JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 6, 2012

Pp. 650–661
DOI: 10.1089/cmb.2012.0033

Efficient Simulation and Likelihood Methods for Non-Neutral Multi-Allele Models

PAUL JOYCE,[1] ALAN GENZ,[2] and ERKAN OZGE BUZBAS[3]

ABSTRACT

Throughout the 1980s, Simon Tavaré made numerous significant contributions to population genetics theory. As genetic data, in particular DNA sequence, became more readily available, a need to connect population-genetic models to data became the central issue. The seminal work of Griffiths and Tavaré (1994a, 1994b, 1994c) was among the first to develop a likelihood method to estimate the population-genetic parameters using full DNA sequences. Now, we are in the genomics era where methods need to scale-up to handle massive data sets, and Tavaré has led the way to new approaches. However, performing statistical inference under non-neutral models has proved elusive. In tribute to Simon Tavaré, we present an article in spirit of his work that provides a computationally tractable method for simulating and analyzing data under a class of non-neutral population-genetic models. Computational methods for approximating likelihood functions and generating samples under a class of allele-frequency based non-neutral parent-independent mutation models were proposed by Donnelly, Nordborg, and Joyce (DNJ) (Donnelly et al., 2001). DNJ (2001) simulated samples of allele frequencies from non-neutral models using neutral models as auxiliary distribution in a rejection algorithm. However, patterns of allele frequencies produced by neutral models are dissimilar to patterns of allele frequencies produced by non-neutral models, making the rejection method inefficient. For example, in some cases the methods in DNJ (2001) require 10^9 rejections before a sample from the non-neutral model is accepted. Our method simulates samples *directly* from the distribution of non-neutral models, making simulation methods a practical tool to study the behavior of the likelihood and to perform inference on the strength of selection.

Key words: algorithms, statistics.

1. INTRODUCTION

While genetic variation is ultimately resolved by DNA sequencing, much of our understanding of natural populations continues to be based on the results of allele-frequency analysis. Therefore, developing likelihood-based methods to assess the strength of selection using allele-frequency

[1]Department of Mathematics and Initiative for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho.
[2]Department of Mathematics, Washington State University, Pullman, Washington.
[3]Department of Biology, Stanford University, Stanford, California.

data is of interest. However, evaluating likelihood functions and simulating samples under non-neutral models is computationally difficult, posing challenges in developing likelihood-based inference methods.

In this article, we present computationally efficient methods for approximating likelihood functions and generating samples under a class of non-neutral models. We consider models that describe the distribution of allele frequencies in a sample of chromosomes from diploid organisms at a single non-recombining region of the genome in the following setup. We assume a fixed number of allelic types at a region of the genome. The frequency of a particular allele is defined as the ratio of the number of chromosomes that are identical to the given allele over this region to the total number of chromosomes in the population. Alleles are subject to a random parent-independent mutation process, in which, the type of the new allele from a mutation event is chosen from the types existing in the population, independently of the type of the parent. Selection acts on a diploid genotype of two alleles and each genotype in the population is assigned a selection coefficient representing the relative advantage of that genotype with respect to a fixed arbitrary genotype in the population. Under these asumptions, the distribution of allele frequencies in a population evolving for a long time can be obtained as the stationary distribution in the diffusion limit, which is governed by a mutation parameter and a matrix of selection parameters. Ultimately, we are interested in performing likelihood-based inference on the strength of selection, given a sample of allele frequencies from the population.

Likelihoods arising from the distribution of the allele frequencies in the class of non-neutral models described in the previous paragraph have an unknown normalizing constant that depends on the parameters of the model. Evaluation of likelihood functions and generating samples under the non-neutral models require approximating this unkown constant, either by direct computation or by a stochastic simulation approach that bypassess its direct computation. Before describing our method, we briefly discuss the disadvantages associated with stochastic simulation approaches available to evaluate likelihoods and to generate samples for the class of non-neutral models we study.

In principle, rejection sampling, importance sampling, and Markov chain Monte Carlo (MCMC) methods can be used to simulate allele-frequency data from non-neutral models and to approximate the normalizing constants in non-neutral models. These Monte Carlo methods use an auxiliary distribution under which simulating the allele frequencies is easy and their computational efficiency increase when the auxiliary distribution mimics the target non-neutral model closely. However, designing a good auxiliary distribution for non-neutral models is difficult in practice because of high-dimensional data and potentially high-dimensional parameter space. When a distribution based on a neutral model is used as the auxiliary distribution to generate a sample from a non-neutral model by rejection sampling, the acceptance rate remains low. For example, under the model presented in DNJ (2001), the rejection method with an auxiliary distribution based on the neutral model in some cases requires 10^9 trials before a sample from the distribution under the non-neutral model is accepted. Approximating normalizing constants by importance sampling, by first expressing them as expectations of neutral distributions and using samples generated from neutral models, suffers from the same inefficieny. Distributions based on neutral models make poor auxiliary distributions in stochastic simulation methods targeting non-neutral models, because under appreciable selection, the allele-frequency patterns produced by neutral models differ substantially from the allele-frequency patterns produced by non-neutral models.

MCMC methods have also been used to simulate samples under a variety of non-neutral models. For example, Fearnhead (2001, 2003), developed methods based on perfect simulation (Propp and Wilson, 1996). MCMC methods based on perfect simulation have the advantage to be applicable to parent-dependent mutation schemes, in which, the type of a mutant allele may depend on the type of its parent. Murray et al. (2006) developed a similar approach to sample the posterior distribution of parameters when the model has an unknown normalizing constant dependent on the parameters (Møller et al., 2006). However, perfect simulation is computationally feasible only for small values of the selection parameter and for haploid populations. As the strength of selection increases, one has to look further back in time in the history of the sample and deeper on the genealogical tree to capture the branches on which selection events occur on the genealogical tree. Further, MCMC methods produce correlated samples from the target distribution, which is a disadvantage when independent samples are required. In summary, there is a lack of efficient computational methods that can operate under a wide range of parameter values and high-dimensional data and parameter spaces.

We present numerical analysis methods to compute the parameter-dependent unknown constant efficiently and accurately for a sub-class of non-neutral models in which the matrix of selection parameters is diagonal. In our numerical analysis approach, we break the high-dimensional integral in the normalizing

constant into a series of one-dimensional iteratively defined integrals. In the end, we compute these integrals to produce a series of numerical approximations to a sequence of conditional cumulative distribution functions. Using these cumulative distribution functions, we simulate independent samples directly from distributions of non-neutral models. To simulate samples from non-neutral models with a non-diagonal matrix of selection parameters, our numerical analysis methods allow the use of distributions in non-neutral models with a diagonal matrix of selection parameters as auxiliary distributions in a rejection sampling setup, improving the efficiency of the rejection sampling.

Our method combines the advantages of stochastic approaches with the advantages of numerical approaches. By taking a numerical analysis approach to compute the normalizing constant, we avoid stochastic error. Yet, because we can produce a series of conditional cumulative distribution functions, simulating independent samples directly from the distribution of interest is feasible. Combined with stochastic samplers such as rejection algorithm or MCMC, our numerical methods provide an efficient algorithm to sample the joint posterior distribution of the mutation and selection parameters under non-neutral models.

We demonstrate the application of our methods with two examples using a symmetric balancing selection model, which corresponds to the special case where the matrix of selection parameters is diagonal with equal elements. Balancing selection is a diversity-promoting mode of selection, resulting from a variety of mechanisms, such as heterozygote advantage and negative-frequency-dependence. Applications of balancing selection arise in many areas, from plant genetics to environmental genomics. Estimating the strength of balancing selection and determining whether balancing selection is a prevalent force are key to understanding evolution in these contexts.

In the first example, we investigate the performance of our inferences on the strength of balancing selection using our methods in conjunction with approximate Bayesian computation, based on rejection sampling (Tavaré et al. 1997; Pritchard et al., 1999; Beaumont et al., 2002). Approximate Bayesian computation performs inference on the posterior distribution of the parameters utilizing the data sets simulated under a model. In the second example, we focus on the stationary distribution of the r largest allele frequencies in a population with K total allelic types (Watterson, 1977). Based on this distribution, we investigate the effect of using an incorrect K-allele model in which small-frequency alleles are missing in the sample, on the posterior distribution of the selection parameter, using an MCMC approach.

2. MODELS

2.1. *Neutral* K-*allele model*

In a population with K allelic types, evolving for a long time under the neutral model with parent-independent mutation, the stationary distribution of allele frequencies $(x_1, x_2, \ldots x_K)$, where x_i denotes the frequency of the ith allelic type, is given by the Dirichlet distribution (Wright, 1949):

$$f_N(\mathbf{x}; \theta\nu) = \frac{1}{b(\theta\nu)} x_1^{\theta\nu_1 - 1} x_2^{\theta\nu_2 - 1} \ldots x_K^{\theta\nu_K - 1} I\left\{\sum_{i=1}^{K} x_i = 1\right\}. \tag{1}$$

Here, $\theta = 4N\mu$ is the population-scaled mutation parameter, where N is the effective population size, μ is the mutation rate per gene per generation, $\nu_i\left(\sum_{i=1}^{K} \nu_i = 1\right)$ is the fraction of the mutation rate corresponding to allele i, and $I\{A\}$ is the indicator function taking a value of 1 on set A and 0 otherwise. The normalizing constant $b(\theta\nu)$ is a well-known ratio of gamma functions and is given by

$$b(\theta\nu) = \frac{\prod_{i=1}^{K} \Gamma(\theta\nu_i)}{\Gamma(\theta \sum_{i=1}^{K} \nu_i)}.$$

The sampling distribution $L_N(\mathbf{n})$ of a sample of size n under the multinomial sampling from distribution given in equation (1) is given by

$$L_N(\mathbf{n}) = \frac{n!(\theta\nu_1)_{(n_1)}(\theta\nu_2)_{(n_2)} \cdots (\theta\nu_K)_{(n_K)}}{n_1! n_2! \cdots n_K! (\theta)_{(n)}},$$

where, $\mathbf{n}(n_1, n_2, \ldots, n_K)$, n_i is the number of copies of ith allele in the sample, $n = \sum_{i=1}^{K} n_i$ is the sample size, and $z_{(n)} = z(z+1)\cdots(z+n-1)$ is the rising factorial. The multinomial sampling probabilities in this

case are also given by $b(\theta v + \mathbf{n})/b(\theta v)$. The Dirichlet distribution is conjugate under multinomial sampling (Kotz, 2000), meaning that if the prior probabilities on the population frequencies is Dirichlet distributed, then the posterior probabilities of the population frequencies given a sample is also Dirichlet distributed. The conditional distribution of the population frequencies given a sample $\mathbf{n} = (n_1, n_2, \cdots, n_K)$ from the distribution in equation (1) is Dirichlet with

$$f_{N|\mathbf{n}}(\mathbf{x}; \theta\nu) = \frac{1}{b(\theta\nu + \mathbf{n})} x_1^{\theta\nu_1 + n_1 - 1} x_2^{\theta\nu_2 + n_2 - 1} \ldots x_K^{\theta\nu_K + n_K - 1} I\left\{\sum_{i=1}^{K} x_i = 1\right\}. \tag{2}$$

2.2. K-*allele model with selection*

The stationary distribution of allele frequencies in the presence of selection is related to the stationary distribution under neutrality given in equation (1) in the following way. For a population with allele frequencies x, we define the mean fitness $\bar{\sigma}(\mathbf{x})$ by

$$\bar{\sigma}(\mathbf{x}) = \sum_{i,j=1}^{K} \sigma(A_i, A_j) x_i x_j,$$

where $\sigma(A_i, A_j)$ denotes the parameter for the strength of selection of a genotype with allelic types A_i and A_j. We let Σ be the symmetric matrix whose (i, j)th element is $\sigma(A_i, A_j) = \sigma_{ij}$. Then, the stationary distribution of allele frequencies in the K-allele model with selection (Wright, 1949) can be expressed as

$$\begin{aligned} f_S(\mathbf{x}; \Sigma, \theta\nu) &= \frac{e^{\mathbf{x}'\Sigma\mathbf{x}}}{c(\Sigma, \theta\nu)} x_1^{\theta\nu_1 - 1} x_2^{\theta\nu_2 - 1} \ldots x_K^{\theta\nu_K - 1} I\left\{\sum_{i=1}^{K} x_i = 1\right\} \\ &= \frac{e^{\sum_{i=1}^{K}\sum_{j=1}^{K} \sigma_{ij} x_i x_j}}{c(\Sigma, \theta\nu)} x_1^{\theta\nu_1 - 1} x_2^{\theta\nu_2 - 1} \ldots x_K^{\theta\nu_K - 1} I\left\{\sum_{i=1}^{K} x_i = 1\right\}, \end{aligned} \tag{3}$$

where $c(\Sigma, \theta v)$ is the normalizing constant and x' denotes the transpose of the vector x. The normalizing constant is given by

$$c(\Sigma, \theta\nu) = \int e^{\sum_{i=1}^{K}\sum_{j=1}^{K} \sigma_{ij} x_i x_j} x_1^{\theta\nu_1 - 1} x_2^{\theta\nu_2 - 1} \ldots x_{K-1}^{\theta\nu_{K-1} - 1} x_K^{\theta\nu_K - 1} dx_K \cdots dx_2 dx_1,$$

subject to $\sum_{i=1}^{K} x_i = 1$, where we used the integral sign for multiple integration over the allele-frequency space. An argument in Joyce (1994) uses equation (2) to show that the relationship between the sampling distribution of a sample of size n under selection, $L_S(\mathrm{n})$, and under neutrality, $L_N(\mathbf{n})$, is given by

$$\begin{aligned} L_S(\mathbf{n}) &= L_N(\mathbf{n}) \frac{E_N(e^{\mathbf{x}'\Sigma\mathbf{x}}|\mathbf{n})}{E_N(e^{\mathbf{x}'\Sigma\mathbf{x}})} = L_N(\mathbf{n}) \frac{c(\Sigma, \theta\nu + \mathbf{n})}{c(\Sigma, \theta\nu)} \frac{b(\theta\nu)}{b(\theta\nu + \mathbf{n})} \\ &= \frac{c(\Sigma, \theta\nu + \mathbf{n})}{c(\Sigma, \theta\nu)}, \end{aligned} \tag{4}$$

where $E_N(\cdot)$ is the expectation with respect to the distribution under neutrality given by equation (1) and the last equality follows by the fact that $L_N(\mathbf{n}) = b(\theta v + \mathbf{n})/b(\theta v)$.

Under non-neutral models we describe, the likelihoods are intractable because the normalizing constants $c(\Sigma, \theta v)$ and $c(\Sigma, \theta v + \mathbf{n})$ in equation (4) are difficult to compute. Monte Carlo integration, based on generating many independent copies of $\mathbf{x}_i$ under the neutral model and approximating the expectation $E_N(e^{\mathbf{x}'\Sigma\mathbf{x}})$ in equation (4) by the average

$$\frac{1}{M} \sum_{i=1}^{M} e^{\mathbf{x}_i'\Sigma\mathbf{x}_i} \tag{5}$$

is inefficient when selection is strong, because the allele-frequency patterns under the non-neutral model are substantially different than the allele-frequency patterns under the neutral model. Importance sampling improves the estimation of $E_N(e^{\mathbf{x}'\Sigma\mathbf{x}})$ to some extent, although obtaining many such estimates remains computationally infeasible.

3. COMPUTING LIKELIHOODS AND SIMULATING DATA

In the last section, we argued that the intractability of likelihoods under non-neutral models is due to difficulties in obtaining the normalizing constants, which appear as high-dimensional integrals in the form $c(\Sigma, \theta v)$ and $c(\Sigma, \theta v + \mathbf{n})$. In this section, we present a numerical analysis approach to compute these constants in the special case where Σ is a diagonal matrix. The method extends the results presented in Genz and Joyce (2003) and has some similarities to a method reported by Fearnhead and Meligkotsidou (2004). Below, we give a brief description of how we implement the numerical methods of Joyce and Genz, and focus in detail on how these methods can be utilized in performing statistical inference on the strength of selection. The mathematical properties of the numerical methods we use, such as computational efficiency, algorithmic complexity, accuracy of approximations, and handling of the end-point singularities in the integrals are described in detail in Joyce and Genz (2003).

The main idea behind our numerical analysis approach is to break the K dimensional integral in the normalizing constant into K one dimensional iteratively defined integrals. Each one-dimensional integral is calculated by numerical integration methods, producing an approximation to the value of the normalizing constant and a series of numerical approximations to a sequence of conditional cumulative distribution functions. On one hand, computing the constant with this method allows the evaluation of likelihood functions. On the other hand, using the cumulative distribution functions produced in the process of computing the normalizing constant, allele frequencies can be simulated directly from the non-neutral model. Thus, using efficient numerical methods, we approximate the intractable likelihoods under non-neutral models with a diagonal matrix of selection parameters.

3.1. Computation of the normalizing constant when Σ is diagonal

We let Σ be a diagonal matrix with elements $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_K)$, $\sigma_i < 0, \forall i$, in the diagonal and 0 elsewhere. We implement the numerical methods with $\sigma_i < 0$ because this type of selection matrix characterizes balancing selection, a selection scheme of biological importance as we discuss in our examples in Section 3.2. The normalizing constant for the stationary distribution of allele frequencies under the non-neutral model under this parameterization is given by

$$\begin{aligned} c(\sigma, \theta\nu) = & \int_0^1 x_1^{\theta\nu_1 - 1} \int_0^{1-x_1} x_2^{\theta\nu_2 - 1} \cdots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\theta\nu_{K-2}-1} \\ & \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\theta\nu_{K-1}-1} \int_0^{1-\sum_{i=1}^{K-1} x_i} x_K^{\theta\nu_K - 1} \\ & \times e^{-\sum_{i=1}^{K} \sigma_i x_i^2} dx_K \cdots dx_2 dx_1. \end{aligned}$$

The special structure of the integrand allow computing $c(\sigma, \theta v)$ as a sequence of one-dimensional integrals. In order to demonstrate this, we first separate the inner exponential term into factors which are distributed to the respective outer integrals, and we rewrite $c(\sigma, \theta v)$ as

$$\begin{aligned} c(\sigma, \theta\nu) = & \int_0^1 x_1^{\alpha_1} e^{-\sigma_1 x_1^2} \int_0^{1-x_1} x_2^{\alpha_2} e^{-\sigma_2 x_2^2} \cdots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\alpha_{K-2}} e^{-\sigma_{K-2} x_{K-2}^2} \\ & \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}} e^{-\sigma_{K-1} x_{K-1}^2} g_K\left(1 - \sum_{i=1}^{K-1} x_i\right) dx_{K-1} \cdots dx_2 dx_1, \end{aligned} \tag{6}$$

where $g_K(y) \equiv y^{\alpha K} e^{-\sigma K y^2}$ and $\alpha_i \equiv \theta v_i - 1$. We note that the constant $c(\sigma, \theta v)$ is expressed as a $(K-1)$-dimensional integral in equation (6). This is possible since the frequency of the Kth allele given the frequencies $x_1, x_2, \ldots, x_{K-1}$, can be determined using the constraint $\sum_{i=1}^{K} x_i = 1$ on the allele frequencies, which allows us to write $x_K = 1 - \sum_{i=1}^{K-1} x_i$. Now we consider the integral with respect to x_{K-1} on the right side of equation (6) which we rewrite as

$$\int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\alpha_{K-1}} e^{-\sigma_{K-1} x_{K-1}^2} g_K\left(1 - \sum_{i=1}^{K-2} x_i - x_{K-1}\right) dx_{K-1},$$

and we note that this last expression depends on the variables $x_1, x_2, \cdots, x_{K-2}$ only through the term $1-\sum_{i=1}^{K-2} x_i$. We define $y_{K-1}=1-\sum_{i=1}^{K-2} x_i$ and denote the last integral by the function $g_{K-1}(y_{K-1})$. We have

$$g_{K-1}(y_{K-1})=\int_0^{y_{K-1}} x_{K-1}^{\alpha_{K-1}} e^{-\sigma_{K-1}x_{K-1}^2} g_K(y_{K-1}-x_{K-1})dx_{K-1}. \tag{7}$$

Substituting the right hand side of equation 7 into equation (6), we get

$$c(\sigma, \theta\nu)=\int_0^1 x_1^{\alpha_1} e^{-\sigma_1 x_1^2} \int_0^{1-x_1} x_2^{\alpha_2} e^{-\sigma_2 x_2^2} \cdots \int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\alpha_{K-2}} e^{-\sigma_{K-2}x_{K-2}^2} g_{K-1}\left(1-\sum_{i=1}^{K-2} x_i\right) dx_{K-2} \cdots dx_2 dx_1.$$

Continuing in this manner, we define the successive integrals by

$$g_i(y_i)=\int_0^{y_i} t^{\alpha_i} e^{-\sigma_i t^2} g_{i+1}(y_i-t)dt, \tag{8}$$

for $y_i=1-\sum_{j=1}^{i-1} x_j$, $y_1 = 1$, and $i = K-1, K-2, \ldots, 1$. The required normalizing constant $c(\sigma, \theta v)$ is given by $g_1(1)$.

For numerical computations, we introduce the same mesh k/m, for $k=0, 1, \ldots, m$, for each of the y_i variables. Then the $g_i(y_i)$ values on this mesh are approximated using a trapezoidal integration rule, and $g_1(1)$ is approximated by computing $g_i(y_i)$ mesh approximations for $i = K-1, K-2, \ldots, 2$. The resulting $g_1(1)$ approximation can be accurately and efficiently computed in $O(m^2)$ time (Genz and Joyce, 2003), except in some cases, where the α_i values produce end-point singularities for the $g_i(y_i)$ integrals. In those cases, we use standard numerical methods for ''subtracting out the singularities'' (David and Rabinowitz, 1984), to eliminate the singular integrand behaviors and produce rapidly converging approximations to $g_1(1)$.

3.2. Generating samples when Σ is a diagonal matrix

We define the cumulative distribution functions $F_i(\cdot; z)$ with parameter z, for $i=1, 2, \cdots, K$ as

$$F_i(y; z)=\begin{cases} 0 & y \le 0 \\ \dfrac{\int_0^y t^{\alpha_i} \exp(-\sigma_i t^2) g_{i+1}(z-t)dt}{g_i(z)} & 0 \le y \le z \\ 1 & y > z \end{cases}$$

where $g_i(y)$ is defined by equation (8). We note that $P(X_i \le y | X_{i-1}, \cdots, X_1)=F_i(y; 1-X_1-\cdots-X_{i-1})$. Since $F_i(\cdot;z)$ is a strictly increasing function for $0 < y < z$, it has a well-defined inverse that we denote by $F_i^{-1}(\cdot; z)$. Given these definitions, we now generate a sequence of random variables $\mathbf{X}=(X_1, X_2, \cdots, X_K)$ with joint probability density function given by equation (3) when the matrix of selection parameters is diagonal. The method, which is an application of inversion sampling (i.e., ''the CDF method'') is as follows. We let U_1 be a uniformly distributed random number on [0, 1]. That is, $U_1 \sim$UNIF[0, 1]. We define $X_1=F_1^{-1}(U_1, 1)$. Now we define iteratively

1. Generate $U_i \sim$ UNIF $[0, 1-X_1-X_2-\cdots-X_{i-1}]$,
2. Define $X_i=F_i^{-1}(U_i; 1-X_1-X_2-\cdots-X_{i-1})$.

If the normalizing constant is calculated using the method described in Section 3.1, then approximations for the $g_i(y)$ functions have also been computed, and we can generate independent random vectors of allele frequencies using the above algorithm. Each vector $\mathbf{X}$ is a realization of the (random) population allele frequencies under selection. To simulate a sample of size n with allele counts given by $\mathbf{n}$ and distribution given by equation (3), it is sufficient to draw a multinomial sample conditional on the given population frequencies $\mathbf{X}$.

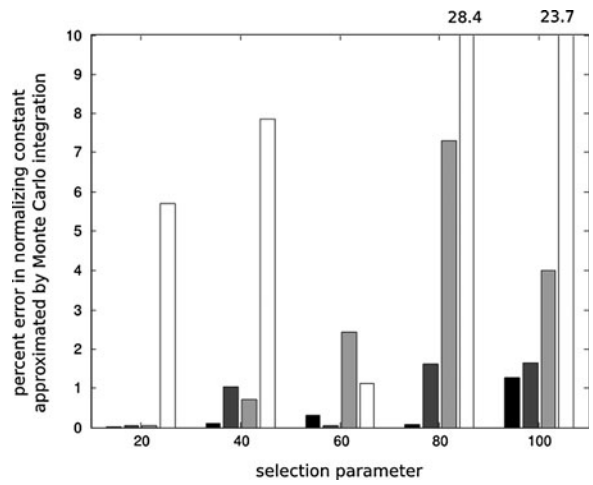

FIG. 1. Percent error in normalizing constant $c(\Sigma, \theta\nu)$ obtained by Monte Carlo integration using samples simulated under the neutral model when the selection matrix in the non-neutral model is diagonal and all values of selection parameters are equal. Monte Carlo averages in calculating the normalizing constants are based on 10^3,10^4,10^5, and 10^6 samples drawn from the neutral model, illustrated from dark to light color respectively. The true value of $c(\Sigma, \theta\nu)$ is assumed to be the value computed by numerical analysis methods. As the strength of selection (given by parameter $\sigma = (20, 40, 60, 80, 100)$) increases, the relative error in $c(\Sigma, \theta\nu)$ obtained by Monte Carlo integration increases on average because of substantial difference between allele frequency patterns in the non-neutral model and allele frequency patterns in the neutral model. Percent error values larger than 10% are given at the top of the corresponding bar numerically.

4. APPLICATIONS OF THE NUMERICAL METHOD FOR INFERENCE UNDER NON-NEUTRAL MODELS

Computation of the normalizing constant $c(\sigma, \theta\nu)$ by methods of Section 4.1 allows us to approximate the intractable likelihoods and to perform inference on the strength of selection under non-neutral models. For comparison purposes, we provide percent error estimates when the normalizing constant $c(\sigma, \theta\nu)$ is obtained by naive Monte Carlo integration using equation (5), with a diagonal matrix of selection parameters, where we assume that $c(\sigma, \theta\nu)$ computed using our numerical methods is the true value (Fig. 1). Accuracy of our numerical methods has been studied in Genz and Joyce (2003). In practice, the normalizing constants can be computed only once on a grid and stored in a lookup table for repeated use. In our tests, computing an approximate likelihood surface with 1000×1000 evaluations for a non-neutral model with $K = 10$ alleles and a diagonal matrix of selection parameters took on average less than minute on a decent desktop computer.

The maximum likelihood estimates for the strength of selection can be obtained by standard maximization methods of likelihood surfaces, computed by our numerical methods. Alternatively, the joint posterior distribution of selection and mutation parameters can be sampled combining our numerical methods with MCMC (Buzbas et al., 2009, 2011).

4.1. Bayesian inference under non-neutral models

By Bayes' Theorem, under the non-neutral model with diagonal matrix of selection parameters, $\Sigma = \mathrm{diag}(\sigma, \sigma, \ldots, \sigma)$, the joint posterior distribution of selection and mutation parameters can be written as

$$\pi(\sigma, \theta\nu|\mathbf{x}) = \frac{f(\mathbf{x}|\sigma, \theta\nu)\pi(\sigma, \theta\nu)}{\int\int f(\mathbf{x}|\sigma, \theta\nu)\pi(\sigma, \theta\nu)d\sigma d\theta\nu}, \tag{9}$$

where $f(\mathbf{x}|\sigma, \theta\nu)$ is the likelihood and $\pi(\sigma, \theta\nu)$ is the joint prior. For simplicity, we assume that mutation is non-preferential among the K alleles and we fix $\nu = (1/K, 1/K, \ldots 1/K)$ although use of other schemes does not add difficulty to the analysis. Further, we assume the prior independence of selection and mutation parameters and write $\pi(\sigma, \theta\nu) = \pi(\sigma)\pi(\theta\nu)$. When sampling the joint posterior distribution in equation (9), the normalizing constant in the denominator need not be computed, since most stochastic samplers require the posterior distribution to be known up to a constant *free* of parameters. However, this is *not* the normalizing constant computed by our methods in Section 3.1. In equation (9), the parameter-dependent normalizing constant computed by our methods is hidden in the likelihood $f(\mathbf{x}|\sigma, \theta\nu)$. We let, $\mathbf{x}_o = (x_{1o}, x_{2o}, \ldots, x_{Ko})$ and $\mathbf{n}_o$ to denote the observed population frequencies and sample frequencies respectively. If the sample frequencies are taken as proxy to

population frequencies and the population distribution is used, the likelihood $f(\mathbf{x}_o|\sigma, \theta\nu)$ can be explicitly written as

$$f(\mathbf{x}_o|\sigma, \theta\nu) = \frac{e^{-\sigma \sum_{i=1}^{K} x_{io}^2}}{c(\sigma, \theta\nu)} x_{1o}^{\theta/K-1} x_{2o}^{\theta/K-1} \ldots x_{Ko}^{\theta/K-1}, \tag{10}$$

and if the sampling distribution is used, it can be written as

$$f(\mathbf{n}_o|\sigma, \theta\nu) = L_N(\mathbf{n}_o) \frac{c(\sigma, \theta\nu + \mathbf{n}_o)}{c(\sigma, \theta\nu)} \frac{b(\theta\nu)}{b(\theta\nu + \mathbf{n}_o)} = \frac{c(\sigma, \theta\nu + \mathbf{n}_o)}{c(\sigma, \theta\nu)}. \tag{11}$$

We now define the statistics

$$F = \sum_{i=1}^{K} x_i^2, \quad G = -\sum_{i=1}^{K} \log x_i \tag{12}$$

and express the density of the non-neutral model with diagonal matrix of selection parameters as

$$\begin{aligned} f(\mathbf{x}|\sigma, \theta\nu) &= \frac{e^{-\sigma \sum_{i=1}^{K} x_i^2}}{c(\sigma, \theta\nu)} x_1^{\theta/K-1} x_2^{\theta/K-1} \ldots x_K^{\theta/K-1} \\ &= \frac{e^{-\sigma F}}{c(\sigma, \theta\nu)} e^{-\log \prod_{i=1}^{K} x_i^{\theta/K-1}} \\ &= \frac{e^{-\sigma F + (\theta/K-1)G}}{c(\sigma, \theta\nu)}. \end{aligned}$$

The right-hand side of the last equality depends on the allele frequencies $\mathbf{x}$ only through the functions of F and G, and the Fisher-Neyman factorization theorem implies that F and G are jointly sufficient for the parameters σ and θ. To sample the joint posterior distribution of the selection and mutation parameters we use an approximate Bayesian computation algorithm with the likelihood based on equation (10):

Algorithm–ABC (with rejection):

1. Simulate a value θ^* from the prior distribution $\pi(\theta)$ and a value σ^* from the prior distribution $\pi(\sigma)$
2. Simulate a vector of population allele-frequencies $\mathbf{x}^*$ from the non-neutral model with parameters θ^*, σ^* using the numerical methods of Section 4.
3. Compute the statistics F^*, G^* by equation (12) using $\mathbf{x}^*$.
4. Accept θ^*, σ^* as a sample from the posterior distribution if

$$||(F^*, G^*) - (F_o, G_o)|| < \delta,$$

where F_o and G_o are the statistics computed from the observed data, $||\cdot||$ is the Euclidean distance, and δ is a suitable tolerance parameter. The posterior sample obtained by Algorithm–ABC is from the correct posterior distribution as δ approaches zero. We also note that, since the summary statistics are sufficient for the parameters under this non-neutral model, there is no approximation associated with reducing the full-data likelihood to a likelihood based on the summary statistics. An example for the performance of inference on the joint distribution of the parameters is given in Figure 2, using the likelihood based on the population frequencies (i.e., equation (10)). We simulated the ''true'' population frequencies used as test data, from a model with 10 replications, from a locus with $K = 10$ alleles, under a selection parameter $\sigma = 10$ and a mutation parameter $\theta = 2$. We generated 10^5 data sets under the model with 10^5 parameter values (σ, θ), drawn from their prior distribution (Step 2 in Algorithm-ABC above), and 1% of the parameters which generated the summary statistics (F^*, G^*) closest to the observed summary statistics (F_o, G_o) in the Euclidean sense, are taken as a sample from the posterior distribution.

4.2. The distribution of the r-largest allele frequencies

In models considered thus far, we assumed that all the allelic types present in the population are observed in the sample, treating K as fixed and known. In many practical problems, this assumption might not be reasonable. For example, in highly-polymorphic immune system genes in humans such as Human

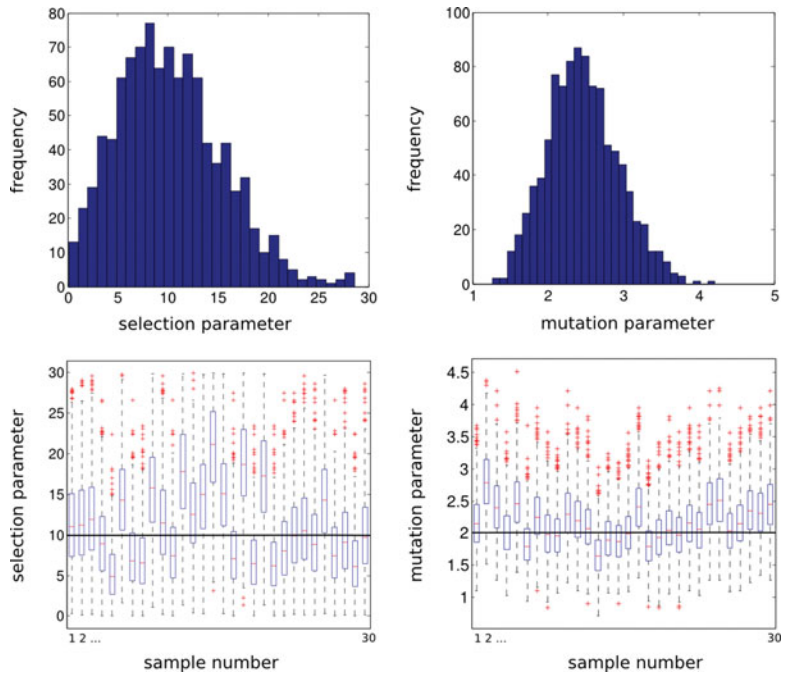

FIG. 2. Posterior distributions **(top)** and boxplots **(bottom)** of selection and mutation parameters. The "true" population frequencies are simulated under a $K = 10$ allele model with 10 replicates (e.g., 10 loci) and a diagonal selection matrix where all elements are equal (i.e., symmetric balancing selection model). The "true" selection parameter is $\sigma = 10$ and the mutation parameter is $\theta = 2$. Top left and right histograms are posterior samples of selection and mutation parameters respectively, both obtained by approximate Bayesian computation as explained in Algorithm-ABC in the text. Bottom left and right series of box plots are from posterior samples for 30 independent runs of "true" population frequencies, obtained by the same procedure as above.

Leukocyte Antigen (HLA), some of the low-frequency alleles in the population will likely be missing in the sample. The stationary distribution of allele frequencies for a diagonal matrix of selection parameters for various cases of missing data and when K is unkown is studied in detail by Watterson (1977). Here, we consider the distribution of the r-largest allele frequencies in a population with K total alleles ($r < K$), to investigate the effect of missing low-frequency alleles in the sample when performing inference on the strength of balancing selection.

The distribution of the r-largest allele frequencies $\mathbf{x} = (x_1, x_2, \ldots, x_r)$, $x_1 > x_2 > \cdots > x_r > \cdots > x_K$, in a population of K total types can be written as (Watterson, 1977),

$$f_S^{(K,r)}(\mathbf{x}; \sigma, \theta\nu) = \frac{K!}{(K-r)!} \times e^{-\sigma \sum_{i=1}^{r} x_i^2} x_1^{\theta\nu_1 - 1} x_2^{\theta\nu_2 - 1} \ldots x_r^{\theta\nu_r - 1} \times \frac{c^{(K,r)}(\sigma, \theta\nu)}{c(\sigma, \theta\nu)}, \tag{13}$$

where

$$c^{(K,r)}(\sigma, \theta\nu) = \int_0^{1-\sum_{i=1}^{r} x_i} x_{r+1}^{\theta\nu_{r+1} - 1} \int_0^{1-\sum_{i=1}^{r+1} x_i} x_{r+2}^{\theta\nu_{r+2} - 1} \ldots \tag{14}$$

$$\int_0^{1-\sum_{i=1}^{K-3} x_i} x_{K-2}^{\theta\nu_{K-2} + n_{K-2} - 1} \int_0^{1-\sum_{i=1}^{K-2} x_i} x_{K-1}^{\theta\nu_{K-1} + n_{K-1} - 1} \tag{15}$$

$$\int_0^{1-\sum_{i=1}^{K-1} x_i} x_K^{\theta\nu_K + n_K - 1} \times e^{-\sigma \sum_{i=r+1}^{K} x_i^2} dx_K \cdots dx_2 dx_{r+1}. \tag{16}$$

The distribution of the r-largest allele frequencies has a normalizing constant that is a ratio of two integrals. The first integral, which appears in the denominator of equation (13) is the normalizing constant of the full model with K-alleles. The second integral, which appears in the numerator of equation (13) is a $(K - r)$-dimensional integral and is a sum over the frequencies of unobserved alleles. The methods of Section 3.1 can be modified to compute the integral in the numerator and hence the normalizing constant in equation (13). Application of our numerical methods, once for the integral in the numerator and once for the integral in the denominator, allows us to evaluate likelihoods in r-largest allele frequencies based on equation (13).

To demonstrate the use of the distribution of r-largest allele frequencies, we consider the following idealized example, where sample frequencies are reasonable proxy for population frequencies. We simulate

a vector of population frequencies under the model with $K = 10$, $\sigma = 100$, and $\theta = 5$ to represent the ''true'' state of the population at a locus. We now consider three analyses. First, we assume that $r = 7$ largest-frequency alleles are observed in a sample with the true frequencies as in the population and analyze these frequencies using a $K = 7$ allele model. This is a case where three alleles are missing in the data set, since the true K is 10, and we are unaware that we are missing the three smallest-frequency alleles, using a model with $K = 7$ to analyze the data. Second, we assume that $r = 7$ largest-frequency alleles are observed in a sample with the true frequencies as in the population and analyze these data using equation (13) where $K = 10$. This is a case where again three alleles are missing in the data set, since the true K is 10, but this time we know that we are missing the three smallest-frequency alleles and use a model with $K = 10$ to analyze the data. Third, we assume all $K = 10$ alleles are observed in a sample with the true frequencies as in the population and analyze these data with a model with $K = 10$ alleles. The posterior distribution obtained from the third analysis represents the ideal situation, in which there are no missing alleles in the sample and the sample frequencies are proxy for the population frequencies. Therefore, the posterior distribution obtained from the third procedure is the gold standard for the posterior distributions obtained from the first and second analyses. We obtain posterior samples for all three scenarios using an MCMC approach with 100,000 iterations and 10,000 burn in steps.

The posterior distribution of the selection and mutation parameters obtained using the first analysis results in biased estimates for selection and mutation parameters (Fig. 3). Low-frequency alleles carry valuable information on mutation and consequently the posterior distribution of the mutation parameter is substantially affected. Using the distribution of the r-largest allele frequencies results in improved estimation, as the model compensates the missing alleles. However, a good estimate of K is required in this case, close to its true value. These results suggest that in a K-allele balancing selection model, it might be important to capture the low-frequency alleles in the sample to obtain good estimates of the strength of selection, especially when K is unknown and no good estimates for K exist.

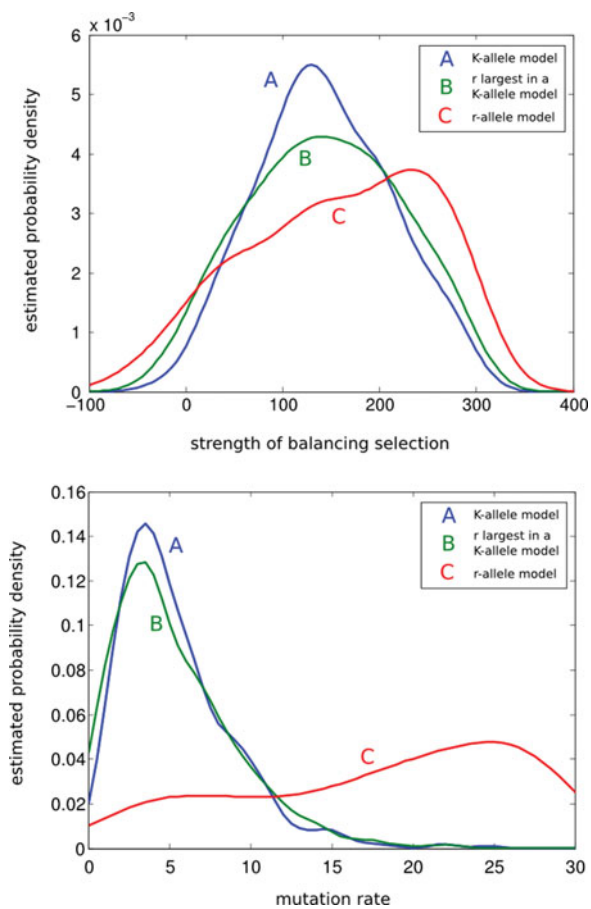

FIG. 3. Kernel density estimates of posterior distributions of selection **(top)** and mutation **(bottom)** parameters obtained analyzing allele-frequencies under three non-neutral models. The models differ in their assumptions on the number of allelic types in the population, K. The ''true'' population has $K = 10$ alleles but only 7 largest-frequency alleles are observed in the sample. Green plots are obtained by analyzing these 7 frequencies under the assumption that $r = 7$ largest-frequency alleles out of $K = 10$ total alleles are observed in the data. Red plots are obtained by analyzing the 7 frequencies under the assumption that these are all the existing allelic types in the population. Hence, the second model corresponds to using a $K = 7$ allele model, whereas the frequencies are actually generated under a model with $K = 10$. Using the second model, the mutation parameter is overestimated (bottom, red plot) because of missing three small-frequency alleles. Consequently, estimates for the selection parameter are not accurate. Blue plots are obtained by assuming ideal conditions where all 10 allelic types are observed and the sample frequencies are proxy for the population frequencies. Hence, the blue plots are obtained under the correct model when there is no missing data and inference in this case is the gold standard for the other two models.

5. CONCLUSION

Both likelihood and Bayesian statistical approaches provide powerful tools for modeling and analyzing genetic polymorphism. However, as we are now well into the genomics era, the argument for or against statistical methods is now more practical than philosophical. The central question is whether statistical methods, especially those that involve averaging over a large amount of missing data, scale appropriately to handle large genomic data sets. Some have argued that the exponential increase in computing power, that we have enjoyed over the past two decades, will continue and effectively solve the problem of scalability of statistical methods for genomic data sets. We now know that increased computing power per se is not enough to provide methods that scale up to handle immense amount of data pouring out of tech-powered wet labs. Therefore, modelers and statisticians have taken two conceptual approaches to address the problem of scalability, both of which aim to maintain benefits of statistical modeling while keeping the methods practical for genomic data sets. The first approach is to abandon the likelihood analysis based on the full-data and settle for an approximate method. Examples include composite likelihood methods and likelihood methods based on summary statistics (Hudson, 2001; Marchini, 2004). While these approaches are suboptimal from a statistical perspective, the increase in algorithmic speed often more than makes up for the slight loss of information. The second approach is to settle for an approximate mechanistic description of the natural system and build a statistical model on which an analysis based on full-data likelihood can be performed (Li and Stephens, 2003). Thus, the first approach uses a simplified statistical method whereas the second approach uses a simplified statistical model. For example, our assumption of using a parent-independent mutation scheme in non-neutral models used in this paper, is a simplification of a more realistic mutation model and therefore an example to the second approach. On the other hand, in our first application in Section 5, we performed inference using an approximate Bayesian computation approach based on the likelihood of summary statistics and not the likelihood of the full-data. Hence, our analysis in that application is in principle an example to the first approach, although in our case summary statistics were sufficient and there was no information loss.

Using a parent-independent mutation model may limit the application of non-neutral models considered in this article when a more realistic mutation model is desired. However, when our numerical methods are embedded in likelihood-based statistical methods, they are scalable to genomic data sets and may provide a reasonable approximate method for detecting selection when the parent-independent mutation assumption is violated.

ACKNOWLEDGMENTS

Funding for this work was provided by NSF EPSCoR, EPS-0132626, NSF Population Biology Panel DEB-0089756, and NIH NCRR grant NIH NCRR 1P20RR016448-01. The research of P.J. is supported in part by the Initiative in Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho. The research of E.B. is supported in part by NIH R01 GM081441, NIH P20 RR016454 from the INBRE Program of the National Center for Research Resources.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Beaumont, M.A., Zhang, W., and Balding, D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035.

Buzbas, E.O., and Joyce, P. 2009. Maximum likelihood estimates under k-allele models with selection can be numerically unstable. *Ann. Appl. Stat.* 3, 1147–1162.

Buzbas, E.O., Joyce, P., and Abdo, Z. 2009. Estimation of selection intensity under overdominance by Bayesian methods. *Stat. Appl. Genet. Mol. Biol.* 8, article32.

Buzbas, E.O., Joyce, P., and Rosenberg, N.A. 2011. Inference on balancing selection for epistatically interacting loci. *Theor. Popul. Biol.* 79, 102–113.

Davis, P.J. and Rabinowitz, P. 1984. *Methods of Numerical Integration*. Academic Press, New York.

Donnelly, P., Nordborg, M., and Joyce, P. 2001. Likelihoods and simulation methods for a class of non-neutral population genetics models. *Genetics* 159, 853–867.

Genz, A., and Joyce, P. 2003. Computation of the normalizing constant for exponentially weighted Dirichlet distribution integrals. *Comput. Sci. Stat.* 35, 181–212.

Fearnhead, P. 2001. Perfect simulation from population genetic model with selection. *Theor. Popul. Biol.* 59, 263–279.

Fearnhead, P. 2003. Ancestral processes for non-neutral models of complex diseases. *Theor. Popul. Biol.* 63, 115–130.

Fearnhead, P., and Meligkotsidou, L. 2004. Exact filtering for partially observed continuous time models. J.R. *Stat. Soc. Ser.* B 66, 771–789.

Griffiths, R.C., and Tavaré, S. 1994a. Ancestral inference in population genetics. *Stat. Sci.* 9, 307–319.

Griffiths, R.C., and Tavaré, S. 1994b. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Ser. B* 344, 403–410.

Griffiths, R.C., and Tavaré, S. 1994c. Simulating probabiity distributions in the coalescent. *Theor. Popul. Biol.* 46, 131–159.

Hudson, R.R. 2001. Two-locus sampling distributions and their application. *Genetics* 159, 1805–1817.

Joyce, P. 1994. Likelihood ratios for the infinite alleles model. *J. Appl. Probabil.* 31, 595–605.

Kotz, S., Balakrishnan, N., and Johnson, N.L. 2000. *Continuous Multivariate Distributions, Volume 1: Models and Applications*, 2nd ed. Wiley, New York.

Li, N., and Stephens, M. 2003. Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* 165, 2213–2233.

Marchini, J., Cardon, L., Phillips, M. et al. 2004. The effects of human population structure on large genetic association studies. *Nat. Genet*. 36, 512–517.

Møller, J., Pettitt, A., Berthelsen, K., et al. 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* 93, 451–458.

Murray, I., Ghahramani Z., and Mackay, D.J.C. 2006. MCMC for doubly-intractable distributions. In *Proc. 22nd Annu. Conf. Uncertainty Artif. Intell*. 359–366.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. et al. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16, 1791–1798.

Propp, J.G., and Wilson, D.B. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct.* Algorithms 9, 223–252.

Tavaré, S., Balding, D.J., Griffiths, R.C., et al. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145, 505–518.

Watterson, G.A. 1977. Heterosis or neutrality? *Genetics* 85, 789–814.

Wright, S. 1949. Adaptation and selection, 365–389. In Jepson, G.L., Simpson, G.G., and Mayr, E., eds. *Genetics, Palaeontology, and Evolution.* Princeton University Press, Princeton, NJ.

Address correspondence to:
Dr. Erkan Ozge Buzbas
Department of Biology
Stanford University
371 Serra Mall
Stanford, CA 94305

E-mail: buzbas@stanford.edu