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 NY, 14853
Running head: Detecting Selection Using SFS Data

Keywords: likelihood ratio test, natural selection, site frequency spectrum, composite likelihood

Abbreviations: CLRT, composite likelihood ratio test; FISHER, forward infinite-sites simulation with selection and recombination; GOF, goodness-of-fit test; MCLE, maximum composite likelihood estimator; PRF, poisson random field; SFS, site frequency spectrum.

Corresponding author:
Carlos D. Bustamante
Department of Biological Statistics and Computational Biology
101 Biotechnology Building, Cornell University
Ithaca, New York 14853
Phone (607)255-1640
Fax (607)255-4698
Email CDB28@CORNELL.EDU.
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, Moriyama and Sawyer (1994) for inference in a Wright-Fisher genic selection model and corrects for non-independence among sites by application of coalescent simulations with recombination. Here, we (1) characterize the distribution of the CLRT statistic \( \Lambda \) as a function of the population recombination rate \( R = 4Nr \), where \( r \) is the population recombination rate per locus per generation; (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 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. Lastly, we demonstrate that the MCLE of the selection parameter has little bias for weak negative selection, and has downward bias for positively selected mutations.
INTRODUCTION

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, Moriyama, and Sawyer 1994; Wakeley 2003; Williamson 2003, Williamson et al, in press.) 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 an 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 oppose to traditional summary statistics of the data such as Tajima’s D (1989), Fay and Wu (2000), and Fu and Li (1993).

Unfortunately, since the PRF model assumes independence among sites, the application of the LRT for 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 in order to evaluate the true likelihood function via the ancestral selection graph (Krone and Neuhauser 1997; Neuhauser and Krone 1997; Slade 2001), (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; Mcvean et al. 2002; Fearnehad and Donnelly 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. (2001)’s result that the LRT proposed by Hartl, Moriyama, and Sawyer (1994) is not robust to deviations from the
assumption of independent 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, Moriyama, and Sawyer’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 $\hat{R}$, an estimate of the population recombination rate, and choosing a cutoff such that $\alpha$ of the simulations are below the critical value. 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 then expected (namely $100^*\alpha\%$ 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 a 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 will 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

In order 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, Moriyama, and Sawyer, 1994): (a) panmictic population of constant size, (b) weak selection with no dominance, (c) equal selective effects of all non-lethal 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,
Fledel-Alon, and Bustamante (2004) have developed PRF models with dominance relaxing assumption (b), Bustamante, Nielsen, and Hartl (2003) and Sawyer, Kulathinal, Bustamante and Hartl (2003) have modeled the effects of a distribution of selective effects among non-lethal mutations (relaxing assumption (c)). The purpose of this article is relaxing assumption (d) for the purpose of inference.

Let \( X = \{X_1, X_2, \ldots, X_{n-1}\} \) represent the site-frequency spectrum for a genomic region of interest such that \( X_k \) 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 paper, Sawyer and Hartl (1992) model 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 \( \gamma = 2Ns \) where \( N \) is the haploid population size; \( s \) is the fitness effect of new mutations such that wildtype fitness is 1, heterozygote fitness is \( 1 + s \), and homozygote fitness for the new mutation is \( 1 + 2s \); and \( \mu \) is the per locus mutation rate for mutations with selective effect \( s \). It is important to note that site frequency spectrum of selected sites is sensitive to assumptions regarding dominance.

The PRF model proposed by Williamson, Fledel-Alon, and Bustamante (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 only consider the genic selection case here. When \( \gamma = 0 \), the population is evolving neutrally; when \( \gamma > 0 \), under positive selection; and when \( \gamma < 0 \), subject to negative selection. According to the results of Sawyer and Hartl (1992), \( X_k \) is a Poisson distributed random variable with mean \( \theta F(k, \gamma) \) where
\[ F(k, \gamma) = \int_0^n \left( \begin{array}{c} n \\ k \end{array} \right) p^k (1 - p)^{n-k} \frac{1 - e^{-2\gamma (1-p)}}{(1 - e^{-2\gamma}) p (1 - p)} dp \]

Since under the model the \( X_k \)'s are independent, the likelihood function \( L(\theta, \gamma \mid X) \) is the product over \( P(X_k = x_k \mid \theta, \gamma) \), namely:

\[
L(\theta, \gamma \mid X) = \prod_{k=1}^{n-1} \frac{e^{-\theta F(k, \gamma)} (\theta F(k, \gamma))^{x_k}}{x_k!}
\] (1)

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

\[
\Lambda = (\frac{-2 \log L(\hat{\theta}_w, 0 \mid X)}) - (\frac{-2 \log L(\hat{\theta}, \hat{\gamma} \mid X)})
\]

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

where \( \hat{\theta}_w \) 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 \( S = \sum_{i=1}^{n-1} x_i \) is the observed number of segregating sites. Under the assumption of independence among sites, \( \Lambda \) 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 a composite likelihood ratio rest (CLRT). Under such a scenario the distribution of the test statistic is no longer $\chi^2_1$, but rather depends on the rate of recombination among sites. We must, therefore, use coalescent simulations with recombination to find the critical value $\Lambda^*$ 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 spectra, $X_{OBS}$, estimate $\hat{\theta}$ and $\hat{\gamma}$ using the one-dimensional optimization described in Bustamante et. al (2001), and calculate the CLRT statistic $\Lambda_{OBS}$ via equation (2).

2. Generate $Q$ replicate data sets $X_1, X_2, \ldots, X_Q$ from a standard neutral model with recombination rate $R$ corresponding to the region of interest and $S$, 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 $\Lambda_1, \Lambda_2, \ldots, \Lambda_Q$. 


3. The $P$-value for the CLRT corresponding to data $X_{OBS}$ is estimated as

$$P(\Lambda_{OBS} \mid H_0) \approx \frac{\sum_{i=1}^{Q} I(\Lambda_{OBS} \leq \Lambda_i)}{Q},$$

where $I()$ is the indicator function which evaluates to 1 if the argument is TRUE and 0 otherwise.

In order to implement our proposed composite-likelihood method, one needs an estimate of the local recombination rate. Given a data set, there are two main approaches that have been employed for estimating $R$. One method is based on observing the frequency of sequence exchange between markers (Ashburner 1989; Bouffard et al. 1997; Nagaraja et al. 1997; True et al. 1996), the other method estimates $R$ from the patterns of sequence variation expected in a random sample of DNA sequences from a population (Griffiths and Marjoram 1996; Hey and Wakeley 1997; Hudson 1987; Hudson and Kaplan 1985; Kuhner et al. 1999; Wakeley 1997; Hudson 2001; 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 meiosis. 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 estimator ($R_{hw}$) 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 $R_h$ over-estimates $R$ (a large proportion of $R_h / R$ greater than 5.0);
while $R_{hw}$ under-estimates $R$ (with majority of $R_{hw} / R$ less than 0.2). In this article, we explore these two estimators since they represent the extreme effects of over- and under-estimating 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 “ms” program (2002) 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 $\chi^2$ 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 program FISHER (Forward Infinite-sites Simulation with $s$Election and $r$ecombination) written in ANSI C by Lan 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 ms program (2002). For each of three sample sizes ($n = 10,50,100$), we simulated 1,000 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 Figure 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-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 a neutral data $X_{OBS}$ with known recombination rate $R$ and apply the CLRT to obtain the test statistic $\Lambda_{OBS}$.

2) For $X_{OBS}$ estimate $R$ by Hudson’s and Hey-Wakeley’s methods using SITES (Hey and Wakeley 1997) and denote the estimates $R_h$ and $R_{hw}$.
3) Generate \( Q = 1000 \) replicate data sets with the same sample size and number of segregating sites as \( X_{OBS} \) under estimated recombination rate, \( R_h \). For each replicate, perform CLRT and keep test statistic \( \Lambda \). The empirical \((1 - \alpha)\) quantile of the distribution of \( \Lambda \) among the 1,000 replicates is the critical value of the test statistic \( \Lambda^* \) at \( \alpha \) level (for all simulations we used \( \alpha = 0.05 \)). Similarly, we can find \( \Lambda^* \) with estimated recombination rate \( R_{hw} \).

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

5) Repeat steps (1-4) 1,000 times. The proportion of the false rejection is the realized size of the CLRT under 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 \( 0 \leq M \leq 15 \) (\( M = 4N_e m \), 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 critical value of CLRT at $\alpha = 0.05$ level as the 9,501 largest value and denote this quantity as $\Lambda^*$.  
2) Sample $n$ sequences with $R = 0$ from a single deme out of $D$ possible demes in the island model with migration for a given level of $M$. Apply the CLRT and retain the observed test statistic, $\Lambda_{OBS}$.  
3) If $\Lambda_{OBS} > \Lambda^*$, reject the neutral hypothesis; otherwise, fail to reject at $\alpha = 0.05$ level.  
4) Repeat steps (2-3) 1,000 times for each parameter combination. The proportion of data sets that reject neutrality (i.e., number of data sets out of 1,000 with $\Lambda_{OBS} > \Lambda^*$ ) 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: $N(t) = N_e \exp(-\beta t)$, where $N_e$ 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 $t_{bs} = 0.025$ or 0.05 (in the unit of $4N_e$ generations) before the current sampling time and recovered to the current population size at $t_{be} = 0.0125$ (in the unit
of $4N_e$ generations). We consider two levels of the population reduction during the bottleneck, i.e. $f \in \{0.1, 0.01\}$ (Figure 7, 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, 10, and 11.

**Algorithm 4: Procedure for Estimating Power of CLRT**

1) Generate a data set of $n$ sequences via FISHER given mutation rate, selection coefficient and recombination rate; apply CLRT to obtain the test statistic $\Lambda_{OBS}$ and corresponding P-value from Algorithm 1.

2) If $P < 0.05$, reject the neutral hypothesis; otherwise, fail to reject at $\alpha = 0.05$ level.

3) Repeat steps (1-2) 500 times and calculate 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 similar effect as negative selection, we would like to 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 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 with 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**

1. **How quickly does the test statistic ($\Lambda$) converge to a $\chi^2_1$ distribution?**

   From Figure 2, we confirm the theoretical prediction that the composite likelihood ratio test statistic $\Lambda$ converges to a $\chi^2_1$ 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 $\chi^2_{1,0.95} = 3.84$ 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 $\chi^2_1$ distribution, the test would not attain the correct Type I error until the rate of recombination was inordinately large. This result is consistent with Bustamante et al. (2001) that the LRT is not robust to deviations from the assumption of independent among site, and highlights the need for developing a statistical method that can deal effectively with linkage among sites.

2. **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 (1997) and Hudson’s (1987) estimator improve as \( R \) increases (Wall 2000). In general, using \( R_h \) to estimate the recombination rate will result in larger Type I error than using \( R_{hw} \). From our study, for \( R \leq 15 \), Hey-Wakeley’s estimator performs better than Hudson’s with size closer to the Type I error (0.05); for \( 15 \leq R \leq 125 \), Hudson’s method actually performs better than Hey and Wakely, and for \( R \geq 125 \), both are overly conservative. Recalling that \( R_h \) is upwardly biased for low levels of \( R \) (Wall 2000), it becomes clear that overestimating the recombination rate leads to a lower \( \Lambda^* \) and hence an increased probability of rejecting the null hypothesis (and therefore larger Type I error). Consistent with this observation is that \( R_{hw} \), which is downwardly biased, leads almost uniformly to a very conservative CLRT.

3. 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 observe 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 towards 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 subpopulation that originated in another deme increases linearly. And as \( M \) tends towards infinity, the fixation index \( F \) will tends to be zero \( (\hat{F} = \frac{1}{1 + 4N_e m}, \text{ at equilibrium}) \), indicating no population structure. Hence a sample of DNA sequences randomly drawn from a subpopulation would be wholly representative of the entire population and the CLRT should have Type I error at the desired level. Indeed, from our simulation study, when \( M \geq 16 \), for number of subpopulations less than 10, all tests we studied have type I error \( \leq 0.05 \). For large number of subpopulations (>10), \( M \) should be greater than 32 in order to have proper size of the CLRT.

There are two possible ways to improve the CLRT vis a vis population structure. One is to modify the critical value of the CLRT by estimating \( M \) from neutral data and thus reducing the Type I error by producing a more sophisticated null model. The second approach is to jointly estimate selection and migration coefficients under various
population structure models. It is important to note that both fixes might also introduce systematic bias in the realized Type I error due to bias in estimation of demographic parameters.

4a. How does recent population expansion affect the type I error?

Another important assumption in the current model is the assumption of constant population size over generations. This assumption does not hold for the vast majority of species which we would like to analyze for evidence of natural selection at the genetic level. From Figure 6, we can say that CLRT is not robust against the assumption of constant population size though it does not do badly for relative tight linkage with low population growth rate. The Type I error increases with the population growth rate. The reason is that population growth causes an increase in the coalescent rate as the process proceed back in time, leading to star-like genealogies which results in an excess of mutations in external branches (i.e., singletons or substitutions present in only one sampled sequence) (Slatkin and Hudson 1991; Tajima 1989). It is difficult to differentiate the site frequency spectrum of population growth data from that under negative selection. The larger the population growth rate, the more singletons and hence more likely 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 less than or equal to 0.05 which means slight changes in the population size do not affect the size of the CLRT. Williamson et al. (in press) have recently developed a method that can jointly estimate selection and population growth assuming independence among sites. 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.

4b. 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 which 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 which may be difficult to distinguish from negative selection. As a consequence, the Type I error of the CLRT is quite high in populations which 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 which measures concordance between the data and a
selective model (Jensen et al, in submission). Alternatively, the genomic distribution of CLRT statistic itself can be used, since a bottleneck would uniformly increase the proportion of loci across the genome that rejects neutrality.

5. How powerful is the CLRT in detecting selection?

To evaluate a statistical test, we not only want to control the Type I error, but also would like to assess the power \(1 - Pr(\text{Type II error})\). Our simulation results (Figure 9) suggest that 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, thus, has little power. When selection is strong and negative \(\gamma<-5\), the site-frequency spectrum is skewed towards 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 sample size from \(n=15\) to \(n=50\) will uniformly increase power (Figure 9). However, increasing 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 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 mutation, and thus reduces
the overall efficacy of natural selection (Robertson 1961; Felsenstein 1974; Hill and Robertson 1966; Comeron and Kreitman 2002).

6. 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 rate in the range of 0.1 ~ 3.2. This suggests the CLRT maybe particularly useful for finding genes that may be subject to moderate negative selection.

7. Is MCLE (Maximum Composite Likelihood Estimator) 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 ($\gamma \approx -1$), composite maximum likelihood estimation performs very well for all parameter combinations considered. Both mutation and recombination affect the accuracy of estimation. The parameter is under-estimated with higher mutation rate or less recombination events. In general, maximum composite likelihood estimator 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 greater than 100 per generation.

MCLE performs rather poorly in estimating the strength of positive selection in the presence of linkage with a large bias towards 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 \( Ns = 5 \), 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 realized size of the test. We predict this will be a general property of CL 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 in 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 test of neutrality) should be interpreted cautiously since there are several putative alternative hypotheses to single null hypothesis. Indeed, aside from the factors explored in this paper, 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 true positives.
LITERATURE CITED


Bouffard, G. G., J. R. Idol, V. V. Braden and et al., 1997  A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79kb. Genome Res. 7: 673-692.


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


Table 1. Table of notations used in the paper

<table>
<thead>
<tr>
<th>Notation</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>number of demes</td>
</tr>
<tr>
<td>$\beta$</td>
<td>population growth rate</td>
</tr>
<tr>
<td>$n$</td>
<td>number of sequences sampled</td>
</tr>
<tr>
<td>$S$</td>
<td>total number of segregating sites in the sampled sequences</td>
</tr>
<tr>
<td>$Q$</td>
<td>number of replicates in the simulation study</td>
</tr>
<tr>
<td>$R_h$</td>
<td>recombination rate estimator from Hudson (1987)</td>
</tr>
<tr>
<td>$R_{hw}$</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>$t_{bs}$</td>
<td>time in the unit of $4N_e$ generations ago that bottleneck happens</td>
</tr>
<tr>
<td>$t_{be}$</td>
<td>time in the unit of $4N_e$ generations ago that population recovers from bottleneck</td>
</tr>
<tr>
<td>$f$</td>
<td>ratio of the population size during bottleneck to the original.</td>
</tr>
</tbody>
</table>
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 1,000 coalescent simulations using ms (Hudson, 2002). “Selection” is the expected SFS under genic selection for the model described by Hartl, Moriyama, and Sawyer (1994). We use a value of $2Ns=1.353$, 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.
Figure 2. Distribution of the test statistics ($\Lambda$) for the test assuming Hartl, Moryama and Sawyer (1994) model as a function of population recombination rate ($R$). Y-axis is quantiles of $\Lambda$’s calculated by CLRT from sampled sequences, X-axis is quantiles of data drawn from $\chi^2_1$ distribution. $\Lambda$ converges to $\chi^2_1$ distribution with large $R$. 1000 replicates of data sets were sampled from Hudson’s “ms” program, each with sample size $n = 50$, fixed number of segregating sites $S = 100$ and various level of recombination rate.
Figure 3. 95% critical value of the test statistic ($\Lambda^*$) converges to $\chi^2_{1,0.95} = 3.84$ (plotted in log scale for both x- and y-axis). Data were drawn from Hudson’s “ms” program with sample size $n \in \{10, 50, 100\}$ and fixed segregating sites $S = 100$. 
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 $n = 50$, fixed segregating sites $S = 100$. Recombination rates were estimated by “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 $R = 0$. 
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 level of recombination rate.
Figure 7. Site frequency spectrum of data from a single population having undergone a recent bottleneck. Bottleneck occurred $0.1N_e$ generations ago, and it lasted $0.05N_e$ generations. Sample size $n = 10$, and 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 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. Data sampling scheme is the same as that described in figure 7-A and 7-B.
Figure 9. Power of the CLRT under varying levels of selection. 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,…, 14/15.
Figure 11. Power of the CLRT in distinguishing negative selection from the population exponentially growing model. Data were simulated by “FISHER” program under the assumption of constant population size with sample size n=50, $\theta=30$, R=100 under forward simulation model with selection coefficient $\gamma = -1, -5, -10$, respectively. X-axis is the growth rate $\beta$, the parameter of the data where the empirical distribution of the test statistics was obtained in order to get the critical value for the test.
Figure 12. $\hat{\gamma}/\gamma$ for data drawn from forward simulation with recombination model (by “FISHER” program). $\hat{\gamma}$ is the maximum likelihood estimator of the selection coefficient, $\gamma$ is the true parameter value under which the data were simulated.