Copyright © 2008 by the Genetics Society of America
DOI: 10.1534/genetics.108.087361

The Polymorphism Frequency Spectrum of Finitely Many Sites Under Selection

Michael M. Desai* and Joshua B. Plotkin†,[1]

**Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544 and †Department of Biology and Program in Applied Mathematics and Computation Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104*

Manuscript received January 21, 2008
Accepted for publication October 6, 2008

ABSTRACT

The distribution of genetic polymorphisms in a population contains information about evolutionary processes. The Poisson random field (PRF) model uses the polymorphism frequency spectrum to infer the mutation rate and the strength of directional selection. The PRF model relies on an infinite-sites approximation that is reasonable for most eukaryotic populations, but that becomes problematic when θ is large ($\theta \gtrsim 0.05$). Here, we show that at large mutation rates characteristic of microbes and viruses the infinite-sites approximation of the PRF model induces systematic biases that lead it to underestimate negative selection pressures and mutation rates and erroneously infer positive selection. We introduce two new methods that extend our ability to infer selection pressures and mutation rates at large θ: a finite-site modification of the PRF model and a new technique based on diffusion theory. Our methods can be used to infer not only a "weighted average" of selection pressures acting on a gene sequence, but also the distribution of selection pressures across sites. We evaluate the accuracy of our methods, as well that of the original PRF approach, by comparison with Wright–Fisher simulations.

MUTATION rates and selective pressures are of central importance to evolution. The number and frequency distribution of genetic polymorphisms within a population carry information about these fundamental processes. Polymorphisms at higher frequencies reflect weaker selective pressures (or positive selection), and vice versa. Similarly, a larger number of polymorphisms indicates a higher mutation rate. Thus we can use the polymorphism frequency spectrum observed in genetic sequences sampled from a population to infer the mutation rate and the selection pressure acting on the sequence.

This intuition can be formalized into a rigorous method for estimating selection pressures and mutation rates by calculating the likelihood of sampled polymorphism data as a function of these parameters. The Poisson random field (PRF) model provides an important and widely used method of doing so. The PRF model assumes a panmictic population of constant size, free recombination, infinite sites, no dominance or epistasis, and equal selection pressures at all sites. Under these assumptions, Sawyer and Hartl (1992) showed that the distribution of frequencies of mutant lineages in a population forms a Poisson random field whose properties depend on the selection pressure and the mutation rate. Hartl *et al.* (1994) and Bustamante *et al.* (2001) developed a maximum-likelihood method of estimating these parameters from data on the polymorphism frequency spectrum. This method has been widely used to study, for example, purifying selection on synonymous (Hartl *et al.* 1994; Akashi and Schaeffer 1997; Akashi 1999) and nonsynonymous (Hartl *et al.* 1994; Akashi 1999) variation, and the evolution of base composition (Lercher *et al.* 2002; Galtier *et al.* 2006).

Closely related to these analyses of polymorphism data are methods that calculate, on the basis of the PRF model, the ratio of the expected number of polymorphisms within species to divergence between species for synonymous and nonsynonymous sites [using the idea behind the McDonald–Kreitman test (McDonald and Kreitman 1991)]. These methods discard some of the available data, as they depend only on the number of polymorphisms and not on their full frequency spectrum. However, they are also less sensitive to assumptions (Sawyer and Hartl 1992; Loewe *et al.* 2006). Such methods have been applied to estimate selection pressures on synonymous variation (Akashi 1995), on nonsynonymous mutations in mitochondrial genomes (Nachman 1998; Rand and Kann 1998; Weinreich and Rand 2000), and on nonsynonymous variation in a variety of nuclear genomes (Bustamante *et al.* 2002; Sawyer *et al.* 2003; Bartolome *et al.* 2005), including humans (Bustamante *et al.* 2005).

Recent theoretical work has focused on relaxing various assumptions of the original PRF method. These include allowing for dominance (Williamson *et al.*

[1]*Corresponding author:* Department of Biology, University of Pennsylvania, Philadelphia, PA 19104. E-mail: jplotkin@sas.upenn.edu

2004), population subdivision (Wakeley 2003), changing population size (Williamson *et al.* 2005), and linkage between sites (Zhu and Bustamante 2005). Several methods for studying the properties of the distribution of selection pressures across sites based on the PRF model have also been developed, using the polymorphism frequency spectrum (Nielsen *et al.* 2005), the ratio of polymorphism to divergence (Sawyer *et al.* 2003; Loewe *et al.* 2006), or several of these methods in conjunction (Bustamante *et al.* 2003; Piganeau and Eyre-Walker 2003; Boyko *et al.* 2008).

A fundamental assumption of the PRF approach is that mutation is irreversible—the infinite-sites assumption. Thus any particular site can be only transiently polymorphic, and there is no steady-state solution to the evolutionary dynamics at any given site. However, since the constant creation of new polymorphic sites is balanced by older polymorphisms fixing or going extinct, the frequency distribution of polymorphisms across sites currently polymorphic does reach an equilibrium. This approach has important advantages. Since it describes the frequency spectrum of new mutations ("derived alleles")—the frequency spectrum relative to the last ancestral state—it provides a natural framework for analyzing polymorphism data when the ancestral state is known. This ancestral state can often be inferred from an appropriate outgroup.

Despite its advantages, the infinite-sites approximation can present problems. The approximation is reasonable for most data currently available from typical eukaryotic populations. However, in many biologically reasonable parameter regimes, particularly those relevant to bacterial and viral populations, more than one mutational event may contribute to polymorphism at a given site. In this article, we show that under these parameter regimes the infinite-sites assumption causes the PRF method to underestimate negative selection pressures and mutation rates by as much as an order of magnitude. In addition, the PRF method often infers that a gene is under strong positive selection when in fact the gene is experiencing weak negative selection. This problem arises both for inferences based on the polymorphism frequency spectrum and for inferences based on the ratio of within-species polymorphism to between-species divergence, but here we focus exclusively on the former.

In this article, we present two methods that relax the infinite-sites assumption of the PRF method, each with its own advantages and drawbacks. Rather than studying mutant lineages across a sequence, our methods focus on explicit models of the evolutionary dynamics at individual sites. We first present a modification of the PRF method that retains the essential framework, but calculates the frequency distribution of mutant lineages at each site rather than across the whole sequence. We next present an alternative method based on well-known diffusion equations in place of the PRF framework. This alternative framework avoids all of the finite-site biases of the PRF, but it cannot make use of knowledge about the ancestral state. Rather than describing a steady-state frequency distribution of derived nucleotides relative to this ancestral state, it describes the frequency distribution of all four nucleotides possible at each site—a fundamentally different steady state. Both of our methods allow us to estimate the selection pressure and the mutation rate from data on the polymorphism frequency spectrum. In addition, these methods also allow us to infer the distribution of selection pressures across sites.

To assess the accuracy of these methods, we generate polymorphism data from simulated Wright–Fisher populations with known selection pressures and mutation rates. By comparing inferences drawn from these simulated data sets, we demonstrate that our methods extend and improve upon the original PRF approach. Throughout this article, we focus on accurately inferring the sign and strength of negative selection, since the most troubling bias in the original PRF method is erroneous inference of positive selection when mutation rates are large. We focus primarily on situations when the ancestral state at each site cannot be reliably inferred, which is the typical situation when the mutation rate is large.

We emphasize that the effects of finite sites are of practical relevance only when the mutation rate is relatively large (θ per site ≥ 0.05). As a result, the methods of inference developed here are not necessary for analyzing human or Drosophila population-genetic data. However, as we shall demonstrate, finite-site effects have significant practical implications when studying the population genetics of viruses, microbes, and some higher eukaryotes, such as sea squirts and starfish, that experience large mutation rates (Drake *et al.* 1998; Lynch and Conery 2003).

THEORY

The Poisson random field model of polymorphisms: We begin by outlining the PRF model of the site-frequency spectrum developed by Sawyer and Hartl (Sawyer and Hartl 1992; Hartl *et al.* 1994). This model assumes that mutations occur in a population of effective size N at a Poisson rate Nu, where u is the *per-sequence* mutation rate, and are all subject to selection of strength s. The fate of each mutant lineage is modeled by a diffusion approximation to the processes of selection and drift. When a new mutant lineage enters the population, it is assumed to arise at a site that has not previously experienced any mutations (the infinite-sites assumption). Each mutant lineage is assumed to be independent of all others (the free-recombination assumption).

Extending earlier work by Wright (1938) and Moran (1959), Sawyer and Hartl (1992) calculated

a steady-state distribution of mutant lineage frequencies. They found that the number of lineages with frequency between x and $x + dx$ is Poisson distributed with mean $f(x)\,dx$, where

$$f(x) = \theta_l \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \frac{1}{x(1-x)}. \quad (1)$$

Here $\gamma \equiv Ns$ is a measure of the strength of selection on the mutant lineages and $\theta_l \equiv 2Nu$ is twice the population per-sequence mutation rate. The function $f(x)$ is referred to as a Poisson random field. In other words, the number of mutant lineages with frequency between x_1 and x_2 is a Poisson random variable with mean $\int_{x_1}^{x_2} f(x)\,dx$. In addition, the number of mutant lineages with frequency in $[x_1, x_2]$ is independent (as a random variable) from the number of mutant lineages with frequency in $[y_1, y_2]$, provided these intervals do not intersect. Note that $f(x)$ is not integrable at $x = 0$. This divergence occurs because the steady state arises from a balance between new mutations constantly occurring and older lineages fixing or going extinct. Thus there is no finite, steady-state expression for the number of lineages that have fixed or gone extinct.

Hartl *et al.* (1994) and Bustamante *et al.* (2001) used Equation 1 as the basis for maximum-likelihood (ML) estimation of the mutation rate θ_l and selection pressure γ from polymorphism data. They imagined sampling n individuals from a population with this steady-state distribution of segregating mutant lineages. They made the infinite-sites assumption that all mutant lineages occur at different sites, consistent with the earlier assumption that each lineage is independent. Since the number of mutant lineages at frequency x in the population is Poisson distributed with mean $f(x)\,dx$, the number of sampled sites containing i mutant nucleotides (we refer to these as i-fold mutant sites) is Poisson distributed with mean

$$F(i) = \theta_l \int_0^1 \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \frac{1}{x(1-x)} \binom{n}{i} x^i (1-x)^{n-i}\,dx. \quad (2)$$

This equation leads immediately to a maximum-likelihood procedure for estimating γ and θ_l (Bustamante *et al.* 2001). A set of sequences from n sampled individuals within a population will contain some number, y_i, of i-fold mutant sites for $0 < i < n$. The set of values $y_1, y_2, \ldots, y_{n-1}$ is called the site-frequency spectrum of the observed data. The probability of a spectrum $\{y_i\}$, given γ and θ_l, is

$$\mathcal{L}_u(\theta_l, \gamma) = \prod_{i=1}^{n-1} e^{-F(i;\gamma,\theta_l)} \frac{[F(i;\gamma,\theta_l)]^{y_i}}{y_i!}. \quad (3)$$

For an observed spectrum $\{y_i\}$ in a particular data set, one maximizes this likelihood over θ_l and γ to estimate the mutation rate and selection pressure.

The likelihood expression above assumes we know which nucleotide is ancestral and which nucleotide is the mutant ("derived") at each polymorphic site. We refer to this situation as the unfolded case. When we do not have this information, we cannot distinguish between an i-fold mutant site and an $(n - i)$-fold mutant site. In this case, a data set will contain some number, y_i, i-fold and/or $(n - i)$-fold mutant sites, where i runs between 1 and the largest integer $\leq n/2$. We refer to this as the folded case. In this situation, the likelihood of a particular data set $\mathcal{L}_f(\theta_l, \gamma)$ is given by the same expression as for $\mathcal{L}_u$, but with the product running from 1 to $n/2$ and with $F(i)$ replaced by $F(i) + F(n-i)$ (except if $i = n/2$).

The infinite-sites approximation: The PRF model makes two key assumptions: that each site is independent of all the others and that two mutant lineages never segregate at the same site. The former assumption is equivalent to assuming free recombination between all segregrating sites, and it has been investigated elsewhere (Akashi and Schaeffer 1997; Bustamante *et al.* 2001); we return to it in the discussion. The main focus of this article, however, is on the second assumption of the PRF, that there are an infinite number of sites, which has not been discussed much in the literature. This assumption can lead to problems that are most apparent when considering how the PRF method treats "multiply polymorphic" sites—those that exhibit more than two types of segregating nucleotides. Polymorphisms of this type are indeed observed in data analyzed by the PRF method (Hartl *et al.* 1994). We refer to the configuration of a particular site as (a, b, c, d), where a, b, c, and d are the numbers of sampled sequences that exhibit each of the four nucleotides. When we have unfolded data, a is the frequency of the ancestral nucleotide and b, c, and d are the frequencies of the three possible mutant nucleotides, in order of decreasing frequency. For folded data, a, b, c, and d are the frequencies of all four possible nucleotides, again in order of decreasing frequency. In the original PRF analysis, a site with a (12, 1, 1, 0) configuration, for example, is treated identically to a site with a (12, 2, 0, 0) configuration. Such a treatment is incorrect: the former configuration can arise only from two mutant lineages, whereas the latter configuration could be caused by a single mutant lineage (presumably at relatively high frequency in the population). Yet the PRF analysis excludes the first possibility and treats both configurations as if they were (12, 2, 0, 0) sites (Hartl *et al.* 1994; Bustamante *et al.* 2001). Similarly, the PRF method treats (10, 2, 2, 0), (10, 3, 1, 0), and (10, 2, 1, 1) sites as if they were in a (10, 4, 0, 0) configuration, etc.

The infinite-sites approximation also affects sites that exhibit only two types of segregrating nucleotides. When multiple lineages are segregating at the same site, an i-fold and a k-fold mutant lineage sampled at the same site can lead to an apparently $(i + k)$-fold mutant site, if the two lineages happen to be mutations to the

same nucleotide. In other words, a (12, 2, 0, 0) site could reflect two low-frequency mutant lineages or one higher-frequency lineage, but the PRF method incorrectly assumes that only the latter is possible. This leads to systematic biases in the estimates of selection and mutation obtained by the PRF method: by disregarding the possibility that an apparently high-frequency mutant lineage is actually several lower-frequency mutant lineages, the PRF method underestimates the mutation rate u and the strength of negative selection $|s|$.

Since a mutant lineage will survive on average $O(\ln[1/|s|])$ generations before fixing or going extinct (Desai and Fisher 2007), and mutations arise at rate $N\mu$ per site, the infinite-sites approximation will be valid only when $N\mu \ln[1/|s|] \ll 1$ (although for sufficiently small samples the condition is weaker, since we may never sample multiple lineages even though they are segregating at the same site). This condition is sometimes violated in real populations, particularly of viruses and microbes (*e.g.*, Hartl *et al.* 1994).

We stress that in parameter regimes relevant to most eukaryotes, including humans and Drosophila, finite-sites biases are negligible. But in parameter regimes relevant to bacteria and viruses, sites with multiple segregrating mutations have a disproportionate weight in estimates of selection and mutation, and thus they can lead to errors of an order of magnitude or more (Figure 1). In such circumstances, the PRF method also erroneously infers positive selection in many situations where selection is actually negative (Figure 1).

A per-site Poisson random field model of polymorphisms: In this section, we describe a method that extends the PRF framework to the case of finite sites and takes full advantage of the information provided by the frequencies of all possible configurations at a site. The basic idea behind this modified approach is to recast the PRF framework on a *per-site* basis. We describe the steady-state frequency distribution of mutant lineages at a given site. From this, we can calculate the probability that a sample of n individuals will contain any configuration of mutants at that site. As in the original PRF method, we retain the assumption of free recombination, so that the DNA sequence is a collection of independent sites. Thus our per-site analysis leads directly to ML estimation of mutation rate and selection strength.

We begin by recasting the PRF expression for the steady-state distribution of mutant lineages to describe the frequencies of mutant lineages at a single site. At a given site, we have

$$f(x) = \theta_s \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \frac{1}{x(1-x)}, \tag{4}$$

where θ_s is the *per-site* value, $\theta_s = 2N\mu$, where μ is the per-site mutation rate. Using this formula to describe multiple lineages at a single site is somewhat peculiar, because this result assumes that all mutant lineages behave independently of one another. Clearly this is not strictly true, since the mutant lineages are segregating at the same site. However, provided two mutant lineages rarely achieve simultaneous high frequencies in the population, then the assumption of independent mutant lineages is a good approximation. This assumption of noninteracting mutant lineages will often hold even when the other aspects of the infinite-sites approximation are violated.

Analogous to the original PRF method, at *a single site* the number of mutant lineages that are observed i times in a sample of n sequences ("i-fold mutant lineages") is Poisson distributed with mean

$$F(i) = \theta_s \int_0^1 \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \frac{1}{x(1-x)} \binom{n}{i} x^i (1-x)^{n-i} dx. \tag{5}$$

On the basis of this, we can calculate the probability of any particular polymorphism configuration at a site.

We begin by describing this calculation in the unfolded case. The probability that a site is monomorphic is just the probability that no i-fold mutant lineages are found at that site, for all i between 1 and $n - 1$. This is

$$P_{\text{mono}} = P_{(n,0,0,0)} = e^{-F(1)} e^{-F(2)} \ \ldots \ e^{-F(n-1)}. \tag{6}$$

We define $M \equiv P_{\text{mono}}$ to denote this monomorphism probability. The probability that a site exhibits a $(n - 1, 1, 0, 0)$ configuration is the probability that a single 1-fold mutant lineage is sampled, and no 2-fold or higher lineages are found,

$$P_{(n-1,1,0,0)} = F(1)M, \tag{7}$$

where M denotes the probability of monomorphism as above. The probability that a site exhibits an $(n - 2, 2, 0, 0)$ configuration is more complex. This configuration could arise from a single 2-fold sampled lineage (as assumed under the standard PRF method) *or* it could arise from two 1-fold sampled mutant lineages that happen to involve mutations to the same nucleotide. Hence its probability is

$$P_{(n-2,2,0,0)} = \frac{F(1)^2}{2!} M \frac{1}{3} + F(2)M, \tag{8}$$

where the factor of $\frac{1}{3}$ is the probability that two mutations result in the same nucleotide. This expression assumes that mutations between all possible nucleotides are equally likely—the obvious generalization applies when there are mutational biases, which we do not discuss further. Similarly the probability of an $(n - 2, 1, 1, 0)$ configuration is

$$P_{(n-2,1,1,0)} = \frac{F(1)^2}{2!} M \frac{2}{3}. \tag{9}$$

The probabilities of more complex configurations can be calculated in a similar way. The probability of

an $(n-4, 4, 0, 0)$ configuration, for example, is the probability of four 1-fold lineages to the same nucleotide plus the probability of two 1-fold lineages and a 2-fold lineage, plus the probability of two 2-fold lineages, plus the probability of one 1-fold lineage and one 3-fold lineage, plus the probability of a single 4-fold lineage. We have

$$P_{(n-4,4,0,0)} = M\left[\frac{F(1)^4}{4!}\frac{1}{3^3} + \frac{F(1)^2F(2)}{2!}\frac{1}{3^2} + \frac{F(2)^2}{2!}\frac{1}{3} + F(1)F(3)\frac{1}{3} + F(4)\right]. \quad (10)$$

In general, the probability of a particular configuration is given by the sum of the probabilities of all possible partitions of n that lead to that configuration.

In the folded case, these calculations become fairly complex. The probability of a (12, 2, 0, 0) folded configuration, for example, includes the probability of a single 12-fold sampled lineage, as well as two 6-fold sampled lineages, and so on. Thousands of terms may arise in the expression for the probability of a particular configuration, even for moderate values of n. We do not quote any of these expressions here, but rather we have developed a computer program to output symbolic expressions for the probabilities of all possible folded as well as unfolded configurations, for a given sample size n. The algorithm used in this program is described in the appendix, and the program is freely available on request.

These probabilities of site configurations form the basis of maximum-likelihood parameter estimation. The probability of a data set with L total sites, including $y_{a,b,c,d}$ sites in configuration (a, b, c, d), is given by

$$\frac{L!}{\prod y_{a,b,c,d}!}\prod P_{a,b,c,d}^{y_{a,b,c,d}}, \quad (11)$$

where the products are over all possible configurations. Given a particular data set, we maximize this probability over θ_s and γ to find the ML estimate of these parameters. In the original PRF method, this maximization is particularly simple, because the ML estimate for θ can be expressed analytically in terms of the ML estimate for γ, leaving a one-dimensional numerical maximization procedure to estimate γ. In our per-site PRF method, however, a full two-dimensional numerical maximization is required to find the ML estimates of θ_s and γ. We have implemented a computer program to perform this maximization; it is freely available on request.

Our per-site PRF method relaxes most of the consequences of the infinite-sites approximation inherent in the original PRF estimation procedure. We allow for the possibility that multiple mutant lineages contribute to the polymorphism observed at a single site. Thus we avoid the systematic underestimation of θ and γ, and incorrect inference of positive selection, that affect the traditional PRF when mutation rates are large (Figure 1). This new method also uses all of the data available in the sample, including the differences between $(n-2, 2, 0, 0)$ and $(n-2, 1, 1, 0)$ sites and the number of monomorphic sites (note the infinite-sites approximation means that the original PRF cannot predict the number of monomorphic sites and hence makes no use of these data). It does still retain one aspect of the infinite-sites approximation: it assumes that mutant lineages segregating at the same site do not interfere with each other. This is never strictly true, but it is a good approximation unless multiple mutant lineages reach high frequency at a given site at the same time. Note that because of this assumption, the probabilities of all possible configurations (a, b, c, d) described above do not precisely sum to unity, because our approach allows a typically small but nonzero probability of multiple mutant processes adding to more than n sampled individuals. The no-interference assumption is substantially weaker than the infinite-sites assumption, and thus our revised sampling method extends the applicability of the PRF framework.

A per-site diffusion model of polymorphisms: In this section, we describe a method that shifts fundamentally from the PRF framework and avoids all of the problems associated with the infinite-sites assumption. Rather than studying the distribution of the frequencies of mutant lineages, we focus on the evolutionary dynamics at each individual site, without keeping track of individual mutant lineages. We develop this into a maximum-likelihood estimation of γ and θ from polymorphism data, which requires neither the infinite-sites nor the no-interference approximation described above. As in the original PRF method, we assume free recombination between sites.

At an individual site, we imagine that one nucleotide is preferred and the other three have the same fitness disadvantage s ($s < 0$). We assume that mutations occur at rate μ and hence at rate $\mu/3$ between any two specific nucleotides (*i.e.*, no mutational biases). These assumptions simplify the discussion, but are not essential. In fact, one advantage of this approach is that these assumptions can be easily relaxed with obvious generalizations (noted below).

We can analyze the process of mutation, selection, and drift at a single site with a three-dimensional diffusion approximation and calculate the joint steady-state probability distribution of the frequencies of the four possible nucleotides at the site. This then leads naturally to the likelihood of any configuration of polymorphism data at the site as a function of γ and θ_s and hence to ML estimation of these parameters from data. Before exploring this fully, we first analyze a simplification of this model, in which we treat all three disfavored nucleotides as a single class. Such a treatment reduces to a standard one-dimensional diffusion process whose steady-state probability distribution describes the frequency of the

preferred nucleotide *vs.* the sum of the frequencies of the disfavored ones; this treatment is essentially a steady-state version of Williamson *et al.* (2005). This simplified model is not likely to be particularly useful, because it requires us to know *a priori* the identity of the preferred nucleotide, which in practice is part of what we wish to infer. This simplification also discards some of the information in the data [*e.g.*, not making use of the difference between (12, 1, 1, 0) and (12, 2, 0, 0) sites]. However, first studying this analytically and computationally simpler model makes the analysis of the full model more clear. We therefore begin by describing the one-dimensional method and then turn to the three-dimensional method.

The one-dimensional diffusion model: We begin by describing a simplified diffusion approach that calculates the frequency distribution of favored *vs.* disfavored nucleotides, similar in spirit to that introduced by Mustonen and Lassig (2007). As noted above, we for simplicity assume that one nucleotide is preferred and the other three nucleotides are disfavored. We denote the sum of the frequencies of the three disfavored alleles by x; the frequency of the preferred nucleotide is $1 - x$.

We assume that mutation, selection, and random drift occur at each site according to standard Wright–Fisher dynamics. Thus the probability distribution of x can be described by the diffusion equation

$$\frac{\partial}{\partial t}f(x,\,t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}[v(x)f(x,\,t)] - \frac{\partial}{\partial x}[m(x)f(x,\,t)], \quad (12)$$

where $f(x,\,t)$ is the probability that the disfavored nucleotides have frequency x at time t and $m(x)$ and $v(x)$ are given by

$$m(x) = sx(1-x) + \mu(1-x) - \frac{\mu}{3}x \quad (13)$$

$$v(x) = \frac{x(1-x)}{N}, \quad (14)$$

where s is the selection coefficient against the disfavored nucleotides ($s < 0$) and μ is the per-site mutation rate per individual per generation. This diffusion equation is well known (Ewens 2004) and has the steady-state solution derived by a zero-flux boundary condition:

$$f(x) = Cx^{\theta_s - 1}(1-x)^{\theta_s/3 - 1}e^{2\gamma x}. \quad (15)$$

Here C is a (θ_s- and γ-dependent) normalization factor, and as before $\theta_s \equiv 2N\mu$ and $\gamma \equiv Ns$.

If the frequency of disfavored nucleotides at a site equals x, the probability that we find i such nucleotides in a sample of n individuals is $\binom{n}{i}x^i(1-x)^{n-i}$. Averaging over x, the overall probability that we sample i disfavored nucleotides at a given site is

$$F(i) = \binom{n}{i}\int_0^1 Cx^{\theta_s + i - 1}(1-x)^{\theta_s/3 + n - i - 1}e^{2\gamma x}\,dx. \quad (16)$$

This integral, including the calculation of the normalization factor C, can be solved analytically. We find

$$F(i) = \binom{n}{i}\frac{\Gamma(n-i+\theta_s/3)\Gamma(i+\theta_s)\,{}_1F_1(i+\theta_s,\,n+4\theta_s/3,\,2\gamma)}{\Gamma(\theta_s/3)\Gamma(\theta_s)\,{}_1F_1(\theta_s,\,4\theta_s/3,2\gamma)}, \quad (17)$$

where Γ is Euler's Gamma function and ${}_1F_1$ is a hypergeometric function.

The expression above leads immediately to a maximum-likelihood method for estimating γ and θ_s in the unfolded case. Note, however, that in contrast to the original PRF method, this model describes a steady state of the frequency distribution of preferred relative to unpreferred nucleotides, rather than derived *vs.* ancestral nucleotides. This is a very different sort of steady state, a point to which we return in more detail below. Since this model refers only to the preferred and unpreferred states at each site, it makes sense to apply to unfolded data only if we assume that the ancestral state is preferred. In a sample of n sequences each of length L, we count the number of sites at which i disfavored nucleotides are sampled, y_i, for $0 \le i \le n$. Since all sites are assumed independent, each with the polymorphism frequency distribution described above, the likelihood of the data given the parameters is

$$\mathcal{L}_u(\theta_s,\,\gamma) = \frac{L!}{\prod_{i=0}^n y_i!}\prod_{i=0}^n F(i)^{y_i}. \quad (18)$$

For any set of polymorphism data, it is straightforward to maximize this function numerically, producing ML estimates of θ_s and γ. As with the per-site PRF method, this procedure involves a two-dimensional maximization routine. Simulated data and fits confirm that this method produces accurate and unbiased estimates of mutation rate and selection strength (data not shown).

When we do not know which nucleotide is preferred at each site, which will be typical in most microbial applications, we must use the "folded" version of the data. This presents difficulties. Imagine a site with an (a, b, c, d) polymorphism configuration. We might naively suppose that since any of the four nucleotides could be the preferred one, the probability of these data is simply $F(b+c+d) + F(a+c+d) + F(a+b+d) + F(a+b+c)$. However, this is not the case. For example, if a is indeed the preferred nucleotide, then $F(b+c+d)$ does not equal the probability that the three disfavored nucleotides will form a (b, c, d) configuration. Rather, it equals the sum of the probabilities that the three disfavored nucleotides will form a configuration (i, j, k), summed over all i, j, k triplets that sum to $b+c+d$. This is a serious problem, because the difference between $F(b+c+d)$ and the probability of the data depends on $b+c+d$, and hence it is not identical for all four possible preferred nucleotides. Simply assuming that the most common

nucleotide is the preferred one (Hartl *et al.* 1994) is a reasonable approach. But this approach will be inaccurate for sites where the most common nucleotide is not overwhelmingly so; when this situation describes a substantial fraction of sites, the method will fail. As a result, the one-dimensional diffusion framework does not allow for a rigorous ML estimate of parameters with folded data.

To perform rigorous ML fits to folded frequency data we must turn to a three-dimensional diffusion method.

The three-dimensional diffusion model: Rather than considering all disfavored nucleotides as a single class, we can instead keep track of the evolutionary dynamics of all four possible nucleotides at a site. To do so, we assume the standard four-allele Wright–Fisher dynamics, with mutation at rate $\mu/3$ between any two particular nucleotides, and selection acting with strength s against the three disfavored nucleotides. The dynamics can then be described by a three-dimensional diffusion equation for the joint distribution of the frequencies of the three disfavored alleles x_1, x_2, and x_3, $f(x_1, x_2, x_3, t)$ (where the preferred allele has frequency $x_0 = 1 - x_1 - x_2 - x_3$). We have

$$\frac{\partial}{\partial t} f(x_1, x_2, x_3, t) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\partial^2}{\partial x_i \partial x_j}[V_{ij} \times f] - \sum_{i=1}^{3}\frac{\partial}{\partial x_i}[M_i \times f], \quad (19)$$

where

$$M_i = (1+s)x_i\left(1 - \sum_{j=1}^{3} x_j\right) + \frac{\mu(1-4x_i)}{3} \quad (20)$$

$$V_{ii} = \frac{x_i(1-x_i)}{N} \quad (21)$$

$$V_{ij} = -\frac{x_i x_j}{N}, \quad (i \neq j). \quad (22)$$

This is a somewhat less well-known diffusion equation (Wright 1949; Watterson 1977); the steady-state solution is

$$f(x_1, x_2, x_3) = C[x_1 x_2 x_3 (1 - x_1 - x_2 - x_3)]^{\zeta} e^{2\gamma(x_1 + x_2 + x_3)}, \quad (23)$$

where C is a normalization factor, and we have defined $\zeta = \theta_s/3 - 1$.

Given the frequencies x_0, x_1, x_2, and x_3 of nucleotides in the population, the probability of sampling a site in an *unordered* configuration (n_0, n_1, n_2, n_3) in a sample of n individuals (adopting the convention that the first nucleotide listed is the preferred one) is just the multinomial probability

$$\frac{n!}{n_0! n_1! n_2! n_3!}(1 - x_1 - x_2 - x_3)^{n_0} x_1^{n_1} x_2^{n_2} x_3^{n_3}. \quad (24)$$

Averaging over f, we therefore find that the probability of sampling a site in an unordered (n_0, n_1, n_2, n_3) configuration is

$$P_{n_0,n_1,n_2,n_3} = C\frac{n!}{n_0! n_1! n_2! n_3!}\int_0^1\int_0^{1-x_1}\int_0^{1-x_1-x_2} \times e^{2\gamma(x_1+x_2+x_3)} x_1^{\zeta+n_1} x_2^{\zeta+n_2} x_3^{\zeta+n_3}[1 - x_1 - x_2 - x_3]^{\zeta+n_0} \times dx_3\, dx_2\, dx_1. \quad (25)$$

(26)

If we have unfolded data and wish to assume that the ancestral state is preferred, we can use unfolded ML inference. In this case, the probability of finding a site in an *ordered* unfolded configuration (a, b, c, d) (where by convention a is the number of individuals that have the preferred nucleotide and $b \geq c \geq d$) is

$$P^{\mathrm{u}}_{a,b,c,d} = \sum_{\{n_0,n_1,n_2,n_3\}} P_{n_0,n_1,n_2,n_3}, \quad (27)$$

where the sum is over all unordered configurations (n_0, n_1, n_2, n_3) that give rise to the ordered unfolded configuration (a, b, c, d).

If on the other hand we do not have a reliable outgroup (or choose to ignore the information from an outgroup on the ancestral state), we can use folded ML inference. Here, the probability of sampling a site in the ordered configuration (a, b, c, d) is just the sum of the probabilities assuming that each of the four possible nucleotides is preferred. As before, we adopt the convention that ordered folded configurations are written as (a, b, c, d) with $a \geq b \geq c \geq d$. The probability of a folded configuration $P^{\mathrm{f}}_{a,b,c,d}$ is then

$$P^{\mathrm{f}}_{a,b,c,d} = \sum_{\{n_0,n_1,n_2,n_3\}} P_{n_0,n_1,n_2,n_3}, \quad (28)$$

where in this case the sum is over all unordered configurations that give rise to the ordered folded configuration (a, b, c, d).

For either folded or unfolded data, given n samples of a sequence L sites long, with $y_{a,b,c,d}$ sites in an (a, b, c, d) polymorphism configuration, the likelihood of the data is

$$\mathcal{L}(\theta_s, \gamma) = \frac{L!}{\prod y_{a,b,c,d}!}\prod [P_{a,b,c,d}]^{y_{a,b,c,d}}, \quad (29)$$

where the products are taken over all possible configurations (a, b, c, d) and $P_{a,b,c,d}$ is the folded or unfolded probability defined above. We can numerically maximize this function to find the ML estimates of θ_s and γ. We have implemented a computer program to perform this maximization; it is freely available on request.

Variable selection pressures across sites: Both the original PRF method and the per-site methods we have proposed in this article assume that all sites experience the same selective pressure. In reality, we expect that

there is some distribution of selective pressures across sites. Hartl *et al.* (1994) suggest that in this case the ML estimate of γ from their PRF method reflects a weighted average selection pressure across the sites, but the nature of this weighting is not well understood. Almost no weight is given to sites at which $|\gamma| \gg 1$, because these sites will likely be monomorphic and hence ignored by the original PRF method. It is unclear how sites with different values of γ of order 1 will be weighted or how the presence of some effectively neutral sites ($|\gamma| \ll 1$) will change the ML estimate. Piganeau and Eyre-Walker (2003), Nielsen *et al.* (2005), and (Boyko *et al.* 2008) have addressed these questions by introducing procedures that allow for inference of some aspects of the distribution of selective coefficients across sites within the PRF framework. In this section we describe a similar generalization in our models.

The issue of variable γ across sites is of particular concern for the methods we have proposed, because these methods make use of the monomorphism data. If some number L_1 of sites are effectively lethal (*i.e.*, have $|\gamma| \gg 1$), these sites will all be monomorphic, which will tend to depress our ML estimate of θ_s and increase our estimate of $|\gamma|$. Fortunately, our methods are able to use the monomorphism data to investigate the number of lethal sites, L_1, or more generally the full distribution of selection pressures across sites.

Since all of the methods we have proposed are defined at a per-site level, it is straightforward to assume that there are multiple different classes of sites with different values of γ. We can posit that there are k classes of sites. Each class is represented by L_j of L total sites and has its own value of γ, which we call γ_j. The probability that a site is in an (a, b, c, d) configuration is then

$$P_{a,b,c,d} = \sum_{j=1}^{k} \frac{L_j}{L} P^j_{a,b,c,d}(\gamma_j, \theta_s), \tag{30}$$

where $P^j_{a,b,c,d}$ (γ_j, θ_s) is the probability that a site with parameters γ_j and θ_s is in the configuration (a, b, c, d). This expression is correct for both the per-site PRF and the diffusion approaches.

Given our new definition of $P_{a,b,c,d}$, we can construct the folded or unfolded likelihood of the overall polymorphism data set in exactly the same way as before. This likelihood function now depends on $2k + 1$ parameters: θ_s, the γ_j, and the L_j. We can find ML estimates of all of these parameters, using a multidimensional numerical maximization of the likelihood function. By choosing k, we determine the resolution at which we measure the distribution of values of γ across sites. Naturally, the larger the k we choose, and hence the greater the resolution on γ, the more data we require to obtain accurate estimates of the individual L_j and γ_j.

Rather than estimating both the L_j and the γ_j, we could instead posit that there are several classes of mutations with different *prespecified* γ_j and estimate only the values of L_j. In other words, we ask what fraction of sites have different values of selective constraints. We describe here one particularly important example of a hybrid between these two procedures, with two classes of sites ($k = 2$). Rather than fitting an ML estimate of γ to both classes, we assume that one class of sites is unable to evolve: mutations at these sites are lethal (more precisely, they have $|\gamma| \gg 1$). We wish to calculate the number of lethal sites and the average selective pressure on the remaining, nonlethal sites. Thus we have three parameters: the mutation rate θ_s, the number of lethal sites L_2, and the strength of selection γ acting at the other $L_1 = L - L_2$ sites. The probability that a site is monomorphic is given by

$$P_{\text{mono}} = \frac{L_2}{L} + \frac{L_1}{L} P^1_{\text{mono}}. \tag{31}$$

Here P^1_{mono} is the probability that a site with strength of selection γ and mutation rate θ_s will be monomorphic, as defined by either the per-site PRF or the per-site diffusion approach (whichever method we are using). The probability that a site is in a nonmonomorphic (a, b, c, d) configuration is

$$P_{a,b,c,d} = \frac{L_1}{L} P^1_{a,b,c,d}. \tag{32}$$

Now we can write the likelihood of the data in the usual way. This results in a three-dimensional ML problem. However, we can simplify the problem by first maximizing L_2 given γ and θ. We find that the ML estimate of L_2 is

$$\hat{L}_2 = \frac{L_{\text{mono}} - L P^1_{\text{mono}}}{1 - P^1_{\text{mono}}}, \tag{33}$$

where L_{mono} is the number of monomorphic sites in the data and P^1_{mono} is the probability a nonlethal (*i.e.*, an L_1) site is monomorphic. Substituting this value for L_2, we are left with a two-dimensional maximization problem in γ and θ_s, similar to the original situation.

It is worth exploring how this procedure for estimating the number of lethal sites utilizes the data. It turns out that this procedure is equivalent to ignoring the monomorphism data when finding the maximum-likelihood estimates of γ and θ_s for the nonlethal sites. The likelihood of the data ignoring monomorphic sites is

$$\mathcal{L}(\theta_s, \gamma) = \frac{y_p!}{\prod y_{a,b,c,d}} \prod \left[\frac{P^1_{a,b,c,d}}{1 - P^1_{\text{mono}}} \right]^{y_{a,b,c,d}}, \tag{34}$$

where y_p is the total number of nonmonomorphic sites and the products are over all configurations of nonmonomorphic sites. After finding ML estimates of γ and θ from this monomorphism-ignoring likelihood function, the procedure then calculates the number of monomorphic sites that would be expected given γ and θ_s. This is $L_1 P^1_{\text{mono}} = (L - L_2) P^1_{\text{mono}}$. We then

estimate the number of lethal sites L_2 as the difference between the observed number of monomorphic sites L_{mono} and the number that would be predicted if all sites were of the L_1 variety, $\hat{L}_2 = L_{\text{mono}} - (L - \hat{L}_2)P^1_{\text{mono}}$. Rearranging this expression, we see that it is identical to Equation 33 above. And indeed, plugging Equations 31–33 into Equation 30 yields Equation 34.

Thus, this procedure ignores the monomorphism data when calculating γ and θ_s (at the nonlethal sites), and it instead uses the monomorphism data to infer one aspect of the distribution of γ across sites—specifically, the number of lethal sites. Since the original PRF method also ignores monomorphism data, we obtain this information on the distribution of γ for "free," relative to the power of the original method, simply by shifting to the per-site model. If desired, we can also posit that there are a number of sites L_3 that are effectively neutral (*i.e.*, with $|\gamma| \ll 1$) and estimate L_3. This would devote some part of the data describing polymorphisms at intermediate frequencies to estimating L_3. From this procedure we could estimate the number of effectively lethal sites, the number of effectively neutral ones, and the "weighted average" selection pressure acting on the remaining sites. If more resolution is desired, and enough data are available, we can increase the number of classes of sites and obtain ML estimates of the numbers of sites in each class and the selection pressure acting on each class.

METHODS

Simulations and fits: We used Wright–Fisher simulations to test the inferential accuracy of the PRF method as well as the accuracy of our two alternative methods. The Wright–Fisher model (or, more precisely, its diffusion limit) forms the basis of the PRF method, and it is therefore the appropriate simulation framework for testing the method.

All simulations assumed a constant population of $N = 1000$ haploid individuals. Each of $L = 1000$ sites, simulated independently, could assume one of four states: a, c, t, or g. One state is assigned fitness 1, and the other three states fitness $1 + s$ (where $s < 0$). Mutations occurred at rate μ per site. The allele frequencies evolved according to the standard Wright–Fisher Markov chain (Ewens 2004). Each simulation was run for at least $10/\mu$ generations, to ensure relaxation to steady state. At the end of the simulation, $n = 14$ individuals were sampled from the population and the polymorphism frequency spectrum was recorded. We chose to consider samples of size $n = 14$ to facilitate comparison with Hartl *et al.* (1994). This choice does affect the relative accuracy of the techniques; the finite-sites biases in the PRF method increase with n (see results). We have focused our fits on the case of folded frequency spectra, which will typically be the only reliable type of polymorphism data available when θ_s is large.

We performed simulations over a wide range of parameter values relevant to viruses and microbes. We considered five different values of θ_s: 0.05, 0.1, 0.5, 1.0, 5.0. For each value of θ_s, we performed one simulation at each of 17 different values of γ, ranging from $\gamma = -10.0$ to $\gamma = -0.1$. For each set of simulation parameters (γ, θ_s), once the simulated folded polymorphism data had been generated, ML parameter estimates ($\hat{\gamma}$, $\hat{\theta}_s$) were obtained by numerical maximization of the likelihood function, as specified by the original PRF model, the per-site PRF model, or the three-dimensional diffusion model. This maximization can be difficult in the diffusion case (see below). A 95% confidence interval for γ was constructed according to Bustamante *et al.* (2001): the interval includes those values of γ within $0.5\chi^2_{1,0.95}$ log-likelihood units from $\hat{\gamma}$ (note these confidence intervals rely on the assumption of no linkage; see discussion). The estimated parameters shown in Figure 1 are somewhat "jagged," because the inference methods have been applied to a single draw of $n = 14$ sequences for each set of simulation parameters, as opposed to averaging over many such draws. Figure 1 also shows fits obtained from inference techniques applied to data generated under the infinite-sites model at $\theta_s = 0.01$ (*i.e.*, deviates drawn from a Poisson distribution with mean given by Equation 1).

As discussed above, the original PRF model disallows multiple mutant lineages at a site. Therefore, when a site sampled from the simulated data exhibited more than two types of segregating nucleotides, the frequencies of all unpreferred nucleotides were summed to represent the frequency of the "mutant" type for the purposes of fitting using the original PRF model, as suggested by Hartl *et al.* (1994). When fitting folded data using the PRF method, the most common nucleotide was assumed to be the ancestral type, as suggested by Hartl *et al.* (1994). This approach is not entirely accurate, as discussed above, but it is probably the best option available within the original PRF framework.

Numerical maximization procedure: In practice, the ML estimation procedure for the diffusion model is difficult to implement, because of the triple integral in the definition of P_{n_0,n_1,n_2,n_3} (as well as that implicit in the definition of the normalization constant C). This integral cannot be solved exactly, and it is difficult to evaluate numerically because the integrand may diverge (though the integral itself converges) near the boundary of the simplex over which it is integrated. We adopt a hybrid method to simplify the evaluation of this triple integral. Near the boundary of the simplex, we Taylor expand the integrand and integrate it analytically. Away from the boundary, the integrand is well behaved and standard numerical integration has no difficulties.

This approach is most easily achieved by making the substitutions $y = x_3/(1 - x_1 - x_2)$ and $z = x_2/(1 - x_1)$. On doing so, we can rewrite the integral as

$$P_{n_0,n_1,n_2,n_3} = C\frac{n!}{n_0!n_1!n_2!n_3!}\int_0^1\int_0^1\int_0^1 \times \exp[2\gamma(x+y+z-xy-zy-xz+xyz)] \times x^{n_1+\zeta}z^{n_2+\zeta}y^{n_3+\zeta} \times (1-x)^{n_0+n_2+n_3+3\zeta+2}(1-z)^{n_0+n_3+2\zeta+1}(1-y)^{n_0+\zeta}. \quad (35)$$

This expression is much easier to handle than our original expression, because the three integrals can be done in arbitrary order. We divide each of the three integrals into three pieces: one from 0 to δ, one from δ to $1-\delta$, and one from $1-\delta$ to 1. Thus the triple integral is split into 27 total terms. For each of the integrals from 0 to δ or $1-\delta$ to 1, we Taylor expand the integrand in the integration variable, and we solve the integral analytically. All of the remaining integrals, from δ to $1-\delta$, are done numerically. We must choose δ large enough that we can perform the numerical integrals quickly, but not so large that the Taylor expansions used for the analytical parts become invalid. For these Taylor expansions, we need $\delta \ll 1/2\gamma$, $\delta \ll 1$, $\delta \ll 1/(3\zeta+2)$, $\delta \ll 1/(2\zeta+1)$, and $\delta \ll 1/\zeta$. For the computational analysis described in this article, we choose whichever of these conditions is most restrictive and set δ to be one-tenth of the most restrictive requirement. We find that this choice of δ is sufficiently small to provide accuracy in the analytical parts of the integrals, but large enough to enable quick numerical integration on the interior of the simplex.

We have written a computer program implementing this numerical integration and the resulting ML estimation for the diffusion model. This program and the simpler programs for ML estimation using the one-dimensional diffusion model and the per-site PRF are freely available on request.

RESULTS

Using data from the Wright–Fisher simulations described above, we tested the inferential accuracy of the original PRF method, our per-site PRF method, and the three-dimensional diffusion method. Figure 1 compares the accuracy of selection pressures estimated from folded data using each of these three methods. For large θ_s, the original PRF method systematically underestimates the strength of negative selection, by as much as a factor of 10. In addition, the PRF method often erroneously infers strong positive selection when in fact mutants are under negative selection. These problems are more severe when the mutation rate is large. The smallest mutation rate at which such problems occur ($\theta_s = 0.05$) is four times smaller than the mutation rate estimated for bacterial genes (Hartl *et al.* 1994). The per-site version of the PRF method that we have developed improves upon the standard PRF method, but it too exhibits systematic biases, especially when selection is weak and the mutation rate large (Figure 1). The one-dimensional diffusion method provides accurate and unbiased estimates of γ over the full range of selective pressures and mutation rates (not shown). Like its one-dimensional counterpart, the three-dimensional diffusion method also provides accurate and unbiased estimates over the full range of simulated parameters (Figure 1). When selection is weak (*i.e.*, $|\gamma| < 1$), however, the confidence intervals on diffusion-based estimates of γ are appreciably larger. This behavior makes perfect sense: when selection is nearly neutral and the ancestral state is unknown, the frequency distribution does not exhibit sufficient skew to deduce the preferred nucleotide. As a result, the diffusion-based estimator cannot distinguish between weak positive and weak negative selection in the absence of information on the preferred nucleotide. Thus, the confidence intervals obtained under the folded diffusion technique properly reflect our inability to estimate the selection pressure precisely when selection is weak.

As shown in Figure 1, when selection is weakly negative, the original PRF method erroneously infers positive selection. This problem occurs in the parameter regimes that have been estimated from biological data from bacterial and viral populations. For example, on the basis of $n = 14$ sampled sequences each 367 sites long, Hartl *et al.* (1994) estimated $\gamma = -1.34$ and $\theta_s = 0.183$ for silent sites in a bacterial gene. If we simulate 367 Wright–Fisher sites under these parameters and sample $n = 14$ sequences, we find that the most likely parameters fitted using the original PRF method are $(\hat{\gamma}, \hat{\theta}_s) = (+18.45, 0.067)$. That is, in this example using estimated microbial parameter values, the PRF method is strongly biased.

Figure 2 shows the accuracy of estimated mutation rates using the original PRF method, the per-site PRF method, and the three-dimensional diffusion method. The original PRF method systematically underestimates the mutation rate. Estimates obtained using the per-site PRF method are an improvement, but still exhibit some biases. The diffusion-based method provides accurate and unbiased estimates of the mutation rate, across the full range of mutation rates and selection pressures (Figure 2).

In addition to the simulations described above, we also fitted data that had been generated under the infinite-sites model at a small mutation rate, $\theta_s = 0.01$ (Figure 1). In this case, we generated a sampled polymorphism frequency spectrum using Poisson deviates with mean given by Equation 1, for $n = 14$ samples of $L = 10{,}000$ sites. When applied to such data, inferences based on the original PRF method must be unbiased. Figure 1 shows that inferences based on the modified per-site PRF method and the diffusion method are also accurate and unbiased when applied to such data. Thus, the new methods we have developed here perform as

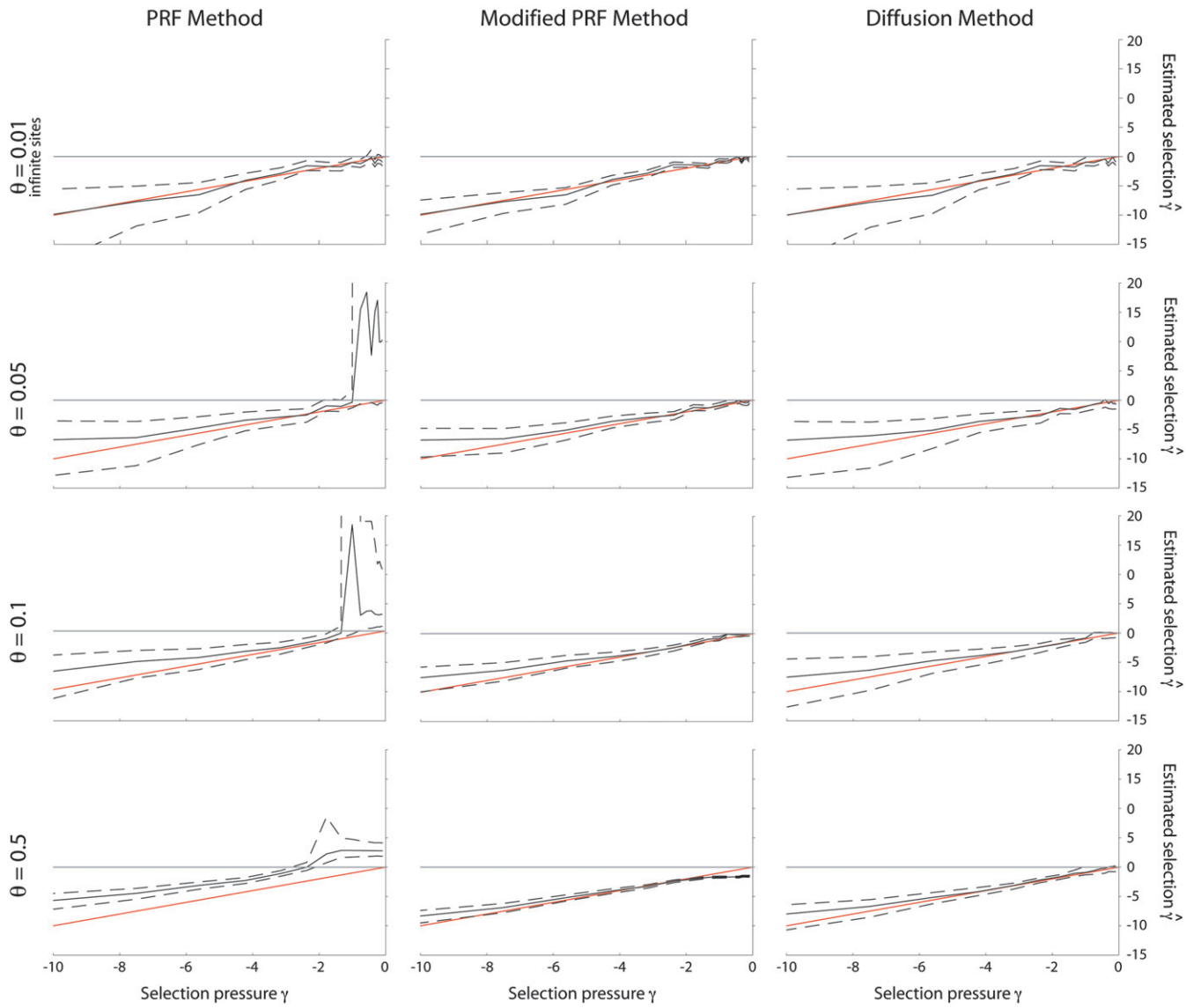

FIGURE 1.—Maximum-likelihood estimates of selection pressures obtained under the PRF method, the modified (per-site) PRF method, and the three-dimensional diffusion method. Selection pressures were estimated from the folded polymorphism frequencies among $n = 14$ sequences sampled from a simulated Wright–Fisher population. In each panel, the simulated selection pressure γ is shown on the *x*-axis and the estimated selection pressure $\hat{\gamma}$ on the *y*-axis. Dashed lines indicate 95% confidence intervals around the ML estimates obtained from the χ^2-distribution (see METHODS). The line $\hat{\gamma} = \gamma$ is shown in red, and the line $\hat{\gamma} = 0$ is shown in gray. The diffusion method and, to a lesser extent, the modified PRF method correct the biases inherent in the original PRF method. When selection is weakly negative, the diffusion method cannot reject positive selection on the basis of folded data, as indicated by the lack of the upper confidence interval for some of the fits. The simulated data in the top panels, $\theta = 0.01$, were generated according to the infinite-sites model (see METHODS).

well as the original PRF method under the assumption of infinite sites and at small mutation rates.

The methods we have developed in this article also allow us to estimate the distribution of selection pressures across sites. In one simple case discussed above, we have presented a procedure for estimating the number of lethal sites and the selective pressure operating on the remaining, nonlethal sites in a gene. This procedure involves estimating γ and θ_s on the basis of polymorphic sites alone and thereafter estimating the proportion of observed monomorphic sites that are lethal. To assess the power and accuracy of this approach, Table 1 shows the error in the predicted number of monomorphic sites in each of our simulations, compared to the number of monomorphic sites actually observed. Across a large range of selective pressures and mutation rates, this approach typically estimates the number of (nonlethal) monomorphic sites within a few percent. As a result, for a gene of length $L = 2000$ sites, at one-half of which mutations are strongly deleterious ($|s| \gg 1/N$), our procedure will accurately predict the number of lethal sites within a few percent, and it will accurately

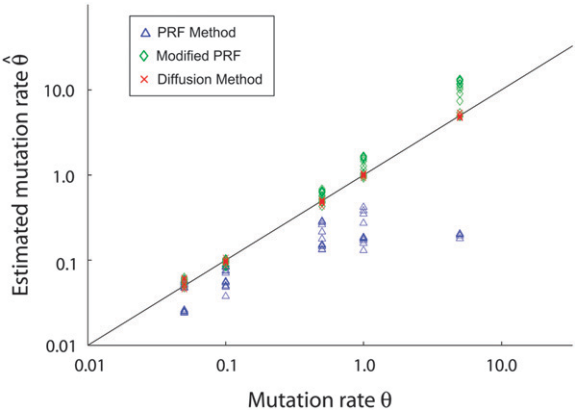

FIGURE 2.—Maximum-likelihood estimates of the mutation rate, $\theta_s = 2N\mu$, obtained under the PRF method (blue), the modified (per-site) PRF method (green), and the diffusion methods (red). Mutation rates were estimated from the folded polymorphism frequencies among $n = 14$ sequences sampled from a simulated Wright–Fisher population. The simulated mutation rate θ_s is shown on the x-axis and the estimated mutation rate $\hat{\theta}_s$ on the y-axis. The line $\hat{\theta}_s = \theta_s$ is shown in black. For each value of θ_s, simulations and fits are shown for 17 different values of γ, ranging from $\gamma = -10.0$ to $\gamma = -0.1$. The PRF method systematically underestimates the mutation rate, especially when selection is weak. The diffusion method provides accurate and unbiased estimates of the mutation rate across the full range of parameters.

predict the selection pressure on the remaining, non-strongly deleterious sites.

All of the simulation results that we have presented thus far have been for the case $n = 14$. This choice affects the relative accuracy of the methods. As n increases, the probability of seeing a high-frequency polymorphism within the sample depends more strongly on the probability that a site has a high-frequency polymorphism in the population, $f(x)$ for large x. This is precisely the part of frequency distribution that is artificially suppressed by the infinite-sites assumption of the PRF method. Thus, as the sample size increases the biases in the PRF method should become more severe. We have tested this expectation by sampling $n = 100$ simulated Wright–Fisher sequences. In the case of $\theta_s = 0.1$, for example, when fitting unfolded polymorphism frequency spectra measured relative to the (known) ancestral state at each site, the PRF method always rejects the true value of γ in favor of a weaker or even positive selection coefficient, throughout the range $\gamma \in [-10, -0.5]$. Inferential biases associated with large sample sizes and mutation rates will become increasingly important as the availability of sequence data improves—in metagenomic surveys of microbial populations, for example.

DISCUSSION

The Poisson random field model (SAWYER and HARTL 1992) and the associated likelihood procedure for estimating parameters (HARTL *et al.* 1994) are perfectly valid when the assumptions underlying the method are met—namely, infinite sites, free recombination, and constant selective pressure across sites. Moreover, the assumption of infinite sites is perfectly reasonable for most eukaryotic populations. It will hold whenever $N\mu \ln[1/|s|] \ll 1$; as we have seen, a rough rule of thumb is that this tends to be true when $\theta_s < 0.05$. We have shown, however, that when the mutation rate is relatively large the PRF method can lead to incorrect inferences. In practice, such concerns arise when studying populations of viruses, microbes, and those eukaryotes that experience large mutation rates ($\theta_s > 0.05$). We have developed two new methods that relax or remove the infinite-sites assumption. These new methods also extend the types of inferences that can be drawn from polymorphism data to include inferences on the distribution of selective pressures across sites.

It may seem surprising that the infinite-sites approximation can lead to substantial errors in the mutation rates and selective strengths inferred by the PRF method. After all, sites at which multiple mutant lineages are sampled are presumably very rare (*e.g.*, HARTL *et al.* 1994). Yet despite their rarity, these sites have a large impact on maximum-likelihood estimation of γ. Because γ enters the likelihood function in the factor $e^{2\gamma x}$, changing γ has a much larger impact on the likelihood of sites with many mutant nucleotides than those with few. In other words, a single site with a high frequency of mutant nucleotides is very strong evidence for positive selection or low $|\gamma|$, whereas a site with a low frequency of mutant nucleotides is not very strong evidence for the opposite. Thus even a few sites at

TABLE 1

Accuracy of the estimated number of monomorphic sites

Actual θ_s	Average error (%)	Median error (%)
0.05	7.6	7.8
0.1	3.3	1.8
0.5	2.3	1.1
1.0	1.4	0.5
5.0	0.1	0.1

Accuracy of estimates for the expected number of monomorphic sites. We calculated maximum-likelihood estimates of γ and θ_s using the one-dimensional diffusion method applied to unfolded simulated data excluding monomorphic sites. From these values, we calculated the expected number of monomorphic sites. Shown are the differences between the observed and the expected number of monomorphic sites (in a simulated gene of length $L = 1000$ sites). For each value of θ_s, we show both the average and the median differences in numbers of monomorphic sites, across simulations with γ ranging from -0.1 to -10. These results imply that we can accurately estimate the number of sites that are monomorphic due to drift. Thus, if a gene contains a similar or larger number of lethal sites, we can also estimate the number of such lethal sites to within the above accuracy.

which multiple mutant lineages are sampled can cause large inaccuracies in the inferred γ when they are incorrectly assumed by the original PRF method to be the result of a single high-frequency mutant lineage. These inaccuracies in γ then force corresponding inaccuracies in the inferred θ_s.

Previous simulation studies have not observed these problems with the PRF method (Bustamante *et al.* 2001). However, such simulations themselves implicitly assumed infinite sites, and hence they cannot be used to test this aspect of the PRF method (to be fair, these simulations were done with typical eukaryotic populations in mind, where these problems are much less severe). In this article, by contrast, we have simulated a finite number of sites that evolve according to the Wright–Fisher model.

The role of linkage: While our focus has been on the infinite-sites approximation, both the original PRF and all of our methods make another key assumption: that each site is independent of all the others (*i.e.*, free recombination). This is essentially an assumption of linkage equilibrium between all polymorphic sites. This potentially crucial assumption is likely to be violated in many real populations, particularly when the method is applied to estimate selective pressure on short stretches of DNA (*e.g.*, a single gene) that are linked over long timescales. This is problematic for both the original PRF and our methods, but it may be particularly relevant in populations in which finite-sites issues are important, because recombination is presumably less common in bacteria than in eukaryotes and because higher θ_s implies a higher density of polymorphic sites, and hence tighter linkage. Ideally, we would address this assumption using theory that allows us to infer the strength of selection acting on a large number of sites with an arbitrary degree of linkage. However, this is a formidable challenge, and no such theory yet exists.

Despite this, models that assume free recombination are still useful for several reasons. First, when selection pressures are weak, sites segregate over long timescales, so recombination may be frequent enough even among intragenic sites (or in bacterial genomes) that segregating sites are unlinked over these timescales. Since both PRF and diffusion-type methods are primarily useful at estimating selection pressures when they are of order $1/N$, the selection pressures our methods can resolve get smaller as θ_s increases. Thus while populations with large θ_s may tend to have lower recombination rates among polymorphic sites, the polymorphic sites we are sensitive to segregate over longer timescales, and recombination has longer to act. Whether this means that linkage can be neglected in any particular population is an empirical question, but fortunately it is usually straightforward to estimate linkage disequilibria directly from the data. Thus in any particular case we can estimate whether or not the assumption of free recombination is valid for the purposes of the original PRF or our modified methods, before applying either approach. When there is no linkage disequilibrium in the data, we can use our methods with confidence. When on the other hand it appears from the data that free recombination is not a reasonable assumption, a model that assumes no linkage may still be useful as a null model and a limiting case, and it can be compared with more complicated possibilities.

Finally, it is important to note that linkage between segregating sites will not bias estimates of the selection pressure γ, provided γ is equal across sites (Akashi and Schaeffer 1997; Bustamante *et al.* 2001). Rather, the primary effect of linkage is to increase the variance in such estimates above the predictions of our method (Bustamante *et al.* 2001). Thus while the predictions of our models (or the original PRF) will tend to have smaller confidence intervals than they should, the estimated values of γ and θ_s should not be systematically biased. It is important to note, however, that when linkage is important we must be cautious about testing hypotheses using these methods; underestimates of the confidence intervals could, for example, lead us to erroneously reject a neutral null. Further, when selection pressures vary significantly across sites, the mean parameter estimates themselves could be biased; this more complex situation remains an important topic for future exploration.

Comparison between the PRF and diffusion methods: Our per-site version of the PRF model relaxes some aspects of the infinite-sites approximation, but it still assumes the mutant lineages do not interact. The diffusion approach removes the infinite-sites approximation altogether. Thus, the diffusion approach contains none of the biases associated with infinite-sites approximation associated with the traditional PRF and, to a lesser degree, the per-site PRF.

The diffusion method is also easily extendable to more complex evolutionary situations. For example, we can explore different selective costs for different nucleotides, or more than one preferred nucleotide. These possibilities lead to obvious modifications of the diffusion equations and their solutions and hence to the maximum-likelihood estimation. Mutational biases are more difficult to explore, as no simple steady-state solution to the three-dimensional diffusion equation exists for biased mutation rates, but computational results can still be obtained. It is also straightforward to investigate balancing selection or the effects of dominance: these lead to well-understood modifications to the diffusion equations and their steady-state solutions (Ewens 2004). In the PRF framework, by contrast, such generalizations are much more complex. In particular, balancing selection is impossible to analyze within the PRF framework, because it leads to mutant lineages reaching stable intermediate frequencies in the population. As a result, the generation of new mutations is not balanced by the extinction or fixation of older ones, and hence no steady-state distribution of lineage frequencies exists.

The methods we have developed also allow us to relax the assumption of constant γ across sites and instead infer aspects of the distribution of selection pressures (Piganeau and Eyre-Walker 2003, Nielsen *et al.* 2005, and Boyko *et al.* 2008 have recently analyzed similar extensions of the original PRF method). It is not yet clear how much data is required to provide adequate power for inferring this distribution to a given resolution. However, we can hope to gain a great deal of insight with only a few additional parameters—say, the number of sites that are neutral, lethal, and negatively selected and the weighted average selection pressure on the latter class. As we have shown, we can estimate the number of lethal sites with no reduction in power relative to the original PRF method, so this proposal would involve only one additional parameter. Additional classes of sites would involve more parameters. Eyre-Walker and Keightley (2007) have recently stressed that the distribution of fitness effects of deleterious mutations is likely to be complex and multimodal. Using our approach, we can choose how to focus the power in the data to investigate the aspects of this complex distribution that are most interesting in a given situation. The appropriate choice of resolution will depend on the context and quality of the data.

However, the diffusion approach is not without drawbacks. The one-dimensional version does not make use of all of the data available in the observed polymorphism spectrum and is likely to be of little use in practice because we typically do not know whether or not the ancestral nucleotide was preferred. The three-dimensional version uses all the data in the folded case, but cannot make use of outgroup data (the unfolded case). In practice, when we do not have an outgroup, the three-dimensional diffusion method provides the most accurate estimates of selection pressures across the full range of parameters. When an outgroup is available, the three-dimensional diffusion approach is still sometimes more accurate than the PRF, despite wasting information on the ancestral state, but only in parameter regimes typical of bacteria or viruses. In other cases the per-site PRF method is best.

The diffusion method also cannot naturally handle positive selection. The steady-state evolutionary dynamics at a site are always dominated by the preferred nucleotide (or nucleotides), with negative selection acting against polymorphisms for the disfavored nucleotides. At the level of an individual site, positive selection is a process that is intrinsically out of steady state: the spread of a favorable nucleotide before it becomes fixed. The PRF method handles positive selection by implicitly positing rather strange dynamics at individual sites. All mutant lineages are assumed to be positively selected—so if a mutant nucleotide fixes at a given site, mutations back to the ancestral nucleotide are again assumed to be positively selected (the analogous strange dynamics also apply to negative selection). While this assumption makes little sense at a per-site level, it allows the PRF model to obtain a steady state *across sites*, provided positive selection is ongoing and not saturated. We could modify our diffusion methods to mimic the PRF treatment of positive selection by changing the boundary conditions in our diffusion equations. Specifically, we would assume that probability flowing into $x = 1$ (*i.e.*, fixation of a mutant nucleotide) is absorbed and moved to $x = 0$ (*i.e.*, "reset" so that new mutations will again be favored). This diffusion equation can be solved exactly, and the solution used as a basis for inferring positive selection using the the per-site diffusion methods we have developed.

Ideally, however, we want to infer positive selection in the context of a realistic and well-defined model of the dynamics at individual sites. Such an approach would necessarily involve full time-dependent solutions to the diffusion equations; Mustonen and Lassig (2007) suggest a method along these lines. We do not pursue this approach here, but see Chen *et al.* (2007) for a step in this direction. Regardless of the methodology, it will always be difficult to discriminate positive selection from negative selection on the basis of the polymorphism frequency spectrum alone, particularly when only folded data are available. Whether selection is positive or negative, mutant lineages drift nearly neutrally when their frequency is between 0 and $1/|\gamma|$. Positively selected lineages then fix relatively quickly once their frequency becomes substantially larger than $1/|\gamma|$, while negatively selected lineages rarely ever reach frequencies larger than $1/|\gamma|$. Thus from the point of view of the polymorphism frequency spectrum, positive selection is similar to random drift on $[0, 1/|\gamma|]$, with the upper bound a roughly absorbing boundary condition. Negative selection, on the other hand, is also similar to random drift on $[0, 1/|\gamma|]$, but with the upper bound a roughly reflecting boundary condition. Although this is relatively crude—selection does in fact have some impact on low-frequency lineages, and the boundary conditions are not exactly absorbing or reflecting—it indicates that the polymorphism frequency spectrum is roughly similar for negative and positive selection at the same $|\gamma|$. Thus power to distinguish positive from negative selection based on the polymorphism frequency spectrum, especially with folded data, will always be relatively limited, regardless of the method used.

Different steady states: The PRF framework and the diffusion methods are based on fundamentally different steady states. The PRF assumes an ancestral state; the steady-state frequency distribution is that of derived nucleotides relative to this ancestral state. By contrast, the diffusion methods assume mutation back and forth between all possible nucleotides at each site. The steady-state frequency distribution is a full mutation-selection-drift balance of all four nucleotides at the site. It has "forgotten" the ancestral state, and it is obtained over a longer timescale than the PRF steady state.

The differences in the nature of the two steady states are most readily apparent when considering their behavior for small θ_s. This is easiest to see by comparing the PRF steady state to the one-dimensional diffusion result (assuming as usual that one nucleotide is preferred and the other three are unpreferred). For small θ_s, the population will typically be fixed for one of the four nucleotides. Occasionally mutations will occur, making the site temporarily polymorphic within the population. Usually these mutant lineages will drift at low frequency for a short while before going extinct, but occasionally they will fix and the population will be fixed for this new nucleotide. The PRF steady state describes the (transient) frequency spectrum of these occasional mutant lineages before they fix or go extinct. This spectrum is always measured relative to the ancestral nucleotide, which is the last mutation that fixed. On the other hand, the diffusion method describes the longer-term steady-state frequency distribution of the nucleotides. This includes some weight near $x = 0$ (the preferred nucleotide is fixed, or nearly so), as well as some weight near $x = 1$ (one of the unpreferred nucleotides is fixed, or nearly so). Because of the selective bias, the population will be fixed or nearly fixed for the preferred nucleotide a fraction $1/(1 + 3e^{2\gamma})$ of the time (Ewens 2004), so $1/(1 + 3e^{2\gamma})$ of the total weight is in the part of the distribution near $x = 0$. The remainder of the time, the population will be fixed or nearly fixed for one of the unpreferred nucleotides; this is the weight in the distribution near $x = 1$.

The discussion above helps to elucidate the relationship between the steady states in the PRF model *vs.* those in the diffusion model. A fraction $1/(1 + 3e^{2\gamma})$ of the time the most recent fixation will have been for the preferred state. This is the ancestral state for the PRF method, and the frequency spectrum of the occasional unpreferred negatively selected mutants is given by the PRF steady state. The remainder of the time the most recent fixation will have been for one of the unpreferred states. Now this is the ancestral state for the PRF method, and the frequency spectrum of occasional mutants for the preferred nucleotide, which are under positive selection, is again given by the PRF steady state, but with the opposite sign of γ. Thus we expect that the diffusion method steady state is just the sum of these two PRF steady states, weighted by the frequency with which each type of nucleotide is the ancestral state, namely

$$f^{\text{diff}}(\gamma, x) \sim p(\gamma) f^{\text{PRF}}(\gamma, x) + [1 - p(\gamma)] f^{\text{PRF}}(-\gamma, 1 - x),$$

where $p(\gamma)$ denotes the chance that the most recent fixation was for the preferred state. This relationship between the two steady states is illustrated in Figure 3. It holds in the limit of small θ_s. For larger θ_s, the population is not always nearly fixed for one state. Mutant lineages occur more frequently and can segregate simultaneously (*i.e.*, finite-sites effects). Since the diffusion approach accounts for this while the PRF does not, the relationship between the steady states breaks down.

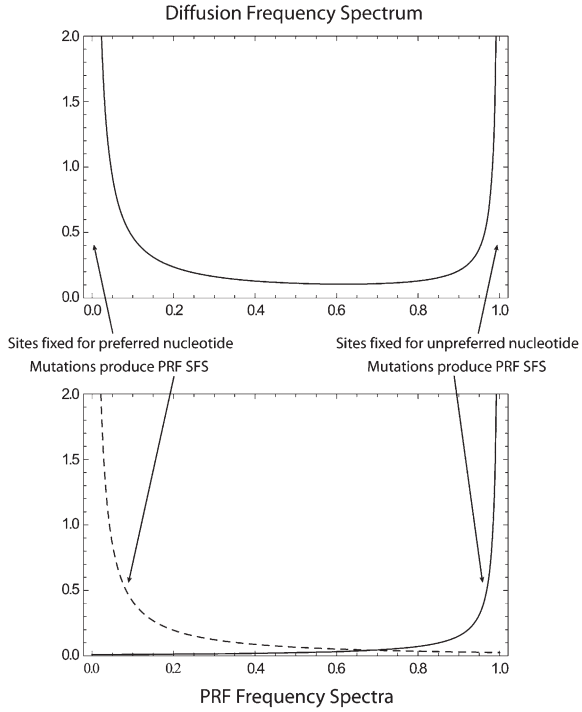

Figure 3.—Relationship between the PRF and diffusion frequency spectra for small θ_s. Shown at the top is the one-dimensional diffusion frequency spectrum for $\theta_s = 10^{-3}$, $\gamma = -0.5$. Most of the time the population is fixed or nearly fixed for the preferred nucleotide (this is the weight near $x = 0$), while sometimes the population is fixed or nearly fixed for an unpreferred nucleotide (the weight near $x = 1$). When the population is nearly fixed for the preferred (unpreferred) nucleotide, mutations produce a PRF site frequency spectrum (SFS) as shown on the left (right) of the bottom graph. For this small value of θ_s, the diffusion frequency spectrum is roughly equal to the sum of these PRF frequency spectra.

The difference between the diffusion and PRF steady states points to one additional bias in the PRF method that we have not yet discussed. Imagine a stretch of sequence that has been under constant purifying selection for an extremely long time. As we have just seen, we expect that at a fraction $3e^{2\gamma}/(1 + 3e^{2\gamma})$ of the sites in this sequence, the most recent ancestor will be an unpreferred state. The PRF frequency spectrum at these sites will be characteristic of positive selection (mutations to the preferred nucleotides are beneficial). Although this will typically represent a minority of the sites, the PRF method weights sites under positive selection more strongly than those under negative selection when inferring a weighted average γ. Thus, the PRF method will be biased toward inferring positive selection—despite the fact the sequence has experienced negative selection for an arbitrarily long time. This effect is present even for small θ_s, and it is most severe when selection is weak but nonzero.

Whether this effect represents a bias in the PRF method is a matter of taste. In some sense, the PRF is

correctly identifying sites under positive selection: after all, the mutations to the preferred nucleotides are in fact beneficial. In another sense, however, the sequence has experienced constant purifying selection throughout its past. The fact that occasionally an unpreferred nucleotide fixes before reverting to the preferred state does not reflect any change in selection pressures. Thus, if we define positive selection as a pressure to innovate, or a shift in the selective landscape, then such a reversion does not reflect positive selection. We have explored this effect through simulation and found that in microbial parameter regimes it is a detectable but minor effect and is much less important in practice than the other biases we have studied. However, it would be straightforward to modify either the original or the per-site PRF method to correct for this (the diffusion methods already account for this effect). We simply change the expected spectrum to reflect the fact that a fraction $1/(1 + 3e^{2\gamma})$ of sites will exhibit the spectrum characteristic of negative selection while a fraction $3e^{2\gamma}/(1 + 3e^{2\gamma})$ will exhibit that characteristic of positive selection (with equal but opposite γ). This leads to a simple modification of the PRF spectrum, with no added parameters, which we can use for ML estimation as before.

Different timescales: Because the PRF frequency spectrum is measured relative to the last fixed state, it reaches steady state in at most of order N generations, the coalescent timescale. The diffusion method takes longer to reach steady state—in the neutral case, mutations fix every $1/\mu$ generations, so it takes of order $1/\mu$ generations to reach steady state. When selection is operating, the diffusion steady state can arise either more quickly or more slowly, depending on the initial state of the population. Note that the ratio of the PRF to the diffusion timescale is of order θ_s. Thus for the small θ_s relevant for most eukaryotic populations, there is a big gap between the two timescales, while for the larger θ_s relevant to bacteria and viruses the two timescales will be more similar.

These two timescales determine whether the presence of information from an outgroup is likely to improve our inferential power. Imagine we sample polymorphism from some population A and one sequence from an outgroup species. If the outgroup diverged from the population A less than N generations ago, population A has not reached the PRF steady state in the time since the divergence. Thus the frequency spectrum of population A is not independent from the frequency spectrum in the outgroup. The outgroup does not provide much more information than simply sampling another individual from population A (though it does provide some additional information). This situation has recently been analyzed in the neutral case by Chen *et al.* (2007) and applied to human polymorphism data using a Neanderthal outgroup.

If instead the outgroup diverged from population A more than N but less than $1/\mu$ generations ago, it can be used to infer the most recent ancestral state. Since the time of this ancestor, the PRF frequency spectrum will have reached equilibrium, and it can be used to estimate parameters accounting for this knowledge (though finite-sites issues can still matter when θ_s is large, and diffusion methods may be more accurate despite the information from the outgroup). But the closer the time since the outgroup diverged is to $1/\mu$ generations, the more unreliable the inference of the ancestral state becomes, and ancestral state misidentification becomes more of a problem. This problem and possible ways to account for it have recently been examined by Hernandez *et al.* (2007).

If the outgroup diverged more than $1/\mu$ generations ago, the population will have forgotten the ancestral state since divergence, and the outgroup will not be useful. Of course, in this case we would probably not even be able to align the outgroup sequence to the sequences from population A. In this case, only the folded frequency data are available, and the diffusion methods are most appropriate.

We thank Daniel Fisher, Marc Feldman, and Warren Ewens for helpful comments on the manuscript. M.M.D. acknowledges support from Center grant P50GM071508 from the National Institute of General Medical Science to the Lewis–Sigler Institute. J.B.P. acknowledges support from the Burroughs Wellcome Fund, the James S. McDonnell Foundation, and the Defense Advanced Research Projects Agency Fun-Bio Program (HR0011-05-1-0057).

LITERATURE CITED

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Akashi, H., and S. W. Schaeffer, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295–307.

Bartolome, C., X. Maside, S. Yi, A. L. Grant and B. Charlesworth, 2005 Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*. Genetics **169:** 1495–1507.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. **4:** e1000083.

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

Bustamante, C. D., R. Nielsen, S. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531–534.

Bustamante, C. D., R. Nielsen and D. L. Hartl, 2003 Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor. Popul. Biol. **63:** 91–103.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Chen, H., R. E. Green, S. Paabo and M. Slatkin, 2007 The joint allele-frequency spectrum in closely related species. Genetics **177:** 387–398.

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics **176:** 1759–1798.

Drake, J. W., B. Charlesworth, W. Charlesworth and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667–1686.

Ewens, W. J., 2004 *Mathematical Population Genetics: I. Theoretical Introduction*. Springer, New York.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. **8:** 610–618.

Galtier, N., E. Bazin and N. Bierne, 2006 Gc-biased segregation of noncoding polymorphisms in Drosophila. Genetics **172:** 221–228.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

Hernandez, R. D., S. H. Williamson and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. **24:** 1792–1800.

Lercher, M. J., N. G. C. Smith, A. Eyre-Walker and L. Hurst, 2002 The evolution of isochores: evidence from SNP frequency distributions. Genetics **162:** 1805–1810.

Loewe, L., B. Charlesworth, C. Bartolome and V. Noel, 2006 Estimating selection on nonsynonymous mutations. Genetics **172:** 1079–1092.

Lynch, M., and J. S. Conery, 2003 The origins of genome complexity. Science **21:** 1401–1404.

McDonald, J., and M. Kreitman, 1991 Adaptive protein evolution at the adh locus in Drosophila. Nature **351:** 652–654.

Moran, P. A. P., 1959 The survival of a mutant gene under selection. ii. J. Aust. Math. Soc. **1:** 485–491.

Mustonen, V., and M. Lassig, 2007 Adaptations to fluctuating selection in Drosophila. Proc. Natl. Acad. Sci. USA **104:** 2277–2282.

Nachman, M. W., 1998 Deleterious mutations in animal mitochondrial DNA. Genetica **102/103:** 61–69.

Nielsen, R., C. D. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. **3:** 976–985.

Piganeau, G., and A. Eyre-Walker, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. USA **100:** 10335–10340.

Rand, D. M., and L. M. Kann, 1998 Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. Genetica **102/103:** 393–407.

Sawyer, S., R. J. Kulathinal, C. D. Bustamante and D. L. Hartl, 2003 Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. J. Mol. Evol. **57:** S154–S164.

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

Wakeley, J., 2003 Polymorphism and divergence for island-model species. Genetics **163:** 411–420.

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

Weinreich, D. M., and D. M. Rand, 2000 Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics **156:** 385–399.

Williamson, S., A. Fledel-Alon and C. D. Bustamante, 2004 Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics **168:** 468–475.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA **102:** 7882–7887.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA **24:** 253–259.

Wright, S., 1949 Adaptation and selection, pp. 365–389 in *Genetics, Paleontology, and Evolution*, edited by G. L. Jepson, G. G. Simpson and E. Mayr. Princeton University Press, Princeton, NJ.

Zhu, L., and C. D. Bustamante, 2005 A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics **170:** 1411–1421.

Communicating editor: M. K. Uyenoyama

APPENDIX: ALGORITHM TO COMPUTE THE LIKELIHOOD EXPRESSIONS UNDER THE PER-SITE PRF

To implement an inference procedure based on our per-site version of the PRF model we need to compute likelihood expressions for any possible unfolded configuration of nucleotides, an example of which is shown in Equation 10.

Our algorithm starts by initializing the expressions associated with each configuration to zero. We loop over all partitions of the sample size n written as the sum of nonzero integers, $n = n_1 + n_2 + \ldots + n_z + n_w$. In this partition, numbers n_1 through n_z represent the number of samples of distinct mutant lineages, and n_w represents the number of samples with the ancestral state. The sum of all such sample numbers equals the total sample size, n. Such a sample will correspond to a particular folded configuration (a, b, c, d) that depends upon the identity of the mutant nucleotides associated with the z mutant lineages. There are 3^z distinct possible assignments of mutant nucleotide types to the mutant lineages. We loop over each of these possible assignments, and for each we add a term of the form

$$(1/3)^z \frac{F(1)^{n_1}}{n_1!} \times \frac{F(2)^{n_2}}{n_2!} \cdots \frac{F(n_z)^{n_z}}{n_z!}$$

to the likelihood expression for the folded configuration (a, b, c, d) that arises from the assignment of nucleotide types. After looping over all possible numbers of mutant lineages and over all possible assignments of mutant nucleotide types to such lineages, we obtain complete expressions for the likelihoods of all unfolded configurations.

To compute likelihood expressions for folded configurations, we perform the exact same algorithm as above with a small modification. In this case, once we have assigned nucleotide types to each of the z mutant lineages, we determine the associated folded configuration by simply sorting the number of lineages (including the wild type) of each mutant nucleotide type in descending order.

A computer implementation of this algorithm, for either folded or unfolded likelihood expressions, is freely available upon request.