J. Math. Biol. (1995) 33: 602–618

Journal of Mathematical Biology

The distribution of rare alleles

Paul Joyce[1], **Simon Tavaré**[2]

[1] Department of Mathematics and Statistics, University of Idaho, Moscow, ID 83843, USA
[2] Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113

Received 1 December 1993; received in revised form 30 August 1994

Abstract. Population geneticists have long been interested in the behavior of rare variants. The definition of a rare variant has been the subject of some debate, centered mainly on whether alleles with small relative frequency should be considered rare, or whether alleles with small numbers should be. We study the behavior of the counts of rare alleles in samples taken from a population genetics model that allows for selection and infinitely-many-alleles mutation structure. We show that in large samples the counts of rare alleles – those represented once, twice, . . . – are approximately distributed as a Poisson process, with a parameter that depends on the total mutation rate, but not on the selection parameters. This result is applied to the problem of estimating the fraction of neutral mutations.

Key words: Selection – Poisson approximation – Ewens sampling formula – Functional central limit theorem

1 Introduction

A question of some importance in studies of genetic variation is: What is a rare allele? Traditionally, rare alleles have been defined in terms of their *relative frequencies*. For example, Kimura (1983a) defines a rare variant as an allele with a relative frequency of less than q, for some small pre-specified value of q such as 0.01. This definition treats as equally rare an allele arising twice in a sample of size 100 and an allele arising 200 times in a sample of size 10,000. Presumably, though, these two sample configurations would be interpreted very differently (Thompson et al., 1992).

Consider a sample of n genes taken from a single locus in a large stationary population. We define a *rare allele* as one that arises at most b times in the sample. We are interested in the behavior of such rare variants

when the sample size n is large and b is chosen either to be fixed or perhaps to tend to infinity with n in such a way that $b/n \to 0$. This definition of rare alleles focuses directly on the *counts* of alleles, rather than their relative frequencies.

In this paper we study the asymptotic structure of the frequencies of rare variants under the effects of mutation and selection. We study a particular population genetic model, variants of which have been discussed by several authors including Li (1979), Ewens and Li (1980), Kingman (1980), Griffiths (1983), Watterson (1987, 1993) and Ethier and Kurtz (1994). Under this model the alleles at the locus in question are divided into d classes, and the fitness of a genotype is determined by the classes to which the two constituent genes belong.

Related to the issue of rare variants is the problem of estimating the fraction of neutral mutations, P_{neut}. Suppose the fractions in the d classes are $f_1, f_2, \ldots, f_d$. If we think of class 1 as being neutral and the others selective, then $P_{\text{neut}} = f_1$. Kimura (1983a) proposed an estimator of P_{neut} based on data from several loci. Roughly speaking, his estimator uses the number of rare variants (defined in terms of relative frequencies of alleles) to estimate θ_T, the overall mutation rate, and the heterozygosity to estimate the effective neutral mutation rate θ_1; then $f_1 = \theta_1/\theta_T$. The statistical behavior of this estimator was discussed by Watterson (1987). Watterson (1993) assessed the possibility of estimating P_{neut} by maximum likelihood methods from data at a single locus, and Joyce (1994a) established that such an MLE could not be consistent. However, it is the case that θ_T can be estimated consistently, as the results of this paper show.

In his paper, Kimura (1983a) commented that

> "It can be shown mathematically (see, e.g., Kimura 1983b, p. 227), in the neighborhood of $x = 0$, the population behavior of alleles in general, including those having mild selective advantage or disadvantage, is essentially the same as that of selectively neutral mutants."

This statement refers to alleles with small relative frequency. It is the purpose of this paper to establish a rigorous analog of this statement for rare variant allele counts.

Specifically, we show that if $C_j \equiv C_j(n)$ is the number of alleles represented j times in a sample of size n, then the joint distribution of $(C_1, \ldots, C_b)$ is well approximated by the joint law of independent Poisson random variables Z_j with parameters $EZ_j = \theta_T/j$ as long as $b/n \to 0$. We also provide an estimate of the total variation distance between the two distributions. One interesting consequence of this result is that the rare variants have approximately the Ewens sampling distribution (Ewens, 1972) with parameter θ_T, *regardless of the values of the selection parameters*. This may be viewed as an analog of Kimura's observation in the context of rare variant counts.

The results are established by deriving approximations for the allele counts in different classes using the Poisson approximation methods for the Ewens sampling formula provided by Arratia et al. (1992). In addition, we

obtain functional central limit theorems for the allele counts, together with a rate of convergence.

1.1 The Ewens sampling formula

Here we review some of the structure of the Ewens sampling formula. Consider a random sample of n genes taken from a stationary, infinitely-many-neutral-alleles model. Let $C_j^*(n)$ denote the number of alleles represented j times in the sample, and define

$$\boldsymbol{C}_b^*(n) \equiv (C_1^*(n), \ldots, C_b^*(n)),$$

for $1 \leqq b \leqq n$. Define $\boldsymbol{Z}_+ \equiv \{0, 1, \ldots\}$, and let $\boldsymbol{a} \in \boldsymbol{Z}_+^n$. According to the Ewens sampling formula, the probability that $\boldsymbol{C}_n^*(n) = \boldsymbol{a}$ is given by

$$P_n(\theta; \boldsymbol{a}) = 1\left\{\sum_{i=1}^{n} i a_i = n\right\} \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!} \tag{1}$$

where

$$x_{(n)} = x(x+1)\cdots(x+n-1), \quad x_{(0)} = 1,$$

$1\{A\}$ is the indicator of A, and $\theta = 4Nu$ is the neutral mutation rate.

There has recently been a resurgence of interest in Poisson process approximations for combinatorial objects, among them the Ewens sampling formula (cf. Barbour, Holst and Janson (1992), Arratia et al. (1992, 1994)). The main idea of these methods is to approximate the behavior of a complicated stochastic process by a process with simpler dependence structure. The adequacy of the approximation is conveniently measured in terms of total variation distance, about which we recall some basic facts.

Let X and Y be discrete random variables, and let $d_{TV} \equiv d_{TV}(X, Y)$ denote the total variation distance between the law of X and the law of Y:

$$d_{TV}(X, Y) \equiv \sup_A |\boldsymbol{P}(X \in A) - \boldsymbol{P}(Y \in A)|. \tag{2}$$

An equivalent definition of d_{TV} is

$$d_{TV}(X, Y) = \frac{1}{2} \sum_a |\boldsymbol{P}(X = a) - \boldsymbol{P}(Y = a)|. \tag{3}$$

Further,

$$d_{TV}(X, Y) = \inf \boldsymbol{P}(X \neq Y), \tag{4}$$

the infimum being taken over all couplings of X and Y on the same probability space.

For the Ewens sampling formula, the approximating process is a sequence of independent Poisson random variables $Z_1, Z_2, \ldots$, with means $\boldsymbol{E}Z_j = \theta/j$. Define $\boldsymbol{Z}_b \equiv (Z_1, \ldots, Z_b)$, and let $d_b^\theta(n)$ be the total variation distance between $\boldsymbol{C}_b^*(n)$ and $\boldsymbol{Z}_b$:

$$d_b^\theta(n) \equiv d_{TV}(\boldsymbol{C}_b^*(n), \boldsymbol{Z}_b). \tag{5}$$

Arratia et al. (1992) established the following results.

Lemma 1. *For $1 \leqq b \leqq n$, let $d_b^\theta(n)$, defined in (5), be the total variation distance between the laws of $C_b^*(n)$ and Z_b. There is a constant $c(\theta) > 0$ such that*

$$d_b^\theta(n) \leqq c(\theta)\frac{b}{n}. \tag{6}$$

The constant $c(\theta)$ is given by

$$c(\theta) = \begin{cases} \theta(\theta+1) & \theta \geqq 1 \\ \max(2, \theta(\theta+2)), & 0 \leqq \theta < 1. \end{cases} \tag{7}$$

Lemma 2. *For any $A \subseteq \{1,2,\ldots,n\}$,*

$$\sum_{j\in A} EC_j^*(n) \leqq 1 + \sum_{j\in A} EZ_j.$$

2 The selection models

In this section we describe a class of models that play an important role in estimating the fraction of neutral mutants. The description follows Ethier and Kurtz (1987, 1994).

Consider a diploid population of N individuals in which each of the $2N$ genes is assigned a type $x \in [0,1]$. Suppose the genes in the current generation are labelled $x_1, \ldots, x_{2N}$. We form the next generation of genes as follows. To produce each new gene, a pair of genes is selected at random from the population. The probability that the ith and jth genes are selected is

$$\frac{w_N(x_i, x_j)}{\sum_{1\leqq l,m\leqq 2N} w_N(x_l, x_m)},$$

where $w_N(x,y) \geqq 0$ is symmetric and measures the fitness of an individual with genotype (x,y). Next, one of the individual's genes is chosen at random and subjected to mutation. A gene labelled z mutates to a gene with label in $A \in \mathscr{B}[0,1]$, the Borel subsets of $[0,1]$, according to the transition function $P_N(z,A)$. Given the types in the current generation, the $2N$ genes in the next generation are independently and identically distributed. We are interested in the case in which alleles are classified into one of $d \geqq 2$ classes, the selection function being constant for genes in a given class. To this end, let $f_1, \ldots, f_d > 0$ satisfy $f_1 + \cdots + f_d = 1$, and define intervals $I_j \subseteq [0,1]$ by $I_1 = [0, f_1)$, $I_d = [1 - f_d, 1]$ and for $j = 2, \ldots, d-1$

$$I_j = [f_1 + \cdots + f_{j-1}, f_1 + \cdots + f_j).$$

We suppose that

$$w_N(x,y) = 1 + \frac{1}{2N}\sigma(x,y),$$

where

$$\sigma(x,y) = \sum_{r,s=1}^{d} \sigma_{rs} 1_{I_r \times I_s}(x,y), \tag{8}$$

and $\Sigma \equiv (\sigma_{lm})$ is a symmetric $d \times d$ matrix. We also assume an infinitely-many-alleles mutation structure in which

$$P_N(z, A) = \left(1 - \frac{\theta_T}{4N}\right)\delta_z(A) + \frac{\theta_T}{4N}\lambda(A), \tag{9}$$

where λ is the uniform distribution on $[0, 1]$, δ_z is point mass at z, and $u \equiv \theta_T/(4N)$ is the mutation rate per gene per generation.

It is conventional to examine the behavior of such models in the limit as $N \to \infty$. In this case, one keeps track of the fraction of genes of different types, perhaps using the theory of measure-valued diffusions. The details may be found in Ethier and Kurtz (1994). For our purposes, it is enough to record the structure of the stationary measure of the limit process. We denote the fraction of genes in selective classes $1, \ldots, d$ by $(P_1, \ldots, P_d)$, a random variable in the simplex $\Delta_d \equiv \{(p_1, \ldots, p_d): p_i \geqq 0, 1 \leqq i \leqq d; p_1 + \cdots + p_d = 1\}$. The distribution μ of $(P_1, \ldots, P_d)$ is given by Ethier and Kurtz (1994, (4.63)) as

$$\mu(dx) = Cx_1^{\theta_1 - 1} \cdots x_d^{\theta_d - 1} \exp\left(\sum_{l,m=1}^{d} x_l \sigma_{lm} x_m\right) dx_1 \cdots dx_{d-1}, \tag{10}$$

where C is a normalizing constant, and

$$\theta_i \equiv \theta_T f_i, \quad i = 1, \ldots, d. \tag{11}$$

See Wright (1949, p. 383) for early applications of (11), and Watterson (1978) for an application having $\theta_1 = \cdots = \theta_d$. Ethier and Kurtz (1994) also show that, conditional on the class frequencies, the (renormalized) allele frequencies in decreasing order in classes $1, \ldots, d$ are independently distributed with Poisson–Dirichlet distributions with parameters $\theta_1, \ldots, \theta_d$ respectively. The case of genic selection ($\sigma_{lm} = \sigma_l + \sigma_m$) was treated by Griffiths (1983), following on from work of Li (1979), Ewens and Li (1980), and Kingman (1980).

3 Poisson approximations

In this section we use some of the results of Arratia et al. (1992) to analyze the selection model in Sect. 2. Consider a sample of n genes taken from a d-class model at stationarity, and define $C_{ij}(n)$ to be the number of alleles from selective class i represented j times in the sample, $i = 1, \ldots, d$. It is not usually known which genes belong to which selective classes, and the observable counts are just

$$C_j(n) \equiv \sum_{i=1}^{d} C_{ij}(n), \quad j = 1, 2, \ldots, n. \tag{12}$$

For $i = 1, \ldots, d$, let Y_{in} be the number of genes of class i in the sample, and write $\boldsymbol{Y}_n \equiv (Y_{1n}, \ldots, Y_{dn})$. Conditional on the population class frequencies $(P_1, \ldots, P_d)$, $\boldsymbol{Y}_n$ has a multinomial distribution with parameters n and $(P_1, \ldots, P_d)$. Conditional on $\boldsymbol{Y}_n = (y_1, \ldots, y_d)$, the genes in different selective

classes are distributed as independent samples of sizes $y_1, y_2, \ldots, y_d$ from Poisson–Dirichlet distributions with parameters $\theta_1, \theta_2, \ldots, \theta_d$ respectively. It follows that the counts $\{C_{ij}(n), j = 1, \ldots, n\}$ are distributed as independent Ewens distributions (1) with parameters θ_i and y_i, $i = 1, \ldots, d$.

3.1 A total variation estimate

The first lemma shows that the counts of allele frequencies in the $d \geqq 2$ classes are asymptotically independent, with Poisson distributions. The result is a special case of Corollary 1 below, and the proof is omitted. For $i = 1, \ldots, d$, define

$$C_{i\infty}(n) = (C_{i1}(n), \ldots, C_{in}(n), 0, \ldots) .$$

Lemma 3. *As* $n \to \infty$,

$$(C_{i\infty}(n), i = 1, \ldots, d) \Rightarrow (Z_{i\infty}, i = 1, \ldots, d) , \tag{13}$$

where $\Rightarrow$ *denotes convergence in distribution and* $Z_{i\infty} \equiv (Z_{i1}, Z_{i2}, \ldots)$, $i = 1, \ldots, d$ *are mutually independent Poisson processes on* $N \equiv \{1, 2, \ldots\}$, *with means given by*

$$EZ_{ij} = \frac{\theta_i}{j} . \tag{14}$$

Remark. This result says that in large samples the counts of rare alleles have approximately independent Poisson distributions with parameters that do not depend on the selection scheme; the effects of selection are washed out in the limit process.

It is now of some interest to assess the quality of the approximation in Lemma 3. To this end, define

$$C_{ib_i}(n) \equiv (C_{i1}(n), C_{i2}(n), \ldots, C_{ib_i}(n)) ,$$

and

$$Z_{ib_i} \equiv (Z_{i1}, \ldots, Z_{ib_i}) ,$$

for $1 \leqq b_i \leqq n$, $i = 1, \ldots, d$. We wish to estimate the total variation distance

$$d^*_{b_1, \ldots, b_d}(n) \equiv d_{TV}((C_{ib_i}(n), i = 1, \ldots, d), (Z_{ib_i}, i = 1, \ldots, d))$$

between the counts of alleles in class i with at most $b_i \equiv b_i(n)$ representatives ($1 \leqq i \leqq d$), and the corresponding counts for the limiting Poisson processes. We use the following lemma.

Lemma 4. *Let* $(X_1, \ldots, X_d)$, $(Y_1, \ldots, Y_d)$ *and* V *(possibly vector-valued) be discrete random variables. Assume that* $Y_1, \ldots, Y_d$ *are independent, and that conditional on* V *the random variables* $X_1, \ldots, X_d$ *are independent. Then*

$$d_{TV}((X_1, \ldots, X_d), (Y_1, \ldots, Y_d)) \leqq \sum_{i=1}^{d} \sum_{m} P(V = m) d_{TV}(X_i | V = m, Y_i) ,$$

where $X_i | V = m$ *denotes a random variable with distribution given by the conditional law of* X_i, *given* $V = m$.

Proof.

$$
\begin{aligned}
&d_{TV}((X_1,\ldots,X_d),(Y_1,\ldots,Y_d)) \\
&\quad=\frac{1}{2}\sum_{a_1,\ldots,a_d}|P(X_1=a_1,\ldots,X_d=a_d)-P(Y_1=a_1,\ldots,Y_d=a_d)| \\
&\quad=\frac{1}{2}\sum_{a_1,\ldots,a_d}\left|\sum_m P(V=m)\prod_{i=1}^d P(X_i=a_i|V=m)-\prod_{i=1}^d P(Y_i=a_i)\right| \\
&\quad=\frac{1}{2}\sum_{a_1,\ldots,a_d}\left|\sum_m P(V=m)\left(\prod_{i=1}^d P(X_i=a_i|V=m)-\prod_{i=1}^d P(Y_i=a_i)\right)\right| \\
&\quad\leqq\frac{1}{2}\sum_{a_1,\ldots,a_d}\sum_m P(V=m)\left|\prod_{i=1}^d P(X_i=a_i|V=m)-\prod_{i=1}^d P(Y_i=a_i)\right| \\
&\quad=\frac{1}{2}\sum_{a_1,\ldots,a_d}\sum_m P(V=m)\left|\sum_{i=1}^d\left(\prod_{l=1}^{i-1}P(X_l=a_l|V=m)\right)\left(\prod_{l=i+1}^d P(Y_l=a_l)\right)\right. \\
&\qquad\left.\times(P(X_i=a_i|V=m)-P(Y_i=a_i))\right| \\
&\quad\leqq\frac{1}{2}\sum_{a_1,\ldots,a_d}\sum_m P(V=m)\sum_{i=1}^d\left(\prod_{l=1}^{i-1}P(X_l=a_l|V=m)\right)\left(\prod_{l=i+1}^d P(Y_l=a_l)\right) \\
&\qquad\times|P(X_i=a_i|V=m)-P(Y_i=a_i))| \\
&\quad=\sum_{i=1}^d\sum_m P(V=m)\sum_{a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_d}\left(\prod_{l=1}^{i-1}P(X_l=a_l|V=m)\right) \\
&\qquad\times\left(\prod_{l=i+1}^d P(Y_l=a_l)\right)\frac{1}{2}\sum_{a_i}|P(X_i=a_i|V=m)-P(Y_i=a_i)| \\
&\quad=\sum_{i=1}^d\sum_m P(V=m)d_{TV}(X_i|V=m,Y_i). \qquad\square
\end{aligned}
$$

The next theorem provides a bound on the total variation distance d^*. For $i=1,\ldots,d$, let Y_{in} be the number of genes in the sample of size n that belong to selective class i, and let P_i be the fraction of the population that is in selective class i. While we are particularly interested in the case where the P_i are distributed according to (10), the next result is somewhat more general. Let F_i be the distribution function of P_i. The main result of the paper is:

Theorem 1. *For $1\leqq b_i\leqq n$, $i=1,\ldots,d$, we have*

$$
d^*_{b_1,\ldots,b_d}\leqq\sum_{i=1}^d\tau_i(b_i)\,, \tag{15}
$$

where

$$
\begin{aligned}
\tau_i(b)&=P(Y_{in}<b)+c(\theta_i)P(P_i\leqq(b+1)/(n+1)) \\
&\quad+c(\theta_i)\frac{b+1}{n+1}\int_{(b+1)/(n+1)}^1 p^{-1}F_i(dp)\,,
\end{aligned} \tag{16}
$$

the constant $c(\theta_i)$ being given in equation (7).

Proof. Let $d_b^\theta(r)$ be the total variation distance, defined in (5), between the law of the first b components of counts following the Ewens sampling formula with sample size r and parameter θ, and the law of the first b components of the corresponding Poisson process in which the jth component has mean θ/j. From Lemma 4, it follows that

$$d^*_{b_1,\ldots,b_d}(n) \leqq \sum_{i=1}^{d} \sum_{y_1,\ldots,y_d} P(Y_n = (y_1,\ldots,y_d)) d^{\theta_i}_{b_i}(y_i)$$

$$= \sum_{i=1}^{d} \sum_{y_i} P(Y_{in} = y_i) d^{\theta_i}_{b_i}(y_i) \tag{17}$$

We estimate the size of the ith term on the right of (17). From Lemma 1 we see that

$$\sum_{y=0}^{n} P(Y_{in} = y) d^{\theta_i}_{b_i}(y) \leqq \sum_{y=0}^{b_i-1} P(Y_{in} = y) 1 + c(\theta_i) \sum_{y=b_i}^{n} P(Y_{in} = y) \frac{b_i}{y}$$

$$= P(Y_{in} < b_i) + c(\theta_i) \sum_{y=b_i}^{n} P(Y_{in} = y) \frac{b_i}{y}. \tag{18}$$

Conditional on $(P_1,\ldots,P_d)$, $Y_{in} \sim \mathrm{Bin}(n, P_i)$, where $\mathrm{Bin}(n, P_i)$ denotes a binomial random variable with parameters n and P_i. Conditional on $P_i = p$, we have

$$\sum_{y=b_i}^{n} P(Y_{in} = y | P_i = p) \frac{b_i}{y} = \sum_{y=b_i}^{n} \binom{n}{y} p^y (1-p)^{n-y} \frac{b_i}{y}$$

$$= \sum_{y=b_i}^{n} \left(\binom{n+1}{y+1} p^{y+1} (1-p)^{n-y} \right) \frac{b_i}{p(n+1)} \frac{y+1}{y}$$

$$\leqq \frac{b_i+1}{p(n+1)} \sum_{y=b_i+1}^{n+1} \binom{n+1}{y} p^y (1-p)^{n+1-y}$$

$$= \frac{b_i+1}{p(n+1)} P(\mathrm{Bin}(n+1,p) \geqq b_i+1).$$

Averaging over the distribution of P_i and using Markov's inequality, we have

$$\frac{b_i+1}{n+1} \int_0^1 p^{-1} P(\mathrm{Bin}(n+1,p) \geqq b_i+1) F_i(dp)$$

$$\leqq \int_0^{(b_i+1)/(n+1)} \frac{b_i+1}{p(n+1)} \frac{(n+1)p}{b_i+1} F_i(dp) + \frac{b_i+1}{n+1} \int_{(b_i+1)/(n+1)}^1 p^{-1} F_i(dp)$$

$$= P(P_i \leqq (b_i+1)/(n+1)) + \frac{b_i+1}{n+1} \int_{(b_i+1)/(n+1)}^1 p^{-1} F_i(dp).$$

This completes the proof of the theorem. □

We are particularly interested in the case where the distribution μ of $(P_1, \ldots, P_d)$ is given by (10). The following corollary gives the rate of convergence in this case.

Corollary 1. *Suppose that* $d \geqq 2$ *and that* $(P_1, \ldots, P_d)$ *has density given by* (10). *Define*

$$r(b,n;\theta) = \begin{cases} b/n & \theta \geqq 1\,; \\ b/n^{\theta} & \theta < 1,\ b < n^{\theta/(2-\theta)}\,; \\ (b/n)^{\theta/2} & \theta < 1,\ b \geqq n^{\theta/(2-\theta)}\,. \end{cases} \tag{19}$$

Then if $b_i = o(n)$ *as* $n \to \infty$

$$\tau_i(b_i) = O(r(b_i, n; \theta_i))$$

for $i = 1, \ldots, d$, *and if* $b^* \equiv \max_{1 \leqq i \leqq d} b_i = o(n)$, *and* $\theta \equiv \min_{1 \leqq i \leqq d} \theta_i > 0$,

$$d^*_{b_1, \ldots, b_d} = O(r(b^*, n; \theta))\,.$$

Proof. Choose a particular value of i, and, for typographical simplicity, write $b = b_i$. Assume that $b = o(n)$. Define $\theta_i' = \sum_{j \neq i} \theta_j$, and note that from (10) the marginal density of P_i is bounded above by a constant times a beta density $B(\theta_i, \theta_i')$ where

$$B(\alpha, \beta)(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1\,.$$

Therefore there exists a constant $C \in (0, \infty)$ such that

$$\begin{aligned} P(Y_{in} < b) &= \int_0^1 P(Y_{in} < b \mid P_i = p) F_i(dp) \\ &\leqq C \int_0^1 \sum_{j=0}^{b-1} \binom{n}{j} p^j (1-p)^{n-j} p^{\theta_i - 1} (1-p)^{\theta_i' - 1}\, dp \\ &\leqq C \sum_{j=0}^{b-1} \binom{n}{j} \int_0^1 p^{j+\theta_i - 1} (1-p)^{n-j-1}\, dp\,. \end{aligned} \tag{20}$$

In case $\theta_i \geqq 1$, the right hand side of (20) is

$$\begin{aligned} &\leqq C \sum_{j=0}^{b-1} \binom{n}{j} \int_0^1 p^j (1-p)^{n-j-1}\, dp \\ &= C \sum_{j=0}^{b-1} \frac{1}{n-j} \\ &\leqq C \frac{b}{n-b+1} \\ &= C \frac{b}{n} \left(1 - \frac{b-1}{n}\right)^{-1} = O(b/n)\,. \end{aligned}$$

On the other hand if $\theta_i < 1$, then the right hand side of (20) is

$$= C \sum_{j=0}^{b-1} \binom{n}{j} \frac{\Gamma(n-j)\Gamma(j+\theta_i)}{\Gamma(n+\theta_i)}$$

$$= C \sum_{j=0}^{b-1} \frac{\Gamma(n+1)}{\Gamma(n+\theta_i)} \frac{\Gamma(j+\theta_i)}{\Gamma(j+1)} \frac{1}{n-j}$$

Using the fact that $\lim_{n\to\infty} n^{y-x}\Gamma(n+x)/\Gamma(n+y) = 1$, we see that there exists another constant C' such that for n sufficiently large

$$\begin{aligned} \boldsymbol{P}(Y_{in} < b) &\leqq C' \sum_{j=0}^{b-1} n^{1-\theta_i} \frac{\Gamma(j+\theta_i)}{\Gamma(j+1)} \frac{1}{n-j} \\ &\leqq C' \frac{1}{n^{\theta_i}} \left(1 - \frac{b-1}{n}\right)^{-1} \sum_{j=0}^{b-1} \frac{\Gamma(j+\theta_i)}{\Gamma(j+1)} \\ &\leqq C' \frac{b}{n^{\theta_i}} \left(1 - \frac{b-1}{n}\right)^{-1} = O(b/n^{\theta_i}) . \end{aligned}$$

We have established that if $\theta_i \geqq 1$ then $\boldsymbol{P}(Y_{in} < b) = O(b/n)$, while if $\theta_i < 1$ then $\boldsymbol{P}(Y_{in} < b) = O(b/n^{\theta_i})$. However, if b grows faster than n^{θ_i} then the bound on $\boldsymbol{P}(Y_{in} < b)$ does not tend to zero. In this case we need a more subtle large deviation argument to establish a rate of $(b/n)^{\theta_i/2}$. However, this rate does not hold when b grows *too slowly*. This is why we have combined the two rates in (19). Note that when $b = n^{\theta_i/(2-\theta_i)}$ the two rates in (19) are the same.

Now consider the case where $b \geqq n^{\theta_i/(2-\theta_i)}$. Let $a = b/n$ and choose $\alpha \in (a, 1)$. Hoeffding's (1963, Theorem 1) bound for binomial tail probabilities gives

$$\boldsymbol{P}(Y_{in} < b | P_i) \leqq \mathrm{e}^{-2(a-P_i)^2 n}$$

provided $P_i \geqq \alpha$. Therefore

$$\begin{aligned} \boldsymbol{P}(Y_{in} < b) &= \int_0^1 \boldsymbol{P}(Y_{in} < b | P_i = p) F_i(dp) \\ &= \int_0^\alpha \boldsymbol{P}(Y_{in} < b | P_i = p) F_i(dp) + \int_\alpha^1 \boldsymbol{P}(Y_{in} < b | P_i = p) F_i(dp) \\ &\leqq \boldsymbol{P}(P_i \leqq \alpha) + \int_\alpha^1 \mathrm{e}^{-2(a-p)^2 n} F_i(dp) \\ &\leqq C\alpha^{\theta_i}(1-\alpha)^{-1} + \mathrm{e}^{-2(a-\alpha)^2 n}, \end{aligned}$$

the constant C occurring before (20).

Now choose $\alpha = \sqrt{a}$. To establish that $\boldsymbol{P}(Y_{in} < b) = O((b/n)^{\theta_i/2})$, we show that for $\theta_i < 1$

$$\mathrm{e}^{-2(a-\alpha)^2 n} \leqq C\alpha^{\theta_i} \tag{21}$$

for n sufficiently large. To establish (21), we show that

$$2(a-\alpha)^2 n + \theta_i \log \alpha \to \infty \tag{22}$$

as $n \to \infty$. Substituting $\alpha = \sqrt{a}$ into the left side of (22), we see that

$$2(a-\alpha)^2 n + \theta_i \log \alpha = 2a(1-\sqrt{a})^2 n + \frac{\theta_i}{2}\log a$$

$$= 2b(1-\sqrt{b/n})^2 - \frac{\theta_i}{2}(\log n - \log b) \to \infty$$

as long as b grows faster than $\log n$. This completes the proof that $P(Y_{in} < b) = O(r(b,n;\theta_i))$.

We now consider the second and third terms on the right of (16). As above,

$$P(P_i \leqq (b+1)/(n+1)) \leqq C\left(1 - \frac{b+1}{n+1}\right)^{-1} \int_0^{(b+1)/(n+1)} x^{\theta_i - 1}\, dx = O((b/n)^{\theta_i}) .$$

Note that if $\theta_i > 1$ *then* $E(P_i^{-1}) < \infty$, and the third term in (16) is at most

$$c(\theta_i)\frac{b+1}{n+1} E(P_i^{-1}) .$$

If $\theta_i \leqq 1$, L'Hopital's rule shows that

$$\lim_{x \to 0} x \int_x^1 p^{-1} F_i(dp) = \lim_{x \to 0} x f_i(x) ,$$

where f_i is the density of P_i. Therefore

$$\left(\frac{b+1}{n+1}\right) \int_{(b+1)/(n+1)}^1 p^{-1} F_i(dp) = O\left(\left(\frac{b}{n}\right)^{\theta_i}\right) . \qquad \square$$

4 Functional central limit theorems

In this section, we use the total variation estimates to provide a functional central limit theorem for the counts of alleles in each of the selective classes. We use the approach of Arratia et al. (1993, 1994). We begin with

Lemma 5. *For any* $1 \leqq b \leqq n$, *and for* $i = 1, \ldots, d$

$$\sum_{j=b+1}^{n} E C_{ij}(n) \leqq 1 + \theta_i \log\left(\frac{n}{b}\right), \tag{23}$$

Proof. By conditioning on the value of Y_{in}, we obtain

$$\sum_{j=b+1}^{n} E C_{ij}(n) = \sum_{y=0}^{n} E\left(\sum_{j=b+1}^{n} C_{ij}(n) \mid Y_{in} = y\right) P(Y_{in} = y) .$$

Since samples of size y cannot contribute to the outer sum if $y \leqq b$, or to the inner sum if $j > y$, the last expression may be written

$$\sum_{j=b+1}^{n} E C_{ij}(n) = \sum_{y=b+1}^{n} E\left(\sum_{j=b+1}^{y} C_{ij}(n) \mid Y_{in} = y\right) P(Y_{in} = y) .$$

Given $Y_{in} = y$, the number of alleles with frequency j has the Ewens distribution with parameter θ_i and sample size y. Hence from the estimate in Lemma 2 we have

$$\begin{aligned}\sum_{j=b+1}^{n} EC_{ij}(n) &\leqq \sum_{y=b+1}^{n}\left(1+\sum_{j=b+1}^{y}\frac{\theta_i}{j}\right)P(Y_{in}=y)\\ &\leqq \left(1+\sum_{j=b+1}^{n}\frac{\theta_i}{j}\right)\sum_{y=b+1}^{n}P(Y_{in}=y)\\ &= \left(1+\sum_{j=b+1}^{n}\frac{\theta_i}{j}\right)P(Y_{in}>b)\\ &\leqq 1+\theta_i\log\left(\frac{n}{b}\right).\end{aligned}$$

We exploit Lemma 5 in the following lemma. □

Lemma 6. *For $n \geqq 1$, there exists a coupling of $\{C_{ij}(n), 1 \leqq i \leqq d\}$ and $\{Z_{ij}, 1 \leqq i \leqq d\}$ for $j = 1, \ldots, n$ such that, if*

$$R_{in} \equiv \sum_{j=1}^{n}\frac{|C_{ij}(n) - Z_{ij}|}{\sqrt{\theta_i \log n}},$$

then for $i = 1, \ldots, d$

$$E(R_{in} \wedge 1) = O\left(\frac{\log\log n}{\sqrt{\log n}} \vee \tau\left(\left\lfloor \frac{n}{\log n}\right\rfloor\right)\right)$$

where

$$\tau(b) \equiv \sum_{i=1}^{d}\tau_i(b). \tag{24}$$

Proof. For any $1 \leqq b \leqq n$, and any $i = 1, \ldots, d$

$$\begin{aligned}E(R_{in} \wedge 1) &= E[(R_{in} \wedge 1)1\{(C_{i1}(n), \ldots, C_{ib}(n)) = (Z_{i1}, \ldots, Z_{ib})\}]\\ &\quad + E[(R_{in} \wedge 1)1\{(C_{i1}(n), \ldots, C_{ib}(n)) \neq (Z_{i1}, \ldots, Z_{ib})\}]\\ &\leqq P[(C_{i1}(n), \ldots, C_{ib}(n)) \neq (Z_{i1}, \ldots, Z_{ib})]\\ &\quad + (\theta_i \log n)^{-1/2}\sum_{j=b+1}^{n}(E(C_{ij}(n)) + E(Z_{ij}))\end{aligned} \tag{25}$$

If we take $b = \lfloor n/\log n \rfloor$ then by Theorem 1 and (4), the first term on the right of (25) is bounded above by $\tau(\lfloor n/\log n \rfloor)$. It follows from Lemma 5 that the second term in (25) is $O(\log\log n/\sqrt{\log n})$. □

For $i = 1, \ldots, d$ define random elements B_{in} of $D_R[0,1]$ by

$$B_{in}(t) = (\theta_i \log n)^{-1/2}\left(\sum_{j=1}^{\lfloor n^t \rfloor} C_{ij}(n) - \theta_i t \log n\right) \tag{26}$$

Theorem 2. *It is possible to construct* $\boldsymbol{B}_n \equiv (B_{1n}, \ldots, B_{dn})$ *and a d-dimensional standard Brownian motion* $\boldsymbol{B} \equiv (B_1, \ldots, B_d)$ *on the same probability space in such a way that*

$$E\left\{\sup_{0 \leqq t \leqq 1} \|\boldsymbol{B}_n(t) - \boldsymbol{B}(t)\| \wedge 1\right\} = O\left(\frac{\log\log n}{\sqrt{\log n}} \vee \tau\left(\left\lfloor \frac{n}{\log n} \right\rfloor\right)\right).$$

Proof. Let $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_d$ be d independent Poisson processes constructed to satisfy

$$\mathscr{P}_i(s_i(\lfloor n^t \rfloor)) = \sum_{j=1}^{\lfloor n^t \rfloor} Z_{ij}$$

where

$$s_i(j) = \theta_i \sum_{r=1}^{j} \frac{1}{r}.$$

From the inequality $\log(1+j) \leqq \sum_{r=1}^{j} 1/r \leqq 1 + \log j$, valid for any integer $j \geqq 1$, it follows that

$$\sup_{0 \leqq t \leqq 1} |\theta_i t \log n - s_i(\lfloor n^t \rfloor)| \leqq \theta_i \tag{27}$$

Using a result of Kurtz (1978), we may construct d independent standard Brownian motions $\tilde{\boldsymbol{B}} \equiv (\tilde{B}_1, \ldots, \tilde{B}_d)$ in such a way that for $i = 1, \ldots, d$

$$\sup_{t \geqq 0} \frac{|\mathscr{P}_i(t) - t - \tilde{B}_i(t)|}{\log(t \vee 2)} = K_i < \infty$$

where $Ee^{\lambda K_i} < \infty$ for some $\lambda > 0$; in particular, $E(K_i) < \infty$. With this construction

$$|\mathscr{P}_i(s_i(\lfloor n^t \rfloor)) - s_i(\lfloor n^t \rfloor) - \tilde{B}_i(s_i(\lfloor n^t \rfloor))| \leqq K_i(\log(2 + s_i(n))),$$
$$0 \leqq t \leqq 1. \tag{28}$$

By the triangle inequality

$$\begin{aligned} |\sqrt{\theta_i \log n} B_{in}(t) - \tilde{B}_i(\theta_i t \log n)| \leqq\ & |\mathscr{P}(s_i(\lfloor n^t \rfloor)) - s_i(\lfloor n^t \rfloor) \\ & - \tilde{B}_i(s_i(\lfloor n^t \rfloor))| + \left|\sum_{j=1}^{\lfloor n^t \rfloor} (C_{ij} - Z_{ij})\right| \\ & + |s_i(\lfloor n^t \rfloor) - \theta_i t \log n| \\ & + |\tilde{B}_i(s_i(\lfloor n^t \rfloor)) - \tilde{B}_i(\theta_i t \log n)|. \end{aligned}$$

For $i = 1, \ldots, d$, define

$$B_i(t) = \frac{\tilde{B}_i(\theta_i t \log n)}{\sqrt{\theta_i \log n}}.$$

It follows from (28) and (27) that

$$\sup_{0 \le t \le 1} |B_{in}(t) - B_i(t)| \le \frac{K_i(\log(2 + \theta_i + \theta_i \log n))}{\sqrt{\theta_i \log n}} + R_{in}$$
$$+ \frac{\theta_i}{\sqrt{\theta_i \log n}} + \sup_{0 \le t \le 1} \left| \frac{\tilde{B}_i(s_i(\lfloor n^t \rfloor)) - \tilde{B}_i(\theta_i t \log n)}{\sqrt{\theta_i \log n}} \right| \quad (29)$$

We now use Lemma 1.2.1 from Csörgő and Révész (1981) to see that for any fixed c

$$E\left\{ \sup_{\substack{0 < u, v \le \theta_i \log n + c \\ |u - v| \le c}} |\tilde{B}_i(u) - \tilde{B}_i(v)| \right\} = O\left(\sqrt{\log \log n} \right). \quad (30)$$

It follows from (30), (29) and Lemma 6 that

$$E\left\{ \sup_{0 \le t \le 1} |B_{in}(t) - B_i(t)| \wedge 1 \right\} = O\left(\frac{\log \log n}{\sqrt{\log n}} \vee \tau\left(\left\lfloor \frac{n}{\log n} \right\rfloor \right) \right).$$

Finally we note that

$$\sup_{0 \le t \le 1} \|\boldsymbol{B}_n(t) - \boldsymbol{B}(t)\| \wedge 1 \equiv \left(\sup_{0 \le t \le 1} \sum_{i=1}^{d} |B_{in}(t) - B_i(t)| \right) \wedge 1$$
$$\le \sum_{i=1}^{d} \left(\sup_{0 \le t \le 1} |B_{in}(t) - B_i(t)| \wedge 1 \right). \quad \square$$

The following corollary is an immediate consequence of Theorem 2 and Corollary 1.

Corollary 2. *If* $(P_1, P_2, \ldots, P_d)$ *has density given by* (10) *with* $\theta = \min_{1 \le i \le d} \theta_i > 0$, *then*

$$E\left(\sup_{0 \le t \le 1} \|\boldsymbol{B}_n(t) - \boldsymbol{B}(t)\| \wedge 1 \right) = O\left(\frac{\log \log n}{\sqrt{\log n}} \vee \frac{1}{(\log n)^{\theta/2}} \right)$$

4.1 The approximate distribution of rare alleles

In this section, we collect together the analogous results for the observable allele counts $C_j(n)$ defined in (12). Defining $\boldsymbol{C}(n) = (C_1(n), \ldots, C_n(n), 0, \ldots)$, the analog of Lemma 3 is

Corollary 3. *As* $n \to \infty$,

$$\boldsymbol{C}(n) \Rightarrow \boldsymbol{Z}, \quad (31)$$

where $\boldsymbol{Z} \equiv (Z_1, Z_2, \ldots)$ *is a Poisson processes on* N *with means given by*

$$EZ_j = \frac{\theta_T}{j}. \quad (32)$$

To estimate the rate of convergence, define $C_b(n) \equiv (C_1(n), \ldots, C_b(n))$, and $Z_b \equiv (Z_1, \ldots, Z_b)$. The analog of Lemma 1 is

Lemma 7. *Let* $d_b(n) \equiv d_{TV}(C_b(n), Z_b)$. *Then*

$$d_b(n) = O(\tau(b))$$

Proof. Since

$$d_{TV}\left(\sum_{i=1}^{d} C_{ib}(n), \sum_{i=1}^{d} Z_{ib}\right) \leqq d^*_{b,\ldots,b}(n)\,,$$

the result follows from Theorem 1. □

Remark. Suppose that $C^*_n(n) = (C^*_1(n), \ldots, C^*_n(n))$ is distributed according to the Ewens sampling formula (1) with parameter θ_T, and $C(n) = (C_1(n), \ldots, C_n(n))$ are the observable allele counts for the model with selection. It follows from Lemma 1 and Lemma 7 that

$$d_{TV}(C^*_b(n), C_b(n)) = O(\tau(b))\,.$$

Thus the counts of rare allele are approximately distributed according to the Ewens sampling formula with parameter θ_T, regardless of the values of the selection parameters. For the model defined by (10), the rate $\tau(b)$ is $O(b/n)$ if $\theta_T \geqq 1$.

The functional central limit theorem for the allele counts follows from Theorem 2. Define an element W_n of $D_R[0,1]$ by

$$W_n(t) = \frac{\sum_{j=1}^{\lfloor nt \rfloor} C_j(n) - \theta_T t \log n}{\sqrt{\theta_T \log n}}, \quad 0 \leqq t \leqq 1\,.$$

Corollary 4. *We can construct* W_n *and a standard Brownian motion* W *on the same probability space in such a way that*

$$E\left\{\sup_{0 \leqq t \leqq 1} |W_n(t) - W(t)| \wedge 1\right\} = O\left(\frac{\log\log n}{\sqrt{\log n}} \vee \tau\left(\left\lfloor \frac{n}{\log n} \right\rfloor\right)\right)$$

where $\tau(\cdot)$ *is defined by* (24).

Proof. Construct $\boldsymbol{B}_n$ and $\boldsymbol{B}$ as in Theorem 2, and define

$$W_n = \sum_{i=1}^{d} \sqrt{\frac{\theta_i}{\theta_T}}\, B_{in}\,,$$

and

$$W = \sum_{i=1}^{d} \sqrt{\frac{\theta_i}{\theta_T}}\, B_i\,.$$

The result now follows from Theorem 2 and the triangle inequality. □

Remark. The functional central limit theorem for the Ewens sampling formula is due to Hansen (1990). This can be established using Poisson approximation methods, as shown by Arratia and Tavaré (1992). The rate in that

setting is $O(\log\log n/\sqrt{\log n})$, as shown by Arratia et al. (1994). In the d-class setting the rate can be much slower, due primarily to the effects of selection.

These results show that θ_T can be estimated consistently from counts at a single locus. In particular, the total number of alleles in the sample is $K(n) = \sum_{j=1}^{n} C_j(n)$. It follows from Lemma 4 that $K(n)$ has asymptotically a normal distribution:

Corollary 5. *As* $n \to \infty$,

$$\frac{K(n) - \theta_T \log n}{\sqrt{\theta_T \log n}} \Rightarrow \mathrm{N}(0,1)\,.$$

Kimura's estimator of θ_T is based on the statistic $K(\lfloor nq \rfloor) = \sum_{j=1}^{\lfloor nq \rfloor} C_j(n)$, where $q > 0$ is a prescribed small number. Lemma 5 shows that

$$E \sum_{j=b+1}^{n} C_j(n) \leqq d + \theta_T \log(n/b)\,, \tag{33}$$

which in turn implies that

$$\frac{K(\lfloor nq \rfloor) - \theta_T \log(nq)}{\sqrt{\theta_T \log(nq)}} \Rightarrow \mathrm{N}(0,1)$$

also. Thus Kimura's estimator of θ_T is also consistent. These two normal laws are essentially the same, because (33) shows that the contribution to the normal law comes from the 'small' counts.

We have shown that the rare allele counts in a class of population genetic models with selection and infinitely-many-alleles mutation structure have approximately independent Poisson distributions. Joyce (1994b) establishes an analogous result, without a rate, for a more general class of models with selection.

Acknowledgements. This research was supported in part by National Science Foundation grants DMS90-05833 and DMS92-07410. The authors thank Richard Arratia and Geoff Watterson for helpful comments on earlier versions of this paper, and two referees for their careful reading of the paper.

5 References

Arratia, R., Barbour, A. D. and Tavaré, S. (1992) Poisson process approximations for the Ewens sampling formula. Ann. Appl. Probab., **2**, 519–535

Arratia, R., Barbour, A. D. and Tavaré, S. (1993) On random polynomials over finite fields. Math. Proc. Camb. Phil. Soc., **114**, 347–368

Arratia, R., Barbour, A. D. and Tavaré, S. (1994) Logarithmic combinatorial structures. Ann. Probab., in preparation

Arratia, R. and Tavaré, S. (1992) Limit theorems for combinatorial structures via discrete process approximations. Rand. Struct. Alg., **3**, 321–345

Barbour, A. D., Holst, L., and Janson, S. (1992) Poisson approximation. Oxford University Press

Csörgő, M. and Révész, P. (1981) Strong approximation in probability and statistics. Academic Press, New York

Ethier, S. N. and Kurtz, T. G. (1987) The infinitely-many-alleles model with selection as a measure-valued diffusion. Lecture Notes in Biomathematics, **70**, 72–86. Springer-Verlag, Berlin

Ethier, S. N. and Kurtz, T. G. (1994) Convergence to Fleming–Viot processes in the weak atomic topology. Stoch. Proc. Applns. 54, 1–27

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

Ewens, W. J. and Li, W. -H. (1980) Frequency spectra of neutral and deleterious alleles in a finite population. J. Math. Biol., **10**, 155–166

Griffiths, R. C. (1983) Allele frequencies with genic selection. J. Math. Biol., **17**, 1–10

Hansen, J. C. (1990) A functional central limit theorem for the Ewens sampling formula. J. Appl. Prob. **27**, 28–43

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc., **58**, 13–30

Joyce, P. (1994a) Likelihood ratios for the infinite alleles model. J. Appl. Prob., 31, 595–605

Joyce, P. (1994b) Robustness of the Ewens sampling formula. J. Appl. Prob., in press

Kimura, M. (1983a) Rare variant alleles in the light of the neutral theory. Mol. Biol. Evol., **1**, 84–93

Kimura, M. (1983b) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kingman, J. F. C. (1980) The mathematics of genetic diversity. CBMS-NSF Regional Conference Series in Applied Mathematics. Volume 34. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania

Kurtz, T. G. (1978) Strong approximation theorems for density dependent Markov chains. Stoch. Proc. Applns., **6**, 223–240

Li, W. -H. (1979) Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. Genetics, **92**, 647–667

Thompson, E. A., Neel, J. V., Smouse, P. E., and Barrantes, R. (1992) Microevolution of the Chibcha-speaking peoples of lower Central America: rare genes in an Amerindian complex. Am. J. Hum. Genet., **51**, 609–626

Watterson, G. A. (1978) The homozygosity test of neutrality. Genetics, **88**, 405–417

Watterson, G. A. (1987) Estimating the proportion of neutral mutants. Genet. Res. Camb., **51**, 155–163

Watterson, G. A. (1993) Estimating the proportion of neutral mutations. New Zealand J. Botany, **31**, 297–306

Wright, S. (1949) Adaptation and selection. In Genetics, Paleontology and Evolution, G. L. Jepson, G. G. Simpson and E. Mayr, eds., pp. 365–389. Princeton University Press, Princeton