DOI: 10.1534/genetics.106.069450

# On the Utility of Linkage Disequilibrium as a Statistic for Identifying Targets of Positive Selection in Nonequilibrium Populations

**Jeffrey D. Jensen,*,[1] Kevin R. Thornton,*,[2] Carlos D. Bustamante[†] and Charles F. Aquadro***

**Department of Molecular Biology and Genetics and [†]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853*

Manuscript received December 12, 2006
Accepted for publication May 19, 2007

ABSTRACT

A critically important challenge in empirical population genetics is distinguishing neutral nonequilibrium processes from selective forces that produce similar patterns of variation. We here examine the extent to which linkage disequilibrium (*i.e.*, nonrandom associations between markers) improves this discrimination. We show that patterns of linkage disequilibrium recently proposed to be unique to hitchhiking models are replicated under nonequilibrium neutral models. We also demonstrate that jointly considering spatial patterns of association among variants alongside the site-frequency spectrum is nonetheless of value. Through a comparison of models of equilibrium neutrality, nonequilibrium neutrality, equilibrium hitchhiking, nonequilibrium hitchhiking, and recurrent hitchhiking, we evaluate a linkage disequilibrium (LD) statistic ($\omega_{max}$) that appears to have power to identify regions recently shaped by positive selection. Most notably, for demographic parameters relevant to non-African populations of *Drosophila melanogaster*, we demonstrate that selected loci are distinguishable from neutral loci using this statistic.

PATTERNS of DNA sequence polymorphism are shaped by both a population's demographic history and natural selection. Uncoupling these two processes is of outstanding importance, as this differentiation will enable evolutionary biologists to quantify the relative importance of adaptive and nonadaptive factors in shaping levels of variation in natural populations. A number of methods have been proposed in recent years to distinguish demography from selection. For the most part, research has focused on identifying patterns of DNA sequence variation that are "unique" to selective sweeps (Maynard-Smith and Haigh 1974; Hudson *et al.* 1987; Kaplan *et al.* 1989; Tajima 1989; Stephan *et al.* 1992; Braverman *et al.* 1995; Fu 1997; Fay and Wu 2000; Przeworski 2002; Kim and Nielsen 2004; Stephan *et al.* 2006). The most commonly used statistics rely on the site-frequency spectrum (SFS) or observed frequencies of DNA polymorphism in the data. Predictions are often tested in the form of likelihood-ratio tests, comparing a selective sweep model of SFS variation against (a) a neutral equilibrium model (*e.g.*, Kim and Stephan 2002; Kim and Nielsen 2004), (b) a neutral nonequilibrium model (*e.g.*, Wright *et al.* 2005), (c) a nonneutral nonequilibrium model (*e.g.*, Teshima *et al.* 2006; Thornton and Jensen 2007), or, with the advent of large-scale genomic data, (d) the background site-frequency spectrum (*e.g.*, Nielsen *et al.* 2005). Related approaches have relied on specifically testing the goodness-of-fit of a given data set to the predictions of a selection model (*e.g.*, Jensen *et al.* 2005). The results of this endeavor have been mixed, with many putatively unique patterns being reproduced by demographic scenarios.

Since SNP data contain information about linkage disequilibrium (LD) in addition to site frequencies, it has been hypothesized that this additional information could be utilized for hypothesis testing and allow for greater discriminatory power. Specifically, a number of theoretical and simulation results have demonstrated that LD is an important signature of a selective sweep (*e.g.*, Parsch *et al.* 2001; Przeworski 2002; Sabeti *et al.* 2002; Wootton *et al.* 2002; Kim and Nielsen 2004; Eberle *et al.* 2006; Stephan *et al.* 2006). Therefore, it is reasonable to think that vital information is being ignored by not considering associations between markers. The extent to which incorporating LD may improve our ability to distinguish selection from demography has been largely unexplored.

Kim and Nielsen (2004) examined the effects of including LD into the Kim and Stephan (2002) likelihood framework. They describe three patterns of LD predicted from a genealogical model that are proposed to be potentially unique to a selective sweep. First, a high level of LD is expected in regions near, but not immediately adjacent, to the target of selection. Second, a high level of LD is expected on both sides of the target,

[1]*Corresponding author:* Section of Ecology, Behavior and Evolution, AP&M 4th Floor Annex, University of California, La Jolla, California 92037. E-mail: jjensen@ucsd.edu

[2]*Present address:* Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697.

but should not span the site of selection. Finally, there is a strong correlation between high-frequency-derived alleles (as measured by Fay and Wu's *H*-statistic) and LD, such that the probability of observing these alleles is greater in regions of strong LD. They thus proposed a new composite-likelihood method designed to incorporate this information. They note, however, that the improvement made by including LD is small, suggesting that most relevant information is efficiently captured by considering only the site-frequency spectrum, owing to the correlation between LD and high-frequency-derived alleles. Importantly, their result pertains specifically to the case of distinguishing between a selective and a neutral equilibrium model.

Stephan *et al.* (2006) analytically studied a three-locus model of genetic hitchhiking in which one locus is under positive selection while the other two are neutral and partially linked. While they further support a number of the conclusions described in Kim and Nielsen (2004) and further generalize their results, they also note that when the direction of LD is polarized with respect to the more common allele at each neutral site, more positive than negative LD is created after a selective sweep. They propose that this pattern may indeed be unique to a selection model, and thus hitchhiking may have a distinctively patterned LD-*reducing* effect near the target of selection. Encouraged by this result, we undertook a simulation study to explore if there were patterns of linkage disequilibrium that are indeed unique to models of positive selection relative to nonequilibrium models, which may aid in the discovery of adaptively important loci.

## METHODS

**Modeling neutrality:** For all neutral simulations we used Hudson's (2002) *ms* program. Specifically, we simulated data under bottleneck scenarios of varying intensity as well as under an island model of population subdivision. We simulated a region of 10-kb-long sequences with a scaled mutation rate of $\theta = 75$ and $4Nr = 100$, where $r$ is the probability per generation of crossing over for the entire simulated region, values roughly corresponding to a typical *Drosophila melanogaster* data set. The bottleneck model has five parameters: the population mutation rate ($\theta = 4N_0\mu$, where $N_0$ is the effective size of the ancestral population), the population recombination rate ($\rho = 4N_0r$), the time at which the derived population recovered from the bottleneck ($t_r$), the duration of the bottleneck ($d$), and the severity of the bottleneck ($f, 0 < f \leq 1$). In the figures, the time of the bottleneck ($t_b$) is often referred to—where $t_b = t_r + d$.

Simulations of population subdivision under an island model are performed with two subpopulations and scaled migration rate, $M = 4Nm$, where $m$ is the fraction of migrants in each subpopulation in each generation. The sampling scheme is denoted by $\mathbf{n} = \{n_1, n_2\}$, where $n_1$ and $n_2$ refer to the numbers of chromosomes sampled from the first and second subpopulations, respectively. In this study, we examine equal and unequal sampling from the subpopulations, for $M =$ 0.1, 1, 4, and 10. To distinguish from bottlenecks and subdivisions, we refer to the model of neutral evolution under random mating and constant size as the equilibrium neutral model.

**Modeling selective sweeps:** We model positive selection using coalescent simulations for a region of $M$ nucleotides. At time $\tau$ in the past (measured in units of $4N$ generations), a beneficial allele has fixed in the population at position $X$. For all single-sweep simulations, $X$ lies in the interval $[1, M]$. The simulation consists of a neutral phase, which is the standard coalescent with recombination (Hudson 1983), and a selective phase (Braverman *et al.* 1995). At time $\tau$ in the past, the simulation enters the selective phase, which is modeled as a structured coalescent process (*e.g.*, Kaplan *et al.* 1988; Braverman *et al.* 1995), and time is incremented in small units, $\delta t$, until the frequency of the beneficial allele first reaches $x(t) < \xi$, at which point the simulation continues in a neutral phase until the most recent common ancestor of the sample is reached. Full details of the single-sweep simulations are found in Thornton and Jensen (2007).

We also considered a model of selective sweeps occurring in the genome at a rate determined by $\lambda$, the expected number of sweeps per recombination unit in the last $4N$ generations (Kaplan *et al.* 1989; Braverman *et al.* 1995). Here we allow for selective sweeps both within the region of $M$ nucleotides as well as at linked sites. We do this because we simulate a relatively large neutral region ($M = 10^4$), and the probability of a sweep within that region may not be negligible for large $\lambda$, assuming a constant $\lambda$ across the genome. In this model, the time until the next selective phase is entered is exponentially distributed with rate $8Ns\lambda/\rho_{bp} + M\lambda$, where $\rho_{bp}$ is the scaled recombination rate between adjacent base pairs. The first half of this rate accounts for sweeps flanking the sequenced region, and the $M\lambda$ accounts for sweeps within the region. Given that a selective phase is entered, the selected site is located within the $M$ nucleotides with probability $M\lambda/(8Ns\lambda/\rho_{bp} + M\lambda)$; otherwise it is located at a linked site up to a maximum genetic distance of $2\alpha$ (where $\alpha = 2Ns$) on either side of the sampled region (see Kaplan *et al.* 1989 and Durrett and Schweinsberg 2004 for details).

Briefly, the expected time between successive hitchhiking events is $E[t_L]$, the expected length of a hitchhiking event, plus $E[t_S]$, the expected time until the next fixation of a selected allele. For the model considered here, this equals $-(\log\xi/\alpha) + 1/(8Ns\Lambda/\rho_{bp} + M\Lambda)$ in units of $4N$ generations. For example, for the case of $\alpha =$ 5000, $\lambda = 10^{-5}$, $\rho = 10$, sweeps are occurring on average every $\approx 0.008$ time units for the ~30-kb region ($2\alpha + 10^4 +$

$2\alpha = 30$ kb). This extrapolates to approximately one sweep per 80 generations somewhere in the 120-Mb euchromatic portion of the *D. melanogaster* genome.

We estimated LD for two sample sizes ($n = 12$ and 50) and 90 parameter combinations generated by considering all combinations of $\theta \in \{10, 75\}$, $\rho \in \{10, 50, 100\}$, $\alpha \in \{100, 500, 1000, 2500, 5000\}$, and $\lambda \in \{10^{-7}, 10^{-6}, 10^{-5}\}$. These parameters cover cases where we expect hitchhiking effects to be minimal ($\lambda = 10^{-7}$, $\alpha = 100$) to those where the effect should be substantial ($\lambda = 10^{-5}$, $\alpha = 5000$). For these simulations, we used $N = 10^6$.

**Statistics:** We evaluate the likelihood-ratio test (comparing a neutral equilibrium model and a single-sweep equilibrium model) proposed by Kim and Nielsen (2004) under all simulated scenarios. We also examine the LD statistic that they proposed to more specifically quantify the extent to which "sweep-like" patterns of LD are being generated under alternative models. This statistic, termed $\omega$, defined as

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1}\left(\sum_{i,j\in L} r_{ij}^2 + \sum_{i,j\in R} r_{ij}^2\right)}{(1/l(S-l))\sum_{i\in L, j\in R} r_{ij}^2},$$

divides the $S$ polymorphic sites in the data set into two groups, one from the first to the $l$th polymorphic site from the left and the other from the $(l+1)$th to the last site ($l = 2, \ldots, S-2$), where $L$ and $R$ represent the left and the right set of polymorphic sites, and $r_{ij}^2$ is the squared correlation coefficient between the $i$th and $j$th sites. Thus, $\omega$ increases with increasing LD within each group and decreasing LD between groups (*i.e.*, the larger the value of the statistic the more sweep-like the underlying pattern). For a data set, the value of $l$ that maximizes $\omega$ ($\omega_{max}$) is found. Singletons were excluded prior to calculation. Because the statistic is two tailed, rejections may be the result of values of $\omega_{max}$ that are either too large or too small relative to the null.

## RESULTS

**Distinguishing single selective sweep models from nonequilibrium neutral models:** As a starting point, the Kim and Nielsen (2004) likelihood-ratio test was used to analyze both neutral nonequilibrium and nonneutral equilibrium data sets. Parameters for these models were chosen both for their relevance to natural populations (particularly for *D. melanogaster*) and to overlap with the space investigated in Jensen *et al.* (2005). When applied to selection data sets, and consistent with Kim and Nielsen (2004), we observe that the probability of rejecting neutrality in favor of selection increases as $\alpha$ ($= 2Ns$) increases (Figure 1A). An island model with two subdivided populations was also evaluated, and we considered a sampling scheme in which all alleles are sampled from one subpopulation, as well as one in which the subpopulations are sampled equally. We find

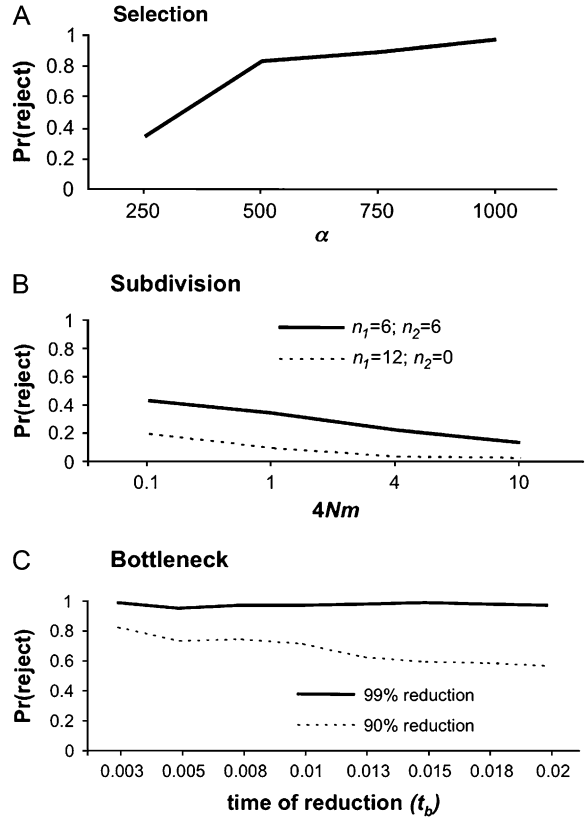


Figure 1.—Results of the Kim and Nielsen (2004) approach incorporating LD when applied to data sets generated under various models. Results are based on 1000 simulations for each data point using the following parameter estimates; $4Nr = 100$, $4N\mu = 75$, $n = 12$, and the length of the region is 10 kb. (A) The time since the sweep is 0.001 in units of $4N$ generations, and $\alpha = 2Ns$. On the *y*-axis is the probability of rejecting neutrality in favor of selection (in this case the proportion of true positives), and on the *x*-axis is the value of the selection coefficient $\alpha = 2Ns$. (B) Values are shown for two scenarios, one in which all alleles are sampled from one subpopulation, the other in which alleles are equally sampled. On the *y*-axis is the probability of rejecting neutrality in favor of selection (false positives), and on the *x*-axis is the value of the migration parameter ($4Nm$). (C) The time at which the population recovers back to its prebottleneck size is 0.0001 in units of $4N$ generations. Values are shown for two scenarios, one in which the population is reduced by 90% during the bottleneck and the other in which it is reduced by 99%. On the *y*-axis is the probability of rejecting neutrality in favor of selection (false positives), and on the *x*-axis is the time of the population crash ($t_b = t_r + d$) in units of $4N$ generations relative to the present.

that the test has a large false positive rate (FPR) (or type I error) under both scenarios when migration is rare, and the FPR gradually decreases as $4Nm$ increases owing to the deterioration of population structure (Figure 1B).

Finally, we examined neutral stepwise bottleneck models in which the time of the population crash ranged from $t_b = 0.0025$ to 0.02 in units of $4N$ generations [where $t_b$ = time of recovery ($t_r$) + duration ($d$); Figure 1C]. False positive rates are observed near 100% for severe bottlenecks (a 99% reduction in population size),

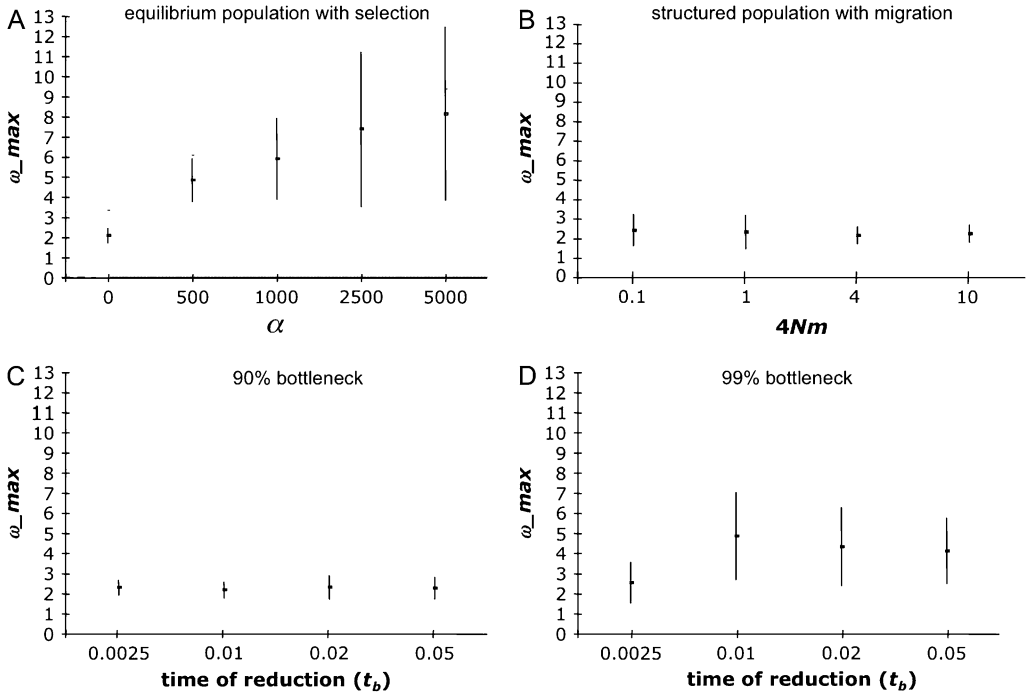


FIGURE 2.—The averages and standard deviations of $\omega_{max}$ under neutral equilibrium, neutral nonequilibrium, and nonneutral equilibrium scenarios. Results are based on 1000 simulations for each data point using the following parameter estimates: $4Nr = 100$, $4N\mu = 75$, the sample size ($n$) = 50, and the length of the region is 10 kb. (A) A sweep occurred at time 0.002 $4N$ generations ago, of intensity 500, 1000, 2500, or 5000 in units of $2Ns$ (where $\alpha = 0$ is the equilibrium neutral model). (B) Two subpopulations are sampled evenly ($n_1 = 25$; $n_2 = 25$) with rates of symmetric migration of 0.1, 1, 4, or 10 in units of $4Nm$. (C) The population size is reduced by 90% at times $t_b = 0.0025$, 0.01, 0.02, or 0.05 $4N$ generations ago and recovered to the prebottleneck size at time 0.0011 $4N$ generations ago. (D) The population size is reduced by 99% at times $t_b = 0.0025$, 0.01, 0.02, or 0.05 $4N$ generations ago and recovered to the prebottleneck size at time 0.0011 $4N$ generations ago. On the $y$-axis is $\omega_{max}$.

with slightly lower FPRs for less severe bottlenecks (a 90% reduction in population size). The performance of this test is thus similar to the CLRT of KIM and STEPHAN (2002; JENSEN *et al.* 2005). The parameters examined are intended to span the non-African bottleneck estimates recently proposed for *D. melanogaster* (THORNTON and ANDOLFATTO 2006), although they are relevant for other recently bottlenecked populations (*e.g.*, humans).

Figure 2 summarizes the averages and standard deviations of $\omega_{max}$ under the equilibrium neutral, nonequilibrium neutral, and equilibrium selection models examined for $n = 50$ ($n = 12$ not shown). There are a number of notable features. First, under the equilibrium neutral model ($\alpha = 0$), $\omega_{max}$-values were observed between 2 and 3 for common ($n = 12$) and large ($n = 50$) sample sizes, with small standard deviations ($n = 50$ shown in Figure 2A). A number of equilibrium selection models produced distinctive distributions of $\omega_{max}$. For large $n$, $\omega_{max}$ is greatest when the selective event was recent and strong. In contrast, for small sample sizes, individual observed values of $\omega_{max}$ may be *reduced* relative to the null for large selection coefficients—owing to the fact that there is very little variation within the 10-kb region following such a severe sweep, an effect that is exacerbated in small sample sizes.

Second, no model of population structure was identified that regularly produced the pattern of two distinctive stretches of strong LD within, but low LD between, and the distributions are largely indistinguishable from the neutral equilibrium model (Figure 2B). Third, modest bottleneck models (90% reduction) returned values of $\omega_{max}$ near that observed under neutral equilibrium conditions even for large sample sizes (Figure 2C), while severe bottlenecks (99% reduction) result in a distribution with large values in the tail. For example, a 99% reduction at time $t_b = 0.01$ $4N$ generations in the past, with a recovery $t_r = 0.0011$ $4N$ generations ago, has an average $\omega_{max}$ near 5 (Figure 2D).

**Distinguishing nonequilibrium selection models from nonequilibrium neutral models:** The THORNTON and ANDOLFATTO (2006) bottleneck model estimated for *D. melanogaster* was singled out for specific analysis. Results are also presented for sweeps in an equilibrium population for comparison (Figure 3, a and b). The distribution of the $\omega_{max}$-statistic is largely overlapping between the neutral and nonneutral scenarios for the bottlenecked population, particularly for $n = 12$ (Figure 3c). Thus, these results, taken with those from Figure 2, strongly suggest that the patterns of linkage disequilibrium proposed to be unique to positive selection are being replicated under realistic demographic models. Moreover, these results highlight the relative difficulty of inferring selection in nonequilibrium *vs.* equilibrium populations. The selection distributions of $\omega_{max}$ observed under nonequilibrium models are considerably less distinctive than those under equilibrium models (*e.g.*, Figure 3c *vs.* Figure 3a and Figure 3d *vs.* Figure 3b).

Nonetheless, comparing the neutral and nonneutral bottleneck models, it is noteworthy that for large sample sizes ($n = 50$; Figure 3d), the distributions of $\omega_{max}$ are

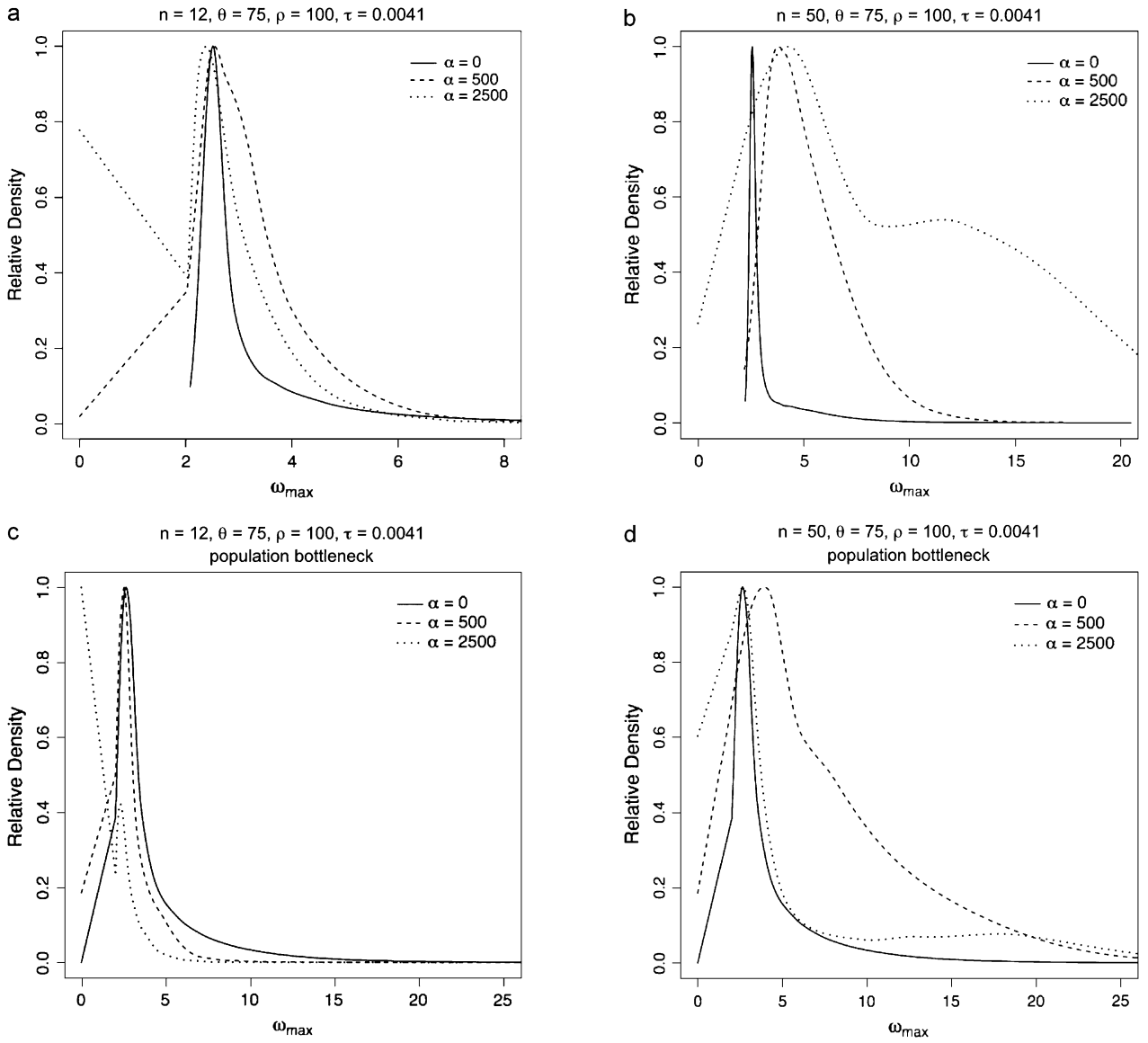


FIGURE 3.—Relative densities of $\omega_{max}$ under three selection coefficients ($\alpha = 0$, $\alpha = 500$, and $\alpha = 2500$) for (a) a sweep in an equilibrium population for $n = 12$, (b) a sweep in an equilibrium population for $n = 50$, (c) a sweep in a bottlenecked population for $n = 12$, and (d) a sweep in the same bottlenecked population for $n = 50$. $4Nr = 100$, $4N\mu = 75$, $\tau = 0.0041$, and the length of the region is 10 kb. The time of reduction, the severity, and the duration of the bottleneck are taken from the THORNTON and ANDOLFATTO (2006) parameter estimates.

partially distinguishable between neutral and selected loci in bottlenecked populations. For $\alpha = 500$, the great majority of replicates are distinguishable from neutrality, with very large values ($\omega_{max} > 5$) being produced with high probability. While it is seemingly counterintuitive, larger values of $\alpha$ result in smaller values of $\omega_{max}$, particularly in the bottlenecked relative to equilibrium populations. For very large values ($\alpha = 2500$), the distribution is still partially distinct from neutrality, particularly in the direction of values of $\omega_{max}$ that are too small. The twofold diversity-reducing effect of a bottleneck plus a strong sweep largely eliminates variation within the 10-kb region. Accounting for departures in both directions, these results indicate that loci that have experienced recent and strong selection may often be identifiable in nonequilibrium populations (at least for the parameter space estimated by THORNTON and ANDOLFATTO 2006), with both small and large values of $\omega_{max}$ being consistent with selection ($\omega_{max} < 2$ and $\omega_{max} > 4$, respectively). This suggests that the $\omega_{max}$-statistic is of value when evaluating both African and non-African sequence data alike.

To better evaluate the utility of the $\omega_{max}$-statistic, we present receiver operating characteristic (ROC) curves. In brief, ROC curves plot power as a function of the false positive rate, where an ideal performance would be a

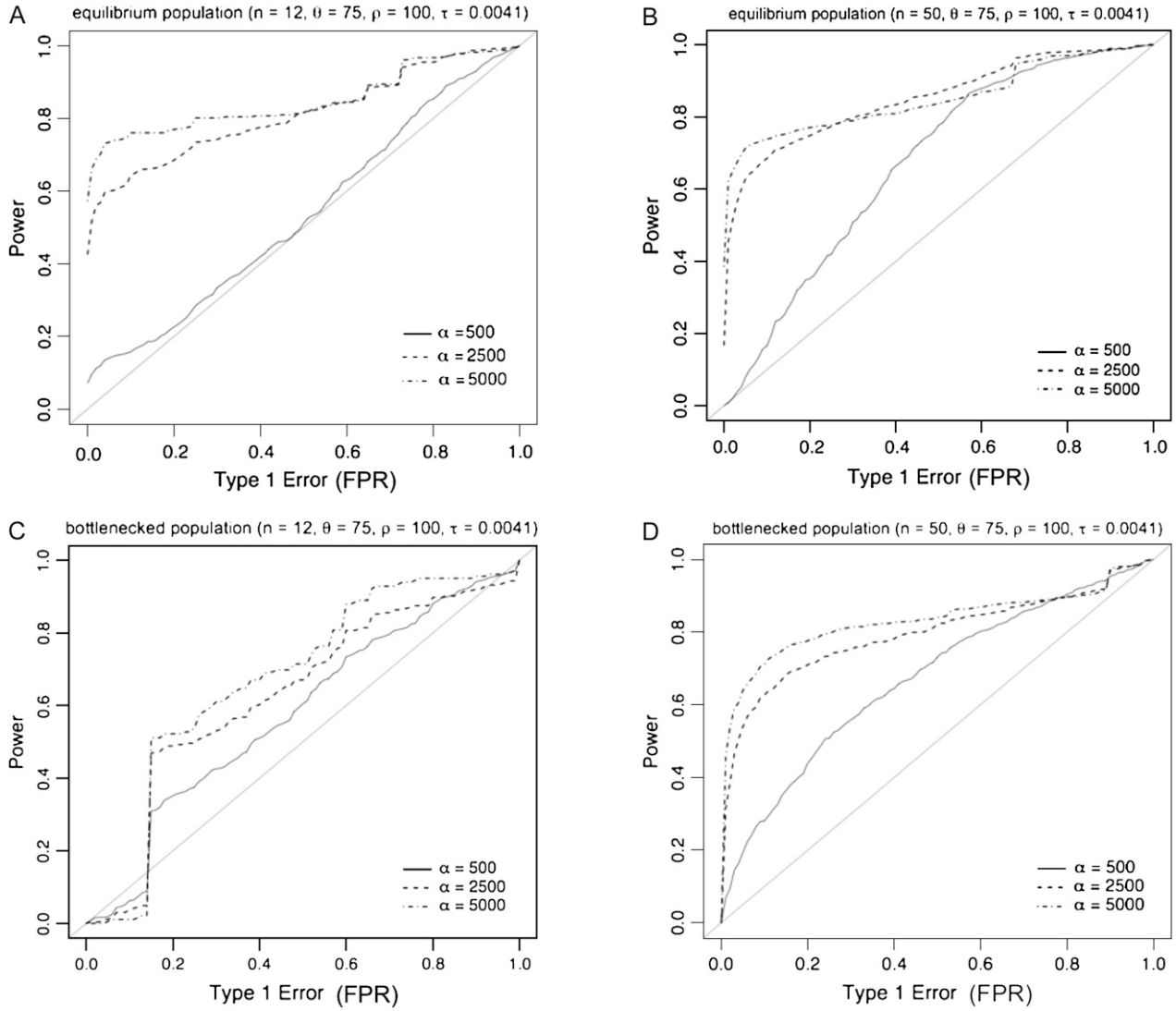


FIGURE 4.—Receiver operating characteristic (ROC) curves for selection in an equilibrium population and a bottlenecked population. The time of reduction, the severity, and the duration of the bottleneck are taken from the THORNTON and ANDOLFATTO (2006) parameter estimates. $4Nr = 100$, $4N\mu = 75$, $\tau = 0.0041$, and the length of the region is 10 kb. We present results for $\alpha = 500$, 2500, and 5000 in an equilibrium population for $n = 12$ (A) and $n = 50$ (B) and in the bottlenecked population for $n = 12$ (C) and $n = 50$ (D). The shaded diagonal line is representative of the situation in which there is an even trade-off between FPR and power.

curve near the left and the top of the graph (*i.e.*, high power is achieved with a very low FPR). The diagonal represents the situation in which there is a linear relationship between power and FPR (*e.g.*, 50% power corresponds to a 50% FPR). ROC curves are ideal for these comparisons, as they do not summarize performance merely at a single arbitrarily selected value, but across all possible values. The ROC curve can be used to evaluate the gain in power achieved by using a type I error rate other than the standard 0.05. In particular, one may prefer to choose a value that balances the probability of misclassification of either class [*i.e.*, the probability of false positives (*i.e.*, type I error) and false negatives (*i.e.*, power)].

Examining ROC curves for our bottleneck with selection data sets, we observe a number of interesting features (Figure 4). Once again, results are also presented for sweeps in an equilibrium population for comparison ($n = 12$, Figure 4A; $n = 50$, Figure 4B). For small sample sizes in a bottlenecked population ($n = 12$, Figure 4C), the $\omega_{max}$-statistic has 50% power to detect strong selection, if an ~15% FPR is accepted. Beyond that point, to increase power, a nearly linear increase in type I error must be accepted. Notably, for a 5% cutoff, the test statistic has almost no power. Reiterating a previous point, because this is a two-tailed test, a number of these rejections are in the direction of too *little* LD relative to the null, particularly for large $\alpha$. For larger

sample sizes ($n = 50$, Figure 4D), a different pattern is observed. A 5% FPR corresponds to ~60% power to detect strong selection. To achieve 80% power the accepted type I error would approach 30%. For weaker selection ($\alpha = 500$), a 5% FPR corresponds to ~20% power. While greater power is achieved with a lower FPR in equilibrium populations, these results indicate that $\omega_{max}$ is a useful statistic in bottlenecked population as well, as long as sample sizes are large ($n = 50$ *vs.* $n = 12$).

**Distinguishing recurrent-selective-sweep models from neutrality:** As an alternative to the single-sweep models discussed above, we also consider patterns of LD produced under recurrent-sweep models. The motivation for considering the recurrent hitchhiking model is that selective sweeps are a mutation-rate limited process, and the simulations of a sweep at a particular time do not account for having to wait for selected mutations to arise in populations. In particular, we examined parameter combinations relevant for both Drosophila ($\theta = 75$, $\rho = 100$) and humans ($\theta = \rho = 10$). Examining both $n = 12$ and $n = 50$ across all values, the resulting $\omega_{max}$-values do not differ significantly from those expected under neutrality (results not shown).

Examining ROC curves better reveals the difficulty of detecting recurrent selection. Figure 5 plots $n = 50$, $\theta = 75$, and $\rho = 10$ or 100 (Figure 5, A and B, respectively) for three rates of sweeps ($\lambda = 10^{-7}$, $10^{-6}$, and $10^{-5}$). For the low-recombination case (Figure 5A), the lowest rate of sweeps is essentially imperceptible, with a 50% FPR corresponding to ~50% power (solid line in Figure 5A). For higher rates of sweeps the situation is scarcely better, with a 5% FPR approaching only 30% power. For high-recombination regions the situation is slightly worse. Occasionally the performance is poorer than that of the null model (*i.e.*, the ROC curve is below the diagonal), owing to the fact that the distribution is contained completely within that of the null. Basically, for rare sweeps (solid line in Figure 5B), there is roughly the same mean, but less variance, in $\omega_{max.}$ Thus, these results indicate that recurrent selection is extremely difficult to detect using this statistic, as it is for other site-frequency spectrum-based approaches—particularly for Drosophila-like recombination parameters. For human-like recombination parameters, the situation is slightly better, primarily owing to the fact that lower recombination rates result in stronger patterns of LD across larger portions of the genome.

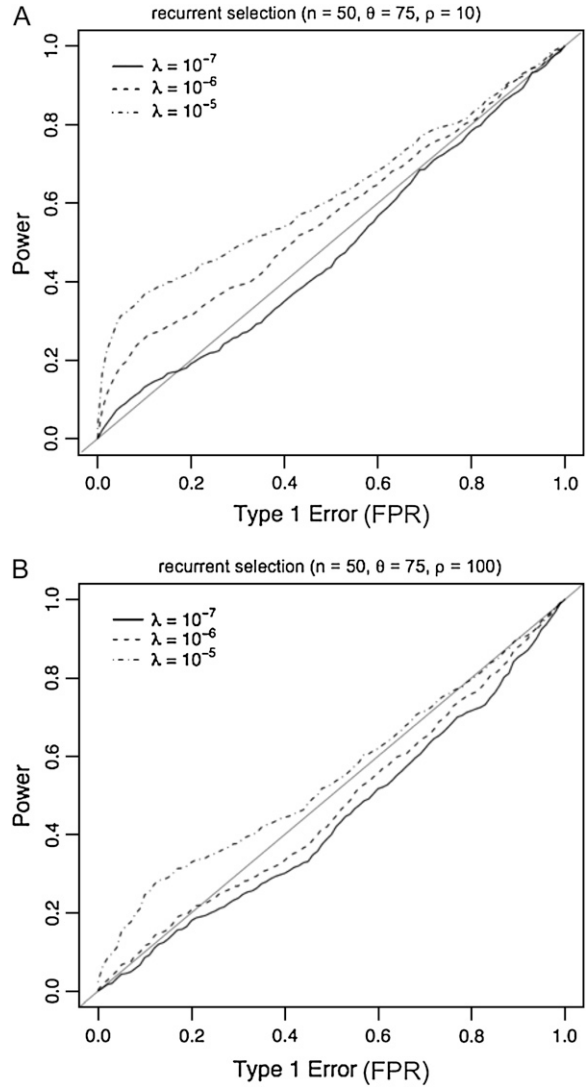


FIGURE 5.—ROC curves for recurrent selective sweeps. We present results for $\lambda = 10^{-7}$, $10^{-6}$, and $10^{-5}$ for $n = 50$, $\theta = 75$, and $\rho = 10$ (A) or $\rho = 100$ (B).

## DISCUSSION

While bottlenecks have been previously demonstrated to replicate other patterns of the site-frequency spectrum that are predicted under selection models, including an excess of high-frequency-derived alleles (*e.g.*, PRZEWORSKI 2002), we here observe that this includes spatial patterns of variation as well. We propose that the correlation between high-frequency-derived alleles and LD observed by KIM and NIELSEN (2004) and STEPHAN *et al.* (2006) is generated under neutral nonequilibrium models as well, with haplotypes that escape the bottleneck generating stretches of strong LD. As the bottleneck estimated for non-African populations of *D. melanogaster* is severe and strongly reduces diversity, it appears as though the target of selection ($l$) in the $\omega_{max}$-calculation is being maximized to regions of reduced or absent variation, and haplotype blocks that coalesce during the bottleneck are found in flanking regions with high probability.

Apart from considering a variety of nonequilibrium neutral, equilibrium-sweep, and non-equilibrium-sweep models, we also examine recurrent-sweep models. Although the fixed-$\tau$ case is of value for quantifying the performance of these approaches, the recurrent model is arguably nearer biological reality. While the true rate

of sweeps in natural populations is unknown, there are distinct challenges for both the high and low values of λ that we have considered. If the rate of sweeps is high, then there may be many recent sweeps across the genome that existing methods will have power to detect. However, if the rate is this great, then there is an appreciable probability that sweeps are occurring on already swept backgrounds. This multiple-sweep effect will result in very different patterns in the site-frequency spectrum, particularly with regard to high-frequency-derived alleles (Kim 2006), and thus linkage disequilibrium, owing to the correlation between the two statistics.

If the rate of sweeps is low, then sweeps will be old on average, and patterns of variability will have recovered (Przeworski 2002). In other words, a low rate implies that there will not be many regions of the genome that have experienced a recent enough sweep to be readily detectable by existing methods. Thus, the fixed-τ single-sweep simulations represent potentially infrequent evolutionary events. This argument of course relies on a uniform rate of sweeps over an organism's recent evolutionary history. Although this assumption may be approximately accurate, it is likely violated in a number of organisms. For example, many domesticated crop species have experienced very recent and extreme artificial selection—although the effect of overlapping sweep patterns will likely be of importance here as well. Additionally, under this "domestication" scenario, models that consider selection on standing variation will likely be more relevant than selection on new mutations, a process that results in very different patterns of variation (Przeworski *et al.* 2005).

To better evaluate the empirical relevance of the $\omega_{max}$-statistic, ROC plots were examined for a number of the most germane scenarios. Most significantly, for the bottleneck parameters inferred by Thornton and Andolfatto (2006), the $\omega_{max}$-statistic appears to have good power to differentiate adaptive loci from neutral loci in bottlenecked populations. For small sample sizes, accepting a 15% type I error corresponds to >50% power to detect selected loci (Figure 4C); and for larger sample sizes, a 5% type I error corresponds to >60% power, when selection is strong (Figure 4D). This result is extremely encouraging, given the difficulty that this bottleneck parameter space presents for existing and commonly used test statistics (*e.g.*, Jensen *et al.* 2005).

For recurrent selective sweeps, the situation appears less encouraging. Even for strong selection, large sample sizes, and low rates of recombination, a 5% FPR corresponds to only ~20% power to detect selected loci (Figure 5A). For other parameter combinations, the results are essentially near the null (Figure 5B). It is important to note that the $\omega_{max}$-statistic is designed for situations in which sequence data span the site of a fixed beneficial mutation. Thus, under a recurrent selection model in which sweeps are occurring across a genomic region that is very large relative to the sampled region, it may not be surprising that this statistic has low power. As such, performance will be maximized when the swept region has been previously localized, as is assumed in the fixed-τ simulations—although it is crucial to account for the ascertainment bias introduced by preselecting regions (Thornton and Jensen 2007). Either way, the challenges presented for both high and low rates of recurrent sweeps discussed above remain.

Given the difficulties observed under recurrent selection models when the target of selection has not been sequenced, we equally anticipate that the $\omega_{max}$-statistic will have limitations for detecting other types of selection. Specifically, while selection from standing variation generates patterns that differ strongly from neutrality (Przeworski *et al.* 2005), the allele will appear swept only if the selective pressure began while it was segregating at very low frequency (see Figure 7 of Stephan *et al.* 2006). Otherwise the expectation of strong LD flanking the target, and reduced LD across the target, described by Kim and Nielsen (2004) and Stephan *et al.* (2006), which the $\omega_{max}$-statistic is designed to detect, will not be created. Additionally, this statistic will likely be inappropriate to detect partial sweeps. Although this model may produce strong linkage disequilibrium, owing to the fact that the beneficial allele has undergone at least part of the rapid increase in frequency, the LD pattern discussed here is expected to be created only at the time of fixation. Other LD-based methods have been proposed that would be more appropriate for the detection of partial sweeps, such as the extended haplotype heterozygosity (EHH) approach (*e.g.*, Voight *et al.* 2005).

A number of important points need to be mentioned. First, these results are of particular relevance to derived populations of species that have experienced a population size reduction associated with colonization. Ancestral populations of these species with stable demographic histories are less likely to be producing spatial patterns of variation that replicate sweep predictions, particularly as population structure was not observed to replicate sweep-like patterns of LD. This suggests that searching for adaptively important loci in these more stable ancestral populations will likely be fruitful. Apart from nonequilibrium considerations, however, and given the recurrent-sweep results, the impact of different rates of recurrent sweeps (λ) needs to be considered when analyzing empirical data—regardless of whether the population is ancestral or derived. However, whether this rate is so great as to obscure individual sweep patterns, in humans or in flies or in any other natural population, remains an open question.

Nevertheless, simulations suggest that identifying adaptively important regions is possible even in bottlenecked populations, despite the fact that we observe neutral bottleneck models to be capable of producing patterns of LD previously proposed to be unique to hitchhiking models. Specifically, loci under strong

selection ($\alpha > 500$) produce a distribution of $\omega_{max}$ that is only partially overlapping with the neutral case—and we demonstrate that by accepting a modest type I error it is possible to achieve significant power. The direction of the rejection, however, differs along with the intensity of selection, with very strong selection rejecting because of too *little* LD relative to the null. These combined results suggest that the $\omega_{max}$-statistic should be used alongside SFS-based methods when analyzing polymorphism data and in particular that it appears to allow for the identification of adaptive loci even in nonequilibrium populations—a challenge that has historically been very difficult to address. If $N_e$ is taken to be on the order of $1 \times 10^6$, selection coefficients on the order of $s = 0.0025$ may be identifiable using existing statistics in ancestral and derived populations alike. Previous analyses suggest that selection coefficients of this magnitude are not unrealistic for natural populations (*e.g.*, Endler 1986). However, the true distribution is unknown, and uncoupling the average strength of sweeps from the average rate of sweeps remains a formidable and an important challenge (Wiehe and Stephan 1993; Kim 2006).

The authors thank Yuseob Kim, Rick Durrett, and two anonymous reviewers for helpful comment and discussion, as well as the Aquadro lab. J.D.J. is supported by National Institutes of Health (NIH) grant GM36431 to C. F. Aquadro and by National Science Foundation grant DMS-0201037 to R. Durrett, C. F. Aquadro, and R. Nielsen. K.R.T. is supported by NIH grant GM065509 to A. G. Clark. C.D.B. is supported by National Science Foundation grant 0516310.

## LITERATURE CITED

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. **66:** 129–138.

Eberle, M. A., M. J. Rieder, L. Kruglyak and D. A. Nickerson, 2006 Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. **2:** e142.

Endler, J. A., 1986 *Natural Selection in the Wild*, edited by R. M. May. Princeton University Press, Princeton, NJ.

Fay, J., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model. Bioinformatics **18:** 337–338.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics **170:** 1401–1410.

Kaplan, N. L., T. Darden and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 "The hitchhiking effect" revisited. Genetics **123:** 887–899.

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics **172:** 1967–1978.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513–1524.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566–1575.

Parsch, J., C. D. Meiklejohn and D. L. Hartl, 2001 Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics **159:** 647–657.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

Przeworski, M., G. Coop and J. D. Wall, 2005 Signatures of positive selection on standing variation. Evolution **59:** 2312–2323.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Stephan, W., Y. S. Song and C. H. Langley, 2006 Hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics **172:** 2647–2663.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis. Genetics **123:** 437–460.

Teshima, K .M., G. Coop and M. Przeworski, 2006 How reliable are empirical genome scans for selective sweeps? Genome Res. **16:** 702–712.

Thornton, K. R., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in non-African populations of *Drosophila melanogaster*. Genetics **172:** 1607–1619.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false positive rate in multi-locus genome scans for selection. Genetics **175:** 737–750.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA **102:** 18508–18513.

Wiehe, T. H., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–854.

Wootton, J. C., X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu *et al.*, 2002 Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature **418:** 320–323.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310–1314.

Communicating editor: D. M. Rand