Inferring Selection in Partially Sequenced Regions

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

Department of Molecular Biology and Genetics, Cornell University

A common approach for identifying loci influenced by positive selection involves scanning large portions of the genome for regions that are inconsistent with the neutral equilibrium model or represent outliers relative to the empirical distribution of some aspect of the data. Once identified, partial sequence is generated spanning this more localized region in order to quantify the site-frequency spectrum and evaluate the data with tests of neutrality and selection. This method is widely used as partial sequencing is less expensive with regard to both time and money. Here, we demonstrate that this approach can lead to biased maximum likelihood estimates of selection parameters and reduced rejection rates, with some parameter combinations resulting in clearly misleading results. Most significantly, for a commonly used sample size in Drosophila population genetics (i.e., $n = 12$), the estimate of the target of selection has a large mean square error and the strength of selection is severely under estimated when the true selected site has not been sampled. We propose sequencing approaches that are much more likely to accurately localize the target and estimate the strength of selection. Additionally, we examine the performance of a commonly used test of selection under a variety of recurrent and single sweep models.

Introduction

There is considerable interest in using population genetic approaches to identify regions of the genome that underlie population- or species-specific adaptations. These approaches can also be used to address basic evolutionary questions such as the relative importance of adaptive and demographic factors in shaping patterns of genome variability. The rapid increase in our ability to survey population level nucleotide variability for larger sample sizes and for larger portions of the genome yields increasing statistical power to distinguish among alternative population genetic models. At the same time, because more tests are being performed by each study (with more power per test), the chance of identifying false positives also increases dramatically.

Methods for identifying regions influenced by positive selection from sequence data rely on the expectation that the substitution of a strongly selected advantageous mutation alters the frequencies of linked neutral variation (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Stephan et al. 1992). These approaches can generally be divided into 2 classes. The first involves a scan in which outlier loci are identified that are not compatible with neutrality under some plausible demographic model (e.g., Schlotterer 2002; Kauer et al. 2003; Storz et al. 2004; Tenaillon et al. 2004; Altshuler et al. 2005; Bauer DuMont and Aquadro 2005; Ometto et al. 2005; Stajich and Hahn 2005; Wright et al. 2005). A related approach to detect selected loci has been to summarize the empirical, genome-wide background site–frequency spectrum from which outliers are identified (e.g., Nielsen et al. 2005; Williamson et al. 2005), though model-based comparisons are often necessary in order to assess significance. An important addition to this framework has been the ability to correct for the ascertainment bias introduced from choosing loci based on the presence of "sweep-like" characteristics (Thornton and Jensen 2007).

By identifying markers with skewed distributions or decreased variation, subsequent sequencing studies may be directed in order to determine if the observed patterns are consistent with a sweep hypothesis (e.g., Bauer DuMont and Aquadro 2005; Beisswanger et al. 2006; Pool et al. 2005; Jensen et al. 2007). Although partial sequencing is often used to quickly screen these large regions identified as being near putative selective sweeps and to better localize the target, the optimal way to sample these identified regions has not been systematically investigated.

We examine both models in which the age of a single selective sweep is fixed and known, as well as a model of recurrent selective sweeps. In the case of the former, we here assume that the departure originally detected, providing the motivation for regional localization, truly represents selection. As such, we suggest that these results be used in combination with the genome scan localization procedure of Thornton and Jensen (2007). We ask how best to sample these localized regions in order to obtain accurate estimates of selection parameters as well as provide available methods with enough information to reject neutrality. In the case of the latter, we assume that a randomly selected region has been sequenced and we determine the power of existing tests to reject neutrality when there is a background rate of selective sweeps in which advantageous mutations are uniformly distributed across a chromosome. We examine a wide range of parameter combinations, including those that are relevant for both Drosophila and humans. This analysis suggests that modifications to current strategies for sampling regions believed to be shaped by a selective sweep can lead to a greater accuracy of parameter estimates.

Methods

Modeling Selective Sweeps

We model positive selection using coalescent simulations for a region of M nucleotides, as described in equations 1–7 of Thornton and Jensen (2007). At time τ in the past (measured in units of $4N$ generations), a beneficial allele has fixed in the population at position X. For cases where the selected site is within the region, $1 \leq X \leq M$. For models of recurrent sweeps (see below), X may lie outside the M nucleotides.

[1] Present address: Department of Ecology, Behavior and Evolution, University of California, San Diego.

[2] Present address: Department of Ecology and Evolution, University of California, Irvine.

Key words: selective sweeps, natural selection, composite likelihood, recurrent selection.

E-mail: jjensen@ucsd.edu.

Mol. Biol. Evol. 25(2):438–446. 2008
doi:10.1093/molbev/msm273
Advance Access publication December 28, 2007

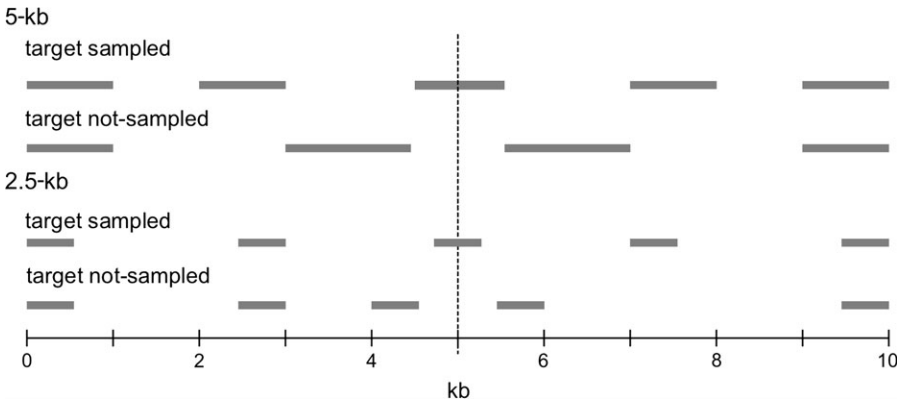

FIG. 1.—The 4 sampling schemes employed in this study; an example of the target of selection being sampled and not sampled for 2.5/10 kb and 5/10 kb. The target is at position $X = 5{,}000$. All partial data are parsed from the complete 10-kb data sets.

In addition to what was implemented in Thornton and Jensen (2007), we also simulate stochastic trajectories of beneficial alleles, conditioning on their reaching fixation in the population (Coop and Griffiths 2004; Przeworski et al. 2005). For a beneficial mutation at frequency x at time t, x jumps to either

$$x \rightarrow \mu(x)\Delta t - \sqrt{x(1-x)\Delta t},$$

or

$$x \rightarrow \mu(x)\Delta t + \sqrt{x(1-x)\Delta t},$$

with equal probability during the interval Δt. The term $\mu(x)$ is the infinitesimal mean change in allele frequency of the conditional process. For the case of genic selection considered here and conditional on the ultimate fixation of the beneficial mutation,

$$\mu(x) = \mu^{+}(x) = 2Nsx(1-x)/\tanh(2Nsx),$$

(Ewens 2004, p.170). In our implementation, we used $\Delta t = \frac{1}{20N}$, and $N = 10^6$.

Recurrent Selective Sweeps

We also considered a model of selective sweeps occurring in the genome at a rate determined by Λ, the expected number of sweeps per recombination unit in the last $4N$ generations (Kaplan et al. 1989; Braverman et al. 1995). Our implementation follows that described in Przeworski (2002), with 2 modifications. First, the allele frequency trajectory of the selected site is determined stochastically, as described above. Second, we allow for the selective sweeps both within the region of M nucleotides as well as at linked sites. We do this because we simulate relatively large neutral regions ($M = 10^4$), and the probability of a sweep within that region may not be negligible for large Λ, assuming a constant Λ across the genome. Similarly, it is important to consider sweeps outside of the M nucleotides as they will impact patterns of variation within the region under investigation. In this model, the time until the next selective phase is entered is exponentially distributed with rate $8Ns\Lambda/\rho_{bp} + M\Lambda$, where ρ_{bp} is the scaled recombination rate between adjacent base pairs. 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 2α on either side of the sampled region (see Kaplan et al. 1989; Durrett and Schweinsberg 2004, for details).

We estimated the power to reject the equilibrium neutral model using 2 sample sizes ($n = 12$ and 50) and 90 parameter combinations generated by considering all combinations of $\theta \in \{10, 75\}$,$\rho \in \{10, 50, 100\}$, $\alpha \in \{100, 500, 1{,}000, 2{,}500, 5{,}000\}$, 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 ($\Lambda = 10^{-5}$, $\alpha = 5{,}000$). For these simulations, we used $N = 10^6$.

For each simulated replicate, we also calculated the power P values for D of Tajima (1989) and H statistics of Fay and Wu (2000), using 1-tailed tests (of the lower tail) for both statistics. In order to make the power estimates of D and H comparable with those from the composite likelihood ratio test (CLRT), we assumed that ρ is known precisely.

Sampling

In order to evaluate the effects of partial data for the single sweep data sets, a number of sampling schemes were evaluated. First, 1,000 replicates of complete 10-kb data sets were simulated for $n = 50$, $2Ns = 0$, 100, 500, and 1,000; $\rho_{bp} = 0.05$ and 0.1; $\theta = 15$ and 75; and $\tau = 0.001$, 0.01, and 0.02. Then, using these data, partial data sets were parsed in 4 configurations: 5 or 2.5 kb of sequence distributed across the 10-kb region, including for each a scenario in which the selected site does and does not fall in a sampled region (fig. 1). In all cases, the target of selection is at position $X = 5{,}000$, and there is sequencing on both sides of the target. These parameters were chosen for their relevance to a significant portion of the *Drosophila melanogaster* genome (e.g., $\rho_{bp} = 0.05$ means a recombination rate of 1.25×10^{-8}/base pair/generation over a 10-kb region for $N_e = 1 \times 10^6$, and $\theta = 75$ means $\mu = 1.87 \times 10^{-8}$/base

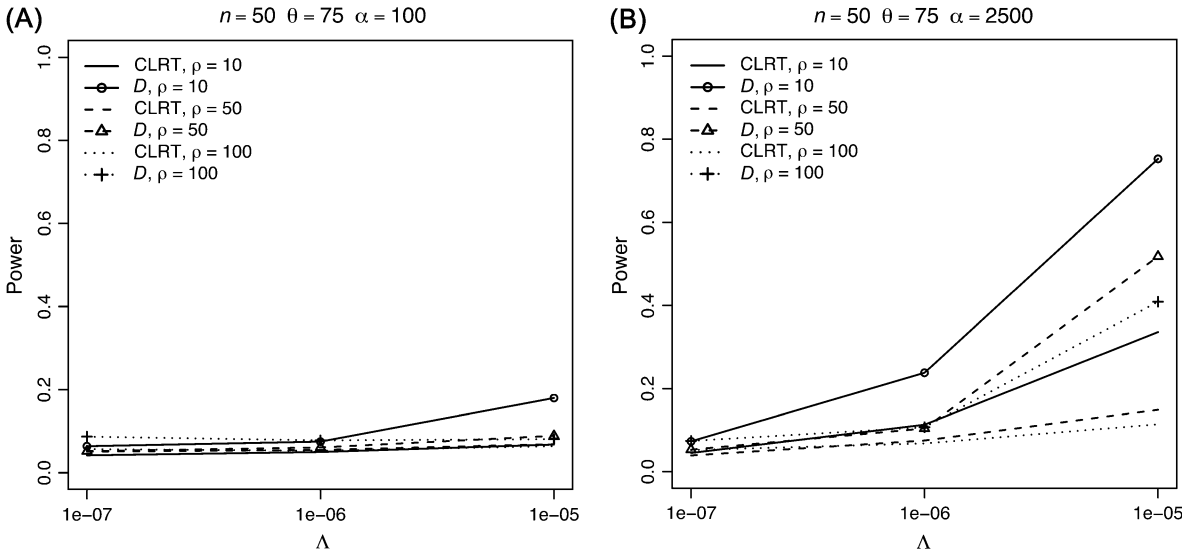

FIG. 2.—Power of the CLRT and Tajima's D under recurrent hitchhiking for $n = 50$, $\theta = 75$, and (*A*) $\alpha = 100$ and (*B*) $\alpha = 2{,}500$.

pair/generation over a 10-kb region for $N_e = 1 \times 10^6$). Additionally, the size of the region (10 kb) was chosen as it encompasses perturbations of the site-frequency spectrum produced after a selective sweep, for the values of $2Ns$ here considered (Kim and Stephan 2002).

In order to replicate a likely empirical approach, we simulated a second round of sequencing by adding data around the predicted target and then reanalyzing the data set. This was done by assuring that the predicted target had at least 0.5 kb on either side, which, depending on where the prediction was made relative to the initial segments, meant adding anywhere between 0.5 and 1 kb of new data.

Statistics

Let $\hat{\alpha}$ and $\hat{X}$ be the maximum likelihood estimates (MLEs) of the strength of the selection parameter ($2Ns$) and target of selection, respectively. These parameter estimates are found via maximization of the composite likelihood function of Kim and Stephan (2002) so that

$$\{\hat{\alpha}, \hat{X}\} = \arg\max_{\alpha, X \in \Theta} L_S(\alpha, X|\text{Data}),$$

where

$$L_S(\alpha, X|\text{Data}) = P(\text{Data}|\alpha, X) = \prod_{i=1}^{L} P(Y_i = y_i|\ \alpha, X),$$

where L is the length of the sequence, y_i (for $i = 1, \ldots, L$) denotes the observed count of the derived nucleotide at the ith site with corresponding random variable $Y_i \in \{0, 1, \ldots, n-1\}$, and $P(Y_i|\alpha, X)$ is given by equation (5) of Kim and Stephan (2002), using $\epsilon = (2\alpha)^{-1}$.

Two statistics were utilized to evaluate the MLEs of X and α. First, in order to measure any biases in the predicted location of selection introduced by partial sampling, relative bias (RB) was determined from 1,000 replicates, conditional on rejecting neutrality, as:

$$\text{RB} = \text{Mean}(\hat{X} - X)/X.$$

Second, in order to measure deviations from the expected values, the relative mean square error (RMSE) was determined as:

$$\text{RMSE} = \text{Mean}\ (\hat{X} - X)^2 \Big/ X^2.$$

The RB and RMSE were also calculated for α in an identical way.

Results

Rejecting Neutrality in Favor of Selection—Single Sweep Model

Applying the CLRT to our partial and complete data sets, we see that with less data the null is rejected less often (supplementary table 1, Supplementary Material online). For high recombination ($\rho_{bp} = 0.1$) and $\theta = 75$, with a complete 10 kb of sequence, the neutral model is rejected in favor of the sweep model in 95–97% of simulated sweep data sets when α is very large (≥ 500), and in approximately 77–82% of cases when $\alpha = 100$ for very recent sweeps ($\tau = 0.001$ in units of $4N$ generations). Predictably, as τ increases, these rejection rates decrease (supplementary table 2, Supplementary Material online [$\tau = 0.01$] and supplementary table 3, Supplementary Material online [$\tau = 0.02$]). In partially sequenced regions when the target of selection has been sampled, rates of rejection are nearly equivalent, except for $\alpha = 100$. When the target has not been sampled, these rejection rates are uniformly lower—rejecting approximately 92%, 83%, and 19% of the time for $\alpha = 1{,}000$, 500, and 100, respectively, for the 5-kb data set, where $n = 50$, $\theta = 75$, and $\tau = 0.001$.The primary factor determining rejection remains whether or not the site of

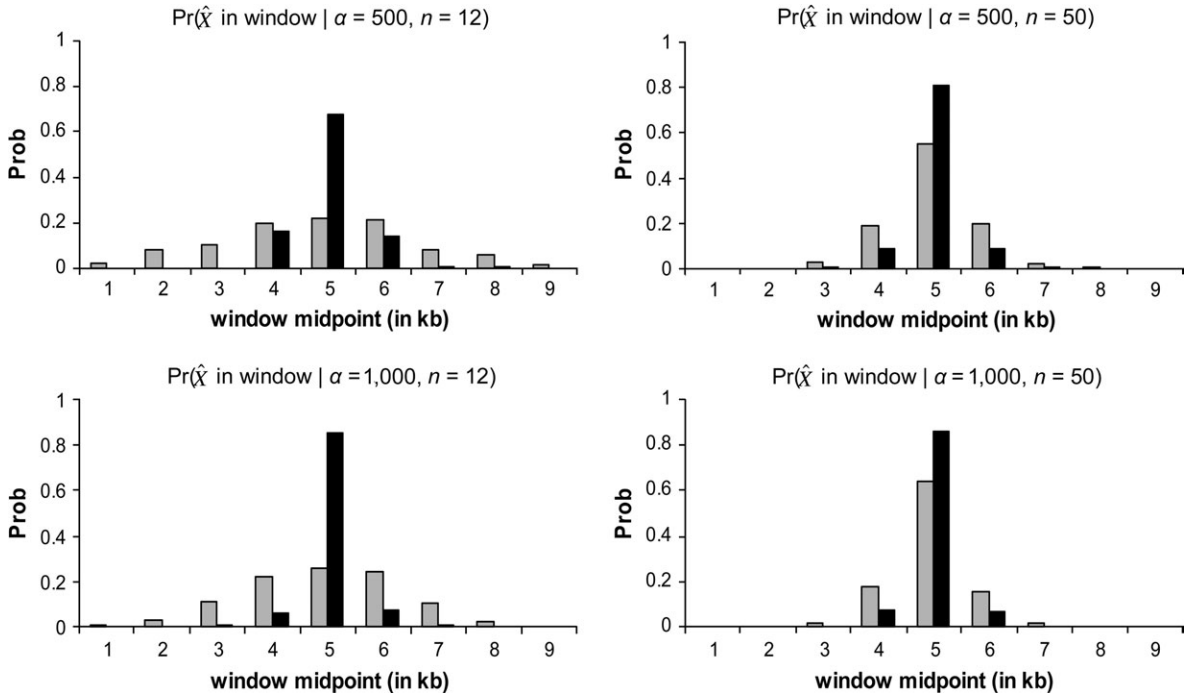

FIG. 3.—Probability of X being placed within a 1-kb window centered around positions 1, 2, 3 kb, etc., for 2 large values of α and common ($n = 12$) and large ($n = 50$) sample sizes, for the 5-kb partial data sets in which the target has (black) and has not (gray) been sampled. In all cases, $\rho_{bp} = 0.1$ and $X = 5{,}000$.

selection has been sequenced (supplementary tables 1–3, Supplementary Material online). Thus, although the ability to detect selection is diminished under all partial sampling schemes, the effect is simply to make the test more conservative with respect to rejecting neutrality.

Rejecting Neutrality in Favor of Selection—Recurrent Sweep Model

The simulation results presented above assume a single selective sweep fixing at time τ in the past. For considering the power of the CLRT when applied to genome scan data, it is appropriate to consider a model where τ is a random variable determined by Λ, the rate of sweeps in the genome (per recombination unit per $4N$ generations), $\alpha = 2Ns$, and $\rho = 4Nr$.

The parameters of this model have important implications. If the rate of sweeps is high, then there may be many recent sweeps across the genome which existing methods could 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 (Kim 2006). If the rate of sweeps is low, then many sweeps will be old enough that patterns of variability will have recovered (Przeworski 2002). As a consequence, the CLRT has low power to reject the null model, unless both Λ and α are large (e.g., fig. 2). Further, Tajima's D was observed to be generally more powerful than the CLRT and the power of Fay and Wu's H was never estimated to be greater than 10% (supplementary table 4, Supplementary Material online). These results are qualitatively similar to those of Przeworski (2002). Further, power was higher in regions of low recombination (fig. 2, supplementary table 4 [Supplementary Material online]) and increased with larger sample size. Parameter combinations for which a test's power is observed to exceed 0.5 are noted in bold on supplementary table 4 (Supplementary Material online).

Inferring the Target of Selection

Among the single sweep data sets that rejected the CLRT in favor of selection, we evaluated the accuracy of target prediction as measured by the RMSE, as well as the RB in the MLEs of the target of selection (as described in the Methods section). When a high recombination ($\rho_{bp} = 0.1$) region is fully sequenced, $\theta = 75$, and the sweep is very recent ($\tau = 0.001$), the estimate of the target is within the correct 1-kb window that encompasses the true target with probability 0.89, 0.87, and 0.84 for $\alpha = 1{,}000$, 500, and 100 for $n = 12$, respectively (representative cases illustrated in fig. 3). In the 5-kb partially sequenced regions in which the target has been sampled, these probabilities are similar except for low α, in which the probability drops to around 0.65, regardless of the sample size. When the target has not been sampled, however, the situation is considerably different. For a commonly used sample size ($n = 12$), very large selection coefficients ($\alpha = 500$, 1,000), recent sweeps, and having sequenced regions immediately flanking the true target, the MLE only has a probability of roughly 1/3 of being within the correct 1-kb window. Figure 3 visualizes these results for a subsample of our data. Full results across all parameter combinations are presented

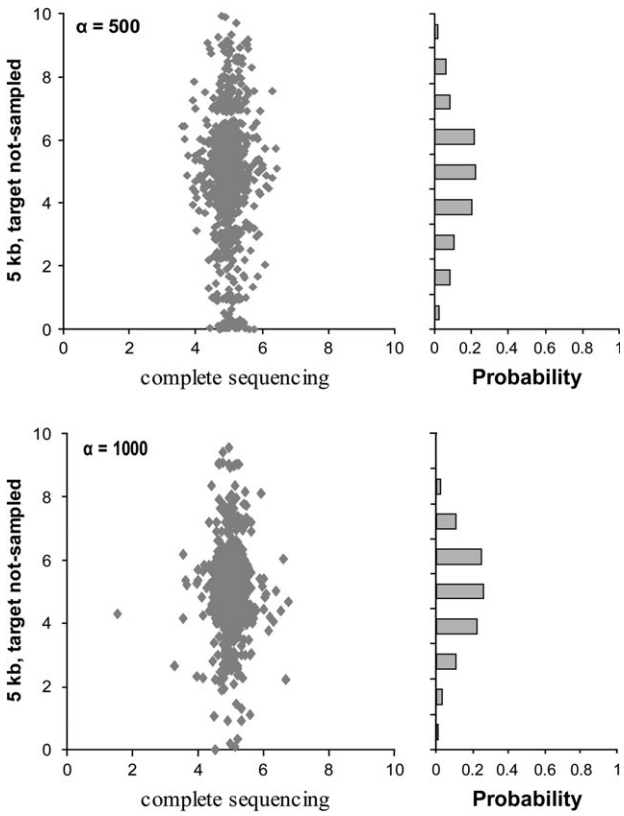

FIG. 4.—An example of the advantage of complete over incomplete sequencing for the MLE of the target of selection for 2 different large values of α. On the *x* axis is the distribution of the target prediction for the complete data set. On the *y* axis the same distribution for the 5-kb data set in which the target has not been sampled. $X = 5{,}000$, $\rho_{bp} = 0.1$, and $n = 12$. Note, the partial data set is parsed from the complete data set. On the parallel is the 5-kb distribution in histogram form (adapted from fig. 3) for clarity.

in supplementary tables 1–3 (Supplementary Material online).

As the partial data sets are simply subsamples of the complete data sets, it is possible to examine directly the benefits of complete versus incomplete sequencing. For example, figure 4 summarizes the improvement in the MLE of the target of selection of a complete, 10-kb data set over a data set in which only half of the region has been sequenced (5 kb), but the true target of selection (at position 5 kb) has not been sampled. Consistent with the RMSEs presented in supplementary tables 1–3 (Supplementary Material online), we see a wide range of target predictions when the site of selection has not been sampled and a relatively small range in the complete data set. In order to further explore this issue, we selected a small number of scenarios and fixed the number of segregating sites between the complete and partial data sets in order to determine if the performance is based simply on the fact that the complete data sets have approximately twice the number of segregating sites as the 5/10 kb data sets. Under this scheme, we observed results that are very similar to our fixed θ results presented above. We note, however, that this example is illustrative only because fixing *S* creates the problem that the $\Pr(S|\theta)$ would be drastically different between the partial and complete data sets. The average number of segregating sites produced under each set of parameters is given in supplementary tables 1–3 (Supplementary Material online).

Examining the relative bias, we observe no significant skew in the prediction of the location of the target under any sampling scenario (supplementary tables 1–3 Supplementary Material online). In order to evaluate whether the performance in these complete 10-kb data sets was being maximized by sequencing symmetrically around the target, we also evaluated otherwise identical data sets with the target at position 1 kb rather than 5 kb. There were no significant differences with regards to either RB or RMSE.

In order to better replicate a typical empirical approach, we examined a sample of the above described scenarios ($n = 12$, $\rho_{bp} = 0.1$, $\theta = 75$, and $\tau = 0.001$) to determine the extent to which target prediction is improved by "resequencing" around the predicted target (fig. 5, supplementary table 5 [Supplementary Material online]). For the data sets consisting of five 0.5-kb regions, we added a sixth fragment encompassing the predicted target (by taking it from the corresponding 10-kb data set), both for scenarios in which the true target has, and has not, been sampled. Note that we simply assure that there is 1 kb of data surrounding the predicted target, so, depending on whether this happens to overlap with an existing fragment, this additional data could represent between 0.5 and 1 kb of new sequence (see Methods).

There are 3 observations of particular interest. First, there is a strong correlation between target predictions between the first and second round of sampling, particularly when the true target was not originally sampled. This is owing to the fact that the second sampling does not represent an independent draw—rather it is simply an addition of a relatively small amount of data. Second, in data sets in which the true target was not originally sampled, this additional sequencing makes a measurable improvement in a proportion of replicates. This is shown clearly in figure 4 by the horizontal grouping centered around 5 kb, demonstrating a wide range of primary target predictions and more accurate secondary MLEs. However, it is worth noting that the improvement seen by resequencing is scarcely comparable with the accuracy associated with complete sequence, where the RMSE for $\hat{X}$ is 0.0548 for the resequenced data set and 0.0083 for the complete data sets, when $\alpha = 500$ (supplementary table 5, Supplementary Material online). Finally, the MLEs are not investigated under the recurrent selection model as localization would not be attempted if the pattern of hitchhiking was not initially detectable. As shown in the power analysis (supplementary table 4, Supplementary Material online), the probability of rejection under recurrent hitchhiking models rarely exceeds 10% for the CLRT. In the cases where rejections do occur, the same limitations of partial sequencing for target site estimation are expected as were described for the single sweep model.

Estimating the Strength of Selection

Evaluating the MLEs for data sets that rejected in favor of the selection model, we determine the RB in the estimated strength of selection (supplementary tables 1–3,

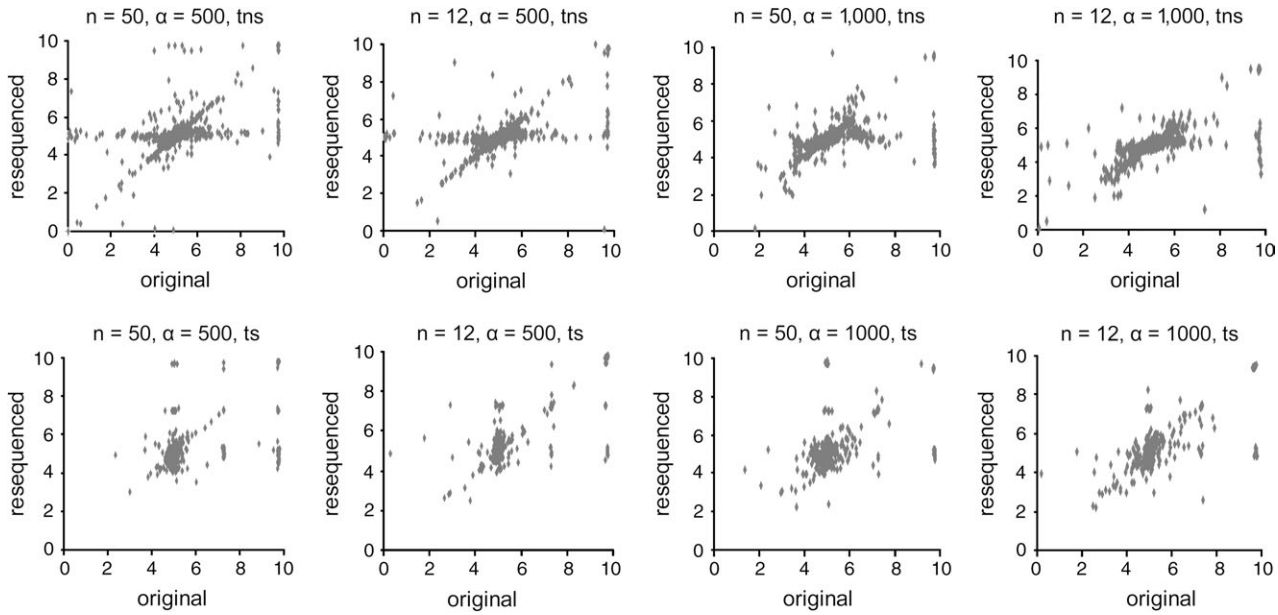

FIG. 5.—An example of a likely empirical approach, in which the original site of the target prediction is sequenced in a follow-up study. On the top are data sets in which the target was not originally sampled; on the bottom data sets in which it was. For each of the 4 cases, we show for comparison the distribution for $n = 12$ and $n = 50$, as well as $\alpha = 500$ and $\alpha = 1{,}000$. In all cases $\rho_{bp} = 0.1$, $X = 5{,}000$, and the original sampling scheme (x axis) is comprised of 2.5 kb of data in the 10-kb region. The resequencing (y axis) involves ensuring that there is at least 0.5 kb of data on both sides of $\hat{X}$ (see Methods). ts = target sequenced; tns = target not sequenced.

Supplementary Material online). We observe a stark underestimate of α under all scenarios of partial sequencing (i.e., the RB for $\hat{\alpha}$ is nearly always negative regardless of recombination rate, whether the target has been sampled, sampling scheme, or sample size). In regions of high recombination for the complete 10-kb data sets, we observe only a small RB in these estimates across all sample sizes and selection coefficients. As τ increases, however, this bias becomes increasingly more negative, owing to the assumption of the CLRT that the sweep has just ended. However, the RMSE on these estimates remains large even in the completely sequenced data sets. We observe similar relative biases across all partial sequencing scenarios, with relatively little difference between samples in which the target has and has not been sampled. The variance of the estimate may be decreased slightly by having a larger sample size (note that the numerator of the RMSE expression is equivalent to the variance of the estimator, thus a smaller RMSE implies a smaller variance). As with rates of rejection and the MLE of $\hat{X}$, the performance is consistently, though mildly, worse across all scenarios when the recombination rate of the region is reduced by half (supplementary tables 1–3, Supplementary Material online).

Application to Data

The challenges associated with target site prediction are illustrated by 2 recent experimental data sets. First is the putative sweep around the *wapl* region of *D. melanogaster*, which was inferred from partial data (roughly 6 kb of total data distributed in 12 fragments across a 110-kb region for a sample size of $n = 12$; Beisswanger et al. 2006). We evaluated the ability of the MLE to accurately estimate the location of the target of selection by generating, via parametric bootstrap, 1,000 sweep replicates using the African parameters given in Beisswanger et al. (2006) (location of sequenced regions, θ, recombination, the selection coefficient [$\hat{\alpha}$], and the target of selection [$\hat{X}$]). We found that target prediction is very poor in this case, with only a 20% chance of the target being placed within the correct 10-kb window and a 2% chance of being in the proper 1-kb window. We note that the 95% confidence intervals (constructed using the percentile method)

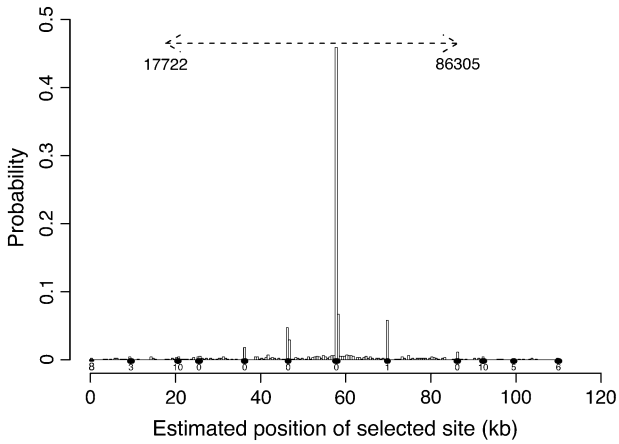

FIG. 6.—The distribution of the MLE of $\hat{X}$ for 1,000 selection simulations obtained via parametric bootstrap using the parameters specified for the *Drosophila melanogaster* Zimbabwe data (the population showing the strongest evidence of a sweep) in Beisswanger et al. (2006). The sweep was simulated as though it had just ended ($\tau = 0$), a value that allows the test to perform much better than using their inferred ancestral sweep value. The lines indicate the 95% confidence interval on their estimate of the target ($\hat{X}$=49.8 kb). Note that we have indicated the positions of their sequenced fragments, as well as the number of segregating sites observed in each region.

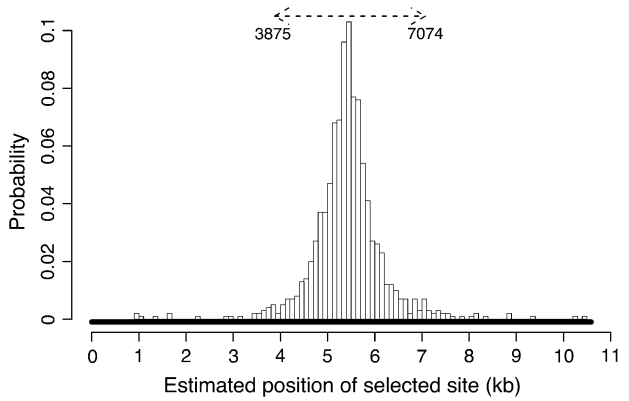

FIG. 7.—The distribution of the MLE of $\hat{X}$ for 1,000 selection simulations obtained via parametric bootstrap using the parameters specified from the *Drosophila melanogaster Notch* region analysis of Bauer DuMont and Aquadro (2005). The lines indicate the 95% confidence interval on their estimate of the target ($\hat{X}$=5,426). Note that the entire 10.5-kb region was completely sequenced in all lines ($n = 15$) for this USA population sample, and 147 segregating sites were observed.

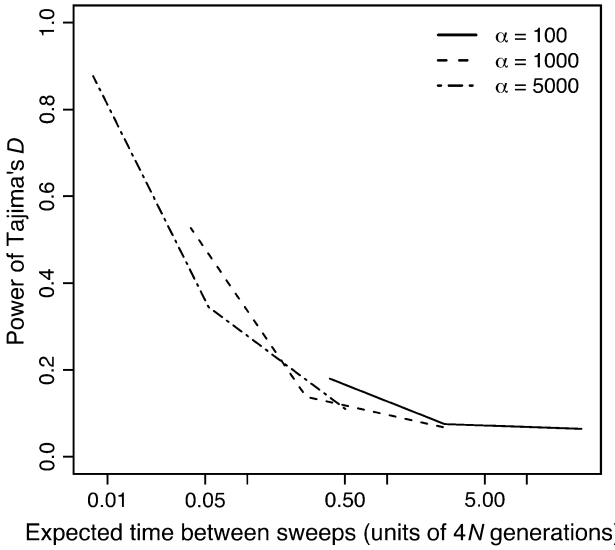

FIG. 8.—Power of Tajima's D as a function of the time between successive hitchhiking events. For the recurrent sweep model considered here, the expected time between sweeps is given by $E\left[t_L\right] + E\left[t_S\right] = -\frac{\log \xi}{\alpha} + \frac{1}{8Ns\Lambda/\rho_{bp}+M\Lambda}$, in units of $4N_e$ generations. The values on the x axis are calculated for $N = 10^6$, $\xi = 1/2N$, $\rho_{bp} = 0.001$, α=$\{100, 1,000, 5,000\}$, and $\Lambda \in \{10^{-7}, 10^{-6}, 10^{-5}\}$. The estimates of power are taken from supplementary table 4 (Supplementary Material online) for the case $n = 50$, $\theta = 75$, and $M = 10$ kb.

on their estimate of $\hat{X}$ spans nearly 68 kb for $\tau = 0$ or 65% of the region (fig. 6). We also see a grouping of target predictions to fragments where sequence data exist, emphasizing that there is no information about X where data are missing. For comparison, we also set the age of the sweep (τ) to its minimum value necessary to be consistent with their ancestral sweep hypothesis ($\tau = 0.019$ in units of $4N$ generations, based on Bayesian estimates of the colonization time presented in Thornton and Andolfatto 2006). In this case, the 95% confidence intervals span 90 kb or approximately 81% of the region examined.

The putative sweep downstream of the *Notch* locus in *D. melanogaster* (Bauer DuMont and Aquadro 2005) provides a second illustrative empirical example, in this case, determined by a complete sequencing approach. Approximately 10.5 kb of contiguous sequence was generated for a sample size of $n = 15$ for a USA population sample after initially identifying the candidate sweep region through a large-scale microsatellite screen. Using all parameter estimates presented in that paper, we see that the target prediction has a 62% chance of being in the proper 1-kb window and the 95% confidence interval spans approximately 3 kb, or 28% of the region examined (fig. 7).

Based on these combined results, we propose that parametric bootstrapping to obtain confidence intervals is appropriate for quantifying uncertainty in parameter estimates and is informative when presenting and interpreting results from the CLRT. We note that as the CLRT is widely used in tandem with a recently proposed goodness-of-fit test (Jensen et al. 2005), the null simulations from that test could be used to construct these confidence intervals.

Discussion

Simulations were used to investigate the effects of different sequencing sampling strategies on the ability to detect signatures of hitchhiking along a recombining chromosome, particularly using the CLRT proposed by Kim and Stephan (2002), which allows a prediction of both the target location and strength of selection. Comparing single sweep versus recurrent sweep models, we found that Tajima's D had much more power to reject neutrality under recurrent hitchhiking models than either the CLRT or Fay and Wu's H (supplementary table 4, Supplementary Material online). Further, under the recurrent hitchhiking model, a deficit of high-frequency derived sites is observed for some parameter combinations (Przeworski 2002; Kim 2006), explaining why we sometimes estimated H to have a power less than 5%, the nominal type I error under the null model (supplementary table 4, Supplementary Material online). Thus, although there is a considerable hitchhiking effect for large Λ and α, the signature of selection observable in the data is a reduction in diversity and an excess of rare alleles, rather than an excess of high-frequency derived alleles. Only very recent sweeps appear to be detectable using the CLRT because for older sweeps, the pattern of variation will have recovered somewhat, as noted by Przeworski (2002).

The cases where D had high power to reject the null model were for high rates of strong sweeps in regions of low recombination. It is useful to consider what the rate of sweeps must be in order for the power to reject the null model to be high. For the case of $\alpha = 5{,}000$, $\Lambda = 10^{-5}$, and $\rho = 10$ (i.e., very strong and very common), sweeps are occurring on average every $\approx$0.008 time units ($4N$ generations) for the 10^7 bp region examined here, and Tajima's D rejects the null model 87.7% of the time (for large sample sizes in regions of relatively low recombination; fig. 8). Such frequent and strong sweeps would have nearly chromosome-wide effects on levels of variability (Braverman et al. 1995), and it remains to be determined if such a large mutation rate to strongly selected mutations is biologically reasonable. A further discussion is found in Thornton et al. (2007).

For single, recent selective sweeps, we found that the Kim and Stephan (2002) CLRT was sensitive to the use of partial sequence data, with the MLEs of the strength and location of selection being potentially biased and widely variable for sparsely sampled regions. We also demonstrate that all aspects of detection are improved in regions of high recombination, though reducing recombination by half only does mildly worse.

Additionally, whereas sampling only partial segments across the region leads to lower rates of rejection and higher RMSEs for both $\hat{\alpha}$ and $\hat{X}$, the principle factor dictating performance remains whether the target has been sampled. Thus, smaller data sets are shown to be undesirable if for no other reason than this effectively decreases the probability of sampling the target. With regard to sample size, although $n = 12$ summarizes the site-frequency spectrum sufficiently to provide accurate MLEs in a complete 10-kb data set, and does reasonably well in partial data sets in which the target has been sampled, there is a marked difference between small and large sample sizes when the target has not been sequenced. Importantly, it is unwise to reason that the target has been placed accurately just because it falls within a sequenced segment.

By adding an additional sequenced fragment encompassing the initially predicted target of selection, we examined the relative benefit of follow-up sequencing aimed at refining the true targets location. We observe a strong correlation between primary and secondary predictions, though we note a marked improvement in a small proportion of the resequenced data sets. Thus, the addition of more data around the first estimated target, $\hat{X}$, particularly when the target was not originally sampled, leads to small improvements but is far less reliable than an initial analysis based on complete sequencing.

Thus, although partial sequencing has oft been employed for reasons both financial and practical, we demonstrate that when regions are localized through initial marker screens, complete sequencing offers far superior results in terms of the probability of rejecting neutrality in favor of selection, as well as in estimating the selection coefficient and target of selection. Although this proposal may seem sequence intensive, we note that the approach in distant second with regard to all of these measures (sequencing half of the region for $n = 50$) represents more than a 2-fold increase in data generation given that complete sequencing performs well for sample sizes of $n = 12$ (e.g., 5 kb for $n = 50$ represents 250 kb of total sequencing vs. 10 kb for $n = 12$ that represents 120 kb).

Supplementary Material

Supplementary tables 1–5 are available at *Molecular Biology Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We appreciate fruitful discussion with the Aquadro lab, particularly Vanessa Bauer DuMont, as well as comments on the manuscript from Yuseob Kim and several reviewers. This research was supported by National Institutes of Health grant GM36431 to C.F.A., National Science Foundation grant DMS-0201037 to R. Durrett, C. F. Aquadro, and R. Nielsen, a Sloan postdoctoral fellowship in Computational Molecular Biology to K.R.T, and a National Science Foundation postdoctoral fellowship in Biological Informatics to J.D.J.

Literature Cited

Bauer DuMont V, Aquadro CF. 2005. Multiple signatures of positive selection downstream of *notch* on the X chromosome in *Drosophila melanogaster*. Genetics. 171:639–653.

Beisswanger S, Stephan W, De Lorenzo D. 2006. Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. Genetics. 172:265–274.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 140:783–796.

Coop G, Griffiths RC. 2004. Ancestral inference on gene trees under selection. Theor Popul Biol. 66:219–232.

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

Ewens W. 2004. Mathematical Population Genetics I. Theoretical Introduction, 2nd. Springer-Verlag, New York.

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

Jensen JD, Bauer DuMont V, Ashmore AB, Gutierrez A, Aquadro CF. 2007. Patterns of variability and divergence at the diminutive gene region of Drosophila melanogaster. Genetics. 177:832–840.

Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170:1401–1410.

Kaplan NL, Hudson RR, Langley CH. 1989. "The hitchhiking effect" revisited. Genetics. 123:887–899.

Kauer MO, Dieringer D, Schlotterer C. 2003. A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of *Drosophila melanogaster*. Genetics. 165:1137–1148.

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

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

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

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22:2119–2130.

Pool JE, Bauer DuMont V, Mueller JL, Aquadro CF. 2005. A scan of molecular variation leads to the narrow localization of a selective sweep affecting both afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics. 172:1093–1105.

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

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

Schlotterer C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics. 160:753–763.

Stajich ES, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 22:63–73.
Stephan W, Wiehe THE, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol. 41: 237–254.
Storz JF, Payseur BA, Nachman MW. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of African. Mol Biol Evol. 21:1800–1811.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis. Genetics. 123:437–460.
Tenaillon MI, U'Ren J, Tanaillon O, Gaut BS. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol. 21:1214–1225.
The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature. 437:1299–1320.
Thornton KR, Andolfatto P. 2006. Approximate bayesian inference reveals evidence for a recent, severe, bottleneck in non-African populations of *Drosophila melanogaster*. Genetics. 172:1607–1619.
Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 175: 737–750.
Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity. 98:340–348.
Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA. 102:7882–7887.
Wright SI, Bi IV, Gaut BS. 2005. The effect of artificial selection on the maize genome. Science. 308:1310–1314.

Michael Nachman, Associate Editor

Accepted December 1, 2007