Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2013, Article ID 963831, 10 pages
http://dx.doi.org/10.1155/2013/963831

Research Article

Modeling Neutral Evolution Using an Infinite-Allele Markov Branching Process

Xiaowei Wu[1] and Marek Kimmel[2,3]

[1] *Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*
[2] *Departments of Statistics and Bioengineering, Rice University, Houston, TX 77005, USA*
[3] *Systems Engineering Group, Silesian University of Technology, 44-100 Gliwice, Poland*

Correspondence should be addressed to Xiaowei Wu; xwwu@vt.edu

Received 6 November 2012; Revised 14 February 2013; Accepted 18 February 2013

Academic Editor: Peter Olofsson

We consider an infinite-allele Markov branching process (IAMBP). Our main focus is the frequency spectrum of this process, that is, the proportion of alleles having a given number of copies at a specified time point. We derive the variance of the frequency spectrum, which is useful for interval estimation and hypothesis testing for process parameters. In addition, for a class of special IAMBP with birth and death offspring distribution, we show that the mean of its limiting frequency spectrum has an explicit form in terms of the hypergeometric function. We also derive an asymptotic expression for convergence rate to the limit. Simulations are used to illustrate the results for the birth and death process.

1. Introduction

The infinite-allele branching process was first introduced by Griffiths and Pakes [1]. As a special type of branching process, this process allows individuals to mutate into infinitely many allelic variants, each of which is "new" in the sense of being different from all previously existing variants. This idealization is approximately correct for rare point mutations in long DNA sequences. Fundamental results for the discrete-time case (simple branching process) and for the continuous-time case (Markov branching process) have been obtained in [1, 2]. These include the number of alleles at a given generation or time, the generation number or time of the last mutation, and the limiting frequency spectrum. There exists an analogy between the results for the discrete-time and the continuous-time cases; however, the characteristics in the continuous case are relatively easier to derive [2]. Many evolutionary processes may be considered time continuous, and frequently we assume Markov property in modeling. A classical example is the discrete-time Wright-Fisher model, which is typically either approximated by a continuous-time diffusion or replaced by a continuous-time Markov chain, the so-called continuous-time Moran process [3]. Therefore, the time-continuous infinite-allele Markov branching process (TCIAMBP, or simply IAMBP) seems to be appropriate for modeling evolution in population genetics.

Consider a Markov branching process with neutral mutations. Suppose that the process starts from a group of individuals carrying the same allele, and individuals can mutate into new allelic variants. We assume that the mutation is independent of the previous history of the process, and the offspring distribution is independent of the allelic type, that is, the selection is neutral for all alleles. The process can be described as an "infinitely-many-alleles" model (IAM). Whenever a mutation happens, it yields a new allele, which differs from all the previously existing ones. In this paper, we are interested in the frequency spectrum of the IAMBP, which may be defined as the number or proportion of alleles present in a given number of individuals at a specified time point. Frequency spectrum in this paper refers to random allele frequencies, not their expected value as in Griffiths and Pakes [1] and in Pakes [2], since we also consider the variance of the allele frequency later on. Unless specified otherwise, we will use terms "mean frequency spectrum" and "variance frequency spectrum" in the remainder of this paper to denote the expected value and variance of the allele frequency. The frequency spectrum plays an important role in many genetic processes, such as DNA sequence evolution. As an example, Kimmel and

Mathaes [4] modeled the Alu sequence data using an infinite-allele simple branching process with linear-fractional offspring distribution, and the goodness of fit testing suggested that Alu sequences do not evolve neutrally and might be under selection. It has to be noted that the concept of the frequency spectrum is in some sense similar to the Ewens' sampling formula [5] in population genetics. We will return to this subject in the discussion, although analysis of the analogies and differences transcends the scope of the present paper.

The paper is organized as follows. In Section 2, we rigorously define the IAMBP and the mean frequency spectrum of the IAMBP. Then, we provide explicit expressions for the special case of the birth and death process. In Section 3, we derive the variance frequency spectrum and discuss its use in interval estimation for process parameters. We perform simulations to illustrate the results using the birth and death process example in Section 4. Section 5 is a summary.

2. IAMBP and Its Limiting Mean Frequency Spectrum

2.1. Definition and Basic Properties in the Supercritical Case. Let us consider a continuous-time Markov branching process consisting of individuals with exponential life spans with mean a^{-1}. Let us assume that upon death, each individual produces a random number of offspring. As usually assumed, the offspring counts are identically distributed according to probability generating function (pgf) $f(s)$, and they are independent conditional on the past process. The mean $f'(1^-)$ of the offspring distribution is m, regardless of the allelic type. We further assume that a newborn individual mutates into a new allelic type with probability μ independently of the previous history of the process. Let us denote by $h(s) = f(\mu+(1-\mu)s)$ the offspring pgf in a clone, started by the overall ancestor or any of mutants, containing only the like-type individuals. The entire process is a union over all individual types of such clones. The theory of the IAMBP has been developed by Griffiths and Pakes [1] in the discrete-time case and then by Pakes [2] in the continuous-time case. We will assume $m > 1$ and $M = h'(1^-) > 1$, although some results can be proved without this latter assumption.

Let $\alpha_t(j)$ be the number of alleles present in j individuals at time t and $\phi_{i,t}(j) = E_i[\alpha_t(j)]$, where subscript i indicates that the process begins with i individuals carrying the same allele. It has been shown that [2]

$$\phi_{i,t}(j) = q_{ij}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx, \quad j \geq 0, \tag{1}$$

where $\lambda = a(m-1)$ is the Malthusian parameter of the overall process and $q_{ij}(t)$ is the probability of observing j individuals ($j \geq 1$) carrying the parental allele at time t when starting from i individuals with the parental allele at time $t = 0$. Consequently, for the number K_t of alleles at time t, we have

$$E_i[K_t] = \sum_{j=1}^{\infty}\phi_{i,t}(j) = 1 - q_{i0}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}\left[1 - q_{10}(x)\right]dx. \tag{2}$$

Let $G_j = \int_0^\infty e^{-\lambda t} q_{1j}(t)dt$, $j \geq 0$. If we define $\psi_{ij}(t) = \phi_{i,t}(j)/E_i[K_t]$ and

$$\psi_j = \lim_{t\to\infty}\psi_{ij}(t) = \lim_{t\to\infty}\frac{q_{ij}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx}{1 - q_{i0}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}\left[1 - q_{10}(x)\right]dx} \tag{3}$$

as the limiting mean frequency spectrum, that is, the expected proportion of alleles present in j individuals as $t \to \infty$, then we see that for the supercritical process such that $\lambda > 0$,

$$\psi_j = \frac{\lambda G_j}{1 - \lambda G_0}, \quad j \geq 1. \tag{4}$$

If $M > 1$, then the process of the like-type clones is supercritical, and as it is known [6], $q_{10}(t) \uparrow q_{10}(\infty) < 1$ and $q_{1j}(t) \to 0$, $j \geq 1$, as $t \to \infty$. Therefore, $e^{\lambda t}\left|\int_0^t e^{-\lambda x} q_{10}(x)dx - q_{10}(\infty)/\lambda\right| \to 0$ and $e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)dx \to 0$ as $t \to \infty$, for $j \geq 1$. This yields the following asymptotic equivalence:

$$\psi_{ij}(t) - \psi_j \underset{t\to\infty}{\sim} \frac{\lambda G_j\left[1 - q_{10}(\infty) - \left(\lambda\left(1 - q_{i0}(\infty)\right)/iam\mu\right)\right]}{\left(1 - \lambda G_0\right)^2} e^{-\lambda t}. \tag{5}$$

Details of the proof are omitted, since they appear elementary.

2.2. IAMBP with Birth and Death Offspring Distribution. For the IAMBP with birth and death offspring distribution $f(s) = \alpha + \beta s^2$, $\alpha + \beta = 1$, we are able to obtain an explicit form for G_j, $j \geq 0$; therefore, the limiting mean frequency spectrum ψ_j, $j \geq 1$ can be derived. The offspring pgf of the like-type individuals clone in the birth and death IAMBP is written as

$$h(s) = f\left(\mu + (1-\mu)s\right) = \alpha + \beta\left[\mu + (1-\mu)s\right]^2, \tag{6}$$

where α, β and μ stand for the death, birth, and mutation probabilities for every individual and $\alpha + \beta = 1$. Note that under another parameterization where the two newborn individuals die, live, and mutate independently, this pgf may be formulated differently as $h(s) = [\alpha + \beta\mu + \beta(1-\mu)s]^2$. Under either parameterization, $\lambda = a(2\beta - 1)$. If, as assumed, $M = m(1-\mu) > 1$, then parameters α and μ are subject to a constraint

$$(1-\alpha)(1-\mu) > \frac{1}{2}. \tag{7}$$

Let us write $A^2 = \alpha + \beta\mu^2$ and $B^2 = \beta(1-\mu)^2$ (note, for the other formulation, $A^2 = (\alpha+\beta\mu)^2$ and $B^2 = \beta^2(1-\mu)^2$). The explicit form of G_j can be written as

$$\begin{aligned} G_0 &= \frac{1}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2+(\lambda/c))} \\ &\quad \times F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), \\ G_j &= \frac{1}{c}\left(1-\frac{A^2}{B^2}\right)^2 \frac{\Gamma(1+(\lambda/c))\,\Gamma(j)}{\Gamma(j+1+(\lambda/c))} \\ &\quad \times F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), \quad j \geq 1, \end{aligned} \tag{8}$$

where $c = a(B^2 - A^2) = a[2\beta(1-\mu) - 1]$ is the Malthusian parameter of the like-type clone and $F(\cdot,\cdot;\cdot;\cdot)$ is the Gauss hypergeometric function [7], defined as

$$\begin{aligned} &F(a,b;c;z) \\ &\quad = \frac{\Gamma(c)}{\Gamma(b)\,\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt, \\ &\qquad\qquad c > b > 0. \end{aligned} \tag{9}$$

For a detailed derivation, see Appendix A. Note that the supercritical condition also guarantees that the argument of the hypergeometric function remains within its region of definiteness.

It follows that

$$\begin{aligned} \psi_j &= \frac{\lambda G_j}{1-\lambda G_0} \\ &= \left(\frac{\lambda}{c}\left(1-\frac{A^2}{B^2}\right)^2 \right. \\ &\quad \times \left(\frac{\Gamma(1+\lambda/c)\,\Gamma(j)}{\Gamma(j+1+\lambda/c)}\right) \\ &\quad \left. \times F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)\right) \\ &\quad \times \left(1-\frac{\lambda}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2+\lambda/c)} \right. \\ &\quad \left. \times F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)\right)^{-1}, \quad j \geq 1. \end{aligned} \tag{10}$$

Figure 1 shows an example of the limiting mean frequency spectrum for the birth and death process with parameters $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$, based on formula (10). To see how the spectrum varies with different parameter settings, we plot in Figure 2(a), the 3-D surface of a major component of the spectrum, ψ_1, for different α's and μ's. Figures 2(b) and 2(c) illustrate the effect of one parameter on ψ_1 given a fixed value of the other parameter.

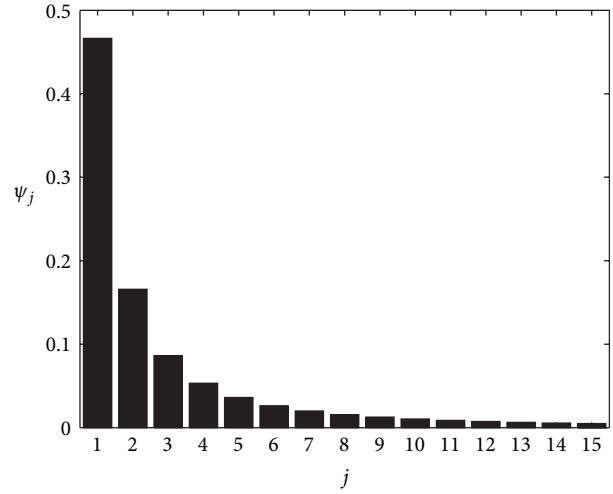

Figure 1: Limiting mean frequency spectrum of the infinite-allele birth and death process with $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$.

We see that for fixed α, increasing μ causes an increase of ψ_1. This can be intuitively explained by the offspring pgf $h(s)$ of the like-type clone. From the pgf expression $h(s) = \alpha + \beta[\mu + (1-\mu)s]^2$, we see that the probability of obtaining one like-type individual in the offspring is $2(1-\alpha)\mu(1-\mu)$, which is an increasing function of μ for a given α, under the constraint $(1-\alpha)(1-\mu) > 1/2$. Therefore, increasing μ will finally lead to an increase of ψ_1. The effect of α on ψ_1 when fixing μ is not so obvious, but we notice that when fixing μ very close to 0, as α approaches 1/2, the process is approximately critical binary fission; therefore, ψ_1 drops down because of almost sure extinction of the process, as seen from the tail behavior of the solid thick line in Figure 2(c).

Arguably, the frequency spectrum can only be observed in finite time. The finite-time mean frequency spectrum can be obtained by computing $G_j(t) = \int_0^t e^{-\lambda x} q_{1j}(x)dx$, $j \geq 0$ numerically. For the birth and death process, this involves the computation of the incomplete hypergeometric function. The following is a valid question in this context. In order to safely use the limiting mean frequency spectrum, how long should the process history be? Figure 3(a) compares the limiting mean frequency spectrum with some long-term mean frequency spectra, for the birth and death process with parameters $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$. We see that under this setting, the long-term mean frequency spectrum is almost identical to the limiting mean frequency spectrum when $t \geq 28$. In general, this result depends strongly on parameters a, α, and μ, for example, small μ leads to longer t. This provides us with some intuitions concerning the sufficiently large t for approximating the limiting mean frequency spectrum. Figure 3(b) illustrates the difference between the finite-time mean frequency spectrum and the limiting mean frequency spectrum as a function of t, for large t, $t \in [15, 35]$ and for $j = 1, 2$, where lines represent the true difference and markers represent the asymptotic approximation by formula (5). To emphasize the agreement for t large, this figure is plotted in semilogarithmic scale. We

see that the true difference drops exponentially fast, and the asymptotic approximation is good for large t.

Given the observed long-term mean frequency spectrum, the parameters θ of the IAMBP, such as α, μ in the birth and death process, can be estimated by equating the observed long-term mean frequency spectrum ψ_{obs} from the sample to the expected limiting mean frequency spectrum ψ_{exp} from formula (3) and solving for the process parameters. In the case of the birth and death process, we may estimate α and μ for example by solving

$$\begin{aligned}
&\left(\Gamma(j_1)\,\Gamma\left(k_1+1+\frac{\lambda}{c}\right)\right.\\
&\quad\times F\left(j_1+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.j_1+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)\\
&\quad\times\left(\Gamma(k_1)\,\Gamma\left(j_1+1+\frac{\lambda}{c}\right)\right.\\
&\qquad\times F\left(k_1+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.k_1+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)^{-1}\\
&=\frac{\psi_{\text{obs}}(j_1)}{\psi_{\text{obs}}(k_1)}\\
&\left(\Gamma(j_2)\,\Gamma\left(k_2+1+\frac{\lambda}{c}\right)\right.\\
&\quad\times F\left(j_2+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.j_2+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)\\
&\quad\times\left(\Gamma(k_2)\,\Gamma\left(j_2+1+\frac{\lambda}{c}\right)\right.\\
&\qquad\times F\left(k_2+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.k_2+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)^{-1}\\
&=\frac{\psi_{\text{obs}}(j_2)}{\psi_{\text{obs}}(k_2)}
\end{aligned}\tag{11}$$

for positive integers $j_1 \neq k_1$, $j_2 \neq k_2$, where λ/c and A^2/B^2 are both functions of α and μ.

There is no explicit solution for such estimator, but numerical search according to some criteria is feasible. Another possibility is to minimize the distance (such as the l^2 norm) between the observed long-term mean frequency spectrum and the expected limiting mean frequency spectrum, that is, $\widehat{\theta} = \arg\min_\theta \|\psi_{\text{obs}} - \psi_{\text{exp}}(\theta)\|_2$.

The estimated parameters can be used to check the goodness of fit of the IAMBP model. Another interesting problem is to test whether two sets of parameters are identical, given two observed mean frequency spectra. A simple approach is to use Pearson's χ^2 test, such as in Kimmel and Mathaes [4]. However, there may be restrictions to applying the χ^2 test, such as small cell counts and inappropriateness due to the finite length of the observed spectrum. This motivates us to develop an interval estimator for the IAMBP parameters.

3. Variance of the Frequency Spectrum

Moment estimators based on the mean frequency spectrum only give point estimates of the process parameters. In order to quantify the uncertainty of point estimates, an interval estimator is needed, which requires more information about the distribution of the statistic $\alpha_t(j)$. First, it can be seen that [2]

$$\alpha_t(j) = I_{0,j}(t) + \sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t - T_n), \tag{12}$$

where $T_1, T_2, \ldots$ are the successive split times of the process, $I_{0,j}(t)$, $I_{n,k,j}(t)$ are two indicators, and $I_{0,j}(t) = 1$ if there are j individuals alive at time t carrying the parental allele, and $I_{n,k,j}(t) = 1$, for $n, k \geq 1$ if the kth individual born at time T_n ($T_n < t$) mutates to a novel allelic type and further produces j individuals carrying this allele t time units later. N_t is the number of split times in $(0, t]$, and U_n is the number of offspring produced at time T_n. Obtaining the distribution of $\alpha_t(j)$ is not elementary. However, it may still be possible to define a confidence interval (CI) based on the first and second moments of $\alpha_t(j)$.

Let $\eta_{i,t}(j) = \text{Var}_i(\alpha_t(j))$ be the variance frequency spectrum; by the law of total variance and independence between the indicators in the expression of $\alpha_t(j)$ (details in Appendix B), we have

$$\begin{aligned}
\eta_{i,t}(j) &= q_{ij}(t)\left[1 - q_{ij}(t)\right]\\
&\quad + im^2\mu^2\left[C(t) + (\lambda + a)\,e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx\right]\\
&\quad + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t}\\
&\quad\times\int_0^t e^{-\lambda x}q_{1j}^2(x)\,dx,
\end{aligned}\tag{13}$$

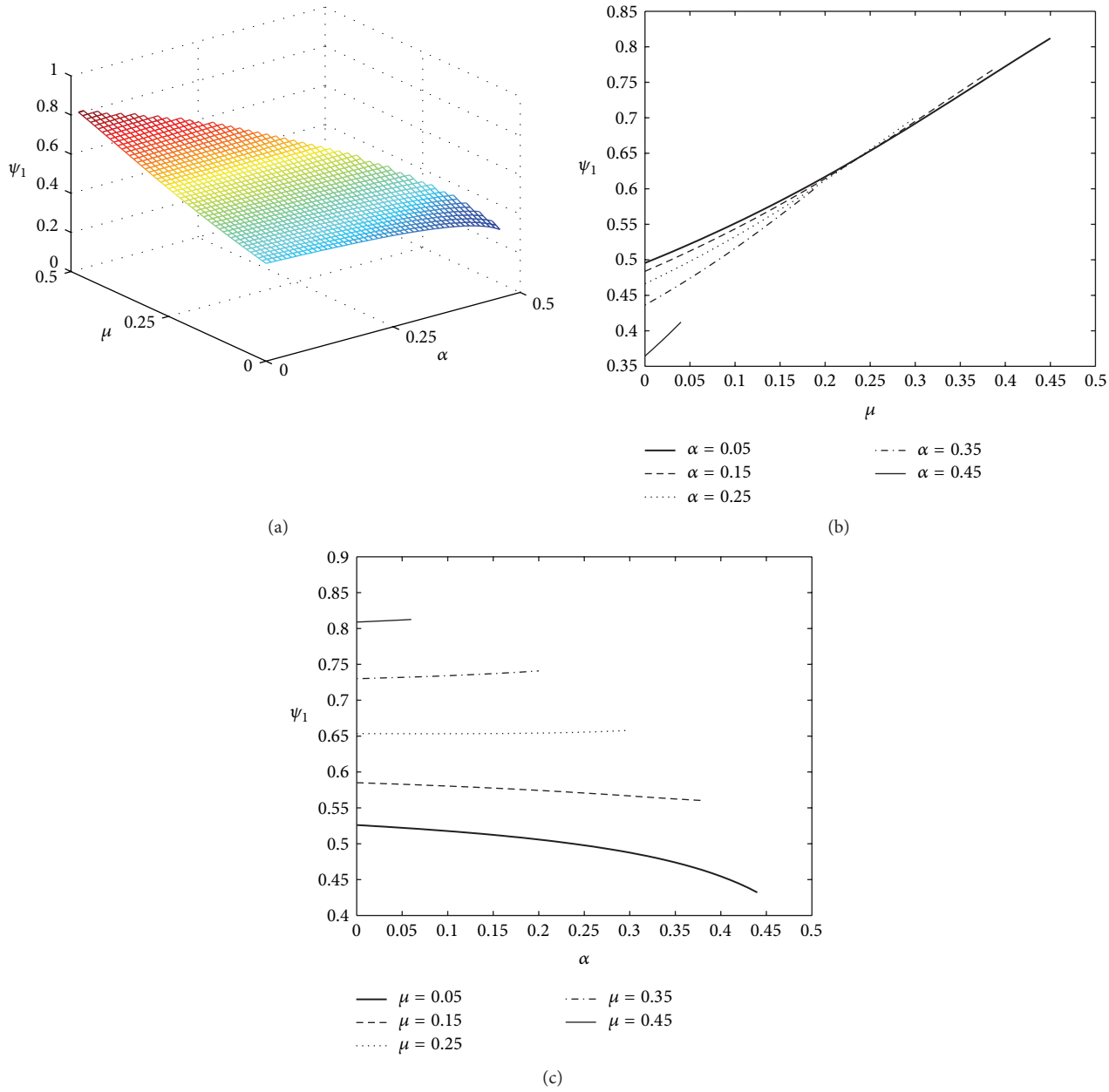

FIGURE 2: (a) Surface of ψ_1 at different α and μ, for the infinite-allele birth and death process. (b) Relation between ψ_1 and μ when fixing α. (c) Relation between ψ_1 and α when fixing μ.

where

$$C(t) = a\int_0^t \left[q_{1j}^2(x) + \left(\sigma^2 + m^2\right)\beta_1^2(x)\right]e^{-a(t-x)}dx - \left[a\int_0^t q_{1j}(x)e^{-a(t-x)}dx\right]^2 - \left[am\int_0^t \beta_1(x)e^{-a(t-x)}dx\right]^2. \quad (14)$$

In Expression (14), $\beta_1(x) = ae^{\lambda x}\int_0^x e^{-\lambda u}q_{1j}(u)du$, and σ^2 is the variance of the offspring distribution, regardless of the allelic types.

Similarly as in Expression (3), we may define a limiting variance frequency spectrum $\xi_j = \lim_{t\to\infty}\eta_{i,t}(j)/(E_i[K_t])^2$. Expression (13) is complicated and usually does not assume an explicit form, even for the special case of the birth and death process. Therefore, we will only give numerical solutions for the finite-time variance frequency spectrum. Figure 4 shows an example of the "2σ"-bands of the finite-time frequency spectrum for the infinite-allele birth and death

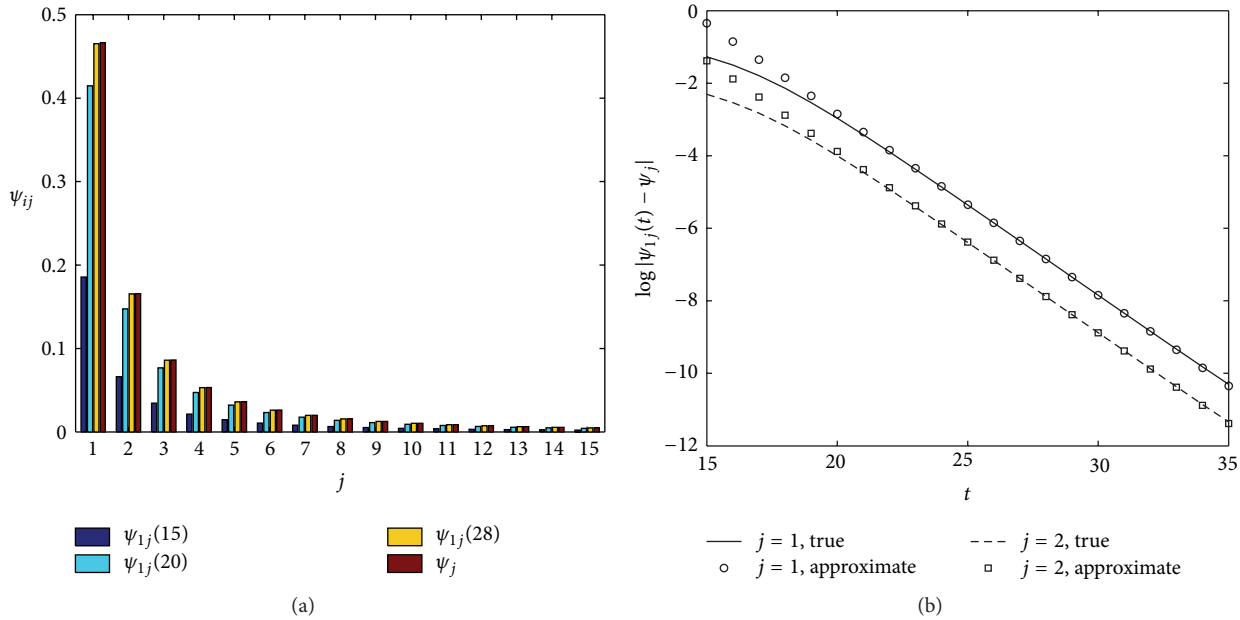

FIGURE 3: Comparison between the finite-time mean frequency spectrum and the limiting mean frequency spectrum of the infinite-allele birth and death process. (a) $\psi_{1j}(t)$ and ψ_j for $1 \leq j \leq 15$, $t = 15, 20, 28$ and ∞. (b) Difference between $\psi_{1j}(t)$ and ψ_j as a function of t, $t \in [15, 35]$ for $j = 1, 2$. Lines represent the true difference and markers represent asymptotic approximations.

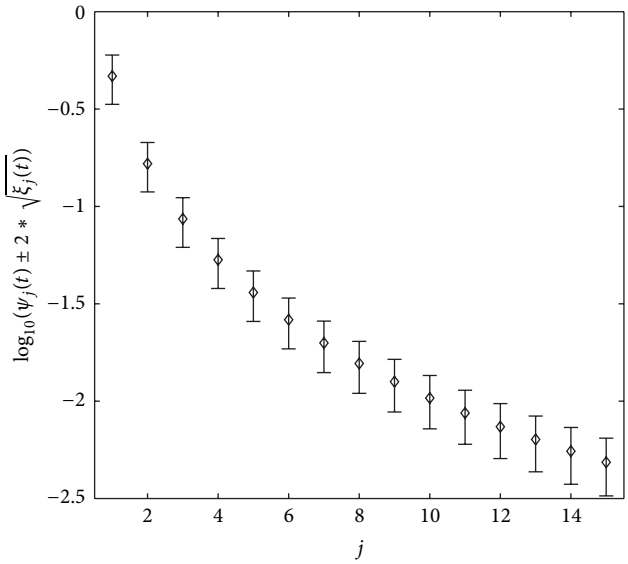

FIGURE 4: "2σ"-bands in semi-logarithmic scale of the finite-time ($t = 28$) frequency spectrum for the infinite-allele birth and death process with $i = 100$, $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$.

process with $i = 100$, $\alpha = 0.25$, $\mu = 10^{-4}$, $a = 1$, and $t = 28$. To emphasize the tail probabilities, we draw this plot in semilogarithmic scale.

From the finite-time variance frequency spectrum, it is possible to define a CI $[\theta_l, \theta_u]$ where the upper and lower bounds can be written as

$$\arg\min_{\theta} \left\| \psi_{\text{obs}}(t) - \left[\psi_{\text{exp}}(\theta, t) \pm 2\sqrt{\xi_{\text{exp}}(\theta, t)} \right] \right\|_2. \quad (15)$$

This CI is useful for checking model validity and for testing whether two observed mean frequency spectra are from the same IAMBP model.

4. Simulation Study

We perform a simulation study of the birth and death process to illustrate the finite-time mean and variance frequency spectra. First we generate samples (genealogical trees) from an IAMBP with birth and death offspring distribution starting from 100 individuals carrying the same parental allele. Due to memory restrictions caused by forward simulation, we limit our simulations to 12 generations and a relatively large mutation probability $\mu = 0.01$. The other parameters of the process are set to be $a = 1$ and $\alpha = 0.25$. At time $t = 2$, we record the number of alleles $\alpha_t(j)$ represented by j copies, for $j = 1, 2, \ldots$. Repeating the simulation 1000 times, we then obtain the simulated finite-time mean and variance frequency spectra from the replicates.

Figure 5 shows side-by-side bar plots of the simulated and expected finite-time mean and variance frequency spectra, for $j = 1, \ldots, 10$. We plot the mean frequency spectrum and the variance frequency spectrum in semi-logarithmic scale to emphasize the tail probabilities. In each bar plot, the first black bar represents the expected finite-time mean frequency spectrum ψ_j or variance frequency spectrum ξ_j. The remaining ten white bars represent ten replicates of the simulated finite-time mean or variance frequency spectrum as described above. We see that for some classes, the expected mean or variance frequency spectrum is slightly different from the simulated spectrum. Beside sampling bias, this may be caused by the small scale of the simulations. For small mutation probability μ, we have to set large initial population

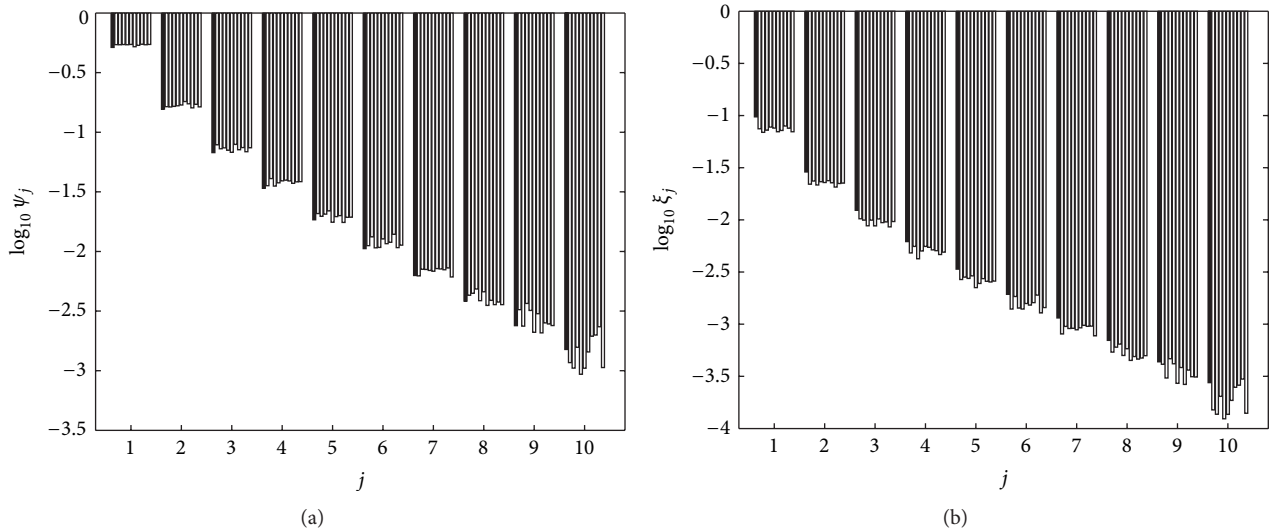

FIGURE 5: Comparison between the simulated and expected finite-time ($t = 2$) frequency spectra of the infinite-allele birth and death process with $i = 100$, $a = 1$, $\alpha = 0.25$, and $\mu = 0.01$. In each class j, $1 \le j \le 10$, the first black bar represents the expected frequency spectrum and the rest 10 white bars represent 10 replicates of the simulated frequency spectrum. (a) mean frequency spectrum ψ_j in semi-logarithmic scale; (b) variance frequency spectrum ξ_j in semi-logarithmic scale.

size i and a long time t to obtain acceptable values of $\psi_{i,t}(j)$ and $\xi_{i,t}(j)$ from the simulated genealogical trees. We note that if one tries to use a naive method to calculate the variance frequency spectrum, that is, assume the proportions ψ_j of alleles having j representatives to be the mean of some independent Bernoulli random variables (they are not independent) and employ $\xi_j = \psi_j(1 - \psi_j)$, such method performs much worse than ξ_j based on the derivation of the variance frequency spectrum.

5. Summary

In this paper, we consider the frequency spectrum of the IAMBP of Pakes [2]. We develop an explicit expression for the limiting mean frequency spectrum for the special case of the birth and death process, which can be stated in terms of the hypergeometric function. We also derive an asymptotic expression for the rate of convergence of the finite-time mean frequency spectrum to the limiting mean frequency spectrum and illustrate the convergence using the birth and death process. We further state and prove a theorem concerning the variance frequency spectrum of the IAMBP, which helps to quantify uncertainty in parameter estimation and hypothesis testing. We illustrate the results using simulations of the birth and death process case.

As noted in the introduction, the frequency spectrum is similar to the Ewens' sampling formula [5] in population genetics, since they both concern the count or frequency spectrum based on an infinite-allele model under neutral selection. However, they differ in several aspects. (1) Our frequency spectrum describes population property under a branching process, whereas the Ewens' sampling formula describes allelic class count probabilities caused by a sampling procedure and further requires the sample size n to be small compared to the size of the whole population which is assumed constant. (2) Our results concerning the frequency spectrum only provide the first and second moments and not the distribution function of the proportion of alleles having a given number of copies at a specified time point, whereas the Ewens' sampling formula gives the joint probability of all allelic classes. We also note that the Poisson-Dirichlet process [8] is usually used to describe the equilibrium behavior of the neutral infinite-allele model. Study of the relation between variations of the frequency spectrum under different models is of our future interest.

The question of validity of the Wright-Fisher and Moran models of population genetics [3], as compared to stochastic population processes such as the IAMBP or O'Connell process [9], has importance for estimation of parameters based on genetic data. As an example, Cyran and Kimmel [10] compared estimates of the age of the Mitochondrial Eve based on the various versions of the Wright-Fisher model with those based on various branching process models. The outcomes showed differences of about 10–15%.

Appendices

A. Derivation of G_j, $j \ge 0$ for the Birth and Death Process

For the birth and death process, the offspring pgf of the like-type individuals clone assumes either the form $h_1(s) = \alpha + \beta[\mu + (1 - \mu)s]^2$ or the form $h_2(s) = [\alpha + \beta\mu + \beta(1 - \mu)s]^2$. In both cases, the backward Kolmogorov equation gives a unified expression for the process pgf $F(s, t)$:

$$\begin{aligned} \frac{\partial F}{\partial t} &= a\left[B^2F^2 - \left(A^2 + B^2\right)F + A^2\right] \\ &= a\left(A^2 + B^2\right)\left[\frac{1}{A^2 + B^2}\left(A^2 + B^2F^2\right) - F\right], \end{aligned} \tag{A.1}$$

under different parameterizations, where for offspring pgf h_1, $A^2 = \alpha + \beta\mu^2$ and $B^2 = \beta(1-\mu)^2$, whereas for offspring pgf h_2, $A^2 = (\alpha + \beta\mu)^2$ and $B^2 = \beta^2(1-\mu)^2$. We see that this process is equivalent to a birth and death process with $\widetilde{a} = a(A^2 + B^2)$ and offspring pgf $\widetilde{h}(s) = (1/(A^2 + B^2))(A^2 + B^2 s^2)$. Using the known result of the birth and death process pgf [6], we obtain

$$F(s,t) = \frac{A^2(1-s) - \left(A^2 - B^2 s\right)e^{-ct}}{B^2(1-s) - \left(A^2 - B^2 s\right)e^{-ct}}, \tag{A.2}$$

where $c = a(B^2 - A^2) > 0$.

To obtain an explicit form for G_j, we may use two approaches. The first approach is to start from finding the pgf of G_j, which then leads to G_j. The second approach is to find $q_{1j}(t)$ directly and then obtain G_j. Both approaches lead to the same result. Here, we give derivation for the second approach only.

From the explicit form of $F(s,t)$, we can directly read $q_{10}(t)$ and $q_{1j}(t)$, $j \geq 1$. Consider

$$\begin{aligned} F(s,t) &= \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct} + B^2\left(e^{-ct}-1\right)s} \\ &\quad + \frac{\left(B^2e^{-ct} - A^2\right)s}{B^2 - A^2e^{-ct} + B^2\left(e^{-ct}-1\right)s} \\ &= \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct}} \cdot \frac{1}{1 - \left(B^2\left(1-e^{-ct}\right)/\left(B^2 - A^2e^{-ct}\right)\right)s} \\ &\quad + \frac{B^2e^{-ct} - A^2}{B^2 - A^2e^{-ct}} \cdot \frac{s}{1 - \left(B^2\left(1-e^{-ct}\right)/\left(B^2 - A^2e^{-ct}\right)\right)s}. \end{aligned} \tag{A.3}$$

Let $w_1 = A^2(1-e^{-ct})/(B^2 - A^2)e^{-ct}$, $w_2 = (B^2e^{-ct} - A^2)/(B^2 - A^2)e^{-ct}$, and $\widetilde{p} = (B^2 - A^2)e^{-ct}/(B^2 - A^2e^{-ct})$, the above expression becomes

$$F(s,t) = w_1\frac{\widetilde{p}}{1 - s\left(1-\widetilde{p}\right)} + w_2\frac{s\widetilde{p}}{1 - s\left(1-\widetilde{p}\right)}, \tag{A.4}$$

and $w_1 + w_2 = 1$.

This is a mixture of two geometric pgf's with the same parameter $\widetilde{p}$ but different supports, one is the set $\{0, 1, \ldots\}$, the other is the set $\{1, 2, \ldots\}$. Therefore,

$$q_{10}(t) = w_1\widetilde{p} = \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct}}, \tag{A.5}$$

$$\begin{aligned} q_{1j}(t) &= w_1\widetilde{p}(1-\widetilde{p})^j + w_2\widetilde{p}(1-\widetilde{p})^{j-1} \\ &= \frac{\left(B^2 - A^2\right)^2\left[B^2\left(1-e^{-ct}\right)\right]^{j-1}e^{-ct}}{\left(B^2 - A^2e^{-ct}\right)^{j+1}}, \quad j \geq 1. \end{aligned} \tag{A.6}$$

Hence,

$$\begin{aligned} G_0 &= \int_0^\infty e^{-\lambda t}q_{10}(t)\,dt \\ &= \int_0^\infty \frac{A^2\left(1-e^{-ct}\right)}{\left(B^2 - A^2e^{-ct}\right)}e^{-\lambda t}dt \\ &= \frac{1}{c}\frac{A^2}{B^2}\int_0^1 \frac{v^{(\lambda/c)-1}(1-v)}{1 - \left(A^2/B^2\right)v}dv \\ &= \frac{1}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2 + (\lambda/c))}F\left(1, \frac{\lambda}{c};\ 2 + \frac{\lambda}{c}; \frac{A^2}{B^2}\right), \end{aligned} \tag{A.7}$$

$$\begin{aligned} G_j &= \int_0^\infty e^{-\lambda t}q_{1j}(t)\,dt \\ &= \int_0^\infty \frac{\left(B^2 - A^2\right)^2\left[B^2\left(1-e^{-ct}\right)\right]^{j-1}}{\left(B^2 - A^2e^{-ct}\right)^{j+1}}e^{-(\lambda+c)t}dt \\ &= \frac{\left(B^2 - A^2\right)^2}{cB^4}\int_0^1 \frac{v^{\lambda/c}(1-v)^{j-1}}{\left(1 - \left(A^2/B^2\right)v\right)^{j+1}}dv \\ &= \frac{1}{c}\left(1 - \frac{A^2}{B^2}\right)^2\frac{\Gamma(1 + (\lambda/c))\,\Gamma(j)}{\Gamma\left(j + 1 + (\lambda/c)\right)} \\ &\quad \times F\left(j+1, 1 + \frac{\lambda}{c}; j + 1 + \frac{\lambda}{c}; \frac{A^2}{B^2}\right), \quad j \geq 1. \end{aligned} \tag{A.8}$$

B. Derivation of the Variance Frequency Spectrum for IAMBP

Let $T_1, T_2, \ldots$ be the successive split times, let N_t be the number of split times till time t, and let U_n be the number of offspring produced at split time T_n. Consider that at time t, the alleles which are represented by j individuals are from two sources: the initial allele or the mutant alleles. Correspondingly, we define two indicator functions. $I_{0,j}(t) = 1$ if there are j individuals carrying the initial allele alive at time t, and $I_{n,k,j}(t) = 1$, for $n, k \geq 1$ if the kth individual born at time T_n ($T_n < t$) mutates to a new allelic type and further produces j individuals carrying this allele t units later. Then

$$\alpha_t(j) = I_{0,j}(t) + \sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}\left(t - T_n\right). \tag{B.1}$$

For each n, $I_{n,k,j}(t)$ is independent of U_n and T_n, as well as $I_{0,j}(t)$, and it can be seen that $E_i[I_{0,j}(t)] = q_{ij}(t)$, $\mathrm{Var}_i(I_{0,j}(t)) = q_{ij}(t)[1 - q_{ij}(t)]$, $E[I_{n,k,j}(t) \mid U_n, T_n] = \mu q_{1j}(t)$, $\mathrm{Var}(I_{n,k,j}(t) \mid$

$U_n, T_n) = \mu q_{1j}(t)[1-\mu q_{1j}(t)]$. By the law of total variance, the variance frequency spectrum takes the form

$$\begin{aligned}
&\eta_{i,t}(j)\\
&\quad = \mathrm{Var}_i\left(\alpha_t(j)\right)\\
&\quad = \mathrm{Var}_i\left(I_{0,j}(t)\right) + \mathrm{Var}_i\left(\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\right)\\
&\quad = q_{ij}(t)\left[1-q_{ij}(t)\right]\\
&\qquad + \mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right)\\
&\qquad + E_i\left[\mathrm{Var}\left(\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right)\right]
\end{aligned} \tag{B.2}$$

For the second term, we know from independence among the indicator functions conditional on N_t that

$$\begin{aligned}
&\mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right)\\
&\qquad = m^2\mu^2\mathrm{Var}_i\left(\sum_{n=1}^{N_t} q_{1j}(t-T_n)\right).
\end{aligned} \tag{B.3}$$

The variance on the right hand side can be obtained from the following theorem, which is an analogue to Lemma 3.1.1 in [2]: $\beta_i(t) = E_i[\sum_{n=1}^{N_t}\alpha(t-T_n)] = iae^{\lambda t}\int_0^t e^{-\lambda x}\alpha(x)dx$.

Theorem B.1. *Let $\alpha(t)$ be a bounded continuous function. Then*

$$\begin{aligned}
\gamma_i(t) &= \mathrm{Var}_i\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\right)\\
&= iC(t) + i(\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx,
\end{aligned} \tag{B.4}$$

where

$$\begin{aligned}
C(t) &= a\int_0^t\left[\alpha^2(x) + \left(\sigma^2+m^2\right)\beta_1^2(x)\right]e^{-a(t-x)}dx\\
&\quad - \left[a\int_0^t \alpha(x)e^{-a(t-x)}dx\right]^2\\
&\quad - \left[am\int_0^t \beta_1(x)e^{-a(t-x)}dx\right]^2.
\end{aligned} \tag{B.5}$$

Proof. By independence of family lines, $\gamma_i(t) = i\gamma_1(t)$. By the law of total variance,

$$\begin{aligned}
\gamma_1(t) &= \mathrm{Var}_1\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\right)\\
&= \mathrm{Var}\left(E\left[\sum_{n=1}^{N_t}\alpha(t-T_n)\mid T_1, U_1\right]\right)\\
&\quad + E\left[\mathrm{Var}\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\mid T_1, U_1\right)\right]\\
&= \mathrm{Var}\left(\alpha(t-T_1) + U_1 E\left[\sum_{n=1}^{N_t'}\alpha\left(t-T_1-T_n'\right)\mid T_1\right]\right)\\
&\quad + E\left[U_1\,\mathrm{Var}\left(\sum_{n=1}^{N_t'}\alpha\left(t-T_1-T_n'\right)\mid T_1\right)\right],
\end{aligned} \tag{B.6}$$

where $T_n' = T_n - T_1$ and N_t' is the number of split times in $(T_1, t]$. The right hand side can be further written as

$$\begin{aligned}
&E\left[\alpha^2(t-T_1)\right] - E^2\left[\alpha(t-T_1)\right] + \left(\sigma^2+m^2\right)E\left[\beta_1^2(t-T_1)\right]\\
&\quad - m^2E^2\left[\beta_1(t-T_1)\right] + mE\left[\gamma_1(t-T_1)\right]\\
&\quad = a\int_0^t\left[\alpha^2(t-x) + \left(\sigma^2+m^2\right)\beta_1^2(t-x)\right.\\
&\qquad \left. + m\gamma_1(t-x)\right]e^{-ax}dx\\
&\qquad - \left[a\int_0^t \alpha(t-x)e^{-ax}dx\right]^2 - \left[am\int_0^t \beta_1(t-x)e^{-ax}dx\right]^2\\
&\quad = am\int_0^t \gamma_1(t-x)e^{-ax}dx + C(t).
\end{aligned} \tag{B.7}$$

Differentiating both sides and solving the resulting differential equation, we obtain

$$\begin{aligned}
\gamma_1(t) &= e^{\lambda t}\int_0^t e^{-\lambda x}\left[C'(x) + aC(x)\right]dx\\
&= C(t) + (\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx.
\end{aligned} \tag{B.8}$$

□

Replacing $\alpha(x)$ by $q_{1j}(x)$ and replacing $\beta_1(x)$ by $ae^{\lambda x}\int_0^x e^{-\lambda u}q_{1j}(u)du$, we see that (B.3) becomes

$$\begin{aligned}
&\mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right)\\
&\qquad = im^2\mu^2\left[C(t) + (\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx\right].
\end{aligned} \tag{B.9}$$

For the third term, we have the expression inside the expectation as

$$\begin{aligned}
&\sum_{n=1}^{N_t} \mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right)\right) \\
&\quad = \sum_{n=1}^{N_t}\left\{\mathrm{Var}\left(E\left[\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right) \mid U_n, T_n\right]\right)\right. \\
&\qquad \left. + E\left[\mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right) \mid U_n, T_n\right)\right]\right\} \\
&\quad = \sum_{n=1}^{N_t}\left[m\mu q_{1j}\left(t-T_n\right) + \left(\sigma^2 - m\right)\mu^2 q_{1j}^2\left(t-T_n\right)\right],
\end{aligned} \tag{B.10}$$

where σ^2 is the variance of the offspring distribution, regardless of the allelic types. Therefore,

$$\begin{aligned}
&E_i\left[\sum_{n=1}^{N_t} \mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right)\right)\right] \\
&\quad = m\mu E_i\left[\sum_{n=1}^{N_t} q_{1j}\left(t-T_n\right)\right] \\
&\qquad + \left(\sigma^2 - m\right)\mu^2 E_i\left[\sum_{n=1}^{N_t} q_{1j}^2\left(t-T_n\right)\right] \\
&\quad = iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t} \\
&\qquad \times \int_0^t e^{-\lambda x} q_{1j}^2(x)\,dx
\end{aligned} \tag{B.11}$$

By (B.2), (B.9) and (B.11), the final expression of $\eta_{i,t}(j)$ is then

$$\begin{aligned}
\eta_{i,t}(j) &= q_{ij}(t)\left[1 - q_{ij}(t)\right] \\
&\quad + im^2\mu^2\left[C(t) + (\lambda + a)e^{\lambda t}\int_0^t e^{-\lambda x} C(x)\,dx\right] \\
&\quad + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t} \\
&\quad \times \int_0^t e^{-\lambda x} q_{1j}^2(x)\,dx.
\end{aligned} \tag{B.12}$$

Acknowledgments

The authors thank an anonymous referee for suggesting an asymptotic expression, which following some revisions, became (5). Part of M. Kimmel's work has been carried out in Fall 2011 when he was visiting the Institute of Advance Study at the Warwick University, supported by grant from IAS and EPSCR. His work was also supported by the Polish NCN Grant NN519579938.

References

[1] R. C. Griffiths and A. G. Pakes, "An infinite-alleles version of the simple branching process," *Advances in Applied Probability*, vol. 20, no. 3, pp. 489–524, 1988.

[2] A. G. Pakes, "An infinite alleles version of the Markov branching process," *Journal of Australian Mathematical Society A*, vol. 46, no. 1, pp. 146–169, 1989.

[3] W. J. Ewens, *Mathematical Population Genetics*, vol. 9, Springer, New York, NY, USA, 1979.

[4] M. Kimmel and M. Mathaes, "Modeling neutral evolution of Alu elements using a branching process," *BMC Genomics*, vol. 11, supplement 1, article S11, 2010.

[5] W. J. Ewens, "The sampling theory of selectively neutral alleles," *Theoretical Population Biology*, vol. 3, pp. 87–112, 1972.

[6] K. B. Athreya and P. E. Ney, *Branching Processes*, Springer, Berlin, Germany, 1972.

[7] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 1972.

[8] J. F. C. Kingman, S. J. Taylor, A. G. Hawkes et al., "Random discrete distribution," *Journal of the Royal Statistical Society B*, vol. 37, pp. 1–22, 1975.

[9] N. O'Connell, "The genealogy of branching processes and the age of our most recent common ancestor," *Advances in Applied Probability*, vol. 27, no. 2, pp. 418–442, 1995.

[10] K. A. Cyran and M. Kimmel, "Alternatives to the Wright-Fisher model: the robustness of mitochondrial Eve dating," *Theoretical Population Biology*, vol. 78, no. 3, pp. 165–172, 2010.