mathbb{Z}_+^3\setminus\{\mathbf{0},\mathcal{S}\}} n'\int_0^{\infty}\mathbb{P}_{l,m,n,0}[(a(u), b(u), c(u)) = (l', m', n')]\,\mathrm{d}u,\end{aligned}$$

where the last equality follows by Fubini's theorem.

Remark 5.3. A recursion of $\mathbb{E}_{l,m,n,0}[s(\tau)]$ is given in Theorem 6 of [7].

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

Corollary 5.2. *The probability that no recombination events occur on an ARG until a sample has a common ancestor at both of the two loci is*

$$\mathbb{P}_{l,m,n,0}[s(\tau)=0]=\mathbb{E}_{l,m,n,0}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\right].$$

Theorem 5.2. *The probability that no recombination events occur on an ARG with the initial condition* $(0,0,n)$ *until the sample has a common ancestor at both of the two loci is*

$$\frac{(n-1)!}{(\rho+1)_{n-1}}.$$

Proof. By (3.1) we see that $a_0(t)=b_0(t)=0$ for all t, and the marginal process $\{c_0(t);\ t\geq 0\}$ is a death process with death rate $i(i-1)/2$ when $c_0(t)=i$. Consider the hitting time $\gamma=\inf\{s\geq 0;\ a(s)=n-1\}$. By Corollary 5.2,

$$\begin{aligned}\mathbb{P}_{0,0,n,0}[s(\tau)=0]&=\mathbb{E}_{0,0,n,0}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\right]\\&=\mathbb{E}_{0,0,n,0}\left[\mathbb{E}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\,\middle|\,\gamma\right]\right]\\&=\mathbb{E}_{0,0,n,0}\left[\mathbb{E}_{0,0,n-1,0}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\right]\mathrm{e}^{-n\rho\gamma/2}\right]\\&=\mathbb{E}_{0,0,n-1,0}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\right]\frac{n-1}{n-1+\rho},\end{aligned}$$

where the third equality follows by the strong Markov property, with the boundary condition

$$\mathbb{E}_{0,0,1,0}\left[\exp\left\{-\frac{\rho}{2}\int_0^{\tau_0}c_0(u)\,\mathrm{d}u\right\}\right]=1.$$

Solving this recursion completes the proof.

Theorem 5.3. *The probability that no recombination events occur on an ARG with the initial condition* $(n,0,1)$ *until the sample has a common ancestor at both of the two loci is*

$$\prod_{i=0}^{n}\frac{i(i+1)}{i(i+1)+\rho}.$$

Proof. By (3.1) we see that $b_0(t)=0$ and $c_0(t)=1$ for all t, and the marginal process $\{a_0(t);\ t\geq 0\}$ is a death process with death rate $i(i+1)/2$ when $a_0(t)=i$. By Corollary 5.2, we have

$$\mathbb{P}_{n,0,1}[s(\tau)=0]=\mathbb{E}_n[\mathrm{e}^{-\rho\tau_0/2}].$$

The expression follows by an argument similar to that used in the proof of Theorem 5.2.

Finally, let us consider the limit $\rho\to\infty$. We introduce two processes: a diffusion process $\{x_\infty(t),\,y_\infty(t);\ t\geq 0\}$ in $[0,1]^2$ with generator

$$\mathcal{L}_\infty=\frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}+\frac{y(1-y)}{2}\frac{\partial^2}{\partial y^2}$$

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

and $(x_\infty(0), y_\infty(0)) = (p, q)$, and a Markov chain $\{a_\infty(t), b_\infty(t); t \geq 0\}$ in $\mathbb{Z}_+^2 \setminus \mathbf{0}$ whose transition rates are

$$(a, b) \to \begin{cases} (a-1, b), & a(a-1)/2, \\ (a, b-1), & b(b-1)/2, \end{cases}$$

with $(a_\infty(0), b_\infty(0)) = (l, m)$. Let $\tau_\infty = \{s \geq 0; (a_\infty(s), b_\infty(s)) \in \{(1,0), (0,1), (1,1)\}\}$.

Theorem 5.4. (Theorems 1 and 2 of [4].) *If $(x_\infty(0), y_\infty(0)) = (p, q)$ then $\{x(t), y(t); t \geq 0\}$ converges weakly in $C([0,\infty), [0,1]^2)$ to $\{x_\infty(t), y_\infty(t); t \geq 0\}$, and $\{z(t) - d\mathrm{e}^{-\rho t/2}; t \geq 0\}$ converges weakly in $C([0,\infty), \mathbb{R})$ to the zero process in $\mathbb{R}$ as $\rho \to \infty$. The two function spaces are given the topology of uniform convergence on compact intervals.*

Corollary 5.3. *The probability generating function for the distribution of the number of ancestral lineages in a section of an ARG with the initial condition $(a(0), b(0), c(0)) = (l, m, n)$ has the limit*

$$\mathbb{E}_{l,m,n}[p^{a(t)} q^{b(t)} g^{c(t)}] \to \mathbb{E}_{l+n}[p^{a_\infty(t)}] \mathbb{E}_{m+n}[q^{b_\infty(t)}] \quad \text{as } \rho \to \infty.$$

Proof. From Lemma 3.1 we have

$$\mathbb{E}_{l,m,n}[p^{a(t)} q^{b(t)} g^{c(t)}] = \mathbb{E}_{p,q,g}[x(t)^{l+n} y(t)^{m+n}] + \sum_{i=1}^{n} \frac{n!}{(n-i)!\, i!} \mathbb{E}_{p,q,g}[x(t)^{l+n-i} y(t)^{m+n-i} z(t)^i].$$

It follows from Theorem 5.4 and Lebesgue's convergence theorem that

$$\mathbb{E}_{p,q,g}[x(t)^{l+n-i} y(t)^{m+n-i} z(t)^i] \leq \mathbb{E}_{p,q,g}[z(t)^i] \to 0 \quad \text{as } \rho \to \infty$$

for $t > 0$ and $i = 1, 2, \ldots, n$, while

$$\mathbb{E}_{p,q,g}[x(t)^{l+n} y(t)^{m+n}] \to \mathbb{E}_p[x_\infty(t)^{l+n}] \mathbb{E}_q[y_\infty(t)^{m+n}] = \mathbb{E}_{l+n}[p^{a_\infty(t)}] \mathbb{E}_{m+n}[q^{b_\infty(t)}] \quad \text{as } \rho \to \infty.$$

The last equality is a result of the duality between $\{x_\infty(t); t \geq 0\}$ and $\{a_\infty(t); t \geq 0\}$, and between $\{y_\infty(t); t \geq 0\}$ and $\{b_\infty(t); t \geq 0\}$.

Corollary 5.3 shows that all AB gametes in a sample instantaneously split into an $A-$ and $-B$ gametes pair in the limit $\rho \to \infty$. Therefore, the length of $\mathcal{C}$ lineages in an ARG goes to 0 in this limit.

Theorem 5.5. *The expected lengths of $\mathcal{C}$ lineages and whole lineages of an ARG with the initial condition (l, m, n) until the sample has a common ancestor at both of the two loci are*

$$\mathbb{E}_{l,m,n}\left[\int_0^\tau c(u)\,\mathrm{d}u\right] \to \frac{2}{\rho} \mathbb{E}_{l,m}\left[\int_0^{\tau_\infty} a_\infty(u) b_\infty(u)\,\mathrm{d}u\right] + \frac{2n}{\rho} \tag{5.2}$$

and

$$\mathbb{E}_{l,m,n}\left[\int_0^\tau (a(u) + b(u) + c(u))\,\mathrm{d}u\right] \to \mathbb{E}_{l,m}\left[\int_0^{\tau_\infty} (a_\infty(u) + b_\infty(u))\,\mathrm{d}u\right], \tag{5.3}$$

respectively.

https://doi.org/10.1239/jap/1363784437 Published online by Cambridge University Press

Proof. Let $\eta_{l,m,n} = \lim_{\rho\to\infty} \mathbb{E}_{l,m,n,0}[s(\tau) = 0]$ and $\lambda_{l,m} = \eta_{l,m,0}$. For $(l, m, n) \in \mathbb{Z}_+^3 \setminus \mathbf{0}$, we have $\eta_{l,m,n} = n + \lambda_{l+n,m+n}$ for $n \geq 1$ and

$$0 = lm - \frac{l(l-1)+m(m-1)}{2}\lambda_{l,m} + \frac{l(l-1)}{2}\lambda_{l-1,m} + \frac{m(m-1)}{2}\lambda_{l,m-1},$$

with the boundary condition $\lambda_{1,0} = \lambda_{0,1} = \lambda_{1,1} = 0$. This boundary-value problem is uniquely solved by means of the Feynman–Kac formula. That is,

$$\lambda_{l,m} = \mathbb{E}_{l,m}\left[\int_0^{\tau_\infty} a_\infty(u) b_\infty(u)\,\mathrm{d}u\right].$$

Equation (5.2) then follows from Corollary 5.1.

Let

$$\eta'_{l,m,n} = \lim_{\rho\to\infty} \mathbb{E}_{l,m,n,0}\left[\int_0^{\tau_\infty} (a(u) + b(u) + c(u))\,\mathrm{d}u\right]$$

and $\lambda'_{l,m} = \eta'_{l,m,0}$. For $(l, m, n) \in \mathbb{Z}_+^3 \setminus \mathbf{0}$, we have $\eta'_{l,m,n} = \lambda'_{l,m}$ for $n \geq 1$ and

$$0 = l + m - \frac{l(l-1)+m(m-1)}{2}\lambda'_{l,m} + \frac{l(l-1)}{2}\lambda'_{l-1,m} + \frac{m(m-1)}{2}\lambda'_{l,m-1},$$

with the boundary condition $\lambda'_{1,0} = \lambda'_{0,1} = \lambda'_{1,1} = 0$. Solving this boundary problem leads to (5.3).

References

[1] Bailey, W. N. (1935). *Generalized Hypergeometric Series.* Cambridge University Press.
[2] Crow, J. F. and Kimura, M. (1971). *An Introduction to Population Genetics Theory.* Harper and Low, New York.
[3] Erdély, A. (ed.) (1953). *Higher Transcendental Functions*, Vol. I. McGraw-Hill, New York.
[4] Ethier, S. N. (1979). A limit theorem for two-locus diffusion models in population genetics. *J. Appl. Prob.* **16,** 402–408.
[5] Ethier, S. N. and Griffiths, R. C. (1990). The neutral two-locus model as a measure-valued diffusion. *Adv. Appl. Prob.* **22,** 773–786.
[6] Ethier, S. N. and Griffiths, R. C. (1990). On the two-locus sampling distribution. *J. Math. Biol.* **29,** 131–159.
[7] Griffiths, R. C. (1991). The two-locus ancestral graph. In *Selected Proceedings of the Sheffield Symposium on Applied Probability* (Sheffield, 1989; IMS Lecture Notes Monogr. **18**), eds I. V. Basawa and R. L. Taylor, Institute of Mathematical Statistics, Hayward, CA, pp. 100–117.
[8] Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. London* B **344,** 403–410.
[9] 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.
[10] Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Acad. Sci. USA* **41,** 144–150.
[11] Kimura, M. (1955). Stochastic process and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symp. Quantitative Biol.* **20,** 33–53.
[12] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.
[13] Krone, S. M. and Neuhauser, C. (1997). Ancestral process with selection. *Theoret. Pop. Biol.* **51,** 210–237.
[14] Liggett, T. M. (1985). *Interacting Particle Systems.* Springer, Berlin.
[15] Littler, R. A. (1972). Multidimensional stochastic models in genetics. Doctoral Thesis, Monash University.
[16] Malécot, G. (1948). *Les Mathematiques de l'Hérédité.* Masson et Cie, Paris.
[17] Mano, S. (2005). Random genetic drift and gamete frequency. *Genetics* **171,** 2043–2050.
[18] Mano, S. (2009). Duality, ancestral and diffusion processes in models with selection. *Theoret. Pop. Biol.* **75,** 164–175.
[19] Mano, S. (2013). Ancestral graph with bias in gene conversion. *J. Appl. Prob.* **50,** 239–255.
[20] Ohta, T. and Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genet. Res.* **13,** 47–55.
[21] Shiga, T. (1981). Diffusion processes in population genetics. *J. Math. Kyoto Univ.* **21,** 133–151.
[22] Tavaré, S. (2004). Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics* (Lecture Notes Math. **1837**), ed. J. Picard, Springer, Berlin, pp. 1–188.