The Annals of Applied Probability
2002, Vol. 12, No. 2, 637–663

APPROXIMATE EWENS FORMULAE FOR SYMMETRIC OVERDOMINANCE SELECTION[1]

BY MARK N. GROTE AND TERENCE P. SPEED

University of California, Davis and University of California, Berkeley

We derive a family of approximate sampling distributions for the symmetric overdominance model of population genetics. The distributions are selective versions of the Ewens Sampling Formula, which gives sample likelihoods under a model of neutral evolution. We draw on basic results for the general selection model of Ethier and Kurtz, and use mathematical tools well-suited for calculating expectations of symmetric functions of Poisson–Dirichlet atoms. We conclude by briefly examining a Human Leukocyte Antigen data set, in light of a distribution conditional on the number of sample atoms.

1. Introduction.

1.1. *Ewens distributions and infinite alleles models.* Ewens distributions arise naturally in a number of sampling problems in the biological and physical sciences [see, e.g., Johnson, Kotz and Balakrishnan (1997)]. In population genetics, Ewens distributions have been used to describe samples at genetic loci which follow the "infinitely-many neutral alleles" model [see Kimura and Crow (1964), Ewens (1972, 1990)]. The basic features of this model are: (i) alleles (distinct versions of a gene) are equivalent with respect to natural selection, and (ii) mutation, which occurs in any gene copy with probability u each generation, transforms the copy into a completely new allele.

In a common formulation, a random sample of n genes from a population of size N is described by the "frequency spectrum" $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, where a_i is the number of alleles in the sample represented exactly i times, $\sum_{i=1}^{n} i a_i = n$, and $k = \sum_{i=1}^{n} a_i$ is the total number of alleles in the sample. If A_n and K_n are respectively the random variables for the frequency spectrum and number of alleles, under the stationary neutral infinite alleles model

$$\Pr(A_n = \mathbf{a},\ K_n = k) \to \frac{n!}{a_1! \cdots a_n!\, 1^{a_1} \cdots n^{a_n}} \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \theta^k \tag{1.1}$$

as $N \to \infty$, where $\lim_{N\to\infty, u\to 0} 4Nu = \theta$ [Ewens (1972, 1990)]. The right hand side of (1.1) is known as the Ewens Sampling Formula (ESF). Although the value

Received September 2000; revised July 2001.

[1]Supported by NIH Grants GM-35326 and 5 T32 GM-07127.

AMS 2000 subject classifications. Primary 92D15; secondary 62F10.

Key words and phrases. Ewens Sampling Formula, Poisson–Dirichlet distribution, general selection model, size-biased permutation, symmetric overdominance selection, Human Leukocyte Antigen (HLA) loci.

of K_n is determined entirely by $(a_1, a_2, \ldots, a_n)$, we choose to make $K_n = k$ explicit for clarity. The ESF can be viewed as a distribution on permutation cycles of $\mathbb{N}_n = \{1, 2, \ldots, n\}$ [Ewens (1990)], but the ESF has an equivalent formulation as a distribution on partitions of the integer n [Pitman (1995)]. In other biological or physical settings, the equivalence classes in $\mathbb{N}_n$ determined above by allelic identity may instead consist of individuals of the same species, or objects of the same size or type.

Ewens (1972) also gives a version of (1.1) conditional on K_n:

$$\Pr(A_n = \mathbf{a} | K_n = k) \to \frac{n!}{|S_n^k| a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \tag{1.2}$$

as $N \to \infty$, where S_n^k is a Stirling number of the first kind. The lack of dependence on θ in the conditional ESF implies that the statistic K_n is sufficient for θ.

Watterson (1977, 1978) obtained approximate sampling formulae, analogous to (1.1) and (1.2), for two genetic models of natural selection: the symmetric overdominance and deleterious alleles models. Based on the approximations, Watterson showed that the sample homozygosity statistic

$$f_n = \frac{1}{n^2} \sum_{j=1}^{n} j^2 a_j \tag{1.3}$$

can be used to detect departures from the neutral model in the direction of either of these selective alternatives; but for use as selective analogues to (1.1) and (1.2), Watterson's approximations are appropriate only when the values of the selective parameters are very small.

1.2. *The general selection model.* We require a brief description of a class of infinite alleles models with selection, and the statement of a few fundamental results. A complete mathematical treatment is given by Ethier and Kurtz (1987, 1994) and Joyce (1994, 1995). The finite-population model is a Markov process which has as its state-space the set of Borel probability measures on [0, 1] [Ethier and Kurtz (1994)]. The Markov process records the frequencies of exchangeably labeled alleles as follows: in generation t, the individual genes in the population are indexed by $1, 2, \ldots, 2N$ arbitrarily, and the ith gene occupies a position y_i in the interval [0, 1], called its "label." Genes with the same label are said to be of the same allele, and the allele frequency is obtained by dividing the number of genes with the given label by $2N$. In order to form generation $t + 1$ for a neutral locus, $2N$ genes are sampled at random with replacement from the genes of generation t. Each sampled gene may then experience a mutation. A gene with label y chosen at random mutates to a gene with a label in B (a Borel subset of [0, 1]) with probability $u\lambda(B)$, where λ is Lebesgue measure on [0, 1]; hence the probability that this gene has a label in B is

$$P_{2N}(y, B) = (1 - u)\mathbf{1}_B(y) + u\lambda(B). \tag{1.4}$$

The population at generation $t+1$ is described by the labels of the sampled genes after mutation.

The long-run behavior of the Markov process is most conveniently studied by analyzing an appropriate diffusion approximation: the allele frequencies are ordered from largest to smallest at each generation, and in the limit as $N \to \infty$, the frequencies evolve as a diffusion process on the infinite ordered simplex

$$\nabla = \left\{ (x_1, x_2, \ldots) : \sum_{i=1}^{\infty} x_i = 1, x_1 \geq x_2 \geq \cdots \geq 0 \right\}.$$

The stationary distribution μ for the neutral infinite alleles diffusion is the Poisson–Dirichlet(θ) distribution on ∇ [see Kingman (1977), Ethier and Kurtz (1986)].

The selection model introduces a modification of the simple sampling scheme: to form generation $t+1$, N *pairs* of genes are selected with replacement from the genes of generation t. The probability that the ith and jth genes are selected is

$$\frac{w_N(y_i, y_j)}{\sum_{1 \leq l,m \leq 2N} w_N(y_l, y_m)}, \tag{1.5}$$

where

$$w_N(x, y) = 1 + \frac{1}{2N}\sigma(x, y) \geq 0 \tag{1.6}$$

is a symmetric function giving the relative fitness of an individual having genotype (x, y), with $\sigma(x, y)$ continuous $\lambda \times \lambda$-a.e. and $\sigma(x, x)$ continuous λ-a.e. The N pairs chosen to form generation $t+1$ are then split into $2N$ individual genes, and each gene may experience a mutation as in (1.4). The population at generation $t+1$ is then described by the labels of the sampled genes after mutation.

A model of genetic selection with the infinite alleles mutation scheme (1.4) and a fitness function satisfying (1.5) and (1.6) will be called a "general selection model." Fundamental results for the diffusion process on ∇ corresponding to the general selection model have been obtained by Ethier and Kurtz (1987, 1994). The "genic selection" model, which generalizes the deleterious alleles model examined by Watterson (1978), has been fairly extensively analyzed in the context of this general theory [see Ethier and Kurtz (1994), Joyce (1994, 1995), Joyce and Tavaré (1995)]. The symmetric overdominance model is a general selection model with

$$\sigma(x, y) = \sigma \mathbf{1}_{(x \neq y)}(x, y), \qquad \sigma \geq 0. \tag{1.7}$$

In the standard population-genetics formulation, $\lim_{N \to \infty, s \to 0} 2Ns = \sigma$, where $1+s$ is the fitness of heterozygotes, relative to homozygotes, all of which have fitness 1.

1.3. *The Radon–Nikodym derivative $d\nu/d\mu$ and sample probabilities.* In the following, $\mathbf{X} = (X_1, X_2, \ldots)$ will denote a random vector in ∇, and $\mathbf{x} = (x_1, x_2, \ldots)$ a particular version of $\mathbf{X}$. Let μ be the Poisson–Dirichlet(θ) distribution on ∇, and let ν be the stationary distribution for the diffusion process on ∇ obtained as $N \to \infty$ in the general selection model. Ethier and Kurtz (1994) showed that ν is absolutely continuous with respect to μ, and provided a general formula for the Radon–Nikodym derivative $d\nu/d\mu$. We use a slightly simplified expression due to Joyce (1994):

$$\frac{d\nu}{d\mu}(x_1, x_2, \ldots) = \frac{E[\exp\{\sum_{i,j}^{\infty} \sigma(U_i, U_j) x_i x_j\}]}{E[\exp\{\sum_{i,j}^{\infty} \sigma(U_i, U_j) X_i X_j\}]},$$

where $\mathbf{X} = (X_1, X_2, \ldots)$ is distributed according to μ, and $U_1, U_2, \ldots$ are i.i.d. uniform random variables on $(0, 1)$ independent of $(X_1, X_2, \ldots)$. For the symmetric overdominance model, the numerator of $d\nu/d\mu$ is

$$E\left[\exp\left\{\sum_{i=1}^{\infty}\sum_{j<i} 2\sigma \mathbf{1}_{(U_i \neq U_j)}(U_i, U_j) x_i x_j\right\}\right];$$

as the random variable

$$Z_{ij} = 2\sigma \mathbf{1}_{(U_i \neq U_j)}(U_i, U_j) x_i x_j$$

equals $2\sigma x_i x_j$ on a set of probability 1, the expectation above is simply

$$\exp\left\{\sum_{i=1}^{\infty}\sum_{j<i} 2\sigma x_i x_j\right\} = \exp\{\sigma(1 - f)\},$$

where $f = \sum_{i=1}^{\infty} x_i^2$ is the realized (non-random) population homozygosity. Applying similar reasoning to the denominator of $d\nu/d\mu$, one obtains

$$\frac{d\nu}{d\mu}(x_1, x_2, \ldots) = \frac{\exp\{\sigma - \sigma f\}}{E[\exp\{\sigma - \sigma \mathcal{F}\}]} = \frac{\exp\{-\sigma f\}}{E[\exp\{-\sigma \mathcal{F}\}]} \tag{1.8}$$

for the symmetric overdominance model, where $\mathcal{F} = \sum_{i=1}^{\infty} X_i^2$ is the (random) population homozygosity and the expectation is with respect to Poisson–Dirichlet(θ).

When h is a bounded, measurable function on ∇ and $g(\mathbf{X}) = \frac{d\nu}{d\mu}(X_1, X_2, \ldots)$, we may write the "change of measure" equation

$$E_\nu[h(\mathbf{X})] = E_\mu[g(\mathbf{X})h(\mathbf{X})], \tag{1.9}$$

where E_ν and E_μ are expectations with respect to ν and μ, respectively [see Billingsley (1986), Theorem 16.10, also Griffiths (1983)]. In the population genetics context, Joyce (1994) has shown that sample likelihoods and likelihood

ratios may be written as special cases of (1.9): the probability of observing a particular sample ($A_n = \mathbf{a},\ K_n = k$) under a general selection model is

$$\begin{aligned} P_\nu(A_n = \mathbf{a}, K_n = k) &= E_\nu[\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})] \\ &= E_\mu[g(\mathbf{X}) \Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})], \end{aligned} \tag{1.10}$$

and the sample likelihood ratio is

$$\begin{aligned} \Lambda &= \frac{E_\nu[\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}{E_\mu[\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]} \\ &= \frac{E_\mu[g(\mathbf{X}) \Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}{E_\mu[\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}, \end{aligned} \tag{1.11}$$

where the right hand side of (1.11) is by definition $E_\mu[g(\mathbf{X})|A_n = \mathbf{a}, K_n = k]$. From (1.11), a second formula for sample probabilities under the general selection model may be deduced:

$$\begin{aligned} P_\nu(A_n = \mathbf{a}, K_n = k) &= P_\mu(A_n = \mathbf{a}, K_n = k) \\ &\quad \times E_\mu[g(\mathbf{X})|A_n = \mathbf{a}, K_n = k], \end{aligned} \tag{1.12}$$

where $P_\mu(A_n = \mathbf{a}, K_n = k)$ is given by the ESF on the right hand side of (1.1).

1.4. *Main results and synopsis.* We wish to obtain explicit sampling formulae for the symmetric overdominance model, suitable for likelihood-based data analysis. In Section 2, we describe a rather general weak-convergence approach which can be used when the Radon–Nikodym derivative g is bounded and continuous. This approach formalizes and extends methods of Watterson (1977, 1978) based on limits of finite-alleles models. The weak-convergence calculation yields the approximate formula

$$\begin{aligned} &P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k) \\ &\quad \approx \gamma^{-1} \frac{n!}{a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \theta^k \exp\left\{-\sigma \left(\frac{n}{n+\theta}\right)^2 f_n\right\}, \end{aligned} \tag{1.13}$$

where $\gamma = \gamma(\theta, \sigma)$ is the denominator of the Radon–Nikodym derivative (1.8) and f_n is the sample homozygosity statistic (1.3).

In Section 3, we use the "size-biased" permutation $(\tilde{X}_1, \tilde{X}_2, \ldots)$ of the Poisson–Dirichlet atoms [see Pitman (1995, 1996)] to obtain lower and upper bounds on the sample probability:

$$\begin{aligned} &P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k) \\ &\quad \ge \gamma^{-1} \frac{n!}{a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \theta^k \\ &\qquad \times \exp\left\{-\sigma n^2 \frac{f_n + 1/n + \theta/n^2}{(\theta+n+1)(\theta+n)}\right\} \\ &\quad \equiv p^* \end{aligned} \tag{1.14}$$

and

$$P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k) < \left(1 + \frac{\sigma^2 e^{\sigma}}{2} \frac{n^4}{((n+\theta+1)(n+\theta))^2}\left(\frac{4}{n} + O(n^{-2})\right)\right) p^*. \tag{1.15}$$

In Section 4, we use the size-biased permutation and Monte Carlo averaging to obtain a numerical approximation for the expectation in (1.12), and give a "rejection rule" algorithm which generates i.i.d. samples $(a_1, \ldots, a_n)$ under the symmetric overdominance model.

In Section 5, we consider conditional sampling formulae, analogous to expression (1.2). In the conditional setting, the weak-convergence approximation (1.13) yields

$$P_{\theta,\sigma}(A_n = \mathbf{a} | K_n = k) \approx \frac{n!}{|S_n^k| a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \frac{\exp\{-\sigma (n/(n+\theta))^2 f_n\}}{E[\exp\{-\sigma (n/(n+\theta))^2 F_n\} | K_n = k]}, \tag{1.16}$$

where F_n is the sample homozygosity viewed as a random variable, and the expectation in the denominator is with respect to the conditional ESF (1.2). The conditional formula (1.16) depends only weakly on θ when $\theta \ll n$. In Section 5, we also examine a Human Leukocyte Antigen (HLA) data set in the context of the overdominance sampling theory.

2. Weak convergence methods.

2.1. *Mathematical preliminaries.* We begin by putting the expectation on the right hand side of (1.10) in a form more suitable for calculation. Conditional on $\mathbf{x}$, the probability of the sample $(A_n = \mathbf{a}, K_n = k)$ is given by the multinomial sampling formula

$$\Phi_{\mathbf{a}}(\mathbf{x}) = C_{\mathbf{a}} \sum_{\mathbf{m}} x_1^{m_1} x_2^{m_2} \cdots,$$

where

$$C_{\mathbf{a}} = \frac{n!}{(1!)^{a_1} \cdots (n!)^{a_n}},$$

and the sum is over all distinct arrays $\mathbf{m} = (m_1, m_2, \ldots)$ consistent with

$$a_j = \#\{i : m_i = j\}, \qquad j = 1, 2, \ldots, n$$

and $\sum_1^n a_j = k$ [see Kingman (1977), Joyce (1994)]. Using

$$\Pr(A_n = \mathbf{a}, K_n = k | \mathbf{x}) = \Phi_{\mathbf{a}}(\mathbf{x}),$$

we write the right hand expectation of (1.10) as

$$E_\mu\big[g(\mathbf{X})\Pr(A_n=\mathbf{a},K_n=k|\mathbf{X})\big]=\int_{\mathbf{x}\in\nabla}g(\mathbf{x})\Phi_{\mathbf{a}}(\mathbf{x})d\mu(\mathbf{x}). \tag{2.1}$$

Watterson (1977, 1978) evaluated expressions similar to that on the right of (2.1) by taking limits of finite-dimensional integrals. We wish to establish a weak-convergence context for Watterson's approach.

Kingman (1975, 1977) showed that the symmetric K-dimensional ordered Dirichlet distribution with parameter $\frac{\theta}{K-1}$ has the Poisson–Dirichlet(θ) distribution as its natural weak limit as $K\to\infty$ (convergence in distribution). The symmetric ordered Dirichlet is a distribution on a set of K-dimensional elements

$$D=\left\{(x_1,x_2,\ldots,x_K):x_1\geq x_2\geq\cdots\geq x_K\geq 0,\sum_{i=1}^{K}x_i=1\right\}$$

with probability density

$$f(x_1,x_2,\ldots,x_K)=K!\frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K}x_1^{\varepsilon-1}\cdots x_K^{\varepsilon-1} \tag{2.2}$$

for $\frac{\theta}{K-1}=\varepsilon$. It will be convenient to imbed the support D of the ordered Dirichlet distribution in ∇; that is,

$$D=\{\mathbf{x}\in\nabla:x_{K+1}=x_{K+2}=\cdots=0\}.$$

Let μ_K be the probability measure on ∇ with density (2.2) on $\mathbf{x}\in D$, and density equal to zero on $\nabla-D$. Since $\mu_K\to\mu$,

$$\int_\nabla h(\mathbf{x})d\mu(\mathbf{x})=\lim_{K\to\infty}\int_\nabla h(\mathbf{x})d\mu_K(\mathbf{x}) \tag{2.3}$$

for any bounded, continuous function $h:\nabla\to\mathbf{R}$ [see Kingman (1977)]. We would like to show that (2.3) holds for $h(\mathbf{x})=g(\mathbf{x})\Phi_{\mathbf{a}}(\mathbf{x})$, with $g=d\nu/d\mu$ bounded and continuous. On the hypothesis that $\Phi_{\mathbf{a}}(\mathbf{x})$ is lower semi-continuous on ∇, Kingman (1977) showed that (2.3) holds for $h(\mathbf{x})=\Phi_{\mathbf{a}}(\mathbf{x})$, using special properties of $\Phi_{\mathbf{a}}$. Indeed, one can show that (2.3) holds for $h(\mathbf{x})=g(\mathbf{x})\Phi_{\mathbf{a}}(\mathbf{x})$ using an argument modelled on Kingman's. However, Paul Joyce has communicated a result that considerably simplifies the development:

PROPOSITION (Joyce). $\Phi_{\mathbf{a}}(\mathbf{x})$ *is continuous on* ∇.

PROOF. For a given $\mathbf{x}\in\nabla$, $\Phi_{\mathbf{a}}(\mathbf{x})$ is a probability distribution on the finite set of samples

$$\boldsymbol{\alpha}=\left\{\mathbf{a}=(a_1,a_2,\ldots,a_n):\sum_{1}^{n}ia_i=n,a_i\geq 0\right\}$$

so that

$$\Phi_{\mathbf{a}}(\mathbf{x}) = 1 - \sum_{\mathbf{b}\in\boldsymbol{\alpha},\mathbf{b}\neq\mathbf{a}} \Phi_{\mathbf{b}}(\mathbf{x}).$$

Lower semi-continuity of $\Phi_{\mathbf{b}}$ on ∇ implies upper semi-continuity of $-\Phi_{\mathbf{b}}$, so the right hand side above, and hence $\Phi_{\mathbf{a}}$, must be upper semi-continuous on ∇. $\Phi_{\mathbf{a}}$ is then both upper and lower semi-continuous, therefore continuous on ∇. $\square$

COROLLARY. *If $g:\nabla\to\mathbb{R}$ is bounded and continuous, then*

$$\int_{\nabla} g(\mathbf{x})\Phi_{\mathbf{a}}(\mathbf{x})d\mu(\mathbf{x}) = \lim_{K\to\infty}\int_{\nabla} g(\mathbf{x})\Phi_{\mathbf{a}}(\mathbf{x})d\mu_K(\mathbf{x}). \tag{2.4}$$

2.2. *Expectation via the K-dimensional ordered Dirichlet distribution.* The integral on the right hand side of (2.4) is equivalent to the finite-dimensional integral

$$\int_D g_K(\mathbf{x})\Phi_{K,\mathbf{a}}(\mathbf{x})d\mu_K(\mathbf{x}),$$

where for $\mathbf{x}\in D$,

$$g_K(\mathbf{x}) = \gamma^{-1}\exp\left\{-\sigma\sum_{i=1}^{K} x_i^2\right\},$$

$$\Phi_{K,\mathbf{a}}(\mathbf{x}) = C_{K,\mathbf{a}}\sum_{\mathbf{m}_K} x_1^{m_1}x_2^{m_2}\cdots x_K^{m_K},$$

$$C_{K,\mathbf{a}} = \frac{n!}{(0!)^{a_0}(1!)^{a_1}\cdots(n!)^{a_n}},$$

and the sum in $\Phi_{K,\mathbf{a}}(\mathbf{x})$ is over all distinct arrays $\mathbf{m}_K = (m_1, m_2, \ldots, m_K)$ consistent with

$$a_j = \#\{i : m_i = j\}, \qquad j = 0, 1, \ldots, n$$

and $\sum_1^n a_j = k$. We assume $K \geq k$ to avoid complications arising from the use of $\Phi_{K,\mathbf{a}}(\mathbf{x})$ for $\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{x})$, and specify $a_0 = K - k$.

The integral is then explicitly

$$\begin{aligned}\int_D & g_K(\mathbf{x})\Phi_{K,\mathbf{a}}(\mathbf{x})d\mu_K(\mathbf{x}) \\ &= \int_D \gamma^{-1}\exp\left\{-\sigma\sum_{1}^{K} x_i^2\right\} C_{K,\mathbf{a}}\sum_{\mathbf{m}_K} x_1^{m_1}\cdots x_K^{m_K} \\ &\quad \times K!\frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K} x_1^{\varepsilon-1}\cdots x_K^{\varepsilon-1}dx_1\cdots dx_{K-1},\end{aligned} \tag{2.5}$$

using the ordered Dirichlet density (2.2) with parameter $\varepsilon = \frac{\theta}{K-1}$. Substituting the Taylor expansion

$$\exp\left\{-\sigma \sum_{1}^{K} x_i^2\right\} = \sum_{l=0}^{\infty} \frac{(-\sigma \sum_{i=1}^{K} x_i^2)^l}{l!},$$

the right hand side of (2.5) takes the form

$$\int_D \sum_{l=0}^{\infty} h_l(\mathbf{x}) d\mathbf{x},$$

with

$$h_l(\mathbf{x}) = \gamma^{-1} \frac{(-\sigma \sum_{i=1}^{K} x_i^2)^l}{l!} \Phi_{K,\mathbf{a}}(\mathbf{x}) K! \frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K} x_1^{\varepsilon-1} \cdots x_K^{\varepsilon-1}.$$

The usual conditions allowing for the exchange of integral and sum can be verified. Making the further substitution

$$\left(\sum_{i=1}^{K} x_i^2\right)^l = \sum_{l_1 \ldots l_K} \frac{l!}{l_1! \cdots l_K!} x_1^{2l_1} \cdots x_K^{2l_K},$$

where the arrays $(l_1, \ldots, l_K)$ satisfy $\sum_1^K l_i = l$, $l_i \geq 0$, integrating and collecting terms, expression (2.5) is found to be

$$\begin{aligned} \int_D g_K(\mathbf{x}) \Phi_{K,\mathbf{a}}(\mathbf{x}) d\mu_K(\mathbf{x}) &= \gamma^{-1} C_{K,\mathbf{a}} \frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K} \sum_{\mathbf{m}_K} \sum_{l=0}^{\infty} \frac{(-\sigma)^l}{l!} \\ &\quad \times \sum_{l_1,\ldots,l_K} \frac{l!}{l_1! \cdots l_K!} \frac{\prod_{i=1}^{K} \Gamma(m_i + 2l_i + \varepsilon)}{\Gamma(n + 2l + K\varepsilon)}. \end{aligned} \tag{2.6}$$

We approximate the gamma functions of (2.6) as

$$\Gamma(m_i + 2l_i + \varepsilon) \approx (m_i + \varepsilon)^{2l_i} \Gamma(m_i + \varepsilon)$$

and

$$\Gamma(n + 2l + K\varepsilon) \approx (n + \varepsilon)^{2l} \Gamma(n + \varepsilon),$$

leaving questions about the accuracy of the resulting expression to later sections. Expression (2.6) is then approximately

$$\begin{aligned} \int_D g_K(\mathbf{x}) \Phi_{K,\mathbf{a}}(\mathbf{x}) d\mu_K(\mathbf{x}) &\approx \gamma^{-1} C_{K,\mathbf{a}} \frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K} \sum_{\mathbf{m}_K} \sum_{l=0}^{\infty} \frac{(-\sigma)^l}{l!} \\ &\quad \times \sum_{l_1,\ldots,l_K} \frac{l!}{l_1! \cdots l_K!} \frac{\prod_{i=1}^{K} (m_i + \varepsilon)^{2l_i} \Gamma(m_i + \varepsilon)}{(n + K\varepsilon)^{2l} \Gamma(n + K\varepsilon)} \end{aligned}$$

$$= \gamma^{-1} C_{K,\mathbf{a}} \frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K \Gamma(n+K\varepsilon)} \sum_{\mathbf{m}_K} \left(\prod_{i=1}^{K} \Gamma(m_i+\varepsilon) \right)$$
$$\times \exp\left\{ \frac{-\sigma}{(n+K\varepsilon)^2} \sum_{i=1}^{K} (m_i+\varepsilon)^2 \right\}.$$

On the set $\mathbf{m}_K$,

$$\prod_{i=1}^{K} \Gamma(m_i+\varepsilon) = \prod_{j=0}^{n} \Gamma(j+\varepsilon)^{a_j},$$
$$\sum_{i=1}^{K} (m_i+\varepsilon)^2 = \sum_{j=0}^{n} a_j (j+\varepsilon)^2,$$

and we can re-write the expression of interest as

$$\gamma^{-1} C_{K,\mathbf{a}} \frac{\Gamma(K\varepsilon)}{[\Gamma(\varepsilon)]^K \Gamma(n+K\varepsilon)} \left(\prod_{j=0}^{n} \Gamma(j+\varepsilon)^{a_j} \right)$$
$$\times \exp\left\{ \frac{-\sigma}{(n+K\varepsilon)^2} \sum_{j=0}^{n} a_j (j+\varepsilon)^2 \right\} \mathcal{M},$$

where

$$\mathcal{M} = \frac{K!}{a_0! \cdots a_n!} = \frac{K!}{(K-k)!\, a_1! \cdots a_n!}$$

is the cardinality of $\mathbf{m}_K$. Having removed the dependence on the variables $m_1, \ldots, m_K$, the limit as $K \to \infty$ can be readily evaluated.

As $K \to \infty$, we have $K\varepsilon \to \theta$, $\varepsilon = \theta/(K-1) \to 0$, and

$$\frac{[\Gamma(\varepsilon)]^{K-k}}{[\Gamma(\varepsilon)]^K} \frac{K!}{(K-k)!} \to \theta^k$$

[Watterson (1976)]. Finally,

$$C_{K,\mathbf{a}} \prod_{j=1}^{n} \Gamma(j+\varepsilon)^{a_j} \to \frac{n!}{1^{a_1} \cdots n^{a_n}}$$

as $K \to \infty$, the $j=0$ term of the product having been used as $[\Gamma(\varepsilon)]^{K-k}$ above. Putting the terms together, we have:

APPROXIMATION 1 (Weak convergence).

$$P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k) = \lim_{K\to\infty} \int_D g_K(\mathbf{x}) \Phi_{K,\mathbf{a}}(\mathbf{x}) d\mu_K(\mathbf{x}) \tag{2.7}$$

$$\approx \gamma^{-1} \frac{n!}{a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \theta^k \times \exp\left\{-\sigma \left(\frac{n}{n+\theta}\right)^2 f_n\right\}.$$

Some care is required in the interpretation of Approximation 1, for although the right hand side of (2.7) is non-negative, it would not in general sum to unity over the sample space $\boldsymbol{\alpha}$. It is convenient to view the right hand side of (2.7) as an un-normalized family which approximates $P_{\theta,\sigma}$. It is worth noting that the approximation comes not from the limiting operation directly, but rather from the need to approximate the finite-dimensional integral (2.6).

3. Size-biased permutation methods.

3.1. *The residual allocation model.* A useful formulation of the conditional expectation in (1.12) arises by considering the "size-biased" permutation $(\tilde{X}_1, \tilde{X}_2, \ldots)$ of the Poisson–Dirichlet atoms $(X_1, X_2, \ldots)$. Here, an explicit description of the $\tilde{X}_i$ leads to analytical approximations of (1.12) that do not rely on limiting arguments. A method for numerical approximation of the conditional expectation also follows.

To form the size-biased permutation $(\tilde{X}_1, \tilde{X}_2, \ldots)$, we construct the sample one gene at a time, beginning with the first gene sampled, and define sample atoms $\tilde{n}_1, \ldots, \tilde{n}_k$ in order of the appearance of distinct allelic types. $\tilde{n}_i > 0$ is the number of representatives of the ith distinct allele to appear in the sample, and the size-biased atom $\tilde{X}_i$ is the element of $(X_1, X_2, \ldots)$ associated with $\tilde{n}_i$. When $K_n = k$, the size-biased atoms $(\tilde{X}_{k+1}, \tilde{X}_{k+2}, \ldots)$ correspond to alleles not observed in the sample. Pitman (1996) has given an explicit representation of the $\tilde{X}_i$ conditional on the sample counts $(\tilde{n}_1, \ldots, \tilde{n}_k)$:

$$\begin{aligned} &\tilde{X}_1 \stackrel{\mathcal{D}}{=} W_1, \\ &\tilde{X}_i \stackrel{\mathcal{D}}{=} (1 - W_1) \cdots (1 - W_{i-1}) W_i, \qquad i \geq 2, \end{aligned} \tag{3.1}$$

where $W_1, W_2, \ldots$ are independent and

$$W_i \sim \begin{cases} \text{beta}\left(\tilde{n}_i, \theta + \sum_{j=i+1}^{k} \tilde{n}_j\right), & i = 1, \ldots, k, \\ \text{beta}(1, \theta), & i > k. \end{cases} \tag{3.2}$$

In the following, we will refer to (3.1) and (3.2) as the residual allocation model or RAM.

The Radon–Nikodym derivative $g(\mathbf{X}) = \gamma^{-1} \exp\{-\sigma \sum_{i=1}^{\infty} X_i^2\}$ is a symmetric function of the Poisson–Dirichlet atoms, so we may write an expression equivalent

to the conditional expectation of (1.12) in terms of the size-biased permutation $(\tilde{X}_1, \tilde{X}_2, \ldots)$:

$$
\begin{aligned}
&E_\mu\left[\gamma^{-1}\exp\left\{-\sigma\sum_{i=1}^{\infty}X_i^2\right\}\Big|A_n=\mathbf{a},\,K_n=k\right]\\
&\qquad=E_{\tilde{\mu}}\left[\gamma^{-1}\exp\left\{-\sigma\sum_{i=1}^{\infty}\tilde{X}_i{}^2\right\}\right],
\end{aligned}
\tag{3.3}
$$

where E_μ is expectation with respect to the Poisson–Dirichlet(θ) distribution, and $E_{\tilde{\mu}}$ is expectation with respect to the distribution of the $\tilde{X}_i$ given by (3.1) and (3.2). In the right hand expectation, the conditioning is incorporated directly into the distribution of the $\tilde{X}_i$.

3.2. *Lower and upper bounds.* We will require some moment expressions. For $r\in\mathbb{Z}^+$, straightforward calculation using (3.1) and (3.2) yields

$$
E_{\tilde{\mu}}[\tilde{X}_i{}^r]=\begin{cases}\dfrac{\Gamma(\theta+n)}{\Gamma(\theta+r+n)}\dfrac{\Gamma(\tilde{n}_i+r)}{\Gamma(\tilde{n}_i)}, & i\le k,\\[2ex] r!\dfrac{\Gamma(\theta+n)}{\Gamma(\theta+r+n)}\left(\dfrac{\theta}{\theta+r}\right)^{i-k}, & i>k.\end{cases}
\tag{3.4}
$$

We will also require moments of the form $E_{\tilde{\mu}}[\tilde{X}_i{}^r\tilde{X}_j{}^q];\ j<i$. This calculation is made easier by setting $\tilde{X}_i^r\tilde{X}_j^q=YZ$, where

$$
\begin{aligned}
Y&=(1-W_1)^{q+r}\cdots(1-W_{j-1})^{q+r},\\
Z&=W_j^q(1-W_j)^r\cdots(1-W_{i-1})^rW_i^r
\end{aligned}
$$

and Y and Z are independent. There are three cases to consider, depending on where i and j lie with respect to k, the number of alleles in the sample. By calculations similar to the above,

$$
E_{\tilde{\mu}}[\tilde{X}_i^r\tilde{X}_j^q]=\begin{cases}\dfrac{\Gamma(\theta+n)}{\Gamma(\theta+q+r+n)}\dfrac{\Gamma(\tilde{n}_j+q)}{\Gamma(\tilde{n}_j)}\dfrac{\Gamma(\tilde{n}_i+r)}{\Gamma(\tilde{n}_i)}, & j<i\le k,\\[2ex] r!\dfrac{\Gamma(\theta+n)}{\Gamma(\theta+q+r+n)}\dfrac{\Gamma(\tilde{n}_j+q)}{\Gamma(\tilde{n}_j)}\left(\dfrac{\theta}{\theta+r}\right)^{i-k}, & j\le k<i,\\[2ex] q!\,r!\dfrac{\Gamma(\theta+n)}{\Gamma(\theta+q+r+n)}\\ \qquad\times\left(\dfrac{\theta}{\theta+q+r}\right)^{j-k}\left(\dfrac{\theta}{\theta+r}\right)^{i-j}, & k<j<i.\end{cases}
\tag{3.5}
$$

Using (3.4) with $r = 2$, one finds

$$E_{\tilde{\mu}}[\mathcal{F}] = E_{\tilde{\mu}}\left[\sum_{i=1}^{\infty} \tilde{X}_i{}^2\right] = \frac{\sum_{i=1}^{k}(\tilde{n}_i^2 + \tilde{n}_i) + 2\sum_{i=k+1}^{\infty}(\theta/(\theta+2))^{i-k}}{(\theta+n+1)(\theta+n)} = n^2\frac{f_n + 1/n + \theta/n^2}{(\theta+n+1)(\theta+n)}. \tag{3.6}$$

Viewed as a function of $\mathcal{F}$, the Radon–Nikodym derivative

$$g(\tilde{\mathbf{X}}) = \gamma^{-1}\exp\left\{-\sigma\sum_{1}^{\infty}\tilde{X}_i^2\right\} = \gamma^{-1}\exp\{-\sigma\mathcal{F}\}$$

is convex in $\mathcal{F} \in [0, 1]$, so we may use (1.12) and (3.6), along with the Jensen inequality to obtain:

APPROXIMATION 2 (Jensen lower bound).

$$\begin{aligned} &P_{\theta,\sigma}(A_n = \mathbf{a},\ K_n = k) \\ &\quad = \gamma^{-1}\frac{n!}{a_1!\cdots a_n!1^{a_1}\cdots n^{a_n}}\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\theta^k \\ &\qquad \times E_{\tilde{\mu}}[\exp\{-\sigma\mathcal{F}\}] \\ &\quad \geq \gamma^{-1}\frac{n!}{a_1!\cdots a_n!1^{a_1}\cdots n^{a_n}}\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\theta^k \\ &\qquad \times \exp\left\{-\sigma n^2\frac{f_n + 1/n + \theta/n^2}{(\theta+n+1)(\theta+n)}\right\}. \end{aligned} \tag{3.7}$$

An upper bound related to (3.7) can be obtained using a Taylor expansion near zero for the centered variable

$$\mathcal{H} = \mathcal{F} - E_{\tilde{\mu}}[\mathcal{F}].$$

For each realization h of $\mathcal{H}$, by Taylor's Theorem there is a number c between h and zero such that

$$e^{-\sigma h} = 1 - \sigma h + \frac{\sigma^2}{2}e^{-\sigma c}h^2.$$

The value $c = -1$ makes $e^{-\sigma c}$ maximal within the range of $\mathcal{H}$, giving the global upper bound

$$e^{-\sigma h} \leq 1 - \sigma h + \frac{\sigma^2}{2}e^{\sigma}h^2, \qquad h \in [-1, 1].$$

Elementary properties of the expectation then give

$$\begin{aligned} E_{\tilde{\mu}}[e^{-\sigma\mathcal{H}}] &\leq 1 - \sigma E_{\tilde{\mu}}[\mathcal{H}] + \frac{\sigma^2}{2}e^{\sigma}E_{\tilde{\mu}}[\mathcal{H}^2] \\ &= 1 + \frac{\sigma^2}{2}e^{\sigma}E_{\tilde{\mu}}[\mathcal{H}^2]. \end{aligned}$$

As

$$e^{-\sigma\mathcal{F}} = e^{-\sigma\mathcal{H}} e^{-\sigma E_{\tilde{\mu}}[\mathcal{F}]},$$

we have an upper bound for $E_{\tilde{\mu}}[e^{-\sigma\mathcal{F}}]$ of the form

$$E_{\tilde{\mu}}[e^{-\sigma\mathcal{F}}] \leq \left(1 + \frac{\sigma^2}{2} e^{\sigma} E_{\tilde{\mu}}[\mathcal{H}^2]\right) v \tag{3.8}$$

where $v = \exp\{-\sigma\ E_{\tilde{\mu}}[\mathcal{F}]\}$ and $E_{\tilde{\mu}}[\mathcal{F}]$ is given by (3.6). As $E_{\tilde{\mu}}[\mathcal{H}^2] \equiv \mathrm{Var}_{\tilde{\mu}}[\mathcal{F}]$, the previous moment expressions can be used to make (3.8) explicit. After some calculation, conveniently writing polynomial functions of $(\tilde{n}_1, \ldots, \tilde{n}_k)$ in terms of f_n, one obtains

$$\begin{aligned} E_{\tilde{\mu}}[\mathcal{F}^2] = \frac{\Gamma(\theta+n)}{\Gamma(\theta+4+n)} \Bigg(& n^4 {f_n}^2 + 2n^3 f_n + 4\sum_{i=1}^{k} \tilde{n}_i^3 \\ & + n^2(1 + 10 f_n + 2\theta f_n) + 2n(3+\theta) + \theta^2 + 6\theta \Bigg). \end{aligned} \tag{3.9}$$

The substitution of

$$\big((n+\theta+1)(n+\theta)\big)^{-2}$$

for $\Gamma(\theta+n)/\Gamma(\theta+4+n)$ in (3.9) leads to a common factor for $E_{\tilde{\mu}}[\mathcal{F}^2]$ and $(E_{\tilde{\mu}}[\mathcal{F}])^2$, and to a simpler expression, but also makes the inequality strict:

$$\begin{aligned} E_{\tilde{\mu}}[\mathcal{H}^2] &= E_{\tilde{\mu}}[\mathcal{F}^2] - (E_{\tilde{\mu}}[\mathcal{F}])^2 \\ &< \frac{n^4}{((n+\theta+1)(n+\theta))^2} \left(\frac{4}{n^4} \sum_{i=1}^{k} \tilde{n}_i^3 + \frac{10 f_n}{n^2} + \frac{6}{n^3} + \frac{6\theta}{n^4} \right) \\ &\leq \frac{n^4}{((n+\theta+1)(n+\theta))^2} \left(\frac{4}{n} + \frac{10 f_n}{n^2} + \frac{6}{n^3} + \frac{6\theta}{n^4} \right). \end{aligned} \tag{3.10}$$

We then have:

APPROXIMATION 3 (Upper bound).

$$\begin{aligned} & P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k) \\ & < \left(1 + \frac{\sigma^2\, e^{\sigma}}{2} \frac{n^4}{((n+\theta+1)(n+\theta))^2} \left(\frac{4}{n} + \frac{10 f_n}{n^2} + \frac{6}{n^3} + \frac{6\theta}{n^4} \right) \right) p^*, \end{aligned} \tag{3.11}$$

where p^* is given by the right hand side of (3.7). Because $0 \leq f_n \leq 1$, the upper and lower bounds can be made arbitrarily close by taking $n \gg \theta$ and $\log n \gg \sigma$; if instead σ is large compared to n, the e^{σ} term of (3.11) may make the upper bound rather crude. Straightforward comparison shows that the weak-convergence approximation (2.7) always lies above the lower bound (3.7), although these approximations can be made arbitrarily close by taking $n \gg \theta$.

Alternative approximations similar to 1–3 can be obtained by truncating the sum on the right hand side of (3.3), using an expression of the form

$$\sum_{i=1}^{\infty} \tilde{X}_i^2 \approx \sum_{i=1}^{s} \tilde{X}_i^2 + \tau, \tag{3.12}$$

with $s \geq k$ and

$$\tau = \tau(s) = E_{\tilde{\mu}}\left[\sum_{i=s+1}^{\infty} \tilde{X}_i^2\right] = \frac{\theta(\theta/(\theta+2))^{s-k}}{(\theta+n+1)(\theta+n)}. \tag{3.13}$$

The contribution of the truncation constant can be made as small as desired by increasing s and taking more size-biased atoms $\tilde{X}_i$ into the sum. Our main interest in truncation is its convenience in a fully numerical approach.

4. Numerical methods.

4.1. *Approximation of the basic functional.* Using the truncation scheme in (3.12) and (3.13) along with Monte Carlo integration, it is possible to obtain a numerical approximation of $E_{\tilde{\mu}}[\exp\{-\sigma \sum_{i=1}^{\infty} \tilde{X}_i^2\}]$ to a desired level of accuracy. For a given sample $(\tilde{n}_1, \ldots, \tilde{n}_k)$, pseudo-random beta variates conforming to (3.2) are combined using the RAM to obtain realizations of the array $(\tilde{X}_1, \ldots, \tilde{X}_s)$. For approximation of the expectation, the random variable of interest is $S_m = \sum_{i=1}^{m} Z_i$, where

$$Z_i = \exp\left\{-\sigma\left(\sum_{j=1}^{s} \tilde{X}_j^2 + \tau(s)\right)\right\}, \tag{4.1}$$

and the $Z_1, \ldots, Z_m$ are obtained from independent realizations of $(\tilde{X}_1, \ldots, \tilde{X}_s)$. The Monte Carlo estimate S_m/m converges in probability to $E_{\tilde{\mu}}[\exp\{-\sigma(\sum_{j=1}^{s} \tilde{X}_j^2 + \tau(s))\}]$, with a standard deviation proportional to $1/\sqrt{m}$ [see Ripley (1987)]. Two points are worth noting: first, although the constant γ which appears in (3.3) could also be approximated, it is more convenient to leave γ undetermined and use methods tailored to un-normalized distributions. Second, S_m/m estimates the integral involving the truncated sum $\sum_{j=1}^{s} \tilde{X}_j^2 + \tau(s)$ rather than the full expectation (3.3); however, by choosing s appropriately one can, in principle, obtain close agreement between the truncated and full expressions.

We use a method for choosing s that makes the truncated term $\sum_{i=s+1}^{\infty} \tilde{X}_i^2$ close to its expectation $\tau(s)$ with high probability. By the Chebychev inequality, for a given $\varepsilon > 0$,

$$\Pr\left(\left|\sum_{i=s+1}^{\infty} \tilde{X}_i^2 - \tau\right| \geq \varepsilon\right) \leq \frac{v}{\varepsilon^2},$$

where $v = v(s) = \mathrm{Var}[\sum_{i=s+1}^{\infty} \tilde{X}_i^2]$. Clearly $v(s)$ decreases to zero as s increases, since

$$\lim_{s\uparrow\infty} \sum_{i=s+1}^{\infty} \tilde{X}_i^2 = 0.$$

For any particular ε then, one can choose $s \geq k$ so that $v(s)/\varepsilon^2 \leq p$ (say). Then for this s,

$$\begin{aligned} p &\geq \Pr\left(\left| \sum_{i=s+1}^{\infty} \tilde{X}_i^2 - \tau \right| \geq \varepsilon \right) \\ &= 1 - \Pr\left(-\varepsilon < \sum_{i=1}^{\infty} \tilde{X}_i^2 - \left(\sum_{i=1}^{s} \tilde{X}_i^2 + \tau \right) < \varepsilon \right) \\ &= 1 - \Pr\left(e^{-\sigma\varepsilon} < \frac{\exp\{-\sigma(\sum_{i=1}^{s} \tilde{X}_i^2 + \tau)\}}{\exp\{-\sigma \sum_{i=1}^{\infty} \tilde{X}_i^2\}} < e^{\sigma\varepsilon} \right). \end{aligned} \tag{4.2}$$

For a given σ and probability p, $\varepsilon > 0$ is determined so that the interval $(e^{-\sigma\varepsilon}, e^{\sigma\varepsilon})$ is as narrow as desired, and the integer $s \geq k$ is then determined so that $v(s)/\varepsilon^2 \leq p$.

It remains only to calculate $v(s)$. Using (3.4), for $i > k$,

$$\mathrm{Var}_{\tilde{\mu}}[\tilde{X}_i^2] = \frac{4!\Gamma(\theta+n)}{\Gamma(\theta+4+n)} \left(\frac{\theta}{\theta+4} \right)^{i-k} - \frac{4\Gamma^2(\theta+n)}{\Gamma^2(\theta+2+n)} \left(\frac{\theta}{\theta+2} \right)^{2(i-k)}.$$

For $i > j > k$, using (3.4) and (3.5), one obtains after some calculation

$$\begin{aligned} \mathrm{cov}_{\tilde{\mu}}[\tilde{X}_i^2, \tilde{X}_j^2] = {} & \frac{4\,\Gamma(\theta+n)}{\Gamma(\theta+4+n)} \left(\frac{\theta}{\theta+4} \right)^{j-k} \left(\frac{\theta}{\theta+2} \right)^{i-j} \\ & - \frac{4\Gamma^2(\theta+n)}{\Gamma^2(\theta+2+n)} \left(\frac{\theta}{\theta+2} \right)^{i+j-2k}. \end{aligned}$$

Using properties of the geometric series, one then obtains, for $s \geq k$,

$$\begin{aligned} v(s) = {} & \frac{(\theta+4)(\theta+6)\Gamma(\theta+n)}{\Gamma(\theta+4+n)} \left(\frac{\theta}{\theta+4} \right)^{s-k+1} \\ & - \frac{(\theta+2)^2\,\Gamma^2(\theta+n)}{\Gamma^2(\theta+2+n)} \left(\frac{\theta}{\theta+2} \right)^{2(s-k+1)}. \end{aligned}$$

We can now state:

ALGORITHM 1 (Monte Carlo approximation of the basic integral).

```
set n, θ, σ, ε, p
repeat {
        generate (ñ_1, ..., ñ_{K_n})
        determine s so that v(s)/ε^2 ≤ p
        repeat {
                generate (X̃_1, ..., X̃_s) using the RAM
                store exp{−σ(Σ_{i=1}^s X̃_i^2 + τ)}
        }
        calculate the average of exp{−σ(Σ_{i=1}^s X̃_i^2 + τ)}
}
```

For a given value of θ, i.i.d. realizations of the samples $(\tilde{n}_1, \ldots, \tilde{n}_{K_n})$ are generated using the "Chinese Restaurant Construction" of Dubins and Pitman [see Aldous (1985)]. The parameters of the beta distributions for W_i, $i \le k$, may be large enough that common methods for generating beta variates perform poorly. For simulation purposes, we used integer values of θ, and took as the beta variate an appropriate order statistic from an i.i.d. uniform (0, 1) sample [see Stuart and Ord (1994), Ripley (1987)].

Each of the analytical approximations (2.7), (3.7) and (3.11) has a main component which approximates the symmetric functional

$$E_\mu\left[\exp\left\{-\sigma\sum_{i=1}^{\infty}X_i^2\right\}\Big|A_n=\mathbf{a},K_n=k\right]. \tag{4.3}$$

The different approximations to (4.3) can be calculated for each sample $(\tilde{n}_1, \ldots, \tilde{n}_{K_n})$ as part of the outermost loop of Algorithm 1, allowing for comparison. The y-coordinate values in Figure 1 were obtained by subtracting the lower bound for (4.3),

$$\exp\{-\sigma n^2(f_n+1/n+\theta/n^2)/(\theta+n+1)(\theta+n)\},$$

from the corresponding weak-convergence (upper row) and Monte Carlo (lower row) approximations, over i.i.d. realizations $(\tilde{n}_1, \ldots, \tilde{n}_{K_n})$. For values of σ used here, the upper bound is of a different order of magnitude and is not shown. Judging from the graphs in Figure 1, the weak-convergence and Monte Carlo approximations tend to differ most when θ is large, with weak-convergence apparently over-estimating the basic functional (4.3) in samples with low homozygosities. In general, the Monte Carlo approximation tends to be intermediate to the weak-convergence and lower bound approximations. Although the Monte Carlo approximation is subject to stochastic variation, a second source of variability in these graphs originates from samples having different numbers of alleles, but the same homozygosity value. We favor use of the Monte Carlo approximation for numerically-based work, provided computing resources are not severely limiting.

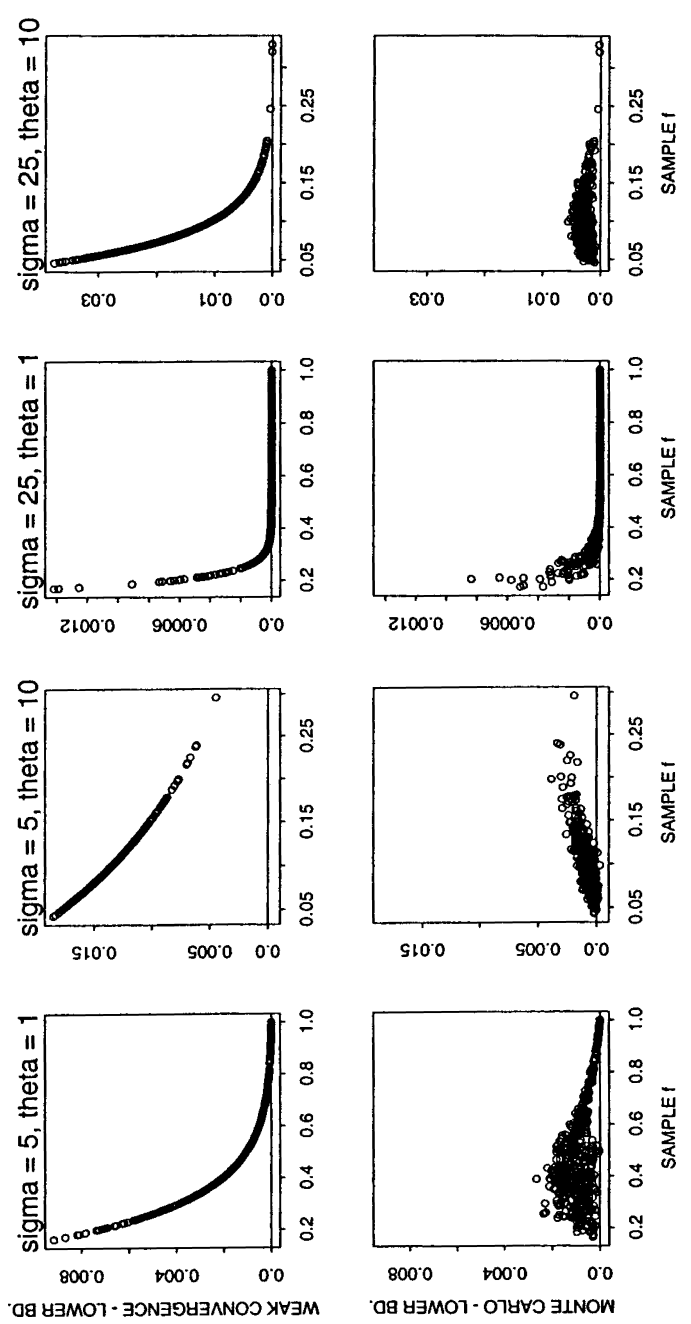

FIG. 1. *Comparison of analytical and numerical approximations of the basic functional* (4.3)*, as a function of the sample homozygosity* f_n*. Following Algorithm* 1*,* 500 *independent samples of size* $n = 200$ *were generated using the Chinese Restaurant Construction. For each sample, the Monte Carlo approximation* S_m/m *was calculated using* $m = 10,000$*,* $\varepsilon = 10^{-5}$ *and* $p = 10^{-3}$*. The weak-convergence and lower bound approximations for* (4.3) *were also calculated for each sample. The lower bound is subtracted from the weak-convergence (upper row) and Monte Carlo (lower row) approximations to give the y-coordinate value for each sample.*

4.2. *The distribution of F_n.* A rejection rule algorithm [see Ripley (1987)] can be used to simulate values of F_n under the various approximate formulae. As before, we generate candidate samples $(\tilde{n}_1, \ldots, \tilde{n}_{K_n})$ via the Chinese Restaurant Construction, and calculate an approximation to (4.3) for each sample. Rejection or acceptance of a candidate sample depends on the ratio of the sample probability under the target and proposal distributions, respectively $P_{\theta,\sigma}$ and the ESF [see Ripley (1987) for a general description]. For the symmetric overdominance model, the ratio is

$$\rho = \frac{P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k)}{P_{\theta,0}(A_n = \mathbf{a}, K_n = k)} = \gamma^{-1} E_\mu \left[\exp\left\{ -\sigma \sum_1^\infty X_i^2 \right\} \Big| A_n = \mathbf{a}, K_n = k \right].$$

In practice, γ is unknown, and we use an approximation $\hat{E}_\mu$ for the conditional expectation; the key ratio is then

$$\hat{\rho} = \hat{E}_\mu \left[\exp\left\{ -\sigma \sum_1^\infty X_i^2 \right\} \Big| A_n = \mathbf{a}, K_n = k \right],$$

which has a (perhaps crude) upper bound $\hat{\rho} \leq 1$. A rejection rule algorithm for generating i.i.d. samples $(a_1, \ldots, a_n)$ using the Monte Carlo approximation for (4.3) is:

ALGORITHM 2 (Rejection rule sampling).

```
set n, θ, σ, ε, p
repeat {
    repeat {
        generate (ñ_1, ..., ñ_{K_n})
        generate U ~ u(0, 1)
        determine s so that v(s)/ε^2 ≤ p
        repeat {
            generate (X̃_1, ..., X̃_s) using the RAM
            store e^{-σ(Σ_{i=1}^s X̃_i^2 + τ)}
        }
        calculate Ê = the average of e^{-σ(Σ_{i=1}^s X̃_i^2 + τ)}
    until U ≤ Ê
    }
    store (a_1, ..., a_n)
}
```

The interior loop of the algorithm is required to calculate the Monte Carlo approximation. Alternatively, $\hat{E}$ could be calculated using one of the approximations for (4.3) contained in formulae (2.7), (3.7) or (3.11).

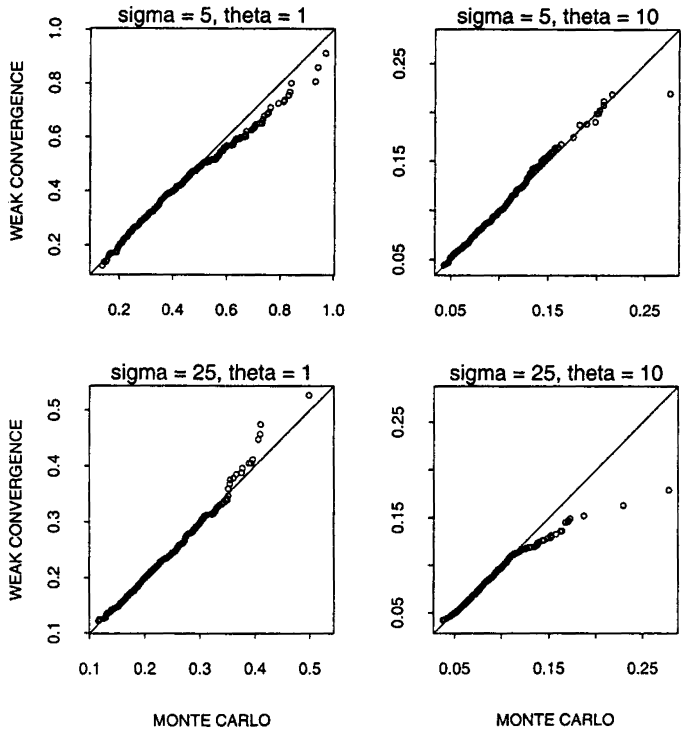

FIG. 2. *Quantile–quantile plots of the distribution of F_n, for samples generated by the fully numerical rejection method of Algorithm* 2 (*x-coordinate*) *and by rejection with the weak-convergence approximation* (2.7) (*y-coordinate*). *Following Algorithm* 2, *independent samples of size* $n = 200$ *were generated using the Chinese Restaurant Construction. For the Monte Carlo quantiles, the acceptance ratio*

$$\hat{\rho} = \max\Big\{S_m/m, \exp\Big(-\sigma n^2(f_n + 1/n + \theta/n^2)/(\theta + n + 1)(\theta + n)\Big)\Big\},$$

with $m = 100, \varepsilon = 10^{-3}$ *and* $p = 10^{-2}$, *was calculated for each candidate sample until* 500 *samples had been accepted. The weak-convergence quantiles are from independent runs of the rejection rule algorithm, using* $\hat{\rho} = \exp\{-\sigma(n/(n+\theta))^2 f_n\}$ *in the rejection step.*

Figure 2 shows quantile–quantile plots of the sample homozygosity statistic F_n, based on rejection rule sampling from approximate $P_{\theta,\sigma}$. In each graph, we plot F_n quantiles based on the weak-convergence approximation (2.7) against fully numerical quantiles obtained from Algorithm 2. Recalling from Figure 1 that $\hat{E} = S_m/m$ in Algorithm 2 can be less than its theoretical lower bound, we have set $\hat{E}$ equal to the lower bound when this occurs. As shown in the figure, the weak-convergence and fully numerical quantiles tend to differ in the upper tail of the F_n distribution, and the discrepancies appear to be magnified by increasing both σ and θ. The accuracy of the weak-convergence approximation cannot be assured unless the upper and lower bounds which enclose it are close together; yet for $\sigma = 25$, a very large sample would be required to do this. Although our coarse

sampling of rare values in the upper tail most likely contributes to the apparent discrepancies, we view the Monte Carlo quantiles as probably the more accurate representation of the distribution of F_n.

5. Conditional sampling formulae and an HLA data set.

5.1. *Conditional formulae.* Proceeding as in (1.10), the conditional probability of a sample $\mathbf{a} = (a_1, \ldots, a_n)$, given $\sum_{i=1}^n a_i = k$, is

$$\begin{aligned} P_\nu(A_n = \mathbf{a} | K_n = k) &= \frac{P_\nu(A_n = \mathbf{a}, K_n = k)}{P_\nu(K_n = k)} \\ &= \frac{E_\nu[\Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}{E_\nu[\Pr(K_n = k|\mathbf{X})]} \\ &= \frac{E_\mu[g(\mathbf{X}) \Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}{E_\mu[g(\mathbf{X}) \Pr(K_n = k|\mathbf{X})]}. \end{aligned} \tag{5.1}$$

We can make use of approximate expressions for the numerator by writing the above as

$$P_\nu(A_n = \mathbf{a} | K_n = k) = \frac{E_\mu[g(\mathbf{X}) \Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}{\sum_{\mathbf{a} \in \boldsymbol{\alpha}_k} E_\mu[g(\mathbf{X}) \Pr(A_n = \mathbf{a}, K_n = k|\mathbf{X})]}, \tag{5.2}$$

where

$$\boldsymbol{\alpha}_k = \left\{ (a_1, a_2, \ldots, a_n) : \sum_1^n i a_i = n, \sum_1^n a_i = k \right\}.$$

Using Approximation 1 in the numerator and denominator above, we find:

APPROXIMATION 4 (Conditional on K_n).

$$\begin{aligned} &P_{\theta,\sigma}(A_n = \mathbf{a} | K_n = k) \\ &\quad \approx \frac{P_{\theta,0}(A_n = \mathbf{a}, K_n = k) \gamma^{-1} \exp\{-\sigma (n/(n+\theta))^2 f_n\}}{\sum_{\mathbf{a} \in \boldsymbol{\alpha}_k} P_{\theta,0}(A_n = \mathbf{a}, K_n = k) \gamma^{-1} \exp\{-\sigma (n/(n+\theta))^2 f_n\}} \\ &\quad = \frac{P_{\theta,0}(A_n = \mathbf{a}, K_n = k) \exp\{-\sigma (n/(n+\theta))^2 f_n\}}{P_{\theta,0}(K_n = k) E[\exp\{-\sigma (n/(n+\theta))^2 F_n\} | K_n = k]} \\ &\quad = P_0(A_n = \mathbf{a} | K_n = k) \frac{\exp\{-\sigma (n/(n+\theta))^2 f_n\}}{E[\exp\{-\sigma (n/(n+\theta))^2 F_n\} | K_n = k]}, \end{aligned} \tag{5.3}$$

where

$$P_0(A_n = \mathbf{a} | K_n = k) = \frac{n!}{|S_n^k| a_1! \cdots a_n! 1^{a_1} \cdots n^{a_n}} \tag{5.4}$$

is the conditional ESF given by the right hand side of (1.2), and the expectation in the denominator is with respect to (5.4). The first-order expansion of (5.3) in powers of σ is Watterson's (1977) approximation for $P_{\theta,\sigma}(A_n = \mathbf{a}|K_n = k)$, provided one uses

$$n^2/(n+\theta)(n+\theta+1) \approx \big(n/(n+\theta)\big)^2 \tag{5.5}$$

in Watterson's expression. A very similar expression can be obtained by conditioning in Approximation 2; indeed, the conditional formulae obtained from Approximations 1 and 2 differ only in the factor (5.5).

Two main advantages of working in the conditional setting are evident, especially if inferences about σ are of primary interest. First, by using the same approximation for $P_{\theta,\sigma}(A_n = \mathbf{a}, K_n = k)$ in the numerator and denominator of (5.3), we have ensured that the denominator of the final expression is an exact normalizing constant. Second, by conditioning on K_n, we obtain sampling formulae which depend only weakly on θ. When $\theta \ll n$, a crude estimate of θ would be adequate in (5.3), and the likelihood could be treated as a one-parameter family. In the neutral model, K_n is a sufficient statistic for θ, and K_n is asymptotically normally distributed, with mean and variance both equal to $\theta \log n$ [Watterson (1974)]. For the general selection model with $d\nu/d\mu$ bounded, Joyce (1995) has shown that K_n is again asymptotically $N(\theta \log n,\ \theta \log n)$; indeed, the asymptotic distribution of K_n distinguishes, in part, the general selection model from Pitman's two-parameter model [see Pitman (1996)]. In light of (5.3), K_n appears to be asymptotically sufficient for θ in the symmetric overdominance model.

Although the undetermined constant γ^{-1} cancelled in the conditional formula (5.3), it has been replaced with $E[\exp\{-\sigma(n/(n+\theta))^2 F_n\}|K_n = k]$. Rejection-rule sampling again appears to be warranted for generating samples from the conditional distribution. Stewart's algorithm [see Fuerst, Chakraborty and Nei (1977)] can be used to generate samples $(a_1, a_2, \ldots, a_n)$; $\sum_1^n a_i = k$, which have the conditional ESF (5.4) as their probability law, and these can be used as candidate samples for rejection rule sampling from (5.3). To implement a conditional version of Algorithm 2, we use a method communicated by Jim Pitman for obtaining size-biased samples from those generated by Stewart's algorithm: suppose the allele labels of the sample are $(y_1, y_2, \ldots, y_n)$ in some particular order, and let π be a uniform random permutation of $\{1, 2, \ldots, n\}$. By exchangeability, the sequence $(y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(n)})$, obtained by permuting the allele labels at random, is equal in distribution to the sequence of labels of a size-biased sample [see Aldous (1985)]. The size-biased sample atom $\tilde{n}_i$ is then the number of representatives of the ith distinct label appearing in $(y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(n)})$. Conditioning on $K_n = k$ in Algorithm 2 offers enough computational savings that upper limits on σ for feasible rejection rule sampling can be extended.

5.2. *Estimation of σ for a Human Leukocyte Antigen sample.* The Human Leukocyte Antigen (HLA) loci, which code for proteins important in the immune response, are thought to be under some form of "balancing" selection (selection favoring a diversity of allelic types) [see Parham and Ohta (1996)]. The symmetric overdominance model is perhaps the simplest model of balancing selection; indeed, the assumption of equal heterozygote fitnesses is severe from a biological standpoint, as is the stationarity assumption underlying all of our formulae. However, efforts to analyze HLA data even under this simple model have been hampered by the absence of sampling formulae that incorporate selection and mutation explicitly.

Table 1 shows estimated HLA-B allele frequencies and counts from a sample of 99 serologically typed Australian Aboriginals. There are $n = 198$ gene copies in the sample and $k = 21$ alleles, taking the single "blank" copy as a distinct allele. The homozygosity statistic for this sample is $f = 0.104$, which lies roughly between the 0.05 and 0.1 quantiles of the neutral distribution of F, suggesting a departure from the neutral model in the direction of overdominance. The effective population size N for Australian Aboriginals is thought to be considerably lower than 1000 [see Cavalli-Sforza, Menozzi and Piazza (1994)], suggesting that $\sigma \approx 2Ns$ might be moderate enough to estimate using a rejection-rule approach. The mutation rate u to new HLA alleles is still the subject of speculation, but is very unlikely to be higher than 10^{-3}, a rate typical of microsatellite loci [see Weber and Wong (1993)], among the most mutable in the human genome. The constraints on u, along with those on N, suggest that $\theta \approx 4Nu$ in this population could hardly be greater than 10, and is probably much lower. For these reasons, and owing to its relative simplicity, we favor estimation of σ based on the conditional distribution.

We have used established Monte Carlo methods [see, e.g., Penttinen (1984), Geyer and Thompson (1992)] based on realizations of samples $(a_1, \ldots, a_n)$ under the conditional model to find approximate maximum-likelihood estimates of σ. Briefly, the log-likelihood ratio comparing sample probabilities at selection intensities σ and ϕ is

$$\begin{aligned} l(\sigma) &= \log \frac{P_{\theta,\sigma}(A_n = \mathbf{a} | K_n = k)}{P_{\theta,\phi}(A_n = \mathbf{a} | K_n = k)} \\ &= \log \frac{E_\mu[\exp\{-\sigma \mathcal{F}\} | A_n = \mathbf{a},\, K_n = k]}{E_\mu[\exp\{-\phi \mathcal{F}\} | A_n = \mathbf{a},\, K_n = k]} \\ &\quad - \log \frac{E_\mu[\exp\{-\sigma \mathcal{F}\} | K_n = k]}{E_\mu[\exp\{-\phi \mathcal{F}\} | K_n = k]}. \end{aligned} \tag{5.6}$$

Using the conditional formula (5.3), the first term of (5.6) is

$$(-\sigma + \phi)\left(\frac{n}{n+\theta}\right)^2 f_n, \tag{5.7}$$

and the second term is estimated as the average of (5.7) over realizations from the conditional distribution $P_{\theta,\phi}(A_n = \mathbf{a} | K_n = k)$. Under the fully numerical

TABLE 1
Estimated HLA-B allele frequencies and counts from a sample of 99 *serologically typed Australian Aboriginals* ($n = 198$ *gene copies*)*

Allele	**Frequency (%)**	**Count ($n = 198$)**
B61	19.6	39
B56	17.2	34
B60	10.0	20
B13	7.5	15
B62	7.1	14
B27	6.6	13
B44	6.6	13
B35	3.5	7
B57	3.5	7
B75	3.4	6
B14	3.0	6
B51	2.5	5
B38	2.0	4
B7	1.5	3
B39	1.5	3
B8	1.0	2
B55	1.0	2
B58	1.0	2
B37	0.5	1
B41	0.5	1
B-Blank	0.4	1

*From the 11th International Histocompatibility Workshop [Tsuji, Aizawa and Sasazuki (1992)]. An individual serological panel giving a positive reading for only one HLA allele could result either from a homozygous genotype, or from a failure in a reaction with a (possibly novel) "blank" HLA protein. Therefore, a standard maximum likelihood method, which resolves these uncertainties by use of the Hardy–Weinberg Law, was used to obtain allele frequency estimates. We converted the estimated frequencies to counts by rounding the product $n \times$ (frequency) to the nearest integer. The only ambiguity that arose was for allele B75, which was given the count "6" to ensure that the total number of gene copies remained at $n = 198$.

approach, the first term of (5.6) is obtained from a single execution of Algorithm 1, using a size-biased version of the sample data in place of the generated sample. To obtain the second term of (5.6) numerically,

$$\hat{E} = \overline{Z(\phi)} = \text{ the average of} \exp\left\{-\phi\left(\sum_{i=1}^{s} \tilde{X}_i^2 + \tau\right)\right\}$$

is used in the rejection step of Algorithm 2, and the ratio $\overline{Z(\sigma)}/\overline{Z(\phi)}$ is averaged over accepted samples, conditional on $K_n = k$. An estimate of the likelihood curve using either the fully numerical or formula-based approach is obtained by varying σ in a neighborhood of ϕ.

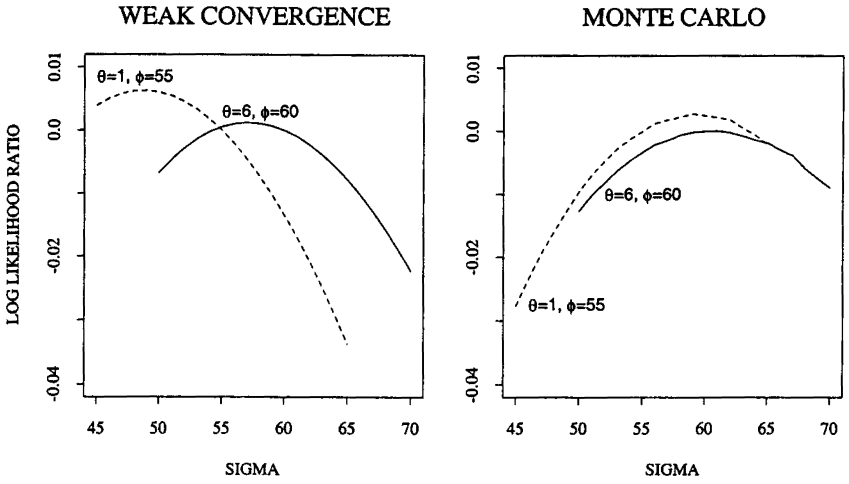

FIG. 3. *Estimated likelihood curves for σ, for the Australian Aboriginal data of Table* 1. *For each curve,* 500 *i.i.d. samples of size $n = 198$ were generated by rejection rule sampling under an approximate conditional $(k = 21)$ distribution, with selection parameter ϕ. Rejection ratios and numerical parameters for the weak-convergence* (*left panel*) *and Monte Carlo* (*right panel*) *approximations are as in Figure* 2, *with ϕ substituted for σ in the Figure* 2 *description. Dashed curves give log-likelihood ratios $\log(P_{\theta,\sigma}/P_{\theta,\phi})$ for $\theta = 1, \phi = 55$* (*with maxima at $\sigma = 49$ in the left panel and $\sigma = 59$ in the right*). *Solid curves give log-likelihood ratios for $\theta = 6, \phi = 60$* (*with maxima at $\sigma = 57$ in the left panel and $\sigma = 61$ in the right*).

Figure 3 shows estimated likelihood curves for the Aboriginal data using the conditional formula (5.3) (left panel) and the fully numerical approach (right panel). For each method, we considered two candidate values of θ: the moderate value $\theta = 1$, and $\theta = 6$, the closest integer value to the solution of

$$E[K_n] = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1} = k,$$

the expected value of K_n under the neutral model [see Ewens (1979)], with $n = 198$ and $k = 21$. For a given value of θ, $E[K_n]$ as shown above is less than the expected value of K_n under symmetric overdominance selection [Ewens (1979)], so $\theta = 6$ is perhaps the largest plausible value supported by the HLA-B data set.

As shown by Figure 3, the two approximate methods and two different values of θ lead to relatively modest differences in $\hat{\sigma}$. Although stochastic variation in $\overline{Z(\sigma)}/\overline{Z(\phi)}$ is evident in the piecewise-smooth property of the Monte Carlo curves, we favor the Monte Carlo method in this case; for $\hat{\sigma}$ appears to be in a region of the parameter space where, given the sample size, the error in the approximate formula (5.3) could be unacceptably large. Clearly, more intensive numerical work is needed before stronger conclusions can be drawn. The value $\sigma \approx 55$, with $N \approx 500$ taken as a crude estimate for the Aboriginal effective population size, suggests selective differentials of a few percent for heterozygotes, in agreement

with values obtained for HLA-B by other methods [see Satta et al. (1994)]. As estimates of selection parameters like σ are known to be inconsistent [see Joyce (1994)], further work of a more purely statistical nature is needed in order to understand the uncertainties associated with these estimates.

Acknowledgments. Over the last several years, M. Grote has had the opportunity to discuss this work with many of the authors cited in the manuscript. We are very grateful for their input and encouragement. In particular, we thank Paul Joyce and Jim Pitman, who read an earlier version of the manuscript and made suggestions which led to considerable revisions. In addition, Steve Evans, Peter Bickel, John Gillespie and an anonymous referee gave valuable technical advice. Finally, we thank Glenys Thomson and Charles Langley for their patience and encouragement.

REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. In *Ecole d'Été de Probabilités de Saint-Flour XIII. Lecture Notes in Math.* **1117**. Springer, Berlin.

BILLINGSLEY, P. (1986). *Probability and Measure.* Wiley, New York.

CAVALLI-SFORZA, L. L., MENOZZI, P. and PIAZZA, A. (1994). *The History and Geography of Human Genes.* Princeton Univ. Press.

ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence.* Wiley, 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, Berlin.

ETHIER, S. N. and KURTZ, T. G. (1994). Convergence to Fleming–Viot processes in the weak atomic topology. *Stochastic Process. Appl.* **54** 1–27.

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

EWENS, W. J. (1979). *Mathematical Population Genetics*. Springer, Berlin.

EWENS, W. J. (1990). Population genetics theory — the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (S. Lessard, ed.) 177–227. Kluwer, Dordrecht.

FUERST, P. A., CHAKRABORTY, R. and NEI, M. (1977). Statistical studies on protein polymorphism in natural populations I: Distribution of single-locus heterozygosity. *Genetics* **86** 455–483.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

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

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions.* Wiley, New York.

JOYCE, P. (1994). Likelihood ratios for the infinite alleles model. *J. Appl. Probab.* **31** 595–605.

JOYCE, P. (1995). Robustness of the Ewens Sampling Formula. *J. Appl. Prob.* **32** 609–622.

JOYCE, P. and TAVARÉ, S. (1995). The distribution of rare alleles. *J. Math. Biol.* **33** 602–618.

KIMURA, M. and CROW, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49** 725–738.

KINGMAN, J. F. C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser. B* **37** 1–22.

KINGMAN, J. F. C. (1977). The population structure associated with the Ewens Sampling Formula. *Theoret. Population Biol.* **11** 274–283.

PARHAM, P. and OHTA, T. (1996). Population biology of antigen presentation by MHC class I molecules. *Science* **272** 67–74.

PENTTINEN, A. (1984). Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jyväskylä Studies in Computer Science, Economics and Statistics* **7**.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158.

PITMAN, J. (1996). Some developments of the Blackwell–MacQueen Urn Scheme. In *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds.). IMS, Hayward, CA.

RIPLEY, B. D. (1987). *Stochastic Simulation.* Wiley, New York.

SATTA, Y., O'HUIGIN, C., TAKAHATA, N. and KLEIN, J. (1994). Intensity of natural selection at the Major Histocompatibility Complex loci. *Proceedings of the National Academy of Sciences of the United States of America* **91** 7184–7188.

STUART, A. and ORD, J. K. (1994). *Kendall's Advanced Theory of Statistics* **1**: *Distribution Theory.* Arnold, London.

TSUJI, K., AIZAWA, M. and SASAZUKI, T. (eds.) (1992). *HLA 1991: Proceedings of the Eleventh International Histocompatibility Workshop and Conference* **1**. Oxford Science Publications, Oxford.

WATTERSON, G. A. (1974). The sampling theory of selectively neutral alleles. *Adv. Appl. Probab.* **6** 463–488.

WATTERSON, G. A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Probab.* **13** 639–651.

WATTERSON, G. A. (1977). Heterosis or neutrality? *Genetics* **85** 789–814.

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

WEBER, J. L. and WONG, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics* **2** 1123–1128.

SECTION OF EVOLUTION AND ECOLOGY
ONE SHIELDS AVENUE
UNIVERSITY OF CALIFORNIA
DAVIS, CA 95616
E-MAIL: mngrote@ucdavis.edu

DEPARTMENT OF STATISTICS
367 EVANS HALL #3860
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720-3860
E-MAIL: terry@stat.berkeley.edu