A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data

Ha My T. Vy* and Yuseob Kim†,1

*Interdisciplinary Program of EcoCreative and †Department of Life Science, Ewha Womans University, Seoul, Korea 120-750

ABSTRACT Adaptive evolution occurs as beneficial mutations arise and then increase in frequency by positive natural selection. How, when, and where in the genome such evolutionary events occur is a fundamental question in evolutionary biology. It is possible to detect ongoing positive selection or an incomplete selective sweep in species with sexual reproduction because, when a beneficial mutation is on the way to fixation, homologous chromosomes in the population are divided into two groups: one carrying the beneficial allele with very low polymorphism at nearby linked loci and the other carrying the ancestral allele with a normal pattern of sequence variation. Previous studies developed long-range haplotype tests to capture this difference between two groups as the signal of an incomplete selective sweep. In this study, we propose a composite-likelihood-ratio (CLR) test for detecting incomplete selective sweeps based on the joint sampling probabilities for allele frequencies of two groups as a function of strength of selection and recombination rate. Tested against simulated data, this method yielded statistical power and accuracy in parameter estimation that are higher than the *iHS* test and comparable to the more recently developed nS_L test. This procedure was also applied to African *Drosophila melanogaster* population genomic data to detect candidate genes under ongoing positive selection. Upon visual inspection of sequence polymorphism, candidates detected by our CLR method exhibited clear haplotype structures predicted under incomplete selective sweeps. Our results suggest that different methods capture different aspects of genetic information regarding incomplete sweeps and thus are partially complementary to each other.

KEYWORDS positive selection; selective sweep; composite likelihood; polymorphism

POSITIVE natural selection is one of the most fundamental driving forces for biological evolution. However, it is known that mutations conferring higher relative fitness to carriers, or beneficial mutations, do not occur frequently at a given gene or genomic region of interest in most natural populations of plants and animals. Even if a beneficial allele is currently under strong directional selection, its direct identification at the sequence level is not easy since the allele frequency change is likely to be too slow to follow over time in typical population genetic surveys unless the generation time is very short and a large amount of serially sampled sequences are available. Therefore, it is extremely difficult to directly follow the random occurrence of beneficial mutations and their spread under selective environments in nature. For this reason, the investigation depends heavily on detecting the signature of past episodes of positive selection, whether the beneficial mutation is already fixed in the population or still on the way to fixation (*i.e.*, ongoing selection for a mutation that occurred in the past but still segregating in the population), from the present-day patterns of within- and between-species genetic variation (reviewed in Nielsen 2005; Sabeti *et al.* 2006; Akey 2009; Stephan 2010). Such signatures of positive selection provide information for reconstructing evolutionary events that happened in the population's history. In addition, signals of positive selection imply functional importance of the loci and thus can be used to identify genetic variation that contributes to phenotypic diversity or annotate the genome functionally (Biswas and Akey 2006).

One of the basic methods for detecting positive selection is to search for the distinct pattern of within-species genetic variation left by a "selective sweep." A selective sweep occurs when a new advantageous mutation increases in frequency quickly in the population and results in a great reduction in variation, a temporary increase in linkage disequilibrium, and

doi: 10.1534/genetics.115.175380

Manuscript received February 5, 2015; accepted for publication April 23, 2015; published Early Online April 24, 2015.

Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175380/-/DC1.

[1] Corresponding author: Department of Life Science, Ewha Womans University, Ewhayeodae-gil 52, Seodaemun-gu, Seoul, Korea 120-750. E-mail: yuseob@ewha.ac.kr

a skew in allele frequency distribution in the nearby region of a recombining chromosome (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Fay and Wu 2000; Kim and Nielsen 2004). A selective sweep may be "complete" when the advantageous mutation goes to fixation and all local variation is removed except those that escaped the sweep by recombination. This type of selective sweep has drawn much attention and a number of statistical tests, mostly based on summary statistics such as Tajima's D, Fu and Li's D and F, and Fay and Wu's H test, were proposed to detect mainly complete positive selection from sequences sampled shortly after the fixation of a beneficial mutation (Tajima 1989; Fu and Li 1993; Fay and Wu 2000). More advanced statistical tests based on composite likelihood were also proposed (Kim and Stephan 2002; Meiklejohn *et al.* 2004; Nielsen *et al.* 2005).

Hudson *et al.* (1994) first observed evidence of an ongoing selective sweep—a subgroup of sampled sequences harboring very low variation due to linkage to the putative beneficial allele that reached an intermediate frequency—at the *Sod* locus in *Drosophila melanogaster*. However, as the availability of population genomic data was limited and discovering rare episodes of recent selective sweeps was considered very difficult in natural populations, capturing such "incomplete" or ongoing selective sweeps must have been considered even more difficult. Therefore, theoretical work mainly focused on inferring selective sweeps that were already completed in the past (Kaplan *et al.* 1989; Barton 1998; Fay and Wu 2000; Kim and Stephan 2002; Przeworski 2002). However, Sabeti *et al.* (2002), in one of the first large-scale population genomic surveys for detecting recent positive selection, showed that the human genome harbors a number of loci with clear signatures of incomplete selective sweeps. Since then, detecting this type of selective sweep soon became an important topic in both empirical and theoretical population genetics (Quesada *et al.* 2003; Meiklejohn *et al.* 2004; Sabeti *et al.* 2006; Saunders *et al.* 2006; Voight *et al.* 2006).

Sabeti *et al.* (2002) introduced a long-range haplotype test based on extended haplotype homozygosity (EHH) that quantifies the residual association between an allele at the core locus and its genetic background (*i.e.*, the linked haplotype at the time of the allele's mutational origin). Under neutrality, a haplotype associated with an allele at higher frequency extends to a shorter distance, thus yielding smaller EHH, since the allele is older (Toomajian *et al.* 2003). A significantly large EHH for a given allele frequency at the focal locus then suggests the hitchhiking effect driven by positive selection. If the ancestral *vs.* derived alleles of a polymorphic site can be distinguished, positive selection is expected to generate a much larger EHH for the derived allele than that for the ancestral allele. This is the rationale of the *iHS* statistic in Voight *et al.* (2006) that is now routinely used in population genomic studies. Recently, Ferrer-Admetlla *et al.* (2014) proposed a new statistic, nS_L, that is similar to *iHS* but is robust to recombination rate variation and exhibits improved power to detect sweeps.

The success and popularity of discovering incomplete sweeps may be attributable to unique haplotype structures that can be relatively easily and reliably captured by a rather simple test statistic such as *iHS*. If the local mutation rate fluctuates, it may create a random region of severely reduced variation that might be taken as a candidate for a complete selective sweep (Kim and Stephan 2002). With an incomplete sweep, the pattern of polymorphism in the haplotype block containing the ancestral allele of the focal locus reflects genetic variation that existed before the start of the selective sweep. Then, this haplotype block is effectively a negative control for the selective sweep that would alleviate the problem of local fluctuation in mutation rate. In the case of local adaptation, the inclusion of sequences from a neighboring deme, where positive selection did not take place, into analysis was shown to increase the statistical power of detecting positive selection (Innan and Kim 2008). The ancestral haplotype block in an incomplete sweep is expected to play a similar role in increasing statistical power to detect selection to that of the neighboring deme for complete sweeps.

The analysis of incomplete selective sweeps therefore provides a great opportunity for understanding positive natural selection in nature. However, as the current methods are not built on an explicit model of selection, information regarding the process of selection underlying the incomplete sweeps was limited. In this study, we obtain an approximate formula for sampling probabilities in a model of an incomplete selective sweep and then build a composite-likelihood-ratio (CLR) test for formal hypothesis testing and parameter estimation. Previously, Meiklejohn *et al.* (2004) proposed a CLR test for detecting an incomplete sweep by extending the sampling probabilities under complete selective sweeps of Kim and Stephan (2002) into cases where the final frequency of the beneficial allele in the population is less than one. However, in this approach the probability of sampling a neutral variant from the entire set of samples was obtained without explicitly specifying the polymorphic site causing an incomplete sweep or the joint configuration of polymorphism in the neutral and the putative selected loci. While a key parameter in their composite-likelihood ratio is the final frequency, β, of a beneficial mutation in the population, the frequency spectrum of the total data contains only a limited amount of information, yielding a very broad peak of the composite-likelihood ratio over the parameter space. Therefore, the joint estimation of β and the location/strength of selection was not accurate and the statistical power to detect selection was much lower compared to that of the *iHS* test. To overcome this difficulty, this study uses an approach to take each single-nucleotide polymorphism (SNP) in data as a putative locus under selection, essentially identical to the *iHS* method above. Namely, the derived alleles at all SNPs are tested to find whether they increased to the current frequencies by strong directional selection, by jointly analyzing the pattern of linked polymorphism surrounding the derived allele of each SNP and that surrounding the ancestral allele. This test is aimed at detecting selection in large-scale population genomic data generated by next-generation sequencing (NGS) methods, which inevitably contain occasional low-quality or missing base calls. The composite-likelihood approach can be straightforwardly applied to such data with missing information. By applying this method to simulated data and a population genomic data set in *D. melanogaster*, we demonstrate that this approach improves our ability to detect clear signatures of incomplete selective sweeps.

Materials and Methods

Sampling probability under an incomplete selective sweep

We aim to detect the signature of an incomplete selective sweep in which a beneficial allele originating from a single event of a point mutation (thus a hard selective sweep) reaches an intermediate frequency in a population. Consider multisite polymorphism observed in the alignment of n randomly sampled homologous chromosomes (Figure 1). It is assumed that neutral alleles segregate at these polymorphic sites, except one under selection (denoted the "S locus") with n_1 copies of the beneficial allele and n_2 ($= n - n_1$) copies of the ancestral allele. The strength of selection for the beneficial allele is given by $\alpha = 2Ns$, where N is the number of diploid individuals in the population and s is the selection coefficient [the relative fitness of the beneficial over the ancestral allele is $1 + s$, assuming codominance ($h = 0.5$)]. At a neutral site that is d nucleotides away from the S locus, let k_1 (k_2) be the count of the derived allele in the subsample of n_1 (n_2) chromosomes carrying the beneficial (ancestral) allele. If d is small enough to generate the hitchhiking effect of the beneficial allele, an increased or decreased frequency of the derived neutral allele due to hitchhiking is reflected by k_1/n_1, while its frequency before hitchhiking is estimated by k_2/n_2, assuming that the frequency of the neutral allele among chromosomes carrying the ancestral allele at the S locus does not change during the sweep (see *Appendix*). Therefore, the hypothesis of an incomplete sweep acting on the (putative) S locus predicts a very distinct joint probability distribution of k_1 and k_2, compared to an alternative (*i.e.*, neutral) hypothesis (Figure 2). Our goal is to build a parametric test based on this joint sampling probability, denoted by $\phi \equiv \phi(k_1, k_2, n_1, n_2, d)$, for detecting an incomplete selective sweep (*i.e.*, identifying the S locus in DNA sequence polymorphism).

We obtained two approximate solutions to such a joint sampling probability, ϕ_{S1} and ϕ_{S2}, by modifying the equivalent solution for complete sweeps in Nielsen *et al.* (2005) and that in Etheridge *et al.* (2006), respectively (*Appendix*). The corresponding probability under the null hypothesis (no selection), ϕ_N, can also be obtained. The primary parameter that determines ϕ_{S1} and ϕ_{S2} is $r/s = R/(2\alpha)$, where $r = r_n d$ (r_n = recombination rate per nucleotide per generation) is the recombination rate between the S locus and the neutral site and $R = 4Nr$. In comparison against simulated data generated by *msms* (Ewing and Hermisson 2010) under the model of an incomplete sweep, ϕ_{S2} approximates the sampling probability much better than ϕ_{S1} for small recombination rates (Supporting Information, Figure S1). However, ϕ_{S2} is not applicable for larger recombination rates ($r/s > 1/\sum_{i=1}^{n-1} 1/i$) (Etheridge *et al.* 2006).

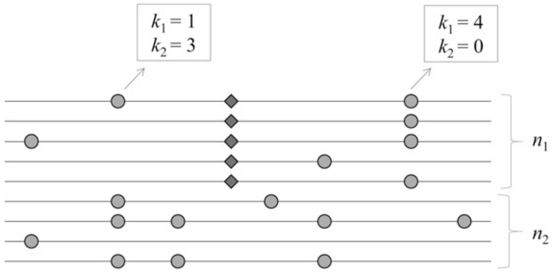


Figure 1 Pattern of DNA sequence polymorphism under an incomplete sweep. Lines indicate individual sequences and circles indicate derived alleles at neutral sites. Diamonds indicate new advantageous mutations spreading in the population.

CLR test

Let x be the position of the putative S locus, which is assumed to be one of the polymorphic sites in the sequence alignment and thus partition the data into subsamples of n_1 and n_2 chromosomes carrying the derived and ancestral alleles at the locus, respectively. n_1 is also denoted as $n_1(x)$ to emphasize that the position x uniquely determines the derived allele frequency of the putative S locus. This partition by the S locus also determines the counts of the derived neutral allele at nucleotide site i in the two subsamples, $k_1^{(i)}$ and $k_2^{(i)}$. Then, for a given x, a maximum-composite-likelihood estimate of the strength of selection,$\hat{\alpha}(x)$, is obtained as a value of α (if R per site is externally given) that maximizes the CLR,

$$\Lambda(x, \alpha) = \log \frac{L_{\mathrm{IS}}(x, \alpha|\mathrm{Data})}{L_{\mathrm{N}}(\mathrm{Data})}, \tag{1}$$

where

$$\begin{aligned} L_{\mathrm{IS}}(x, \alpha|\mathrm{Data}) &= P(\mathrm{Data}|x, \alpha) \\ &= \prod_{i \neq x} \phi_{\mathrm{S}.}\left(k_1^{(i)}, k_2^{(i)}; n_1, n_2, |i - x|\right) \end{aligned} \tag{2}$$

and

$$L_{\mathrm{N}}(\mathrm{Data}) = \prod_i \phi_{\mathrm{N}}\left(k_1^{(i)}, k_2^{(i)}; n_1, n_2\right) \tag{3}$$

are composite likelihoods under the hypotheses of incomplete selective sweep and neutrality, respectively. In the following, unless stated otherwise we use only ϕ_{S1} for Equation 2 despite its error for small r/s. The impact of this error on the performance of our likelihood test is addressed below. Unless stated otherwise, multiplication above is done across all sites in the data, including monomorphic sites ($k_1^{(i)} = k_2^{(i)} = 0$). It is also possible to multiply probabilities over polymorphic sites only (analogous to L_2 and L_4 in Kim and Nielsen 2004), which leads to a composite-likelihood test for detecting selection based on the joint allele frequency spectrum only but not on the patterning of polymorphic sites along the sequence.

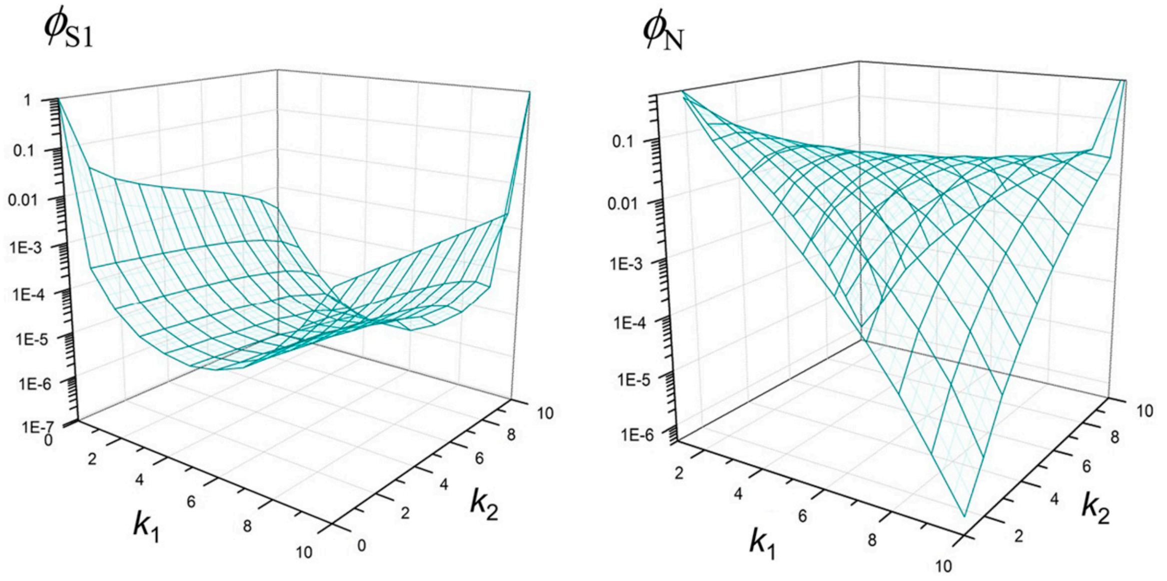


Figure 2 Joint sampling probability as a function of k_1 and k_2 (= 0, . . . , 10) for $n_1 = n_2 = 10$ and $N = 10^5$, under the incomplete sweep model with r/s = 0.04 (left) and the neutral model (right).

It is straightforward to calculate the CLR given in Equation 1 in the presence of missing values in sequence data, for example due to low-quality base calls that are common in NGS data sets. If missing base calls are made at a site on chromosomes in the sample, the sampling probability for this site is calculated after n_1 and/or n_2 are reduced accordingly. If base calls are missing at the core SNP (the putative S locus), entire chromosomes carrying the missing base calls are excluded from the calculation of composite likelihoods.

Next, the maximum-composite-likelihood estimate of the locus under selection is obtained by calculating $\hat{\alpha}(x)$ for all polymorphic sites in a given chromosomal region and then identifying the site (at position $\hat{x}$) that maximizes $\Lambda_x := \Lambda(x, \hat{\alpha}(x))$. This procedure also implies that a test statistic for hypothesis testing would be given by

$$T_0 = \max_x \Lambda_x \tag{4}$$

and we may reject the null hypothesis (neutrality) if T_0 is larger than a certain cutoff value. The null distribution of T_0 is determined by applying the above calculation to polymorphic sites in a large number of data sets simulated under the neutral model. Imposing fixed polymorphic sites (-s option in ms) or fixed scaled mutation rate (-t option) but conditioning on the similar number of polymorphic sites in simulated data led to almost identical distributions (data not shown). However, it was observed that the maximum CLR for a given focal site (Λ_x) is negatively correlated with $n_1(x)$, most likely because a derived allele with smaller allele frequency originated more recently and is thus associated with a longer extended haplotype. If not corrected, this will bias the estimated locus of selection to be a polymorphic site with a lower frequency of the derived (putatively beneficial) allele. A solution to this problem would be to transform Λ_x to remove its correlation with $n_1(x)$. We tried and evaluated various forms of the normalized test statistic. The following procedure yielded the most optimal performance of parameter estimation (see below). Let $m(f)$ and $Q(f, \varepsilon)$ be the mode and the $1 - \varepsilon$ quantile of the distribution of Λ_x obtained from polymorphic sites whose derived allele frequency is f in simulated neutral data sets. Then, we define a new statistic

$$T_1 = \max_x T_{1x} = \max_x \frac{\Lambda_x - m(n_1(x))}{Q(n_1(x), \varepsilon) - m(n_1(x))}. \tag{5}$$

Then, the estimated location of the S locus, $\hat{x}$, is the value of x that achieves the maximum in the above formula. This also leads to the final estimate of the strength of selection, $\hat{\alpha} \equiv \hat{\alpha}(x)$, for a given set of sequences. For a given ε, the null distribution of T_1 is obtained by applying the above procedure to a large number of sequence samples, with an equivalent number of polymorphic sites and a scaled recombination rate, that are generated by neutral simulation. Unless stated otherwise, we reject the null hypothesis of neutrality (no selection) with significance level $P = 0.001$, which resulted in an optimal range of statistical powers with varying parameter values chosen below.

Analysis of D. melanogaster population genomic data

We used 22 primary core genomes in the Rwanda (RG) sample of *D. melanogaster* described in Pool *et al.* (2012), available for download from the DPGP2 project (http://www.dpgp.org/). As the violation of the random sample from unrelated individuals may generate spurious occurrence of long-range haplotype homozygosity, we removed identical-by-descent (IBD) tracts detected by Pool *et al.* (2012): in the sample of 22 sequences, if any pair of chromosome segments are IBD, we treated one of them as a missing observation, replacing it by a sequence of "N" characters. Then, we extracted a phased table of polymorphic sites with their physical locations. Next, the ancestral and derived alleles were inferred using the syntenic assembly of *D. melanogaster* and *D. simulans* (available at www.dpgp.org) and designating the allele observed in *simulans* as ancestral or the table of ancestral allele probability for polymorphic sites calculated for DPGP1 RAL sequences (Chan *et al.* 2012). This procedure could assign the ancestral/derived states for ~85% of polymorphic sites obtained above. The remaining polymorphic sites were not included in input files. We also excluded from analysis the telomeric and centromeric regions of each chromosome arm with low recombination rates: from the midpoint of a chromosome arm we moved toward the telomere and toward the centromere until the points over which the mean genetic distance per megabase first becomes <1 cM, using the best-fitting equations for crossing-over rates on 100-kb windows obtained by Comeron *et al.* (2012).

Composite likelihood is calculated by taking a SNP as the putative *S* locus ("core SNP"): thus the sample is partitioned into n_1 and n_2 sequences as described above. The sample frequency of the focal derived allele is therefore $f = n_1/(n_1 + n_2)$. Note that, as this SNP may contain missing values (N in data) and the corresponding chromosomes are excluded from calculating the composite likelihood, $n_1 + n_2$ can be <22. For computational convenience, we assumed scaled recombination rate $4Nr_\mathrm{n} = 0.012$ per site in the calculation of likelihood for all chromosome arms. As sampling probability under selection is primarily a function of $r/s = R/(2\alpha)$, but only slightly modified by α alone, a deviation of actual recombination rate from the above assumption would lead to a corresponding error in the estimate of α, without affecting the location and value of the maximum-composite-likelihood ratio. Local fluctuation in the scaled mutation rate, θ, was also ignored: we estimated mean θ for each chromosome arm and used it in the calculation of likelihoods for any region within the chromosome arm. Incorrect assumptions of θ were shown to affect minimally the performance of our test (see below).

Joint sampling probabilities were obtained using the approximation proposed in Nielsen *et al.* (2005), *i.e.*, ϕ_{S1}, assuming that the ancestral pattern of polymorphism at the time of the beneficial mutation follows either standard neutral equilibrium (test option A) or the currently observed genome-wide empirical frequency spectrum (test option B) (*Appendix*). The significance of the CLR, maximized with respect to α and then normalized for the derived allele frequency, is assessed as described for T_1 above, however, using the site-wise null distribution of CLR obtained from individual polymorphic sites (5×10^5 SNPs) generated by *msms* under neutrality, with parameters adjusted to match sample size, mean recombination rate, and the mean density of polymorphic sites to those of *Drosophila* genome data. Namely, multiple-test correction, as implemented above by the null distribution of the local maximum of test statistic (T_1) in a window of defined sequence length, is not performed here. Therefore, a *P*-value determined this way cannot be compared to that used for analyzing simulated incomplete sweeps above. We consider sites that yield large normalized CLR, corresponding to $P < 0.001$, as candidate loci under selection. This level of significance is rather arbitrary. However, rather questionable candidates of incomplete sweeps (with unclear haplotype structure upon visual inspection; see *Results* below) are already detected at this level and, therefore, a less stringent level will likely increase the number of such loci.

Haplotype homozygosity tests

We applied two haplotype homozygosity tests, *iHS* (Voight *et al.* 2006) and nS_L (Ferrer-Admetlla *et al.* 2014), to detect incomplete sweep in simulated data as well as *D. melanogaster* data. For the analysis of simulated data, unstandardized *iHS* ($\log[iHH_\mathrm{A}/iHH_\mathrm{D}]$) was calculated for individual polymorphic sites according to Voight *et al.* (2006), using the rehh R package (Gautier and Vitalis 2012) (http://cran.r-project.org/web/packages/rehh/index.html), and unstandardized nS_L was calculated through the program provided by Ferrer-Admetlla *et al.* (2014) at http://cteg.berkeley.edu/~nielsen/. Using the same set of simulated neutral samples as used above for CLR analysis, the $1 - \varepsilon$ quantile and mode of the distribution of the unstandardized *iHS* were obtained for each derived allele frequency and these values were used to define standardized *iHS* by applying the procedure of obtaining T_1 by Equation 5. The test statistic for detecting an incomplete selective sweep in a replicate of a 100-kb sequence sample is therefore the most negative standardized *iHS* among sites and the procedure of obtaining the null distribution and assessing the significance of this statistic is identical to that of T_1 by the CLR method. The same normalization procedure was applied for the nS_L statistic. We also tried the standardization procedure based on the assumption of normal distribution described in Voight *et al.* (2006) for both statistics and discovered that our standardization procedure leads to slightly increased statistical power (data not shown).

For the genomic scan of *D. melanogaster* data below, we first obtained standardized *iHS* and associated *P*-values for individual polymorphic sites in the data according to the procedure described by Voight *et al.* (2006) performed by the rehh package. In the calculation of iHH_A and iHH_D by this package haplotype homozygosity for sequences does not extend from the core SNP if missing base calls are encountered in a subset of the sequences. Namely, missing bases (N) are treated as an allele distinct from A, C, G, or T. We found that this frequently generates very small iHH_A and thus erroneously very negative *iHS* (*i.e.*, false detection of selection). To correct this problem, we wrote our own code that calculates iHH_A and iHH_D while skipping positions of missing bases in extending haplotype homozygosity. We also used the full data including all polymorphic sites rather than the input data used above for the CLR test, in which ~15% of polymorphic sites were excluded as their ancestral/derived alleles could not be determined. Excluding these sites caused frequent false positives since the excluded sites are often clustered to make the region to falsely appear monomorphic and thus inflate iHH_D.

These corrections led to detection of clearer signatures of incomplete sweeps (upon visual inspection of haplotype structures). For the nS_L statistic, since we find it less sensitive to missing data than *iHS*, we used the same input data as used for the CLR test and then performed standardization according to Ferrer-Admetlla *et al.* (2014).

Simulated data under different demographic assumptions

To explore the robustness of the CLR test to demographic assumptions, we generated neutral data sets, using *msms* (Ewing and Hermisson 2010) under three different scenarios: population bottleneck, exponential population growth, and population subdivision. All data sets were generated with equal sequence length (100 kb) and number of polymorphic sites (3000). For the bottleneck model, we simulated a population bottleneck lasting from 0.4*N* to 0.2*N* generations in the past with different severities $c = 0.4$, 0.2, and 0.1 ($c = N_b/N$, where N_b is the population size during the bottleneck). In the case of the exponential growth model, populations start growing exponentially from a population size of 0.4*N*, with three different growth rates $g = 10$, 100, and 500. For the population subdivision model, we simulated a two-island model with symmetric, constant migration rates $M = 0.1$, 1, 10 and then drew all sequences for each sample from one island. We also varied the recombination rate [*R* (per 10 kb) = 4000, 6000, 8000, 10,000, 12,000] in each model to study the effect of altered linkage disequilibrium on the null distribution. For each parameter set, at least 1000 replicates with a sample size of 20 chromosomes were obtained.

Codes and scripts

All source codes developed here for analyzing simulated and actual data are available upon request. Command line scripts for simulations performed above using *ms* and *msms* are provided in File S1.

Results

Statistical power of the composite-likelihood test

To evaluate the performance of the composite-likelihood method described above, we applied it to simulated data sets generated by *msms* (Ewing and Hermisson 2010) under the model of incomplete selective sweeps. In simulation, the beneficial allele of the *S* locus, located in the middle of the 100-kb sequence, reaches frequency $\beta = 0.5$ in the population and a sample of 20 sequences is generated. Then, test statistic T_1 in Equation 5 was determined after the maximum CLRs were calculated over all SNPs in the sample with derived allele frequency ≥ 3 and ≤ 17. The null hypothesis of neutrality was rejected if $T_1 >$ the 99.9th percentile of the null distribution, which was obtained from data sets simulated under the model of neutral equilibrium with the same sequence length and recombination rate. This cutoff value ($P = 0.001$) of T_1 is a function of ε. When various values of ε (0.0006, 0.001, 0.0016, 0.002, and 0.003) were tried, the statistical power fluctuated moderately (~5%) while $\varepsilon = 0.002$ resulted in the best performance in parameter estimation (the largest proportion of replicates in which the correct site, at position 50 kb, yielded the largest T_1). We thus use $\varepsilon = 0.002$ in the following analyses.

The statistical power of the test increased as the final frequency of the beneficial mutation, β, in the population increased: with $R = 2000$, $\alpha = 2000$, and β increasing from 0.3 to 0.7 by 0.1, the statistical powers were 0.45, 0.71, 0.87, 0.95, and 0.98, respectively. This test performed better with larger β presumably because as a larger proportion of individuals (thus sequences in a sample) are affected by selection, the pattern of polymorphism becomes more distinctive from the neutrality, and also because the φ_S component of sampling probability was obtained from the solution obtained for a complete selective sweep. For a fixed value of β, the statistical power increased with increasing strength of selection, as expected (Table 1).

This performance of the composite-likelihood test was compared to that of long-range haplotype methods that use *iHS* and nS_L statistics (Voight *et al.* 2006; Ferrer-Admetlla *et al.* 2014). Instead of using the normal distribution-based standardization of *iHS* and nS_L, we applied the normalization procedures that were used to obtain T_1 above (see *Materials and Methods*), which made it possible to directly compare the performance of the CLR, *iHS*, and nS_L methods. In all parameter sets tested, the statistical power of our composite-likelihood method is higher than that of *iHS* but only slightly better than that of the nS_L method [Table 1; note that results here were obtained assuming that the correct scaled recombination rate of the sequence is available (see below)]. Interestingly, there are a relatively large number of simulated incomplete sweeps detected by either CLR or nS_L only, particularly with weaker strength of selection (Figure 3), suggesting that the CLR and nS_L methods capture different aspects of data as signatures of incomplete sweeps and thus are largely complementary to each other.

Effect of recombination rate and linkage disequilibrium

The above result is based on the null distribution of the test statistic obtained from neutral simulations that used recombination rates identical to those used in the simulation of incomplete sweeps. However, in practice, the correct rate of recombination, scaled or unscaled, for a given genomic region may not be available. This turns out to be a serious problem for our CLR method, as we found that the null distribution of the likelihood ratio is highly sensitive to the scaled recombination rate (Figure 4). It appears that, with decreasing recombination rate, linkage disequilibrium (LD) between adjacent polymorphic sites increases and this inflates the likelihood of an incomplete sweep (L_{IS}) relative to that of neutral evolution (L_N). Therefore, one approach to control the false-positive rate of detection might be to adjust the recombination rate during neutral simulation until the average LD among sites matches that of data under examination (see below for the case of demographic complications). As an incomplete selective sweep generates a high level of LD around the *S* locus (Stephan *et al.*

Table 1 Accuracy of parameter estimates using the composite-likelihood, *iHS*, and nS_L methods

Parameters		Composite likelihood				*iHS*			nS_L		
R^a	α	Statistical power (%)^b	$\hat{\alpha}$ mean ± SD	$\hat{x}$ deviation mean (median)^c	Exact site detected (%)^d	Statistical power (%)^b	$\hat{x}$ deviation mean (median)^c	Exact site detected (%)^d	Statistical power (%)^b	$\hat{x}$ deviation mean (median)^c	Exact site detected (%)^d
		Using both polymorphic and monomorphic sites									
2000	1000	57 (17)	1160 ± 660	3.68 (1.73)	9.23 [10.9]	24	5.67 (2.91)	4.92 [5.29]	41	4.72 (2.23)	11.1 [13.1]
	2000	86 (36)	2070 ± 920	4.11 (2.43)	10.6 [12.7]	63	6.18 (4.05)	3.54 [3.22]	63	5.01 (2.95)	9.4 [11.4]
	4000	97 (30)	3850 ± 1780	6.03 (3.92)	9.07 [10.7]	89	8.58 (6.09)	2.61 [2.23]	76	7.10 (4.45)	6.9 [7.0]
4000	1000	35 (16)	1290 ± 840	2.92 (0.95)	15.6 [20.4]	9.6	5.71 (1.80)	7.37 [9.00]	42	3.36 (1.21)	18.6 [24.5]
	2000	75 (40)	2290 ± 1000	2.51 (1.39)	15.9 [19.2]	41	3.98 (2.30)	5.46 [6.58]	73	2.77 (1.62)	16.3 [17.1]
	4000	94 (48)	4030 ± 1610	3.56 (2.20)	13.8 [16.7]	78	5.12 (3.45)	3.90 [4.29]	87	3.66 (2.35)	12.6 [12.4]
		Using only polymorphic sites									
4000	1000	4.0	860 ± 950	3.62	13.3						
	2000	36	1630 ± 1010	3.21	13.3						
	4000	82	3290 ± 1870	4.91	10.3						

[a] Scaled recombination rate ($4Nr$) across the 100-k-long simulated sequence.

[b] Percentages of simulated samples that yield $P < 0.001$. In parentheses: using an adjusted recombination rate in neutral simulations to yield mean LD identical to that of data to be analyzed.

[c] Mean (median) deviation in kilobases of the estimated location of selection from the true location (x = 50,000): $|\hat{x} - 50,000| \times 10^{-3}$.

[d] Percentages of simulated samples for which $\hat{x}$ = 50,000. Those of samples in which the frequency of the beneficial allele is exactly 10 are given in brackets.

2006), to generate samples with an equivalent level of LD [measured by the average ρ^2 over all pairs of sites, where ρ is the normalized LD as measured by correlation of allele frequencies between two loci (Hill and Robertson 1968)] the recombination rate needs to be greatly reduced in the neutral simulation. When we obtained the null distribution of T_1 from such low-recombination simulation, the statistical power of our CLR method decreased dramatically (Table 1). In contrast, the null distribution of nS_L was affected minimally by recombination rate variation (data not shown), as it was originally proposed to cope with uncertainty in recombination rates (Ferrer-Admetlla *et al.* 2014).

Inferring the strength of selection and the position of the S locus

Because sequences are randomly sampled from a population, the copy number of the beneficial allele, $n_B = n_1(50{,}000)$, in a sample is variable (binomial): with $\beta = 0.5$, $n_B < 3$ or >17 in <0.5% of replicates, which makes it impossible to detect the true locus under selection. In the other replicates, the exact locus is detected if the maximum T_1 is obtained at the correct site (*i.e.*, $\hat{x} = 50{,}000$). Compiling results from all replicates (regardless of whether the correct site is inferred or not), we find that the estimate of the strength of selection $\hat{\alpha}$ is unbiased, although the variance of the estimate is large (Table 1). More than half of replicates yielded $\hat{x}$ within ~1–3 kb from the target of selection. The proportion of replicates in which the exact site is inferred ranges from 9 to 16%, more accurate estimates occurring with higher recombination rates. If the sample frequency of the beneficial allele matches the population frequency (0.5), this proportion significantly increases (Table 1).

In the *iHS* and nS_L methods the estimated location of the *S* locus, $\hat{x}$, is given as the polymorphic site from which the most negative normalized statistic is obtained. Applied to the same sets of simulated data, $\hat{x}$ by *iHS* was less accurate than by either CLR or nS_L (Table 1). The exact position of the *S* locus was correctly inferred about three times more often by CLR than by *iHS* but roughly as often as by nS_L. CLR also yielded the smallest mean deviation of $\hat{x}$ from the true location. Overall, the accuracies of estimates are similar between the CLR and nS_L methods. Surprisingly, however, the three methods are weakly correlated with respect to estimating the exact location of selection (Figure 3): for example, applied to 10,000 replicates of simulation with $\alpha = 4000$, the CLR and nS_L methods detected the correct site under selection in 1390 and 1255 replicates, respectively. However, in only 252 replicates the correct site was detected by both methods. Again, this result suggests that the CLR and *iHS*/nS_L methods capture slightly different information in multisite polymorphism to detect incomplete sweeps and estimate the position of the putative *S* locus. When we define a new estimate as the average over those by the CLR and nS_L methods, its mean deviation from the correct site in kilobases [*i.e.*, $\left|(\hat{x}_{\mathrm{CLR}} + \hat{x}_{nS_L})/2 - 50,000\right| \times 10^{-3}$] is 2.74, 2.42, and 3.25 for α = 1000, 2000, and 4000, respectively (with R = 4000), which is smaller than the deviation obtained by an individual method (Table 1). Therefore, small improvements in the accuracy of position estimates are made by combining the two methods.

Modification of composite likelihoods

So far, sampling probability based on approximation by Nielsen *et al.* (2005), ϕ_{S1}, was used for obtaining composite likelihoods. For small recombination rates ($r/s < 1/\sum_{i=1}^{n-1} 1/i$), we may replace ϕ_{S1} by more accurate approximation, ϕ_{S2} based on Etheridge *et al.* (2006). This, however, did not lead to a significant change in the profile of the CLR (Figure S2). We also examined the effect of not including monomorphic sites in the data. When the CLR is calculated by multiplying joint sampling probabilities over only polymorphic sites in the data, it leads to lower statistical power to detect selection

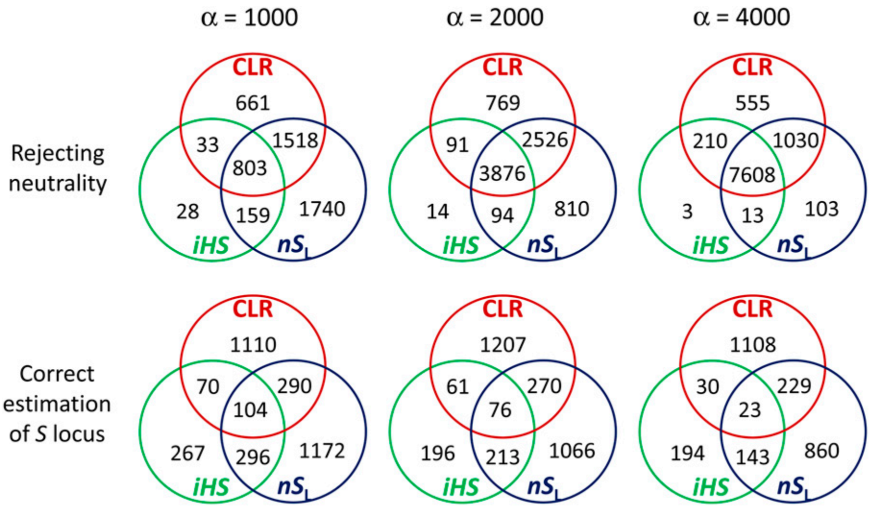


Figure 3 Numbers of simulation replicates (of 10,000) from which incomplete selection sweeps were detected (rejection of null hypothesis at $P < 0.001$) and the correct site under selection was inferred, individually or jointly by the CLR, *iHS*, and nS_L methods. $R = 4000$.

and larger errors in estimating the strength and position of selection than when multiplication was done over all sites (Table 1). This result suggests that not only the (joint) frequency spectrum of polymorphism but also the spatial distribution or density of polymorphic sites contains information regarding incomplete selective sweeps.

Effect of complex demography

Next, to evaluate the robustness of the CLR method to complex demography and population structure, we examined how the null distribution of the test statistic (T_1) changes if it is obtained from data sets simulated under the models of population bottleneck, expansion, and subdivision (see *Materials and Methods*). In each model parameters were chosen to produce a significant deviation of the frequency spectrum from that under neutral equilibrium. The number of polymorphic sites (3000) per sample remained constant for varying models and parameters. First, with a population bottleneck that lasted from 0.4N to 0.2N generations ago, decreasing the size of the bottlenecked population ($c = N_B/N$ decreasing from 0.2, 0.1, to 0.05) dramatically shifted the distribution of the CLR upward (Figure S3A). This shift appears to be explained by a reduction in scaled (population-level) recombination rate due to the bottleneck, which leads to increased LD: when the recombination rate was increased to reduce LD (quantified by mean pairwise ρ^2), the distribution of the CLR shifted back downward (Figure S4A). With matching LD, distributions obtained under the bottleneck ($4Nr_n = 0.1$; mean $\rho^2 = 0.0543$) and under the standard neutral model (a constant-sized panmictic population; $4Nr_n = 0.04$; mean $\rho^2 = 0.0543$) are very similar (Figure S4A). However, the right tail of the distribution is still slightly larger than that of neutral equilibrium.

Similarly, the null distribution of the CLR shifts upward due to rapid exponential growth of population size ($g > 100$) in the expansion model and limited migration ($M < 1$) in the subdivision model (Figure S3, B and C). Again, by increasing the recombination rate in the simulation, thus reducing the average level of LD among sites, these distributions are shifted downward. Similar distributions of the CLR (right tails) are obtained from simulations under the standard and complex demography if the levels of LD match (Figure S4). These results suggest that, in the analysis of a genomic region for which underlying population demography and/or correct recombination rate are not known, the false-positive rate of detecting incomplete sweeps by CLR can be greatly reduced, if not completely, by generating samples with matching LD by standard neutral simulation.

Results above were obtained by calculating the likelihood of incomplete sweeps, assuming the standard neutrality at the time of beneficial mutation [$f_0(p) = \theta/p$; test option A]. We can replace $f_0(p)$ with the empirical distribution of the derived allele frequency observed in the simulations of these demographic models (test option B). The latter option is essentially the approach by Nielsen *et al.* (2005) to minimize the compounding effect of complex demography in detecting the signature of selection. However, it had little effect in correcting the null distribution and did not prevent the inflation of the CLR with increasing LD between segregating sites (Figure S3).

Application to D. melanogaster genomic data

The composite-likelihood method described above was applied to population genomic data of *D. melanogaster* to detect incomplete sweeps. We used 22 haploid genome sequences from Rwanda (the RG sample) described in Pool *et al.* (2012). As the species' ancestral range is known to lie within southern and eastern Africa, the RG sample is likely to satisfy the assumption of equilibrium demography (constant-sized random-mating population before the start of the sweep in our model) better than any other available genomic data sets in *D. melanogaster.* However, when we examined the genome-wide distribution of derived allele frequency, a slight but clear deviation (excess of rare alleles) from the standard neutrality was observed (Figure S5). This is likely due to nonequilibrium demography (mild population bottleneck and recent population growth) that may have affected the RG sample (Pool *et al.* 2012) but might also be due to errors in base calling and ancestral/derived state inference.

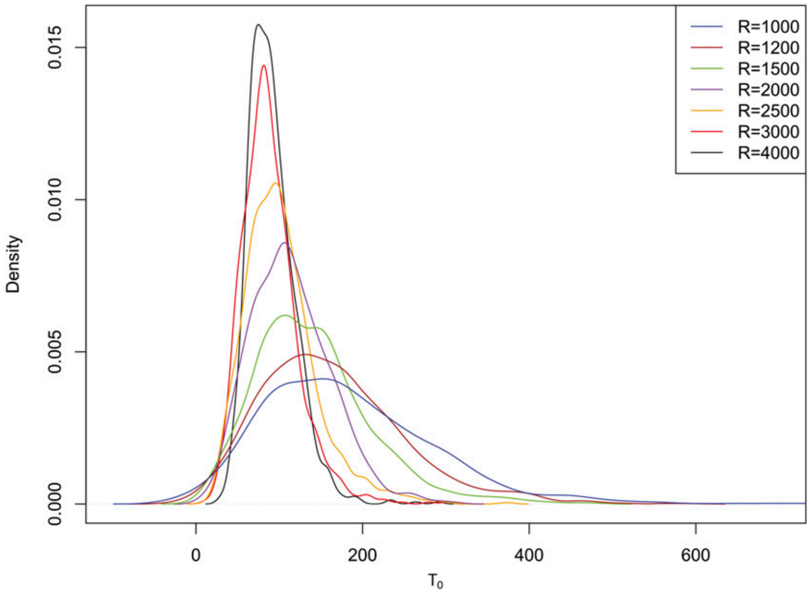


Figure 4 Distribution of maximum CLR, $T_0 = \max_{x \in S[10]} \log(L_{IS}/L_N)$, where the maximum was obtained over the set of polymorphic sites with n_1 = 10 (*S*[10] for each replicate), for samples generated under standard neutral simulation with varying recombination rate.

A genome scan was conducted by sequentially taking all polymorphic sites in the data with derived allele frequencies satisfying $0.35 < f < 0.8$ as core SNPs and calculating composite likelihoods. We observed clear clustering of SNPs yielding a large CLR (Figure 5 for chromosome arm 2R), corresponding to $P < 0.001$ (see *Materials and Methods*), scattered over the five major chromosome arms. We consider each cluster as a footprint of a single episode of an incomplete selective sweep. (Other scattered and isolated sites that yield $P < 0.001$ but do not form clusters were not considered.) A SNP with the largest CLR within a cluster (the "peak") is therefore a candidate position of ongoing selection. There are 42 clusters in total, using test option A, and we identified an annotated gene in FlyBase (version FB2014_03) containing or closest to the peak in each cluster (Table 2). Test options A and B generated very similar profiles of the CLR along the chromosome (Figure S6) and thus led to the detection of almost identical sets of candidate loci in each chromosome arm. When clusters are ranked according to T_1 within each chromosome arm, ranks by options A and B are strongly correlated (Table 2). Upon visual inspection of aligned and sorted sequences, we observed clear segregating patterns of SNPs indicative of incomplete sweeps—far fewer polymorphisms and high linkage disequilibrium among chromosomes containing the derived allele compared to those containing the ancestral allele at the core SNP—at the majority of these candidate loci (Figure 5 and Figure S7).

The calculations above were performed using a uniform value of scaled mutation rate, θ, for each chromosome arm. To examine whether local variation in θ has an effect on the accuracy of inference, we performed the CLR test with the local value of θ calculated from a 10-kb window surrounding each core SNP. This procedure yielded almost the same profile of composite likelihood along the chromosome and the same list of selection candidates (data not shown), presumably because the ratio of composite likelihoods depends weakly on θ: change in θ appears to affect L_{IS} and L_N in Equation 1 to a similar degree.

Patterns similar to the outcome of incomplete selective sweeps may arise by a complete selective sweep: at an appropriate recombination distance from the position of the beneficial mutation that reached fixation, low variation and high frequency of derived alleles would be observed among chromosomes whose linkages to beneficial mutation were not broken by recombination. However, a normal level of variation will be observed among chromosomes that recombined away from the beneficial mutation. We therefore checked whether our candidate regions of incomplete selective sweeps overlap with those of complete selective sweeps in the RG sample detected by Pool *et al.* (2012) (343 regions listed in their table S13). Seventeen of our 42 clusters overlap with the candidate regions of complete sweeps (Table 2).

Next, we calculated *iHS* and nS_L statistics for the same data set for which CLR was obtained above. Even though corrections were made to address the complexity of data (missing base calls and incomplete inference of ancestral/derived alleles; see *Materials and Methods*), many sites yielding very negative *iHS* appear to be false positives because clear haplotype structures predicted under incomplete sweeps are not observed at those loci (Figure S8). On the other hand, sites yielding very negative nS_L are associated with a much clearer haplotype pattern. However, there are still cases of very unclear haplotype patterns detected by nS_L (Figure S8). We could identify clusters of negative *iHS* and those of negative nS_L, similar to clusters of large CLR above. However, the overall pattern of clustering for negative *iHS* or nS_L is not clear, whereas very distinct clusters of large CLR were observed (Figure 5). Many sites generated large negative *iHS* or nS_L by themselves without belonging to any cluster and we did not consider them as candidate loci under selection. We found that these isolated occurrences of large negative iHS/nS_L and other sites with large negative iHS/nS_L but without clear

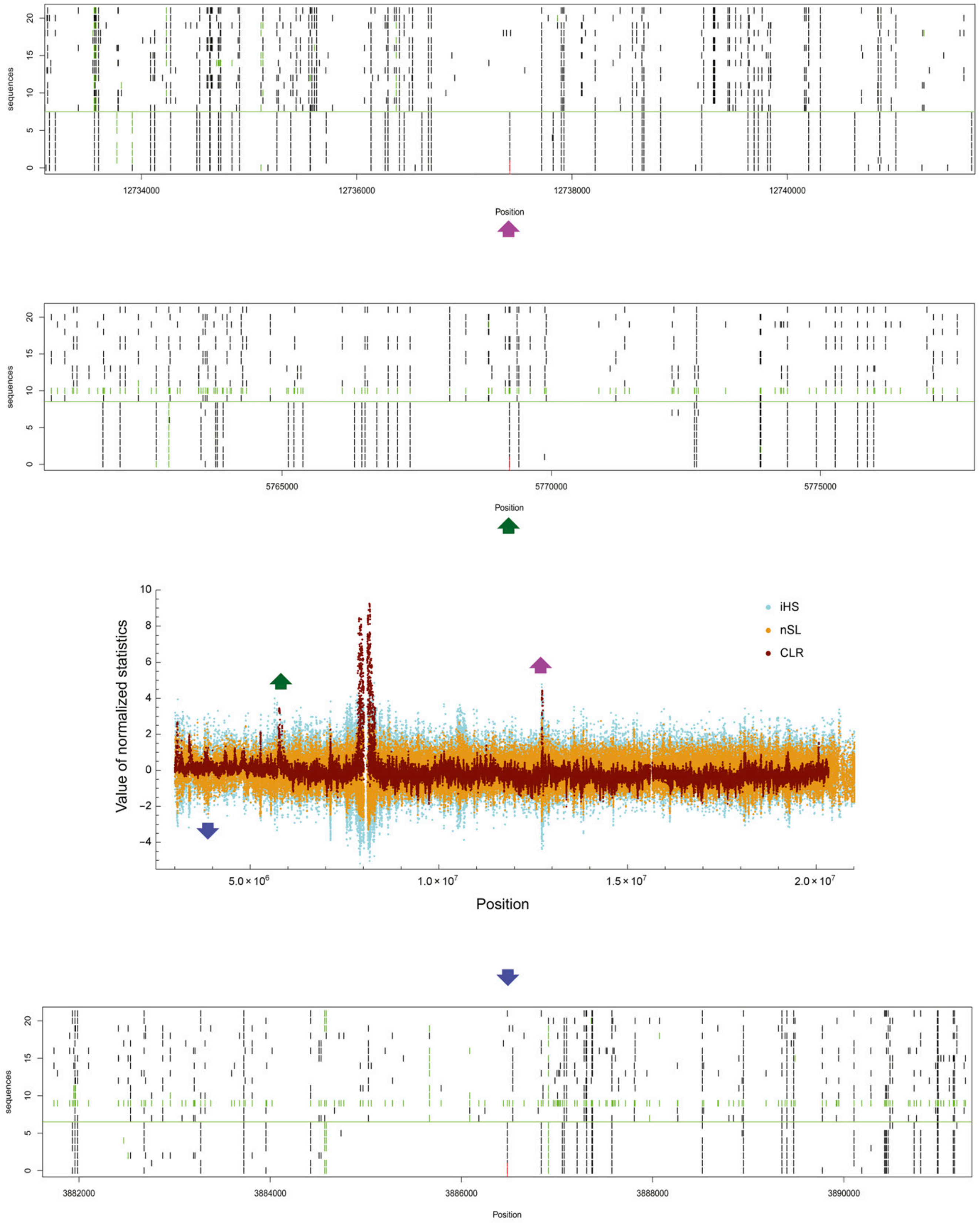


Figure 5 Plots of normalized maximum-composite-likelihood ratio (T_{1x}; dark red) and standardized *iHS* (light blue) and nS_L (orange) along chromosome 2R. Haplotype structures around three putative positions under selection (positions 5,769,223 and 12,737,423 detected by the CLR method and position 3,886,479 detected by the nS_L method) are shown. Aligned sequences were sorted by the position of the putative site, with those carrying

Table 2 List of putative loci under incomplete selective sweeps in *D. melanogaster* Rwanda population detected by the CLR method

Chr. arm	Cluster start to end[a]	Max T_1[e]	Site of max T_1[f]	$\hat{\alpha}$	Sample DAF[g]	Rank by option A (B)[h]	*iHS*[i]	nS_L	Closest gene
2L	1,517,366–1,529,788[b]	1.71	1,527,302	8,000	13/19	11 (12)	−0.662	−0.42	halo
	5,803,333–5,815,486[b,c,d]	2.04	5,805,001	8,000	9/22	5 (3)	−1.87	−1.44	CG11034
	6,650,567–6,657,389[b,d]	1.92	6,652,011	8,000	17/22	7 (13)	−0.577	−0.46	Tango1
	7,409,201–7,413,132	1.81	7,409,825	2,828	11/22	9 (6)	−0.577	−1.96	CG5181
	14,094,223–14,102,937[b]	1.72	14,100,158	8,000	11/21	10 (9)	−1.81	−2.29	nAChRalpha5
	16,001,993–16,020,004[c]	2.05	16,005,369	16,000	8/21	4 (5)	−2.52	−1.88	Beat-Ic
	17,221,968–17,339,080[c,d]	6.66	17,271,945	64,000	8/21	1 (1)	−2.23	−1.58	CG6380
	17,602,170–17,624,875[c,d]	2.28	17,616,351	11,313	11/21	3 (4)	−1.42	−1.69	Sytalpha
	18,446,487–18,454,519	1.55	18,453,145	11,313	8/20	12 (11)	−0.892	−1.30	bsf
	18,993,669–18,999,601	1.93	18,996,657	11,313	13/20	6 (10)	−1.23	−0.97	CG10650
	19,480,924–19,546,720	1.91	19,493,563	22,627	8/21	8 (8)	−1.32	−0.94	swm
	19,729,511–19,790,297[c,d]	3.13	19,756,197	22,627	9/21	2 (2)	−2.14	−1.08	CG10631
2R	3,058,650–3,079,356[d]	2.67	3,073,701	45,255	8/20	4 (5)	−2.88	−2.06	didum
	5,269,193–5,278,463	2.12	5,271,741	11,313	11/20	6 (6)	−2.38	−2.14	CG13954
	5,756,581–5,788,311	3.45	5,769,223	32,000	9/21	3 (3)	−0.44	−0.95	Sec24AB
	7,126,724–7,131,721[c,d]	2.27	7,127,281	5,656	11/22	5 (4)	−3.87	−2.40	CG13215
	7,882,689–8,178,854[c,d]	9.24	8,157,979	11,313	9/22	1 (1)	−3.88	−2.08	otk
	12,727,369–12,759,779[b,c,d]	4.43	12,737,423	22,627	8/22	2 (2)	−2.40	−1.81	IntS8
	20,061,659–20,075,219[b]	1.53	20,073,016	5,656	17/22	7 ()	−1.11	−0.48	Nop60B
3L	3,173,086–3,190,811[b]	2.01	3,175,908	8,000	15/22	5 (5)	−1.48	−0.97	Girdin
	4,478,135–4,479,071[b,c]	1.38	4,478,135	2,828	11/22	8 (8)	−2.44	−2.91	CG7465
	6,100,579–6,157,140[b,c,d]	2.09	6,146,679	11,313	11/21	3 (3)	−0.179	−0.62	Lcp65Ag2
	6,550,102–6,557,978[b,c,d]	1.81	6,551,837	8,000	9/21	6 (4)	−3.92	−3.52	CG18769
	11,825,913–11,831,926[b,d]	1.35	11,829,615	2,000	17/22	9 (9)	−2.11	−2.38	CG43064
	13,425,802–13,431,380[b,d]	1.70	13,430,186	5,656	11/22	7 (7)	−2.72	−3.18	CG10713
	16,084,427–16,136,825[c,d]	2.86	16,106,542	16,000	17/22	2 (2)	−1.79	−2.45	Taf4
	17,733,967–17,741,172	2.04	17,735,433	8,000	16/22	4 (6)	0.0795	−0.49	CG7460
	19,211,270–19,237,070[c,d]	2.89	19,220,338	22,627	8/22	1 (1)	−1.85	−2.74	fz2
3R	3,697,516–3,769,886	2.90	3,727,631	45,255	9/19	5 (7)	−0.147	0.08	mRpS9
	4,155,075–4,182,535	2.68	4,158,518	64,000	11/20	7 (9)	−1.06	−1.03	CG9601
	5,530,419–5,688,202	3.91	5,548,751	90,510	8/20	2 (5)	−1.15	−1.36	CG8478
	8,486,190–8,497,516	3.32	8,497,516	32,000	11/19	3 (4)	1.63	0.53	CG14395
	9,040,956–9,111,809[c]	7.03	9,057,704	64,000	8/19	1 (1)	−2.87	−1.53	Ace
	10,380,467–10,391,240[c,d]	2.73	10,386,839	8,000	10/21	6 (3)	−2.99	−2.56	Pde6
	12,060,488–12,066,143[c,d]	2.48	12,066,090	16,000	17/22	8 (10)	−3.09	−3.01	tara
	16,575,106–16,577,277[b,d]	2.10	16,575,113	11,313	16/22	9 (14)	−1.91	−2.34	CG42322
	17,406,332–17,432,203[b]	3.03	17,414,532	16,000	15/22	4 (2)	−1.67	−0.54	InR
	18,232,347–18,251,667[b]	2.09	18,245,938	11,313	12/22	10 (6)	−1.33	−1.58	IqfR
X	525,809–1,798,033[c]	2.75	1,350,182	32,000	17/22	1 (1)	−1.06	−0.97	MED18
	2,817,759–2,835,897[c,d]	1.73	2,828,033	5,656	10/22	2 (2)	−4.89	−4.64	kirre
	14,156,405–14,160,294[b,d]	1.21	14,157,513	2,828	16/21	4 (4)	−3.86	−4.50	CG1461
	15,607,843–15,628,927[b]	1.43	15,620,351	5,656	14/21	3 (3)	−1.40	−1.90	CG8184

[a] Positions of the first (start) and last (end) sites of significant CLR ($P < 0.001$) within the cluster.

[b] Overlap with a candidate region of complete selective sweep.

[c] Overlap with a cluster detected by the *iHS* test.

[d] Overlap with a cluster detected by the nS_L test.

[e] Maximum CLR (T_1) within the cluster.

[f] The location of maximum T_1 or the putative nucleotide site under selection within the cluster.

[g] The derived allele frequency (DAF) in the data at the putative site under selection.

[h] The rank within the chromosome arm of the maximum T_1 when option A (B) is used for calculating composite likelihoods.

[i] The value of *iHS* and nS_L calculated at the putative site under selection detected by the CLR method.

haplotype structure of incomplete sweep are associated with unusually small iHH_A. Namely, stochastic fluctuation in haplotype structure surrounding the ancestral allele appears to frequently generate false-positive signatures of selection captured by *iHS* or nS_L.

To examine whether the CLR, *iHS*, and nS_L methods detect common candidate loci under selection, we adjusted the *P*-value cutoff of *iHS* or nS_L for each chromosome arm so that the numbers of *iHS* or nS_L clusters match that of the CLR in the same chromosome arm (Table S1 and Table S2).

ancestral alleles on the top and derived alleles on the bottom (divided by a horizontal green line). Derived alleles and missing base calls at polymorphic sites are marked by black and green bars, respectively.

If a CLR cluster and an *iHS* or nS_L cluster are not >50 kb away from each other, they are defined as overlapping candidates of selection. Of 25 CLR clusters that do not overlap with candidates of complete sweeps, 13 overlap with *iHS* clusters (Table 2). Ten of those 13 *iHS* clusters are also nS_L clusters, reflecting a very high level of overlap between the *iHS* and nS_L methods. There is only one case of coincidence between CLR and nS_L peaks not being an *iHS* peak (excluding those overlapping with complete sweeps). Therefore, less than half of CLR peaks were detected also by the nS_L method. Visual inspection of haplotype structures indicates that such candidate loci detected by all three methods tend to exhibit a much clearer pattern of incomplete sweeps than others (Figure S7). However, there are also loci detected by the CLR method only but with clear haplotype patterns (for example, near position 5,770,000 in 2R; Figure 5). We also identified a few peaks of negative nS_L with clear haplotype patterns not overlapping with CLR or *iHS* peaks (for example, near position 3,886,000 in 2R; Figure 5). However, such cases are exceptional: if an nS_L peak is not overlapping with the CLR or *iHS* peaks, it is more likely to show unclear than clear haplotype patterns (Figure S8).

Discussion

We developed a composite-likelihood method for detecting incomplete selective sweeps and inferring the location and strength of positive selection from DNA sequence polymorphism. As this method is built on analytic approximations to sampling probabilities under an explicit model of the evolutionary process, hypothesis testing and parameter estimation can be performed systematically, for example, allowing the estimation of the strength of selection. This approach also has the potential to be extended to incorporate more complex scenarios of incomplete sweeps if the sampling probabilities can be obtained as functions of additional parameters. On the other hand, statistical methods aiming to capture the extended haplotype such as the *iHS* and nS_L tests (Voight *et al.* 2006; Ferrer-Admetlla *et al.* 2014) have an advantage of requiring fewer assumptions about the evolutionary process to be inferred (*i.e.*, how directional selection occurs) and are also easier to implement the procedure and to interpret the result. We thus compared the performance of our CLR method and the extended haplotype method, using both simulated and actual sequence data.

Analysis of simulated data showed that our CLR approach achieves statistical power and accuracy in estimating the location of selection similar to those by the nS_L method (Table 1), however, under the assumption that the true scaled recombination rate of the genomic region is known when generating the null distribution by neutral simulation. If a falsely lower estimate of the scaled recombination rate is used for a genomic region under test, which is likely true if an incomplete selective sweep left a polymorphism with long-range LD, it will greatly reduce the statistical power to detecting it as the cutoff value in the null distribution becomes larger. Such a large sensitivity of the CLR to the recombination rate (the level of linkage disequilibrium) is a major problem that needs to be addressed in future improvement of our approach. However, if local recombination rate or map distance is well estimated in advance over a large genomic region (much larger than typical sizes of sweep-affected areas), scaled recombination at a particular locus might be correctly inferred from observed polymorphism in the neighboring regions, given that LD over a large region is much less affected by local fluctuation, for example by selection. Namely, generating the null distribution with neutral simulation that yields the observed level of LD in data under test, as we suggested to correct the effect of unknown recombination rate, might be an unnecessarily conservative test, if the observed LD is definitely unusual (*i.e.*, higher) compared to that in neighboring regions.

A related problem due to the sensitivity of our statistic to the level of linkage disequilibrium is the increased chance of detecting false-positive incomplete sweeps in the presence of nonstandard demography (Figure S4). Because various demographic processes can inflate the level of LD throughout the genome, which upwardly shifts the distribution of T_1 in the absence of selection, obtaining the null distribution under the assumption of the standard neutral model can lead to erroneous detections of sweeps. Again, if the nature of (complex) demography affecting the data is not known, the false-positive detection might be controlled by the null distribution from simulated samples under the standard neutral model but adjusted to exhibit the level of LD observed in the data.

A more important result in the comparison between the CLR and *iHS*/nS_L tests is that their performances are rather complementary to each other, as their outcomes are not so strongly correlated, especially for weak selection ($\alpha = 1000$; Figure 3). It is probably because the two methods are designed to detect slightly different footprints of incomplete selective sweeps. Our method primarily captures joint frequency spectra at linked neutral loci for the two subsamples divided according to the *S* locus (Figure 2), whereas the *iHS* and nS_L methods target the extended haplotype homozygosity, although these two signatures are obviously closely related through the reduction of polymorphism surrounding the putative beneficial allele.

As it was not as feasible to evaluate statistical significance of CLR tests by generating appropriate null distributions for a large number of genomic regions in *D. melanogaster*, we applied the CLR and *iHS*/nS_L methods as outlier detection approaches. We evaluated the relative performance of the three methods by obtaining similar numbers of outliers (candidate loci) for each chromosome arm and visually inspecting haplotype structures surrounding the putative sites under selection. In general, the clearest haplotype patterns of incomplete selective sweeps were obtained when the loci were detected by all three methods. Candidates detected only by our CLR method exhibited relatively clean patterns compared to those detected by the *iHS* or nS_L method (Figure 5, Figure S7, and Figure S8). Again this can be attributed to the gain of additional information from DNA sequence polymorphism in the CLR approach. Visual inspection also suggests that many false positives are detected by *iHS* because extended homozygosity surrounding the ancestral allele of the core SNP can be randomly reduced to very small values. Namely, while iHH_D captures the hitchhiking effect of the beneficial allele, stochastic fluctuation of iHH_A greatly increases the variance of iHH_A/iHH_D. In addition, if a small number, say n', of sequences containing the derived allele of focal SNP are highly homozygous (*e.g.*, hidden identity by descent) by chance while the other $n_1 - n'$ sequences are heterozygous at the normal level, it can lead to a very large iHH_D. Our approach is not affected by such problems, as our CLR does not simply depend on differences in the levels of variation between the two subsamples of data but compares neutral *vs.* selective scenarios as potential explanations for the subdivided pattern of polymorphism. The stochastic fluctuation of SNP density in the ancestral block appears to be less of a problem for nS_L than for *iHS*, given that much clearer haplotype structures are detected by nS_L than by *iHS*, probably because it does not use genetic map distance but the number of intervening SNPs for measuring the size of the extended haplotype.

As population genomic data are obtained predominantly by NGS platforms, missing or low-quality base calls in data may greatly affect the performance of evolutionary inferences from DNA sequence polymorphism. It is straightforward to calculate sampling probability under both neutral and selective hypotheses given the configuration of missing bases at each site in the data. Therefore, our CLR approach can be applied to data with an arbitrary frequency of missing bases without systematic problems. On the other hand, it is not clear how to handle missing bases in quantifying the extended homozygosity for the *iHS* or nS_L test. We skipped the site containing a missing base in calculating the extension of homozygosity for a pair of sequences because clear haplotype structure of an incomplete sweep could not be identified otherwise. It is not clear how this procedure would affect the performance of the *iHS* test.

In conclusion, we proposed a composite-likelihood method for detecting incomplete selective sweeps and demonstrated that it achieves improvements in parameter estimation and ability to capture clear haplotype patterns compatible with incomplete sweeps compared to long-range haplotype tests. Although it has a disadvantage in not being robust to uncertainty in scaled recombination rates and complex demography, our composite-likelihood ratio provides information that is not captured by an advanced haplotype-based method using nS_L. We thus recommend that both CLR and nS_L be used together to maximize the chance of detecting true targets of selection. As incomplete selective sweeps provide excellent opportunities to estimate the strength and location of selection, due to the presence of ancestral polymorphism in the data, compared to complete sweeps, these methods will contribute to broadening our understanding of adaptive evolution in nature. In the framework of the likelihood-ratio test, we may conceive extension of this approach to study further details of incomplete selective sweeps beyond simple confirmation of positive selection and basic parameter estimation. For example, recent analysis predicted that many beneficial mutations are likely to stall at intermediate frequencies due to heterozygote advantage (Sellis *et al.* 2011). If this process generates sampling probabilities distinct from that left by simple directional selection with incomplete dominance, we may detect it under the current framework of the composite-likelihood test.

Acknowledgments

This research was supported by the Global Top5 Grant of Ewha Womans University 2013 and the National Research Foundation of Korea grants 2012R1A1A2004932 (to Y.K.).

Literature Cited

Akey, J. M., 2009 Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. 19: 711–722.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133.

Biswas, S., and J. M. Akey, 2006 Genomic insights into positive selection. Trends Genet. 22: 437–446.

Chan, A. H., P. A. Jenkins, and Y. S. Song, 2012 Genome-wide fine-scale recombination variation in Drosophila melanogaster. PLoS Genet. 8: e1003090.

Comeron, J. M., R. Ratnappan, and S. Bailin, 2012 The many landscapes of recombination in Drosophila melanogaster. PLoS Genet. 8: e1002905.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger, 2006 An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. 16: 685–729.

Ewing, G., and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.

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

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. 31: 1275–1291.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Gautier, M., and R. Vitalis, 2012 rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28: 1176–1177.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 473–485.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics 136: 1329–1340.

Innan, H., and Y. Kim, 2008 Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics 179: 1713–1720.

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

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.

Meiklejohn, C. D., Y. Kim, D. L. Hartl, and J. Parsch, 2004 Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics 168: 265–279.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. 39: 197–218.

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.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno *et al.*, 2012 Population genomics of sub-Saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet. 8: e1003080.

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

Quesada, H., U. E. Ramírez, J. Rozas, and M. Aguadé, 2003 Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics 165: 895–900.

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.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science 312: 1614–1620.

Saunders, M. A., J. M. Good, E. C. Lawrence, R. E. Ferrell, W.-H. Li *et al.*, 2006 Human adaptive evolution at myostatin (GDF8), a regulator of muscle growth. Am. J. Hum. Genet. 79: 1089–1097.

Sellis, D., B. J. Callahan, D. A. Petrov, and P. W. Messer, 2011 Heterozygote advantage as a natural consequence of adaptation in diploids. Proc. Natl. Acad. Sci. USA 108: 20666–20671.

Stephan, W., 2010 Detecting strong positive selection in the genome. Mol. Ecol. Res. 10: 863–872.

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

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

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

Toomajian, C., R. S. Ajioka, L. B. Jorde, J. P. Kushner, and M. Kreitman, 2003 A method for detecting recent selection in the human genome from allele age estimates. Genetics 165: 287–297.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72.

Communicating editor: R. Nielsen

Appendix

Derivation of ϕ_{S1} and ϕ_{S2}

We consider a constant-sized population of N diploid individuals that reproduce in discrete generations according to the Wright–Fisher model, thus equivalent to a population of $2N$ haploid individuals. Assume that mutation to a beneficial allele occurred at position x of a chromosome at time $T = \tau$ (generations counted backward in time) in the past. At the time of sampling ($T = 0$), this mutant allele reaches an intermediate frequency β in the population. A random sample of n chromosomes is assumed to contain n_1 and $n_2 = n - n_1$ copies of the beneficial and the ancestral allele, respectively, that define the corresponding partition of the sample into two subsamples as illustrated in Figure 1. Let k_1 and k_2 be the counts of the derived allele in respective subsamples at a neutrally evolving site at position $x - d$ or $x + d$. The probability of observing k_1 and k_2 jointly is given by

$$\phi(k_1,k_2;n_1,n_2,d) \approx \int_0^1 \varphi_{\mathrm{S}}(k_1;n_1,p,d)\varphi_{\mathrm{N}}(k_2;n_2,p)f_0(p)dp, \tag{A1}$$

where $f_0(p)$ is the probability density of the derived allele frequency at the time of beneficial mutation ($T = \tau$). $\varphi_{\mathrm{N}}(k_2;n_2,p)$ is the probability of sampling k_2 derived alleles in a sample of n_2 chromosomes in a neutrally evolving population in which frequency of the allele drifted for τ generations starting from p. During the course of a selective sweep, the deterministic change of the linked neutral allele frequency among chromosomes carrying the ancestral allele at the S locus (frequency p_{A} in the "ancestral background") is predicted to be small (Stephan *et al.* 1992; Meiklejohn *et al.* 2004). A moderate deterministic change in p_{A} occurs while $\beta < 0.8$, the range to which our method applies (Figure S9). We, however, ignore this change. We also ignore the change of allele frequency by genetic drift in the ancestral background, assuming $\tau << 2N(1-\beta)$, and obtain

$$\varphi_{\mathrm{N}}(k_2;n_2,p) = \binom{n_2}{k_2}p^{n_2}(1-p)^{n_2-k_2}. \tag{A2}$$

Namely, we assume that the subsample of n_2 chromosomes effectively captures the ancestral polymorphism at the time of beneficial mutation. Next, $\varphi_{\mathrm{S}}(k_1;n_1,d,p)$ is the probability of observing k_1 copies of the derived allele at position d in the subsample of n_1 sequences that carry the beneficial allele. Strictly, this probability must be a function of the frequency of the beneficial allele at the time of sampling. However, as the frequency of the neutral allele among chromosomes carrying the beneficial allele (*i.e.*, in the "beneficial background") is known to change drastically only at the early stage of hitchhiking when the frequency of the beneficial allele is low and then change little until the fixation of the beneficial allele (Stephan *et al.* 1992), we approximate $\varphi_{\mathrm{S}}(k_1;n_1,d,p)$ by sampling probability for the case of the complete selective sweep. We multiply φ_{S} and φ_{N} inside the integral of (A1), assuming that the frequency of linked neutral alleles in the beneficial background is distributed independently of possible stochastic change in allele frequency in the ancestral background and that chromosomes are sampled independently in the two genetic backgrounds. In reality, the "migration" of lineages by recombination during the selective sweep may cause correlated stochastic changes of allele frequencies in the two backgrounds. However, we ignore such complications, as the stochastic fluctuation of p in the ancestral background by genetic drift is ignored in the first place (see above).

Nielsen *et al.* (2005) and Etheridge *et al.* (2006) provided approximate solutions that allow the derivation of the above sampling probability φ_{S} as a function of neutral allele frequency, p, at the time of the beneficial mutation. Using a star-like genealogy approximation, Nielsen *et al.* (2006) obtained the probability of observing k_1 derived alleles at the neutral locus from the sample of n_1 chromosomes after a selective sweep,

$$\varphi_{\mathrm{S}}(k_1;n_1,d,p) = Z_{n_1,n_1}v_{k_1,n_1} + \sum_{i=0}^{n_1-1} Z_{i,n_1}\left(v_{k_1+1-n_1+i,i+1}\frac{k_1+1-n_1+i}{i+1} + v_{k_1,i+1}\frac{i+1-k_1}{i+1}\right), \tag{A3}$$

where $v_{k,n} = \binom{n}{k}p^k(1-p)^{n-k}$ is the probability that k of n distinct ancestral lineages at $T = \tau$ carry the derived mutant alleles and $Z_{k,n} = \binom{n}{k}{z_e}^k(1-z_e)^{n-k}$ is the probability that k of n lineages at $T = 0$ escape the sweep by recombining away from the beneficial allele, with the escaping probability per lineage given by $z_e = 1 - (4Ns)^{-(r_n d/s)} = 1 - (2\alpha)^{-(R/2\alpha)}$.

Replacing (A2) and (A3) into (A1), we obtain

$$\begin{aligned}\phi_{S1}(k_1,k_2;n_1,n_2,d) = {} & Z_{n_1,n_1}\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}}P(k_1+k_2|n_1+n_2) \\ & + \sum_{i=0}^{n_1-1} Z_{i,n_1}\left(\frac{\binom{i+1}{k_1+1-n_1+i}\binom{n_2}{k_2}}{\binom{n_2+i+1}{k_1+1-n_1+i+k_2}} \frac{k_1+1-n_1+i}{(i+1)} P(k_1+1-n_1+i+k_2|n_2+i+1) \right. \\ & \left. + \frac{\binom{i+1}{k_1}\binom{n_2}{k_2}}{\binom{n_2+i+1}{k_1+k_2}}\frac{i+1-k_1}{(i+1)}P(k_1+k_2|n_2+i+1)\right),\end{aligned} \tag{A4}$$

where $P(k|n)$ is the probability of having k derived alleles in a sample of n chromosomes at time τ . $P(k|n)$ can be given by θ/k, assuming the population at this time is under neutral equilibrium, or by the proportion of polymorphic sites with k derived alleles in the data, namely assuming that the distribution of the derived allele frequency at time τ is identical to that observed at present. The latter approach of using the empirical frequency spectrum was suggested by Nielsen *et al.* (2005) to correct for nonequilibrium demography. These two approximations are bases of CLR test options A and B, respectively.

Alternatively, we may derive the sampling probability from the work of Etheridge *et al.* (2006), which showed that n lineages at a linked neutral locus sampled at the time of a beneficial allele's fixation are divided into three parts: l late recombinants, e early recombinants, and $n - l - e$ nonrecombinants. Given the selection coefficient s and recombination rate r, the joint distribution of l and e, $P(l, e)$, follows equation 2.7 of Etheridge *et al.* (2006). However, this result in terms of genealogical structure needs to be translated into sampling probability by considering the transmission of mutant alleles along the lineages. The probability of sampling k derived alleles can be obtained separately in the following four cases.

First, consider the case in which the beneficial allele appears on a chromosome carrying the derived allele at the neutral locus. In addition, the ancestor of early recombinants carries the ancestral allele. Therefore, the sample contains at least $n - l - e$ derived alleles and at least e ancestral alleles. In addition, assume that in l late recombinants, there are l_d derived alleles and $l - l_d$ ancestor alleles. Then, the total number of derived allele in the sample is $k = n – e – (l – l_d)$. Since $l_d = l – (n – e – k)$, the probability for this case is

$$S_1(k) = \sum_{e=0}^{n-k}\sum_{l=n-e-k}^{n-e} P(e,l)\binom{l}{n-e-k}p^{l-(n-e-k)}(1-p)^{n-e-k}, \tag{A5}$$

where p is the initial frequency of the derived allele before hitchhiking. In the case that the ancestor of early recombinants carries the derived allele,

$$S_2(k) = \sum_{e=0}^{k}\sum_{l=n-k}^{n-e} P(e,l)\binom{l}{n-k}p^{l-(n-k)}(1-p)^{n-k}. \tag{A6}$$

Next, the beneficial mutation is now assumed to appear on a chromosome carrying the ancestral allele of the neutral locus. Probabilities that there are k derived alleles in the sample if the ancestor of early recombinants carries the ancestral and the derived allele are, respectively,

$$S_3(k) = \sum_{e=0}^{n-k}\sum_{l=k}^{n-e} P(e,l)\binom{l}{k}p^{k}(1-p)^{l-k} \tag{A7}$$

and

$$S_4(k) = \sum_{e=0}^{k}\sum_{l=k-e}^{n-e} P(e,l)\binom{l}{k-e}p^{k-e}(1-p)^{l-k+e}. \tag{A8}$$

Since these cases are mutually exclusive, the final solution for sampling probability for a complete selective sweep is after the above probabilities are weighted accordingly:

$$
\begin{aligned}
\varphi_S(k;n,p,d) &= p\big((1-p)S_1(k)+pS_2(k)\big)+(1-p)\big((1-p)S_3(k)+pS_4(k)\big)\\
&=\sum_{e=0}^{n-k}\left[\sum_{l=n-k-e}^{n-e}P(e,l)\binom{l}{n-e-k}(1-p)^{n-k-e+1}\,p^{k+e+l-n+1}\right.\\
&\quad\left.+\sum_{l=k}^{n-e}P(e,l)\binom{l}{k}p^{k}(1-p)^{l-k+2}\right]\\
&\quad+\sum_{e=0}^{k}\left[\sum_{l=n-k}^{n-e}P(e,l)\binom{l}{n-k}p^{l-(n-k)+2}(1-p)^{n-k}\right.\\
&\quad\left.+\sum_{l=k-e}^{n-e}P(e,l)\binom{l}{k-e}p^{k-e+1}(1-p)^{l-(k-e)+1}\right].
\end{aligned}
\tag{A9}
$$

Using Equations A2 and A9, Equation A1 is now turned into our second approximation:

$$
\begin{aligned}
\phi_{S2}(k_1,k_2;n_1,n_2,d) \quad &= \theta\binom{n_2}{k_2}\sum_{e=0}^{n_1-k_1}\left[\sum_{l=n_1+k_1-e}^{n_1-e}P(e,l)\,\frac{\binom{l}{n_1-e-k_1}}{\binom{n_2+l+2}{e+l+k_1+k_2+1-n_1}}P(e+l+k_1+k_2+1-n_1|n_2+l+2)\right.\\
&\quad\left.+\sum_{l=k_1}^{n_1-e}P(e,l)\frac{\binom{l}{k_1}}{\binom{n_2+l+2}{k_1+k_2}}P(k_1+k_2|n_2+l+2)\right]\\
&\quad+\sum_{e=0}^{k_1}\left[\sum_{l=n_1-k_1}^{n_1-e}P(e,l)\frac{\binom{l}{n-k_1}}{\binom{n_2+l+2}{k_1+k_2+l+2-n_1}}P(k_1+k_2+l+2-n_1|n_2+l+2)\right.\\
&\quad\left.+\sum_{l=k_1-e}^{n_1-e}P(e,l)\frac{\binom{l}{k_1-e}}{\binom{n_2+l+2}{k_1+k_2+1-e}}P(k_1+k_2+1-e|n_2+l+2)\right].
\end{aligned}
\tag{A10}
$$

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175380/-/DC1

A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data

Ha My T. Vy and Yuseob Kim

DOI: 10.1534/genetics.115.175380

SUPPORTING INFORMATION for

H. T. Vy and Y. Kim, A composite likelihood method for detecting incomplete selective sweep from population genomic data, submitted to Genetics

File S1

Scripts to generate simulated data sets:

1. For testing power of CL, iHS, nSL:

- $R = 4Nr = 2000$

➢ Neutral model:

```
/msms 20 10000 -N 100000 -s 3000  -r 2000 100000
```

➢ Selective sweep model:

```
/msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 1000 -SaA 500 -SF 0 0.5 -Sp 0.5 –
Smark
/msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 2000 -SaA 1000 -SF 0 0.5 -Sp 0.5 –
Smark
/msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 4000 -SaA 2000 -SF 0 0.5 -Sp 0.5 –
Smark
```

- R = 4 = 4000

➢ Neutral model:

```
/msms 20 10000 -N 100000 -s 3000  -r 4000 100000
```

➢ Selective sweep model:

```
/msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 1000 -SaA 500 -SF 0 0.5 -Sp 0.5 –
Smark
/msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 2000 -SaA 1000 -SF 0 0.5 -Sp 0.5 –
Smark
/msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 4000 -SaA 2000 -SF 0 0.5 -Sp 0.5 –
Smark
```

2. For generating neutral data matching the sample size, mean recombination rate, and the mean density of polymorphic sites to those of Drosophila genome data (to calculate T_1 when apply composite likelihood test to Drosophila genomes):

```
/ms 22 20 -t 35000 -r 60000 5000000
```

3. To simulate data under different demographic assumptions:

- Population bottleneck:

➢ With different severities:

```
/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.2 -eN 0.1 1.0
```

```
/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.1 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.05 -eN 0.1 1.0
```

- ➢ With different recombination rates:

```
/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 6000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 8000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 10000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 12000 100000 -eN 0.05 0.05 -eN 0.1 1.0
```

- Exponential population growth:

- ➢ With different growth rates:

```
/ms 20 1000 -s 3000 -r 4000 100000 -G 500 -eG 0.0032 0.0
/ms 20 1000 -s 3000 -r 4000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 4000 100000 -G 10 -eG 0.016 0.0
```

- ➢ With different recombination rates:

```
/ms 20 1000 -s 3000 -r 4000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 6000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 8000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 10000 100000 -G 100 -eG 0.016 0.0
```

- Population subdivision:

- ➢ With different migration rates:

```
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 1.0
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 10
```

- ➢ With different recombination rates:

```
/ms 20 1000 -s 3000 -r 1000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 2000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 6000 100000 -I 2 20 0 0.1
```

Table S1: List of putative loci under incomplete selective sweeps in *D. melanogaster* Rwanda population inferred from *iHS* test.

Chromosome	Cluster start - end	Minimum standardized *iHS*	Site of minimum *iHS*	Derived allele frequency
2L	1547008 - 1557247	-3.75	1547204	15/22
	4824296 - 4860577	-4.03	4824431	10/22
	5810884 - 5825009	-3.34	5815486	8/22
	6020532 - 6055868	-3.34	6055868	9/22
	9509499 - 9831699	-5.90	9582539	10/20
	11022970 - 11036917	-3.82	11036917	9/21
	11866168 - 11892750	-3.52	11892750	9/21
	12804885 - 12840609	-3.76	12840609	10/21
	16019930 - 16020004	-3.17	16019930	9/21
	17230328 - 17297935	-3.84	17297624	12/21
	17602339 - 17603635	-3.32	17603635	13/21
2R	5649989 - 5718051	-3.33	5708056	11/20
	7114975 - 7137571	-3.87	7127281	11/22
	7828412 - 8658752	-5.19	7911386	12/22
	10135366 - 10665005	-3.86	10665005	9/22
	11001899 - 11016632	-3.28	11006196	9/22
	12723745 - 12768255	-4.38	12734718	10/21
	13832525 - 13835190	-3.49	13832748	11/22
3L	3103656 - 3145740	-4.84	3142414	9/22
	4472504 -4490155	-4.35	4490062	8/22
	5960208 - 5974572	-4.04	5966477	8/22
	6072249 - 6129005	-5.13	6109760	10/21
	6537946 - 6595815	-4.26	6548571	12/21
	8126905 - 8182746	-4.18	8134344	10/22
	14430470 - 14434825	-4.12	14431036	11/18
	16070044 - 16095749	-4.48	16095749	12/22
	19210723 - 19236655	-4.14	19210732	8/22
3R	8567616 - 8567660	-3.70	8567660	9/19

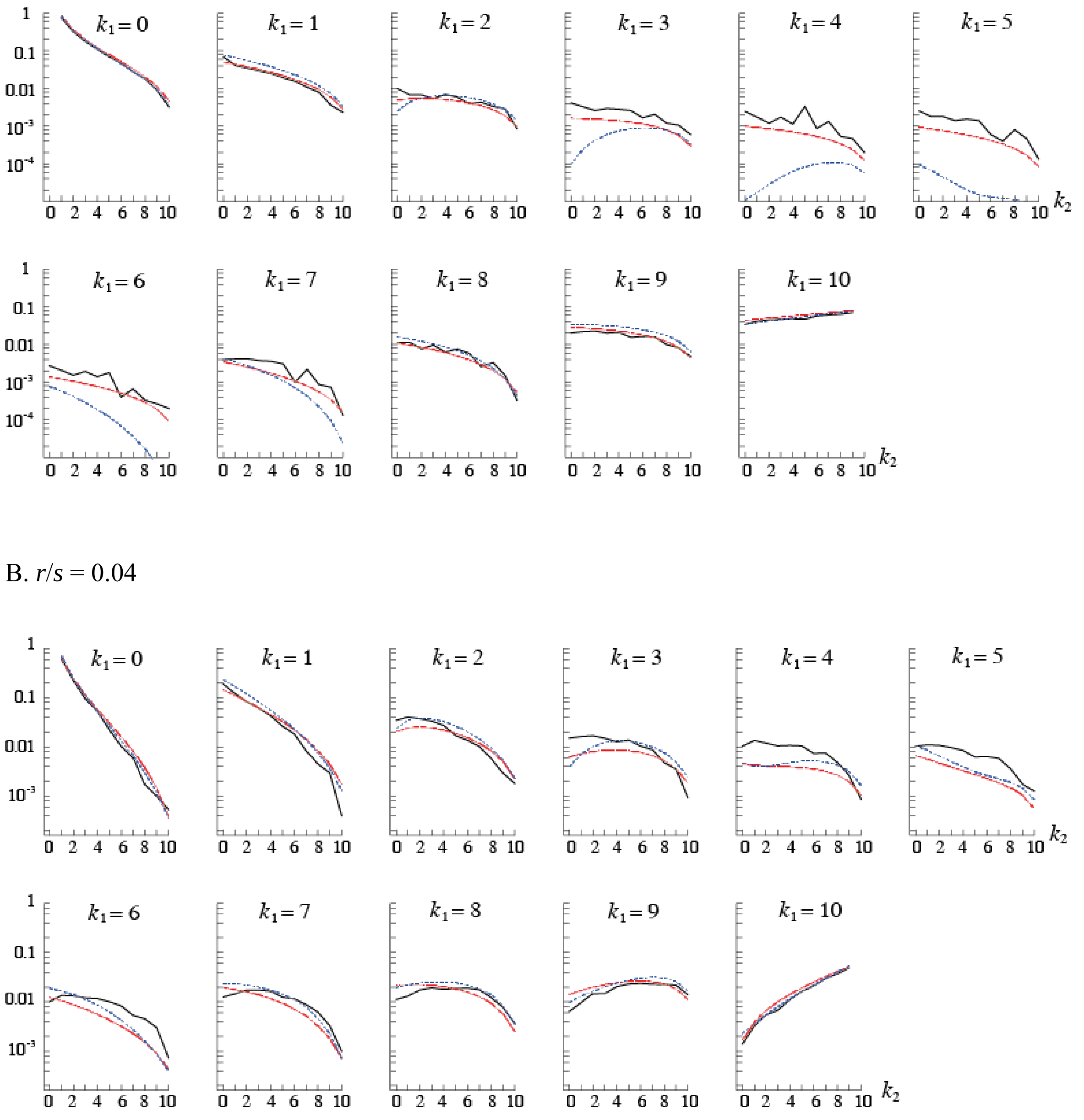
	9033505 - 9157259	-3.63	9125222	8/19
	10333644 - 10389581	-3.46	10386821	10/22
	12063858 - 12066090	-3.12	12063858	14/22
	13905933 - 13911447	-3.84	13910288	12/22
	15427616 - 15503321	-3.42	15442743	13/22
	16254275 - 16259857	-3.40	16254275	11/22
	18973474 - 18973529	-3.49	18973529	13/22
X	737336 - 1229075	-4.50	1081402	8/22
	2814828 - 2836819	-5.68	2832651	9/22
	17230521 - 17234450	-4.88	17230521	8/21
	18636049 - 18639500	-5.23	18636156	8/20
	19077625 - 19104409	-5.80	19093150	8/20

Table S2: List of putative loci under incomplete selective sweeps in *D. melanogaster* Rwanda population inferred from nS_L test.

Chromosome	Cluster start – end	Minimum standardized nS_L	Site of minimum nS_L	Derived allele frequency
2L	1147263-1152901	-3.77	1147263	8/22
	1545621-1548645	-3.68	1545621	14/22
	5812475-5816415	-3.15	5815486	8/22
	6596769-6603746	-3.99	6603091	9/22
	9433919-9610983	-3.33	9433919	17/22
	12372465-12377872	-3.46	12375980	11/21
	12718277-12844111	-3.10	12834789	8/21
	16798951-16894905	-3.35	16894905	12/21
	17234502-17245049	-3.45	17235226	9/21
	17602170-17619087	-3.72	17603916	16/21
	19727261-19768977	-3.44	19730058	8/21
2R	3786646-3886834	-2.42	3886479	8/21
	5254576-5283840	-2.54	5266445	8/21
	5548485-5552974	-2.42	5548485	9/21
	7119351-7133870	-2.47	7126724	12/22
	7816110-8669818	-3.29	8133130	17/22
	12721950-12740626	-2.83	12727369	8/22
	13826522-13834107	-2.46	13826676	8/22
	18089346-18541764	-2.80	18089697	8/22
3L	2997574-2998073	-3.82	2998073	8/22
	3136553-3149179	-4.49	3144835	11/22
	6087588-6129005	-4.51	6129005	16/21
	6538594-6567487	-4.18	6547881	8/21
	11654477-11679598	-3.59	11654544	9/21
	11826677-11839249	-3.88	11833069	12/18

	13427088-13432591	-3.44	13430725	11/22
	16076838-16137484	-4.20	16094714	12/22
	19210376-19236655	-3.80	19210732	8/22
3R	10230101-10904118	-3.14	10333644	14/22
	12062855-12066090	-3.37	12062855	12/22
	13330699-13334656	-3.14	13330699	15/22
	13906691-13911447	-3.41	13908193	10/22
	15432404-15445674	-3.46	15443091	12/22
	16561264-16575140	-3.22	16569519	13/21
	19838638-19884976	-3.17	19839035	12/22
	20872678-20878483	-3.49	20876949	12/22
X	2814198-2838349	-5.45	2833040	9/22
	10210768-10225115	-4.21	10219139	8/22
	14151360-14164063	-4.50	14157513	16/21
	14969055-14996063	-4.16	14975977	13/21
	16948533-17158766	-4.25	17124249	8/19

Figure S1


A. *r*/*s* = 0.01

B. *r*/*s* = 0.04

Figure S1 legend: Joint sampling probability under incomplete selective sweep for $n_1 = n_2 = 10$ and r/s = 0.01 or 0.04. ϕ_{S1} (blue) and ϕ_{S2} (red) are compared against simulation result (black).

Figure S2

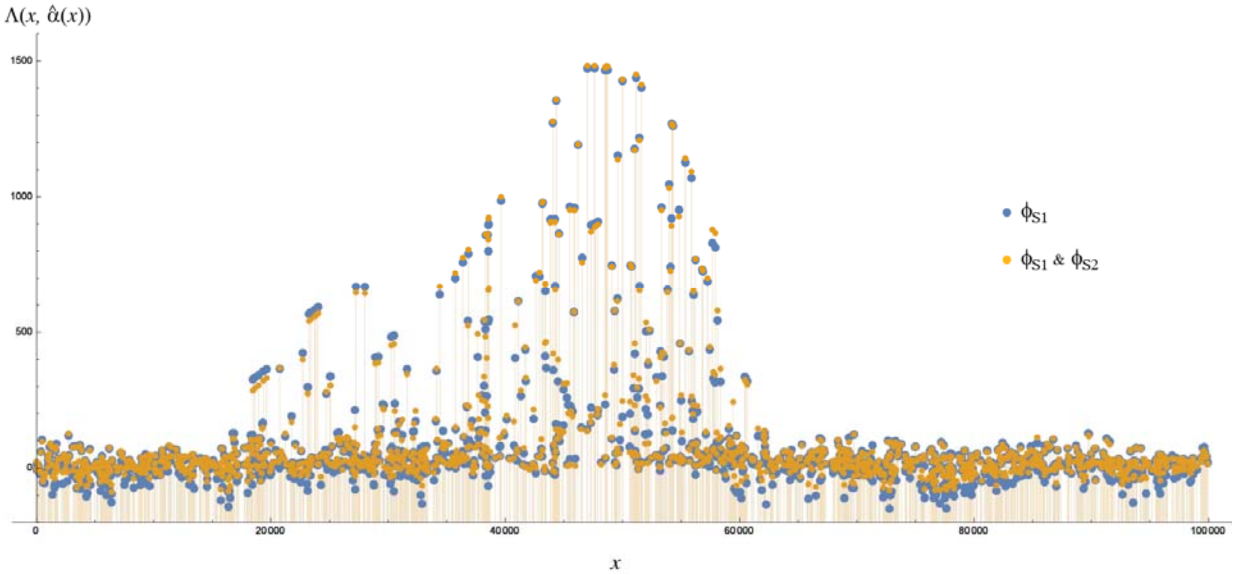


Figure S2 legend: Composite likelihood ratio calculated for a simulated data set of 20 DNA sequences of 100kb long (R = 4,000). Advantageous mutation with α = 4,000 is located in the middle (50kb). Blue dots are CLR calculated using ϕ_{S1}, approximation suggested by Nielsen et al. (2005), and yellow dots are CLR calculated using ϕ_{S1} for $r/s < 0.03$ but ϕ_{S2}, approximation based on Etheridge et al. (2006), for $r/s \geq 0.03$.

Figure S3

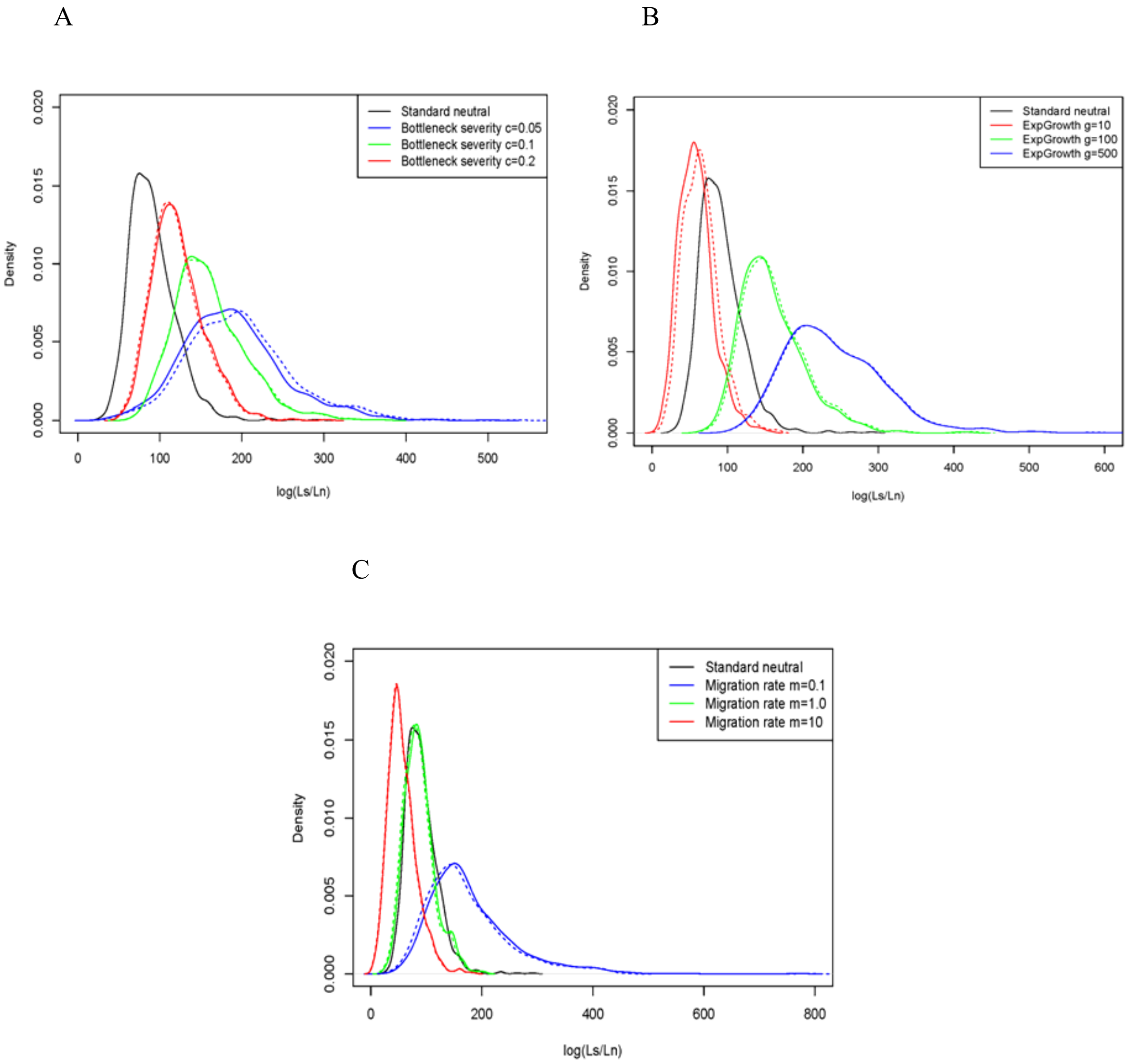


Figure S3 legend: Distributions of maximum CLR, $T_0 = \max_{x \in S[10]} \log(L_{\mathrm{IS}} / L_{\mathrm{N}})$ where the maximum was obtained over the set of polymorphic sites with $n_1 = 10$ (S[10] for each replicate), for samples generated under different demographic models: A, population bottleneck model with different bottleneck severities c = 0.05, 0.1, and 0.2; B, exponential population growth with different growth rates g = 10, 100 and 500; C, population subdivision model with different migration rates m = 0.1, 1, and 10 between 2 subpopulations. Recombination rate $4Nr_{\mathrm{n}} = 0.04$ (R = 4,000) was used to generate all data sets. Distribution of T_0 for standard neutral model is plotted in each figure (black lines) for comparison. Distributions of T_0 calculated from empirical frequency spectrum (option B) are shown by dashed curves.

Figure S4

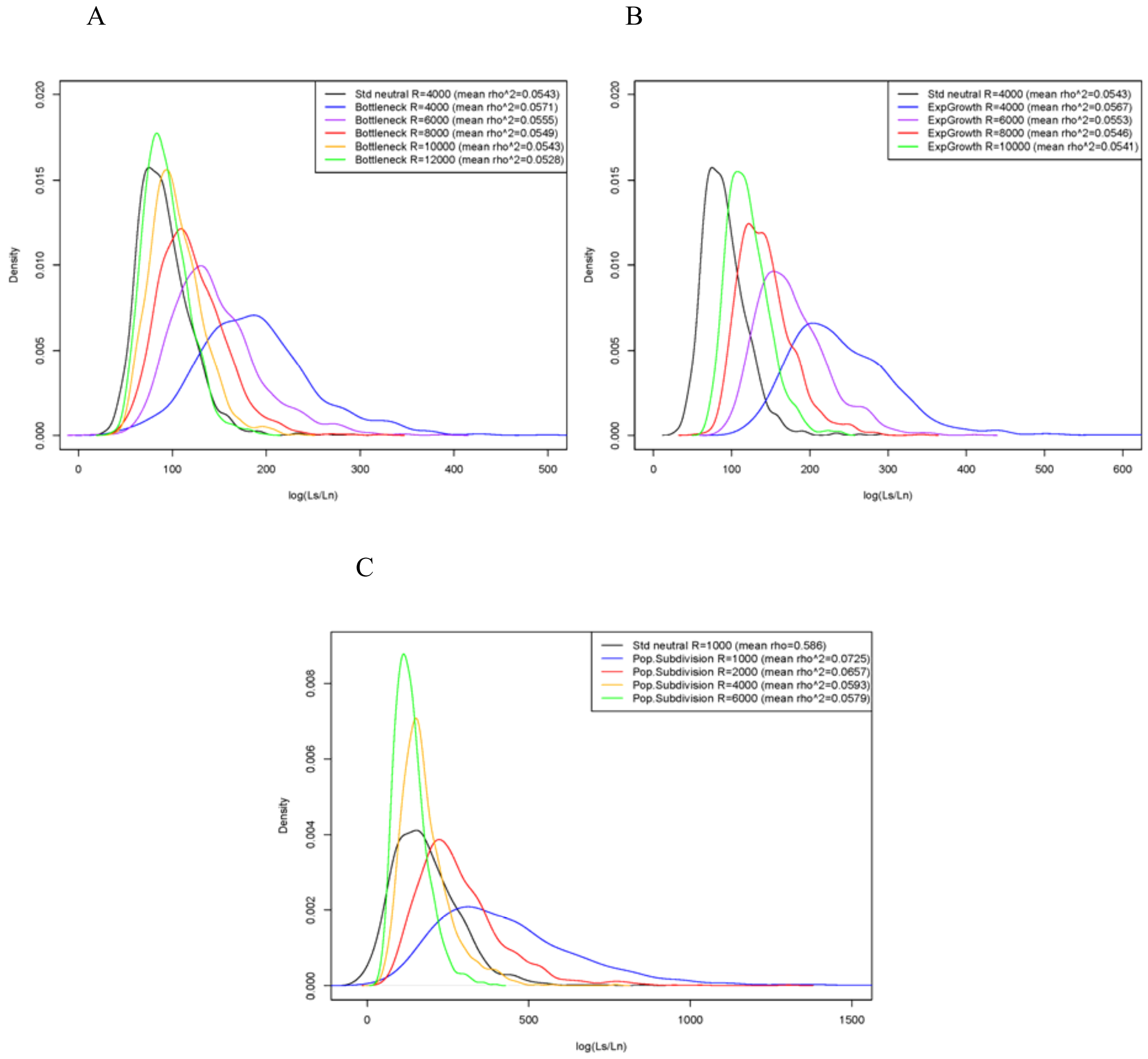


Figure S4 legend: Changes in the distributions of T_0 with varying recombination rates in different demographic models: A, population bottleneck model with bottleneck severity c = 0.05; B, exponential population growth with growth rate g = 500; C, population subdivision model with migration rate between two subpopulations m = 0.1. Mean correlation coefficient of LD among polymorphic sites (average ρ^2) for each model is shown in parenthesis.

Figure S5

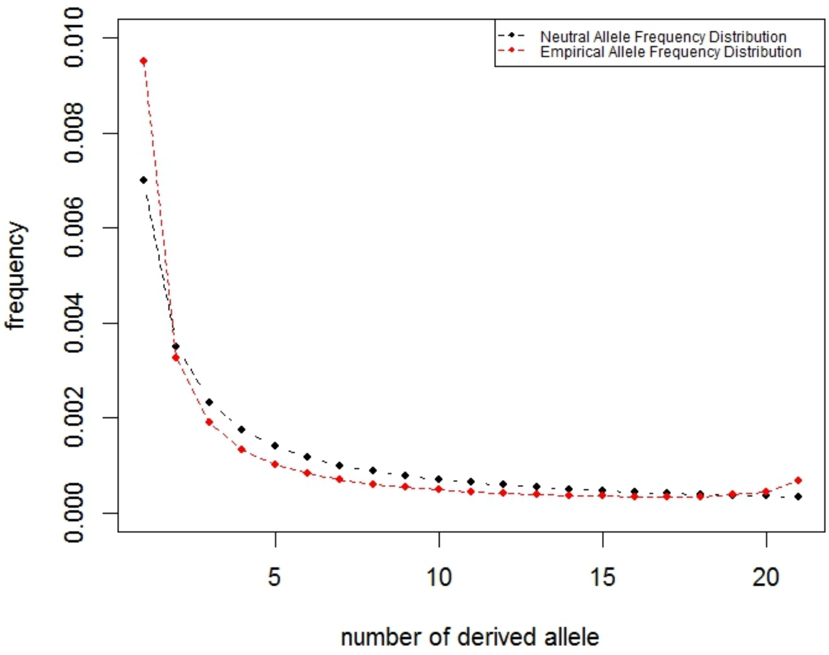


Figure S5 legend: Genome-wide empirical distribution of derived-allele frequency in the Rwanda D. *melanogaster* sample (22 sequences) in comparison with the standard neutral distribution for a sample of same size.

Figure S6

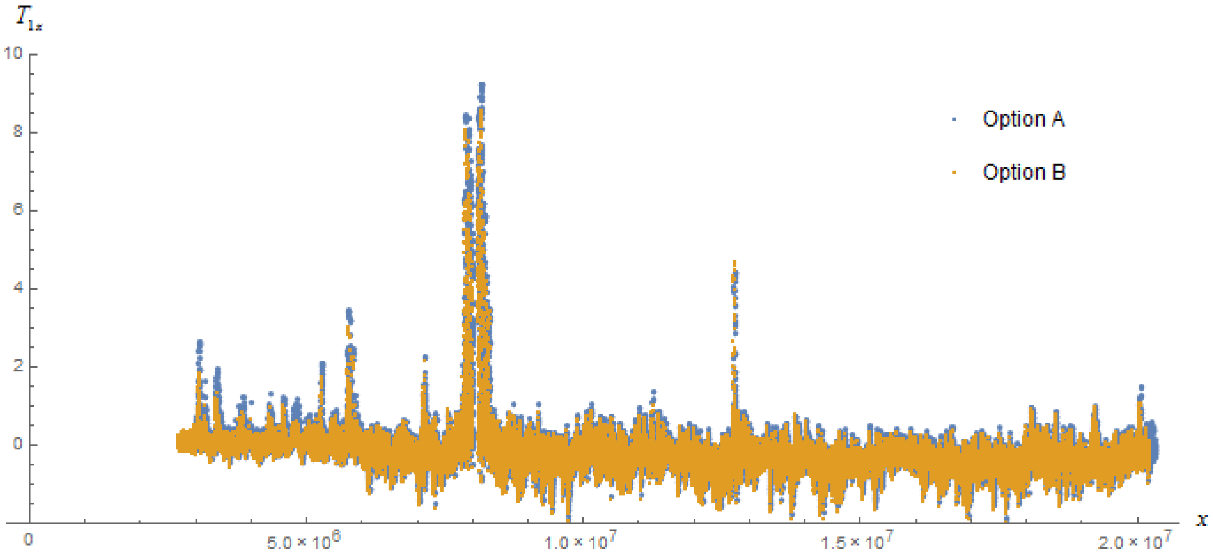
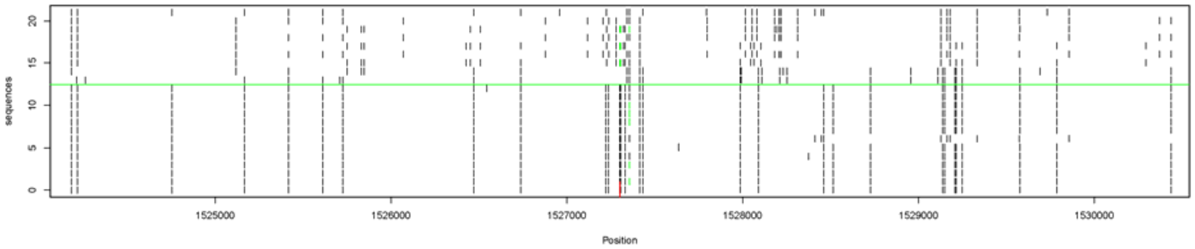


Figure S6 legend: Composite Likelihood Ratio (T_{1x}) calculated for chromosome 2R. T_{1x} was calculated based on sampling probabilities assuming neutral equilibrium (option A) or empirical frequency spectrum (option B) at the start of a selective sweep.

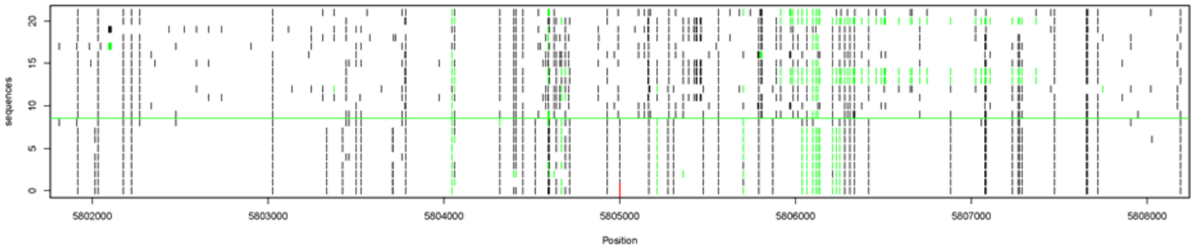
Figure S7

2L Patterns:

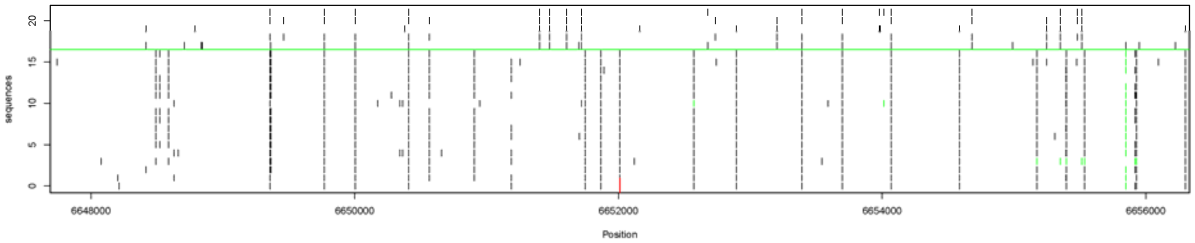
Putative site: 1527302*, closest gene: halo (1517533 – 1518148)



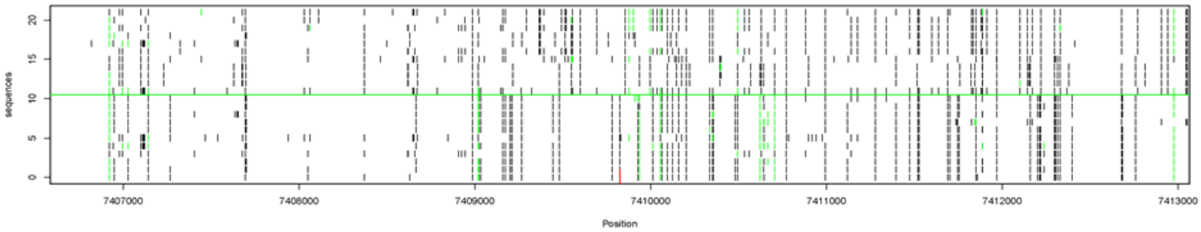
Putative site: 5805001*$^{\#\$}$, closest gene: CG11034 (5805395 – 5809063)



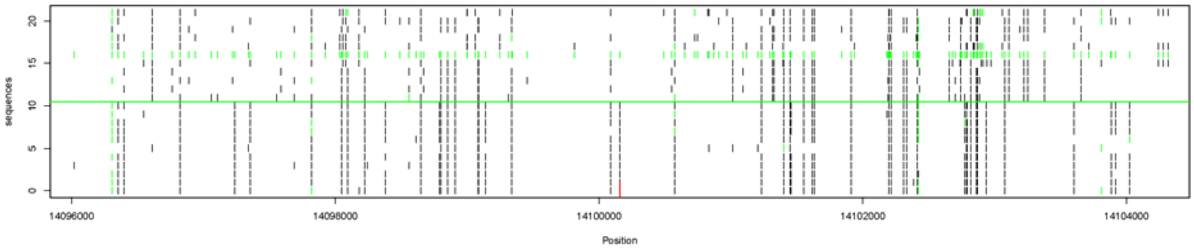
Putative site: 6652011$^{*\$}$, closest gene: Tango1 (6649388 – 6654574)



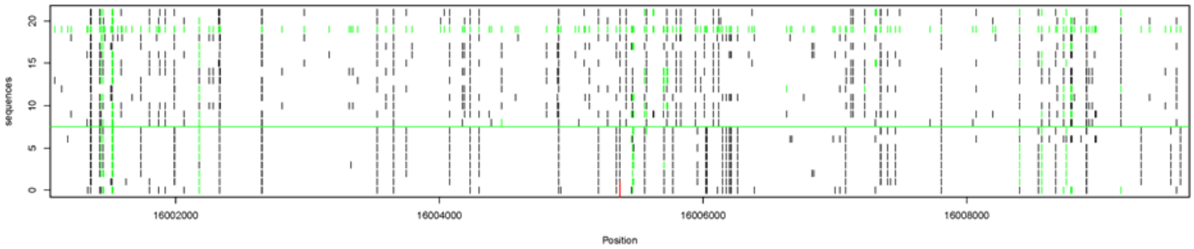
Putative site: 7409825, closest gene: CG5181 (7408533 – 7409809)



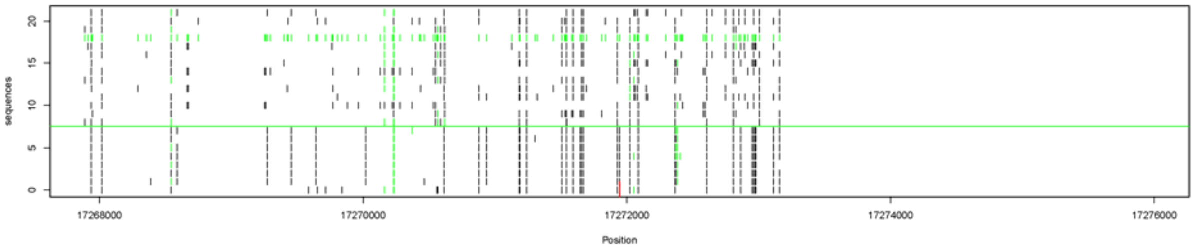
Putative site: 14100158*, closest gene: nAChRalpha5 (14040170 – 14094401)



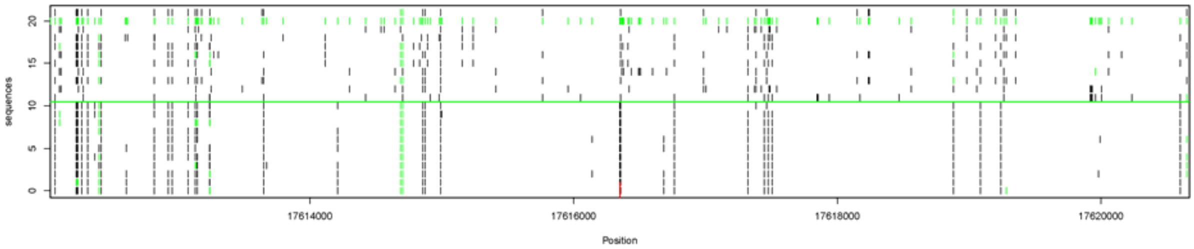
Putative site: 16005369$^{\#}$, closest gene: Beat-Ic (16000291 – 16041703)



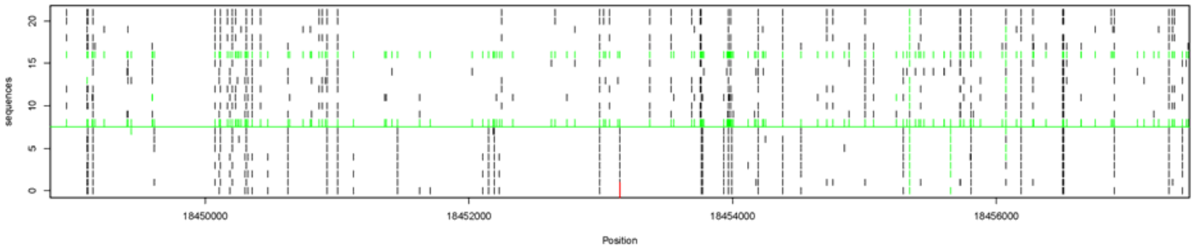
Putative site: 17271945$^{\#\$}$, closest gene: CG6380 (17291075 – 17292202)



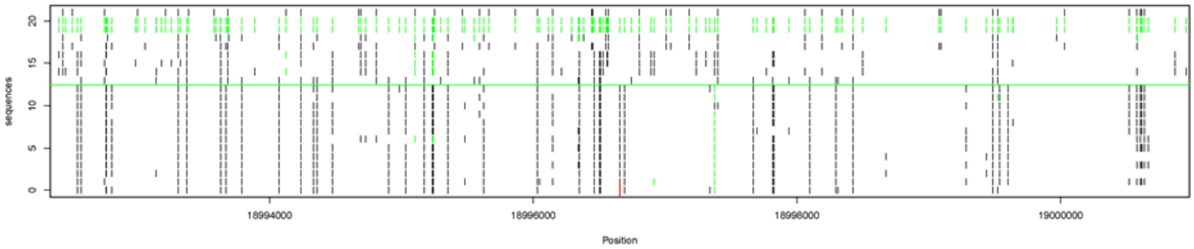
Putative site: 17616351$^{\#\$}$, closest gene: Sytalpha (17592260 – 17604387)



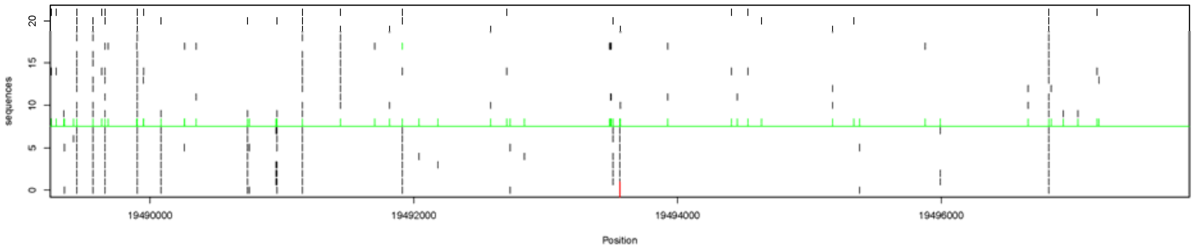
Putative site: 18453145, closest gene: bsf (18449517 – 18454587)



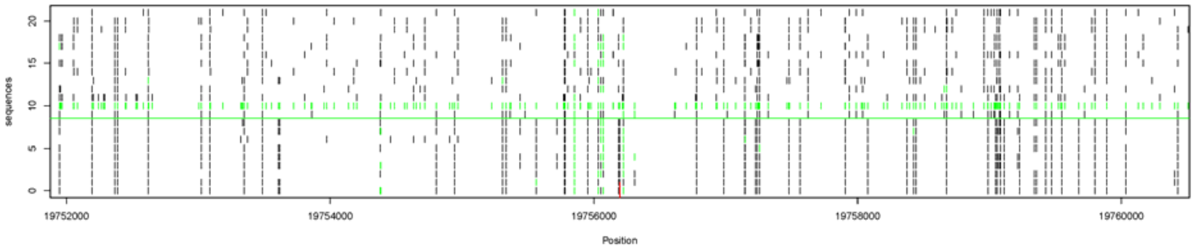
Putative site: 18996657, closest gene: CG10650 (18993360 – 18995934)



Putative site: 19493563, closest gene: swm (19493251 – 19497978)

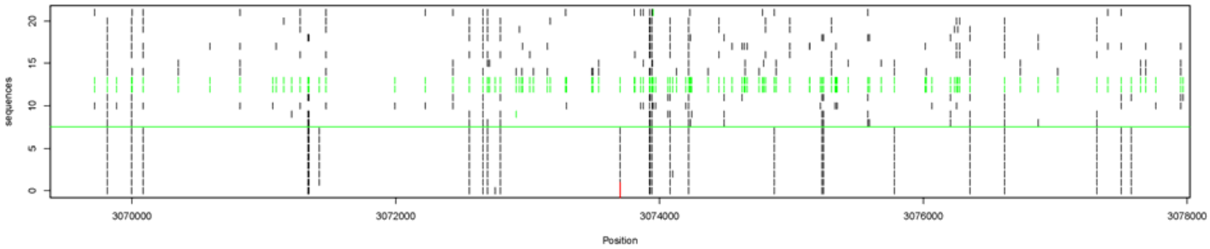


Putative site: 19756197[#$], closest gene: CG10631 (19742817 – 19756904)

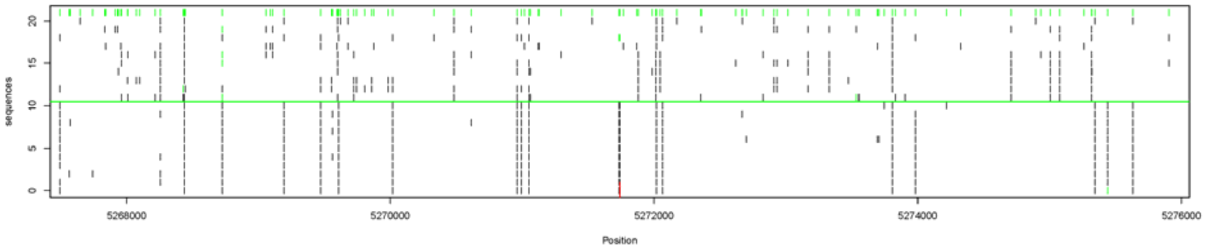


2R Patterns:

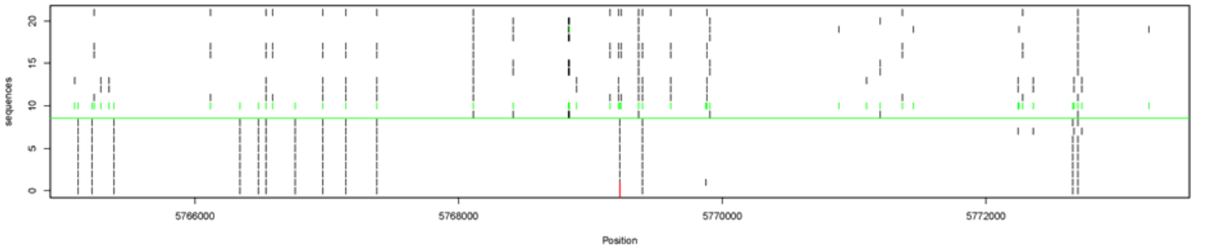
Putative site: 3073701, closest gene: diddum (3387652 – 3396130)



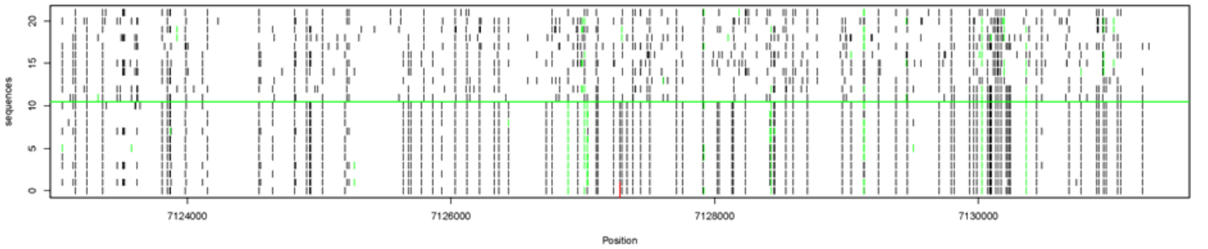
Putative site: 5271741$^{\$}$, closest gene: CG13954 (5196801 – 5276972)



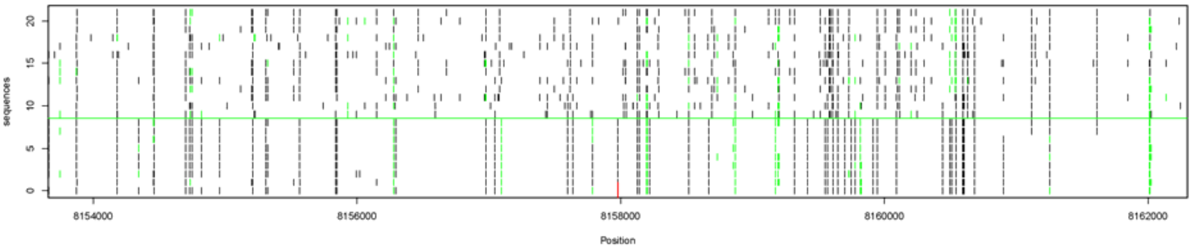
Putative site: 5769223, closest gene: Sec24AB (5763737 – 5769862)



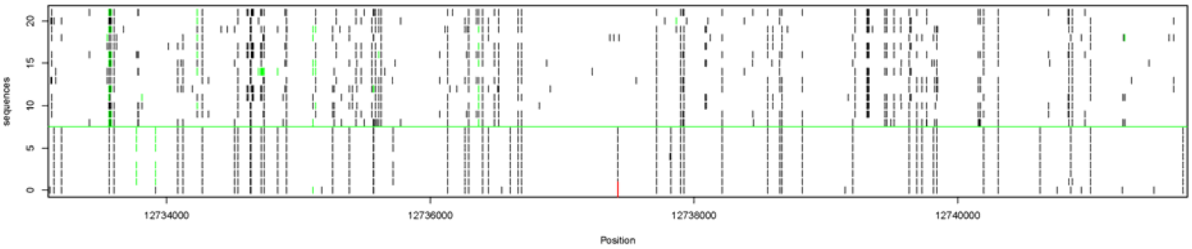
Putative site: 7127281$^{\#\$}$, closest gene: CG13215 (7126999 – 7127619)



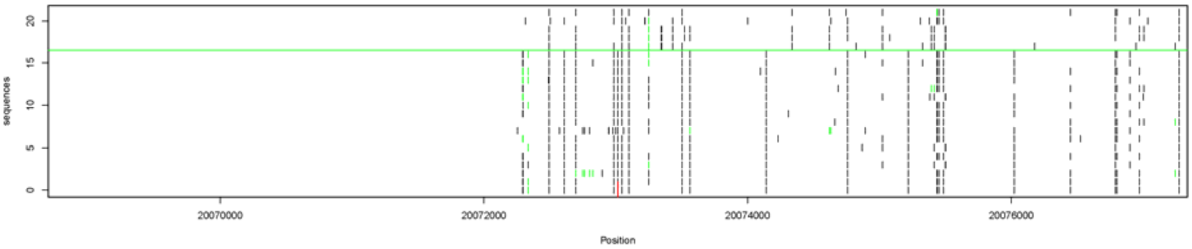
Putative site: 8157979$^{\#\$}$, closest gene: otk (7888978 – 7907351)



Putative site: 12737423*$^{\#\$}$, closest gene: IntS8 (12737942 – 12741609)

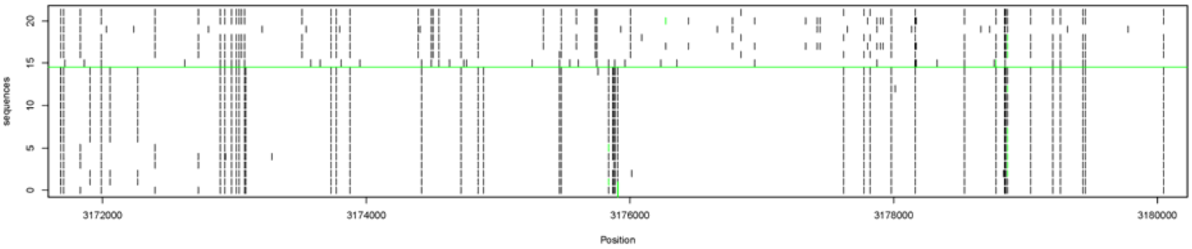


Putative site: 20073016*, closest gene: Nop60B (20062400 – 20073866)

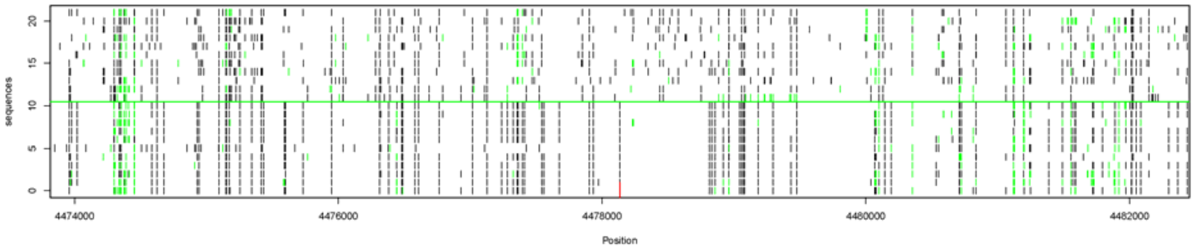


3L Patterns:

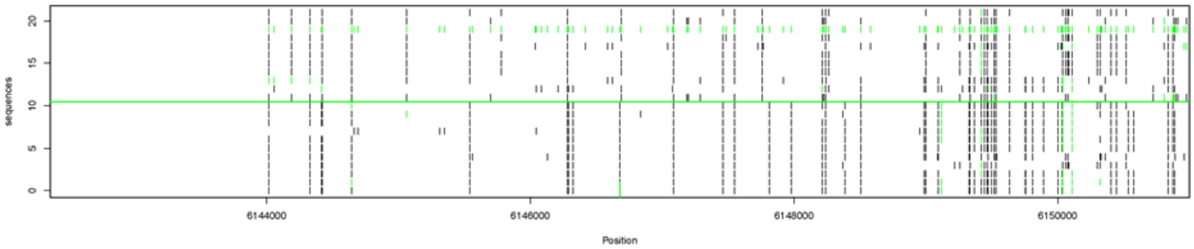
Putative site: 3175908*, closest gene: Girdin (3178930 – 3185287)



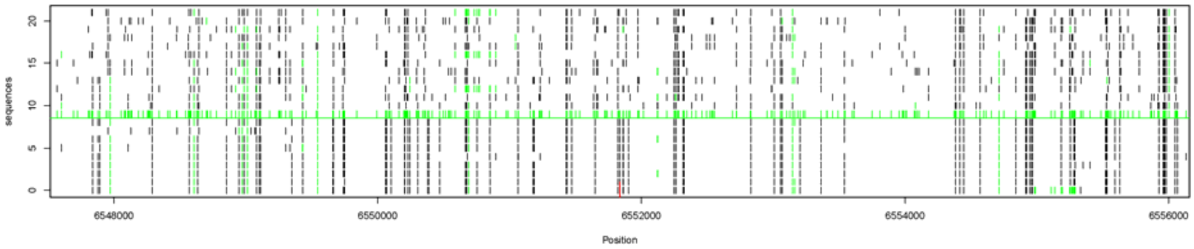
Putative site: 4478135*$^{\#}$, closest gene: CG7465 (4480283 – 4481487)



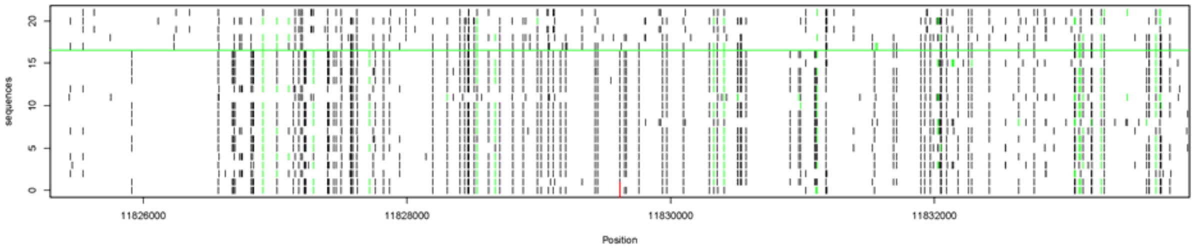
Putative site: 6146679*$^{\#\$}$, closest gene: Lcp65Ag2 (6126090 – 6126693)



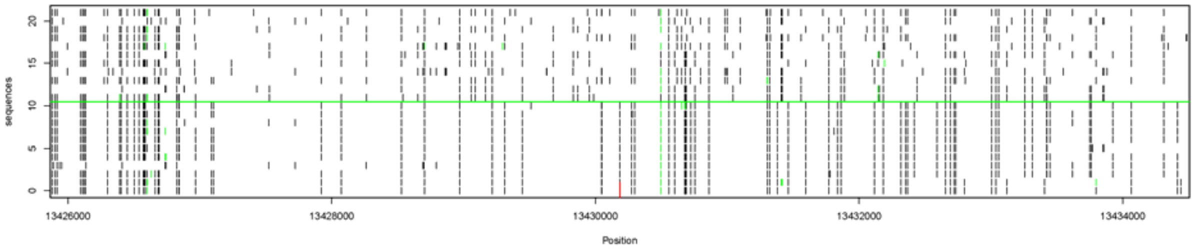
Putative site: 6551837*$^{\#\$}$, closest gene: CG18769 (6543838 – 6587040)



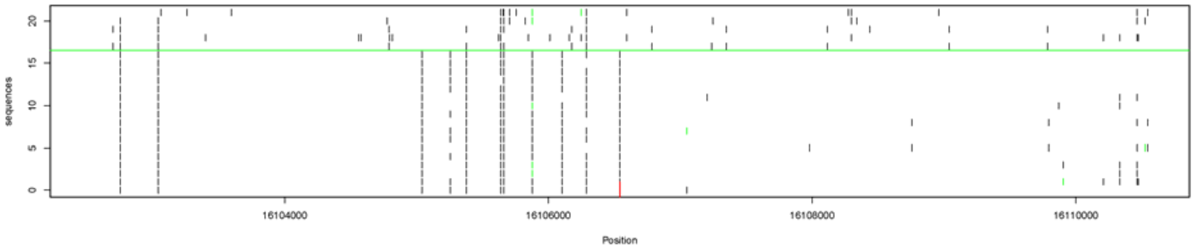
Putative site: 11829615*$^{\$}$, closest gene: CG43064 (11828293 – 11829821)



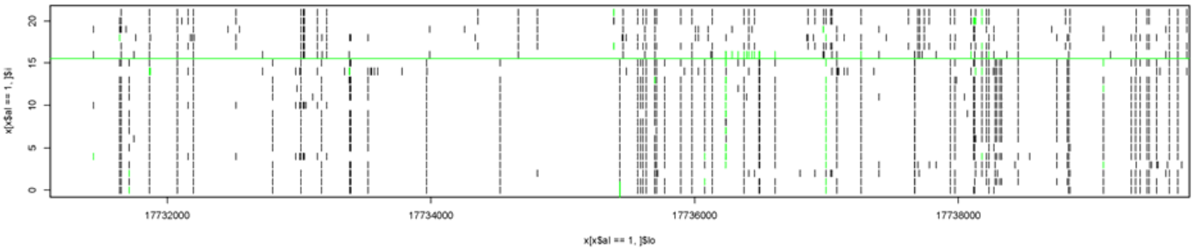
Putative site: 13430186$^{*\$}$, closest gene: CG10713 (13421939 – 13428329)



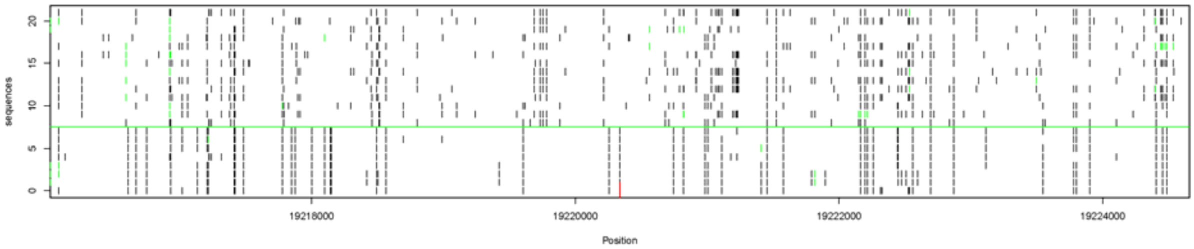
Putative site: 16106542$^{\#\$}$, closest gene: Taf4 (16106312 – 16114751)



Putative site: 17735433, closest gene: CG7460 (17733640 – 17735640)

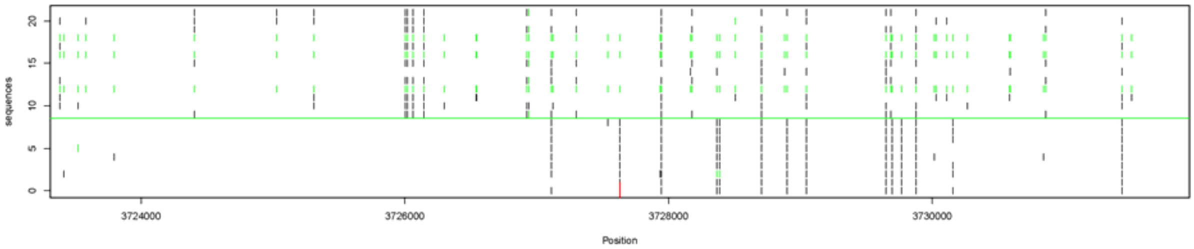


Putative site: 19220338$^{\#\$}$, closest gene: fz2 (19134075 – 19228473)

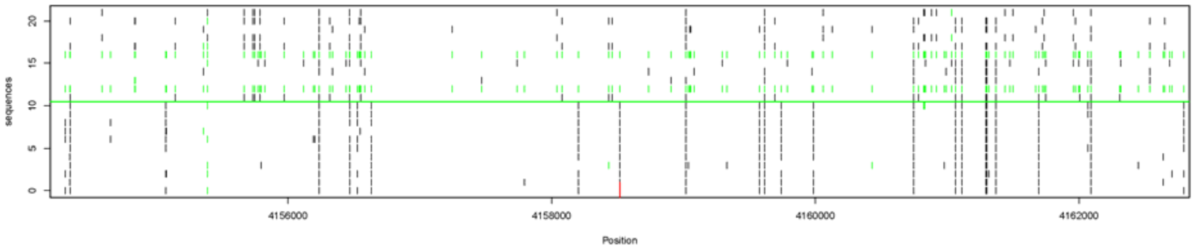


3R Patterns:

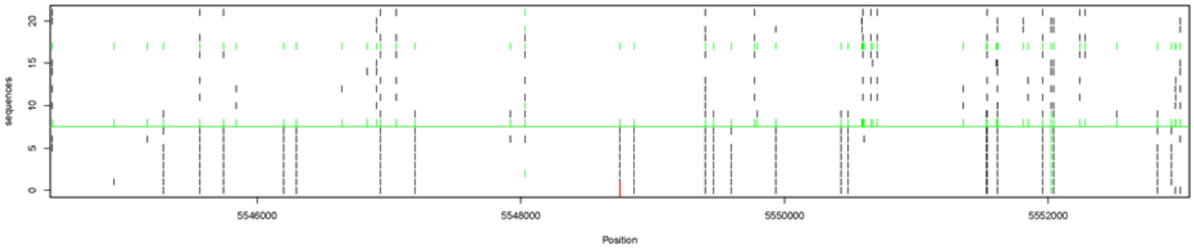
Putative site: 3727631, closest gene: mRpS9 (3714999 – 3728389)



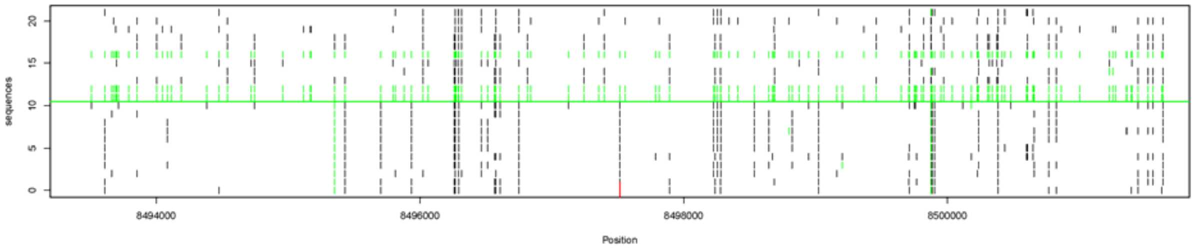
Putative gene: 4158518, closest gene: CG9601 (4167383 – 4169238)



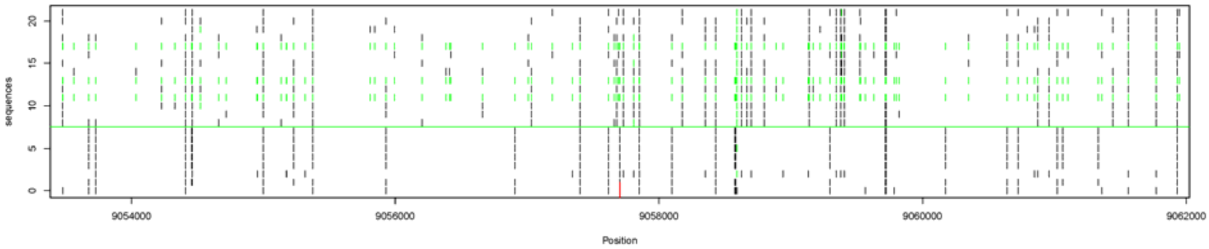
Putative site: 5548751, closest gene: CG8478 (5589372 – 5591857)



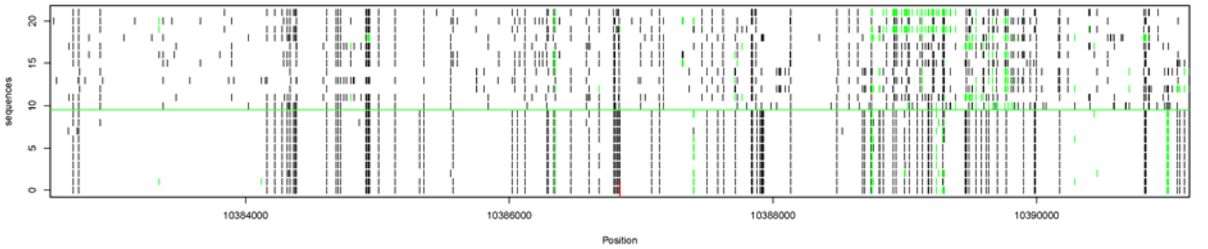
Putative site: 8497516, closest gene: CG14395 (8488553 – 8499681)



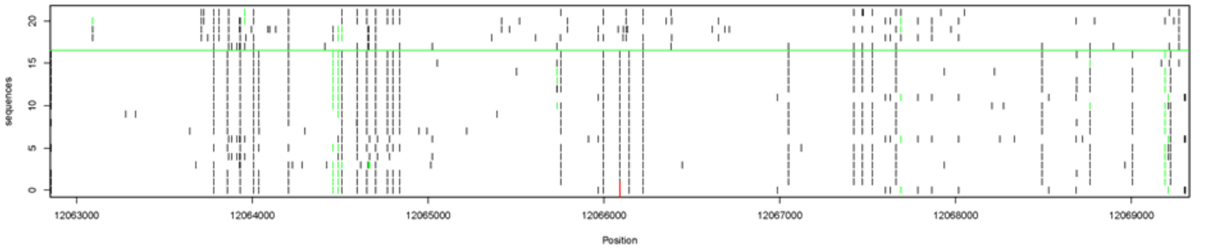
Putative site: 9057704$^{\#}$, closest gene: Ace (9048673 – 9085239)



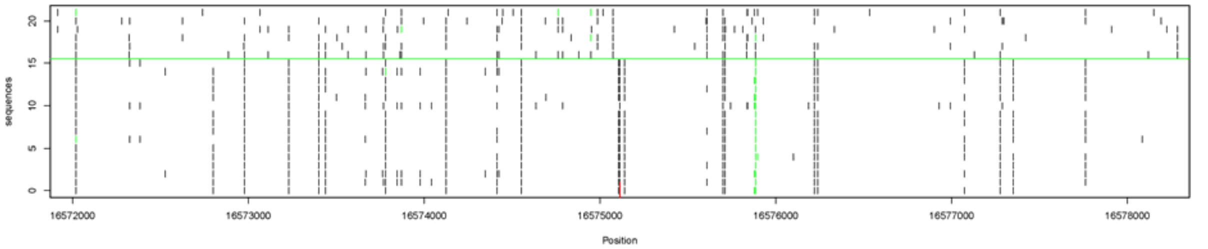
Putative site: 10386839$^{\#\$}$, closest gene: Pde6 (10339623 – 10384026)



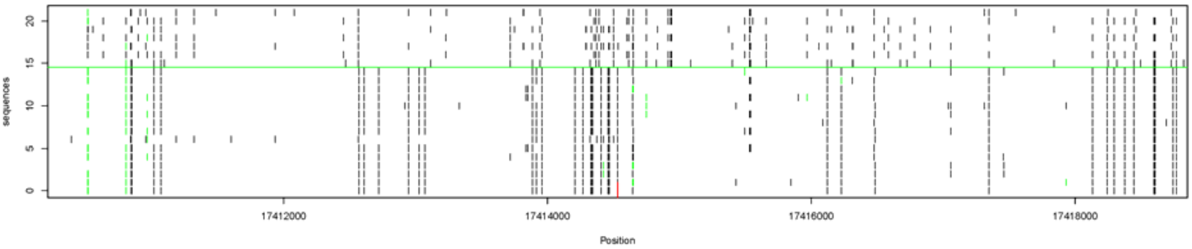
Putative site: 12066090$^{\#\$}$, closest gene: tara (12051373 – 12086051)



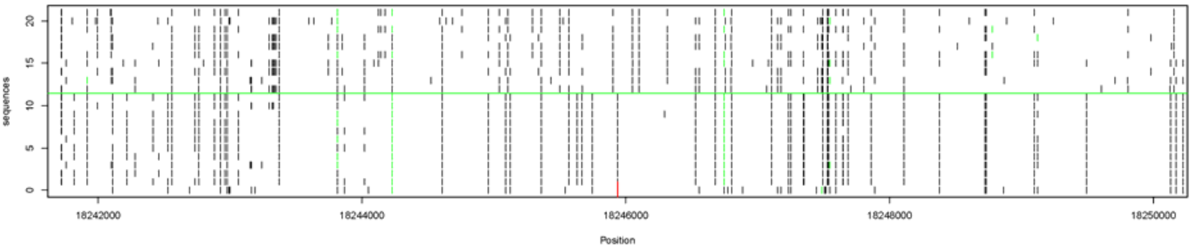
Putative site: 16575113$^{*\$}$, closest gene: CG42322 (16565830 – 16582361)



Putative site: 17414532*, closest gene: InR (17395970 – 17445043)

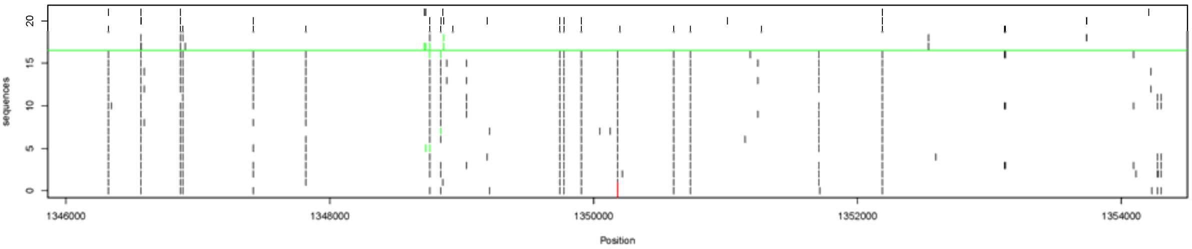


Putative site: 18245938*, closest gene: IqfR (18237023 – 18244773)

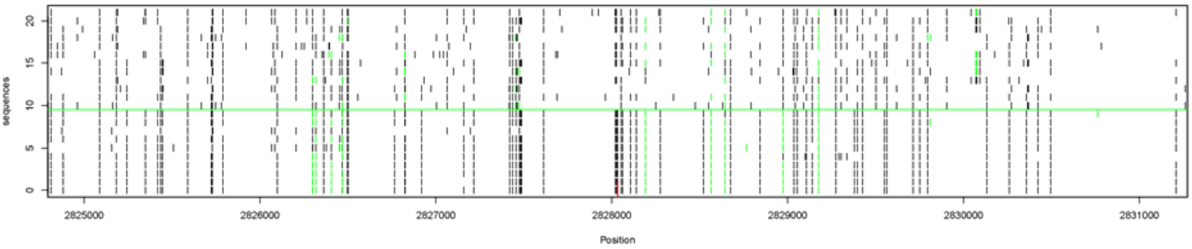


X Patterns:

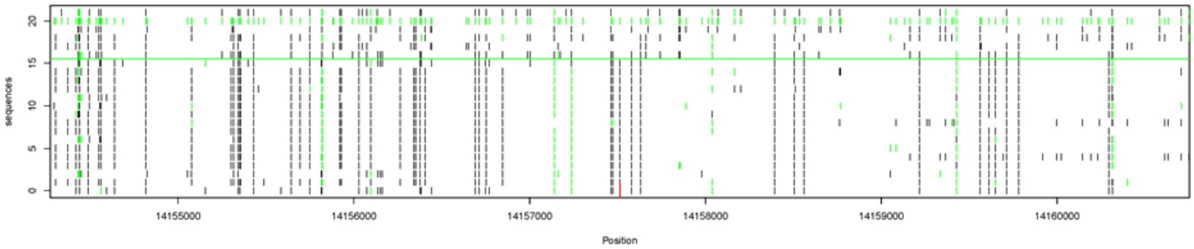
Putative site: 1350182$^{\#}$, closest gene: MED18 (1759942 – 1760920)



Putative site: 2828033$^{\#\$}$, closest gene: kirre (2634417 – 3028565)



Putative site: 14157513*$^{\$}$, closest gene: CG1461 (14155256 – 14159412)



Putative site: 15620351*, closest gene: CG8184 (15606661 – 15625968)

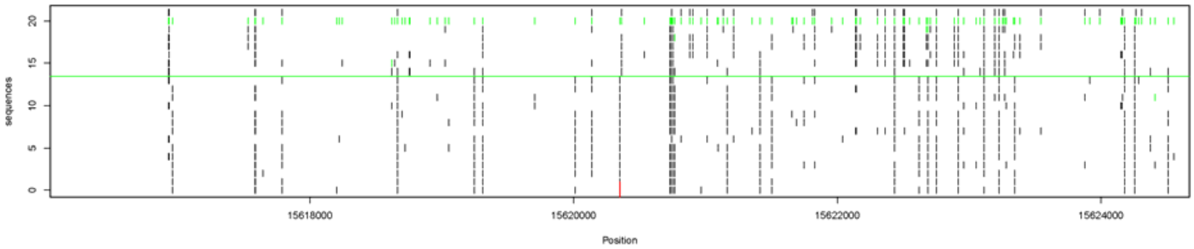
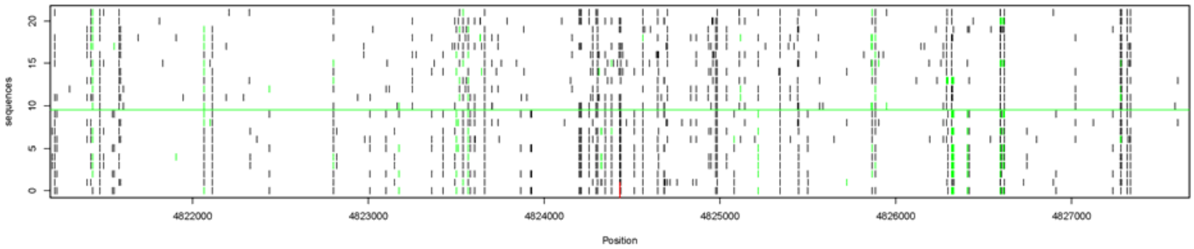


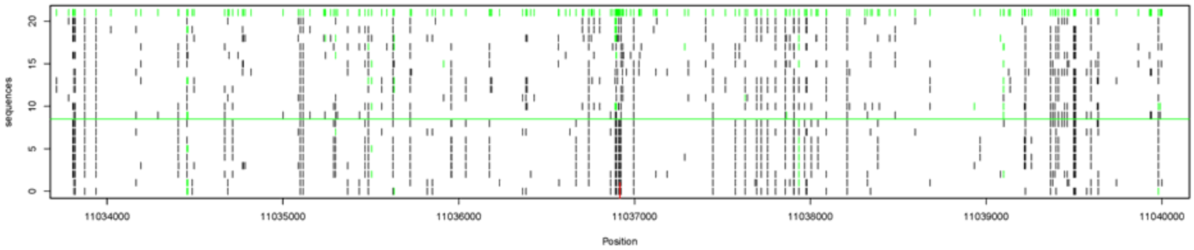
Figure S7 legend: Polymorphism pattern surrounding the site of strongest signal detected by CLR test within each cluster listed in Table 2. In each figure, 22 chromosomes are aligned (number from 0 to 21 vertically) and the putative site under selection (site with strongest signal) is located in the middle (red tick on horizontal axis). Chromosomes are arranged below or above a green line according to allele type (derived or ancestral, respectively) at the putative site. Derived alleles and missing base calls at polymorphic sites are represented by black and green bars, respectively. Whether each region overlaps with a candidate region of complete selective sweep, with a cluster detected by *iHS* test, and by nS_L test are indicated by *, $^{\#}$, and $^{\$}$, respectively.

Figure S8

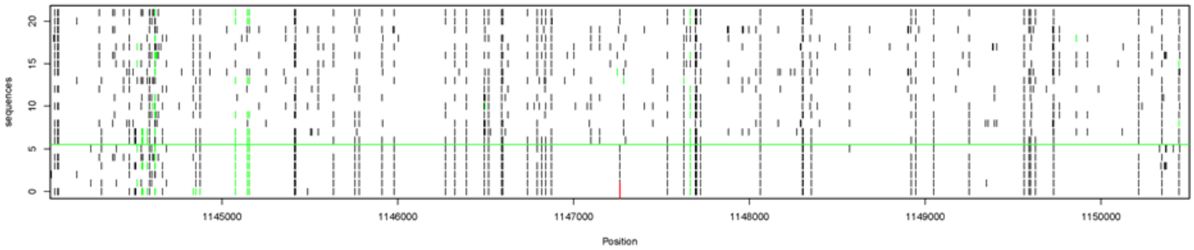
Chromosome: 2L, putative site: 4824431, detected by *iHS* (-4.03)



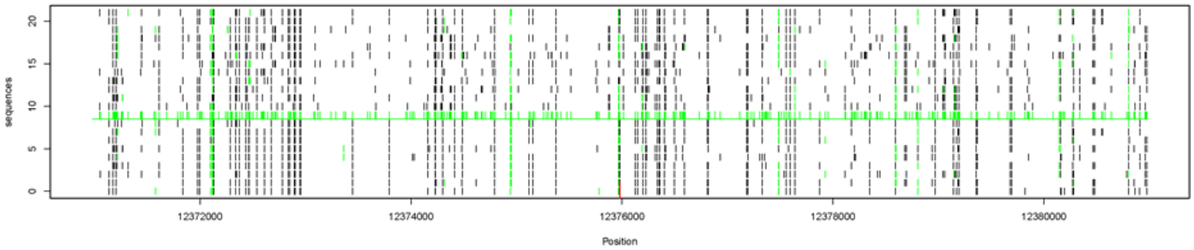
Chromosome: 2L, putative site: 11036917, detected by *iHS* (-3.82)



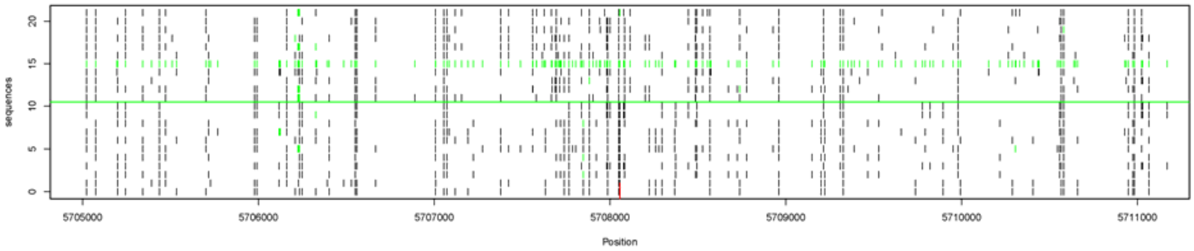
Chromosome: 2L, putative site: 1147263, detected by nS_L (-3.77)



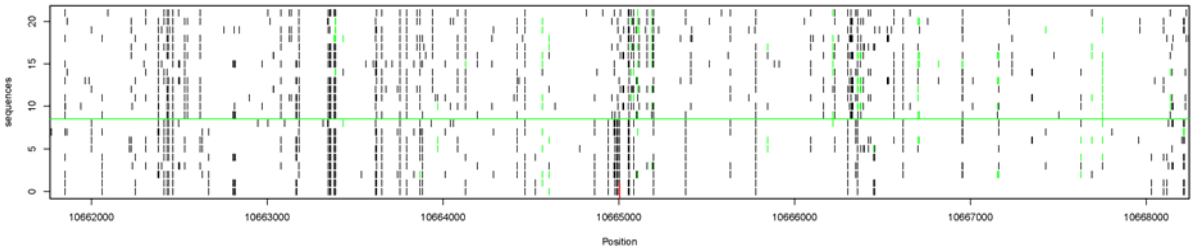
Chromosome: 2L, putative site: 12375980, detected by nS_L (-3.46)



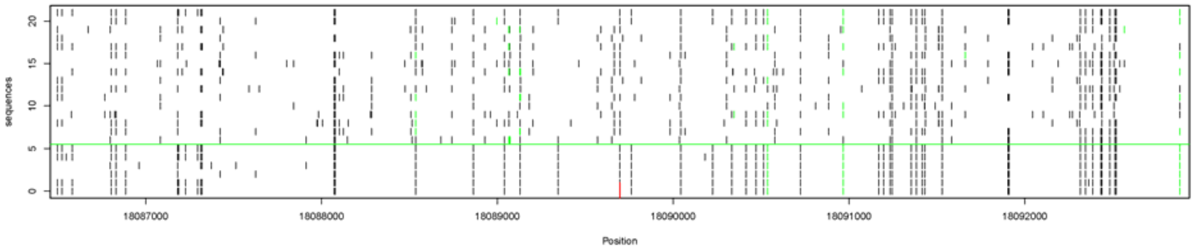
Chromosome: 2R, putative site: 5708056, detected by *iHS* (-3.33)



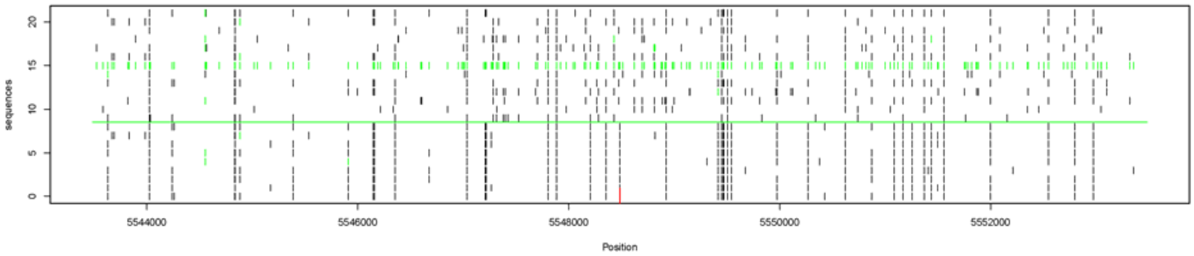
Chromosome: 2R, putative site: 10665005, detected by *iHS* (-3.86)



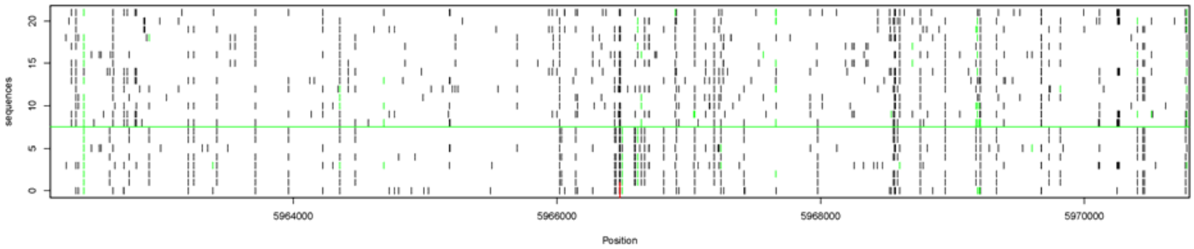
Chromosome: 2R, putative site: 18089697, detected by nS_L (-2.80)



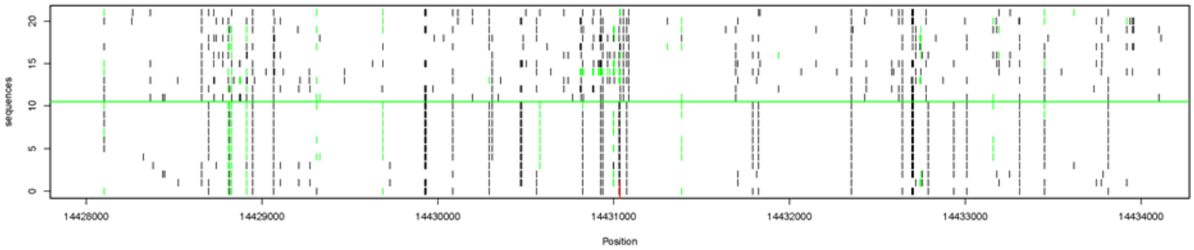
Chromosome: 2R, putative site: 5548485, detected by nS_L (-2.42)



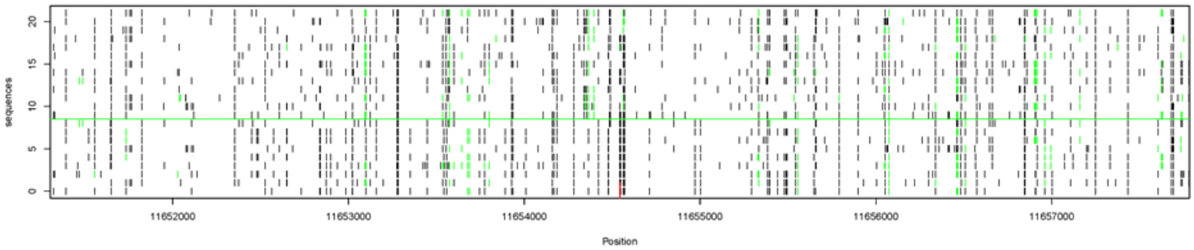
Chromosome: 3L, putative site: 5966477, detected by *iHS* (-4.18)



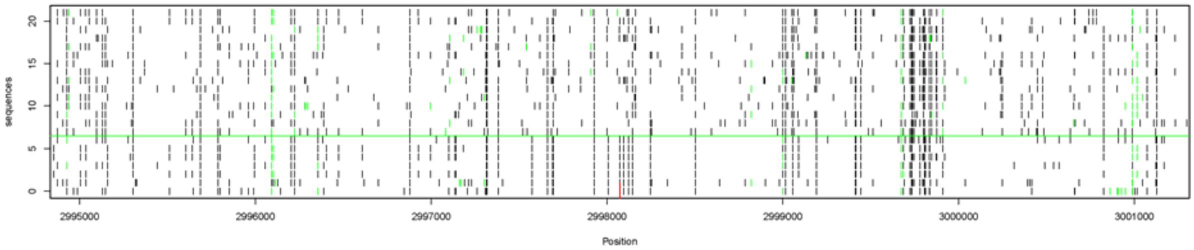
Chromosome: 3L, putative site: 14431036, detected by *iHS* (-4.12)



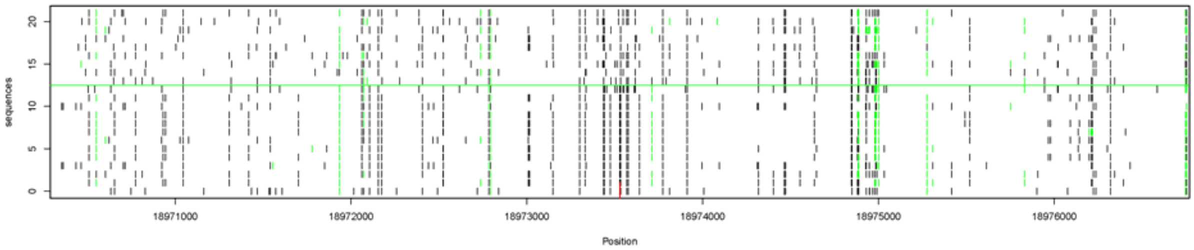
Chromosome: 3L, putative site: 11654544, detected by nS_L (-3.59)



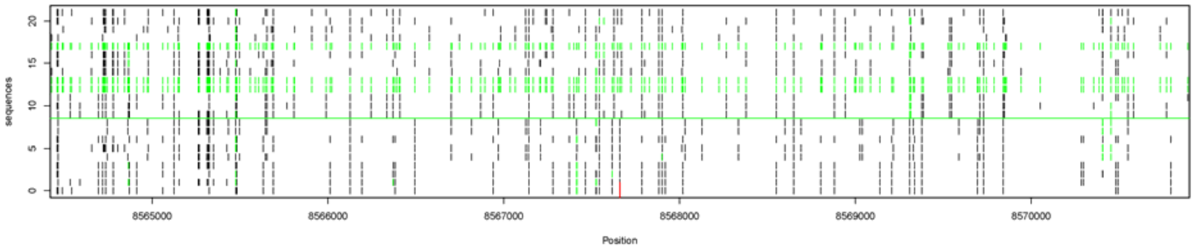
Chromosome: 3L, putative site: 2998073, detected by nS_L (-3.82)



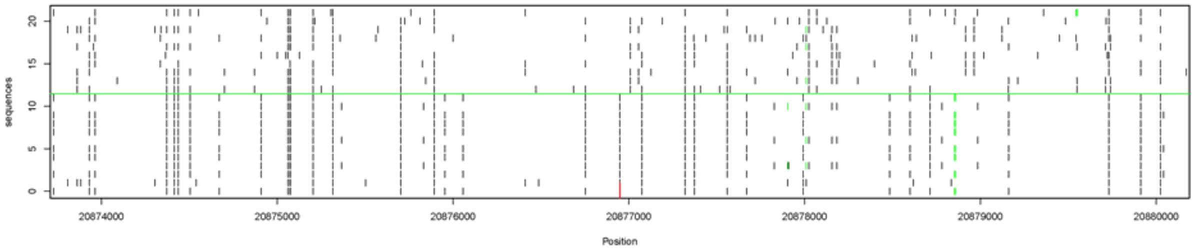
Chromosome: 3R, putative site: 18973529, detected by *iHS* (-3.49)



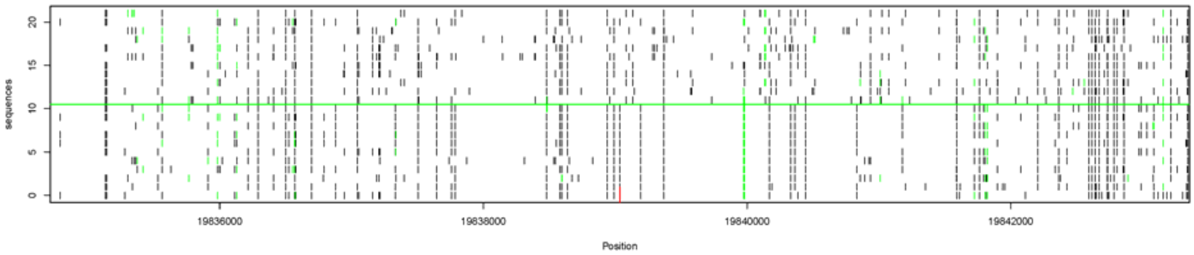
Chromosome: 3R, putative site: 8567660, detected by *iHS* (-3.70)



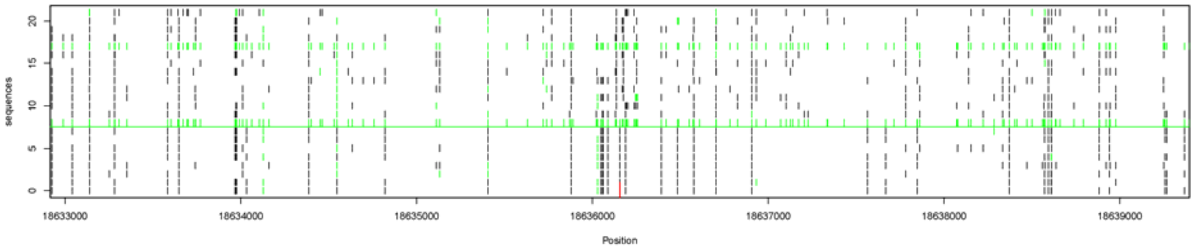
Chromosome: 3R, putative site: 20876949, detected by nS_L (-3.49)



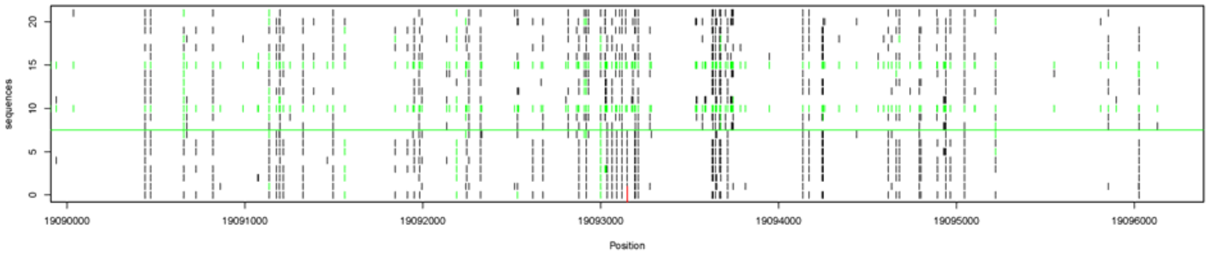
Chromosome: 3R, putative site: 19839035, detected by nS_L (-3.17)



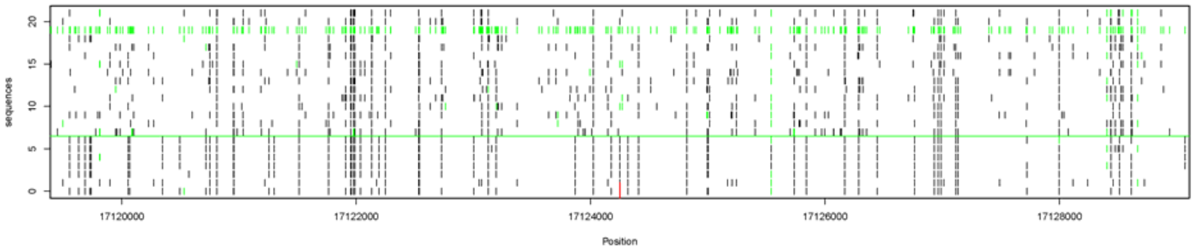
Chromosome: X, putative site: 18636156, detected by *iHS* (-5.23)



Chromosome: X, putative site: 19093150, detected by *iHS* (-5.80)



Chromosome: X, putative site: 17124249, detected by nS_L (-4.25)



Chromosome: X, putative site: 14975977, detected by nS_L (-4.16)

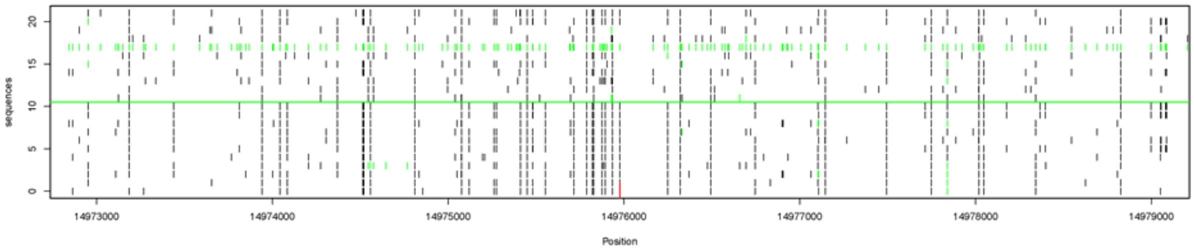


Figure S8 legend: Polymorphism patterns of genome areas surrounding the putative site under selection detected exclusively by *iHS* or *nSL* method. For each chromosome arm, top two candidate loci with strongest signals by each method, however not significant by the other tests, are shown.

Figure S9

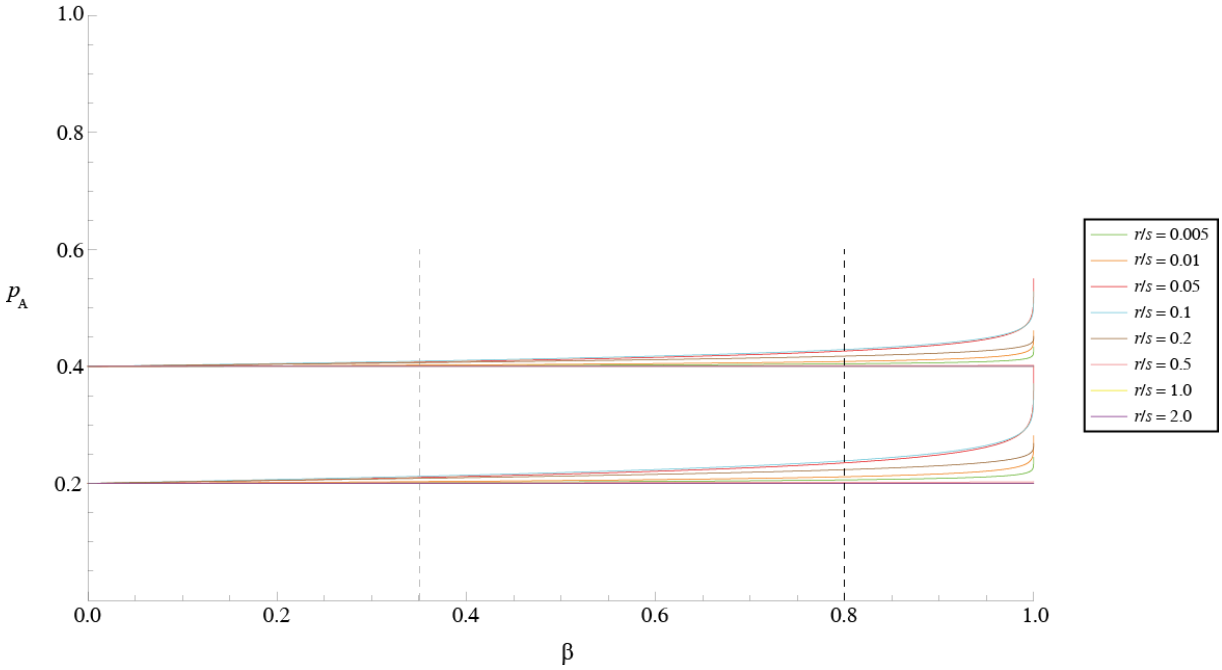


Figure S9 Legend: Deterministic changes in p_A, the frequency of a linked neutral derived allele in the subpopulation of chromosomes carrying the ancestral allele of the *S* locus during the course of a selective sweep. β is the frequency of the beneficial mutation in the population. The frequencies were obtained from equation (12b) of Stephan et al. (1992) for two different values of *p* (frequency of neutral derived allele at the beginning of sweep): 0.2 and 0.4, with different values of *r/s* (recombination rate/selection coefficient). Dashed lines (at β = 0.35 and β = 0.8) mark the interval of beneficial allele frequency at the *S* locus for which composite likelihood test is performed.