The Equilibrium Allele Frequency Distribution for a Population with Reproductive Skew

Ricky Der[1] and Joshua B. Plotkin
Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

ABSTRACT We study the population genetics of two neutral alleles under reversible mutation in a model that features a skewed offspring distribution, called the Λ-Fleming–Viot process. We describe the shape of the equilibrium allele frequency distribution as a function of the model parameters. We show that the mutation rates can be uniquely identified from this equilibrium distribution, but the form of the offspring distribution cannot itself always be so identified. We introduce an estimator for the mutation rate that is consistent, independent of the form of reproductive skew. We also introduce a two-allele infinite-sites version of the Λ-Fleming–Viot process, and we use it to study how reproductive skew influences standing genetic diversity in a population. We derive asymptotic formulas for the expected number of segregating sites as a function of sample size and offspring distribution. We find that the Wright–Fisher model minimizes the equilibrium genetic diversity, for a given mutation rate and variance effective population size, compared to all other Λ-processes.

MANY questions in population genetics concern the role of demographic stochasticity and its interaction with mutation and selection in determining the fates of allelic types. The foundational work of Fisher, Wright, Haldane, Kimura (Wright 1931; Haldane 1932; Fisher 1958; Kimura 1994), and others has been instrumental in shaping our intuition about the powerful role that genetic drift plays in evolution and especially its role in maintaining diversity. This classical theory, which views genetic drift as a strong force, emanates from the Wright–Fisher model of replication and its large-population limit, the Kimura diffusion (Kimura 1955). The diffusion approximation has been particularly well studied, not only because it is mathematically tractable, but also because it is robust to variation in many of the underlying model details. Many discrete population-genetic models, including a large number of Karlin–Taylor and Cannings processes (Karlin and McGregor 1964; Cannings 1974; Ewens 2004), share the same diffusion limit as the Wright–Fisher model, and they therefore exhibit qualitatively similar behavior.

doi: 10.1534/genetics.114.161422
Manuscript received November 12, 2013; accepted for publication January 16, 2014; published Early Online January 28, 2014.
[1]Corresponding author: Carolyn Lynch Laboratories, Department of Biology, University of Pennsylvania, Philadelphia, PA 19103. E-mail: rickyder@sas.upenn.edu

Nevertheless, Kimura's classical diffusion is not appropriate in every circumstance. Its central assumption is the absence of skew in the reproduction process—that is, the assumption that no single individual can contribute a sizable proportion to the composition of the population in a single generation. Recent studies have suggested that this assumption is violated in several species, especially in marine taxa but also including many types of plants (Beckenbach 1994; Hedgecock 1994), whose mode of reproduction involves a heavy-tailed offspring distribution.

While the number of empirical studies on heavy-tailed offspring distributions is limited, there is a rich mathematical theory to describe the dynamics of populations with heavy reproductive skew. Beginning with Cannings' (1974) paper on neutral exchangeable reproduction processes, this literature has led to generalized notions of genetic drift, which subsume the traditional Wright–Fisherian concept of drift. The resulting forward-time continuum limits of such processes generalize the Kimura diffusion. One tractable class of models is the so-called Λ-Fleming–Viot processes, parameterized by a drift measure Λ (Donnelly and Kurtz 1999; Bertoin and Le Gall 2006), which we often refer to as simply "Λ"-processes or Λ-models. The corresponding backward-in-time, or coalescent, theory for such processes leads to the Λ-coalescents, first defined by Pitman and others (Pitman 1999; Sagitov 1999). Two conspicuous features stand out in this more general theory: Λ-processes may

have discontinuous sample paths, which feature "jumps" in the frequency of an allele, in contrast to the continuous sample paths of Kimura's diffusion. These jumps occur precisely at the occasional times when a small group of individuals contribute their genes to a sizeable fraction of the population in the next generation. Likewise, in the corresponding backward-time theory, the coalescents of such processes can exhibit multiple and even simultaneous mergers, instead of the strictly binary mergers of the classical Kingman coalescent.

Although mathematical aspects of such population processes (such as their construction, existence, uniqueness, etc.) have already been described, the specific population-genetic consequences of reproductive skew have only recently begun to be worked out. In many cases, the classical picture of population genetics must be considerably enlarged to accommodate new phenomena—see, for example, Möhle (2006) on generalizations of the Ewens sampling formula, Eldon and Wakeley (2006b) on linkage disequilibrium in processes with skewed offspring distributions, Birkner and Blath (2008) and Birkner *et al.* (2011) on inference and sampling in the Λ-coalescent, and Der *et al.* (2012) on the fixation probability of an adaptive allele in the Λ-process.

The purpose of this article is to study the stationary allele frequency distribution for populations with reproductive skew, under neutrality. When there are a finite number of allelic types subject to mutation, allele frequencies evolve to a unique stationary distribution, and our principle aim is to understand how this distribution depends on the form of reproductive skew, Λ, and how it may depart from the Wright–Fisherian picture.

Whereas a closed-form expression exists for the stationary allele frequency distribution in the (continuum) Wright–Fisher model, very few explicit expressions can be obtained in the general case of an arbitrary Λ-drift measure, corresponding to an arbitrary form of reproductive skew. Instead, we study the stationary distribution indirectly, by deriving a recurrence relation for its moments. This relation provides significant information about how the model parameters (θ, Λ) influence the stationary distribution. In particular, we demonstrate that the mutation parameters θ are identifiable from the stationary distribution; counterexamples exist, however, in which the form of drift measure Λ is not identifiable. We also introduce a simple estimator for θ, given samples from the equilibrium distribution, which we prove is consistent regardless of the skew in the underlying offspring distribution.

We also study how reproductive skew alters the standing genetic diversity in a population at equilibrium. Some numerical experiments of Möhle (2006), as well as some asymptotic results of Berestycki *et al.* (2007) for the β-coalescent, have suggested that the Wright–Fisher model tends to minimize standing diversity, compared to other offspring distributions. To analyze this behavior, we develop a Λ-version of Kimura's infinite-sites model, and we study the mean number of segregating sites, $\mathbb{E}S_n$, in a sample of size n. This measure of genetic diversity is robust in the sense that it is immune to many assumptions of the model and it coincides with the mean number of segregating sites in other infinite-sites models, including Watterson's fully linked infinite-sites model. We demonstrate that the Wright–Fisher model minimizes diversity among all Λ-processes of the same variance-effective population size. In other words, reproductive skew always tends to amplify standing genetic diversity, compared to the classical population-genetic model. We also derive a recursion formula for the mean number of segregating sites, and we use this to obtain asymptotic formulas for the number of segregating sites in large samples.

The remainder of the article is structured as follows. We start by reviewing discrete population models under reproductive skew. We then describe the forward-time continuum limits of such processes, which can be identified as Λ-Fleming–Viot processes. We develop a recursion equation for the moments of the stationary distribution in the two-allele case, and we use this to determine the identifiability of model parameters. To further examine equilibrium diversity we then introduce a two-allele, infinite-sites model with free recombination, and we study the frequency spectrum of samples from this process. This leads to a recursion formula for the mean number of segregating sites and theorems concerning the minimization and maximization of diversity among all Λ-measures. We conclude by providing a simple intuition for our results and by placing them in the context of the large literature on reproductive skew.

Discrete Population Models with Reproductive Skew

The Λ-Fleming–Viot processes are generalizations of the classical Wright–Fisher and Moran processes, which incorporate the possibility of large family sizes in the offspring distribution. The characteristics of these processes are most easily understood by studying them in a continuous-state, continuous-time setting, described below. Nonetheless, we first describe these models and review their properties in a discrete setting, along the lines of the treatment in Eldon and Wakeley (2006a).

We consider a population containing a fixed number N of individuals, each of two types. At every time step, a single individual is chosen uniformly from the population and produces a random number U of offspring, drawn from a distribution of offspring numbers, P_U, so that $P_U(i) = \mathbb{P}(U = i)$. The subsequent generation is then composed of the U offspring from the chosen individual supplemented by $N - U$ other individuals, randomly selected without replacement from the remainder of the population. Only a single individual contributes offspring in each reproduction event—the remaining individuals who neither contribute offspring nor die simply persist to the next time step.

When the offspring distribution P_U is concentrated at two individuals, *i.e.*, $\mathbb{P}(U = 2) = 1$, this model coincides with the

Moran process. More generally, we consider any discrete offspring number distribution P_U supported on the set $\{0, \ldots, N\}$.

To incorporate mutation, an additional stage is appended after reproduction wherein each individual may mutate to the opposing type, independently and identically with a probability that depends upon the individual's type, where the probability of a mutation from type 1 to type 2 is μ_1/N and μ_2/N with the mutation rate $\mu_i \geq 0$. This composite process is graphically depicted in Figure 1. We call this *discrete* process, which is a special case of the Cannings process (Cannings 1974), an "Eldon–Wakeley" model.

For our purposes, in the case of two alleles, it suffices to keep track only of the number of individuals in generation k of type 1, denoted by X_k. Since X_k is a Markov chain on the states $\{0, \ldots, N\}$, it possesses an associated transition matrix. We do not describe the specific form of the Markov transition matrix in this article, but we do focus on one important feature of this matrix: the variance of the allele frequency after one generation, after starting from a single mutant of type 1 in the population:

$$\sigma_N^2 = \mathrm{Var}[X_k | X_{k=1} = 1]. \tag{1}$$

This quantity, called the "offspring variance", determines the time-scaling of the continuum limit (see below), and it is related to the offspring number distribution P_U by

$$\sigma_N^2 = \frac{\mathbb{E}[U(U-1)]}{N-1}. \tag{2}$$

Continuum Approximations of Population Models with Reproductive Skew

Analysis of the Moran or Wright–Fisher model is often made easier by taking a continuum limit, which becomes accurate in the limit of large population size $N \to \infty$ (Kimura 1955; Ewens 2004). As described in Der *et al.* (2011), it is possible to define a significantly larger natural class of discrete population processes (those whose first two conditional moments coincide with that of the Wright–Fisher process), and derive their continuum limits, without restrictions on the offspring distribution P_U. This class of discrete processes contains the Cannings processes and the Eldon–Wakeley processes discussed above as special cases. The continuum theory for the Cannings case has been developed by Möhle (2001). While the limiting continuum processes are not, in general, diffusions with continuous sample paths, they are still characterized by an operator G, the infinitesimal generator of the continuum process, which reduces to the second-order differential equation of Kimura in the case of the Wright–Fisher model.

Continuum approximations involve choosing how to scale time and space, as $N \to \infty$. Such scalings replace the number i of individuals with the frequency $x = i/b_N$, and the generation number k by the time $t = kc_N$, for some choices of sequences $\{b_N\}$, $\{c_N\}$. The continuum limit is then the process

$$\widetilde{X}_t = \lim_{N \to \infty} \frac{1}{b_N} X_{[t/c_N]}. \tag{3}$$

In the classical Moran model we use the scalings $b_N = N$ and $c_N = N^{-2}$. In fact, it can be shown that the relationship between the space-scaling b_N and time-scaling c_N is fixed, in the sense that no other relationship leads to nontrivial limiting processes. We wish to study allele frequencies and hence impose the natural scaling $b_N = N$. The general theory (Der *et al.* 2011; Möhle 2001) then indicates that the time-scaling must be proportional to

$$c_N = \frac{\sigma_N^2}{N}, \tag{4}$$

where σ_N^2 is the offspring variance of the discrete-time process (2).

Once a time-scaling is fixed, then so is the appropriate scaling regime for the mutation rates, to produce a nontrivial balance of mutation and drift. This scaling must satisfy

$$\mu_i = O(\sigma_N^2). \tag{5}$$

In the classical Moran model $\sigma_N^2 = 2/N$, which produces the traditional scaling of mutation rate $\mu_i = O(N^{-1})$. In other models, such as those described in Eldon and Wakeley (2006a), where $\sigma_N^2 = N^{-\gamma+1}$ and $\gamma < 2$, mutation rates must scale faster to compensate for the increased rate of evolution from the drift process.

The limiting continuum process for an Eldon–Wakeley model

By applying the techniques of Möhle (2001) and Der (2010), we may derive the continuum limit for the Eldon–Wakeley process. These limits are characterized by an operator G and an associated Kolmogorov backward equation, analogous to the diffusion equation of Kimura. We consider a sequence of Eldon–Wakeley models, one for each population size N, and each with offspring distribution $P_U^{(N)}$. We assume the time-scaling and mutational constraints of (4) and (5) so that

$$\theta_i = \lim_{N \to \infty} \frac{2\mu_i}{\sigma_N^2} \tag{6}$$

defines the effective population-wide mutation rate. Under appropriate conditions on the sequence of offspring distributions $P_U^{(N)}$, there exists a limiting measure Λ that may be derived from $\left\{P_U^{(N)}\right\}$ as

$$\Lambda = \lim_{N \to \infty} \Lambda_N \quad \text{(weak limit)} \tag{7}$$

$$\Lambda_N\left(\frac{i}{N}\right) = \frac{1}{c_N}\left(\frac{i}{N}\right)^2 P_U^{(N)}(i), \quad i = 0, \ldots, N \tag{8}$$

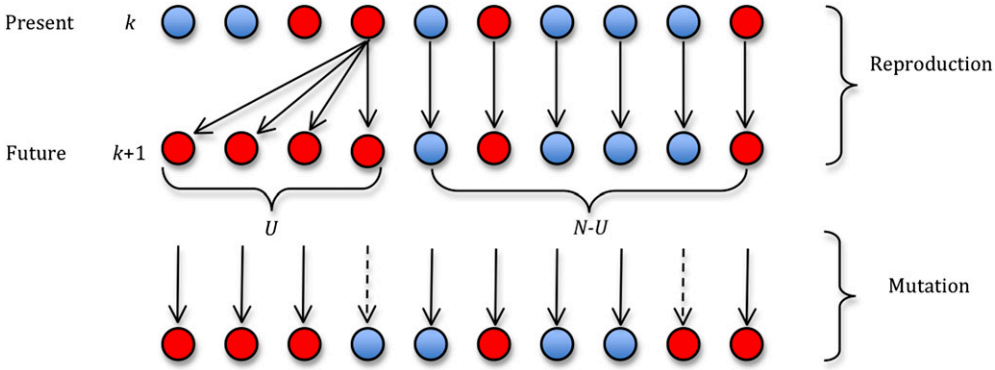

Figure 1 Schematic diagram of a discrete-time population model with reproductive skew.

and that characterizes the continuum limit. Letting $\tilde{X}_t^{(N)} = (1/N)X_{[t/c_N]}$, for $t \geq 0$, denote the time and state rescaled process, one can show then that $\tilde{X}_t^{(N)}$ converges to a limiting process $\tilde{X}_t$ on the allele frequency state-space [0, 1] as $N \to \infty$, which satisfies the backward equation

$$\frac{\partial u(x,t)}{\partial t} = Gu(x,t), \quad u(x,0) = f(x), \tag{9}$$

where

$$\begin{aligned} Gu(x) = {} & \tfrac{1}{2}(-\theta_1 x + \theta_2(1-x))\frac{\partial u}{\partial x} \\ & + \int_0^1 \frac{xu(x+(1-x)\lambda) - u(x) + (1-x)u(x-\lambda x)}{\lambda^2} \\ & \times \; d\Lambda(\lambda) \end{aligned} \tag{10}$$

and where $u(x,t) = \mathbb{E}[f(\tilde{X}_t)|\tilde{X}_0 = x]$.

The Markov process whose generator G is given by (10) is called the forward-time, two-type Λ-Fleming–Viot process (with mutation). In the sequel, we consider without any loss of generality only those processes for which Λ is a probability measure; *i.e.*, $\int_0^1 d\Lambda(\lambda) = 1$. This normalization can be thought of as restricting the space of models to those with the same offspring variance or, equivalently, the same rate of drift.

Intuitive remarks on the generator

As with the decomposition of the discrete Eldon–Wakeley process into a reproduction stage and mutation stage, the generator of (10) splits into two terms: a portion $(1/2)(-\theta_1 x + \theta_2(1-x))(\partial/\partial x)$ that describes mutation, independent of the reproduction measure Λ, and an integral portion that describes genetic drift. The term describing mutation coincides with the standard first-order advection term in Kimura's diffusion equation. The integral term, however, generally differs from the Kimura term, and it depends on the drift measure Λ.

Throughout the remainder of this article we distinguish several important families of Λ-processes. We define the *pure Λ-processes* to be those models for which $\Lambda = \delta_\lambda$, the Dirac measure concentrated at a single point λ, with $0 \leq \lambda \leq 1$. The terminology "pure" has been adopted from functional analysis, where it describes extreme points of a convex set—*i.e.*, the points that cannot be written as nontrivial mixtures of other points in the set.

Since (10) expresses the generator as an integral decomposition over such pure processes, we can think of a Λ-process as being a random mixture of these pure processes. Of particular interest are the extreme cases $\Lambda = \delta_0$ and $\Lambda = \delta_1$—which correspond to the Wright–Fisher process and the so-called "star" processes, respectively. As we will show, these two processes constrain the range of dynamics in Λ-models. Another well-studied family in the coalescent literature are the β-processes, for which Λ has a β-distribution (Berestycki *et al.* 2007).

One can interpret Λ as a jump measure controlling the frequency of large family sizes. If Λ is concentrated near zero, then jump sizes are small. In this regime, the integrand $(xu(x+(1-x)\lambda) - u(x) + (1-x)u(x-\lambda x))/\lambda^2$ behaves like the standard Kimura drift term $(1/2)x(1-x)u''(x)$. For the pure processes, where Λ is concentrated at the point λ, allele frequencies remain constant for an exponential amount of time, until a bottleneck event in which a fraction λ of the population is replaced by a single individual. Such events cause the allele frequency to increase instantaneously by the amount $(1-x)\lambda$ or decrease by λx. In the most general case of an arbitrary measure Λ, these behaviors are mixed, and the jump events occur at exponential times with mean dependent on $\lambda^{-2}\, d\Lambda(\lambda)$ and are associated with jumps of random size λ.

If Λ places a large mass near zero, the process becomes diffusion-like, with sample paths exhibiting frequent, small jumps. On the other hand, if Λ is mostly concentrated away from zero, then allele dynamics are of the "jump and hold" type, with fewer, but more sizable, jumps. Such behavior is most extreme in the star model, whose sample paths are constant until a single jump to absorption.

The Stationary Distribution of Λ-Processes

In the absence of mutation, allele frequencies must eventually fix at 0 or 1, and thus any discrete Eldon–Wakeley model possesses a trivial stationary distribution whose concentration at the absorbing states $\{0, N\}$ depends on the initial condition. When mutation rates μ_i are strictly positive, however, each Eldon–Wakeley process in a population

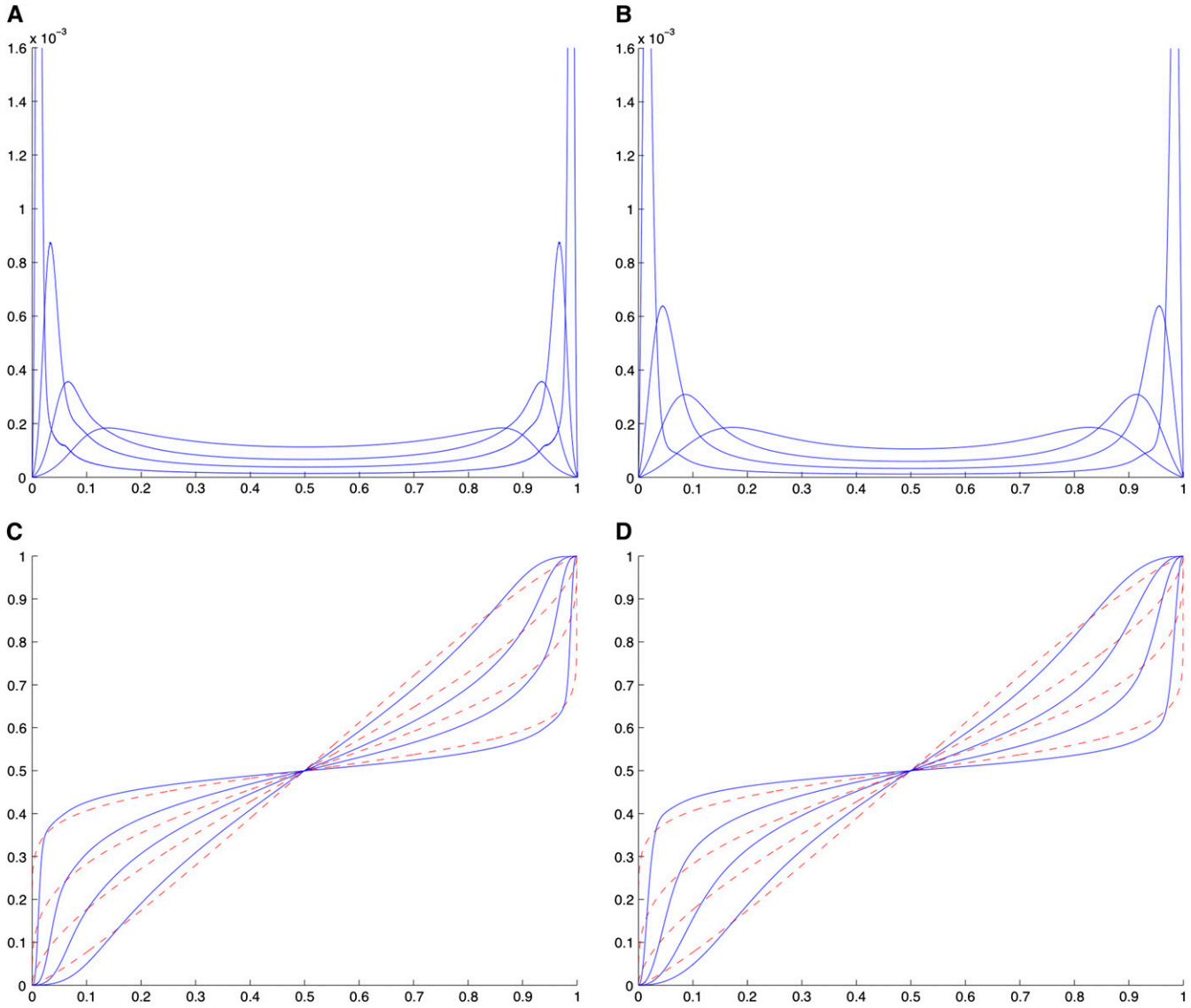

Figure 2 Stationary distributions for the β-process (solid lines) and Wright-Fisher process (dashed lines). (A and B) Stationary densities. (C and D) Stationary cumulative distribution functions. (A and C) Beta process (solid lines) with Beta density parameters $\alpha = 0.7$, $\beta = 1$. (B and D) Beta process (solid lines) with parameters $\alpha = 1.3$, $\beta = 1$. Mutation values are $\theta = 0.1, 0.3, 0.6, 1.2$. Population size $N = 8000$. Note that for a given mutation rate, the Wright–Fisher process concentrates more mass at the boundaries than the β-process.

size N possesses a unique, nontrivial stationary distribution, π_N, to which the process converges, regardless of the initial condition.

In Figure 2, we plot the stationary distributions for a few Λ-processes, in the case of symmetric mutation $\theta_1 = \theta_2$. Generally, these distributions have the same qualitative dependence on the mutation rate as the classical Wright–Fisher stationary distribution: they continuously progress from Dirac singularities at the boundaries to distributions concentrated more in the center of the interval, as the mutation rate increases. It is interesting to observe, however, that the non-Wright–Fisherian processes tend to have more mass at intermediate allele frequencies, and less relative mass near the boundaries, than the Wright–Fisherian model. We study this phenomenon more precisely below.

The continuum limit $\tilde{X}$ of a sequence of Eldon–Wakeley processes also possesses a unique stationary distribution, π. In the *Appendix*, we demonstrate that

$$\pi_N \to \pi, \quad \text{as } N \to \infty. \tag{11}$$

In other words, the sequence of discrete equilibrium measures converges to the continuum equilibrium distribution. As a result, we can use the continuum equilibrium as a good approximation in large populations.

Moments of the stationary distribution

In the case of two alleles, the stationary allele frequency distribution describes the likelihood of finding the mutant allele at any given frequency, if the process started a long

time ago. We study the moments of the stationary distributions for Λ-processes, using a version of the Fokker–Planck equation, analogous to the equation used by Kimura to study the stationary distribution of the Wright–Fisher process. In general, a stationary distribution π of a Markov process with generator G is the solution to its so-called adjoint Fokker–Planck equation, so that

$$\int_0^1 Gu(x)\, d\pi = 0, \tag{12}$$

for every smooth function u on [0, 1]. We take G as the generator for the Λ-process with mutation, given by (10). Although it is difficult to solve for π in general, this equation can nonetheless be used to obtain detailed information about the stationary distribution.

To begin, we develop formulas for the moments of the stationary distribution, which allow us to characterize aspects of standing genetic diversity. This derivation is similar in spirit to the generator approach contained in Birkner and Blath (2009), specialized to the case of two alleles.

Let m_k denote the kth moment of π; that is, $m_k = \int_0^1 x^k\, d\pi(x)$. Setting $u(x) = x$ into (12) yields an equation for the mean value of the equilibrium, so that

$$m_1 = \frac{\theta_2}{\theta_1 + \theta_2}. \tag{13}$$

Next, setting $u(x) = x^2$ into (12) yields a relation between m_2 and m_1 that can be solved to give

$$m_2 = \frac{(1+\theta_2)\theta_2}{(\theta_1+\theta_2)(1+\theta_1+\theta_2)}. \tag{14}$$

This recursive process can be continued, because the generator G of (10) maps polynomials of degree k to polynomials of degree k. Thus, we can derive a system of equations that define the moments of π. In the *Appendix*, we show that this recursion has the form

$$m_k = \frac{\big((k/2)\theta_2 + a_{k-1,k}\big)m_{k-1} + \sum_{j=1}^{k-2} a_{jk} m_j}{(k/2)(\theta_1+\theta_2) + a_{kk}}, \tag{15}$$

where the coefficients $\{a_{jk}\}$, $1 \le j \le k$, $k = 1, 2, \ldots,$ are functions of Λ and are given by

$$a_{kk} = \int_0^1 \frac{1-(1-\lambda)^k - k\lambda(1-\lambda)^{k-1}}{\lambda^2} d\Lambda(\lambda) \tag{16}$$

$$a_{j,k} = \binom{k}{j-1} \int_0^1 \lambda^{k-j-1}(1-\lambda)^{j-1} d\Lambda(\lambda), \quad j = 1, \ldots, k-1. \tag{17}$$

Initializing this system by (13), and observing that $a_{kk} > 0$, we see that (15) uniquely determines the moments of π, and indeed this equation can be used to solve for any specific moment of the stationary distribution. While it does not appear that the moments m_k can be solved explicitly to produce simple, closed-form expressions as functions of the Λ-measure, it is clear that the coefficients a_{jk} are all linear combinations of moments of Λ. Moreover, each moment m_k is always a ratio of polynomials in θ_1 and θ_2.

Identifiability of parameters from equilibrium

One of the most important questions about the stationary allele frequency distribution is what population-genetic parameters can be identified from it—that is, which parameters of the population can be uniquely determined from data sampled in equilibrium? In the case of Λ-processes, the parameters we might wish to infer are the mutation rates, θ_1 and θ_2, as well as the (high-dimensional) drift measure, Λ, which describes the offspring distribution.

The first two moments of the stationary distribution are given by (13) and (14), and they are independent of the drift measure, Λ. Thus, the first- and second-order moments of the stationary distribution for all Λ-processes and any function of these moments, such as the second-order heterozygosity $\int_0^1 x(1-x)d\pi(x)$, must coincide with those of the classic Wright–Fisher model. In the case of symmetric mutation $\theta_1 = \theta_2 = \theta$, the *third* moment is also, remarkably, constant across the Λ-processes and has the value

$$m_3 = \frac{2+\theta}{4+8\theta}. \tag{18}$$

The constancy of the first two moments with respect to Λ, and the fact that the mapping from the first two moments to the two mutation parameters $(m_1, m_2) \mapsto (\theta_1, \theta_2)$ given by (13) and (14) is one-to-one, allows us to conclude that, regardless of the underlying reproductive process, the mutation rates (θ_1, θ_2) are always identifiable from the equilibrium distribution. This is a productive result—because it means that we can always infer mutation rates from sampled data, even when the offspring distribution of a species is unknown to us.

Conversely, we may ask whether we can identify the form of reproductive process without knowledge of mutation rates—that is, Is Λ always uniquely identifiable from the stationary distribution alone? It turns out that the answer is negative, as we demonstrate with the following simple example.

Consider the star process $\Lambda = \delta_1$ with mutation parameters $\theta_1 = \theta_2 = \theta$. The generator for this process is

$$Gu(x) = \frac{1}{2}\theta(1-2x)u'(x) + (1-x)u(0) - u(x) + xu(1). \tag{19}$$

The associated stationary distribution π_1 is easily derived (see *Appendix*), and it has a density $d\pi_1/dx$ given by

$$\frac{d\pi_1}{dx} = \frac{1}{\theta}|1-2x|^{(1-\theta)/\theta}. \tag{20}$$

For comparison, the Kimura diffusion has a Dirichlet-type stationary distribution π_0:

$$\frac{d\pi_0}{dx} = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2} x^{\theta-1}(1-x)^{\theta-1}. \tag{21}$$

Note that both distributions (20) and (21) coincide when $\theta = 1$, despite the enormous difference between the drift measures of these processes. Hence, the map from a given two-allele Λ-process to its stationary distribution is not one-to-one, and consequently the drift measure Λ cannot always be identified from the stationary distribution, *e.g.*, when θ is near unity. Nonetheless, the Λ-measure is often identifiable under an infinite-sites model with a small per-site mutation rate, as discussed below.

An estimator for θ from the equilibrium distribution

Here we introduce a method to estimate the mutation rate, θ, from data sampled from the two-way equilibrium distribution. Because we know that the first two moments of the equilibrium distribution do not depend on the underlying Λ-measure, we will be able to construct a simple and robust estimate for the mutation rate that is consistent regardless of the underlying skew in the offspring distribution.

Specifically, let us consider the case of a Λ-process undergoing symmetric two-way mutation, $\theta = \theta_1 = \theta_2$, and suppose we are provided with i.i.d. samples $X_1, \ldots, X_T$ drawn from π, its equilibrium distribution. The first two moments of π are

$$m_1 = \frac{1}{2} \tag{22}$$

$$m_2 = \frac{1+\theta}{2(1+2\theta)} \tag{23}$$

and thus the variance of the equilibrium distribution is

$$v(\theta) = m_2 - m_1^2 = \frac{1}{4(1+2\theta)}, \tag{24}$$

which gives a map between θ and the variance, v. Therefore, a natural method-of-moments estimator for θ is

$$\hat{\theta} = g\left(\frac{1}{T}\sum_{i=1}^{T}\left(X_i - \frac{1}{2}\right)^2\right), \tag{25}$$

where $g(v) = \frac{1}{8v} - \frac{1}{2}$.

In the *Appendix*, we prove the following two theorems concerning the consistency and the asymptotic variance of this estimator:

Theorem 1. *The estimator $\hat{\theta}$ is consistent for θ.*

Theorem 2. *The deviation $\sqrt{T}(\hat{\theta} - \theta)$ is asymptotically ($T \to \infty$) normally distributed with mean zero and variance*

$$\left(\frac{1}{8}\log(4+8\theta)\right)^2(\mu_4 - \mu_2^2), \tag{26}$$

where μ_i is the ith central moment of the equilibrium distribution.

The critical point, here, is that the estimator $\hat{\theta}$ is consistent, regardless of the nuisance parameter Λ—allowing us to infer mutation rates in the absence of information on the form of reproductive skew in the population. Nonetheless, Theorem 2 shows that the rate of convergence of this estimator does depend on the fourth moment of equilibrium distribution and hence on the reproduction measure Λ.

An Infinite-Sites Model for the Λ-Processes

To study how reproductive skew influences standing genetic diversity, we now develop an infinite-sites version of the Λ-process and study its equilibrium behavior. This model generalizes the infinite-sites approach of Desai and Plotkin (2008) and RoyChoudhury and Wakeley (2010), for the Wright–Fisher model. We study the sampled site frequency spectrum of our model, under two-way mutation. Our analysis allows us to quantify our previous observation that the Wright–Fisher model minimizes the amount of standing genetic diversity, among all Λ-processes. The site frequency spectrum that we describe in this section, for independent sites, differs from the Watterson-type spectrum for fully linked sites; but our approach nonetheless yields information in that case as well.

We consider an evolving population of large size N, following the reproduction dynamics of a neutral forward-in-time Λ-process, for a fixed Λ-measure. We keep track of L sites along the genome, each with two possible allelic types under symmetric two-way mutation at rates $\theta = \theta_1 = \theta_2$. The allele dynamics at each site are described by a two-type Λ-process; and the site processes are assumed independent of one another (that is, we assume free recombination).

Let π_θ denote the two-allele stationary distribution for the Λ-model, given by (12), where the subscript denotes the explicit dependence on the mutation rate. We imagine sampling n individuals from the population at equilibrium, assuming $n \ll N$. We let Y_i, $1 \le i \le L$ represent the (random) number of sampled individuals carrying a particular type at site i, so that their joint distribution has the form

$$P(Y_1 = y_1, \ldots, Y_L = y_L) = \prod_{i=1}^{L}\int_0^1 \binom{n}{y_i} x^{y_i}(1-x)^{n-y_i} d\pi_\theta(x). \tag{27}$$

The sampled site frequency spectrum (Sawyer and Hartl 1992; Bustamante *et al.* 2001) is defined as the vector $(Z_0, \ldots, Z_n)$

$$Z_k = \sum_{i=1}^{L} 1_{Y_i = k}, \quad k = 0, \ldots, n. \tag{28}$$

The variables Z_k record the number of sites with precisely k (of n) sampled individuals of a given allelic type. In this sense, the sampled site frequency spectrum represents

a discretized version of the stationary distribution π_θ. Equation 28 implies that the variables $(Z_0, \ldots, Z_n)$ are distributed multinomially on the simplex $\sum_{k=0}^{n} Z_k = L$. The sites $Z_1, \ldots, Z_{n-1}$ are called the *segregating sites*, representing locations where there is diversity observed in the the sample. Conversely, the sum $Z_0 + Z_n$ represents the number of monomorphic sites in the sample.

The infinite-site limit and its Poisson representation

To study the sampled site frequency spectrum we take the limit of an infinite number of sites, $L \to \infty$, and we apply a Poisson approximation. We define the genome-wide mutation rate as $\Theta_L = L \cdot \theta$, and we assume that this mutation rate approaches a constant in the limit of many sites: $\Theta_L \to \Theta < \infty$. In the *Appendix*, we show that the segregating site variables $(Z_1, \ldots, Z_{n-1})$ then converge, as $L \to \infty$, to a sequence of independent Poisson random variables with means $(c_1\Theta, \ldots, c_{n-1}\Theta)$, given by

$$c_j(n) = \lim_{\theta \to 0} \frac{1}{\theta} \int_0^1 \binom{n}{j} x^j (1-x)^{n-j} d\pi_\theta(x), \quad j = 1, \ldots, n-1. \tag{29}$$

The behavior of the integral term in the above is studied in Lemma 1, Equation A20. There, it is shown that for small θ, the integral has linear order near zero, and so the above limit is well defined. The numbers c_j may be interpreted as an infinite-sites sample frequency spectrum. From (29), it is apparent that the means c_j depend on the heterozygotic moments of π and thus also on the moments of Λ. This representation is thus a generalization of a result of RoyChoudhury and Wakeley (2010) for the two-allele Wright–Fisher independent-sites model, where $\Lambda = \delta_0$, and where the spectrum c_j has the form $c_j = (1/2)(n/\, j(n-j))$ for $j = 1, \ldots, n-1$, which can be easily verified by direct substitution of the known β-distribution equilibrium for the Wright–Fisher case into (29).

For the Wright–Fisher model, RoyChoudhury and Wakeley (2010) have shown that the number of segregating sites in the sample of size n, $S_n = \sum_{i=1}^{n-1} Z_i$, is a sufficient statistic for Θ, under the independent-sites assumption. This is an important result because the number of segregating sites vastly compresses the information in the frequency spectrum, yet nonetheless contains no loss of information for the purposes of inferring the mutation rate. The Poisson representation of the sample frequency spectrum we have derived shows that S_n remains Poisson distributed even in the general Λ-infinite-sites case—under the assumption of site independence. Thus, the sufficiency of S_n for Θ remains true, and consequently S_n possesses desirable qualities for robust estimation of Θ.

Diversity amplification and the number of segregating sites

The number of segregating sites in a sample is a classic and powerful method to quantify genetic diversity in a population. Here we study how S_n depends on the form of reproduction—that is, on the form of the drift measures Λ. In particular, we show that the Wright–Fisher model minimizes the expected number of segregating sites in a sample, compared to all other Λ-processes. Thus, large family sizes in the offspring distribution will tend to amplify the amount of diversity in a population.

As usual, we consider a Λ-Fleming–Viot process such that Λ is a probability measure. Under the infinite-sites Poisson approximation, the number of segregating sites S_n in a sample of size n is Poisson distributed, and its expected value is

$$\mathbb{E}S_n = \mathbb{E}\sum_{j=1}^{n-1} Z_j = \Theta \sum_{j=1}^{n-1} c_j(n), \tag{30}$$

where $c_j(n)$ are the coefficients in (29). Applying the binomial theorem to the sum above shows that

$$\frac{\mathbb{E}S_n}{\Theta} = \lim_{\theta \downarrow 0} \frac{1}{\theta} \int_0^1 [1 - x^n - (1-x)^n] d\pi_\theta(x) \tag{31}$$

and so we may interpret the expected number of segregating sites as a type of higher-order heterozygosity statistic of the stationary distribution. According to (29), $c_j(n)$ is a linear combination of moments of π, of order at most n. It follows that the average number of segregating sites may be evaluated by the recursion (15) and it can be expressed as rational functions of moments of Λ. The first several such expressions are listed below:

$$\mathbb{E}S_2 = \Theta \tag{32}$$

$$\mathbb{E}S_3 = \frac{3}{2} \cdot \Theta \tag{33}$$

$$\mathbb{E}S_4 = \frac{\int_0^1 \left(5\lambda^2 - 14\lambda + 11\right) d\Lambda(\lambda)}{\int_0^1 \left(6 - 8\lambda + 3\lambda^2\right) d\Lambda(\lambda)} \cdot \Theta \tag{34}$$

$$\mathbb{E}S_5 = \frac{5}{2} \frac{\int_0^1 \left(5 - 6\lambda + 2\lambda^2\right) d\Lambda(\lambda)}{\int_0^1 \left(6 - 8\lambda + 3\lambda^2\right) d\Lambda(\lambda)} \cdot \Theta \tag{35}$$

$$\mathbb{E}S_6 = \frac{1}{2} \frac{\int_0^1 \left(2608\lambda^2 - 1558\lambda + 411 - 2428\lambda^3 + 1312\lambda^4 - 388\lambda^5 + 49\lambda^6\right) d\Lambda(\lambda)}{\int_0^1 (6 - 8\lambda + 3\lambda^2)(15 - 40\lambda + 45\lambda^2 - 24\lambda^3 + 5\lambda^4)\, d\Lambda(\lambda)} \cdot \Theta \tag{36}$$

$$\vdots$$

These expressions for the expected number of segregating sites become extremely complex for larger sample sizes n. Nevertheless, we can use asymptotic methods to study how diversity is expected to behave in large sample sizes. We address two primary questions. First, how does the expected number of segregating sites, $\mathbb{E}S_n$, grow as a function of the sample size, for a given drift-measure Λ? And

second, which reproduction processes Λ maximize and minimize $\mathbb{E}S_n$, for fixed Θ?

In the *Appendix*, we use the moment recursion (15) to derive the following recursion for the sequence $\{\mathbb{E}S_n\}$, $n = 2, 3, \ldots,$

$$\mathbb{E}S_n = \frac{n\Theta}{2a_{nn}} + \sum_{j=1}^{n-1} \frac{a_{jn}}{a_{nn}} \mathbb{E}S_j, \tag{37}$$

with the initial value $\mathbb{E}S_1 = 0$, and where $\{a_{jn}\}$ are given by (16) and (17). We can use this relation to obtain detailed information about $\mathbb{E}S_n$ both as a function of the sample size n and as a function of the underlying Λ-measure.

Consider first the pure Λ-processes $\Lambda = \delta_\lambda$, in which a single individual may replace a given fixed fraction $0 < \lambda \leq 1$ of the population. Then we can prove from (37) that, for every $p > 1$(see *Appendix*),

$$\mathbb{E}S_n = \frac{\lambda\Theta}{2} n + O\left(n^{1/p}\right), \quad n \to \infty. \tag{38}$$

Two features are of interest in this asymptotic expression. First, the average number of segregating sites grows linearly with sample size in a pure Λ-process, as opposed to logarithmically as in the Wright–Fisher case. Second, the rate of linear growth depends on the jump fraction, λ, so that asymptotically, diversity is maximized for large replacement fractions λ and correspondingly minimized when this fraction is small.

Equation 38 can be generalized to a larger class of Λ-measures. If Λ is any probability measure whose support excludes a neighborhood of zero [*i.e.*, there exists an interval $I = [0, \varepsilon]$, for $\varepsilon > 0$, such that $\Lambda(I) = 0$], then we have the asymptotic formula, for every $P > 1$,

$$\mathbb{E}S_n = C(\Lambda)\Theta \cdot n + O\left(n^{1/p}\right), \quad n \to \infty, \tag{39}$$

where $C(\Lambda) = (2\int_0^1 \lambda^{-1}\, d\Lambda)^{-1}$. This equation shows that linear growth of the expected number of segregating sites is characteristic of any Λ-process whose drift measure is bounded away from zero—that is, any Λ-process that does not contain a component of the Wright–Fisher process. This result allows us to determine which reproduction processes Λ maximize and minimize the average number of segregating sites $\mathbb{E}S_n$, for a given value of Θ. In the *Appendix*, we prove the following optimization principle: *for each sample size n, the diversity maximizing and minimizing processes within the class of all Λ processes must in fact be pure Λ processes, i.e., where Λ is concentrated at a single point*. It follows then from (38) that, asymptotically, the Wright–Fisher model ($\lambda = 0$) minimizes and the star-model ($\lambda = 1$) maximizes, respectively, the mean number of segregating sites among all Λ-processes.

Although these results apply in the limit of large sample sizes, we conjecture that the Wright–Fisher and star models are also the extremal diversity processes for any sample size, n. From the optimization principle stated above, it suffices to check this statement within the restricted class of *pure* Λ-processes. In Figure 3, we show $\mathbb{E}S_n/\Theta$ as a function of the jump-size parameter λ for the pure models, for a few values of n. These results confirm that the Wright–Fisher model minimizes $\mathbb{E}S_n$, whereas the star model maximizes $\mathbb{E}S_n$, over all Λ-models. We have conducted numerical studies that support this proposition more generally, even for very small sample sizes. In this sense, the Wright–Fisher model and star models are extremal processes and, for a given effective variance population size, respectively minimize and maximize the expected genetic diversity in any sample.

Discussion

We have studied the stationary distribution of a very general class of population models, under recurrent mutation. We have focused on understanding the interaction between the form of the offspring distribution and the resulting form of genetic drift it engenders, as well as the shape of the stationary distribution. We have demonstrated that the mutation rate can always be uniquely identified from the stationary distribution, even when the drift measure is unknown (as it always will be, in practice). In addition, we have provided a simple example in which the drift measure Λ cannot be uniquely identified from the equilibrium properties of the process.

While inference of the Λ-measure has been studied for particular subclasses of processes (*e.g.*, the pure processes and β-processes) in the literature, the general question of identifiability from equilibrium properties does not appear to have a definitive answer. Our results suggest there are several facets to the problem. The counterexample to uniqueness provided above shows that, even when restricted to the class of pure processes, mutation rates exist at which the diallelic equilibria of different Λ-processes are indistinguishable. However, identifiability is recovered in this class of processes when the per-site mutation rates are small, as shown by our infinite-sites analysis: the asymptotic formula for the mean number of segregating sites indicates that the pure Λ-processes can be uniquely discriminated by the rate of linear growth in diversity as a function of sample size. It remains unclear, however, even in the infinite-sites framework, whether Λ can be identified in general. The fact that $\mathbb{E}S_n$ grows linearly for a large class of processes implies that this statistic may be insufficient to discriminate among processes whose coalescents are highly non-Wright–Fisherian—for example, processes whose coalescents do not come down from infinity. In these cases dynamic data, such as allele-frequency time series, would allow researchers to infer the form of the offspring distribution.

The stationary allele frequency distribution of the Wright–Fisher process is extremal, in a sense, within the class of Λ-processes. Specifically, the Wright–Fisher model exhibits greater probability mass near very high and low allele frequencies. This observation was formalized by analyzing a Λ-infinite-sites model, in which we found that the mean

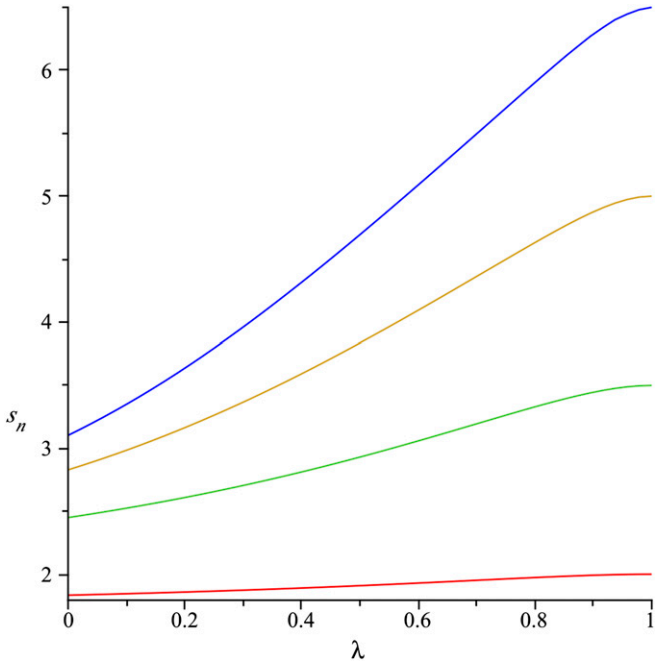

Figure 3 Diversity $s_n = \mathbb{E}S_n/\Theta$ *vs.* jump-size λ in the pure Λ-processes, for $n = 4, 7, 10, 13$ (bottom to top). These functions attain their extrema at the endpoints 0 and 1, implying that the Wright–Fisher and star processes minimize and maximize diversity for these sample sizes.

number of segregating sites in a sample is indeed minimized by the Wright–Fisher process, among all Λ-processes of the same offspring variance σ_N^2 and with the same per-generation mutation rate.

Our results can be placed in the context of a nascent literature that views the Wright–Fisher process as an extremal model within the large space of possible population processes. Aside from the diversity-minimization property we have demonstrated here, it has previously been observed, for instance, that among Λ-processes, the Wright–Fisher model minimizes the fixation probability of an adaptive allele (Der *et al.* 2012), minimizes the time to absorption for new mutants (Der *et al.* 2011), and, among generalized coalescents, possesses the fastest rate of "coming down from infinity" (Berestycki *et al.* 2010). The basic intuition behind all these results revolves around the type of sample paths possessed by different processes. A typical sample path in the Kimura diffusion undergoes a high frequency of small jumps (in fact, is continuous), and thus new mutants persist for only $O(\log N)$ generations before being eliminated by genetic drift. By contrast, in a general Λ-model with the same variance effective population size, large jumps in the sample path may occur, but with lower frequency, thereby lengthening the absorption time—for example, up to order N generations in the pure Λ-processes. Since the mean number of segregating sites in the entire population is the product of the genomic mutation rate and the expected absorption time for a new mutant, standing genetic diversity must increase when reproductive skew is present.

Although we have presented results only within the class of Λ-processes, many of our formulas—for example (15)—can be generalized to the set of all Cannings models. We expect the diversity-minimization property of the Wright–Fisher model will hold even within this larger family.

The infinite-sites model of the Λ-process we have developed here differs from the Watterson infinite-sites model typically encountered in coalescent theory, in two respects. First, we have assumed free recombination and hence independent sites, whereas in Watterson's model sites are tightly linked. Second, we assume two-way mutation between alternative alleles at each site, whereas Watterson's model features one-way mutation at each site away from the existing type. Nonetheless, some of the results derived for our site-independent, infinite-sites model extend to the Watterson, linked infinite-sites Λ-processes as well.

In general, the (random) number of segregating sites S_n in a sample is a function of the dependency structure among sites. For example, in the simple Wright–Fisherian case, independence of sites gives rise to a Poisson distribution for S_n, compared to a sum of geometric random variables in the case of no recombination (Ewens 2004). However, as Watterson (1975) has already remarked, the *mean* value of S_n is generally robust to the recombination structure of an infinite-sites model. If $Y_1, \ldots, Y_L$ denote the allelic distributions at L sites, then (28) shows that the expected number of segregating sites is a function only of the marginal distributions of Y_i, instead of their joint distribution. Thus the expected number of segregating sites in a sample is unaffected by linkage. Likewise, the distinction between one-way and two-way mutation (and folded and unfolded spectra) does not alter the mean number of segregating sites other than by a possible overall scaling.

Because S_n is such a common measure of genetic diversity, our results have some connections to the literature on Λ-coalescents. Recently, Berestycki *et al.* (2012) (see their Theorem 3) showed that, for those Λ-measures whose coalescent comes down from infinity, the (random) number of segregating sites S_n in a sample of size n for the Watterson model has the asymptotic law

$$\frac{S_n}{\int_0^n q\psi^{-1}(q)dq} \to \Theta, \tag{40}$$

where ψ is the Laplace exponent of the Λ-measure, defined as

$$\psi(q) = \int_0^1 \frac{\exp(-q\lambda) - 1 + q\lambda}{\lambda^2} d\Lambda. \tag{41}$$

The authors conjectured that (40) holds more generally, even when Λ does not come down from infinity. In this respect, our asymptotic result (39) for $\mathbb{E}S_n$—derived for Λ-measures bounded away from zero (and thus always failing to come down from infinity)—is evidence in favor of their more general conjecture, in the case not covered by

the hypotheses of their theorem. For under such assumptions, the Laplace exponent has the expression

$$\psi(q) \sim q \int_0^1 \lambda^{-1}\, d\Lambda, \tag{42}$$

which implies from (40) that

$$\int_0^n q\psi^{-1}(q)\, dq \sim n \cdot \left(\int_0^1 \lambda^{-1} d\Lambda \right)^{-1}, \tag{43}$$

which is proportional to our Equation 39 for $\mathbb{E}S_n$. Finally, returning to the case of independent sites, developed in this article, it is also true that the distributional convergence of (40) holds, a fact that follows from the Poisson representation for S_n.

In our analysis of the expected number of segregating sites, we have concentrated on the two extreme cases—the Wright–Fisher case, for which $\mathbb{E}S_n$ is known to grow logarithmically in the sample size n, and the case of pure Λ-processes (and more generally those Λ-processes whose drift measure support excludes zero), for which we have demonstrated linear growth of $\mathbb{E}S_n$. Nevertheless, the recurrence relation (37) can be used to analyze intermediary cases as well, for example the β-processes, in which the density of Λ behaves like a power law in the vicinity of zero. Based on our stated optimization principle, we conjecture that for such reproduction measures, growth in diversity with sample size will lie somewhere between the logarithmic and the linear cases.

Acknowledgments

The authors thank Warren Ewens and Charles Epstein for many fruitful discussions. J.B.P. acknowledges funding from the Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the Foundational Questions in Evolutionary Biology Fund (RFP-12-16), and the U.S. Army Research Office (W911NF-12-1-0552).

Literature Cited

Beckenbach, A., 1994 Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models, pp. 188-198 in *Non-neutral Evolution: Theories and Molecular Data*, Springer, USA.

Berestycki, J., J. Berestycki, and J. Schweinsberg, 2007 Beta-coalescents and continuous stable random trees. Ann. Probab. 35: 1835–1887.

Berestycki, J., N. Berestycki, and V. Limic, 2010 The Λ-coalescent speed of coming down from infinity. Ann. Probab. 38: 207–233.

Berestycki, J., N. Berestycki, and V. Limic, 2012 Asymptotic sampling formulae for Λ-coalescents. arXiv: 1201.6512.

Bertoin, J., and J. F. Le Gall, 2006 Stochastic flows associated to coalescent processes. Illinois J. Math. 50: 147–181.

Birkner, M., and J. Blath, 2008 Computing likelihoods for coalescents with many collisions in the infinitely many sites model. J. Math. Biol. 57: 435–465.

Birkner, M., and J. Blath, 2009 Measure-valued diffusions, general coalescents and population genetic inference. Trends Stoch. Anal. 351: 329–363.

Birkner, M., J. Blath, and M. Steinrucken, 2011 Importance sampling for Lambda-coalescents in the infinitely many sites model. Theor. Popul. Biol. 79: 155–173.

Bustamante, C. D., J. Wakeley, S. Sawyer, and D. L. Hartl, 2001 Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv. Appl. Prob. 6(2): 260–290.

Der, R., 2010 A theory of generalised population processes. Ph.D. Thesis, University of Pennsylvania, Philadelphia.

Der, R., C. L. Epstein, and J. B. P. Plotkin, 2011 Generalized population processes and the nature of genetic drift. Theor. Popul. Biol. 80: 80–99.

Der, R., C. Epstein, and J. Plotkin, 2012 Dynamics of neutral and selected alleles when the offspring distribution is skewed. Genetics 191: 1331–1344.

Desai, M. M., and J. B. Plotkin, 2008 The polymorphism frequency spectrum of finitely many sites under selection. Genetics 180(4): 2175–2191.

Donnelly, P., and T. Kurtz, 1999 Particle representations for measure-valued population models. Ann. Probab. 27: 166–205.

Eldon, B., and J. Wakeley, 2006a Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172: 2621–2633.

Eldon, B., and J. Wakeley, 2006b Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178: 1517–1532.

Ethier, S. N., and T. G. Kurtz, 1986 *Markov Processes: Characterization and Convergence*. Wiley Interscience, Hoboken, NJ.

Ewens, W. J., 2004 *Mathematical Population Genetics*, Ed. 2. Springer-Verlag, New York.

Fisher, R. A., 1958 *The Genetical Theory of Natural Selection*. Dover, New York.

Haldane, J., 1932 A mathematical theory of natural and artificial selection, part IX, rapid selection. Proc. Camb. Philos. Soc. 28: 244–248.

Hedgecock, D., 1994 Does variance in reproductive success limit effective population size of marine organisms? pp. 122-134 in *Genetics and Evolution of Aquatic Organisms*, Springer, USA.

Karlin, S., and J. McGregor, 1964 Direct product branching processes and related Markov chains. Proc. Natl. Acad. Sci. USA 51: 598–602.

Kimura, M., 1955 Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41: 144–150.

Kimura, M., 1994 *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers*. University of Chicago Press, Chicago.

McDonald, D., 1980 On the Poisson approximation to the multinomial distribution. Can. J. Stat. 8: 115–118.

Möhle, M., 2001 Forward and backward diffusion approximations for haploid exchangeable population models. Stoch. Proc. Appl. 95: 133–149.

Möhle, M., 2006 On sampling distributions for coalescent processes with simultaneous multiple collisions. Bernoulli 1: 35–53.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.

RoyChoudhury, A., and J. Wakeley, 2010 Sufficiency of the number of segregating sites in the limit under finite-sites mutation. Theor. Popul. Biol. 78: 118–122.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Communicating editor: J. Hermisson

Appendix

The Stationary Distribution of Processes with Reproductive Skew

Let $X^{(N)}$ be a sequence of discrete Eldon–Wakeley processes, one for each population size N, converging to a continuum Λ-process $\tilde{X}$, under the state and time renormalization of (4) and (5). If we suppose that each discrete process operates under strictly positive mutation rates $\mu_i^{(N)}$, then it is easily verified that the associated forward-time transition matrices $\mathbf{P}^{(N)}$ for the discrete processes possess strictly positive entries, and thus, from the Perron–Frobenius theorem, a unique stationary distribution π_N exists for each process $X^{(N)}$. A standard argument, using the fact that the sequence π_N is tight, shows that there is a subsequence π_{N_k} converging to a probability measure π that is a stationary distribution for $\tilde{X}$ (see Ethier and Kurtz 1986, for example). This argument indeed demonstrates that any weak limit point of π_N is a stationary distribution $\tilde{X}$; below, through the moment recursion, this distribution is uniquely characterized, and hence every weakly convergent subsequence of π_N converges to π; thus $\pi_N \to \pi$.

Derivation of a Recursion for the Moments of the Stationary Distribution

Let π be the stationary distribution for the two-type Λ-process, which satisfies (12). In this section we obtain a recursion formula for the moments of π.

Define the operator $Lu(x) = xu(x + (1 - x)\lambda) - u(x) + (1 - x)u(x - \lambda x)$. Setting $u(x) = x^k$, $k \geq 0$, we have

$$\begin{aligned} Lx^k &= \Big((1-\lambda)^k - 1\Big)x^k - (1-\lambda)^k x^{k+1} + x(x + (1-x)\lambda)^k = \Big((1-\lambda)^k - 1\Big)x^k - (1-\lambda)^k x^{k+1} + x^{k+1}(1-\lambda)^k \\ &\quad + kx^k(1-\lambda)^{k-1}\lambda + \sum_{j=0}^{k-2}\binom{k}{j}x^{j+1}(1-\lambda)^j\lambda^{k-j} = \Big((1-\lambda)^k - 1 + k\lambda(1-\lambda)^{k-1}\Big)x^k + \sum_{j=1}^{k-1}\binom{k}{j-1}x^j(1-\lambda)^{j-1}\lambda^{k-j+1}. \end{aligned} \tag{A1}$$

Separating the Λ-generator (10) into the mutation and pure-drift portions, we define the latter to be the operator

$$G_D u(x) = \int_0^1 \frac{1}{\lambda^2} Lu(x) d\Lambda(\lambda). \tag{A2}$$

If we write

$$G_D x^k = \sum_{j=1}^{k} b_{jk} x^j, \tag{A3}$$

then substituting (A1) into (A2) and comparing the coefficients to (A3), we have

$$b_{kk} = \int_0^1 \frac{(1-\lambda)^k + k\lambda(1-\lambda)^{k-1} - 1}{\lambda^2}\, d\Lambda(\lambda) \tag{A4}$$

$$b_{jk} = \binom{k}{j-1} \int_0^1 \lambda^{k-j-1}(1-\lambda)^{j-1}\, d\Lambda(\lambda), \quad j < k. \tag{A5}$$

Let π be any stationary distribution of the process. Then according to (12),

$$\int_0^1 \left[\frac{1}{2}(-\theta_1 x + \theta_2(1-x))kx^{k-1} + G_D x^k\right] d\pi(x) = 0. \tag{A6}$$

Using the expansion for $G_D x^k$ above, we derive

$$\left(b_{kk} - \frac{k}{2}(\theta_1 + \theta_2)\right)m_k + \left(\frac{k}{2}\theta_2 + b_{k-1,k}\right)m_{k-1} + \sum_{j=1}^{k-2} b_{jk} m_j = 0, \tag{A7}$$

where m_j is the jth moment of π. This is equivalent to formulas (15)–(17), where $b_{kk} = -a_{kk}$, and $b_{jk} = a_{jk}$ for $j < k$.

Derivation of the star-process stationary distribution

Consider the probability measure μ on [0, 1], with density

$$\frac{d\mu}{dx} = \frac{1}{\theta}|1-2x|^{(1-\theta)/\theta}. \tag{A8}$$

The generator for the star process undergoing symmetric mutation is $Gu = (1/\ 2)\theta(1-2x)u'(x) + (1-x)u(0) - u(x) + xu(1)$, and the space of twice continuously differentiable functions $C^2[0, 1]$ is a core for G. Noting that the density $d\mu/dx$ satisfies the equation $-d/dx((\theta/\ 2)(1-2x)d\mu/dx) - d\mu/dx = 0$ and integrating by parts, it is readily verified that $\int_0^1 Gu\ d\mu = 0$ for every $u \in C^2$. Thus μ is a stationary distribution for the process and is further unique as established by the moment recursion (A7).

Properties of the θ-estimation from the equilibrium distribution

We establish here Theorems 1 and 2 for the estimator $\hat{\theta}$. From the strong law of large numbers, $(1/\ T)\sum_{i=1}^{T}(X_i-(1/2))^2$ converges almost surely to v, the variance of the equilibrium. Therefore $\hat{\theta} \to g(v) = \theta$, which proves Theorem 1.

Next we study the asymptotic variance of the estimator $\hat{\theta}$. The asymptotic mean from the consistency result above, is, of course, θ. Since $(1/\ T)\sum_{i=1}^{T}(X_i-1/2)^2$ concentrates around $v(\theta)$, we consider a Taylor expansion around that point:

$$g(v(\theta)+x) = g(v) + g'(v)x + O\left(x^2\right) \tag{A9}$$

$$= \theta + \frac{1}{8}\log v(\theta)\cdot x + O\left(x^2\right). \tag{A10}$$

Let $Y = (1/\ T)\sum_{i=1}^{T}(X_i-1/2)^2$. Then setting $x = Y - v(\theta)$,

$$g(Y) = g(v(\theta)) + \frac{1}{8}\log v(\theta)\cdot(Y - v(\theta)) + O(Y-v(\theta))^2 \tag{A11}$$

$$= \theta + \frac{1}{8}\log v(\theta)\cdot(Y - v(\theta)) + O(Y-v(\theta))^2. \tag{A12}$$

It follows that

$$\mathrm{Var}\left(\hat{\theta}\right) = E(g(Y)-\theta)^2 = \frac{1}{64}(\log v(\theta))^2 E(Y-v(\theta))^2 + O\left(E(Y-v(\theta))^3\right) \tag{A13}$$

$$= \frac{1}{64}(\log v(\theta))^2 \cdot \mathrm{Var}(Y) + O\left(E(Y-v(\theta))^3\right). \tag{A14}$$

Now,

$$\mathrm{Var}(Y) = \frac{1}{T}\,\mathrm{Var}\left(X_i-\frac{1}{2}\right)^2 \tag{A15}$$

$$= \frac{1}{T}\left(E\left(X_i-\frac{1}{2}\right)^4 - \left(E\left(X_i-\frac{1}{2}\right)^2\right)^2\right). \tag{A16}$$

Thus in summary, for large T,

$$\mathrm{Var}\left(\hat{\theta}\right) \sim \frac{C}{T}, \tag{A17}$$

where

$$C = \frac{1}{64}(\log v)^2\left(\mu_4 - \mu_2^2\right) \tag{A18}$$

$$v = \frac{1}{4(1+2\theta)}, \tag{A19}$$

and μ_i is the ith central moment of the equilibrium distribution.

Using the Taylor estimate (A12) and combining with the central limit theorem, this establishes Theorem 2.

Poisson representation of the infinite-sites model

In this section we show that the segregating site variables $(Z_1, \ldots, Z_{n-1})$ in the independent-sites Λ-model converge to a sequence of Poisson random variables with means $(c_1\Theta, \ldots, c_{n-1}\Theta)$, given by (29). We make use of the structure of the moments of the stationary distribution as found in (15). First, we require a preliminary lemma.

Lemma 1. *Let π_θ be the stationary distribution of a Λ process undergoing symmetric mutation θ. Then constants $c_j \geq 0$ exist, for $1 \leq j \leq n-1$,*

$$p_j(\theta) \equiv \int_0^1 \binom{n}{j} x^j (1-x)^{n-j}\, d\pi_\theta(x) = c_j\theta + o(\theta), \quad \theta \downarrow 0. \tag{A20}$$

Proof. Lemma 1 is equivalent to saying $p_j(0) = 0$ and that the derivative of p_j exists at $\theta = 0$, taken as a limit from positive values. First, observe from (15) that under $\theta_1 = \theta_2$, all the moments $m_k(\theta)$ of the stationary distribution are differentiable in θ for all $\theta \geq 0$, and hence $p_j(\theta)$ is differentiable everywhere. Also

$$\left|p_j(\theta)\right| \leq \binom{n}{j} \int_0^1 x(1-x)\, d\pi_\theta(x) = \binom{n}{j} \frac{2\theta}{1+2\theta}. \tag{A21}$$

Hence $p_j(\theta) \to 0$ as $\theta \downarrow 0$.

Now to establish the Poisson representation, it is enough to apply the well-known Poisson approximation to the multinomial distribution. We use the following:

Theorem 3 (McDonald 1980). *If $(Z_0, \ldots, Z_n)$ is multinomial with parameters $(L, p_0, \ldots, p_n)$, and $(V_1, \ldots, V_{n-1})$ are independent Poissons with means Lp_j, then*

$$||(Z_1, \ldots, Z_{n-1}) - (V_1, \ldots, V_{n-1})|| \leq 2L \left(\sum_{j=1}^{n-1} p_j \right)^2, \tag{A22}$$

where $||\cdot||$ is the total variation norm of measures.

The Poisson representation is now obvious, since $p_j(\theta) = O(\theta) = O(1/L)$ by Lemma 1 and therefore the right-hand side of (A22) goes to zero as $L \to \infty$. Since $(V_1, \ldots, V_{n-1})$ are converging to a sequence of independent Poisson distributions with finite means $c_j\Theta$, where c_j are as in Lemma 1, so must $(Z_1, \ldots, Z_{n-1})$.

The average number of segregating sites

In this section, we study the number of segregating sites S_n in the infinite-sites Λ-model, deriving the recursion (37) for the average diversity measure $s_n = \mathbb{E}S_n/\theta$, and use it to obtain asymptotic expressions for diversity.

Let π_θ be the stationary distribution of the Λ-process under two-way symmetric mutation θ. Define $H_n(\theta)$ as the heterozygosity measure

$$H_n(\theta) = \sum_{j=1}^{n-1} \int_0^1 \binom{n}{j} x^j (1-x)^{n-j}\, d\pi_\theta(x). \tag{A23}$$

Applying the binomial theorem,

$$H_n(\theta) = \int_0^1 \left(1 - x^n - (1-x)^n\right) d\pi_\theta(x) = \int_0^1 (1 - 2x^n) d\pi_\theta(x), \tag{A24}$$

where the second equality follows from symmetry of the stationary distribution. Now, define the diversity measure $s_n = \mathbb{E}S_n/\Theta$. We have from (31)

$$s_n = \lim_{\theta \downarrow 0} \frac{1}{\theta} H_n(\theta). \tag{A25}$$

Under symmetric mutation $\theta = \theta_1 = \theta_2$, the recursion formulas for moments (15) read

$$m_n(\theta) = \frac{(n\theta/2)\cdot m_{n-1}(\theta) + \sum_{j=1}^{n-1} a_{jn} m_j(\theta)}{n\theta + a_{nn}}. \tag{A26}$$

Observe that by the binomial theorem, one has the relation

$$a_{nn} = \sum_{j=1}^{n-1} a_{jn}. \tag{A27}$$

Taking the limit as $\theta \downarrow 0$ in (A26), and using (A27) and $m_1(\theta) = 1/2$, it is easy to show that $\lim_{\theta\to 0} m_n(\theta) = 1/2$ for every n. Since (A26) also demonstrates that m_n is analytic in θ in a neighborhood of 0, there are numbers $\{v_n\}$ such that $m_n(\theta) = 1/2 + v_n\theta + O(\theta^2)$. Inserting this into the right-hand side of (A26), and expanding the quotient (A26) in a Taylor series, we obtain, by equating the first-order coefficients, a recursion for v_n:

$$v_n = \frac{-n/4 + \sum_{j=1}^{n-1} a_{jn} v_j}{a_{nn}}. \tag{A28}$$

The equations (A24) and (A25) imply that

$$\lim_{\theta\downarrow 0} \frac{1 - 2m_n(\theta)}{\theta} = s_n. \tag{A29}$$

And using the Taylor expansion for $m_n(\theta)$, we find that $s_n = -2v_n$. Thus the corresponding recursion for $\{s_n\}$ is

$$s_n = \frac{n/2 + \sum_{j=1}^{n-1} a_{jn} s_j}{a_{nn}}, \tag{A30}$$

where we initialize $s_1 = 0$.

From the relation (A27), the numbers a_{jn}/a_{nn} define a probability measure on the set $j \in \{1, \ldots, n-1\}$. By studying this measure and the recurrence relation defining s_n, we may derive the asymptotics for s_n.

Now suppose that the underlying Λ-process is associated with a Λ-measure with support bounded away from zero. Then from (16), $a_{nn} = \int_0^1 \lambda^{-2} d\Lambda + O(\gamma^n)$, for some $0 < \gamma < 1$. Therefore, $a_{nn}^{-1} = \left(\int_0^1 \lambda^{-2} d\Lambda\right)^{-1} + O(\gamma^n)$. Using this estimate in the recursion (A30), and defining $A = \left(\int_0^1 \lambda^{-2} d\Lambda\right)^{-1}$,

$$s_n = \frac{An}{2} + A\sum_{j=1}^{n-1} a_{jn} s_j + O(\gamma^n). \tag{A31}$$

Now, we prove the following:

Theorem 4. *For the Λ-Fleming–Viot infinite-sites model under symmetric mutation and where Λ is a measure whose support is bounded away from zero, the mutation-normalized mean number of segregating sites s_n satisfies, for every $p > 1$,*

$$s_n = Cn + O\left(n^{1/p}\right), \quad n \to \infty, \tag{A32}$$

where

$$C = \frac{A}{2 - 2A\int_0^1 \lambda^{-2}(1-\lambda)\, d\Lambda} = \frac{1}{2\int_0^1 \lambda^{-1} d\Lambda}. \tag{A33}$$

Proof. Let $s_n = C(n-1) + g_n$; we derive a recurrence relation for the residual error g_n. From the recursion for s_n, we find, using the explicit expressions (17),

$$s_n = \frac{An}{2} + A\sum_{j=1}^{n-1} \int_0^1 \lambda^{-2} \binom{n}{j-1} \lambda^{n-(j-1)} (1-\lambda)^{j-1} d\Lambda \cdot \Big(C(j-1) + g_j\Big) + O(\gamma^n) \tag{A34}$$

for some $0 < \gamma < 1$.

Denote the right-hand side of the above by RHS, and to ease the notation define the coefficients

$$b_{n,j}(\lambda) = \binom{n}{j-1}\lambda^{n-(j-1)}(1-\lambda)^{j-1}. \tag{A35}$$

Then

$$\begin{aligned}\text{RHS} =& \frac{An}{2} + A\sum_{j=1}^{n+1}\int_0^1 C(j-1)\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda \\ & - A\sum_{j=n}^{n+1}\int_0^1 C(j-1)\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda \\ & + A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)d\Lambda + O(\gamma^n)\end{aligned} \tag{A36}$$

$$= \frac{An}{2} + A\sum_{j=1}^{n+1}\int_0^1 C(j-1)\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda + A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)d\Lambda + O(\gamma_1^n) \tag{A37}$$

for some $0 < \gamma_1 < 1$, with the last equality holding since Λ has support bounded away from zero. Using the formula for the mean of a binomial random variable with parameters $(n, 1-\lambda)$ in the series, this is

$$\text{RHS} = \frac{An}{2} + An\int_0^1 C(1-\lambda)\lambda^{-2}d\Lambda + A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)d\Lambda + O(\gamma_1^n). \tag{A38}$$

Inserting this expression into (A34) and using the definition of C, we find that g_n satisfies the relation

$$g_n = A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda + O(1). \tag{A39}$$

We now use this relation to show that, for every $p > 1$, $g_n = O(n^{1/p})$. To do this we prove that any g_n satisfying (A39) must satisfy the upper bound $g_n \le Mn^{1/p}$ for some M; then the statement follows because $-g_n$ also satisfies (A39) and hence $-g_n \le M'n^{1/p}$.

Fix a $p > 1$. From (A39), there exists a B such that

$$\left| g_n - A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda \right| \le B \tag{A40}$$

for all $n \ge 1$. We demonstrate that $g_n \le Bn^{1/p}$ by induction. First, a base case for g_k can be established on any fixed initial set $k \in \{1, \ldots, k_0\}$ by simply enlarging B sufficiently. Now assume the statement is true up to index $n-1$, for $n \ge k_0 + 1$. Then we have

$$g_n \le A\sum_{j=1}^{n-1}\int_0^1 g_j\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda + B \tag{A41}$$

$$\le A\sum_{j=1}^{n-1}\int_0^1 Bj^{1/p}\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda + B \tag{A42}$$

$$\le A\sum_{j=1}^{n+1}\int_0^1 Bj^{1/p}\lambda^{-2}b_{n,j}(\lambda)\,d\Lambda + B. \tag{A43}$$

There is a probabilistic interpretation to this series. If we define random variables $X_\lambda = 1 + Y_\lambda$, where Y_λ is binomially distributed with parameters $(n, (1 - \lambda))$, then $b_{n,j}(\lambda) = \mathbb{P}(X_\lambda = j)$. Define further a random variable X that is a mixture of X_λ under the mixing probability density $dQ(\lambda) = A\lambda^{-2}d\Lambda(\lambda)$. Then (A43) may be written as an expectation over X, as

$$g_n \leq B \cdot \mathbb{E}\left[X^{1/p}\right] + B. \tag{A44}$$

Applying Jensen's inequality,

$$g_n \leq B(\mathbb{E}X)^{1/p} + B \tag{A45}$$

$$= B\left(1 + A\int_0^1 \lambda^{-2}(n(1-\lambda))d\Lambda\right)^{1/p} + B. \tag{A46}$$

Let $\varepsilon > 0$ be such that $\Lambda([0, \varepsilon]) = 0$. Then

$$g_n \leq B\left(1 + A\int_0^1 \lambda^{-2}(n(1-\varepsilon))\, d\Lambda\right)^{1/p} + B \tag{A47}$$

$$= B\left((1 + (1-\varepsilon)n)^{1/p} + 1\right). \tag{A48}$$

Therefore $g_n \leq Bn^{1/p}$ if the base case is established for k_0 sufficiently large, and Theorem 4 is proved.

An optimization principle for the average number of segregating sites

In this section we prove the following theorem:

Theorem 5. *For every sample size n, and for fixed mutation rate Θ, the minimum and maximum values of $\mathbb{E}S_n$ over the class of Λ-processes are achieved within the class of pure Λ-Fleming–Viot processes, that is, where $\Lambda = \delta_\lambda$, for $0 \leq \lambda \leq 1$.*

Proof. It is evident from the recursion for $\mathbb{E}S_n$ in (A30) that the average number of segregating sites must have the form

$$\frac{\mathbb{E}S_n}{\Theta} = \frac{\int_0^1 f_n(\lambda)\, d\Lambda}{\int_0^1 g_n(\lambda)\, d\Lambda} \tag{A49}$$

for some continuous functions (in fact, polynomials) f_n, g_n. Because $\mathbb{E}S_n$ is positive for every pure Λ-process (where $\Lambda = \delta_\lambda$), we can without loss of generality assume that f_n, $g_n \geq 0$. For such functions, we have the following lemma:

Lemma 2. *For positive continuous functions f, g defined on* [0, 1], *we have the inequalities*

$$\min_\lambda \frac{f(\lambda)}{g(\lambda)} \leq \frac{\int_0^1 f(\lambda)\, d\Lambda}{\int_0^1 g(\lambda)\, d\Lambda} \leq \max_\lambda \frac{f(\lambda)}{g(\lambda)}. \tag{A50}$$

Proof. We prove the lower bound; the upper bound is established in the same way. Say Λ is a two-point measure: $\Lambda = p_1\delta_{\lambda_1} + p_2\delta_{\lambda_2}$. Then elementary manipulations show

$$\frac{p_1 f(\lambda_1) + p_2 f(\lambda_2)}{p_1 g(\lambda_1) + p_2 g(\lambda_2)} \geq \min\left\{\frac{f(\lambda_1)}{g(\lambda_1)}, \frac{f(\lambda_2)}{g(\lambda_2)}\right\}. \tag{A51}$$

By induction one easily generalizes to measures concentrated at any finite number of points. Finally, the full case is obtained by taking weak limits of measures concentrated at a finite number of points.

Returning to the proof of the main theorem, Theorem 5, one sees that Lemma 2 immediately implies the result, since $f_n(\lambda)/g_n(\lambda)$ is precisely $\mathbb{E}S_n/\Theta$ for the case $\Lambda = \delta_\lambda$.

The optimization principle can be refined. Let Λ be any probability measure whose support excludes a neighborhood of zero, and suppose that $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest values, respectively, on which Λ is supported; *i.e.*, $\lambda_{\min} = \inf$ supp Λ, and $\lambda_{\max} = \sup$ supp Λ. Then

$$\frac{\lambda_{\min}}{2} \leq C(\Lambda) \leq \frac{\lambda_{\max}}{2}. \tag{A52}$$

In conjunction with (39), this shows that the asymptotic growth rate in $\mathbb{E}S_n$ for a Λ-process whose drift measure is supported on the interval $[\lambda_{\min}, \lambda_{\max}]$ can be lower and upper bounded by the rates of growth in $\mathbb{E}S_n$ of two pure processes, with parameters $\lambda_{\min}$ and $\lambda_{\max}$, respectively. In other words, the effect of mixing any two pure Λ-processes always results in a process whose equilibrium diversity is intermediate relative to the diversities of the pure models.