Adv. Appl. Prob. **38,** 405–429 (2006)
Printed in Northern Ireland

COALESCENT LINEAGE DISTRIBUTIONS

ROBERT C. GRIFFITHS,* *University of Oxford*

Abstract

We study identities for the distribution of the number of edges at time t back (i.e. measured backwards) in a coalescent tree whose subtrees have no mutations. This distribution is important in the infinitely-many-alleles model of mutation, where every mutation is unique. The model includes, as a special case, the number of edges in a coalescent tree at time t back when mutation is ignored. The identities take the form of expected values of functions of $Z_t = \mathrm{e}^{\mathrm{i}X_t}$, where X_t is distributed as standard Brownian motion. Associated identities are also found for the distributions of the time to the most recent common ancestor, the time until loss of ancestral lines by coalescence or mutation, and the age of a mutation. Hypergeometric functions play an important role in the identities. The identities are of mathematical interest, as well as potentially being formulae to use for numerical integration or simulation to compute distributions that are usually expressed as alternating-sign series expansions, which are difficult to compute.

Keywords: Age of a mutation; coalescent lineage distribution; coalescent process; hypergeometric function; infinitely-many-alleles model; Jacobi polynomial; line of descent

2000 Mathematics Subject Classification: Primary 92D15
Secondary 60G99

1. Introduction

A *coalescent tree* is a random binary tree representing the ancestry of a sample of genes in mathematical genetics models. Kingman (1982) is the seminal paper on the coalescent. Griffiths (1980) is an earlier paper with many coalescent ideas, and later well-written expository papers are Tavaré (1984), Watterson (1984), Hudson (1991), and Nordborg (2001). Edges in the coalescent tree coalesce exchangeably at a rate of 1 per unordered pair, with a total coalescence rate of $\binom{k}{2}$, k being the number of leaf edges in the tree. Coalescence is fast enough to begin with an entrance boundary of an infinite number of leaves, representing the entire population of genes. In the *infinitely-many-alleles model* of mutation, mutations occur according to a Poisson process along the edges of the coalescent tree at rate $\theta/2$, each mutation producing a unique allele type. The distribution of the configuration of alleles in the leaves of an n-coalescent tree is Ewens' sampling formula (Ewens (1972)), and the relative frequency distribution in an infinite-leaf coalescent tree is Poisson–Dirichlet (Kingman (1993, Chapter 9)).

There are a number of post-coalescent proofs of Ewens' sampling formula; a direct combinatorial proof with connections to this paper can be found in Griffiths and Lessard (2005). The distribution back in time of edges in a coalescent tree which have no mutations in their subtended family subtrees is important. Once a mutation occurs in an edge with a nonmutant family, it then determines the allele type of the family. This process of how mutant families arise

Received 5 December 2005; revision received 13 February 2006.

* Postal address: Department of Statistics, University of Oxford, 1 South Parks Rd, Oxford OX1 3TG, UK.
Email address: griff@stats.ox.ac.uk

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

in the history of a sample of genes is determined by the forest obtained by tracing edges in the coalescent tree from the leaves up to first mutations, which become roots in the forest subtrees. The *infinitely-many-sites model* with no recombination is a refinement of the infinitely-many-alleles model where mutations occur at positions never before mutant in DNA segments. The distribution of numbers of genes in families of different types in this model is identical to that in the infinitely-many-alleles model. Let $A_n^\theta(t)$ be the number of nonmutant lineages at time t back (i.e. measured backwards) in a coalescent tree of a sample of n genes. Griffiths (1980) and Tavaré (1984) studied the distribution theory of $A_n^\theta(t)$, which is now summarised. We have

$$\mathrm{P}(A_n^\theta(t) = j) = \sum_{k=j}^{n} \rho_k^\theta(t)(-1)^{k-j} \frac{(2k+\theta-1)(j+\theta)_{(k-1)} n_{[k]}}{j!\,(k-j)!\,(n+\theta)_{(k)}} \tag{1.1}$$

for $j = 0, 1, \dots, n$, where $\rho_k^\theta(t) = \mathrm{e}^{-k(k+\theta-1)t/2}$. We write $a_{(j)} = a(a+1)\cdots(a+j-1)$ and $b_{[j]} = b(b-1)\cdots(b-j+1)$. Equation (1.1) also holds for $n = \infty$, formally setting $n_{[k]}/(n+\theta)_{(k)} = 1$, in which case

$$\mathrm{P}(A_\infty^\theta(t) = j) = \sum_{k=j}^{\infty} \rho_k^\theta(t)(-1)^{k-j} \frac{(2k+\theta-1)(j+\theta)_{(k-1)}}{j!\,(k-j)!}. \tag{1.2}$$

The process $\{A_n^\theta(t),\ t \geq 0\}$ is a death process with edges lost by coalescence or mutation at rates $j(j+\theta-1)/2,\ j = n, n-1, \dots, 1$. If $\theta = 0$ then $A_n^0(t)$ is the number of edges at time t back in the coalescent tree. The distribution (1.1) satisfies the forward equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{P}(A_n^\theta(t) = j) = -\frac{j(j+\theta-1)}{2}\mathrm{P}(A_n^\theta(t) = j) + \frac{(j+1)(j+\theta)}{2}\mathrm{P}(A_n^\theta(t) = j+1) \tag{1.3}$$

for $j = 0, 1, \dots, n-1$, with

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{P}(A_n^\theta(t) = n) = -\frac{n(n+\theta-1)}{2}\mathrm{P}(A_n^\theta(t) = n).$$

It is straightforward to verify that the distribution (1.1) satisfies the forward equations (1.3): by equating coefficients of $\rho_k^\theta(t)$ in (1.3) it is sufficient to show that

$$[k(k+\theta-1) - j(j+\theta-1)]\frac{(j+\theta)_{k-1}}{j!\,(k-j)!} = [(j+1)(j+\theta)]\frac{(j+\theta+1)_{(k-1)}}{(j+1)!\,(k-j-1)!},$$

which follows from writing

$$k(k+\theta-1) - j(j+\theta-1) = (k-j)(j+k+\theta-1)$$

and simplifying.

The lth falling factorial moment of $A_n^\theta(t)$ is

$$\mathrm{E}(A_n^\theta(t)_{[l]}) = \sum_{k=l}^{n} \rho_k^\theta(t)(2k+\theta-1)\binom{k-1}{l-1}(\theta+k)_{(l-1)}\frac{n_{[k]}}{(n+\theta)_{(k)}},$$

with a similar equation,

$$\mathrm{E}(A_\infty^\theta(t)_{[l]}) = \sum_{k=l}^{\infty} \rho_k^\theta(t)(2k+\theta-1)\binom{k-1}{l-1}(\theta+k)_{(l-1)},$$

for the population lines. There are Poisson and normal limit theorems in Griffiths (1984) for the distribution of $A_n^\theta(t)$ when $A_n^\theta(t) \to \infty$ as $t \to 0$ and $n \to \infty$, and $A_\infty^\theta(t)$ as $t \to 0$.

The time while the number of nonmutant ancestor lines is greater than or equal to k is distributed as $S_{n,k}^\theta = T_n + \cdots + T_k$, $1 \le k \le n$, where T_j are independent exponential random variables with rates $j(j+\theta-1)/2$, $j = n, \dots, 1$. The distribution function of $S_{n,k}^\theta$ satisfies

$$\mathrm{P}(S_{n,k}^\theta < t) = \mathrm{P}(A_n^\theta(t) \le k-1),$$

and the density satisfies

$$f_{n,k}^\theta(t) = \frac{k(k+\theta-1)}{2}\,\mathrm{P}(A_n^\theta(t) = k), \qquad t > 0.$$

The time to loss of the last line is of particular interest. If $\theta > 0$ then all lines are lost by coalescence or mutation until the last line, which is then lost by mutation at time $S_{n,1}^\theta$. If $\theta = 0$ then all lines are lost by coalescence, and $S_{n,2}^0$ is the time to the most recent common ancestor in the tree. The Laplace transform of $S_{n,k}^\theta$ is

$$\mathrm{E}[\mathrm{e}^{-\phi S_{n,k}^\theta}] = \prod_{j=k}^{n}\biggl[1 + \frac{2\phi}{j(j-1+\theta)}\biggr]^{-1}.$$

If $\theta \to \infty$ then all lines are lost by mutation, and the limit distributions of $\theta T_j/2$ are exponential with rates j. It follows directly that the limit density of $\theta S_{n,k}^\theta/2$ is

$$\frac{n!}{(k-1)!\,(n-k)!}\mathrm{e}^{-ks}(1-\mathrm{e}^{-s})^{n-k}, \qquad s > 0,$$

and that the limit distribution of $A_n^\theta(2s/\theta)$ is the binomial distribution

$$\binom{n}{k}\mathrm{e}^{-ks}(1-\mathrm{e}^{-s})^{n-k}, \qquad k = 0, \dots, n.$$

In this paper identities for the distributions of $A_n^\theta(t)$ and $A_\infty^\theta(t)$ are studied. The identities take the form of expected values of functions of $Z_t = \mathrm{e}^{\mathrm{i}X_t}$, where X_t is distributed as standard Brownian motion. The functions of X_t are periodic with period 4π, and X_t can be replaced by the wrapped Brownian motion $X_t \bmod 4\pi$. Associated identities are also found for the distributions of the time to the most recent common ancestor, the time until loss of ancestor lines by mutation, and the age of a mutation. Example identities follow.

The distribution of the number of nonmutant ancestor lineages in the population at time t back is, from (2.14),

$$\mathrm{P}(A_\infty^\theta(t) = j) = \mathrm{e}^{t/8}\frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}\,\mathrm{E}\biggl[\frac{Z_t^{-1/2}(\rho Z_t)^j(1-\rho Z_t)}{(1+\rho Z_t)^{2j+\theta}}\biggr] \tag{1.4}$$

for $j = 0, 1, \dots,$ where $Z_t = \mathrm{e}^{\mathrm{i}X_t}$ and $\rho = \mathrm{e}^{-\theta t/2}$.

The distribution of the time to the most recent common ancestor of the population, T°, is, from (2.34),

$$\mathrm{P}(T^\circ < t) = \mathrm{e}^{t/8}\,\mathrm{E}\biggl[\frac{Z_t^{-1/2}(1-\mathrm{e}^{-t}Z_t)}{(1+\mathrm{e}^{-t}Z_t)^2}\biggr], \tag{1.5}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

and the distribution of the time to the most recent common ancestor of a sample, T_n°, is, from (2.38),

$$\mathrm{P}(T_n^\circ < t) = \mathrm{e}^{t/8}\,\mathrm{E}[Z_t^{1/2}(1 - Z_t)(1 - VZ_t)^{n-2}], \tag{1.6}$$

where V is independent of Z_t with a beta$(2, n-2)$ distribution.

The distribution of the age of a mutation, ξ_p, observed to have frequency p in the current population is, from (2.49),

$$\mathrm{P}(\xi_p \le t) = \frac{\mathrm{e}^{t/8}}{2(1-p)}\,\mathrm{E}\biggl[Z_t^{-1/2}\frac{(1 - Z_t^2)}{R(Z_t)}\biggr], \tag{1.7}$$

where

$$R(Z_t) = [(1 + Z_t)^2 - 4(1 - p)Z_t]^{1/2}.$$

Hypergeometric functions play an important role in the identities. For example, the distribution of the number of nonmutant ancestor lineages in a sample of n genes at time t back is, from (2.5),

$$\begin{aligned}\mathrm{P}(A_n^\theta(t) = j) = {}& \frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j}\mathrm{e}^{(1/8)(\theta-1)^2 t} \\ &\times \mathrm{E}[Z_t^{(2j+\theta-1)/2}(1 - Z_t)F(-n + j + 1, \theta + 2j; n + j + \theta; Z_t)]\end{aligned} \tag{1.8}$$

for $j = 0, 1, \dots, n$. The identities are of mathematical interest as well as potentially being formulae to use for numerical integration or simulation to compute distributions that are usually expressed as alternating-sign series expansions such as (1.1), which are difficult to compute. Integrals where the normal density of X_t is replaced by the wrapped normal density of X_t mod 4π are easiest to use for numerical integration. This is discussed in Subsection 2.6. Each of the representations (1.4)–(1.8), for example, has the form $\mathrm{E}[g(t, X_t)]$, for a function g, and can be calculated using coupled simulation for different values of t by averaging values of $g(t, \sqrt{t}Y_j)$, $j = 1, \dots, r$, where $Y_1, \dots, Y_r$ are independent, identically $\mathrm{N}(0, 1)$-distributed random variables, and r is the number of replicates. The hypergeometric function in (1.8) can be evaluated by simulation from the representation (2.10), instead of formally using a series expansion.

The distributions of the number of nonmutant lineages at time t back in a coalescent tree of m genes, $A_m^\theta(t)$, and in a coalescent tree of a subsample of $n < m$ genes, $A_n^\theta(t)$, are shown to satisfy

$$\begin{aligned}\mathrm{P}(A_n^\theta(t) = j \mid A_m^\theta(t) = l) = {}& \frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)} \\ &\times \frac{(m-n)_{[l-j]}(m+\theta)_{(j)}}{m_{[l]}}\end{aligned} \tag{1.9}$$

for $j = l - (m - n), \dots, l$. In a subsample of n genes from the infinite-leaf coalescent tree,

$$\mathrm{P}(A_n^\theta(t) = j \mid A_\infty^\theta(t) = l) = \frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)} \tag{1.10}$$

for $j = 0, \dots, l$. Formulae (1.9) and (1.10) generalise the corresponding formulae for lineages in a subtree of a coalescent tree studied in Saunders *et al.* (1984). The latter formulae can be obtained by setting $\theta = 0$ in (1.9) and (1.10).

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

2. Complex integral formulae for lineage distributions

A series expansion of the form

$$\sum_{k=0}^{\infty} a_k \mathrm{e}^{-ck^2+dk}$$

with $c > 0$ and $d \in \mathbb{R}$ can be expressed as a convolution of the discrete series $\{a_k\}$ and a normal distribution. This gives a way to represent the series by an inversion formula for its Fourier transform. A lemma phrased in terms of probability distributions and characteristic functions that allows this is the following.

Lemma 2.1. *The convolution of a discrete distribution* $\{p_k\}$ *and an* $\mathrm{N}(0, \sigma^2)$ *distribution has a continuous density*

$$\sum_{k=0}^{\infty} p_k \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(1/2)(y-k)^2/\sigma^2}, \quad \textit{for } y \in \mathbb{R}. \tag{2.1}$$

Let $\phi(\zeta)$ *be the characteristic function* $\phi(\zeta) = \sum_{k=0}^{\infty} p_k \mathrm{e}^{\mathrm{i}k\zeta}$. *An inversion formula is then*

$$\sum_{k=0}^{\infty} p_k \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(1/2)(y-k)^2/\sigma^2} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}y\zeta} \phi(\zeta) \mathrm{e}^{-(1/2)\sigma^2\zeta^2} \,\mathrm{d}\zeta. \tag{2.2}$$

Proof. The density of the convolution is (2.1) with characteristic function

$$\phi(\zeta)\mathrm{e}^{-(1/2)\sigma^2\zeta^2}.$$

Identity (2.2) then follows from the inversion theorem for characteristic functions.

Corollary 2.1. *Let* $\{a_k\}$ *be an absolutely convergent complex series and let*

$$\phi(\zeta) = \sum_{k=0}^{\infty} a_k \mathrm{e}^{\mathrm{i}k\zeta}.$$

Then

$$\begin{aligned}\sum_{k=0}^{\infty} a_k \mathrm{e}^{-(t/2)k(k+\beta-1)} &= \mathrm{e}^{(t/8)(\beta-1)^2} \int_{-\infty}^{\infty} \mathrm{e}^{(\beta-1)\mathrm{i}\zeta/2} \phi(\zeta) \frac{1}{\sqrt{2\pi t}} \mathrm{e}^{-(1/2)\zeta^2/t} \,\mathrm{d}\zeta \\ &= \mathrm{e}^{(t/8)(\beta-1)^2}\, \mathrm{E}[\mathrm{e}^{(\beta-1)\mathrm{i}X_t/2} \phi(X_t)],\end{aligned} \tag{2.3}$$

where $\beta \in \mathbb{R}$ *and* $\{X_t,\ t \geq 0\}$ *is a standard Brownian motion process.*

Proof. It is clear that the inversion formula (2.2) continues to hold with $\{p_k\}$ replaced by a general absolutely convergent series $\{a_k\}$. Multiplying both sides of (2.2) by $\sqrt{2\pi}\sigma \mathrm{e}^{(1/2)y^2/\sigma^2}$ and taking $y = (1-\beta)/2$ and $\sigma^2 = 1/t$ gives (2.3). In the last line of (2.3), X_t is $\mathrm{N}(0, t)$-distributed.

Remark 2.1. A second proof of Lemma 2.1 is based on a form of the Parseval relation in harmonic analysis. Let U and F be probability distributions with respective characteristic functions ω and φ. Then

$$\int_{-\infty}^{\infty} \omega(\eta) F\{\mathrm{d}\eta\} = \int_{-\infty}^{\infty} \varphi(\zeta) U\{\mathrm{d}\zeta\}. \tag{2.4}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

See, for example Feller (1971, Chapter XIX). Choose U to be the $\mathrm{N}(0, 1/\sigma^2)$ distribution, with characteristic function $\omega(\eta) = \mathrm{e}^{-(1/2)\eta^2/\sigma^2}$, and F to be a discrete distribution with atoms $\{p_k\}$ at positions $\{-y + k\}$ and with characteristic function $\varphi(\zeta) = \mathrm{e}^{-\mathrm{i}y\zeta}\phi(\zeta)$. Apart from an overall factor of $\sqrt{2\pi}/\sigma$, (2.4) is identical to (2.2) after evaluating the left-hand side as a series.

Corollary 2.1 is now applied to find a representation for the lineage distribution $\mathrm{P}(A_n^\theta(t) = j)$ from the series (1.1).

Theorem 2.1. *The distribution of the number of nonmutant ancestor lineages at time t back in a coalescent tree of a sample of n genes has a complex integral representation*

$$\begin{aligned}\mathrm{P}(A_n^\theta(t) = j) = {} & \frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j}\mathrm{e}^{(1/8)(\theta-1)^2 t} \\ & \times \int_{-\infty}^{\infty} \mathrm{e}^{(1/2)(2j+\theta-1)\mathrm{i}x}(1-\mathrm{e}^{\mathrm{i}x}) \\ & \qquad \times F(-n+j+1, \theta+2j; n+j+\theta; \mathrm{e}^{\mathrm{i}x})\frac{1}{\sqrt{2\pi t}}\mathrm{e}^{-(1/2)x^2/t}\,\mathrm{d}x, \quad (2.5)\end{aligned}$$

for $j = 0, 1, \dots, n$ *if* $\theta > 0$ *and* $j = 1, 2, \dots, n$ *if* $\theta = 0$. *Here*

$$\begin{aligned}& F(-n+j+1, \theta+2j; n+j+\theta; z) \\ & \quad = B(2j+\theta, n-j)^{-1}\int_0^1 v^{2j+\theta-1}[(1-v)(1-vz)]^{n-j-1}\,\mathrm{d}v \qquad (2.6)\end{aligned}$$

is a hypergeometric function, with B denoting the beta function. If $j = n$ *then*

$$\mathrm{P}(A_n^\theta(t) = n) = \mathrm{e}^{-n(n+\theta-1)t/2}.$$

Proof. The representation holds from (2.3), with $\phi(\zeta)$ being the series obtained by replacing $\rho_k^\theta(t)$ in (1.1) by $\mathrm{e}^{\mathrm{i}k\zeta}$. Consider the generating function $z^j G_n(z)$ found by replacing $\rho_k^\theta(t)$ in (1.1) by z^k. After a shift of the index of summation in (1.1) by j and simplification, the resulting series is

$$\begin{aligned}G_n(z) &= \sum_{k=0}^{n-j} \frac{2k+2j+\theta-1}{j!\,k!}\frac{\Gamma(k+2j+\theta-1)}{\Gamma(j+\theta)} \\ &\qquad \times \frac{\Gamma(n+\theta)}{\Gamma(k+n+j+\theta)}\frac{n!}{(n-k-j)!}(-z)^k \\ &= \sum_{k=0}^{n-j} \frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{n}{j}\binom{n-j}{k}\frac{\Gamma(k+2j+\theta-1)}{\Gamma(k+n+j+\theta)} \\ &\qquad \times \frac{(n-j-k)(k+2j+\theta-1)+k(k+n+j+\theta-1)}{n-j}(-z)^k \\ &= \sum_{k=0}^{n-j} \frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{n}{j}[A_k + A_{k-1}](-z)^k \\ &= \sum_{k=0}^{n-j} \frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\frac{\Gamma(2j+\theta)}{\Gamma(n+j+\theta)}\binom{n}{j}[B_k - B_{k-1}]z^k,\end{aligned}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

where

$$A_k = \binom{n-j-1}{k}\frac{\Gamma(k+2j+\theta)}{\Gamma(k+n+j+\theta)}$$

and

$$B_k = \frac{(-n+j+1)_{(k)}(2j+\theta)_{(k)}}{(n+j+\theta)_{(k)}k!}.$$

Referring to the series expansion for hypergeometric functions (A.1), we see from this that

$$\begin{aligned} G_n(z) = (1-z)\binom{n}{j}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\frac{\Gamma(2j+\theta)}{\Gamma(n+j+\theta)} \\ \times F(-n+j+1, 2j+\theta; n+j+\theta; z). \end{aligned} \tag{2.7}$$

Substituting $\phi(\zeta) = \mathrm{e}^{\mathrm{i}j\zeta}G_n(\mathrm{e}^{\mathrm{i}j\zeta})$ into (2.3) with $\beta = \theta$ completes the proof.

Remark 2.2. There is a proof of Theorem 2.1 sketched in Section A.2 which uses the Brownian motion generator

$$L = \frac{1}{2}\frac{\partial^2}{\partial x^2}.$$

This is an interesting proof, though no easier than the main proof.

Remark 2.3. A general form of (2.5) is

$$\begin{aligned} \mathrm{P}(A_n^\theta(t) = j) = \frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j}\mathrm{e}^{(t/8)(\beta-1)^2} \\ \times \mathrm{E}[{U_t}^j(1-U_t)\mathrm{e}^{(\beta-1)\mathrm{i}X_t/2} \\ \times F(-n+j+1, \theta+2j; n+j+\theta; U_t)], \end{aligned} \tag{2.8}$$

for $j = 0, 1, \dots, n$ if $\theta > 0$ and $j = 1, 2, \dots, n$ if $\theta = 0$. Here $U_t = \mathrm{e}^{-\alpha t/2 + \mathrm{i}X_t}$, $\alpha + \beta = \theta$, and X_t is $\mathrm{N}(0, t)$-distributed. Equation (2.5) is recovered by setting $\alpha = 0$ and $\beta = \theta$.

This can be proved similarly to Theorem 2.1, absorbing $\mathrm{e}^{-(t/2)k\alpha}$ into the kth term of $G_n(z)$ and using the identity (2.3).

Remark 2.4. The relation

$$\begin{aligned} \sum_{j=0}^{n}\frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j}(1-z)z^j \\ \times F(-n+j+1, \theta+2j; n+j+\theta; z) = 1 \end{aligned} \tag{2.9}$$

holds pointwise for $z \in \mathbb{C}$ because the distribution of $A_n^\theta(t)$ sums to unity for $t \geq 0$ and the coefficients of the powers of U_t in (2.8) are unique when $\alpha = \theta - 1$ and $\beta = 1$.

Remark 2.5. An identity is

$$F(-n+j+1, 2j+\theta; n+j+\theta; z) = \mathrm{E}[(1-Vz)^{n-j-1}], \tag{2.10}$$

where V has a beta$(2j+\theta, n-j)$ distribution.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.6. The hypergeometric function in Theorem 2.1 has the special form $F(a, b; a-b+1; z)$. Identities for this form are given in Abramowitz and Stegun (1972, Equations 15.3.26–28). In particular, alternative expressions for (2.6) are found from (A.5):

$$\begin{aligned}
&F(-n+j+1, \theta+2j; n+j+\theta; z)\\
&\quad = (1+z)^{-(\theta+2j)} F\biggl(\frac{\theta}{2}+j, \frac{\theta+1}{2}+j; n+j+\theta; 4z(1+z)^{-2}\biggr) \qquad (2.11)\\
&\quad = \frac{(n+\theta)_{(j)}}{(\theta/2)_{(j)}((\theta+1)/2)_{(j)}}(1+z)^{-(\theta+2j)} F^{(j)}\biggl(\frac{\theta}{2}, \frac{\theta+1}{2}; n+\theta; 4z(1+z)^{-2}\biggr)\\
&\quad = \frac{\Gamma(n+\theta+j)}{\Gamma(n+\theta)}\frac{\Gamma(\theta)}{\Gamma(2j+\theta)} 2^{2j}(1+z)^{-(\theta+2j)} F^{(j)}\biggl(\frac{\theta}{2}, \frac{\theta+1}{2}; n+\theta; 4z(1+z)^{-2}\biggr).
\end{aligned}$$

Here $z = \mathrm{e}^{\mathrm{i}x}$ and $F^{(j)}$ denotes the jth derivative of F.

Corollary 2.2. *An identity is*

$$\mathrm{P}(A_n^{\theta}(t) = j) = \frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(n+j+\theta)\Gamma(j+\theta)}\binom{n}{j}\mathrm{e}^{-j(j+\theta-1)t/2}\,\mathrm{P}(A_{n-j}^{2j+\theta}(t) = 0), \qquad (2.12)$$

for $j = 0, 1, \dots, n$ *if* $\theta > 0$ *and* $j = 1, 2, \dots, n$ *if* $\theta = 0$. *By definition,* $A_0^{\theta}(t) \equiv 1$.

Proof. Identity (2.12) is a direct consequence of (2.5).

Corollary 2.3. *An identity is*

$$\mathrm{P}(A_{\infty}^{\theta}(t) = j) = \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}\mathrm{e}^{-j(j+\theta-1)t/2}\,\mathrm{P}(A_{\infty}^{\theta+2j}(t) = 0), \qquad (2.13)$$

for $j = 0, 1, \dots$ *if* $\theta > 0$ *and* $j = 1, 2, \dots$ *if* $\theta = 0$.

Proof. The corollary follows directly by taking the limit as $n \to \infty$ in (2.12).

Theorem 2.2. *The distribution of the number of nonmutant ancestor lineages at time t back in the coalescent tree of the population has a complex integral representation, if* $\theta > 0$, *given by*

$$\mathrm{P}(A_{\infty}^{\theta}(t) = j) = \mathrm{e}^{t/8}\frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}\int_{-\infty}^{\infty}\frac{\rho^j \mathrm{e}^{\mathrm{i}x(2j-1)/2}(1-\rho\mathrm{e}^{\mathrm{i}x})}{(1+\rho\mathrm{e}^{\mathrm{i}x})^{2j+\theta}}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \qquad (2.14)$$

where $j = 0, 1, \dots$ *and* $\rho = \mathrm{e}^{-\theta t/2}$.

Proof. An identity, found by taking the limit as $n \to \infty$ in (2.7), is

$$\begin{aligned}
G_{\infty}(z) &= \sum_{k=0}^{\infty}\frac{2k+2j+\theta-1}{j!\,k!}\frac{\Gamma(k+2j+\theta-1)}{\Gamma(j+\theta)}(-z)^k\\
&= (1-z)\frac{\Gamma(2j+\theta)}{j!\,\Gamma(j+\theta)}\sum_{k=0}^{\infty}\frac{(2j+\theta)_{(k)}}{k!}(-z)^k\\
&= (1-z)\frac{\Gamma(2j+\theta)}{j!\,\Gamma(j+\theta)}(1+z)^{-(2j+\theta)}.
\end{aligned}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

This is proved similarly to Theorem 2.1, by using the identity (2.3) formally with $\beta = 0$, i.e.

$$\begin{aligned}\sum_{k=0}^{\infty} a_k \mathrm{e}^{-(t/2)k(k-1)} &= \mathrm{e}^{t/8} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\zeta/2} \phi(\zeta) \frac{1}{\sqrt{2\pi t}} \mathrm{e}^{-\zeta^2/2t}\, \mathrm{d}\zeta \\ &= \mathrm{e}^{t/8}\, \mathrm{E}[\mathrm{e}^{\mathrm{i}X_t/2} \phi(X_t)], \end{aligned} \tag{2.15}$$

and letting

$$a_k = \frac{2k+2j+\theta-1}{j!\,k!} \frac{\Gamma(k+2j+\theta-1)}{\Gamma(j+\theta)} (-\rho)^k \quad \text{for } k = 0, 1, \dots .$$

Letting $\phi(\zeta) = \mathrm{e}^{\mathrm{i}j\zeta} G_\infty(\rho \mathrm{e}^{\mathrm{i}j\zeta})$ then completes the proof.

Remark 2.7. The distribution of $A_\infty^\theta(t)$ also has the representation

$$\mathrm{P}(A_\infty^\theta(t) = j) = \mathrm{e}^{(1/8)(\theta-1)^2 t} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta) j!} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{(1/2)\mathrm{i}x(2j+\theta-1)}(1-\mathrm{e}^{\mathrm{i}x})}{(1+\mathrm{e}^{\mathrm{i}x})^{2j+\theta}} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\, \mathrm{d}x. \tag{2.16}$$

The integral exists as an inversion formula for a Fourier transform, considering $(1+\mathrm{e}^{\mathrm{i}x})^{-(2j+\theta)}$ to be a transform of a signed measure with atoms of $(-1)^k(\theta+2j)_{(k)}/k!$ at points $k = 0, 1, \dots,$ and the normal transform

$$\mathrm{e}^{-x^2/2t} = \frac{1}{\sqrt{2\pi t^{-1}}} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}yx} \mathrm{e}^{-ty^2/2}\, \mathrm{d}y.$$

We prefer the representation (2.14) as it is a well-behaved integral whose integrand is bounded in absolute value and does not have singularities.

Formula (2.14) also holds in the sense of a Fourier inverse when $\theta = 0$. That is,

$$\mathrm{P}(A_\infty^0(t) = j) = \mathrm{e}^{t/8} \frac{(2j-1)!}{j!\,(j-1)!} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{(1/2)\mathrm{i}x(2j-1)}(1-\mathrm{e}^{\mathrm{i}x})}{(1+\mathrm{e}^{\mathrm{i}x})^{2j}} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\, \mathrm{d}x,$$

where $j = 1, \dots$.

Remark 2.8. Let $z \in \mathbb{C}$, $|z| < 1$, and $\theta \in \mathbb{R}$, $\theta \geq 0$. An identity is

$$\sum_{j=0}^{\infty} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta) j!} z^j (1-z)(1+z)^{-(2j+\theta)} = 1.$$

This identity holds because of (2.15) and because the distribution of $A_\infty^\theta(t)$ sums to unity. It can also be shown by letting $n \to \infty$ in (2.9). A direct algebraic proof is given in Lemma A.4. An important identity proved in Lemma A.3 is

$$(1-z)(1+z)^{-\alpha} = 1 + \sum_{k=1}^{\infty} (-1)^k (2k+\alpha-1) \frac{\alpha_{(k-1)}}{k!} z^k. \tag{2.17}$$

Another proof of Theorem 2.2 is to show (1.2) by termwise integration of the integrand in (2.14) based on the expansion (2.17).

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.9. The analogue of (2.14) for $A_n^\theta(t)$ is, from (2.8) with $\beta = 0$ and $\alpha = \theta$ and (2.11),

$$\begin{aligned}\mathrm{P}(A_n^\theta(t) = j) = {} & \frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j} \\ & \times \mathrm{e}^{t/8}\,\mathrm{E}\biggl[\mathrm{e}^{(1/2)\mathrm{i}X_t} Z_t^j(1-Z_t)(1+Z_t)^{-(\theta+2j)} \\ & \qquad \times F\biggl(\frac{\theta}{2}+j, \frac{\theta+1}{2}+j; n+j+\theta; 4Z_t(1+Z_t)^{-2}\biggr)\biggr] \qquad (2.18)\end{aligned}$$

for $j = 0, 1, \dots, n$, where $Z_t = \rho\mathrm{e}^{\mathrm{i}X_t}$, $\rho = \mathrm{e}^{-\theta t/2}$, and X_t is $\mathrm{N}(0, t)$-distributed.

Theorem 2.2 can be proved directly by taking the limit as $n \to \infty$ in (2.18), noting that then the hypergeometric function tends to 1 and the first factor tends to $\Gamma(2j+\theta)/\Gamma(j+\theta)j!$.

Corollary 2.4. *We have*

$$\mathrm{P}(A_\infty^\theta(t) = j) = \mathrm{e}^{-(1/2)j(j+\theta-1)t+(t/8)}\frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-(1/2)\mathrm{i}x}(1-\beta\mathrm{e}^{\mathrm{i}x})}{(1+\beta\mathrm{e}^{\mathrm{i}x})^{2j+\theta}}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \qquad (2.19)$$

where $\beta = \mathrm{e}^{-(1/2)(2j+\theta)t}$. *This holds for* $j = 0, 1, \dots$ *if* $\theta > 0$ *and* $j = 1, 2, \dots$ *if* $\theta = 0$.

Proof. The proof follows from (2.13) and (2.14).

2.1. Identities between lineage distributions of different sample sizes, and family sizes subtended by ancestor lineages

An identity between the distributions of $A_n^\theta(t)$ and $A_m^\theta(t)$, $n < m$, follows from the integral representation for the distribution, by algebraic proof. There is also a deeper probabilistic interpretation relating the two distributions. It is of interest to understand both approaches, so we derive our main identity in two ways.

Theorem 2.3. *For* $j = 0, 1, \dots, n$ *and* $n \le m$,

$$\begin{aligned}\mathrm{P}(A_n^\theta(t) = j) = {} & \frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\sum_{l=j}^{m-n+j}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)} \\ & \times \frac{(m-n)_{[l-j]}(m+\theta)_{(j)}}{m_{[l]}}\mathrm{P}(A_m^\theta(t) = l). \qquad (2.20)\end{aligned}$$

The limit form of (2.20) as $m \to \infty$ *is*

$$\mathrm{P}(A_n^\theta(t) = j) = \frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\sum_{l=j}^{\infty}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)}\mathrm{P}(A_\infty^\theta(t) = l). \qquad (2.21)$$

Proof. Make the following substitutions in (A.4): $a \leftarrow \theta/2 + j$, $b \leftarrow (\theta+1)/2 + j$, $c \leftarrow m+j+\theta$, and $m \leftarrow m-n$ (in the sense that, for example, $c-m \leftarrow m+j+\theta-(m-n) = n+j+\theta$). Then

$$\begin{aligned}\frac{a_{(k)}b_{(k)}}{c_{(k)}(c-m)_{(k)}} & := \frac{(\theta/2+j)_{(k)}((\theta+1)/2+j)_{(k)}}{(m+j+\theta)_{(k)}(n+j+\theta)_{(k)}} \\ & = 2^{-2k}\frac{\Gamma(2j+2k+\theta)}{\Gamma(\theta+2j)}\frac{\Gamma(m+j+\theta)}{\Gamma(m+j+\theta+k)}\frac{\Gamma(n+j+\theta)}{\Gamma(n+j+\theta+k)},\end{aligned}$$

by using the duplication formula (A.8). Identity (A.4) becomes

$$\begin{aligned} F\Big(\frac{\theta}{2}+j,\ &\frac{\theta+1}{2}+j;\theta+n+j;z\Big)\\ &=\sum_{k=0}^{m-n}\binom{m-n}{k}\frac{\Gamma(2j+2k+\theta)}{\Gamma(\theta+2j)}\frac{\Gamma(m+j+\theta)}{\Gamma(m+j+\theta+k)}\frac{\Gamma(n+j+\theta)}{\Gamma(n+j+\theta+k)}\\ &\qquad\times\Big(\frac{z}{4}\Big)^k F\Big(\frac{\theta}{2},\frac{\theta+1}{2};\theta+n+k;z\Big). \end{aligned} \tag{2.22}$$

Let

$$Q^{\theta}_{m,j}=\frac{\Gamma(m+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(m+j+\theta)}\binom{m}{j}F\Big(\frac{\theta}{2}+j,\frac{\theta+1}{2}+j;\theta+m+j;z\Big).$$

Then, by (2.22), we have

$$Q^{\theta}_{n,j}=\binom{n}{j}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\sum_{k=0}^{m-n}\frac{(m-n)_{[k]}(m+\theta)_{(j)}}{m_{[j+k]}}\frac{(j+k)!}{k!}\frac{\Gamma(k+j+\theta)}{\Gamma(n+j+k+\theta)}\Big(\frac{z}{4}\Big)^k Q^{\theta}_{m,j+k}. \tag{2.23}$$

The limit form of this identity as $m\to\infty$ is

$$Q^{\theta}_{n,j}=\binom{n}{j}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\sum_{k=0}^{\infty}\frac{(j+k)!}{k!}\frac{\Gamma(j+k+\theta)}{\Gamma(j+k+n+\theta)}\frac{\Gamma(2j+2k+\theta)}{\Gamma(j+k+\theta)j!}\Big(\frac{z}{4}\Big)^k.$$

A representation following from (2.11) and (2.5) is

$$\mathrm{P}(A^{\theta}_n(t)=j)=\mathrm{E}[(1-U_t)U_t^{j+(\theta-1)/2}(1+U_t)^{-(\theta+2j)}Q^{\theta}_{n,j}(V_t)],$$

where U_t is $\mathrm{N}(0,t)$-distributed and $V_t=4U_t(1+U_t)^{-2}$. Substituting (2.23) into this and shifting the summation index by j shows that (2.20) holds. The limit form (2.21) of (2.20) clearly holds.

Remark 2.10. Equations (2.20) and (2.21) have probabilistic interpretations. If a sample of n genes is taken from the leaves of a coalescent tree of $m\geq n$ genes, then the distributions of nonmutant lines in the two trees at time t back are related by the equation

$$\mathrm{P}(A^{\theta}_n(t)=j)=\sum_{l=j}^{m-n+j}\mathrm{P}(A^{\theta}_n(t)=j\mid A^{\theta}_m(t)=l)\,\mathrm{P}(A^{\theta}_m(t)=l).$$

The interpretation in (2.20) is that, for $j=l-(m-n),\dots,l$,

$$\begin{aligned}\mathrm{P}(A^{\theta}_n(t)=j\mid A^{\theta}_m(t)=l)&=\frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)}\\ &\quad\times\frac{(m-n)_{[l-j]}(m+\theta)_{(j)}}{m_{[l]}},\end{aligned} \tag{2.24}$$

and in (2.21) that, for $j=0,\dots,l$,

$$\mathrm{P}(A^{\theta}_n(t)=j\mid A^{\theta}_{\infty}(t)=l)=\frac{n!}{(n-j)!}\frac{\Gamma(n+\theta)}{\Gamma(j+\theta)}\binom{l}{j}\frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)}. \tag{2.25}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

The distribution of the number of lineages in a subtree of size n of a coalescent tree of m genes at time t back was studied in Saunders *et al.* (1984) and corresponds to the distribution (2.24) when $\theta = 0$.

We now describe the sampling of n genes from the leaves of a coalescent tree. This provides detail about the family size distribution subtended by ancestral lines in the sample, as well as independent proofs of (2.24) and (2.25).

Recall that $A_\infty^\theta(t)$ is the number of nonmutant lines at time t back in a coalescent tree with an infinite number of leaves. Given that $A_\infty^\theta(t) = l$, the relative family size distribution of the nonmutant families subtended by the l edges $(V_1, \dots, V_l)$ is Dirichlet with parameters $1, 1, \dots, 1, \theta$, with density

$$\frac{\Gamma(\theta + l)}{\Gamma(\theta)}\biggl(1 - \sum_{i=1}^{l} v_i\biggr)^{\theta-1}, \qquad 0 < v_1, \dots, v_l < 1,\ \sum_{i=1}^{l} v_i < 1. \tag{2.26}$$

The total relative frequency of mutant families is $1 - \sum_{i=1}^{l} V_i$. The relative sizes of new mutant families are distributed as are the atoms in a Poisson–Dirichlet point process $\mathrm{PD}(\theta)$, independent of $V_1, \dots, V_l$ (Griffiths (1980), Watterson (1984), Donnelly and Tavaré (1987)). Then (given that $A_\infty^\theta(t) = l$) a sample of n genes from the infinite-leaf coalescent tree has its own n-coalescent tree in which the number of nonmutant lines at time t back is $A_n^\theta(t)$, which is the number of family types from $1, 2, \dots, l$ represented in a multinomial sample from $V_1, V_2, \dots, V_l$. Let the labelled configuration of the number of genes in j nonmutant families in a sample of size n be $Q = (Q_1, \dots, Q_j)$. Thus,

$$\begin{aligned}
&\mathrm{P}(Q = q, A_n^\theta(t) = j \mid A_\infty^\theta(t) = l)\\
&\quad= \frac{n!}{q_1! \cdots q_j!\,(n - |q|)!}\binom{l}{j}\mathrm{E}\biggl[V_1^{q_1} \cdots V_j^{q_j}\biggl(1 - \sum_{r=1}^{l} V_r\biggr)^{n-|q|}\biggr]\\
&\quad= \frac{n!}{(n - |q|)!}\binom{l}{j}\frac{\theta_{(n-|q|)}}{(\theta + l)_{(n)}}.
\end{aligned} \tag{2.27}$$

Summing over the partition q of n into j nonzero parts with $j \le |q| \le n$, we have

$$\begin{aligned}
\mathrm{P}(|Q| = |q|,\ A_n^\theta(t) = j \mid A_\infty^\theta(t) = l) &= \binom{|q| - 1}{j - 1}\frac{n!}{(n - |q|)!}\binom{l}{j}\frac{\theta_{(n-|q|)}}{(\theta + l)_{(n)}}\\
&= \frac{\Gamma(j + \theta)}{(j - 1)!\,\Gamma(\theta)}\binom{n - j}{|q| - j}B(|q|, \theta + n - |q|)\\
&\quad\times \frac{n!}{(n - j)!}\frac{\Gamma(n + \theta)}{\Gamma(j + \theta)}\binom{l}{j}\frac{\Gamma(l + \theta)}{\Gamma(n + l + \theta)}.
\end{aligned} \tag{2.28}$$

Interpreting (2.28) as

$$\begin{aligned}
&\mathrm{P}(|Q| = |q|, A_n^\theta(t) = j \mid A_\infty^\theta(t) = l)\\
&\quad= \mathrm{P}(|Q| = |q| \mid A_n^\theta(t) = j)\,\mathrm{P}(A_n^\theta(t) = j \mid A_\infty^\theta(t) = l)
\end{aligned}$$

immediately gives

$$\mathrm{P}(|Q| = |q| \mid A_n^\theta(t) = j) = \frac{\Gamma(j + \theta)}{(j - 1)!\,\Gamma(\theta)}\binom{n - j}{|q| - j}B(|q|, \theta + n - |q|),$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

and (2.25) holds. It follows from (2.27), (2.28), and (2.25) that

(a) the sample configuration Q conditional on $A_n^\theta(t) = j$ is independent of the number of population lines $A_\infty^\theta(t)$, and

(b) the sample configuration Q conditional on $A_n^\theta(t) = j$ and $|Q| = |q|$ is uniformly distributed on the $\binom{|q|-1}{j-1}$ labelled partitions with $q_1 + \cdots + q_j = |q|$.

The analogue of (2.26) for a sample of size n is that, for $j \le |q| \le n$,

$$\begin{aligned}\mathrm{P}(Q = q \mid A_n^\theta(t) = j) &= \frac{\Gamma(j+\theta)}{(j-1)!\,\Gamma(\theta)}\binom{n-j}{|q|-j}B(|q|, \theta + n - |q|)\binom{|q|-1}{j-1}^{-1} \\ &= \frac{(n-j)!}{(n-|q|)!}\,\frac{\theta_{(n-|q|)}}{(\theta+j)_{(n-|q|)}}. \end{aligned} \tag{2.29}$$

Distribution (2.29) was derived in Watterson (1984). The distribution of the configuration of sizes of mutant families conditional on $|Q| = |q|$ is distributed according to Ewens' sampling formula in a sample of $n - |q|$ genes.

The distribution of Q when $\theta = 0$ (and, hence, $|Q| = n$) is the labelled ancestral partition distribution of Kingman (1982), where the n sample genes are partitioned into j classes with a uniform probability

$$\mathrm{P}(Q = q \mid A_n^\theta(t) = j) = \binom{n-1}{j-1}^{-1}, \qquad 1 \le j \le n.$$

Let Q be the family size distribution in a sample of m genes of l nonmutant lines at time t back. To have $A_n^\theta(t) = j$ in a subsample of n genes from the original m we require from the n genes a configuration $a = (a_1, \ldots, a_j)$ within families $i_1, \ldots, i_j$ of sizes $q_{i_1}, \ldots, q_{i_j}$. This probability is the same for all $\binom{l}{j}$ choices of families. Thus,

$$\mathrm{P}(A_n^\theta(t) = j \mid A_m^\theta(t) = l) \tag{2.30}$$

$$= \binom{l}{j} \sum_{a>0,\, q>0} \frac{\binom{q_1}{a_1}\cdots\binom{q_j}{a_j}\binom{m-|q|}{n-|a|}}{\binom{m}{n}}\,\mathrm{P}(Q = q \mid A_m^\theta(t) = l)$$

$$= \frac{\binom{l}{j}}{\binom{m}{n}}\,\frac{(m-l)!}{(\theta+l)_{(m-l)}}\,\frac{\theta_{(n-|a|)}}{(n-|a|)!} \tag{2.31}$$

$$\times \sum_{a>0,\, q>0} \binom{q_1}{a_1}\cdots\binom{q_j}{a_j}\frac{(\theta+n-|a|)_{(m-|q|-(n-|a|))}}{(m-|q|-(n-|a|))!}, \tag{2.32}$$

where summation is over nonzero entries of a and q, with $|a| \le n$ and $|q| \le m$. For a fixed a, the sum of (2.32) over $q > 0$ is recognised as the coefficient of s^m in the generating function

$$\left[\prod_{k=1}^{j} s^{a_k}(1-s)^{-(a_k+1)}\right]\left[\prod_{k=j+1}^{l} s(1-s)^{-1}\right]s^{n-|a|}(1-s)^{-(\theta+n-|a|)} = s^{n+l-j}(1-s)^{-(\theta+l+n)}.$$

That is, the sum is equal to

$$\frac{(\theta+l+n)_{(m-n-(l-j))}}{(m-n-(l-j))!}. \tag{2.33}$$

Also,

$$\sum_{a>0} \frac{\theta_{(n-|a|)}}{(n-|a|)!} = \frac{(\theta+j)_{(n-j)}}{(n-j)!},$$

which is the coefficient of s^n in $s^j(1-s)^{-j}(1-s)^{-\theta}$. To evaluate (2.30), collecting terms from (2.31) and (2.33) yields

$$\begin{aligned}
&\mathrm{P}(A_n^\theta(t) = j \mid A_m^\theta(t) = l)\\
&\quad = \frac{\binom{l}{j}}{\binom{m}{n}} \frac{(m-l)!}{(\theta+l)_{(m-l)}} \frac{(\theta+j)_{(n-j)}}{(n-j)!} \frac{(\theta+l+n)_{(m-n-(l-j))}}{(m-n-(l-j))!}\\
&\quad = \frac{n!}{(n-j)!} \frac{\Gamma(n+\theta)}{\Gamma(j+\theta)} \binom{l}{j} \frac{\Gamma(l+\theta)}{\Gamma(n+l+\theta)} \frac{(m-n)_{[l-j]}(m+\theta)_{(j)}}{m_{[l]}},
\end{aligned}$$

which is identical to (2.24).

2.2. Time to the most recent common ancestor

Corollary 2.5. *Let T° be the time to the most recent common ancestor of the population. The distribution function of T° is*

$$\mathrm{P}(T^\circ < t) = \mathrm{e}^{t/8} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-(1/2)\mathrm{i}x}(1-\beta \mathrm{e}^{\mathrm{i}x})}{(1+\beta \mathrm{e}^{\mathrm{i}x})^2} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \tag{2.34}$$

where $\beta = \mathrm{e}^{-t}$.

The density of T° is

$$3\mathrm{e}^{-(7/8)t} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-(1/2)\mathrm{i}x}(1-\beta \mathrm{e}^{\mathrm{i}x})}{(1+\beta \mathrm{e}^{\mathrm{i}x})^4} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \tag{2.35}$$

where $\beta = \mathrm{e}^{-2t}$.

Proof. The proof follows from $\mathrm{P}(T^\circ < t) = \mathrm{P}(A_\infty^0(t) = 1)$, the fact that the density of T° is $\mathrm{P}(A_\infty^0(t) = 2)$, and (2.19) with $\theta = 0$.

Corollary 2.6. *Let T^θ be the time when the last nonmutant ancestral line is lost from the population. The distribution function of T^θ is*

$$\mathrm{P}(T^\theta < t) = \mathrm{e}^{t/8} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-(1/2)\mathrm{i}x}(1-\beta \mathrm{e}^{\mathrm{i}x})}{(1+\beta \mathrm{e}^{\mathrm{i}x})^\theta} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \tag{2.36}$$

where $\beta = \mathrm{e}^{-\theta t/2}$.

The density of T^θ is

$$\tfrac{1}{2}\theta(\theta+1)\mathrm{e}^{-(\theta/2)t+t/8} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-(1/2)\mathrm{i}x}(1-\beta \mathrm{e}^{\mathrm{i}x})}{(1+\beta \mathrm{e}^{\mathrm{i}x})^{2+\theta}} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \tag{2.37}$$

where $\beta = \mathrm{e}^{-(1/2)(\theta+2)t}$.

Proof. The proof follows from $\mathrm{P}(T^\theta < t) = \mathrm{P}(A_\infty^\theta(t) = 0)$ and the fact that the density of T^θ is $(\theta/2)\,\mathrm{P}(A_\infty^\theta(t) = 1)$.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.11. Formulae analogous to (2.34)–(2.37) for a sample are easily derived from (2.5) and (2.10). If T_n° and T_n^θ are sample analogues then

$$\begin{aligned}\mathrm{P}(T_n^\circ < t) = \mathrm{e}^{t/8} \int_{-\infty}^{\infty} & \mathrm{e}^{(1/2)\mathrm{i}x}(1-\mathrm{e}^{\mathrm{i}x}) \\ & \times n(n-1)\int_0^1 v(1-v)^{n-2}(1-v\mathrm{e}^{\mathrm{i}x})^{n-2}\,\mathrm{d}v \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x \end{aligned} \tag{2.38}$$

and

$$\begin{aligned}\mathrm{P}(T_n^\theta < t) = \mathrm{e}^{(1/8)(\theta-1)^2 t} \int_{-\infty}^{\infty} & \mathrm{e}^{(\theta-1)\mathrm{i}x/2}(1-\mathrm{e}^{\mathrm{i}x}) \\ & \times \frac{\Gamma(\theta)\Gamma(n)}{\Gamma(n+\theta)}\int_0^1 v^{\theta-1}(1-v)^{n-1}(1-v\mathrm{e}^{\mathrm{i}x})^{n-1}\,\mathrm{d}v \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x. \end{aligned} \tag{2.39}$$

2.3. The probability generating function of $A_\infty^\theta(t)$

Corollary 2.7. *The probability generating function of* $A_\infty^\theta(t)$*,*

$$G_{A_\infty^\theta(t)}(s) = \mathrm{E}[s^{A_\infty^\theta(t)}],$$

is given by

$$G_{A_\infty^\theta(t)}(s) = 2^{\theta-1}\mathrm{e}^{t/8}\int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x}(1-\rho\mathrm{e}^{\mathrm{i}x})K^{-1}[1+\rho\mathrm{e}^{\mathrm{i}x}+K]^{-(\theta-1)}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x \tag{2.40}$$

for $\theta \geq 0$*, where* $\rho = \mathrm{e}^{-\theta t/2}$ *and* $K = [(1+\rho\mathrm{e}^{\mathrm{i}x})^2 - 4\rho s\mathrm{e}^{\mathrm{i}x}]^{1/2}$.

Proof. The proof follows from integral representation (2.14) of the distribution of $A_\infty^\theta(t)$ and the identity (A.12). Substituting

$$w = \frac{s\rho\mathrm{e}^{\mathrm{i}x}}{(1+\rho\mathrm{e}^{\mathrm{i}x})^2}$$

into the identity and simplifying gives (2.40).

Corollary 2.8. *The probability generating function of* A_∞^0 *is given by*

$$G_{A_\infty^0(t)}(s) = \mathrm{e}^{t/8}\int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x}\frac{1-\mathrm{e}^{2\mathrm{i}x}}{K^0}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \tag{2.41}$$

where $K^0 = [(1+\mathrm{e}^{\mathrm{i}x})^2 - 4s\mathrm{e}^{\mathrm{i}x}]^{1/2}$.

Proof. Substituting $\theta = 0$ into (2.40), simplifying, and noting that

$$\int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x}(1-\rho\mathrm{e}^{\mathrm{i}x})\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x = 0$$

gives (2.41).

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.12. The generating function for the Jacobi polynomials $P_n^{(\alpha,\beta)}(z)$ is

$$\begin{aligned} F^{(\alpha,\beta)}(z,r) &= \sum_{n=0}^{\infty} r^n P_n^{(\alpha,\beta)}(z) \\ &= 2^{\alpha+\beta} R^{-1}(1-r+R)^{-\alpha}(1+r+R)^{-\beta}, \end{aligned} \tag{2.42}$$

where $R = [1-2zr+r^2]^{1/2}$. The first two polynomials are

$$P_0^{(\alpha,\beta)}(z) = 1 \quad \text{and} \quad P_1^{(\alpha,\beta)}(z) = \tfrac{1}{2}[2(\alpha+1)+(\alpha+\beta+2)(z-1)].$$

The Jacobi polynomials are orthogonal on the distribution of $Z = 2X-1$, where X has a beta$(\alpha+1, \beta+1)$ distribution. The Legendre polynomials $\{P_n(z)\}$ are a special case with $\alpha = \beta = 0$ and generating function R^{-1}, and are orthogonal on the uniform distribution on $[-1, 1]$.

The generating function $G_{A_\infty^\theta(t)}(s)$, (2.40), is related to the Jacobi polynomial generating function (2.42). If we make the substitutions $\alpha \leftarrow 0$, $\beta \leftarrow \theta - 1$, $r \leftarrow \rho \mathrm{e}^{\mathrm{i}x}$, and $z \leftarrow 2s-1$, then

$$R = [1-2(2s-1)\rho \mathrm{e}^{\mathrm{i}x} + \rho^2 \mathrm{e}^{2\mathrm{i}x}]^{1/2} = [(1+\rho \mathrm{e}^{\mathrm{i}x})^2 - 4\rho s \mathrm{e}^{\mathrm{i}x}]^{1/2}$$

and we see that

$$\begin{aligned} G_{A_\infty^\theta(t)}(s) &= \mathrm{e}^{t/8} \int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x}(1-\rho \mathrm{e}^{\mathrm{i}x}) F^{(0,\theta-1)}(2s-1, \rho \mathrm{e}^{\mathrm{i}x}) \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x \\ &= \mathrm{e}^{t/8} \int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x}(1-\rho \mathrm{e}^{\mathrm{i}x}) 2^{\theta-1} R^{-1}(1+\rho \mathrm{e}^{\mathrm{i}x}+R)^{-(\theta-1)} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x \\ &= \sum_{n=0}^{\infty} [\mathrm{e}^{-n(n+\theta-1)t/2} - \mathrm{e}^{-(n+1)(n+\theta)t/2}] P_n^{(0,\theta-1)}(2s-1) \\ &= 1 + \sum_{n=1}^{\infty} \mathrm{e}^{-n(n+\theta-1)t/2}[P_n^{(0,\theta-1)}(2s-1) - P_{n-1}^{(0,\theta-1)}(2s-1)]. \end{aligned} \tag{2.43}$$

The first two terms in (2.43) evaluate to

$$1 - (\theta+1)\mathrm{e}^{-\theta t/2} + (\theta+1)\mathrm{e}^{-\theta t/2} s,$$

using

$$P_1^{(0,\theta-1)}(2s-1) = -\theta + (\theta+1)s.$$

If $\theta = 0$ then

$$\begin{aligned} G_{A_\infty^0(t)}(s) &= \frac{\mathrm{e}^{t/8}}{2} \int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}x} \frac{1-\mathrm{e}^{2\mathrm{i}x}}{R^0} \frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x \\ &= s + \frac{1}{2}\sum_{n=2}^{\infty} \mathrm{e}^{-n(n-1)t/2}[P_n(2s-1) - P_{n-2}(2s-1)], \end{aligned}$$

where $R^0 = [(1+\mathrm{e}^{\mathrm{i}x})^2 - 4s\mathrm{e}^{\mathrm{i}x}]^{1/2}$.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

From the identity (A.7) with $\alpha = 0$, we have

$$n(P_n^{(0,\theta-1)}(2x-1) - P_{n-1}^{(0,\theta-1)}(2x-1)) = -2\left(n + \frac{\theta-1}{2}\right)(1-x)P_{n-1}^{(1,\theta-1)}(2x-1), \quad (2.44)$$

allowing terms in the expansion (2.43) to be expressed differently.

An alternative form for the probability generating function of $A_\infty^\theta(t)$, based on (2.16), is

$$G_{A_\infty^\theta(t)}(s) = 2^{\theta-1}\mathrm{e}^{(1/8)(\theta-1)^2 t}\int_{-\infty}^{\infty}\mathrm{e}^{(1/2)(\theta-1)\mathrm{i}x}(1-\mathrm{e}^{\mathrm{i}x})K^{-1}[1+\mathrm{e}^{\mathrm{i}x}+K]^{-(\theta-1)}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x, \quad (2.45)$$

where now $K = [(1+\mathrm{e}^{\mathrm{i}x})^2 - 4s\mathrm{e}^{\mathrm{i}x}]^{1/2}$. The generating function is important in deriving both the distribution of the time to loss or fixation of an allele and the distribution of the age of a mutation.

2.4. The time to loss or fixation of an allele

In a diffusion process model for the frequency of an allele subject to loss by nonreversible mutation at rate $\theta/2$, the generator is

$$L_\theta = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} - \frac{1}{2}\theta x\frac{\partial}{\partial x}.$$

Let T_θ be the time to loss of an allele of initial frequency $1-p$. Then

$$\mathrm{P}(T_\theta \le t) = \sum_{j=0}^{\infty}\mathrm{P}(A_\infty^\theta(t) = j)p^j,$$

with a corresponding density of

$$f_{T_\theta}(t) = \frac{1}{2}\sum_{j=1}^{\infty} j(j+\theta-1)p^{j-1}(1-p)\,\mathrm{P}(A_\infty^\theta(t) = j).$$

The distribution is derived by arguing that either there are no nonmutant lines from the initial population, or the roots of the nonmutant lines in the initial population do not belong to the allele under consideration (Griffiths and Li (1983), Tavaré (1984), Ethier and Griffiths (1993)). The probability generating function of $A_\infty^\theta(t)$ immediately gives two representations, a complex integral from (2.45) and a Jacobi polynomial expansion from (2.43), for the distribution function of T_θ.

Theorem 2.4. *The distribution function of T_θ is*

$$\begin{aligned}\mathrm{P}(T_\theta \le t) &= G_{A_\infty^\theta(t)}(p)\\ &= 2^{\theta-1}\mathrm{e}^{(1/8)(\theta-1)^2 t}\int_{-\infty}^{\infty}\mathrm{e}^{(1/2)(\theta-1)\mathrm{i}x}(1-\mathrm{e}^{\mathrm{i}x})\\ &\qquad\qquad \times K^{-1}(1+\mathrm{e}^{\mathrm{i}x}+K)^{-(\theta-1)}\frac{\mathrm{e}^{-x^2/2t}}{\sqrt{2\pi t}}\,\mathrm{d}x\\ &= 1 + \sum_{n=1}^{\infty}\mathrm{e}^{-n(n+\theta-1)t/2}[P_n^{(0,\theta-1)}(2p-1) - P_{n-1}^{(0,\theta-1)}(2p-1)],\end{aligned}$$

where $K = [(1+\mathrm{e}^{\mathrm{i}x})^2 - 4p\mathrm{e}^{\mathrm{i}x}]^{1/2}$.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.13. The usual way to find $u(x, t) = \mathrm{P}(T_\theta \le t)$ from an initial frequency x is to solve the differential equation

$$\frac{\partial}{\partial t} u = L_\theta u, \tag{2.46}$$

with $u(0, t) = 1$. The solution to (2.46) is easily verified by using the identity for forward equations (1.3) (with $n = \infty$) on the left-hand side. The Jacobi polynomial solution can be derived by using an eigenvalue–eigenvector approach using a differential equation property of the Jacobi polynomials.

With

$$y_n = P_n^{(0,\theta-1)}(2x - 1) - P_{n-1}^{(0,\theta-1)}(2x - 1),$$

we have

$$L_\theta y_n = -\tfrac{1}{2} n(n + \theta - 1) y_n. \tag{2.47}$$

To verify (2.47), first consider the identity (2.44); then verify that the right-hand side of (2.44) satisfies (2.47) by obtaining an identical differential equation for $(1 - x) P_{n-1}^{(1,\theta-1)}(2x - 1)$ from (A.6).

Kimura derived eigenfunction expansions for transition distributions and associated absorption time distributions. A review of Kimura's research can be found in Watterson (1996). Kimura's solution of (2.47) is discussed in Tavaré (1984).

If $\theta = 0$ then there are two allele types subject to random drift. The probability that the allele type of initial frequency $1 - p$ is lost by time t is the probability that the allele of frequency p is fixed by time t, namely $G_{A_\infty^0(t)}(p)$.

2.5. The age of a mutation

The age, ξ_x, of a mutation observed to be at a current frequency x in a neutral model is known to have a distribution

$$\mathrm{P}(\xi_x \le t) = \sum_{j=1}^{\infty} (1 - x)^{j-1} \, \mathrm{P}(A_\infty^0(t) = j), \tag{2.48}$$

with a corresponding density of

$$f_{\xi_x}(t) = \frac{1}{2} x \sum_{j=2}^{\infty} j(j - 1)(1 - x)^{j-2} \, \mathrm{P}(A_\infty^0(t) = j)$$

(Griffiths and Tavaré (1998)).

An integral representation for the distribution of ξ_x follows directly from (2.48) and Theorem 2.4 with the substitutions $0 \leftarrow \theta$ and $1 - x \leftarrow p$.

Corollary 2.9. *We have*

$$\begin{aligned}\mathrm{P}(\xi_x \le t) &= \frac{\mathrm{e}^{t/8}}{2(1 - x)} \int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}y} \frac{1 - \mathrm{e}^{2\mathrm{i}y}}{R(x)} \frac{\mathrm{e}^{-(1/2t)y^2}}{\sqrt{2\pi t}} \, \mathrm{d}y \\ &= 1 + \frac{1}{2(1 - x)} \sum_{n=2}^{\infty} \mathrm{e}^{-n(n-1)t/2} [P_n(1 - 2x) - P_{n-2}(1 - 2x)],\end{aligned} \tag{2.49}$$

where $R(x) = [(1 + \mathrm{e}^{\mathrm{i}y})^2 - 4(1 - x)\mathrm{e}^{\mathrm{i}y}]^{1/2}$.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

The density of ξ_x *is*

$$f_{\xi_x}(t;x) = \frac{x}{2} G''_{A^0_\infty(t)}(1-x)$$
$$= 6\mathrm{e}^{t/8} x \int_{-\infty}^{\infty} \mathrm{e}^{(3/2)\mathrm{i}y} \frac{1-\mathrm{e}^{2\mathrm{i}y}}{R(x)^5} \frac{\mathrm{e}^{-(1/2t)y^2}}{\sqrt{2\pi t}}\,\mathrm{d}y.$$

Remark 2.14. The distribution of the age of a mutation observed to be of frequency b in a sample of n genes is, from Griffiths (2003, Equation (25)),

$$\frac{(n-2)!}{(b-2)!\,(n-b-1)!} \int_0^1 x^{b-1}(1-x)^{n-b}\,\mathrm{P}(\xi_x \le t)\,\mathrm{d}x.$$

Denoting the age by ξ_b^n, we have, directly from (2.49), that

$$\mathrm{P}(\xi_b^n \le t) = \frac{(n-2)!}{(b-2)!\,(n-b-1)!} \frac{\mathrm{e}^{t/8}}{2} \int_{-\infty}^{\infty} \mathrm{e}^{-(1/2)\mathrm{i}y}(1-\mathrm{e}^{2\mathrm{i}y}) \times \int_0^1 x^{b-1}(1-x)^{n-b-1} R(x)^{-1}\,\mathrm{d}x \frac{\mathrm{e}^{-(1/2t)y^2}}{\sqrt{2\pi t}}\,\mathrm{d}y.$$

2.6. Wrapped normal integral representations

Complex integral representations of real functions in this paper have the form $\mathrm{E}[G(\mathrm{e}^{\mathrm{i}X_t})]$, where X_t is $\mathrm{N}(0,t)$-distributed and $G(\zeta) = \zeta^{-1/2}H(\zeta)$, where $H(\zeta)$ has a Laurent series expansion for $|\zeta| \le 1$. The function $G(\mathrm{e}^{2\mathrm{i}x})$ is thus periodic with period 2π. Alternative integral representations are found using a wrapped normal distribution of $X_t/2 \bmod 2\pi$. The identities in (2.50) and (2.51) follow from the Poisson summation formula in complex analysis. The alternative forms are suitable for numerical evaluation: (2.50) for small values of v and (2.51) for larger values of v. The accuracy of truncated series can be found by making an integral comparison with the series tail and then applying the upper inequality from Equation (7.1.13) of Abramowitz and Stegun (1972):

$$\frac{1}{x+\sqrt{x^2+2}} < \mathrm{e}^{x^2} \int_x^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t \le \frac{1}{x+\sqrt{x^2+(4/\pi)}}, \qquad x \ge 0.$$

Lemma 2.2. *Let* X *be* $\mathrm{N}(0,v)$*-distributed and let* $Y = X \bmod 2\pi$*. Then* Y *has a density, for* $0 < y < 2\pi$*, given by*

$$q(y;v) = \frac{1}{\sqrt{2\pi v}} \sum_{k=-\infty}^{\infty} \mathrm{e}^{-(1/2)(y+2\pi k)^2/v} \tag{2.50}$$
$$= \frac{1}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \cos(ky)\mathrm{e}^{-k^2 v/2}, \qquad 0 < y \le 2\pi. \tag{2.51}$$

Corollary 2.10. *Let* X_t *be* $\mathrm{N}(0,t)$*-distributed. For a function,* $G(\zeta)$*, with domain* $\{z \in \mathbb{C} \colon |\zeta| \le 1\}$ *such that* $G(\mathrm{e}^{2\mathrm{i}x})$ *is periodic with period* 2π*, we have*

$$\mathrm{E}[G(\mathrm{e}^{\mathrm{i}X_t})] = \int_0^{2\pi} G(\mathrm{e}^{2\mathrm{i}y})q(y;t/4)\,\mathrm{d}y. \tag{2.52}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Remark 2.15. If

$$G(z) = z^{-1/2}(1-z)\sum_{l=0}^{\infty} a_l z^l,$$

then

$$\int_0^{2\pi} G(\mathrm{e}^{2\mathrm{i}y})q(y;t/4)\,\mathrm{d}y = \sum_{j=0}^{\infty}(a_{j+1}-a_j)\mathrm{e}^{-(2j+1)^2t/8},$$

and $q(y;t/4)$ can be replaced by

$$\tilde{q}(y;t/4) = \frac{1}{\pi}\sum_{j=1}^{\infty}\cos((2j+1)y)\mathrm{e}^{-(2j+1)^2t/8}$$

in (2.52), since the integral of terms of the form

$$\cos(2jy)\mathrm{e}^{-\mathrm{i}y}(1-\mathrm{e}^{2\mathrm{i}y})\mathrm{e}^{2\mathrm{i}ly}$$

vanishes for $j, l = 0, 1, \dots$.

Appendix A.

A.1. Hypergeometric function properties

Reference numbers AB are to equations in Abramowitz and Stegun (1972).

Definition A.1. (*AB 15.1.1*)

$$F(a,b;c;z) = \sum_{k=0}^{\infty}\frac{a_{(k)}b_{(k)}}{c_{(k)}}\frac{z^k}{k!}. \tag{A.1}$$

The series terminates at $k=m$ if $a=-m$.

An integral representation for $c > b > 0$ is (AB 15.3.1)

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\,\mathrm{d}t.$$

The hypergeometric function satisfies the differential formula (AB 15.2.2)

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k}F(a,b;c;z) = \frac{a_{(k)}b_{(k)}}{c_{(k)}}F(a+k,b+k;c+k;z) \tag{A.2}$$

and the differential equation (AB 15.5.1)

$$z(1-z)F'' + [c-(a+b+1)z]F' - abF = 0.$$

Lemma A.1. *(AB 15.2.4) The hypergeometric function satisfies*

$$\frac{\mathrm{d}^n}{\mathrm{d}z^n}[z^{c-1}F(a,b;c;z)] = (c-n)_{(n)}z^{c-n-1}F(a,b;c-n;z), \tag{A.3}$$

which is equivalent to

$$F(a,b;c-n;z) = \sum_{k=0}^{n}\binom{n}{k}\frac{a_{(k)}b_{(k)}}{c_{(k)}(c-n)_{(k)}}z^kF(a+k,b+k;c-n+k;z). \tag{A.4}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Proof. Equation (A.4) follows directly by expanding the derivative of the product of terms on the left-hand side of (A.3), applying the identity (A.2), and then simplifying the coefficient

$$\frac{(c-1)_{[n-k]}}{(c-n)_{(n)}}\frac{a_{(k)}b_{(k)}}{c_{(k)}} = \frac{a_{(k)}b_{(k)}}{c_{(k)}(c-n)_{(k)}}.$$

Remark A.1. Hypergeometric function identities for $F(a, b; a-b+1; z)$ are given in AB 15.3.26–28. The identity AB 15.3.26 is

$$F(a, b; a-b+1; z) = (1+z)^{-a}F(\tfrac{1}{2}a, \tfrac{1}{2}a+\tfrac{1}{2}; a-b+1; 4z(1+z)^{-2}).$$

By applying this identity we obtain

$$\begin{aligned}&F(-n+j+1, \theta+2j; n+j+\theta; z)\\ &\quad= F(\theta+2j, -n+j+1; n+j+\theta; z)\\ &\quad= (1+z)^{-(\theta+2j)}F\left(\frac{\theta}{2}+j, \frac{\theta+1}{2}+j; n+j+\theta; 4z(1+z)^{-2}\right).\end{aligned} \tag{A.5}$$

A limit that is seen directly from (A.5) is

$$\lim_{n\to\infty} F(-n+j+1, \theta+2j; n+j+\theta; z) = (1+z)^{-(\theta+j)},$$

because all the terms in the series expansion of the hypergeometric function on the right-hand side of (A.5) tend to 0, apart from the constant term, 1.

(AB 22.6.1). Let $y = P_n^{(\alpha,\beta)}$. The Jacobi polynomials satisfy the differential equation

$$(1-x^2)y'' + (\alpha-\beta-(\alpha+\beta+2)x)y' + n(n+\alpha+\beta+1)y = 0. \tag{A.6}$$

(AB 22.7.15). Furthermore,

$$\left(n+\frac{\alpha}{2}+\frac{\beta}{2}+1\right)(1-x)P_n^{(\alpha+1,\beta)}(x) = (n+\alpha+1)P_n^{(\alpha,\beta)}(x) - (n+1)P_{n+1}^{(\alpha,\beta)}(x). \tag{A.7}$$

(AB 6.1.18). The gamma function satisfies the following duplication formula:

$$\Gamma(2z) = (2\pi)^{-1/2}2^{2z-1/2}\Gamma(z)\Gamma(z+\tfrac{1}{2}).$$

Lemma A.2. *We have*

$$\frac{\Gamma(2j+\theta)}{\Gamma(\theta)} = 2^{2j}\frac{\Gamma(j+\theta/2)}{\Gamma(\theta/2)}\frac{\Gamma(j+(\theta+1)/2)}{\Gamma((\theta+1)/2)}. \tag{A.8}$$

Lemma A.3. *Let $z \in \mathbb{C}$ with $|z| < 1$ and $\alpha \in \mathbb{R}$ with $\alpha > 0$. An identity is*

$$(1-z)(1+z)^{-\alpha} = 1 + \sum_{k=1}^{\infty}(-1)^k(2k+\alpha-1)\frac{\alpha_{(k-1)}}{k!}z^k. \tag{A.9}$$

Proof. For $k \geq 1$, the coefficient of z^k in (A.9) is

$$\begin{aligned}(-1)^k\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!} - (-1)^{k-1}\frac{\Gamma(\alpha+k-1)}{\Gamma(\alpha)(k-1)!} &= (-1)^k\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}(\alpha+k-1+k)\\ &= (-1)^k(2k+\alpha-1)\frac{\alpha_{(k-1)}}{k!}.\end{aligned}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Lemma A.4. *Let $z \in \mathbb{C}$ with $|z| < 1$ and $\theta \in \mathbb{R}$ with $\theta \geq 0$. An identity is*

$$\sum_{j=0}^{\infty} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!} z^j (1-z)(1+z)^{-(2j+\theta)} = 1. \tag{A.10}$$

Proof. The constant term in the power series expansion of (A.10) is unity, evaluated from the term in the series on the left-hand side with $j = 0$, setting $z = 0$.

For $k \geq 1$, the coefficient of z^k in (A.10) is

$$\begin{aligned}
&\sum_{j=0}^{k} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!} \left\{ (-1)^{k-j} \frac{\Gamma(2j+\theta+k-j)}{\Gamma(2j+\theta)(k-j)!} - (-1)^{k-j-1} \frac{\Gamma(2j+\theta+k-j-1)}{\Gamma(2j+\theta)(k-j-1)!} \right\} \\
&\quad = \sum_{j=0}^{k} \frac{\Gamma(j+\theta+k-1)}{\Gamma(j+\theta)(k-j)!\,j!} (-1)^{k-j} \{(2j+\theta+k-j-1)+(k-j)\} \\
&\quad = \frac{(2k+\theta-1)}{k!} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} (j+\theta)_{(k-1)} \qquad (\text{A.11}) \\
&\quad = 0.
\end{aligned}$$

The sum (A.11) equals 0 because $(j+\theta)_{(k-1)}$ is a polynomial of degree $k-1$ in j which can be expressed as a linear sum in the basis elements $(j_{[l]},\ 0 \leq l \leq k-1)$. For any of the basis elements we have

$$\sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} j_{[l]} = k_{[l]} (1-1)^{k-l} = 0.$$

Lemma A.5. *Let $w \in \mathbb{C}$ with $|w| < \frac{1}{4}$ and $\theta \in \mathbb{R}$ with $\theta \geq 0$. An identity is*

$$\sum_{j=0}^{\infty} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!} w^j = 2^{\theta-1} (\sqrt{1-4w})^{-1} (1+\sqrt{1-4w})^{-(\theta-1)}. \tag{A.12}$$

Proof. Substitute

$$w = \frac{z}{(1+z)^2} \tag{A.13}$$

into (A.10). The left-hand side of (A.12) is then equal to

$$\frac{(1+z(w))^{\theta}}{1-z(w)}, \tag{A.14}$$

where

$$z(w) = \frac{1-\sqrt{1-4w}}{1+\sqrt{1-4w}}$$

is the solution to (A.13) such that $z(0) = 0$. Equation (A.14) simplifies to the right-hand side of (A.12).

Lemma A.6. *Let $w \in \mathbb{C}$ with $|w| < \frac{1}{4}$ and $\theta \in \mathbb{R}$ with $\theta \geq 0$. An identity is*

$$\sum_{j=0}^{\infty} \frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!} w^j = F\left(\frac{\theta}{2}, \frac{\theta+1}{2}; \theta; 4w\right). \tag{A.15}$$

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

Proof. By direct calculation,

$$\begin{aligned}\sum_{j=0}^{\infty}\frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}w^j &= \frac{\Gamma(\theta)}{\Gamma(\theta/2)^2}\sum_{j=0}^{\infty}2^{2j}w^j\frac{\Gamma(j+\theta/2)\Gamma(\theta/2)}{\Gamma(j+\theta)}\frac{\Gamma(j+(\theta+1)/2)}{\Gamma((\theta+1)/2)j!}\\ &= \sum_{j=0}^{\infty}\frac{(\theta/2)_{(j)}((\theta+1)/2)_{(j)}}{(\theta)_{(j)}}\frac{(4w)^j}{j!},\end{aligned}$$

as required.

Remark A.2. There are alternative hypergeometric function expressions for (A.15); see AB 15.3.15 to AB 15.23.18. The expressions follow from Kummer's quadratic transformation formulae for $F(a,b;2b;z)$. Equations (A.15) and (A.12) are seen to be identical from the identity AB 15.3.17,

$$F(a,b;2b;z) = \left(\frac12+\frac12\sqrt{1-z}\right)^{-2a}F\left[a,a-b+\frac12;b+\frac12;\left(\frac{1-\sqrt{1-z}}{1+\sqrt{1-z}}\right)^2\right],$$

with $a=(\theta+1)/2$, $b=\theta/2$, and $z=4w$. Since $a-b+\frac12=1$, $a=b+\frac12$, and $F(a,1;a;v)=(1-v)^{-1}$, (A.12) follows.

A.2. Alternative proof of Theorem 2.1

The generator corresponding to standard Brownian motion is

$$L=\frac12\frac{\partial^2}{\partial x^2}.$$

Thus, for suitable functions $Q(X_t)$, whose second derivatives exist, we have

$$\frac{\partial}{\partial t}\mathrm{E}[Q(X_t)] = \mathrm{E}[LQ(X_t)]. \tag{A.16}$$

Write (2.5) as

$$\mathrm{P}(A_n^\theta(t)=j) = \mathrm{e}^{(1/8)(\theta-1)^2t}\,\mathrm{E}[H_{n,j}^\theta(\mathrm{e}^{\mathrm{i}X_t})], \tag{A.17}$$

where X_t is N$(0,t)$-distributed. In (2.5), $\mathrm{P}(A_n^\theta(0)=j)=\delta_{n,j}$, the Kronecker delta, because, as $t\to 0$,

$$\mathrm{E}[H_{n,j}^\theta(\mathrm{e}^{\mathrm{i}X_t})] = \lim_{z\to1}(1-z)F(-n+j+1,\theta+2j;n+j+\theta;z) = \delta_{n,j}, \tag{A.18}$$

since the hypergeometric function in (A.18) is bounded as $z\to1$ if $j<n$ and equal to $(1-z)^{-1}$ if $j=n$. Applying (A.16) yields

$$\frac{\partial}{\partial t}\mathrm{e}^{(1/8)(\theta-1)^2t}\,\mathrm{E}[H_{n,j}^\theta(\mathrm{e}^{\mathrm{i}X_t})] = \mathrm{e}^{(1/8)(\theta-1)^2t}\,\mathrm{E}[G_{n,j}^\theta(\mathrm{e}^{\mathrm{i}X_t})],$$

where

$$G_{n,j}^\theta(z) = \tfrac18(\theta-1)^2H_{n,j}^\theta(z) - \tfrac12 zH_{n,j}^{\theta\prime}(z) - \tfrac12 z^2H_{n,j}^{\theta\prime\prime}(z). \tag{A.19}$$

It is enough to show the equivalence of (A.19) and

$$G_{n,j}^\theta(z) = -\frac{j(j+\theta-1)}{2}H_{n,j}^\theta(z) + \frac{(j+1)(j+\theta)}{2}H_{n,j+1}^\theta(z),$$

because then (1.3) holds, and the generator claimed for $A_n^\theta(t)$ is correct. We can equivalently show that

$$\frac{(j+1)(j+\theta)}{2}H_{n,j+1}^\theta(z)=\frac{1}{8}(\theta+2j-1)^2H_{n,j}^\theta(z)-\frac{1}{2}zH_{n,j}^{\theta\prime}(z)-\frac{1}{2}z^2H_{n,j}^{\theta\prime\prime}(z).\qquad\text{(A.20)}$$

Let

$$K_{n,j}^\theta(z)=z^{(\theta+2j-1)/2}(1-z)F(-n+j+1,\theta+2j;n+j+\theta;z).$$

Then

$$H_{n,j}^\theta=\frac{\Gamma(n+\theta)\Gamma(2j+\theta)}{\Gamma(j+\theta)\Gamma(n+j+\theta)}\binom{n}{j}K_{n,j}^\theta$$

and (A.20) is equivalent to

$$\frac{(2j+\theta)(2j+\theta+1)(n-j)}{n+j+\theta}K_{n,j+1}^\theta(z)=\frac{1}{4}(\theta+2j-1)^2K_{n,j}^\theta(z)-zK_{n,j}^{\theta\prime}(z)-z^2K_{n,j}^{\theta\prime\prime}(z).\qquad\text{(A.21)}$$

Now, $K_{n,j}^\theta(z)=K_{n-j,0}^{\theta+2j}$ because of its functional form, and the coefficients in (A.21) are also functions of $n-j$ and $\theta+2j$. It is therefore sufficient to show that (A.21) holds when $j=0$, because then we can replace n by $n-j$ and θ by $\theta+2j$ to obtain the general form. That is, it is sufficient to show that

$$\frac{n\theta(\theta+1)}{n+\theta}K_{n,1}^\theta(z)=\frac{1}{4}(\theta-1)^2K_{n,0}^\theta(z)-zK_{n,0}^{\theta\prime}(z)-z^2K_{n,0}^{\theta\prime\prime}(z).\qquad\text{(A.22)}$$

The proof now uses the hypergeometric identity (2.11) to express $K_{n,1}^\theta(z)$ (on the left-hand side) in terms of $K_{n,0}^\theta(z)$ and $K_{n,0}^{\theta\prime}(z)$. This is based on the equations

$$K_{n,0}^\theta(z)=z^{(\theta-1)/2}(1-z)(1+z)^{-\theta}F\biggl(\frac{\theta}{2},\frac{\theta+1}{2};n+\theta;\frac{4z}{(1+z)^2}\biggr),$$
$$K_{n,1}^\theta(z)=z^{(\theta+1)/2}(1-z)(1+z)^{-(\theta+2)}F\biggl(\frac{\theta}{2}+1,\frac{\theta+1}{2}+1;n+\theta+1;\frac{4z}{(1+z)^2}\biggr),$$

and

$$F'\biggl(\frac{\theta}{2},\frac{\theta+1}{2};n+\theta;\frac{4z}{(1+z)^2}\biggr)$$
$$=\frac{\theta(\theta+1)}{4(n+\theta)}\frac{4(1-z)}{(1+z)^2}F\biggl(\frac{\theta}{2}+1,\frac{\theta+1}{2}+1;n+\theta+1;\frac{4z}{(1+z)^2}\biggr).$$

Using the hypergeometric differential equation

$$z(1-z)F''+[n+\theta+(n-\theta-2)z]F'+(n-1)\theta F=0,$$

with $F=F(-n+1,\theta;n+\theta;z)$, we also express $K_{n,1}^{\theta\prime\prime}(z)$ (on the right-hand side) in terms of $K_{n,0}^\theta(z)$ and $K_{n,0}^{\theta\prime}(z)$, to show that (A.22) holds. Much algebraic manipulation shows that both sides of (A.22) are equal to

$$-nz(1-z)^{-1}(1+z)\biggl[\frac{\theta-1}{2}z^{-1}-(1-z)^{-1}-\theta(1+z)^{-1}\biggr]K_{n,0}^\theta(z)+nz(1-z)^{-1}(1+z)K_{n,0}^{\theta\prime}(z),\qquad\text{(A.23)}$$

completing the proof.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*. Dover, New York.

Donnelly, P. J. and Tavaré, S. (1987). The population genealogy of the infinitely many neutral alleles model. *J. Math. Biol.* **25,** 381–391.

Ethier, S. N. and Griffiths, R. C. (1993). The transition function of a Fleming–Viot process. *Ann. Prob.* **21,** 1571–1590.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3,** 87–112.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edn. John Wiley, New York.

Griffiths, R. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoret. Pop. Biol.* **17,** 37–50.

Griffiths, R. C. (1984). Asymptotic line-of-descent distributions. *J. Math. Biol.* **21,** 67–75.

Griffiths, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoret. Pop. Biol.* **64,** 241–251.

Griffiths, R. C. and Lessard, S. (2005). The Ewens sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theoret. Pop. Biol.* **68,** 167–177.

Griffiths, R. C. and Li, W.-H. (1983). Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. *Theoret. Pop. Biol.* **32,** 19–33.

Griffiths, R. C. and Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Stoch. Models* **14,** 273–295.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, eds D. Futuyama and J. Antonovics, Vol. 7, 2nd edn, Oxford University Press, pp. 1–44.

Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

Kingman, J. F. C. (1993). *Poisson Processes* (Oxford Stud. Prob. **3**). Clarendon Press, Oxford.

Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics*, eds D. J. Balding, M. Bishop and C. Cannings, John Wiley, Chichester, pp. 179–208.

Saunders, I. W., Tavaré, S. and Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* **16,** 471–491.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their application in population genetics models. *Theoret. Pop. Biol.* **26,** 119–164.

Watterson, G. A. (1984). Lines of descent and the coalescent. *Theoret. Pop. Biol.* **26,** 239–253.

Watterson, G. A. (1996). Motoo Kimura's use of diffusion theory in population genetics. *Theoret. Pop. Biol.* **49,** 154–188.

https://doi.org/10.1239/aap/1151337077 Published online by Cambridge University Press