A Composite-Likelihood Approach for Detecting Directional Selection From DNA Sequence Data

Lan Zhu and Carlos D. Bustamante

Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853

Manuscript received August 17, 2004
Accepted for publication March 30, 2005

ABSTRACT

We present a novel composite-likelihood-ratio test (CLRT) for detecting genes and genomic regions that are subject to recurrent natural selection (either positive or negative). The method uses the likelihood functions of Hartl et al. (1994) for inference in a Wright-Fisher genic selection model and corrects for nonindependence among sites by application of coalescent simulations with recombination. Here, we (1) characterize the distribution of the CLRT statistic (A) as a function of the population recombination rate ($R = 4N_e r$); (2) explore the effects of bias in estimation of $R$ on the size (type I error) of the CLRT; (3) explore the robustness of the model to population growth, bottlenecks, and migration; (4) explore the power of the CLRT under varying levels of mutation, selection, and recombination; (5) explore the discriminatory power of the test in distinguishing negative selection from population growth; and (6) evaluate the performance of maximum composite-likelihood estimation (MCLE) of the selection coefficient. We find that the test has excellent power to detect weak negative selection and moderate power to detect positive selection. Moreover, the test is quite robust to bias in the estimate of local recombination rate, but not to certain demographic scenarios such as population growth or a recent bottleneck. Last, we demonstrate that the MCLE of the selection parameter has little bias for weak negative selection and has downward bias for positively selected mutations.

The evolutionary fate of a mutation is governed by genetic drift, demographic history, and natural selection acting directly on the mutation or indirectly through the effect of mutations at linked loci. A central goal of population genetics is to quantify the role of each of these factors in the evolution of particular loci in particular populations (Lewontin 1974). The Poisson random field (PRF) approach (Sawyer and Hartl 1992; Hartl et al. 1994; Wakeley 2003; Williamson 2003, Williamson et al. 2005) has proven quite useful in estimating mutation and selection parameters in various population genetics settings when DNA mutations are unlinked. The inference rationale behind the approach is that natural selection will alter the site-frequency spectrum [SFS; i.e., the number of mutations at a frequency 1 out of $n$, 2 out of $n$, . . . ($n - 1$) out of $n$, where $n$ is the number of sequences sampled], making it possible to estimate the strength of selection needed to explain observed deviations from the neutral SFS expectations. A likelihood-ratio test (LRT) derived from the PRF has shown to be quite powerful and maximum-likelihood estimation of mutation and selection parameters performs very well when the ancestral states of all mutations in the sample are known (Bustamante et al. 2001). The PRF approach also has the advantage of using all of the information in the SFS regarding natural selection as opposed to traditional summary statistics of the data such as Tajima’s (1989) $D$ and the methods of Fay and Wu (2000) and Fu and Li (1993).

Unfortunately, since the PRF model assumes independence among sites, the application of the LRT to most genetic data is quite limited unless the assumption of free recombination among sites can somehow be relaxed. One can imagine two potential solutions to the problem: (a) explicitly modeling natural selection and recombination to evaluate the true likelihood function via the ancestral selection graph (Krone and Neuhauser 1997; Neuhauser and Krone 1997; Slade 2001), and (b) taking a “composite-likelihood” approach by continuing to treat sites as independent and then correcting parameter estimates and critical values for the LRT via simulation. From a statistical point of view, the former approach is more desirable, since the likelihood function contains all the information about natural selection available in the data (e.g., distribution of haplotypes, patterns of linkage disequilibrium). Unfortunately, full-likelihood inference is so computationally costly as to be out of reach for practical sample sizes at single loci and certainly out of reach for genome-wide analyses. Composite likelihood has recently been used in population genetics to reduce the computational complexity of various inference problems, for example, estimating recombination rate (Hudson 2001; Fearn-
TABLE 1
Notations used

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_e$</td>
<td>Effective population size</td>
</tr>
<tr>
<td>$r$</td>
<td>Per-locus recombination rate per generation; $R = 4N_e r$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Per-locus mutation rate per generation; $\theta = 4N_e \mu$</td>
</tr>
<tr>
<td>$s$</td>
<td>Relative fitness of the mutant; $\gamma = 2N_e s$</td>
</tr>
<tr>
<td>$m$</td>
<td>Migration rate (proportion of migrants in the subpopulation per generation); $M = 4N_e m$</td>
</tr>
<tr>
<td>$D$</td>
<td>No. of demes</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Population growth rate</td>
</tr>
<tr>
<td>$n$</td>
<td>No. of sequences sampled</td>
</tr>
<tr>
<td>$S$</td>
<td>Total no. of segregating sites in the sampled sequences</td>
</tr>
<tr>
<td>$Q$</td>
<td>No. of replicates in the simulation study</td>
</tr>
<tr>
<td>$R_h$</td>
<td>Recombination rate estimator from Hudson (1987)</td>
</tr>
<tr>
<td>$R_{he}$</td>
<td>Recombination rate estimator from Hey and Wakeley (1997)</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Test statistic of the CLRT</td>
</tr>
<tr>
<td>$\Lambda^*$</td>
<td>95% critical value of the CLRT</td>
</tr>
<tr>
<td>$t_{bw}$</td>
<td>Time in the unit of $4N_e$ generations ago that the bottleneck happens</td>
</tr>
<tr>
<td>$t_{bw}'$</td>
<td>Time in the unit of $4N_e$ generations ago that the population recovers from the bottleneck</td>
</tr>
<tr>
<td>$f$</td>
<td>Ratio of the population size during bottleneck to the original size</td>
</tr>
</tbody>
</table>

**head and Donnelly** 2002; McVean et al. 2002; Wall 2004), variation in recombination rate (Li and Stephens 2003; McVean et al. 2004), and detecting local signature of hitchhiking (Kim and Stephan 2002; Kim and Nielsen 2004; Innan and Kim 2004). Therefore, due to practical motivations, the latter approach is investigated here, since the composite-likelihood solution for a single locus can easily scale to genome-wide levels and can be expanded to include increasingly realistic demographic scenarios.

In this article, we set out to investigate the performance of a composite-likelihood-ratio test for recurrent directional selection under varying levels of selection, mutation, and recombination while relaxing the assumption of independence among sites. The initial motivation for this project was Bustamante et al.’s (2001) result that the LRT proposed by Hartl et al. (1994) is not robust to deviations from the assumption of independence among sites (i.e., the test has a much higher type I error than expected). The basic idea behind our new test is that by modifying the critical value of the LRT statistics a proper test is constructed with desired size (type I error, $\alpha$). Specifically, one can correct the critical value of the LRT statistic by applying Hartl et al.’s (1994) test to a large number of simulated replicate neutral data sets generated conditional on $S$, the observed number of segregating sites in the original data, and $R$, an estimate of the population recombination rate, and choosing a cutoff such that $\alpha$ of the simulations are below the critical value (Table 1). We refer to this test as the composite-likelihood-ratio test (CLRT) to distinguish it from the LRT designed for independent data and to signify that we are not dealing with the true likelihood function of the data under recombination and selection, but rather an approximate likelihood function. If the data come from a population with the same demography we have used for our neutral simulations (e.g., a panmictic population of constant size) and our estimate of the recombination rate is accurate, such a test would be guaranteed not to reject neutrality more often than expected (namely, $100 \times \alpha\%$ of the time).

A potential pitfall of all “tests of neutrality” is that there are several putative alternative hypotheses to the single null hypothesis of a neutrally evolving panmictic population of constant size. In other words, certain demographic scenarios affect the site-frequency spectrum in ways that may be indistinguishable from natural selection, leading to spurious rejection of the null hypothesis of neutral evolution. For example, low levels of migration among subpopulations will elevate the proportion of observed high-frequency-derived mutations above their neutral expectation in a panmictic population much in the same way as positive selection (Nielsen 2001; see also Figure 1). Likewise, a sample of DNA sequences drawn from an exponentially growing population can look like a sample drawn from a population of constant size subject to weak negative selection (i.e., both scenarios lead to an excess of low-frequency variants vis-à-vis neutrality). Therefore, it is imperative to investigate the robustness of our proposed CLRT to demographic history and population structure. This can be accomplished via standard coalescent simulations with recombination (Hudson 2002). In this article we consider a wide range of demographic scenarios, including exponential population growth, an island model of migration, and bottlenecks of varying severity. Likewise, it is important to explore the power of the CLRT (the probability of rejecting the null hypothesis when it is false) as well as the bias of the maximum composite-likelihood estimates (MCLE) of mutation rates and strength of selection. We use a forward simulation program with recombination and selection to address these last two issues. We also investigate the discriminatory power of the CLRT to detect weak negative selection when the null model is exponential growth instead of a panmictic population of constant size.

**MATERIALS AND METHODS**

**Theory**

To model the effects of natural selection on the site-frequency spectrum, several assumptions are made within the standard PRF models (Sawyer and Hartl 1992; Hartl et al. 1994): (a) panmictic population of constant size, (b) weak selection with no dominance, (c) equal selective effects of all nonlethal mutations, (d) free recombination among segregating sites, (e) infinite-sites mutation model, and (f) no epistatic
effect among mutations. Wakeley (2003) has developed models that relax assumption (a) by considering an infinite-demes population structure; Williamson et al. (2004) have developed PRF models with dominance, relaxing assumption (b); and Bustamante et al. (2003) and Sawyer et al. (2003) have modeled the effects of a distribution of selective effects among nonlethal mutations [relaxing assumption (c)]. The purpose of this article is to relax assumption d for the purpose of nonlethal mutations [relaxing assumption (c)]. The purpose of this article is to relax assumption d for the purpose of nonlethal mutations [relaxing assumption (c)]. The purpose of this article is to relax assumption d for the purpose of nonlethal mutations [relaxing assumption (c)].

Let $X = \{X_1, X_2, \ldots, X_{n-1}\}$ represent the site-frequency spectrum for a genomic region of interest such that $X_i$ is the number of sites along the sequence that have $k$ derived mutations and $n - k$ ancestral mutations, where $n$ is the number of sampled sequences (throughout we assume the directionality of mutation is known). In their original article, Sawyer and Hartl (1992) modeled the site-frequency spectrum as a collection of independent Poisson-distributed random variables, governed by the mutation parameter $\theta = 4N_\mu$ and a selection parameter $\gamma = 2N_\mu s$, where $N_c$ is the haplot population size; $s$ is the fitness effect of new mutations such that wild-type fitness is 1, heterozygote fitness is 1 + $s$, and homozygote fitness for the new mutation is 1 + 2$s$; and $\mu$ is the per locus mutation rate for mutations with selective effect $s$. It is important to note that the site-frequency spectrum of selected sites is sensitive to assumptions regarding dominance. The PRF model proposed by Williamson et al. (2004) for analyzing the site-frequency spectrum under dominance and selection is also amenable to the type of modification we propose here. For simplicity and ease of computation, we consider only the genic selection case here. When $\gamma = 0$, the population is evolving neutrally; when $\gamma > 0$, it is under positive selection; and when $\gamma < 0$, it is subject to negative selection. According to the results of Sawyer and Hartl (1992), $X_i$ is a Poisson-distributed random variable with mean $\theta F_k(\gamma)$, where

$$F(k, \gamma) = \int_0^1 \frac{1}{k!} p^{k}(1 - p)^{n-k-1} \frac{1 - e^{-\gamma(1-p)}}{(1 - e^{-\gamma})} (1 - p) \, dp.$$ 

Since under the model the $X_i$’s are independent, the likelihood function $L(\theta, \gamma|X)$ is the product over $P(X_i = x_i|\theta, \gamma)$, namely

$$L(\theta, \gamma|X) = \prod_{i=1}^n \frac{e^{-\theta F_k(\gamma)} (\theta F_k(\gamma))^{x_i}}{x_i!}.$$ 

Our null hypothesis is that the population evolves neutrally, $H_0: \gamma = 0$, while the alternative hypothesis is the complement, $H_1: \gamma \neq 0$. The likelihood-ratio test statistic proposed by Hartl et al. (1994) and investigated by Bustamante et al. (2001) for comparing these hypotheses is

$$L = (-2 \log L(\hat{\theta}_0, \hat{\gamma}|X)) - (-2 \log L(\hat{\theta}, \hat{\gamma}|X))$$

$$= 2 \left( \sum_{i=1}^{n-1} x_i \log F(i, \hat{\gamma}) - S \log \sum_{i=1}^{n-1} F(i, \hat{\gamma}) + \sum_{i=1}^{n-1} \log i \right).$$

where $\hat{\theta}$ is the MLE of $\theta$ under the neutrality, which turns out to be Ewens’ (1974) and Watterson’s (1975) estimator of $\theta$; $\hat{\theta}$ and $\hat{\gamma}$ are the maximum-likelihood estimates of $\theta$ and $\gamma$, respectively, under the full model with selection, found by maximizing the profile log-likelihood function as described in Bustamante et al. (2001), and $S = \Sigma x_i s_i$; is the observed number of segregating sites. Under the assumption of independence among sites, $L$ is asymptotically $\chi^2$ distributed (Kendall 1987).

It has been shown by simulation that if sites are not evolving independently, the $\chi^2$-approximation is too liberal and the LRT will have an unacceptably high type I error (Bustamante et al. 2001). The reason for this is that the likelihood of the data in the presence of linkage is not simply the product of the likelihood across SNPs. That is, if sites are linked, Equation 1 is not the true-likelihood function of the data, but rather a composite-likelihood function, and the LRT statistic no longer corresponds to a true likelihood-ratio test, but rather to a CLRT. Under such a scenario the distribution of the test statistic is no longer $\chi^2$, but rather depends on the rate of recombination among sites. We must, therefore, use coalescent simulations with recombination to find the critical value $L^*$ for the test statistic whenever we wish to analyze data with linkage among SNPs. While the LRT has been shown to have excellent power and $\hat{\theta}$ and $\hat{\gamma}$ have been shown to have little bias under the independence assumption (Bustamante et al. 2001), nothing is known about the statistical properties of the CLRT or the composite maximum-likelihood estimates of $\theta$ and $\gamma$. The algorithm we employ for calculating the CLRT is as follows:

**Algorithm 1: Composite-likelihood-ratio test:**

1. Given an observed site-frequency spectrum, $X_{\text{obs}}$, estimate $\hat{\theta}$ and $\hat{\gamma}$ using the one-dimensional optimization described in Bustamante et al. (2001), and calculate the CLRT statistic $L_{\text{obs}}$ via Equation 2.

2. Generate $Q$ replicate data sets $X_1, X_2, \ldots, X_Q$ from a standard neutral model with recombination rate $R$ corre-

FIGURE 1.—Comparison of expected site-frequency spectra for three scenarios. “Neutral” is the expected SFS under the standard neutral model (see Hudson 1990). “Population structure” is the expected site-frequency spectrum for neutral mutations in a two-deme model with low symmetric migration rate ($4Nm = 0.2$) found via 1000 coalescent simulations using ms (Hudson 2002). “Selection” is the expected SFS under genic selection for the model described by Hartl et al. (1994). We use a value of $2N_\mu = 1.353$, which maximizes the likelihood of the expected population structure data under the selected model. As one can see, the site-frequency spectrum under population structure can look similar to that under recurrent positive selection.
spending to the region of interest and $S$ to the observed number of segregating sites in $X_{OBS}$. Apply the optimization in step 1 to each of the replicate data sets and generate the replicate CLRT statistics $\Lambda_1, \Lambda_2, \ldots, \Lambda_{10}$.

3. The $P$-value for the CLRT corresponding to data $X_{OBS}$ is estimated as $P(\Lambda_{OBS} | H_0) \approx (\sum_{i=1}^{10} 1(\Lambda_{OBS} \leq \Lambda_i)) / Q$, where $I(\cdot)$ is the indicator function that evaluates to 1 if the argument is true and 0 otherwise.

To implement our proposed composite-likelihood method, one needs an estimate of the local recombination rate. Given a data set, two main approaches have been employed for estimating $R$. One method is based on observing the frequency of sequence exchange between markers (Ashburner 1989; True et al. 1996; Bouffard et al. 1997; Nagaraja et al. 1997), and the other method estimates $R$ from the patterns of sequence variation expected in a random sample of DNA sequences from a population (Hudson and Kaplan 1985; Hudson 1987, 2001; Griffiths and Marjoram 1996; Hey and Wakeley 1997; Wakeley 1997; Kuhner et al. 1999; McVean et al. 2002; Wall 2004). Since here we are interested in the local recombination rate, only the latter approach is applicable since a fine-scale estimate requires hundreds of thousands of meioses. Assuming a constant population size model and no selection, Hudson (1987) proposed a method-of-moments approach that transforms the observed variance in pairwise nucleotide differences into an estimate of the population recombination rate. Hey and Wakeley (1997) derived a method of estimating the recombination rate via coalescent theory based on considering multiple subsets of four sequences. Simulation study (Wall 2000) shows that both Hudson’s ($R_h$) and Hey and Wakeley’s estimators ($R_{HW}$) perform well with large sample size (e.g., $n = 50$) and improve as the mutation rate increases. However, comparing these two estimators with eight others, Wall (2000) demonstrated that $R_h$ overestimates $R$ (a large proportion of $R_h / R > 5.0$); while $R_{HW}$ underestimates $R$ (with the majority of $R_{HW} / R < 0.2$). In this article, we explore these two estimators since they represent the extreme effects of over- and underestimating the local recombination rate and are computationally very fast.

Simulations

To explore the statistical properties of the CLRT as well as the MCLE of the selection and mutation parameters, we simulated five different types of data (Hudson’s 2002 “ms” program was used for all coalescent simulations). The first type of data is neutral from a population of constant size. These data were used to explore how quickly the CLRT statistic $\Lambda$ converges to a $\chi^2$-distribution as a function of $R$ and to compare the effect of different estimators of recombination rates on the realized size of the test. The second, third, and fourth types of data were neutral data from (a) a single subpopulation in an island model, (b) a panmictic population that had recently expanded in size, and (c) a panmictic population that had undergone a single bottleneck. These data were used to explore the effect of these demographic factors on the type I error of the test. The fifth type of data was generated by the forward infinite-sites simulation with selection and recombination (FISHER) program written in ANSI C by L. Zhu. FISHER was used to generate polymorphism data with recurrent selection and recombination under an infinite-sites model assuming constant population size. We ran FISHER with $10N_e$ generations of burn-in and replicate data sets sampled every $2N_e$ generations. These data were used to explore the power of the test under varying levels of mutation and selection, as well as recombination.

Robustness simulations: To explore the null distribution of the CLRT statistic, we generated neutral data from a population of constant size for seven levels of recombination, $R \in \{0, 1, 5, 10, 50, 100, 1000\}$, using Hudson’s (2002) ms program. For each of these sample sizes ($n = 10, 50, 100$), we simulated 1000 replicate data sets with a fixed number of segregation sites ($S = 100$) and constant recombination rate. For each replicate, we apply the CLRT and retain the test statistic $\Lambda$. The distribution of the CLRT statistic and the trend that $\Lambda$ varies greatly with $R$ are plotted in Figures 2 and 3, respectively.

In practice, the true recombination rate for sampled sequences is unknown and must be estimated from data. We were interested in investigating the effect of estimation bias in the recombination rate on the type I error of the CLRT. As Wall (2000) showed, there is no single best estimator of $R$ and, in practice, most estimators do poorly if $R$ is close to 0. Here we explored Hudson’s (1987) and Hey and Wakeley’s (1997) estimators since they tend to overestimate and underestimate $R$, respectively, for a broad range of values. Since Hudson’s estimator has low reliability if data sets are not very large (Hudson 1987), we simulated data with sample size $n = 50$ and fixed segregating sites at $S = 100$. The detailed algorithm is as follows and the results of our analysis are summarized in Figure 4.

Algorithm 2: Estimating realized type I error of CLRT when $R$ is estimated from data:

1. Generate neutral data $X_{OBS}$ with known recombination rate $R$ and apply the CLRT to obtain the test statistic $\Lambda_{OBS}$. For each replicate, perform CLRT and keep test statistic $\Lambda$. The empirical (1 - $\alpha$) quantile of the distribution of $\Lambda$ among the 1000 replicates is the critical value of the test statistic $\Lambda^*$ at a-level (for all simulations we used $\alpha = 0.05$). Similarly, we can find $\Lambda^*$ with estimated recombination rate $R_{est}$.

3. Generate $Q = 1000$ replicate data sets with the same sample size and number of segregating sites as $X_{OBS}$ under the estimated recombination rate $R_i$. For each replicate, perform CLRT and compute the test statistic $\Lambda$. The distribution of $\Lambda$ offers the realized size of the test statistic $\Lambda^*$ at a-level.

4. If $\Lambda_{OBS} > \Lambda^*$, reject the neutral hypothesis; otherwise, fail to reject at the $\alpha = 0.05$ level.

5. Repeat steps 1–4 1000 times. The proportion of the false rejection is the realized size of the CLRT under the PRF model when the recombination rate is not known.

In the current model, we assume no population structure to the data. We are interested in investigating how well the CLRT performs when this assumption is violated. We simulated the second type of data with sample size $n = 50$ and fixed number of segregating sites ($S = 100$), $R = 0$ under the island model for $D \in \{2, 5, 10, 20, 50\}$ ($D$ is the number of demes), and $0 \leq M \leq 15$ ($M = 4N_m$, where $m$ is the fraction of each deme made up of new migrants each generation). The reason for fixing the number of segregating sites is that the distribution of the number of segregating sites changes with the migration rate if we fix the overall mutation rate of the entire population (Wakeley 2001). When we explore the effect of the migration rate on the size of the test, we want to control for the effect that is caused by the difference in the number of segregating sites. The detailed procedure is as follows and the results of this analysis are shown in Figure 5.

Algorithm 3: Procedure for estimating realized type I error of CLRT in the presence of population structure/demographic history.

1. Generate $Q = 10,000$ data sets with $S$ segregating sites from a panmictic population of constant size. Estimate the
4. Repeat steps 2 and 3 1000 times for each parameter combination—i.e., the effects of exponential growth—given by $N(t) = N_0 \exp(-\beta t)$, where $N_0$ is the present population size, $t$ is the time before present, measured in units of $4N_e$ generations, and $\beta$ is the growth rate. To explore the realized type I error of the CLRT, the proportion of data sets that reject neutrality (i.e., number of data sets out of 1000 with $\Lambda_{\text{obs}} > \Lambda^*$) is the assumed constant population size. To explore the realized type I error of the CLRT, the proportion of data sets that reject neutrality (i.e., number of data sets out of 1000 with $\Lambda_{\text{obs}} > \Lambda^*$) is the realized type I error of the CLRT.

Another assumption of the PRF model that may be problematic is the assumption of constant population size. To explore the effects of exponential growth [i.e., the population size is given by $N(t) = N_0 \exp(-\beta t)$, where $N_0$ is the present population size, $t$ is the time before present, measured in units of $4N_e$ generations, and $\beta$ is the growth rate], we modify step 2 of the above algorithm and generate data within ms for rates of growth $\beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ (Figure 6). For a bottleneck, we simulate data for $n \in \{10, 100\}$ with various recombination rates $R \in \{0, 10, 100\}$, assuming the bottleneck happened at $t_b = 0.025$ or $0.05$ (in the unit of $4N_e$ generations before the current sampling time and recovered to the current population size at $t_b = 0.0125$ (in the unit of $4N_e$ generations). We consider two levels of the population reduction during the bottleneck, i.e., $f \in \{0.1, 0.01\}$ (Figures 7 and 8).

**Power simulations:** To evaluate the power of the CLRT (i.e., probability of rejecting a false null hypothesis of neutrality), we wrote a forward simulation program, FISHER, to simulate a genomic region under recurrent selection and recombination using an infinite-sites model of mutation assuming constant population size. Power was estimated as the proportion of replicates generated under selection for which the null hypothesis was rejected by the CLRT. The detailed algorithm is as follows and results of the FISHER simulations can be found in Figures 9–11.

Algorithm 4: Procedure for estimating the power of the CLRT:

1. **Generate a data set of $n$ sequences via FISHER given mutation rate, selection coefficient, and recombination rate; apply the CLRT to obtain the test statistic $\Lambda_{\text{obs}}$ and corresponding $P$-value from algorithm 1.**
2. **If $P < 0.05$, reject the neutral hypothesis; otherwise, fail to reject at the $\alpha = 0.05$ level.**
3. **Repeat steps 1 and 2 500 times and calculate the power of the test as the proportion of rejections.**

The power of the test above is based on estimating the $P$-value assuming a constant population size. Since population growth may have an effect similar to negative selection, we examine how powerful the CLRT is in distinguishing negative selection from an exponentially growing population model. For these simulations, the data sets were simulated via FISHER given the selection coefficient, mutation rate, and recombination rate. The critical value of the CLRT was determined assuming the population has been growing exponentially, and mutations were neutral. We sampled 50 sequences with mutation parameter $\theta = 30$, recombination rate $R = 100$, and selection coefficient $\gamma \in \{-1, -5, -10\}$. We analyzed the power of the CLRT in distinguishing negative selection from the exponential growth rate $\beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$. To achieve this, we modify step 2 in algorithm 1 and simulate data under an exponential growth model. All other steps remain unchanged. (Results are shown in Figure 11.)

**RESULTS AND DISCUSSION**

How quickly does the test statistic ($\Lambda$) converge to a $\chi^2$-distribution? From Figure 2, we confirm the theoretical prediction that the composite-likelihood-ratio test statistic $\Lambda$ converges to a $\chi^2$-distribution as recombination rate increases; unfortunately, the convergence rate is very slow. From Figure 3, we can see that the 95% critical value of $\Lambda$ (denoted as $\Lambda^*$) does not attain the expected cutoff of $\chi^2_{0.95} = 3.84$ under the independence model until $R > 1000$ for all three levels of sample size considered ($n = 10, 50, 100$). Were we to test the neutral hypothesis using the CLRT and assume $\Lambda$ followed a $\chi^2$-distribution, the test would not attain the correct type I error until the rate of recombination was inordinately large. This result is consistent with that of Bustamante et al. (2001) that the LRT is not robust to deviations from the assumption of independence among sites and highlights the need for developing a statistical method that can deal effectively with linkage among sites.

**How does bias in estimation of the recombination rate affect the realized size of the CLRT?** We see from Figure 4 that the realized type I error of the CLRT decreases with increasing recombination rate for both estimators studied. This is consistent with the fact that both Hey and Wakeley’s (1997) and Hudson’s (1987) estimators improve as $R$ increases (Wall 2000). In general, using $R_0$ to estimate the recombination rate will result in larger type I error than using $R_{bs}$. From our study, for $R = 15$, Hey and Wakeley’s estimator performs better than Hudson’s with size closer to the type I error.
Figure 3.—95% critical value of the test statistic ($A^*$) converges to $X^2_{0.05} = 3.84$ (plotted in log scale for both x- and y-axes). Data were drawn from Hudson’s “ms” program with sample size $n \in \{10, 50, 100\}$ and fixed segregating sites $S = 100$.

(0.05); for $15 \leq R \leq 125$, Hudson’s method actually performs better than Hey and Wakely’s, and for $R \geq 125$, both are overly conservative. Recalling that $R_w$ is upwardly biased for low levels of $R$ (Wall 2000), it becomes clear that overestimating the recombination rate leads to a lower $A^*$ and hence an increased probability of rejecting the null hypothesis (and therefore larger type I error). Consistent with this observation is that $R_{\text{true}}$, which is downwardly biased, leads almost uniformly to a very conservative CLRT.

Figure 4.—Effect of the bias of the recombination rate estimator on the size of the CLRT. Data were drawn from Hudson’s “ms” program with sample size $n = 50$ and fixed segregating sites $S = 100$. Recombination rates were estimated by the “SITES” program (Hey and Wakeley 1997).

Figure 5.—Effect of the population structure on the size of the CLRT. Data were drawn from the island model using Hudson’s “ms” program with given number of demes, $D \in \{2, 5, 10, 20, 50\}$ with $R = 0$.

How does undetected migration affect the size of the CLRT? Even if we had a “perfect” estimator of $R$, we might not attain a realized type I error of $\alpha = 0.05$ due to other factors, such as population history. We see from Figure 5 that an island model of population subdivision is such a scenario. For all levels of $D$ examined, the observed pattern is very similar: the type I error of the CLRT is 0.05 at $M = 0$; it then increases sharply for $0 < M < 1$ and then decreases slowly to 0.05 as $M$ increases toward infinity. In a structured population with $M = 0$, all subpopulations are completely isolated and within each subpopulation, individuals undergo random mating. Since sequences subject to the CLRT are all sampled from one subpopulation assumed to be at equilibrium, it is not surprising to see that the realized size of the CLRT for data with $M = 0$ is at the proper level for all levels of $D$.

Slightly increasing the migration rate will impact the site-frequency spectrum by reducing the relative proportion of low-frequency SNPs and increasing the relative proportion of high-frequency SNPs. This is due to the fact that if $M$ is “small but not too small,” a sample of DNA sequences from a single subpopulation will often contain a single migrant from another deme. This migrant will, more often than not, be involved in the last coalescent event of the genealogy, since the rate of migration is small relative to the rate of coalescent for $M < 1$. This will cause an overrepresentation of gene genealogies that are stretched near the root and compressed near the external nodes. The site frequency spectrum will, thus, look similar to what is expected under positive selection, predicting an increase in the type I error of the CLRT for neutrality.

As $M$ gets larger, the proportion of a given subpopula-
Detecting Selection Using SFS Data

Figure 6.—Effect of the population size changes on the size of the CLRT. Data were drawn from the population exponentially growing model by Hudson’s “ms” program with sample size \( n = 50 \), fixed segregating sites \( S = 100 \), growth rate \( \beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\} \), and various levels of recombination rate.

The larger the population growth rate, the more singletons and hence the more likely it is to make false rejections. It is expected that recombination substantially affects the size of the CLRT, which is shown to be true in Figure 6. For small population growth rate (\( \beta < 0.1 \)), CLRT still performs very well with type I error \( \leq 0.05 \), which means slight changes in the population size do not affect the size of the CLRT.

Williamson et al. (2005) have recently developed a method that can jointly estimate the site-frequency spectrum of population growth data from that under negative selection. For that model, one can also perform the CLRT, conditioning on the maximum-likelihood estimate of the growth parameter from the rest of the genome and an estimate of the local recombination rate to simulate the critical value of the test statistic for a given gene.

How does a recent population bottleneck affect type I error? Simulation study reveals that the effect of population bottlenecks on the patterns of SFS is very complicated (Figure 7). Moderate bottlenecks (Figure 7A) result in less low-frequency SNPs and more medium- and high-frequency SNPs than under neutrality. Strong bottlenecks (Figure 7B) function in the opposite direction, namely, more rare SNPs than expected under the constant population size model. The reason for this is that rate of coalescence increases during the bottleneck period and depending on parameter values can look like either positive or negative selection (Galtier et al. 2000).

For example, a recent weak bottleneck can lead to disproportionately longer internal branches as several lineages make it back into the ancestral population and thus contribute to high-frequency-derived mutations that can look like positive selection. Alternatively, a very severe recent bottleneck will likely lead to the most recent common ancestor event during the bottleneck period and thus to star-like external branches that may be difficult to distinguish from negative selection. As a consequence, the type I error of the CLRT is quite high in populations that have experienced a recent bottleneck event (Figure 8). Increasing sample size and mutation rates leads to even higher type I error (results not shown).

While it is clear that the CLRT is not robust to the effects of a recent bottleneck, it may be possible to distinguish whether the rejection of the test is due to natural selection or the effect of the recent population bottlenecks. One approach is to use a composite-likelihood goodness-of-fit statistic that measures concor-
Figure 8.—Effect of recent population bottleneck on the size of the CLRT. \( f \) is the ratio of population size during bottleneck to the original size. The data sampling scheme is the same as that described in Figure 7.

Figure 7.—Site frequency spectrum of data from a single population having undergone a recent bottleneck. Bottleneck occurred 0.1\( N_e \) generations ago, and it lasted 0.05\( N_e \) generations. Sample size \( n = 10 \), with fixed segregating sites \( S = 100 \). \( f \) is the ratio of population size during bottleneck to the original size. \( \alpha \) is the type I error of the CLRT. (A) Moderate bottleneck with \( f = 0.1 \); (B) strong bottleneck with \( f = 0.01 \).

Figure 9.—Site-frequency spectrum of data from a single subject to weak positive selection is similar to the SFS population having undergone a recent bottleneck. Bottleneck from a neutral population (Figure 10). One biological reason for this phenomenon is that increasing the sample size from \( n = 10 \) to \( n = 50 \) will uniformly increase the power (Figure 9). However, increasing the mutation rate from \( \theta = 30 \) to \( \theta = 75 \), paradoxically, decreases the power for detecting positive selection. The statistical reason for this is that the site-frequency spectrum of data with a high mutation rate and tightly linked sites subject to weak positive selection is similar to the SFS from a neutral population (Figure 10). One biological reason for this phenomenon is that increasing the mutation rate (or reducing the recombination rate) increases interference among selected mutations and thus reduces the overall efficacy of natural selection (Robertson 1961; Hill and Robertson 1966; Felsenstein 1974; Comeron and Kreitman 2002).

Evaluate a statistical test, we want not only to control the type I error, but also to assess the power \( [1 - \text{Pr(type II error)}] \). Our simulation results (Figure 9) suggest that the CLRT has relatively good power to detect negative selection and moderate power to detect positive selection, if the population recombination rate is on the order of the mutation rate and there is moderately strong selection.

If natural selection is very weak (\( |\gamma| < 1 \)) and sites are tightly linked, selection has little effect on the SFS and the CLRT and, thus, has little power. When selection is strong and negative (\( \gamma < -5 \)), the site-frequency spectrum is skewed toward rare alleles and the CLRT performs very well even for small sample size irrespective of the mutation or recombination rates. In detecting weak positive selection (\( \gamma > 5 \)), the CLRT has medium power for moderate levels of recombination relative to mutation (\( R > 5; \theta = 30 \)). We find that increasing the sample size from \( n = 15 \) to \( n = 50 \) will uniformly increase power (Figure 9). However, increasing the mutation rate from \( \theta = 30 \) to \( \theta = 75 \), paradoxically, decreases the power for detecting positive selection. The statistical reason for this is that the site-frequency spectrum of data with a high mutation rate and tightly linked sites subject to weak positive selection is similar to the SFS from a neutral population (Figure 10). One biological reason for this phenomenon is that increasing the mutation rate (or reducing the recombination rate) increases interference among selected mutations and thus reduces the overall efficacy of natural selection (Robertson 1961; Hill and Robertson 1966; Felsenstein 1974; Comeron and Kreitman 2002).

Can the CLRT distinguish negative selection from the effect of population growth? As we see from Figure 11, the CLRT does not have much power in distinguishing very weak negative selection (\( \gamma = -1 \)) from exponential growth. However, for moderately strong negative selection (\( \gamma = -5 \)), the CLRT has very high power to differentiate selection from exponential growth with growth...
Detecting Selection Using SFS Data

Figure 9.—Power of the CLRT under varying levels of selection. The $x$-axis is the value of the selection parameter in the PRF model under which the data were simulated.

Figure 10.—Site-frequency spectrum under recurrent negative selection, neutral, and positive selection with varying levels of mutation and recombination rates. The $y$-axis is the proportion of SNP sites that were found at frequencies $1/15$, $2/15$, $3/15$, $4/15$, $5/15$, $6/15$, $7/15$, $8/15$, $9/15$, $10/15$, $11/15$, $12/15$, $13/15$, and $14/15$.

Figure 11.—Power of the CLRT in distinguishing negative selection from the population exponentially growing model. Data were simulated by the “FISHER” program under the assumption of constant population size with sample size $n = 50$, $\theta = 30$, $R = 100$ under the forward simulation model with selection coefficient $\gamma = -1$, $-5$, $-10$, respectively. The $y$-axis is the growth rate $\beta$, the parameter of the data where the empirical distribution of the test statistics was obtained to get the critical value for the test.

Figure 12.—$\hat{\gamma}/\gamma$ for data drawn from forward simulation with the recombination model (by the “FISHER” program). $\hat{\gamma}$ is the maximum-likelihood estimator of the selection coefficient, and $\gamma$ is the true parameter value under which the data were simulated.

Is the MCLE a good estimator of selection coefficients? If the assumption of independence among sites is met, maximum-likelihood estimation of the selection and mutation rate parameters performs very well (Bustamante et al. 2001). We are interested to know whether the estimator is still reliable when we relax the assumption of independence among sites. In Figure 12, we summarize the ratio of the MCLE of the selection parameter to the true selection coefficient as a function of $\gamma$. We can see that for weak negative selection ($\gamma = -1$), composite-maximum-likelihood estimation performs very well for all parameter combinations considered. Both mutation and recombination affect the accuracy of estimation. The parameter is underestimated with higher mutation rate or less recombination events. In general, the maximum-composite-likelihood estima-
tor does not deviate far away from the true parameter value under which the data were simulated for negative selection with moderate mutation rate and have total recombination events $>100$ per generation.

The MCLE performs rather poorly in estimating the strength of positive selection in the presence of linkage with a large bias toward underestimation. The main reason is likely to be reduction in the effectiveness of selection because of interference among selected mutations (see Comeron and Kreitman 2002). That is, even if each mutation that enters the population has a selective advantage of, say, $N_s = 5$, because there are few chromosomes that lack positively selected mutations, there will be only small fitness difference among chromosomes. As a result, mutations have a smaller realized effect on the site-frequency spectrum than predicted under the independence assumption.

**Conclusions:** The composite-likelihood method presented here for inferring natural selection from DNA sequence data has reasonably good performance, in terms of power and robustness. One advantage over previous PRF tests is proper control of type I error if DNA sites are linked. As expected, the method used to estimate the local recombination rate can have profound effects on the size of the test. We predict this will be a general property of composite-likelihood methods that aim to infer selection from standing patterns of genetic variation, and very little is known about the accuracy of methods for estimating recombination in the presence of recurrent selection.

We also find that undetected population structure, population growth, and/or bottlenecks can all inflate the realized type I error of the test above its nominal level. One possible solution is to explicitly model selection and demography in future incarnations of the CLRT. In particular, by analyzing several unlinked loci simultaneously one may be able to estimate common shared parameters (such as expansion rate or time since bottleneck), while allowing for locus-specific selection parameters. Likewise, it is known that variation in selection among sites as well as dominance can have strong effects on the SFS (Bustamante et al. 2003; Williamson et al. 2004). We hope to incorporate these factors into future versions of the test.

Our simulation study shows that the composite-likelihood-ratio test has excellent power to detect negative selection and moderate power to detect positive selection. However, for weak selection $|\gamma| < 1$ and tight linkage $R < 5$, the method does not perform well, presumably due to interference selection. We have also shown that mutation rate and recombination rate profoundly influence the power of the CLRT.

It should be pointed out that a significant result of the CRLT (as with all tests of neutrality) should be interpreted cautiously since there are several putative alternative hypotheses to a single null hypothesis. Indeed, aside from the factors explored in this article, processes such as population shrinking, inbreeding, and a single selective sweep could also produce genealogies that are consistent with some form of recurrent natural selection. Functional information will ultimately be needed to sort the false from the true positives.

We thank members of the Bustamante lab for discussion. We also thank Feng Feng for his kind assistance with the source code. The comments of two anonymous reviewers greatly improved the article. This research was supported by National Science Foundation award 0319553 to M. Purugganan, S. McCouch, R. Nielsen, and C.D.B. as well as by a U.S. Department of Agriculture-Hatch grant (NY 151411) to C.D.B.

**LITERATURE CITED**


Jensen, J. D., Y. Kim, V. Bauer-DuMonth, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps...
Detecting Selection Using SFS Data

1421


Nagaraja, R., S. Macmillan, J. Kere, C. Jones, S. Griffin et al., 1997 X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res. 7: 210–222.


Nagaraja, R., S. Macmillan, J. Kere, C. Jones, S. Griffin et al., 1997 X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res. 7: 210–222.


Communicating editor: D. Rand