Directional Selection and the Site-Frequency Spectrum

Carlos D. Bustamante,* John Wakeley,* Stanley Sawyer† and Daniel L. Hartl*

**Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138 and †Department of Mathematics, Washington University, St. Louis, Missouri 63130*

Manuscript received April 24, 2001
Accepted for publication August 20, 2001

ABSTRACT

In this article we explore statistical properties of the maximum-likelihood estimates (MLEs) of the selection and mutation parameters in a Poisson random field population genetics model of directional selection at DNA sites. We derive the asymptotic variances and covariance of the MLEs and explore the power of the likelihood ratio tests (LRT) of neutrality for varying levels of mutation and selection as well as the robustness of the LRT to deviations from the assumption of free recombination among sites. We also discuss the coverage of confidence intervals on the basis of two standard-likelihood methods. We find that the LRT has high power to detect deviations from neutrality and that the maximum-likelihood estimation performs very well when the ancestral states of all mutations in the sample are known. When the ancestral states are not known, the test has high power to detect deviations from neutrality for negative selection but not for positive selection. We also find that the LRT is not robust to deviations from the assumption of independence among sites.

THE Poisson random field (PRF) models of Sawyer and Hartl (1992) and Hartl *et al.* (1994) afford a general likelihood framework for estimating mutation and selection parameters in various population genetic settings. This is currently the only method available for estimating selection directly from DNA sequence data in population genetics. The PRF model has proven quite useful in studying selection on "silent" site polymorphism in Drosophila (Akashi 1995; Akashi and Schaeffer 1997) and *Escherichia coli* (Hartl *et al.* 1994) as well as amino acid variation in *E. coli* and *Salmonella enterica* (Hartl *et al.* 2000). It has also been used to study the power of various tests of neutrality on the basis of polymorphism and divergence data (Akashi 1999).

Since the PRF model assumes independence among sites, it is a limiting case for population genetic models that incorporate mutation, selection, and recombination. For typical DNA sequence data, in which many sites segregate simultaneously, the assumption of independence among sites is equivalent to the assumption that there is free recombination between segregating nucleotides. Studying statistical inference in the PRF framework can therefore illuminate efforts to build models that relax the assumption of independence among sites. The reason for this is that any model that hopes to estimate mutation, selection, and recombination parameters can do as well as or worse than the PRF if the data conform to the PRF model.

It would be of great interest to know, therefore, how well maximum-likelihood methods for point estimation and hypothesis testing in the PRF framework perform under varying levels of selection and mutation. One property of interest is the probability of rejection neutrality if the locus is truly under selection, which addresses the power of the likelihood-ratio tests (LRTs) of neutrality. Another statistical issue of interest concerns the robustness of these models to deviations from the assumption of independently evolving sites, since this assumption may be unrealistic for certain types of data.

This article treats in detail the PRF model for site-frequency data proposed by Hartl *et al.* (1994) for both folded and unfolded site configurations. In particular, we (1) derive the asymptotic variances and covariance of the maximum-likelihood estimates (MLEs) of the selection and mutation parameters, (2) explore the power of the likelihood-ratio test of no selection under varying levels of mutation and selection, (3) explore the coverage of two types of confidence intervals for the selection parameter, and (4) explore the robustness of the model to deviations from the assumption of independence among sites.

THEORY

Poisson random field model: First let us consider a simple example with which to illustrate the type of data being modeled. The data in Table 1 represent a sample of six individuals sequenced at five variable DNA sites. Each column is an aligned DNA site and each row is an individual sampled from the population. At each site where the individual carries the *derived* nucleotide, the resulting entry in the matrix is 1; otherwise the entry is 0. For each site, then, the sum of the column gives

Corresponding author: Daniel L. Hartl, Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138. E-mail: dhartl@oeb.harvard.edu

TABLE 1

A simple example of the type of data modeled in this article

	DNA site				
	1	2	3	4	5
Individual 1	0	0	1	0	0
Individual 2	0	0	1	0	1
Individual 3	0	0	0	1	0
Individual 4	0	1	1	0	0
Individual 5	1	1	1	0	0
Individual 6	0	0	0	0	0
Total	1	2	4	1	1

The site-frequency spectrum for this sample is (3, 1, 0, 1, 0) as explained in the text.

the number of individuals in the sample that carry the derived form of the mutation.

Define the *unfolded* site-frequency spectrum as the random vector $X = (X_1, X_2, \ldots, X_{n-1})$ of sample configurations X_i, where X_i represent the number of sites that have $n - i$ ancestral and i derived nucleotides for $1 \leq i \leq n - 1$ among the n aligned nucleotide sequences. For the example in Table 1 the unfolded site-frequency spectrum is (3, 1, 0, 1, 0) since there are three sites where one individual carries the derived mutation, one site where two individuals carry the derived mutation, no sites where three individuals or five individuals carry the derived mutations, and one site where four individuals carry the derived mutation.

So far we have assumed that for each site we know which state is derived and which is ancestral. If this information is unknown, we would model the *folded* site-frequency spectrum, X', defined as the number of sites where 1 individual or $(n - 1)$ individuals carry a mutation, 2 individuals or $(n - 2)$ individuals carry a mutation, etc., so that $X'_i = X_i + X_{n-i}$ for $i < n - i$ and $X'_{n/2} = X_{n/2}$ for $i = n/2$. For the example above, the folded site-frequency spectrum is (3, 2, 0).

We wish to model the site-frequency spectrum for DNA sites in an infinite-sites (*i.e.*, irreversible mutation) independent (*i.e.*, freely recombining) model with (weak) directional haploid selection. We assume that all mutations that enter the population have the same selective effect and that there are no epistatic interactions among mutations (*i.e.*, independent fitness effects of all mutations). We also assume that there is no population structure to the data and the population has achieved statistical stationarity.

Fisher (1930) and Wright (1938, 1969) derived a formula for what Wright called the "transient distribution" for the frequency of a single newly arisen mutation under selection. Letting $f(q, \gamma)$ be the density of the frequency of a single mutation, q, in the range dq, their formula is

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

where γ is the product of the selection coefficient (a in Fisher's notation, usually s in Wright's notation) and the effective haploid population size. Here s is the per-generation increase in $\log(p/q)$ so that $\Delta q = s\, p(1 - p)$ if s is small.

Sawyer and Hartl (1992) showed that for recurrent mutation at a locus with independently evolving sites, the expected density of a random field composed of the individual frequencies of the derived mutations is $(\theta/2)f(q, \gamma)$, where $\theta/2$ is the *per-locus* mutation rate scaled by the effective population size. They also showed that this random field is a PRF, which means that the number of sites whose derived mutation frequency q satisfies $0 < a < q < b < 1$ has a Poisson distribution for given a, b and also that the number of sites for nonoverlapping intervals (a, b) is independent (Kingman 1993).

For a particular site, given the frequency q of the mutation in the population, the probability of sampling i sequences of one type and $n - i$ of another is binomially distributed with parameters n and q. If mutations enter the population as a Poisson process with rate $\theta/2$, the X_i's are independent Poisson-distributed random variables with mean, $\theta F(i, \gamma)$, where

$$\begin{aligned} F(i, \gamma) &= \int_0^1 \frac{f(q, \gamma)}{2} \cdot \Pr(i|q)\, dq \\ &= \int_0^1 \frac{1 - e^{-2\gamma(1-q)}}{1 - e^{-2\gamma}} \frac{1}{q(1 - q)} \binom{n}{i} q^i (1 - q)^{n-i} dq. \end{aligned} \tag{2}$$

Therefore, the probability of observing x_i sites that have i derived and $(n - i)$ ancestral mutations is

$$p(X_i = x_i|\theta, \gamma) = e^{-\theta F(i,\gamma)} \frac{(\theta F(i, \gamma))^{x_i}}{x_i!}. \tag{3}$$

The PRF model outline above leads directly to a likelihood-ratio test of neutrality, which compares the null hypothesis $\gamma = 0$ with the alternative hypothesis that $\gamma \neq 0$. Since the X_i's are independent, the likelihood function is the product of the individual $p(X_i = x_i|\theta, \gamma)$. Letting $L_f(\theta, \gamma \mid x)$ be the likelihood function for the folded model and L_u be the likelihood function for the unfolded model,

$$L_u(\theta, \gamma|x) = \prod_{i=1}^{n-1} e^{-\theta F(i,\gamma)} \frac{(\theta F(i, \gamma))^{x_i}}{x_i!} \tag{4}$$

and

$$L_f(\theta, \gamma|x') = \prod_{i=1}^{\lfloor n/2 \rfloor} e^{-\theta G(i,\gamma)} \frac{(\theta G(i, \gamma))^{x'_i}}{x'_i!}, \tag{5}$$

where $\lfloor i \rfloor$ is the greatest integer $\leq i$ and

$$G(i, \gamma) = \begin{cases} F(i, \gamma) + F(n - i, \gamma) & \text{if } i < n - i, \\ F\left(\frac{n}{2}, \gamma\right) & \text{if } i = \frac{n}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

To perform the likelihood-ratio test, we need to max-

imize (4) or (5) for θ and γ under (A) the unconstrained state space and (B) the constraint that $\gamma = 0$. Letting $\hat{\gamma}$ and $\hat{\theta}$ be the unconstrained MLEs of γ and θ, and $\hat{\theta}_W$ the MLE of θ under neutrality, the likelihood-ratio test statistic, Λ, is

$$\Lambda = \frac{L(\hat{\theta}, \hat{\gamma}|x)}{L(\hat{\theta}_W, 0|x)}. \tag{7}$$

Appealing to large sample results (Kendall 1987), $2\ln(\Lambda)$ is $\sim\chi^2_1$ distributed since the neutral and selected models differ by 1 d.f.

Maximum-likelihood estimation and asymptotic results: Finding the maximum-likelihood estimates of θ and γ for both the folded and unfolded site frequency models is relatively straightforward. Let l be the log-likelihood function for the unfolded model

$$l(\theta, \gamma|x) \propto \sum_{i=1}^{n-1} x_i \log(\theta F(i, \gamma)) - \theta \sum_{i=1}^{n-1} F(i, \gamma). \tag{8}$$

The maximum-likelihood estimates of γ and θ (*i.e.*, $\hat{\gamma}$ and $\hat{\theta}$) are usually found by setting the partial derivatives of the log-likelihood function equal to zero. Letting S represent the total number of segregating sites ($S = \sum_{i=1}^{n-1} x_i$),

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{n-1} F(i, \gamma) + \frac{S}{\theta} \tag{9}$$

$$\frac{\partial l}{\partial \gamma} = -\theta \sum_{i=1}^{n-1} F'(i, \gamma) + \sum_{i=1}^{n-1} x_i \frac{F'(i, \gamma)}{F(i, \gamma)}, \tag{10}$$

where

$$F'(i, \gamma) = \int_0^1 \frac{f'(q, \gamma)}{2} \binom{n}{i} q^i (1 - q)^{n-i} dq \tag{11}$$

and

$$f'(q, \gamma) = \frac{4e^{-2\gamma(1-q)}}{(1 - e^{-2\gamma})q} - \frac{4e^{-2\gamma}(1 - e^{-2\gamma(1-q)})}{(1 - e^{-2\gamma})^2 q(1 - q)}. \tag{12}$$

An efficient approach to solving for the MLEs is to maximize the profile likelihood that has the same global maximum but requires only maximization in one dimension. The log-profile-likelihood function, $l^*(\gamma)$, is $l(\tilde{\theta}, \gamma)$, where $\tilde{\theta}$ is the maximum-likelihood estimate of θ given γ. By setting (9) to zero and solving for $\tilde{\theta}$ one can easily show that

$$\tilde{\theta} = \frac{S}{\sum_{i=1}^{n-1} F(i, \gamma)}. \tag{13}$$

It follows from (2) that

$$\lim_{\gamma \to 0} F(i, \gamma) = \frac{1}{i}, \tag{14}$$

which proves that under neutrality

$$\tilde{\theta} = \hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} 1/i}. \tag{15}$$

Here $\hat{\theta}_W$ is Watterson's (1975) estimator of θ. This shows that the total number of segregating sites is a sufficient statistic for θ given γ. The two preceding results hold, regardless of the folding of the site configurations since $\sum_{i=1}^{n-1} F(i, \gamma) = \sum_{i=1}^{[n/2]} G(i, \gamma)$.

We can now maximize the log-profile-likelihood function $l^*(\gamma)$ numerically, using a standard optimization technique such as Newton-Raphson iteration (Lange 1999). To implement Newton-Raphson we need to find the first and second derivatives of $l^*(\gamma)$ with respect to γ. To do this we first substitute (13) into (8) and take the necessary derivatives. The results are

$$\frac{dl^*}{d\gamma} = \sum_{i=1}^{n-1} x_i \frac{F'(i, \gamma)}{F(i, \gamma)} - S \frac{\sum_{i=1}^{n-1} F'(i, \gamma)}{\sum_{i=1}^{n-1} F(i, \gamma)} \tag{16}$$

and

$$\frac{d^2 l^*}{d\gamma^2} = \sum_{i=1}^{n-1} x_i \frac{F''(i, \gamma)F(i, \gamma) - F'(i, \gamma)^2}{F(i, \gamma)^2} - S \frac{T_0 T_2 - T_1^2}{T_0^2}, \tag{17}$$

where

$$T_0(\gamma) = \sum_{i=1}^{n-1} F(i, \gamma) \tag{18}$$

$$T_1(\gamma) = \sum_{i=1}^{n-1} F'(i, \gamma) \tag{19}$$

$$T_2(\gamma) = \sum_{i=1}^{n-1} F''(i, \gamma), \tag{20}$$

and

$$F''(i, \gamma) = \int_0^1 \frac{f''(i, \gamma)}{2} \binom{n}{i} q^i (1 - q)^{n-i} dq, \tag{21}$$

where

$$f''(q, \gamma) = \left(\frac{1 - e^{-2\gamma(1-q)}}{q(1 - q)}\right)\left(\frac{16e^{-4\gamma}}{(1 - e^{-2\gamma})^3} + \frac{8e^{-2\gamma}}{(1 - e^{-2\gamma})^2}\right) - \frac{16e^{-2\gamma-2\gamma(1-q)}}{(1 - e^{-2\gamma})^2 q} - \frac{8(1 - q)e^{-2\gamma(1-q)}}{(1 - e^{-2\gamma})q}. \tag{22}$$

To find the equivalent results for the folded model, substitute n by $[n/2]$ and F, F', and F'' by G, G', and G'', respectively, in the equations above, where

$$G'(i, \gamma) = \begin{cases} F'(i, \gamma) + F'(n - i, \gamma) & \text{if } i < n - i, \\ F'\left(\frac{n}{2}, \gamma\right) & \text{if } i = \frac{n}{2}, \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

$$G''(i, \gamma) = \begin{cases} F''(i, \gamma) + F''(n - i, \gamma) & \text{if } i < n - i, \\ F''\left(\frac{n}{2}, \gamma\right) & \text{if } i = \frac{n}{2}, \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

To find the asymptotic covariance matrix of the maximum-likelihood estimates, we compute the Fisher information matrix $\boldsymbol{I}(\theta, \gamma)$, which is minus the expected value

of the matrix of second derivations of the log-likelihood function. Since the components of $\boldsymbol{I}$ are linear in x_i and S, this is equivalent to replacing x_i by $E(X_i) = \theta F(i, \gamma)$ and S by $\theta T_0(\gamma)$. The asymptotic covariance matrix of $\hat{\theta}$ and $\hat{\gamma}$ is the inverse of the empirical Fisher information matrix $\boldsymbol{I}(\hat{\theta}, \hat{\gamma})$.

The components of the Fisher information matrix $\boldsymbol{I}(\theta, \gamma)$ are easily shown to be

$$\boldsymbol{I}_{11} = -E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = \frac{T_0(\gamma)}{\theta} \tag{25}$$

$$\boldsymbol{I}_{12} = \boldsymbol{I}_{21} = -E\left[\frac{\partial^2 l}{\partial \gamma \partial \theta}\right] = T_1(\gamma) \tag{26}$$

$$\boldsymbol{I}_{22} = -E\left[\frac{\partial^2 l}{\partial \gamma^2}\right] = \theta(T_2(\gamma) - t(\gamma)), \tag{27}$$

where

$$t(\gamma) = \sum_{i=1}^{n-1} \frac{F''(i, \gamma)F(i, \gamma) - F'(i, \gamma)^2}{F(i, \gamma)}. \tag{28}$$

The asymptotic covariance of the MLEs is the inverse matrix of $\boldsymbol{I}(\hat{\theta}, \hat{\gamma})$, which by Cramer's rule equals

$$\boldsymbol{M} = \frac{1}{\boldsymbol{I}_{11}\boldsymbol{I}_{22} - \boldsymbol{I}_{12}^2}\begin{pmatrix} \boldsymbol{I}_{22} & -\boldsymbol{I}_{12} \\ -\boldsymbol{I}_{12} & \boldsymbol{I}_{11} \end{pmatrix} \tag{29}$$

evaluated at $(\hat{\theta}, \hat{\gamma})$. It then follows that

$$\mathrm{VAR}(\hat{\theta}) = \boldsymbol{M}_{11} = \frac{\hat{\theta}}{T_0(\hat{\gamma}) - [T_1(\hat{\gamma})^2/(T_2(\hat{\gamma}) - t(\hat{\gamma}))]} \tag{30}$$

$$\mathrm{COV}(\hat{\gamma}, \hat{\theta}) = \boldsymbol{M}_{12} = \frac{1}{[\mathrm{T}_0(\hat{\gamma})(\mathrm{T}_2(\hat{\gamma}) - t(\hat{\gamma}))]/T_1(\hat{\gamma}) - T_1(\hat{\gamma})} \tag{31}$$

$$\mathrm{VAR}(\hat{\gamma}) = \boldsymbol{M}_{22} = \frac{1}{\hat{\theta}(T_2(\hat{\gamma}) - t(\hat{\gamma}) - T_1(\hat{\gamma})^2/T_0(\hat{\gamma}))}. \tag{32}$$

Note that the negative inverse of the second derivative of the log-profile-likelihood function (17), with x_i replaced by $E(X_i) = \theta F(i, \gamma)$ and θ by S/T_0, is formally the same as (32). this is consistent with the fact that both expressions give the asymptotic variance of $\hat{\gamma}$.

From general likelihood theory, we expect $(\hat{\theta}, \hat{\gamma})$ in large samples to be distributed as a multivariate normal with mean (θ, γ) and variance-covariance matrix $\boldsymbol{M}$. Using this result and the fact that $2\ln(\Lambda)$ is approximately χ_1^2 distributed, we can define two types of confidence sets for γ at given confidence level $1 - \alpha$.

The first confidence set consists of γ in the interval $\hat{\gamma} \pm Z(1 - \alpha/2)\sqrt{\mathrm{VAR}(\hat{\gamma})}$, where $Z(1 - \alpha/2)$ is the standard normal quantile corresponding to level $1 - \alpha/2$ (*e.g.*, $Z(0.975) = 1.96$). The second confidence set corresponds to the set of values of γ for which we would not reject the likelihood-ratio test: γ such that $l^*(\gamma) \geq l^*(\hat{\gamma}) - 0.5\chi^2_{1,1-\alpha}$ (Kendall 1987).

SIMULATIONS

The two properties of the LRT that we are interested in exploring are the size and power of the test.

The size, confidence level, or probability of type I error of a test is defined as the probability of rejecting a true null hypothesis and is denoted α. The power of a test is the probability of rejecting a false null hypothesis and is denoted $1 - \beta$, where β is the probability of not rejecting a false null hypothesis (*i.e.*, the probability of a type II error).

In particular, we explore whether the LRT maintains the proper size even if the assumption of independence among segregating sites is not met. If the test is robust to the assumption of free recombination and the null hypothesis is true, then in repeated sampling we should reject the null hypothesis $(100 \cdot \alpha)\%$ of the time. If the test is not robust to the assumption of free recombination, the realized size of the test will vary with the rate of recombination. We might also expect the realized size of the test to vary with the mutation rate if the test is not robust to recombination, since a larger mutation rate will produce a test with a higher probability of rejecting the null hypothesis. It would also be of interest to know how the power of the LRT changes as the strength of selection changes for data that conform to the independence assumption.

To explore these two issues, as well as the coverage of the two different types of confidence intervals for $\hat{\gamma}$, we simulated data under two types of models. The first type of data was generated using coalescent simulations for neutral data under varying rates of recombination and mutation (R. Hudson, personal communication). These data sets were used to explore the robustness of the test to deviations from independence among sites. The second type of data was generated using the PRF model outlined above by sampling independent Poisson random variables with rates given by mutation and selection parameters of interest. These data were used to explore the power of the test under varying levels of mutation and selection, as well as the sampling distributions of $\hat{\theta}_W$, $\hat{\theta}$, and $\hat{\gamma}$. All of the data sets generated had

TABLE 2

Proportion of LRTs that reject neutrality at the 5% level for neutral loci evolving under varying levels of recombination

	Unfolded		Folded	
R	$\theta = 50$	$\theta = 10$	$\theta = 50$	$\theta = 10$
0	0.55	0.27	0.74	0.36
1	0.55	0.26	0.69	0.34
2	0.52	0.23	0.68	0.31
5	0.47	0.17	0.67	0.26
10	0.36	0.15	0.60	0.18
25	0.24	0.11	0.49	0.12
50	0.22	0.07	0.39	0.08
100	0.15	0.05	0.31	0.08
500	0.08	0.05	0.12	0.04
1000	0.05	0.05	0.06	0.04

R refers to the population recombination rate, $4N_e r$, set in the coalescence simulation.

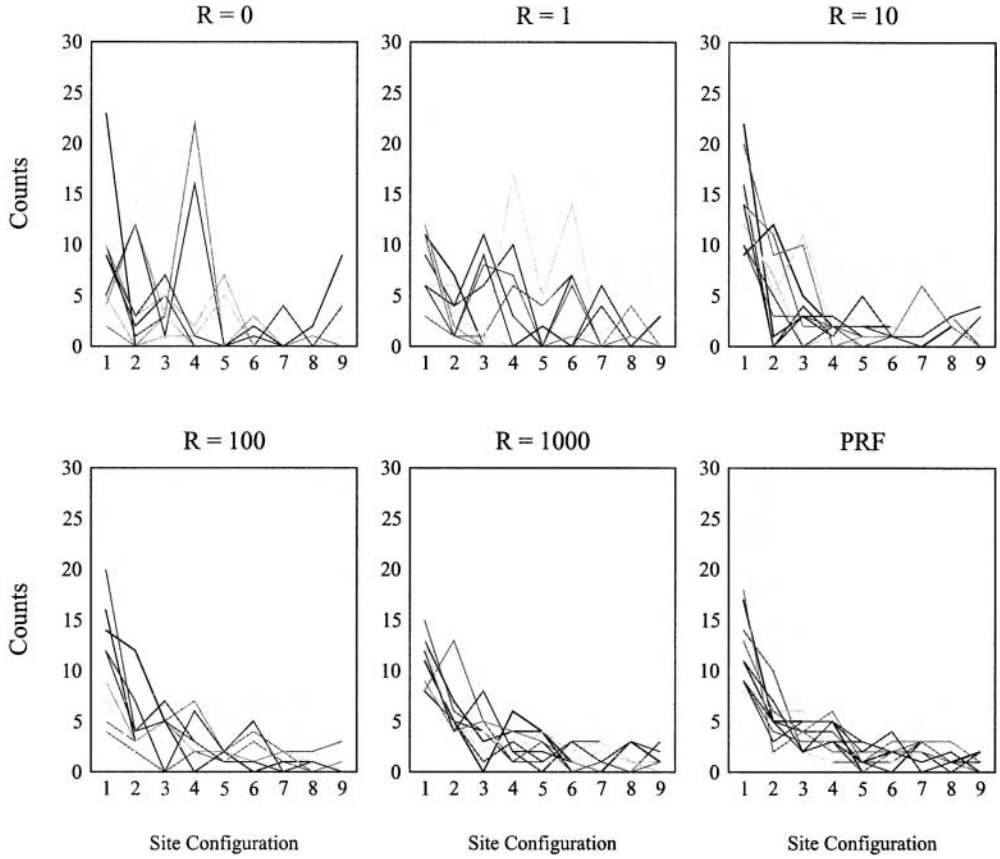

FIGURE 1.—Sampling distribution for unfolded site configurations under varying levels of recombination. The *x*-axis for each graph is the number of sites that were found at frequencies 1/10, 2/10, . . . , 9/10. Each line is a replicate data set of the 500 simulated data sets for each level of recombination.

$n = 10$ sequences and consisted of 500 replicates. The same data were used for all folded and unfolded analysis, with the folded site configurations generated by summing 1 and $(n - 1)$, 2 and $(n - 2)$, etc.

All numerical integration was done using Romberg integration, maximization by Newton-Raphson iteration with bisection, and random number generation using standard numerical algorithms (PRESS *et al.* 1988). Software for calculating the PRF likelihood-ratio test is available from the authors upon request.

RESULTS AND DISCUSSION

Size and power of the LRT: In Table 2 we summarize the results for the analysis of the size of the PRF test for both folded and unfolded site configurations under two different mutation rates and 10 levels of recombination. From this analysis we see that the Poisson random field LRT is extremely sensitive to the assumption of independence among sites. If the test were not sensitive to this assumption, we would expect the entries in the table to fluctuate ~5% but we see rejection levels that reach upward of 50% for tight linkage. While the strength of the effect declines quickly as the rate of recombination increases, the tests do not attain level α until the rate of recombination is more than an order of magnitude greater than the mutation rate.

The reason for this becomes clear when one examines the effect of recombination rate on the distribution of the site-frequency spectrum, *X*. In Figure 1 we present representative replicates of *X* from the coalescent simulations ($\gamma = 0$) under varying levels of recombination for $\theta = 10$. The expected value of the number of sites at a given frequency i/n is the same for each of the six graphs presented $[E(X_i) = \theta/i]$. The variance of X_i, though, increases as the recombination rate tends to

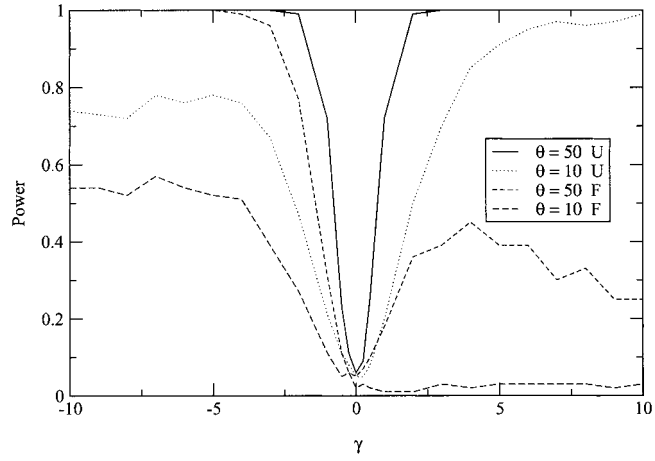

FIGURE 2.—Power of the LRT under varying levels of selection and mutation for folded (F) and unfolded (U) data configurations. On the *x*-axis we plot γ, the value of the selection parameter in the PRF model under which the data were simulated, and the lines connect the proportion of data sets ($n = 500$ for each point) that reject $\gamma = 0$.

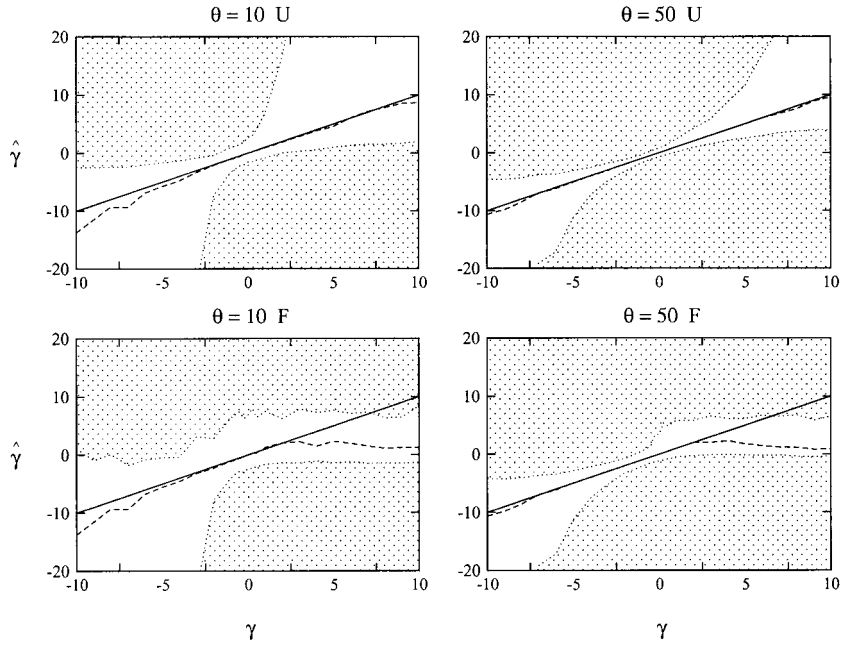

FIGURE 3.—Distributions for the maximum-likelihood estimates of γ ($\hat{\gamma}$) as a function of the simulated value of γ for folded (F) and the unfolded (U) frequency data at two levels of mutation. The open regions correspond to the space containing 95% of the observed $\hat{\gamma}$'s and the dashed lines connect the median estimates across values of γ.

zero. The reason, therefore, that we reject neutrality more than expected in Table 2 with increasing linkage is that we underestimated the variance for the site frequencies. Likewise, if neutral sites are linked to selected sites there will also be an increase in the variance in the distribution of the *X*. These results suggest that the analysis of single genes using this method is invalid if the rate of recombination is not extremely large, since one rejects the null hypothesis (which of course includes the assumption of free recombination) if the data are neutral but not independent.

Performance of maximum likelihood for point estimation in the PRF framework: While the method may not be applicable for linked regions, the analysis of genome-wide single-nucleotide polymorphism may have the quasi-independence required for the test to maintain level α. It would be desirable, therefore, to know how well the method performs in moderate sample sizes when the data conform to the PRF model. For the remainder of this article, all of the analysis is performed on data generated under the PRF model.

In Figure 2 we summarize the results for the analysis of the power of the PRF test for both folded and unfolded site configurations under the same mutation rates above for $\gamma \in [-10, 10]$. We see from that that the LRT has very good power to detect deviations from

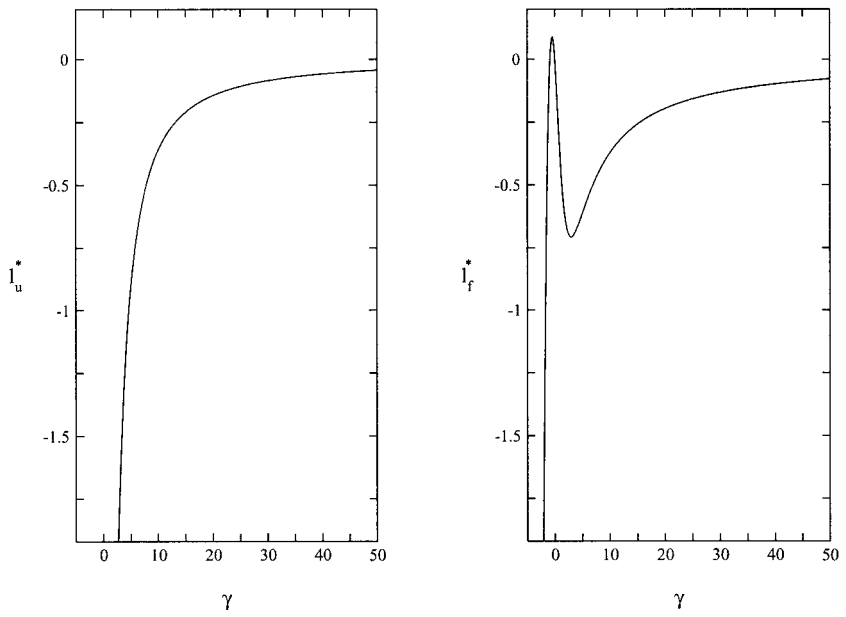

FIGURE 4.—The log-profile-likelihood functions of γ for the folded and unfolded site configurations of the data set in the text with peaks at $\gamma = \infty$ for the unfolded model and at $\gamma = -0.4734$ and $\gamma = \infty$ for the folded model (true $\gamma = 10$ as explained in the text).

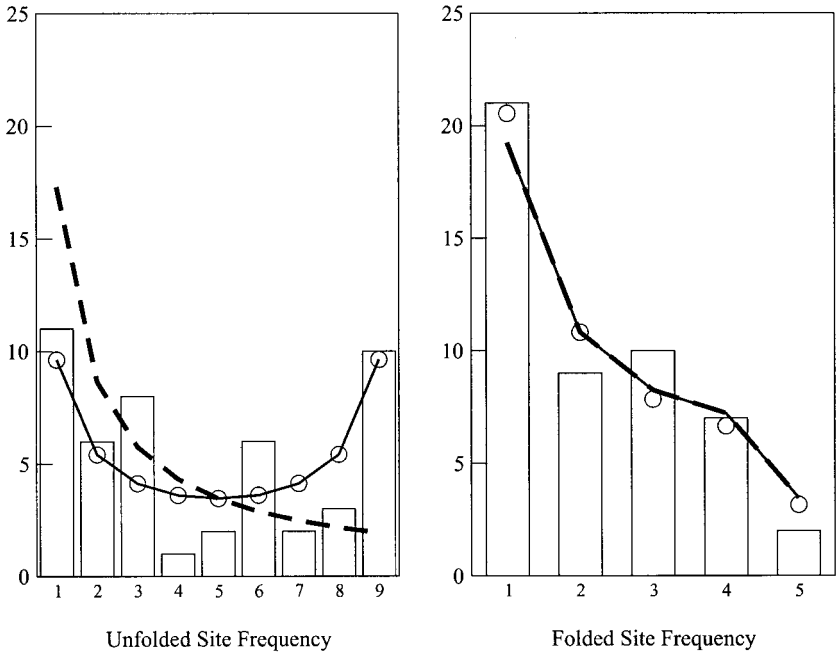

FIGURE 5.—Data and fitted values for the example discussed in the text. The bar graph is the data and the line graphs are the fitted values under the neutral model (dashed line), infinitely strong positive selection model (solid line), and under the MLEs of the PRF model (circles).

neutrality even in moderate sample sizes when the ancestral states of variable sites are known (*i.e.*, for unfolded site configurations). A surprising result is that for folded site configurations the test has strong power to detect negative selection but very weak power to detect positive selection. The reason for this is related to the issue of point estimation in general in the PRF framework.

From general-likelihood theory, we expect the maximum-likelihood estimate for a parameter under certain regularity conditions to be consistent, unbiased, and efficient as the sample size used to estimate the parameter tends to infinity. By "consistent" we mean that the parameter estimate will converge in probability to the true value of the parameter (*i.e.*, the variance in the parameter estimate will tend toward 0), by "unbiased" we mean that the expected value of the parameter estimate is centered on the true value of the parameter, and by "efficient" we mean that the estimator has the lowest possible variance of all possible estimators. In practice,

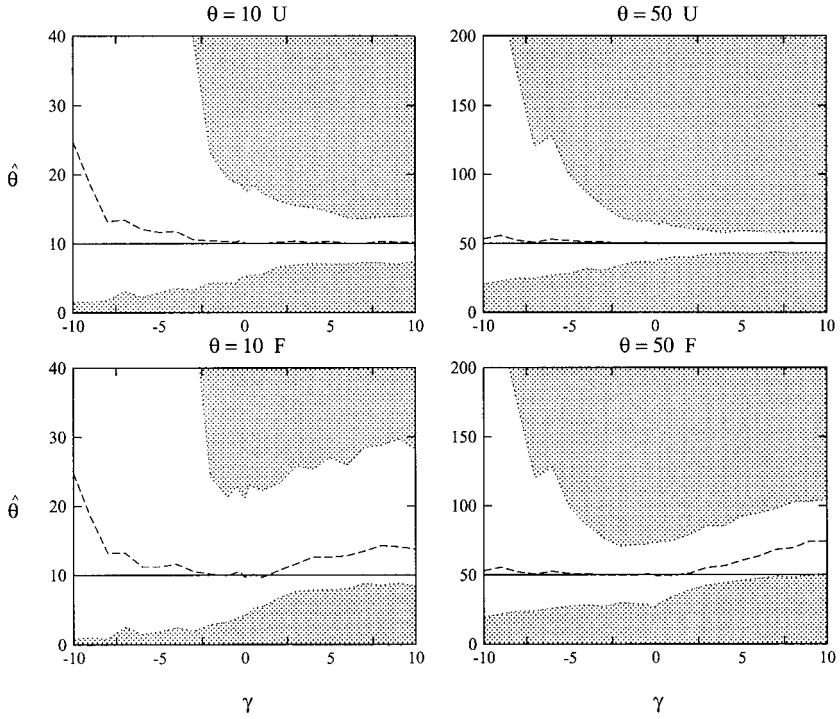

FIGURE 6.—Distribution for the maximum-likelihood estimates of θ as a function of the simulated value of γ. The open region corresponds to the space containing 95% of the observed θ̂'s.

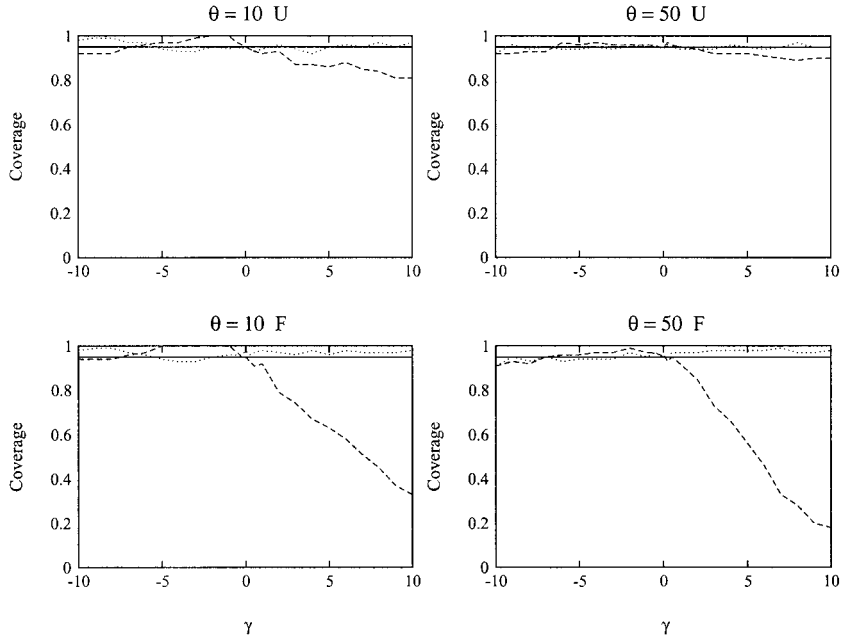

FIGURE 7.—Coverage of 95% confidence intervals for γ as a function of the simulated value of γ for two mutation levels for both folded (F) and unfolded (U) site-frequency spectra. Dotted lines connect the percentage of confidence intervals that contain the true γ for confidence intervals based on l^* and dashed lines connect the percentage of confidence intervals that contain the true value for confidence intervals based on the normal approximation and asymptotic variance.

though, maximum-likelihood estimation can often behave very poorly if the sample size is not large enough or if the regularity conditions are not met.

For the levels of mutation examined in our generated data, we find that the distributions for $\hat{\gamma}$ are reasonably centered around the true value of γ for the unfolded model (Figure 3) with increasing variance as one moves away from neutrality. We note that in regions of extreme selection ($|\gamma| > 5$) the distribution for $\hat{\gamma}$ becomes quite skewed with $\hat{\gamma} = \infty$ or $\hat{\gamma} = -\infty$ for certain replicate data sets. The reason for this is that the log-likelihood function becomes very flat when the magnitude of γ is very large. In the case of negative selection, samples consisting of only singletons are consistent with extreme negative selection and very high mutation, so that $\hat{\gamma}$ becomes indistinguishable from $-\infty$ and we show shortly why the likelihood function asymptotes to ∞ for any given data set. These results indicate that these extrema represent either a local maximum or minimum that needs to be checked once an MLE for γ has been estimated. If it is found that the likelihood at one of these two extremes is higher than the estimate $\hat{\gamma}$, this is problematic for point estimation but not for hypothesis testing since we can still test whether the observed sample configuration is consistent with a neutral model (for most of the cases we examined where $\hat{\gamma} = \infty$ or $\hat{\gamma} = -\infty$ for unfolded data, the LRT is significant).

While maximum-likelihood estimation works reasonably well for point estimation using unfolded data, the method performs very poorly for folded data in the region of positive selection as alluded to above in the section on power. As we see in Figure 3, for the folded data the method performs well in terms of estimating γ for the region of negative selection, performs reasonably well for a small region of positive selection, and then asymptotes to neutrality for arbitrarily large γ.

For a better understanding of this phenomenon, let us focus on one of the replicate data sets we generated under $\theta = 10$ and $\gamma = 10$, $x = (11, 6, 8, 1, 2, 6, 2, 3, 10)$. For these data, $\hat{\theta} = 8.66$, $\hat{\gamma} = \infty$, and LRT = 18.08 ($P < 0.0001$) for the unfolded model. We note that the profile-likelihood function for the unfolded data is monotonic with a peak at ∞ (Figure 4). As expected from the LRT, we see in Figure 5 that the predictions from the infinite selection model are quite different from the predictions of the neutral model. For the folded model, $\hat{\theta} = 20.03$, $\hat{\gamma} = -0.473$, and LRT = 0.18 ($P = 0.328$). As expected from the LRT, we see in Figure 5 that the folded site-frequency spectrum is fitted very well by a neutral model. What may be surprising is that the data are equally well fit by a model with infinitely strong positive selection! Another interesting property of these data is that the likelihood function for folded data configurations has a bimodal nature with a region in between that can be significantly different from either peak, implying that the data are consistent with either neutrality or strong selection but not weak selection (Figure 4).

The reason for this phenomenon is that when the site configurations are folded, the limiting distribution as $\gamma \to \infty$ is indistinguishable from neutral evolution with a mutation rate that is twice as large as the true mutation rate. This can be shown by noting that the log-likelihood function for a given data set will reach finite asymptotic value for $\gamma \to \infty$ since

$$\lim_{\gamma \to \infty} F(i, \gamma) = \frac{n}{(n - i)\, i}. \tag{33}$$

Recalling that $G(i, \gamma) = F(i, \gamma) + F(n - i, \gamma)$ and the result derived earlier that $F(i, 0) = 1/i$, we can easily see that

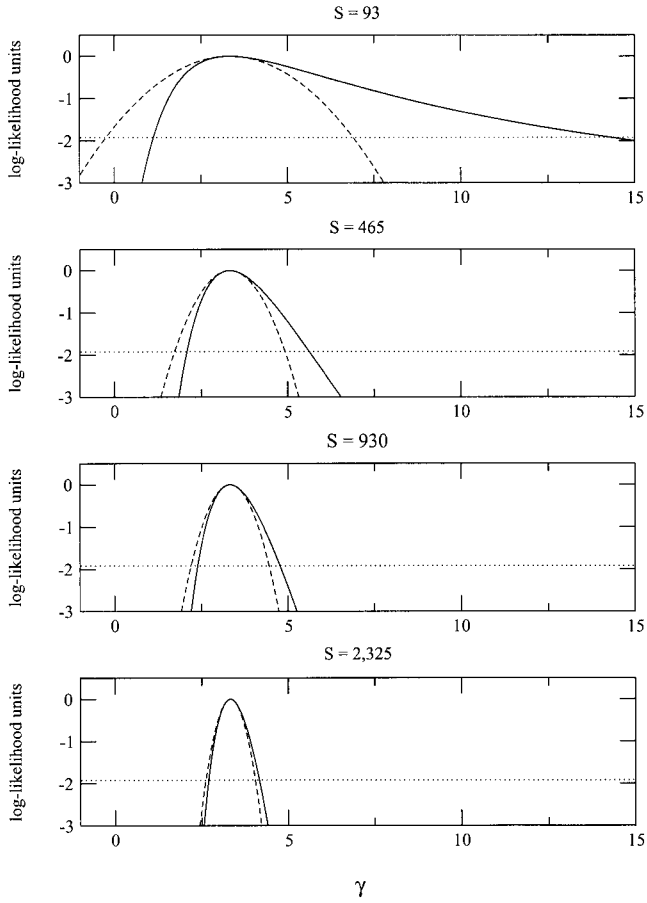

FIGURE 8.—Log-profile-likelihood function (solid line) and its normal approximation (dashed line) under varying sample size. (*S* represents the total number of segregating sites in the sample). The dotted line corresponds to 1.92 log-likelihood units from the maximum. The intersection of this line and the log-profile likelihood or its approximation corresponds to the 95% confidence interval constructed from the method.

$$G(i, \infty) = \frac{2n}{(n - i)i} \tag{34}$$

$$G(i, 0) = \frac{n}{(n - i)i}, \tag{35}$$

leading to a bimodal log-likelihood function in the extreme case with maxima near 0 and ∞ as seen in the example above.

It is important to note that this is a general problem for methods that attempt to detect selection using folded site configurations [*e.g.*, TAJIMA's (1989) *D*-statistic]. This also explains the odd result reported in AKASHI (1999) that Tajima's *D*-statistic has little to no power to detect positive selection. Since this is a general problem for estimating selection from population genetics data, the inclusion of a parameter for recombination would make the problems of inference even worse, since in this case we know that the data conform to the assumptions of the model.

One interesting point related to this issue is the effect of the estimation of γ on the estimation of θ. We see from Figure 6 that estimation of θ is excellent for the unfolded model for regions of positive selection. It may seem paradoxical that while the variance in the estimates of γ increases with increasing level of selection, the variance in the estimates of θ decreases. The reason for this is that once selection becomes large the predictions of the PRF model for the site-frequency spectrum become indistinguishable (this is exactly the reason for the asymptote at ∞ for positive selection), but only a small range of θ is consistent with the observed number of segregating sites. In the case of negative selection, the results are reversed since as selection increases in the negative selection the values for $F(i, \gamma)$ become very small, leading to large variance in the estimate of both selection and mutation.

For the folded model, in Figure 6 we begin to see the effect of the maximum near 0 for $\hat{\gamma}$ on the estimation of mutation. Watterson's estimator of θ will always be larger than the estimated value of θ under positive selection, since $F(i, \gamma)$ increases with increasing γ and reaches its maximum as explained above at twice the value of the partial harmonic sum in Watterson's estimator. This means that since we tend to underestimate selection in the folded model for positive selection, we will tend to overestimate θ.

Confidence sets for γ: The last issue we explore is the construction of confidence intervals for the selection parameter in the PRF framework. As we outlined above our two methods for constructing confidence intervals are (1) the region that contains $(1 - \alpha) \cdot 100\%$ of the area in the normal approximation to likelihood function and (2) the region in the profile likelihood space that is $<0.5\chi^2_{1,1-\alpha}$ likelihood units from the maximum-likelihood point. For the first confidence set, at $\alpha = 0.05$ this corresponds to the area that is 1.96 standard deviations away from $\hat{\gamma}$ where the standard deviation is given by $\sqrt{\mathrm{VAR}(\hat{\gamma})}$. Likewise for the second set, at $\alpha = 0.05$ this corresponds to the region where $l^*(\hat{\gamma}) - l^*(\gamma) < 3.84/2$ likelihood units.

We see from Figure 7 that confidence intervals based on the normal approximation to the likelihood tend to undercover (*i.e.*, they tend to be too small), particularly for the folded model and for strong positive selection. The confidence intervals based on the profile likelihood, however, have excellent coverage regardless of the folding of the data or the strength of selection.

The reason for this is seen in Figure 8, where we plot the log-profile-likelihood function and its normal approximation for a data set with the same $\hat{\gamma}$ but varying sample sizes (in terms of the number of segregating sites for the sample). We see that the number of segregating sites needs to be quite large before the normal approximation begins to have the same coverage as the profile likelihood in a region of 1.92 log-likelihood units. This occurs because a confidence interval based on the normal approximation will, by definition, be symmetric, and the log-likelihood function is quite right

skewed (as we explained above). Since we need an immensely large sample for the region around the maximum-likelihood point to be normally distributed (upward of 1000 segregating sites), confidence intervals based on the normal approximation will tend to undercover.

CONCLUSIONS

The Poisson random field model for site-frequency data holds great potential for untangling the effects of mutation and selection on standing genetic variation. Our analysis suggests, however, that when the data do not conform to the assumptions of very loose linkage, the estimates from this model may be misleading. Likewise, if the data set being analyzed is small, one should be cautious about the asymptotic assumptions implicit in maximum likelihood as a method of point estimation. This is a general problem for methods that attempt to estimate selection using site-frequency data, and methods that attempt to incorporate recombination into point estimation will only make some of these problems worse. The solution may be to focus on genetic variation that is independent (such as the growing number of single nucleotide polymorphisms for many organisms) rather than focusing on methods that detect selection at single loci. We also find that folding the site-frequency spectrum obscures positive selection and suggest that every effort should be made to incorporate information on the ancestry of mutations being studies, if one wishes to estimate selection.

We also point out that simply adding information on the number of fixed differences in the region between the study population and a close species will not aid in point estimation for selection. Such a method would add exactly one data point and one parameter (time since divergence), which can be scaled to fit that point exactly. A method that uses two classes of sites (*e.g.*, silent and replacement) and fixed differences circumvents this problem and will add useful information for the estimation of selection. We have developed a method that uses divergence data and the site-frequency spectrum to estimate selection. Likewise, information on variation in the site-frequency spectrum among different genomic regions also holds great potential for refining point estimation for single regions of interest. We have also developed a method that uses Markov chain Monte Carlo to perform Gibbs and Metropolis sampling for a hierarchical model of this form. We will expand upon these two methods in subsequent publications.

We thank Richard Lewontin, Rasmus Nielsen, and Steve Palumbi for many helpful comments. This work was supported by a Howard Hughes Medical Institute Graduate Fellowship to C.D.B. and National Science Foundation grant DMS9707045 to S.A.S.

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.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

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

Hartl, D., E. F. Boyd, C. D. Bustamante and S. Sawyer, 2000 The glean machine: what can we learn from DNA sequence polymorphism, pp. 37–49 in *Genomics and Proteomics: Functional and Computational Aspects*, edited by S. Suhai. Plenum Press, New York.

Kendall, M., 1987 *Kendall's Advanced Theory of Statistics*, Vol. 2. Oxford University Press, Oxford.

Kingman, J., 1993 *Poisson Processes*. Oxford University Press, Oxford.

Lange, K., 1999 *Numerical Analysis for Statisticians*. Springer-Verlag, New York.

Press, W. H., B. P. Flannery and S. A. T. W. T. Vetterling, 1988 *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.

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

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

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

Wright, S., 1969 *Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Communicating editor: N. Takahata