S/HIC: Robust identification of soft and hard sweeps using machine learning

Daniel R. Schrider[1,*] and Andrew D. Kern[1,2]

[1]Department of Genetics, Rutgers University, Piscataway, NJ 08854

[2]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, NJ 08854

[*]Corresponding author: dan.schrider@rutgers.edu

ABSTRACT

Detecting the targets of adaptive natural selection from whole genome sequencing data is a central problem for population genetics. However, to date most methods have shown sub-optimal performance under realistic demographic scenarios. Moreover, over the past decade there has been a renewed interest in determining the importance of selection from standing variation in adaptation of natural populations, yet very few methods for inferring this model of adaptation at the genome scale have been introduced. Here we introduce a new method, S/HIC, which uses supervised machine learning to precisely infer the location of both hard and soft selective sweeps. We show that S/HIC has unrivaled accuracy for detecting sweeps under demographic histories that are relevant to human populations, and distinguishing sweeps from linked as well as neutrally evolving regions. Moreover we show that S/HIC is uniquely robust among its competitors to model misspecification. Thus even if the true demographic model of a population differs catastrophically from that specified by the user, S/HIC still retains impressive discriminatory power. Finally we apply S/HIC to the case of resequencing data from human chromosome 18 in a European population sample and demonstrate that we can reliably recover selective sweeps that have been identified earlier using less specific and sensitive methods.

INTRODUCTION

The availability of population genomic data has empowered efforts to uncover the selective, demographic, and stochastic forces driving patterns of genetic variation within species. Chief among these are attempts to uncover the genetic basis of recent adaptation (Akey 2009). Indeed, recent advances in genotyping and sequencing technologies have been accompanied by a proliferation of statistical methods for identifying recent positive selection (see Wollstein and Stephan 2015 for recent review).

Most methods for identifying positive selection search for the population genetic signature of a "selective sweep" (Berry et al. 1991), wherein the rapid fixation of a new beneficial allele leaves a valley of diversity around the selected site (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Stephan et al. 1992), about which every individual in the population exhibits the same haplotype (i.e. the genetic background on which the beneficial mutation occurred). At greater genetic distances, polymorphism recovers as recombination frees linked neutral variants from the homogenizing force of the sweep (Kaplan et al. 1989). This process also produces an excess of low- and high-frequency derived alleles (Braverman et al. 1995; Fay and Wu 2000), and increased allelic association, or linkage disequilibrium (LD), on either side of the sweep (Kelly 1997), but not across the two flanks of the sweep (Kim and Nielsen 2004; Stephan et al. 2006). Selective fixation *de novo* beneficial mutations such as described by Maynard Smith and Haigh (1974) are often referred to as "hard sweeps."

More recently, population geneticists have begun to consider the impact of positive selection on previously standing genetic variants (Orr and Betancourt 2001; Hermisson and Pennings 2005). Under this model of adaptation, an allele initially evolves under drift for some time, until a change in the selective environment causes it to confer a fitness advantage and

sweep to fixation. In contrast to the hard sweep model, the selected allele is present in multiple copies prior to the sweep. Thus, because of mutation and recombination events occurring near the selected site during the drift phase, the region containing this site may exhibit multiple haplotypes upon fixation (Pennings and Hermisson 2006a). The resulting reduction in diversity is therefore less pronounced than under the hard sweep model (Innan and Kim 2004; Hermisson and Pennings 2005). For this reason such events are often referred to as "soft sweeps." Soft sweeps will not skew the allele frequencies of linked neutral polymorphisms toward low and high frequencies to the same extent as hard sweeps (Przeworski et al. 2005), and may even present an excess of intermediate frequencies (Teshima et al. 2006). This mode of selection will also have a different impact on linkage disequilibrium: LD will be highest at the target of selection rather than in flanking regions (Schrider et al. 2015). In very large populations, selection on mutations that are immediately beneficial may also produce patterns of soft sweeps rather than hard sweeps, as the adaptive allele may be introduced multiple times via recurrent mutation before the sweep completes (Pennings and Hermisson 2006b, a).

Adaptation could proceed primarily through selection on standing variation if the selective environment shifts frequently relative to the time scale of molecular evolution, and if there is enough standing variation segregating in the population on which selection may act following such a shift (Gillespie 1991; Hermisson and Pennings 2005). However, it is important to note that selection on standing variation may produce a hard sweep of only one haplotype containing the adaptive mutation if this allele is present at low enough frequency prior to sweep (Przeworski et al. 2005; Jensen 2014). In other words, the observation of hard sweeps may be consistent with selection on standing variation as well as selection on *de novo* mutations. For these and other reasons, there is some controversy over whether adaptation will result in soft

sweeps in nature (Jensen 2014). This could be resolved by methods that can accurately discriminate between hard and soft sweeps. To this end, some recently devised methods for detecting population genetic signatures of positive selection consider both types of sweeps (Peter et al. 2012; Ferrer-Admetlla et al. 2014; Garud et al. 2015). Unfortunately, it may often be difficult to distinguish soft sweeps from regions flanking hard sweeps due to the "soft shoulder" effect (Schrider et al. 2015).

Here we present a method that is able to accurately distinguish between hard sweeps, soft sweeps, regions linked to sweeps (or the "shoulders" of sweeps), and regions evolving neutrally. This method incorporates spatial patterns of a variety of population genetic summary statistics across a large genomic window in order to infer the mode of evolution governing a focal region at the center of this window. We combine many statistics used to test for selection using an Extremely Randomized Trees classifier (Geurts et al. 2006), a powerful supervised machine learning classification technique. We refer to this method as Soft/Hard Inference through Classification (S/HIC). By incorporating multiple signals in this manner S/HIC achieves inferential power exceeding that of any individual test. Furthermore, by using spatial patterns of these statistics within a broad genomic region, S/HIC is able to distinguish selective sweeps not only from neutrality, but also from linked selection with much greater accuracy than other methods. Thus, S/HIC has the potential to identify smaller candidate regions around recent selective sweeps, thereby narrowing down searches for the target locus of selection. We also show that S/HIC's reliance on large-scale spatial patterns makes it more robust to non-equilibrium demography than previous methods, even if the demographic model is misspecified during training. This is vitally important, as the true demographic history of a population sample may be unknown. Finally, we demonstrate the utility of our approach by applying it to

chromosome 18 in the CEU sample from the 1000 Genomes dataset (Altshuler et al. 2012), recovering most of the sweeps identified previously in this population through other methods.

METHODS

Combining multiple facets of genetic variation to detect soft and hard sweeps

We sought to devise a method that could not only accurately distinguish among hard sweeps, soft sweeps, and neutral evolution, but also among these modes of evolution and regions linked to hard and soft sweeps, respectively (Schrider et al. 2015). Such a method would not only be robust to the soft shoulder effect, but would also be able to more precisely delineate the region containing the target of selection by correctly classifying unselected but closely linked regions. In order to accomplish this, we sought to exploit the impact of positive selection on spatial patterns of several aspects of variation surrounding a sweep. Not only will a hard sweep create a valley of diversity centered around a sweep, but it will also create a skew toward high frequency derived alleles flanking the sweep and intermediate frequencies at further distances (Braverman et al. 1995; Fay and Wu 2000), reduced haplotypic diversity at the sweep site (Garud et al. 2015), and increased LD along the two flanks of the sweep but not between them (Kim and Nielsen 2004). For soft sweeps, these expected patterns may differ considerably (Przeworski et al. 2005; Pennings and Hermisson 2006a; Schrider et al. 2015), but also depart from the neutral expectation.

While some of these patterns of variation have been used individually for sweep detection (e.g. Kim and Nielsen 2004; Nielsen et al. 2005), we reasoned that by combining spatial patterns of multiple facets of variation we would be able to do so more accurately. To this end, we designed a classifier that leverages spatial patterns of a variety of population genetic

summary statistics in order to infer whether a large genomic window recently experienced a selective sweep at its center. We accomplished this by partitioning this large window into adjacent subwindows, measuring the values of each summary statistic in each subwindow, and normalizing by dividing the value for a given subwindow by the sum of values for this statistic across all subwindows. Thus, for a given summary statistic x, we used the following vector:

$$\left[\frac{x_1}{\sum_i x_i}\ \frac{x_2}{\sum_i x_i}\ \cdots\ \frac{x_n}{\sum_i x_i}\right]$$

where the larger window has been divided into n subwindows, and x_i is the value of the summary statistic x in the i^{th} subwindow. Thus, this vector captures differences in the relative values of a statistic across space within a large genomic window, but does not include the actual values of the statistic. In addition to allowing for discrimination between sweeps and linked regions, this strategy was motivated by the need for accurate sweep detection in the face of a potentially unknown nonequilibrium demographic history, which may grossly affect values of these statistics but may skew their expected spatial patterns to a much lesser extent. In total, we constructed these vectors for each of π (Nei and Li 1979), $\hat{\theta}_w$ (Watterson 1975), $\hat{\theta}_H$ (Fay and Wu 2000), the number of distinct haplotypes, average haplotype homozygosity, H_{12} and H_2/H_1 (Messer and Petrov 2013; Garud et al. 2015), Z_{nS} (Kelly 1997), and ω (Kim and Nielsen 2004). Thus, we represent each large genomic window by the following vector, to which we refer as the feature vector:

$$\left[\frac{\pi_1}{\sum_i \pi_i}\ \frac{\pi_2}{\sum_i \pi_i}\ \cdots\ \frac{\pi_n}{\sum_i \pi_i}\ \frac{\hat{\theta}_{w_1}}{\sum_i \hat{\theta}_{w_i}}\ \frac{\hat{\theta}_{w_2}}{\sum_i \hat{\theta}_{w_i}}\ \cdots\ \frac{\hat{\theta}_{w_n}}{\sum_i \hat{\theta}_{w_i}}\ \frac{\hat{\theta}_{H_1}}{\sum_i \hat{\theta}_{H_i}}\ \frac{\hat{\theta}_{H_2}}{\sum_i \hat{\theta}_{H_i}}\ \cdots\ \frac{\hat{\theta}_{H_n}}{\sum_i \hat{\theta}_{H_i}}\ \cdots\cdots\cdots\ \frac{\omega_1}{\sum_i \omega_i}\ \frac{\omega_2}{\sum_i \omega_i}\ \cdots\ \frac{\omega_n}{\sum_i \omega_i}\right]$$

We sought to discriminate between hard sweeps, regions linked to hard sweeps, soft sweeps, regions linked to soft sweeps, and neutrally evolving regions on the basis of the values of the vectors defined above. Because there is no analytical expectation of the values of these statistics at varying distances from a sweep, we opted to use a supervised machine learning framework, wherein a classifier is trained from regions known to belong to one of these five classes. We trained an Extra-Trees classifier (or extremely randomized forest; Geurts et al. 2006) from coalescent simulations (described below) in order to classify large genomic windows as experiencing a hard sweep in the central subwindow, a soft sweep in the central subwindow, being closely linked to a hard sweep, being closely linked to a soft sweep, or evolving neutrally according to the values of its feature vector (Fig 1).

Briefly, the Extra-Trees classifier is an ensemble classification technique that harnesses a large number classifiers referred to as decision trees. A decision tree is a simple classification tool that uses the values of multiple features for a given data instance, and creates a branching tree structure where each node in the tree is assigned a threshold value for a given feature. If a given data point's (or instance's) value of the feature at this node is below the threshold, this instance takes the left branch, and otherwise it takes the right. At the next lowest level of the tree, the value of another feature is examined. When the data instance reaches the bottom of the tree, it is assigned a class inference based on which leaf it has landed (Quinlan 1986). Typically, a decision tree is built according to an algorithm designed to optimize its accuracy (Quinlan 1986). The Extra-Trees classifier, on the other hand, builds a specified number of semi-randomly generated decision trees. Classification is then performed by simply taking the class receiving the most "votes" from these trees (Geurts et al. 2006), building on the strategy of random forests

(Breiman 2001). While individual decision trees may be highly inaccurate, the practice of aggregating predictions from many semi-randomly generated decision trees has been proved to be quite powerful (Ho 1995).

In the following sections we describe our methodology for training, testing, and applying our Extra-Trees classifier for identifying positive selection. We also experimented with support vector machines (SVMs; Cortes and Vapnik 1995), but found that for this classification problem the Extra-Trees classifier slightly but consistently outperformed SVMs (data not shown).

Coalescent simulations for training and testing

We simulated data for training and testing of our classifier using our coalescent simulator, discoal_multipop (https://github.com/kern-lab/discoal_multipop). As discussed in the Results, we simulated training sets with different demographic histories (Supplemental Table S1), and, for positively selected training examples, different ranges of selection coefficients (α=2Ns, where s is the selective advantage and N is the population size). For each combination of demographic history and range of selection coefficients, we simulated large chromosomal windows that we later subdivided into 11 adjacent and equally sized subwindows. We then simulated training examples with a hard selective sweep whose selection coefficient was uniformly drawn from the specified range, $U(\alpha_{\text{low}}, \alpha_{\text{high}})$. We generated 11,000 sweeps: 1000 where the sweep occurred in the center of the leftmost of the 11 subwindows, 1000 where the sweep occurred in the second subwindow, and so on. We repeated this same process for soft sweeps at each location; these simulations had an additional parameter, the derived allele frequency, f, at which the mutation switches from evolving under drift to sweeping to fixation, which we drew from $U(0.05, 0.2)$. For our equilibrium demography scenario, we drew the fixation time of the selective sweep from

$U(0, 0.2)\times N$ generations ago, while for non-equilibrium demography the sweeps completed more recently (see below). We also simulated 1000 neutrally evolving regions. Unless otherwise noted, for each simulation the sample size was set to 100 chromosomes.

For each combination of demographic scenario and selection coefficient, we combined our simulated data into 5 equally-sized training sets (Fig 1): a set of 1000 hard sweeps where the sweep occurs in the middle of the central subwindow (i.e. all simulated hard sweeps); a set of 1000 soft sweeps (all simulated soft sweeps); a set of 1000 windows where the central subwindow is linked to a hard sweep that occurred in one of the other 10 windows (i.e. 1000 simulations drawn randomly from the set of 10000 simulations with a hard sweep occurring in a non-central window); a set of 1000 windows where the central subwindow is linked to a soft sweep (1000 simulations drawn from the set of 10000 simulations with a flanking soft sweep); and a set of 1000 neutrally evolving windows unlinked to a sweep. We then generated a replicate set of these simulations for use as an independent test set.

Training the Extra-Trees classifier

We used the python scikit-learn package (http://scikit-learn.org/) to train our Extra-Trees classifier and to perform classifications. Given a training set, we trained our classifier by performing a grid search of multiple values of each of the following parameters: max_features (the maximum number of features that could be considered at each branching step of building the decision trees, which was set to 1, 3, $\sqrt{n}$, or n, where n is the total number of features); max_depth (the maximum depth a decision tree can reach; set to 3, 10, or no limit), min_samples_split (the minimum number of training instances that must follow each branch when adding a new split to the tree in order for the split to be retained; set to 1, 3, or 10);

min_samples_leaf. (the minimum number of training instances that must be present at each leaf in the decision tree in order for the split to be retained; set to 1, 3, or 10); bootstrap (whether or not bootstrap sampling is used to create decision trees); criterion (the criterion used to assess the quality of a proposed split in the tree, which is set to either Gini impurity or to information gain). The number of decision trees included in the forest was always set to 100. After performing a grid-search with 10-fold cross validation in order to identify the optimal combination of these parameters, we used this set of parameters to train the final classifier.

Comparisons with other methods

We compared the performance of our classifier to that of various other methods. First, we examined two population genetic summary statistics: Tajima's *D* (1983) and Kim and Nielsen's ω (2004), calculating their values in each subwindow within each large simulated chromosome that we generated for testing (see above). We also used Nielsen et al.'s composite likelihood ratio test, referred to as CLR or SweepFinder (2005), which searches for the spatial skew in allele frequencies expected surrounding a hard selective sweep. When testing SweepFinder's ability to discriminate between modes of evolution within larger regions, we computed the composite-likelihood ratio between the sweep and neutral models for a site located at the center of each of the 11 subwindows of our large simulated test regions. The only training necessary for SweepFinder was to specify the neutral site frequency spectrum.

Next, we used scikit-learn to implement Ronen et al.'s (2013) *SFselect*, a support vector machine classifier that discriminates between selection and neutrality on the basis of a region's binned and weighted site frequency spectrum (SFS). In our implementation we collapsed the SFS into 10 bins as suggested by Ronen et al., and also added soft sweeps as a third class (in addition

to hard sweeps and neutrality), using Knerr et al.'s (1990) method for extending a binary classifier to perform multi-class classification. We trained this classifier from simulated data following the same demographic and selective scenarios used to train our own classifier, and with the same number of simulated training instances, but these simulations encapsulated much smaller regions (equivalent to the size of one of our eleven subwindows). To avoid confusion with the original *SFselect*, which only handles hard sweeps, we refer to this implementation as *SFselect+*.

Finally, we implemented a version of Garud et al.'s (2015) scan for hard and soft sweeps. Garud et al.'s method uses an Approximate Bayesian Computation-like approach to calculate Bayes Factors to determine whether a given region is more similar to a hard sweep or a soft sweep by performing coalescent simulations. For this we performed simulations with the same parameters as we used to train *SFselect+*, but generated 100,000 simulations of each scenario in order to ensure that there was enough data for rejection sampling. We then used two statistics to summarize haplotypic diversity within these simulated data: H_{12} and H_2/H_1 (Messer and Petrov 2013). All simulated regions whose vector [H_{12} H_2/H_1] lies within a Euclidean distance of 0.1 away from the vector corresponding to the data instance to be classified are then counted (Garud et al. 2015). The ratio of simulated hard sweeps to simulated soft sweeps within this distance cutoff is then taken as the Bayes Factor. We also computed Bayes Factors for discriminating between selection and neutrality in the same manner. Note that this step was not taken by Garud et al., who restricted their analysis of the *D. melanogaster* genome to only the strongest signals of positive selection, only asking whether they more closely resembled hard or soft sweeps. When testing the ability of Garud et al.'s method to distinguish selective sweeps from both linked and neutrally evolving regions, we used large simulated windows and examined values of

H_{12} and H_2/H_1 only within the subwindow that exhibited the largest value of H_{12}, in an effort to mimic their strategy of using H_{12} peaks (Garud et al. 2015).

We summarized each method's power using the receiver operating characteristic (ROC) curve, making these comparisons for the following binary classification problems: discriminating between hard sweeps and neutrality, between hard sweeps and soft sweeps, between selective sweeps (hard or soft) and neutrality, and between selective sweeps (hard or soft) and unselected regions (including both neutrally evolving regions and regions linked to selective sweeps). For each of these comparisons we constructed a balanced test set with a total of 1000 simulated regions in each class, so that the expected accuracy of a completely random classifier was 50%, and the expected area under the ROC curve (AUC) was 0.5. Whenever the task involved a class that was a composite of two or more modes of evolution, we ensured that the test set was comprised of equal parts of each subclass. For example, in the selected (hard or soft) versus unselected (neutral or linked selection) test, the selected class consisted of 500 hard sweeps and 500 soft sweeps, while the unselected class consisted of 333 neutrally evolving regions, 333 regions linked to hard sweeps, and 333 regions linked to soft sweeps (and one additional simulated region from one of these test sets randomly selected, so that the total size of the unselected test set was 1000 instances). As with our training sets, we considered the true class of a simulated test region containing a hard (soft) sweep occurring in any but the central subwindow to be hard-linked (soft-linked)—even if the sweep occurred only one subwindow away from the center.

The ROC curve is generated by measuring performance at increasingly lenient thresholds for discriminating between the two classes. We therefore required each method to output a real-valued measure proportional to its confidence that a particular data instance belongs the first of

the two classes. For S/HIC, we used the posterior classification probability from the Extra-Trees classifier obtained using scikit-learn's predict_proba method. For *SFselect*+, we used the value of the SVM decision function. For SweepFinder, we used the composite likelihood ratio. For Garud et al.'s method, we used the fraction of accepted simulations (i.e. within a Euclidean distance of 0.1 from the test instance) that were of the first class: for example, for hard vs. soft, this is the number of accepted simulations that were hard sweeps divided by the number of accepted simulations that were either hard sweeps or soft sweeps. For Tajima's *D* (Tajima 1989) and Kim and Nielsen's ω (Kim and Nielsen 2004), we simply used the values of these statistics.

Simulating sweeps under non-equilibrium demographic models

We simulated training and test datasets from Tennessen et al.'s (2012) European demographic model (Table S1). This model parameterizes a population contraction associated with migration out of Africa, a second contraction followed by exponential population growth, and a more recent phase of even faster exponential growth. Values of θ and ρ=4*Nr* were drawn from prior distributions (Table S1), allowing for variation within the training data, whose means were selected from recent estimates of human mutation (Kong et al. 2012) and recombination rates (Kong et al. 2010), respectively. For simulations with selection, we drew values of α from $U(5.0\times10^3, 5.0\times10^5)$, and drew the fixation time of the sweeping allele form $U(0, 51{,}000)$ years ago (i.e. the sweep completed after the migration out of Africa).

We also generated simulations of Tennessen et al.'s African demographic model, which consists of exponential population growth beginning ~5,100 years ago (Table S1). We generated two sets of these simulations: one where α was drawn from $U(5.0\times10^4, 5.0\times10^5)$, and one with α drawn from $U(5.0\times10^4, 5.0\times10^5)$. The sample size of these simulated data sets was set to 100

chromosomes. These two sets were then combined into a single training set. For these simulations, the sweep was constrained to complete some time during the exponential growth phase (no later than 5,100 years ago). As for simulations with constant population size, when simulating soft sweeps on a previously standing variant, we drew the derived allele frequency at the onset of positive selection from $U(0.05, 0.2)$.

Application to the human chromosome 18 from the 1000 Genomes CEU sample

We applied our method to chromosome 18 from the Phase I data release from the 1000 Genomes project (Altshuler et al. 2012). We restricted this analysis to the CEU population sample (individuals with European ancestry, sampled from Utah), and trained S/HIC using data from the European demographic model described above. After training this classifier, we prepared data from chromosome 18 in CEU for classification. Prior to constructing feature vectors, we first performed extensive filtering for data quality. First, we masked all sites flagged by the 1000 Genomes Project as being unfit for population genetic analyses due to having either limited or excessive read-depth or poor mapping quality (according to the strictMask files for the Phase I data set which are available at http://www.1000genomes.org/). In order to remove additional sites lying within repetitive sequence wherein genotyping may be hindered, we eliminated sites with 50 bp read mappability scores less than one (Derrien et al. 2012) and also sites masked by RepeatMasker (http://www.repeatmasker.org). Finally, we attempted to infer the ancestral state at each remaining site, using the chimpanzee (Mikkelsen et al. 2005) and macaque (Gibbs et al. 2007) genomes as outgroups. For each site, if the chimpanzee and macaque genomes agreed, we used this nucleotide as our inferred ancestral state. If instead only the chimpanzee or the macaque genome had a nucleotide aligned to the site, we used this base as our inferred ancestral

state. For sites that were SNPs, we also required that the inferred ancestral state matched one of the two human alleles. For all cases where these criteria were not met, we discarded the site.

After data filtering, we calculated summary statistics within adjacent 200 kb windows across the entire chromosome. Windows with >50% of sites removed during the filtering processes were omitted from our analysis. For the remaining windows, we used a sliding window approach with a 2.2 Mb window and a 200 kb step size to calculate the feature vector in the same manner as for our simulated data, and then applied S/HIC to this feature vector to infer whether the central subwindow of this 2.2 Mb region contained a hard sweep, a soft sweep, was linked to a hard sweep, linked to a soft sweep, or evolving neutrally. Visualization of candidate regions was performed using the UCSC Genome Browser (Kent et al. 2002).

Software availability

Our classification tool is available at https://github.com/kern-lab/shIC, along with software for generating the feature vectors used in this paper (either from simulated training data or from real data for classification).

RESULTS

S/HIC accurately detects hard sweeps

The most basic task that a selection scan must be able to perform is to distinguish between hard sweeps and neutrally evolving regions, as the expected patterns of nucleotide diversity, haplotypic diversity, and linkage disequilibrium produced by these two modes of evolution differ dramatically (Maynard Smith and Haigh 1974; Fay and Wu 2000; Kim and Nielsen 2004; Jensen et al. 2007; Garud et al. 2015; Schrider et al. 2015). We therefore begin by comparing S/HIC's

power to discriminate between hard sweeps and neutrality to that of several previously published methods: these include SweepFinder (aka CLR; Nielsen et al. 2005), *SFselect* (Ronen et al. 2013), Garud et al.'s haplotype approach using the H_{12} and H_2/H_1 statistics (Garud et al. 2015), Tajima's D (1989), and Kim and Nielsen's ω (2004). We extended *SFselect* to allow for soft sweeps (Methods), and therefore refer to this classifier as *SFselect*+ in order to avoid confusion. We summarize the power of each of these approaches with the receiver operating characteristic (ROC) curve, which plots the method's false positive rate on the x-axis and the true positive rate on the y-axis (Methods). Powerful methods that are able to detect many true positives with very few false positives will thus have a large area under the curve (AUC), while methods performing no better than random guessing are expected to have an AUC of 0.5.

We began by assessing the ability of these tests to detect selection in populations with constant population size and no population structure. First, we used test sets where the selection coefficient $\alpha=2Ns$ was drawn uniformly from $U(2.5\times10^2, 2.5\times10^3)$, finding that S/HIC outperformed every other method except *SFselect*+—both methods had perfect accuracy (AUC=1.0; Fig S1A). We observed the same result with drawing α from $U(2.5\times10^3, 2.5\times10^4)$ (AUC=1.0; Fig S1B). For weaker selection ($\alpha \sim U(25, 2.5\times10^2)$) this classification task is more challenging, and the accuracies of most of the methods we tested dropped substantially. S/HIC, however, performed quite well, with an AUC of 0.9845, slightly better than *SFselect*+, and substantially better than the remaining methods (Fig S1C). Note that Garud et al.'s method leveraging H_{12} and H_2/H_1 performed quite poorly in these comparisons, especially in the case of weak selection. This is likely because the fixation times of the sweeps that we simulated ranged from 0 to $0.2\times N$ generations ago, and the impact of selection on haplotype homozygosity decays quite rapidly after a sweep completes (Schrider et al. 2015). If we repeat this comparison with

sweeps completing immediately prior to sampling, this method performs quite well (consistent with results from Garud et al. 2015), though not as well as our method or *SFselect*+ (data not shown).

For the above comparisons, our classifier, Garud et al.'s method, and *SFselect*+ were trained with the same range of selection coefficients used in these test sets. Thus, these results may inflate the performance of these methods relative to other methods, which do not require training from simulated selective sweeps. If one does not know the strength of selection, one strategy is to train a classifier using a wide range of selection coefficients so that it may be able to detect sweeps of varying strengths (Ronen et al. 2013). We therefore combined the three training sets from the three different ranges of α described above into a larger training set consisting of sweeps of α raging from as low as 25 to as high as 25,000. This step was done not only for S/HIC, but also for *SFselect*+ and our implementation of Garud et al.'s method, and we use this approach for the remainder of the paper when using classifiers trained from constant population size data. When trained on a large range of selection coefficients, S/HIC still detected sweeps with α drawn from $U(2.5\times10^2, 2.5\times10^3)$ with perfect accuracy, as did *SFselect*+ (Fig 2A). For stronger sweeps, we again had excellent accuracy (AUC=0.999; Fig 2B) and outperformed all other methods except *SFselect*+ (AUC=1.0). For weaker sweeps our method had the highest accuracy (AUC=0.9854, versus 0.9697 for *SFselect*+ and lower for other methods; Fig 2C). Thus, S/HIC can distinguish hard selective sweeps of greatly varying strengths of selection from neutrally evolving regions as well as if not better than previous methods.

S/HIC can uncover soft sweeps and distinguish them from hard sweeps

In order to uncover the targets of recent selective sweeps and also determine which mode of positive selection was responsible, one must be able to detect the signatures of soft as well as hard sweeps and to distinguish between them. We therefore assessed the power of each method to distinguish sets of simulated selective sweeps consisting of equal numbers of hard and soft sweeps from neutral simulations, using the same training data (for methods that require it) as for the analysis in Fig 2. For all ranges of selection coefficients, S/HIC has excellent power to distinguish hard and soft sweeps from neutrality; our AUC scores ranging from 0.9636 to 0.9898, and are higher than every other method in each scenario (Fig S2). S/HIC also distinguishes hard sweeps from soft sweeps as well as or better than each other method, except in the case of weak sweeps where *SFselect+* has slightly better power (AUC=0.7998 versus 0.8237, respectively; Fig S3).

Distinguishing sweeps from linked selection to narrow the target of adaptation

The goal of genomic scans for selective sweeps is not merely to quantify the extent to which positive selection impacts patterns of variation, but also to identify the targets of selection in hopes of elucidating the molecular basis of adaption. Unfortunately, hitchhiking events can skew patterns of variation across large chromosomal stretches often encompassing many loci. Furthermore, this problem not only confounds selection scan by obscuring the true target of selection, it may also lead to falsely inferred soft sweeps as a result of the soft shoulder effect (Schrider et al. 2015). Our goal in designing S/HIC was to be able to accurately distinguish among positive selection, linked selection, and neutrality, thereby addressing both of these challenging problems.

In order to assess the ability of our approach and other methods to perform this task, we repeated the test shown in Fig S2, but this time we included regions linked to selective sweeps among the set of neutral test instances. Thus, we ask how well these methods distinguish genomic windows containing the targets of selective sweeps (soft or hard) from neutrally evolving windows or windows closely linked to sweeps. Encouragingly, we find that S/HIC is able confidently distinguish windows experiencing selective sweeps from linked as well as neutrally evolving regions (Fig 3)—S/HIC achieves substantially higher accuracy than each other methods (AUC=0.9628 or higher for all values of α, while no other method has AUC>0.9 for any α). As the selection coefficient increases, S/HIC's performance increase relative to that of other methods is particularly pronounced (Fig 3A, B), which is unsurprising because in these cases the impact of selection on variation within linked regions is much further reaching than for weak sweeps (Fig 3C).

While ROC curves provide useful information about power, a more complete view of our ability to distinguishing among hard sweeps, soft sweeps, linked selection, and neutrality can be obtained by asking how our classifier behaves at varying distances from selective sweeps. We directly compared our method's ability to classify regions ranging from 5 subwindows upstream of a hard sweep to 5 subwindows downstream of a hard sweep to *SFselect+* (the top performer among all other methods we had examined) and Garud et al.'s method using H_{12} and H_2/H_1 (the only other method among those we examined that is designed to detect soft sweeps). For these simulations, each subwindow had a total recombination rate $4Nr$=80, corresponding to 0.2 cM per subwindow when N=10,000. We then counted the fraction of simulations predicted to belong to each of our five classes (hard, hard-linked, soft, soft-linked, and neutral) or the three classes used for *SFselect+*. As shown in Fig 4, we find that when α ranges from 250 to 2500 and there

has been a hard sweep in the central window, that the H_{12} and H_2/H_1 method often detects the sweep, but infers it to be in one of the 10 flanking windows (based on the location of the maximum H_{12} value). S/HIC and *SFSelect+* on the other hand recover the sweep with high frequency when examining the correct window (95.9% and 96.7% accuracy, respectively). However, as we move away from the selected site, a large number of windows are misclassified as hard sweeps by *SFSelect+*. For example, *SFselect+* misclassifies 47% of cases two windows away from the true sweep as hard sweeps, and almost all of the remaining examples as soft sweeps. In contrast, our method classifies <4% of these regions as sweeps, correctly classifying >93% of these windows as hard-linked instead. At a distance of 5 windows away from the sweep, *SFselect+* classifies the majority of windows as soft sweeps and many others as hard sweeps, while we classify >95% of windows as hard-linked, and <1% as sweeps of either mode. For soft sweeps, we have better sensitivity in the sweep window than *SFselect+* (79.1% versus 75.3%). We also narrow the target of selection down to a smaller region, as we classify the majority of flanking windows as soft-linked, while *SFselect+* produces many soft sweep calls in these windows.

The difference between S/HIC and *SFselect+* is amplified when testing these classifiers on stronger hard sweeps (α ranging from 2,500 to 25,000). Our classifier is better able to narrow down the selected region by classifying flanking windows as hard-linked, while *SFselect+* classifies the vast majority of simulations even 5 windows away from the target of selection as hard sweeps (Fig 5). Though here *SFselect+* has more sensitivity to detect hard sweeps when examining the correct window (99.6% versus 87.9%), as S/HIC misclassifies 11.7% of these stronger sweeps as linked-Hard. On the other hand, we recover 80.3% of soft sweeps versus 77.7% for *SFselect+*. We also misclassify relatively few linked regions as sweeps (~13% when

one window away, versus ~50% for *SFselect*+). For weaker sweeps ($\alpha \sim U(25, 250)$), the impact of selection on linked regions is reduced, and *SFselect*+ calls fewer false sweeps in linked regions. However, S/HIC has greater sensitivity than *SFselect*+ to detect both hard and soft sweeps at the correct window, and also misclassify fewer flanking regions as sweeps (Fig S4). Across the entire range of selection coefficients, S/HIC mislabeled fewer neutral simulations as sweeps (~6% for our classifier in each cases, versus >9% for *SFselect*+). Garud et al.'s method also produces relatively few false positives, but has poor sensitivity to detect hard sweeps (especially in the correct window), and even less sensitivity to soft sweeps. In summary, across all selection coefficients S/HIC has greater sensitivity than *SFselect*+ to detect soft sweeps, and also for hard sweeps except when selection is very strong. Importantly, for both types of sweeps S/HIC will identify a smaller candidate region around the selective sweep than *SFselect*+. S/HIC is able to classify far fewer linked windows as selected because it has two classes for this purpose, hard-linked and soft-linked, that *SFselect*+ lacks. Though *SFselect*+ could be improved by incorporating these classes, it may prove difficult to determine whether a window is selected or merely linked to a sweep on the basis of its SFS alone (Schrider et al. 2015), rather than examining larger scale spatial patterns of variation.

The impact of population size change and demographic misspecification

Non-equilibrium demographic histories have the potential to confound population genetic scans for selective sweeps (Simonsen et al. 1995; Wakeley and Aliacar 2001). We therefore sought to assess the power of S/HIC and other methods to detect selection occurring in populations experiencing dramatic changes in population size. To this end we trained and tested our classifiers first on a model of recent exponential population size growth (the African model from

Tennessen et al. 2012; Table S1), and then on a model of recurrent population contraction followed by first slow and then accelerated population growth (Tennessen et al's European model; Table S1). For both of these demographic histories, S/HIC has the highest accuracy of all of those that we examined (ROC curves shown in Fig S5).

A more pessimistic scenario is one where the true demographic history of the population is not known, and therefore misspecified during training. Most demographic events should impact patterns of variation genome-wide rather than smaller regions (but see Przeworski 2002; Jensen et al. 2005). Thus, approaches that search for spatial patterns of polymorphism consistent with selective sweeps may be more robust to demographic misspecification than methods examining local levels of variation only (as demonstrated by Nielsen et al. 2005). To test this, we trained S/HIC and other classifiers on equilibrium datasets, and measured their accuracy on test data simulated under the non-equilibrium demographic models described above. In Fig S6A we show the power of these classifiers to detect selective sweeps occurring under the African model of recent exponential growth. Under this scenario, with $\alpha \sim U(5.0\times10^{3}, 5.0\times10^{4})$ (equivalent to $s \sim U(6.0\times10^{-3}, 6.0\times10^{-2})$ with N=424,000), S/HIC achieves an AUC of 0.7992, while the next-highest performing method is *SFselect+* (AUC=0.7077). Similarly, we perform better than other methods when searching for stronger selection (α ranging from 5.0×10^{4} to 5.0×10^{5}; AUC=0.9846 versus <0.92 for all others; Fig S6B).

Note that the simple summary statistic methods ω and Tajima's D have some power to detect selection even under non-equilibrium demography (Fig S5). However, this result is probably quite optimistic: the ROC curve is generated by repeatedly adjusting the critical threshold and measuring true and false positive rates. In practice, a single critical threshold may be chosen to identify putative sweeps. If this critical value is chosen based on values of the

statistic generated under the incorrect demographic model, then the false positive rate may be quite high. For example, Nielsen et al. (2005) showed that when a threshold for Tajima's *D* is selected based on simulations under equilibrium, 100% of neutral simulations under a population growth model exceed this threshold. In other words, the ROC curve is useful for illustrating a method's potential power if an appropriate threshold is selected, but this may not always be the case in practice.

A more informative approach to evaluating our power may thus be to examine the fraction of regions including sweeps, linked to sweeps at various recombination distances, or evolving neutrally, that were assigned to each class (as done in Figs 4-5 for constant population size). We show this in Fig S7, which better illustrates S/HIC's power and robustness to unknown demographic history. With $\alpha \sim U(5.0\times10^4, 5.0\times10^5)$ we recover 86.7% of hard sweeps versus 85.7% for *SFselect+* and 35.3% for H_{12} and H_2/H_1 (Fig S7D-F). Moreover, while *SFselect+* classifies the vast majority of linked regions as hard or soft sweeps, we classify most of these as hard-linked, and most of the remainder as soft-linked—we classify very few linked regions as selective sweeps. This comparison yields similar results in regions flanking soft sweeps.

In the context of scans for positive selection, the primary concern with non-equilibrium demography is that it will produce a large number of false selective sweep calls. Indeed, when trained on an equilibrium demographic history and tested on the exponential growth model, *SFselect+* classifies roughly one-fourth of all neutral loci as having experienced recent positive selection. In stark contrast, S/HIC does not seem to be greatly affected by this problem: we classify only ~7% of neutrally evolving regions as sweeps. As shown in Fig S7D, we obtain similar results when examining sweeps with $\alpha \sim U(5.0\times10^4, 5.0\times10^5)$.

Next, we examined the impact of demographic misspecification on power to detect selection occurring under Tennessen et al.'s model of the population size history of Europeans following their migration out of Africa (Tennessen et al. 2012) but having trained S/HIC under the standard neutral model. This demographic history presents an even greater challenge for identifying positive selection than the African model, as it is characterized by two population contractions followed by exponential growth, and then a more recent phase of faster population growth (Methods). For this scenario, a single range of selection coefficients was used: $\alpha \sim U(5.0\times10^3,\ 5.0\times10^5)$. Here, we find that, perhaps unsurprisingly, the performance of most methods is lower than in the African scenario. However, S/HIC once again appears substantially more robust to misspecification of the demographic model than other methods (AUC=0.8247 versus 0.5609 for the next best method; Fig 6).

When we take a closer look and examine the proportion of windows at various distances from sweeps that are assigned to each class, we find that while S/HIC classifies hard sweeps with lower sensitivity than under constant population size scenario (63.9% and 12.5% of test examples are classified as hard and soft, respectively), relatively few linked windows are classified as sweeps (Fig 7A). For soft sweeps S/HIC fares less well (16.7% of windows are correctly classified, and 34.5% classified as hard sweeps), though again relatively few false positives are produced in linked regions. In contrast, *SFselect+* classifies the majority of windows as soft sweeps (Fig 7B), including 62.6% of hard sweeps, and 91.5% of neutral regions. Thus, under this model our method remains fairly sensitive to selective sweeps, though it struggles to correctly infer the mode of selection, while avoiding the enormous false positive rates that may plague other methods. Together, the above results lend credence to the idea that

spatial patterns of variation will be far less impaired by misspecification of the demographic model.

Identifying selective sweeps in a human population sample with European ancestry

The results from simulated data described above suggest that our method has the potential to identify selective sweeps and distinguish them from linked selection and neutrality with excellent accuracy. In order to demonstrate our method's practical utility, we used it to perform a scan for positive selection in humans. In particular, we searched the 1000 Genomes Project's CEU population sample (European individuals from Utah) for selective sweeps occurring after the migration out of Africa. We focused this search on chromosome 18, where several putative selective sweeps have been identified in Europeans (Williamson et al. 2007). The steps we took to train our classifier and filter the 1000 Genomes data prior to conducting our scan are described in the Methods.

In total, we examined 345 windows, each 200 kb in length. We classified 17 windows (4.9%) as centered around a hard sweep, 183 (53.0%) as linked to a hard sweep, 38 (11.0%) as centered around a soft sweep, 77 (8.7%) as linked to a soft sweep, and 30 (8.7%) as neutral. Surprisingly, we infer that over 60% of windows lie within regions whose patterns of variation are affected by linked sweeps, even though sweeps appear to be quite rare. This may imply that, given the genomic landscape of recombination in humans, even rare selective events may be strong enough to impact variation across much larger stretches of the genome. However, we cannot firmly draw this conclusion given the difficulty of distinguishing between linked selection and neutrality under the European demographic model (Fig 7).

Encouragingly, our scan recovered 4 of the 5 putative sweeps on chromosome 18 in Europeans identified by Williamson et al. (2007) using SweepFinder. These include *CCDC178* (which we classify as a hard sweep), *DTNA* (which we classify as soft), *CCDC102B* (soft), and the region spanning portions of *CD226* and *RTTN* (hard; shown in Fig 8). In each of these loci, the windows that we predicted to contain the sweep overlapped regions of elevated composite likelihood ratio (CLR) values from SweepFinder (visualized using data from Pybus et al. 2013). Although the CLR statistic is not completely orthogonal to the summary statistics we examine to perform our classifications, the close overlap that we observe between these two methods underscores our ability to precisely detect the targets of recent positive selection. The complete set of coordinates of putative sweeps from this scan is listed in Supplemental Table S2.

Next, we asked whether S/HIC recovered evidence of positive selection on the *LCT* (lactase) locus. Previous studies have found evidence for very recent and strong selection on this gene in the form population differentiation and long-range haplotype homozygosity (Hollox et al. 2001; Bersaglieri et al. 2004; Tishkoff et al. 2007). Moreover, several variants in this region are associated with lactase persistence. Nielsen et al.'s CLR has also identified this region (Nielsen et al. 2005), but not consistently: Williamson et al.'s (Williamson et al. 2007) CLR scan did not detect a sweep at this locus, nor does a recent scan using the 1000 Genomes data (data from Pybus et al. 2013; Fig. S6). This may be expected, as the selection on lactase persistence alleles appears to have not yet produced completed sweeps. Overall, there is very strong evidence of recent and perhaps ongoing selection for lactase persistence in human populations relying on diary for nutrition.

Like the SweepFinder CLR, S/HIC in its current form is also designed to detect completed sweeps. Nonetheless, we applied S/HIC to a 4 Mb region on chromosome 2 spanning

LCT and neighboring loci. Consistent with previous studies, we found an extremely strong signal of recent positive selection, with a 1 Mb region encompassing *LCT* classified as a hard sweep (Fig S8). This result is also in agreement with Peter et al. (2012), whose approximate Bayesian computation (ABC) approach supported a hard sweep over selection on standing variation at *LCT*. Such a large candidate region might be expected after a very strong selective sweep as is believed to have occurred (or may be ongoing) at *LCT*. However, we cannot rule the possibility of additional targets of selection in this region of chromosome 2: our candidate window also overlaps a region identified by Green et al. (2010) as having an excess of derived alleles in the human genome relative to the number observed in Neanderthal.

DISCUSSION

Detecting the genetic targets of recent adaptation and the mode of positive selection acting on them—selection on *de novo* mutations versus previously standing variants—remains an important challenge in population genetics. The majority of efforts to this end have relied on population genetic summary statistics designed to uncover loci where patterns of allele frequency (e.g. Tajima 1989; Fu and Li 1993; Fay and Wu 2000) or linkage disequilibrium (e.g. Kelly 1997; Kim and Nielsen 2004) depart from the neutral expectation. Recently, powerful machine learning techniques have begun to be applied to this problem, showing great promise (Pavlidis et al. 2010; Lin et al. 2011; Ronen et al. 2013; Schrider et al. 2015). Here we have adopted a machine learning approach to develop S/HIC, a method designed to not only uncover selective sweeps, but to distinguish them from regions linked to sweeps as well as neutrally evolving regions, and to identify the mode of selection. This is achieved by examining spatial patterns of a variety of population genetic summary statistics that capture different facets of variation across a

large-scale genomic region. Currently, this method examines the values of nine statistics across eleven different windows in infer the mode of evolution in the central window—this makes for a total of 99 different values considered by the classifier. By leveraging all of this information jointly, our Extra-Trees classifier is able to detect selection with accuracy unattainable by methods examining a single statistic, underscoring the potential of the machine learning paradigm for population genetic inference. Indeed, on simulated datasets with constant population size, S/HIC has power matching or exceeding previous methods when linked selection is not considered (i.e. the sweep site is known *a priori*), and vastly outperforms them under the more realistic scenario where positive selection must be distinguished from linked selection as well as neutrality.

We argue that the task of discriminating between the targets of positive selection and linked but unselected regions is an extremely important and underappreciated problem that must be solved if we hope to identify the genetic underpinnings of recent adaptation in practice. This is especially so in organisms where the impact of positive selection is pervasive, and therefore much of the genome may be linked to recent selective sweeps (e.g. Langley et al. 2012). A method that can discriminate between sweeps and linked selection would have three important benefits. First, it will reduce the number of spurious sweep calls in flanking regions, thereby mitigating the soft shoulder problem (Schrider et al. 2015). Second, such a method would have the potential to narrow down the candidate genomic region of adaptation. Third, such a method would be able to find those regions *least* affected by linked selection, which themselves might act as excellent neutral proxies for inference into demography or mutation. We have shown that S/HIC is able to distinguish among selection, linked selection, and neutrality with remarkable

power, granting it the ability to localize selective sweeps with unrivaled accuracy and precision, demonstrating its practical utility.

While S/HIC performs favorably to other approaches under the ideal scenario where the true demographic history of the population is known, in practice this may not always be the case. However, because our method relies on spatial patterns of variation, we are especially robust to demography: if the demographic model is misspecified, the disparity in accuracy between S/HIC and other methods is even more dramatic. For example, if we train S/HIC with simulated datasets with constant population size, but test it on simulated population samples experiencing recent exponential growth (e.g. the African model from Tennessen et al. 2012), we still identify sweeps and infer the mode of selection with impressive accuracy, and vastly outperform other methods. We also tested S/HIC on a more challenging model with two population contractions followed by slow exponential growth, and more recent accelerated growth (the European model from Tennessen et al. 2012). Though S/HIC performs far better than other tests for selection in this case, power for all methods is far lower than under constant population size, even if the demographic model is properly specified during training. The reason for this is somewhat disconcerting: under this demographic model, the impact of selective sweeps on genetic diversity is blunted, making it far more difficult for any method to identify selection and discriminate between hard and soft sweeps. This underscores a problem that could prove especially difficult to overcome. That is, for some demographic histories all but the strongest selective sweeps may produce almost no impact on diversity for selection scans to exploit.

A second and related confounding effect of misspecified demography is that following population contraction and recovery/expansion, much of the genome may depart from the neutral expectation, even if selective sweeps are rare. By examining the relative levels of various

summaries of variation across a large region, rather than the actual values of these statistics, we are quite robust to this problem (Fig 7 and Fig S7). In other words, while non-equilibrium demography may reduce S/HIC's sensitivity to selection and its ability to discriminate between hard and soft sweeps, we still classify relatively few neutral or even linked regions as selected. Thus, although inferring the mode of positive selection with high confidence may remain extremely difficult in some populations, our method appears to be particularly well suited for detecting selection in populations with non-equilibrium demographic histories whose parameters are uncertain. Indeed, applying our approach to chromosome 18 in a European human population, we detect most of the putative sweeps previously reported by Williamson et al. (2007).

In summary, we have devised a machine learning-based scan for positive selection that possesses not only unparalleled accuracy, but is also exceptionally robust to demography. Adjustments to the feature space can easily be made to better suit a particular study population. For example, if haplotypic phase is unknown, one can replace measures of gametic LD with zygotic LD. Additional classes could also be incorporated into the classifier (e.g. "partial" or incomplete sweeps), along with relevant summary statistics (features) such as *iHS* and nS_L (Voight et al. 2006; Ferrer-Admetlla et al. 2014). Thus, our approach is practical and flexible. As additional population genetic summary statistics and tests for selection are devised, they can be incorporated into our feature space, thereby strengthening an already powerful method which has the potential to illuminate the impact of selection on genomic variation with unprecedented detail.

ACKNOWLEDGEMENTS

We thank Matthew Hahn and Adam Siepel for comments on the manuscript.

REFERENCES

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome research 19: 711-722.

Altshuler DM, Durbin RM, Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

Berry AJ, Ajioka J and Kreitman M. 1991. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 129: 1111-1117.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE and Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. The American Journal of Human Genetics 74: 1111-1120.

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

Breiman L. 2001. Random forests. Machine learning 45: 5-32.

Cortes C and Vapnik V. 1995. Support-vector networks. Machine learning 20: 273-297.

Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R and Ribeca P. 2012. Fast computation and applications of genome mappability. PLoS One 7: e30377.

Fay JC and Wu C-I. 2000. Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

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

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

Garud NR, Messer PW, Buzbas EO and Petrov DA. 2015. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS genetics 11: e1005004.

Geurts P, Ernst D and Wehenkel L. 2006. Extremely randomized trees. Machine learning 63: 3-42.

Gibbs RA, Rogers J, Katze MG, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. science 316: 222-234.

Gillespie JH. 1991. The causes of molecular evolution. Oxford: Oxford University Press.

Green RE, Krause J, Briggs AW, et al. 2010. A draft sequence of the Neandertal genome. science 328: 710-722.

Hermisson J and Pennings PS. 2005. Soft sweeps molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335-2352.

Ho TK editor. Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on. 1995.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI and Swallow DM. 2001. Lactase haplotype diversity in the Old World. The American Journal of Human Genetics 68: 160-172.

Innan H and Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101: 10667-10672.

Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. Nature communications 5.

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

Jensen JD, Thornton KR, Bustamante CD and Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176: 2371-2379.

Kaplan NL, Hudson R and Langley C. 1989. The" hitchhiking effect" revisited. Genetics 123: 887-899.

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. 2002. The human genome browser at UCSC. Genome research 12: 996-1006.

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

Knerr S, Personnaz L and Dreyfus G. 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In. Neurocomputing: Springer. p. 41-50.

Kong A, Frigge ML, Masson G, et al. 2012. Rate of de novo mutations and the importance of father/'s age to disease risk. Nature 488: 471-475.

Kong A, Thorleifsson G, Gudbjartsson DF, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467: 1099-1103.

Langley CH, Stevens K, Cardeno C, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533-598.

Lin K, Li H, Schlötterer C and Futschik A. 2011. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187: 229-244.

Maynard Smith J and Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetical Research 23: 23-35.

Messer PW and Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28: 659-669.

Mikkelsen TS, Hillier LW, Eichler EE, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

Nei M and Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences 76: 5269-5273.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG and Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

Orr HA and Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics 157: 875-884.

Pavlidis P, Jensen JD and Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907-922.

Pennings PS and Hermisson J. 2006a. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Molecular Biology and Evolution 23: 1076-1084.

Pennings PS and Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLOS Genetics 2: e186.

Peter BM, Huerta-Sanchez E and Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. PLOS Genetics 8: e1003011.

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

Przeworski M, Coop G and Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59: 2312-2323.

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H, Bertranpetit J and Engelken J. 2013. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic acids research: gkt1188.

Quinlan JR. 1986. Induction of decision trees. Machine learning 1: 81-106.

Ronen R, Udpa N, Halperin E and Bafna V. 2013. Learning natural selection from the site frequency spectrum. Genetics: doi: 10.1534/genetics.1113.152587.

Schrider DR, Mendes FK, Hahn MW and Kern AD. 2015. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200: 267-284.

Simonsen KL, Churchill GA and Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

Stephan W, Song YS and Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172: 2647-2663.

Stephan W, Wiehe TH and Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology 41: 237-254.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.

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

Tennessen JA, Bigham AW, O'Connor TD, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. science 337: 64-69.

Teshima KM, Coop G and Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Research 16: 702-712.

Tishkoff SA, Reed FA, Ranciaro A, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics 39: 31-40.

Voight BF, Kudaravalli S, Wen X and Pritchard JK. 2006. A map of recent positive selection in the human genome. PLOS Biology 4: e72.

Wakeley J and Aliacar N. 2001. Gene genealogies in a metapopulation. Genetics 159: 893-905.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. Theoretical population biology 7: 256-276.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD and Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLOS genetics 3: e90.

Wollstein A and Stephan W. 2015. Inferring positive selection in humans from genomic data. Investigative genetics 6: 5.

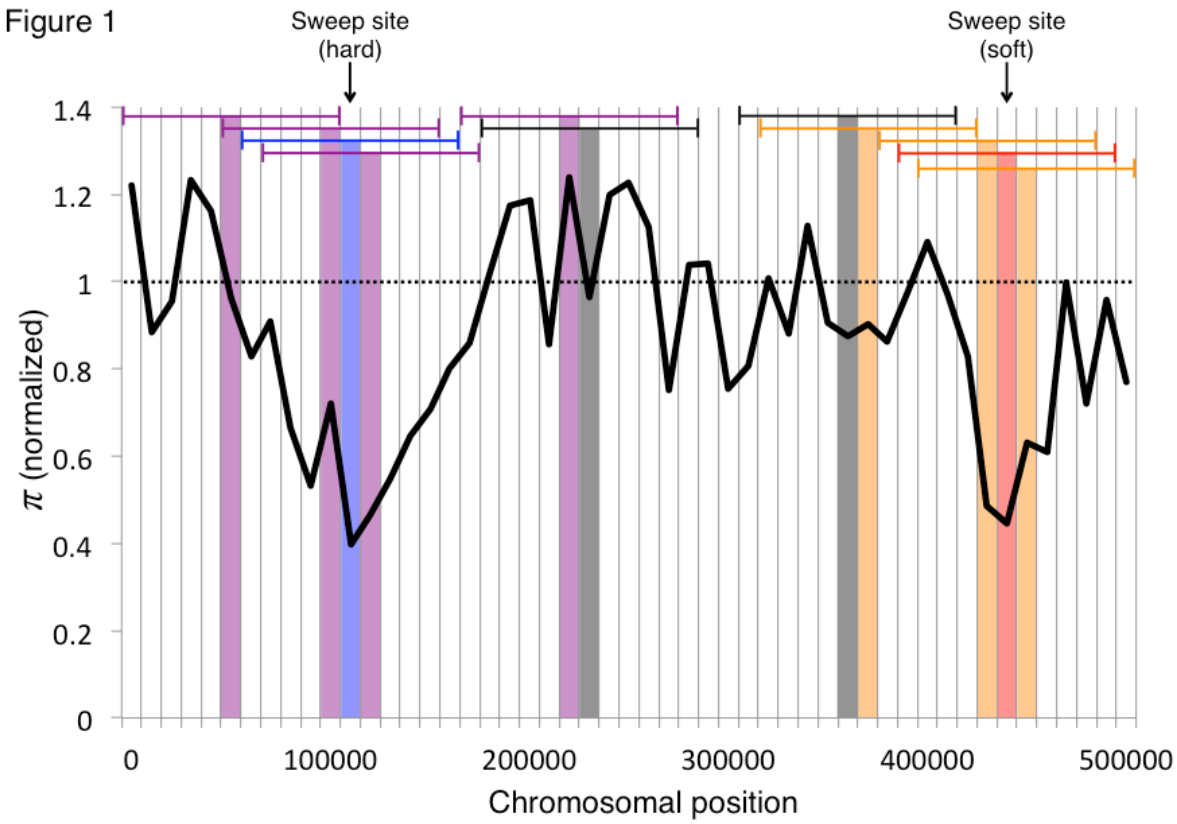

Figure 1: Examples of the five classes used by S/HIC. S/HIC classifies each window as a hard sweep (blue), linked to a hard sweep (purple), a soft sweep (red), linked to a soft sweep (orange), or neutral (gray). This classifier accomplishes this by examining values of various summary statistics in 11 different windows in order to infer the mode of evolution in the central window (the horizontal blue, purple, red, orange, and gray brackets). Regions that are centered on a hard (soft) selective sweep are defined as hard (soft). Regions that are not centered on selective sweeps but have their diversity impacted by a hard (soft) selective sweep but are not centered on the sweep are defined as hard-linked (soft-linked). Remaining windows are defined as neutral. S/HIC is trained on simulated examples of these five classes in order to distinguish selective sweeps from linked and neutral regions in population genomic data.

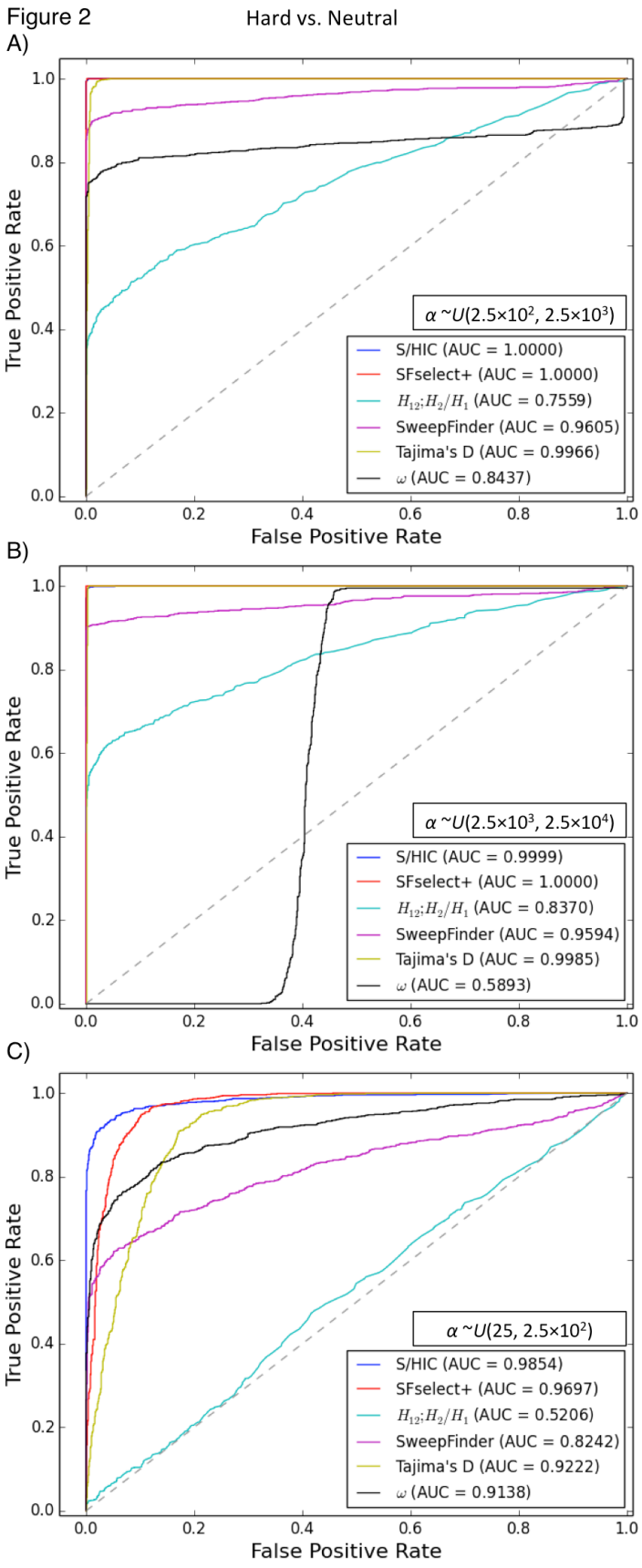

Figure 2: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a hard sweep and neutrally evolving regions. A) For intermediate strengths of selection ($\alpha \sim U(2.5\times10^2, 2.5\times10^3)$). B) For stronger selective sweeps ($\alpha \sim U(2.5\times10^3, 2.5\times10^4)$). C) For weaker sweeps ($\alpha \sim U(2.5\times10^1, 2.5\times10^2)$). Here, and for all other ROC curves unless otherwise noted, methods that require training from simulated sweeps were trained by combining three different training sets: one where $\alpha \sim U(2.5\times10^1, 2.5\times10^2)$, one where $\alpha \sim U(2.5\times10^2, 2.5\times10^3)$, and one where $\alpha \sim U(2.5\times10^3, 2.5\times10^4)$.

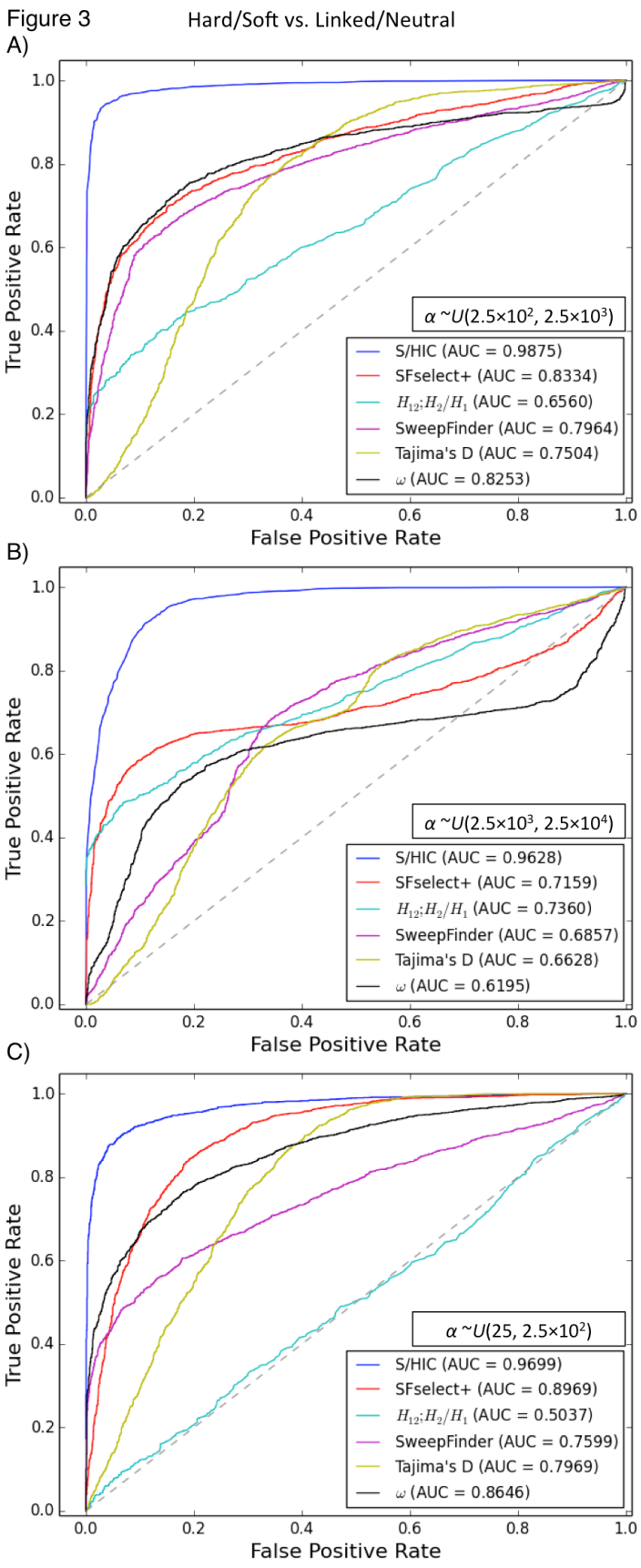

Figure 3: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a sweep (either hard or soft) and unselected regions (either neutral or linked to sweeps). A) For intermediate strengths of selection ($\alpha \sim U(2.5\times10^2, 2.5\times10^3)$). B) For stronger selective sweeps ($\alpha \sim U(2.5\times10^3, 2.5\times10^4)$). C) For weaker sweeps ($\alpha \sim U(2.5\times10^1, 2.5\times10^2)$).

Figure 4

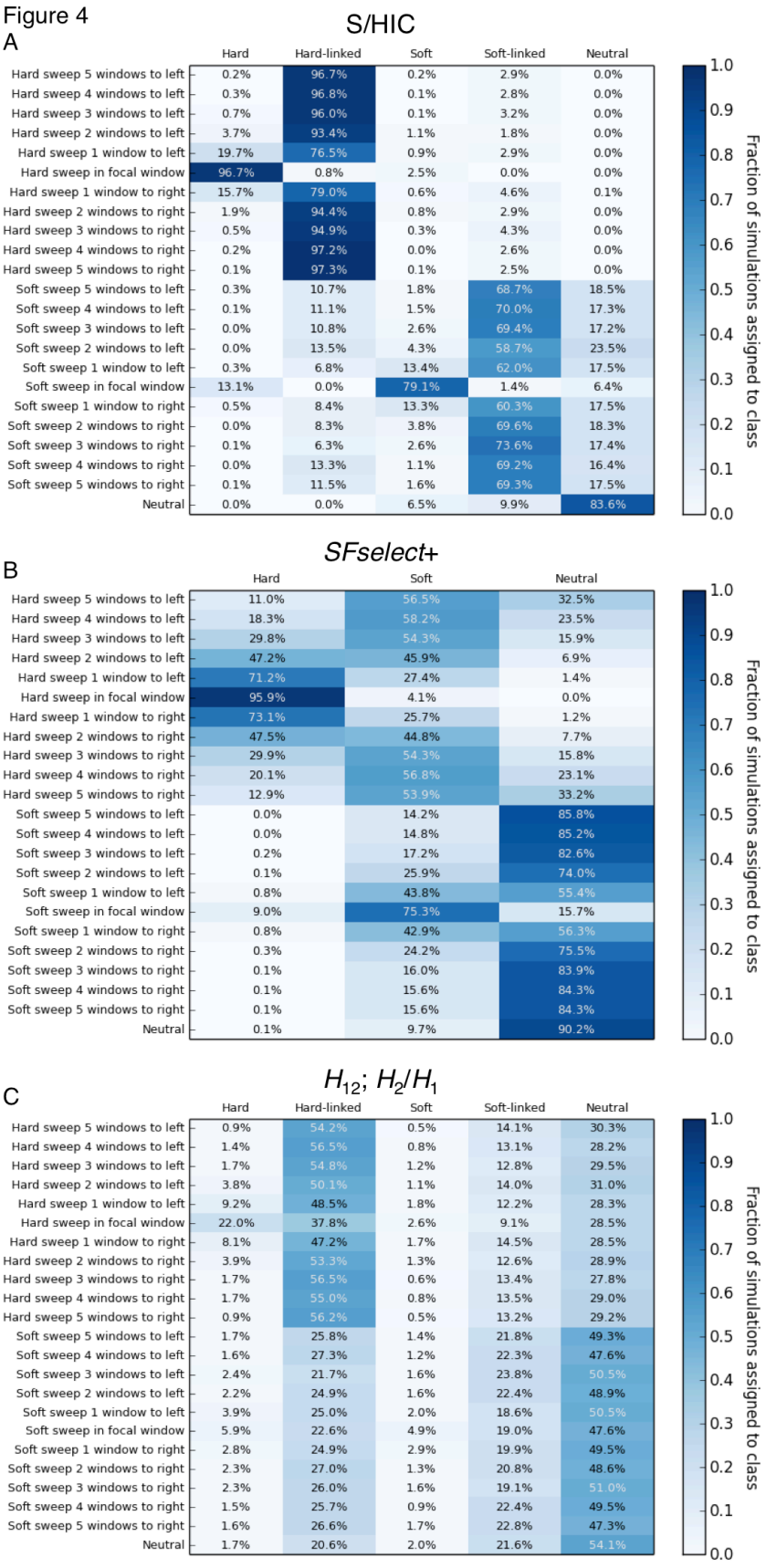

Figure 4: Heatmaps showing the fraction of regions at varying distances from sweeps inferred to belong to each class by S/HIC, *SFselect*+, and Garud et al.'s method using H_{12} and H_2/H_1. The location of any sweep relative to the classified window (or "Neutral" if there is no sweep) is shown on the *y*-axis, while the inferred class on the *x*-axis. Here, $\alpha \sim U(2.5\times10^2, 2.5\times10^3)$. When classifying regions with H_{12} and H_2/H_1, if a hard (soft) sweep was detected, the region was classified as hard (soft) if the maximum H_{12} value was found in the central subwindow, and as hard-linked (soft-linked) otherwise.

Figure 5

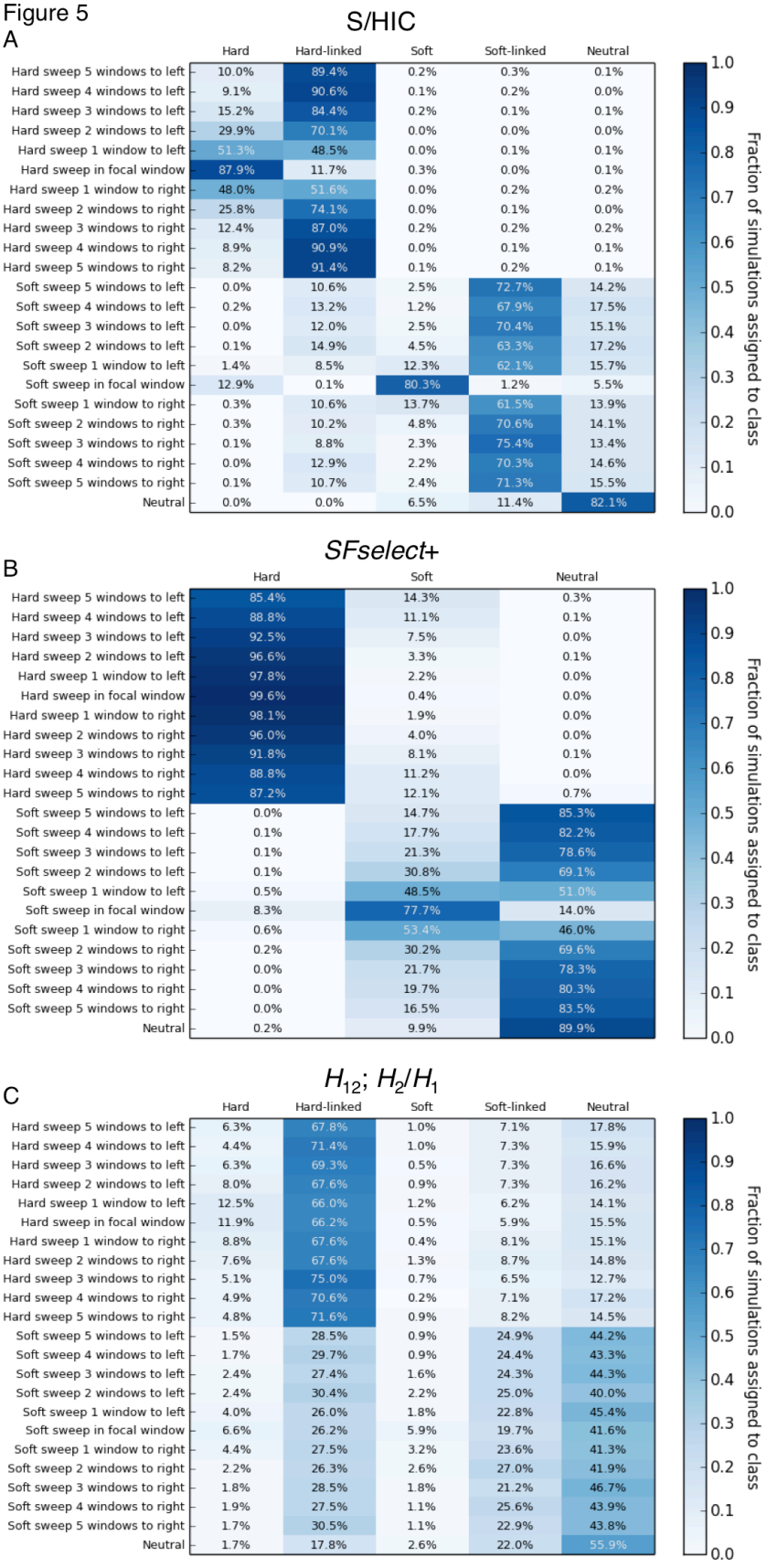

Figure 5: Heatmaps showing the fraction of regions at varying distances from strong sweeps inferred to belong to each class by S/HIC, *SFselect*+, and Garud et al.'s method using H_{12} and H_2/H_1. The location of any sweep relative to the classified window (or "Neutral" if there is no sweep) is shown on the *y*-axis, while the inferred class on the *x*-axis. Here, $\alpha \sim U(2.5\times10^3, 2.5\times10^4)$.

Figure 6 Hard/Soft vs. Linked/Neutral

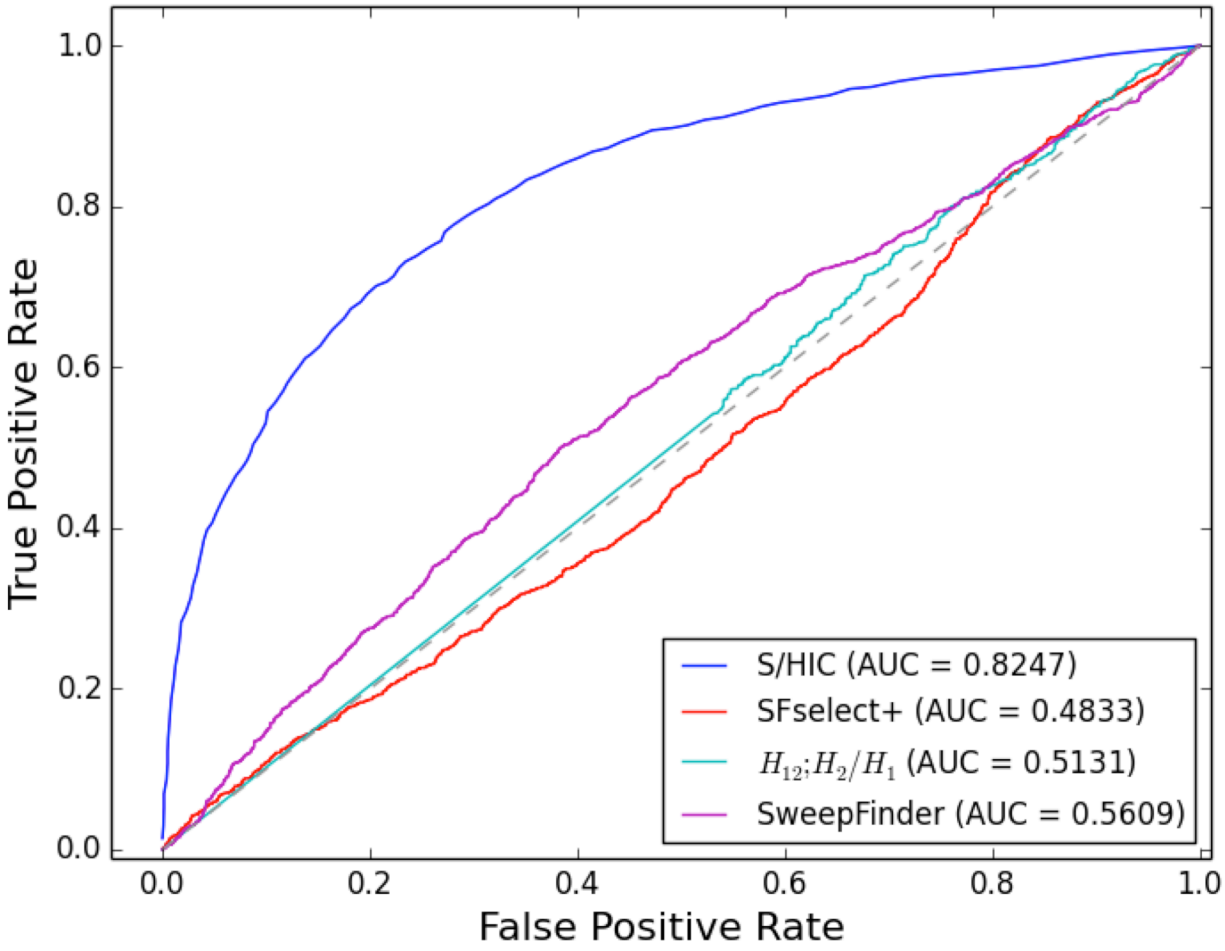

Figure 6: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a sweep (either hard or soft) and unselected regions (either neutral or linked to sweeps) when testing on simulations with Tennessen et al.'s European demographic model. Here, $\alpha \sim U(5\times10^3, 5\times10^5)$, and the methods that require training from simulated sweeps were trained from the same simulations with equilibrium demography as used for Figures 2-7.

Figure 7

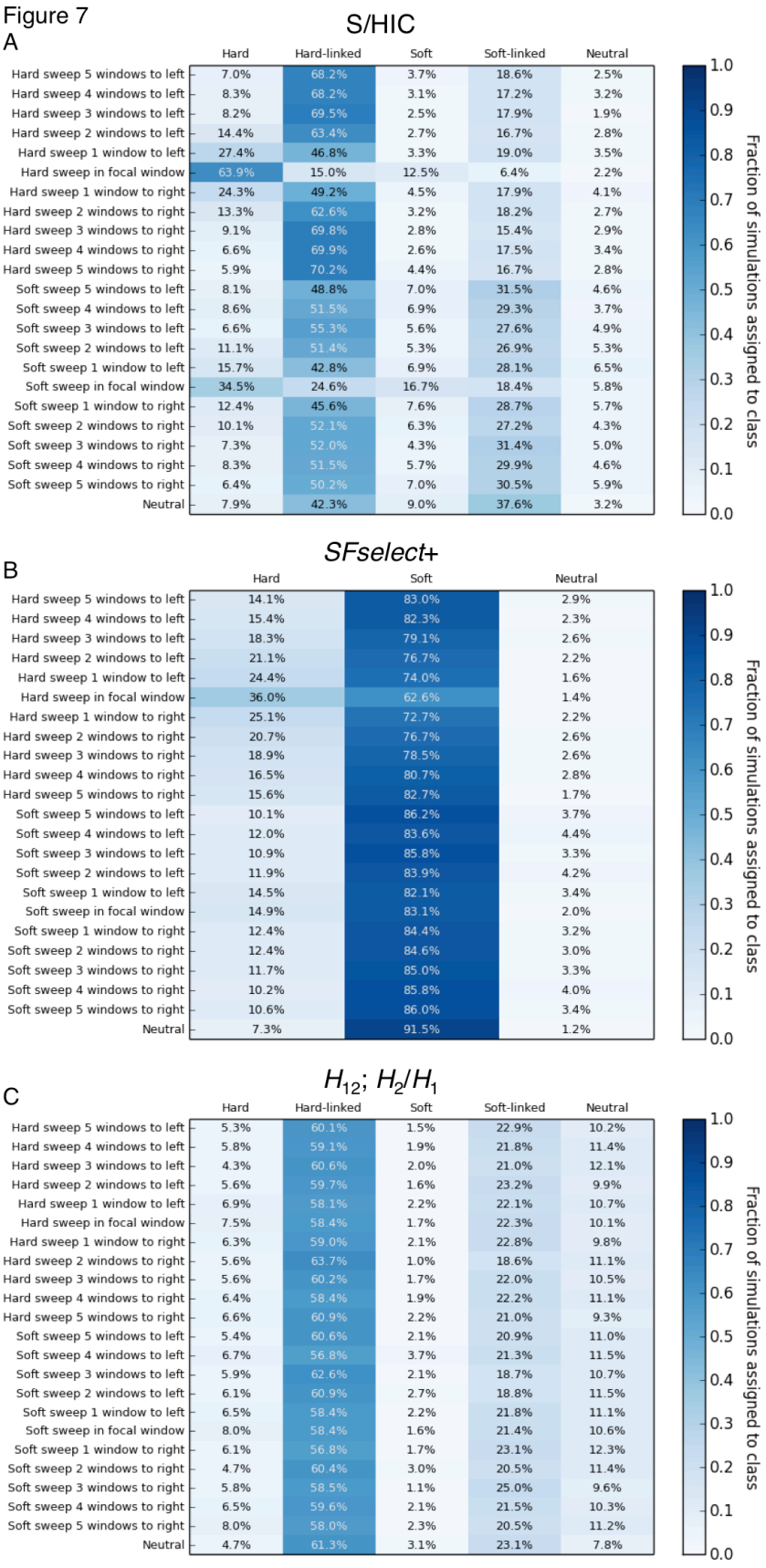

Figure 7: Heatmaps showing the fraction of regions simulated under Tennessen et al.'s European demographic model located at varying distances from sweeps inferred to belong to each class by S/HIC, *SFselect+*, and Garud et al.'s method using H_{12} and H_2/H_1. The location of any sweep relative to the classified window (or "Neutral" if there is no sweep) is shown on the *y*-axis, while the inferred class on the *x*-axis. Here, $\alpha \sim U(5\times10^3, 5\times10^5)$. These three classifiers were trained from simulations with equilibrium demography.

Figure 8

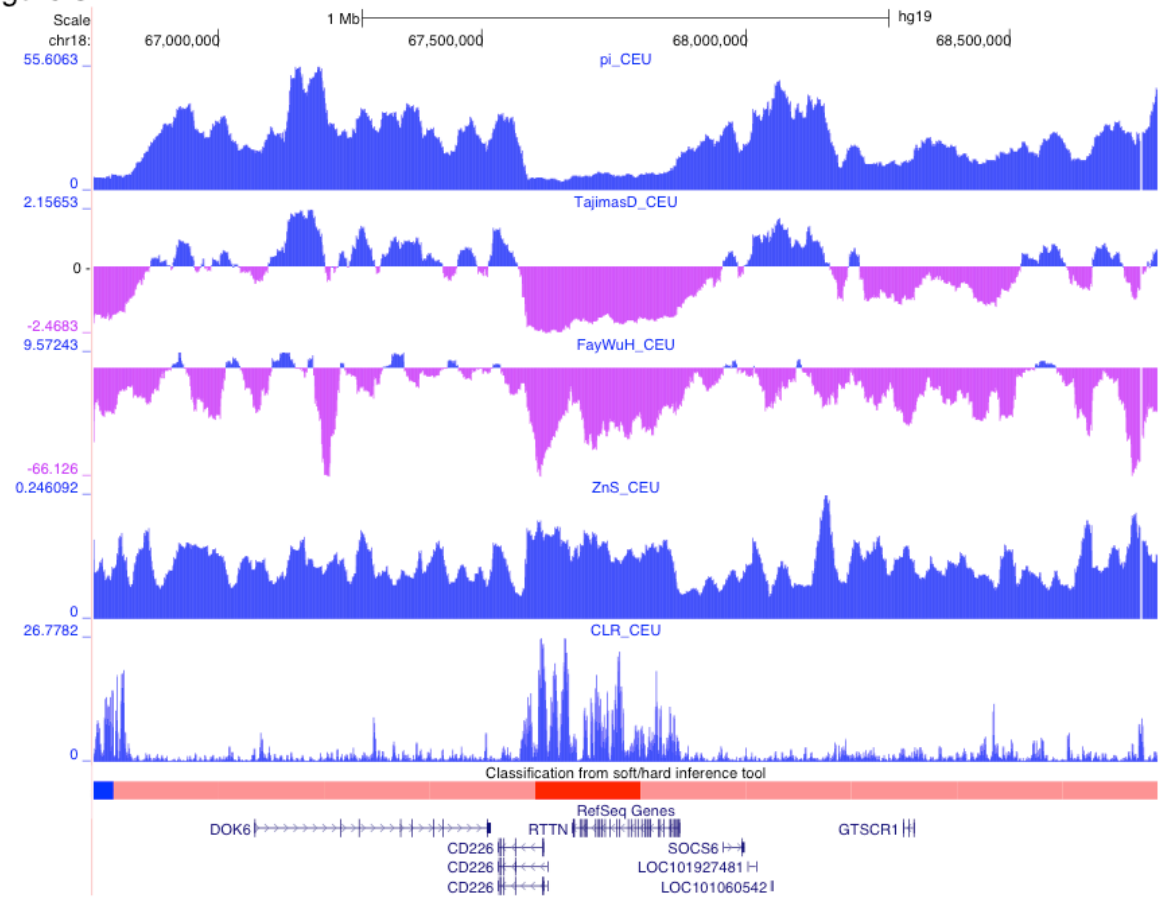

Figure 8: Browser screenshot showing patterns of variation around a putative selective sweep in Europeans near *CD226* and *RTTN* in chr18. Values of π, Tajima's D, Fay and Wu's H, Kelley's Z_{nS}, and Nielsen et al's composite likelihood ratio, all from Pybus et al. (2013), are shown. Beneath these statistics we show the classifications from S/HIC (red: hard sweep; faded red: hard-linked; blue: soft sweep; faded blue: soft-linked; black: neutral). This image was generated using the UCSC Genome Browser (http://genome.ucsc.edu).

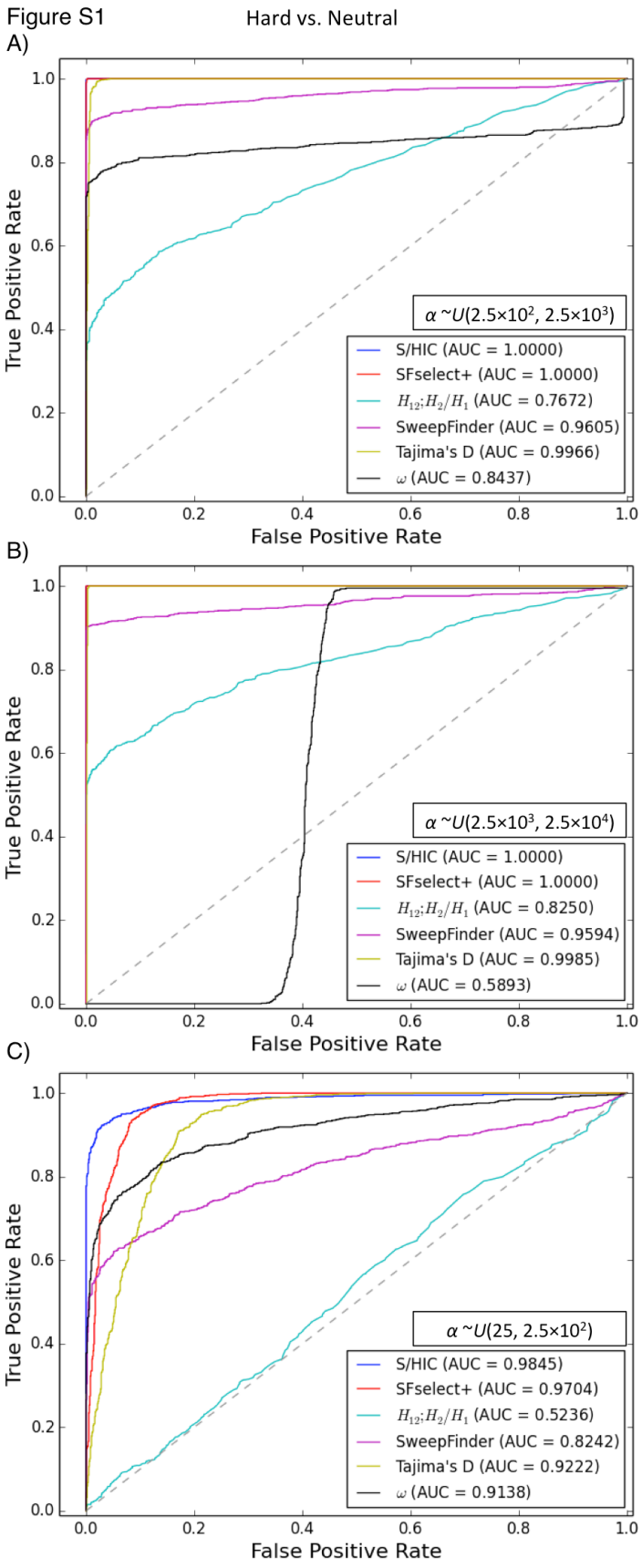

Figure S1: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a hard sweep and neutrally evolving regions. A) For intermediate strengths of selection ($\alpha \sim U(2.5\times10^2, 2.5\times10^3)$). B) For stronger selective sweeps ($\alpha \sim U(2.5\times10^3, 2.5\times10^4)$). C) For weaker sweeps ($\alpha \sim U(2.5\times10^1, 2.5\times10^2)$). Here, the methods that require training from simulated sweeps were trained from a set having the same distribution of selection coefficients as the test set.

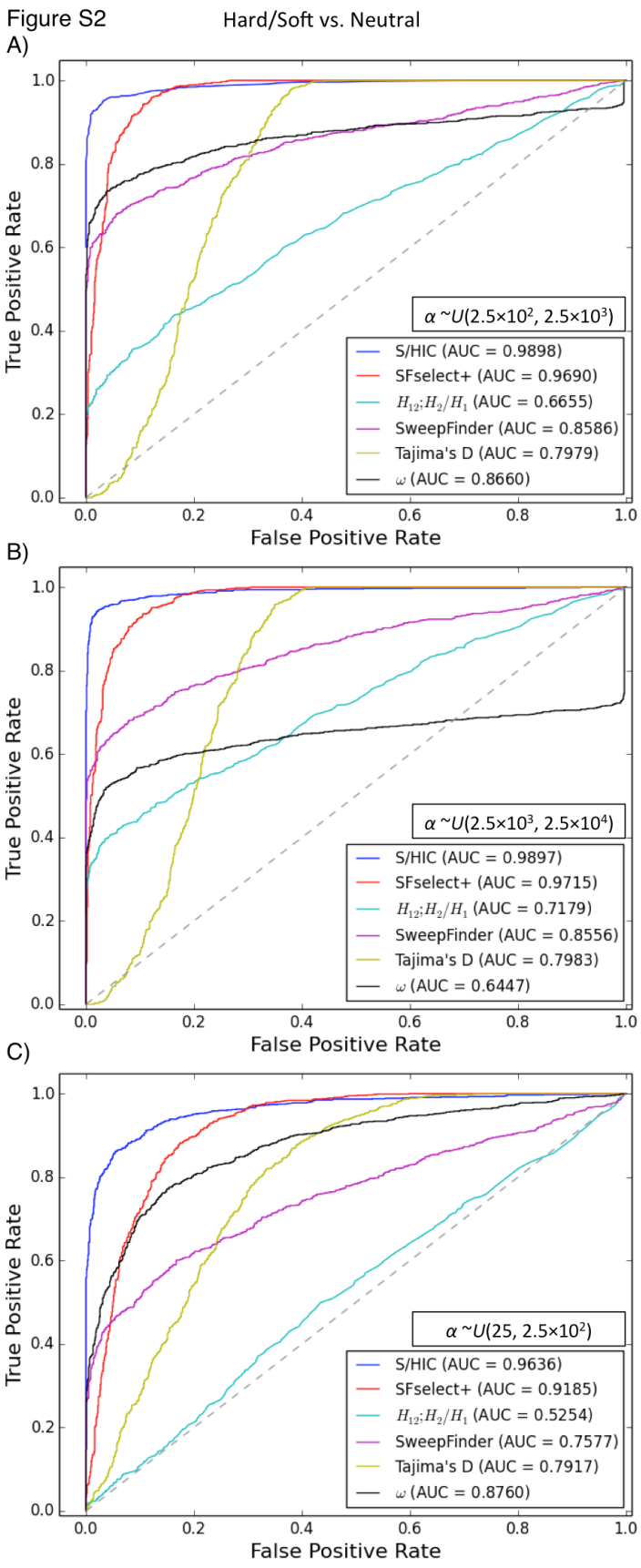

Figure S2: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a sweep (either hard or soft) and neutrally evolving regions. A) For intermediate strengths of selection ($\alpha \sim U(2.5\times10^2, 2.5\times10^3)$). B) For stronger selective sweeps ($\alpha \sim U(2.5\times10^3, 2.5\times10^4)$). C) For weaker sweeps ($\alpha \sim U(2.5\times10^1, 2.5\times10^2)$).

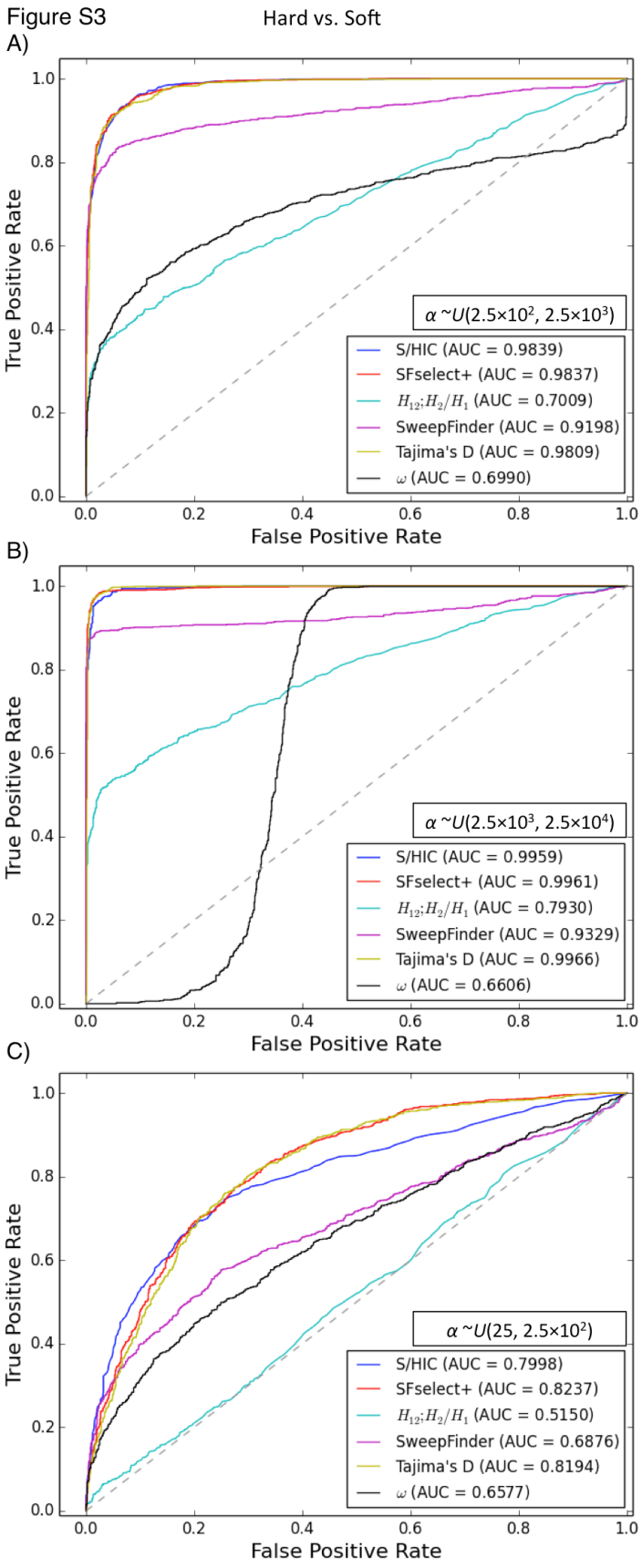

Figure S3: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between hard and soft sweeps. A) For intermediate strengths of selection ($\alpha \sim U(2.5\times10^2, 2.5\times10^3)$). B) For stronger selective sweeps ($\alpha \sim U(2.5\times10^3, 2.5\times10^4)$). C) For weaker sweeps ($\alpha \sim U(2.5\times10^1, 2.5\times10^2)$).

Figure S4

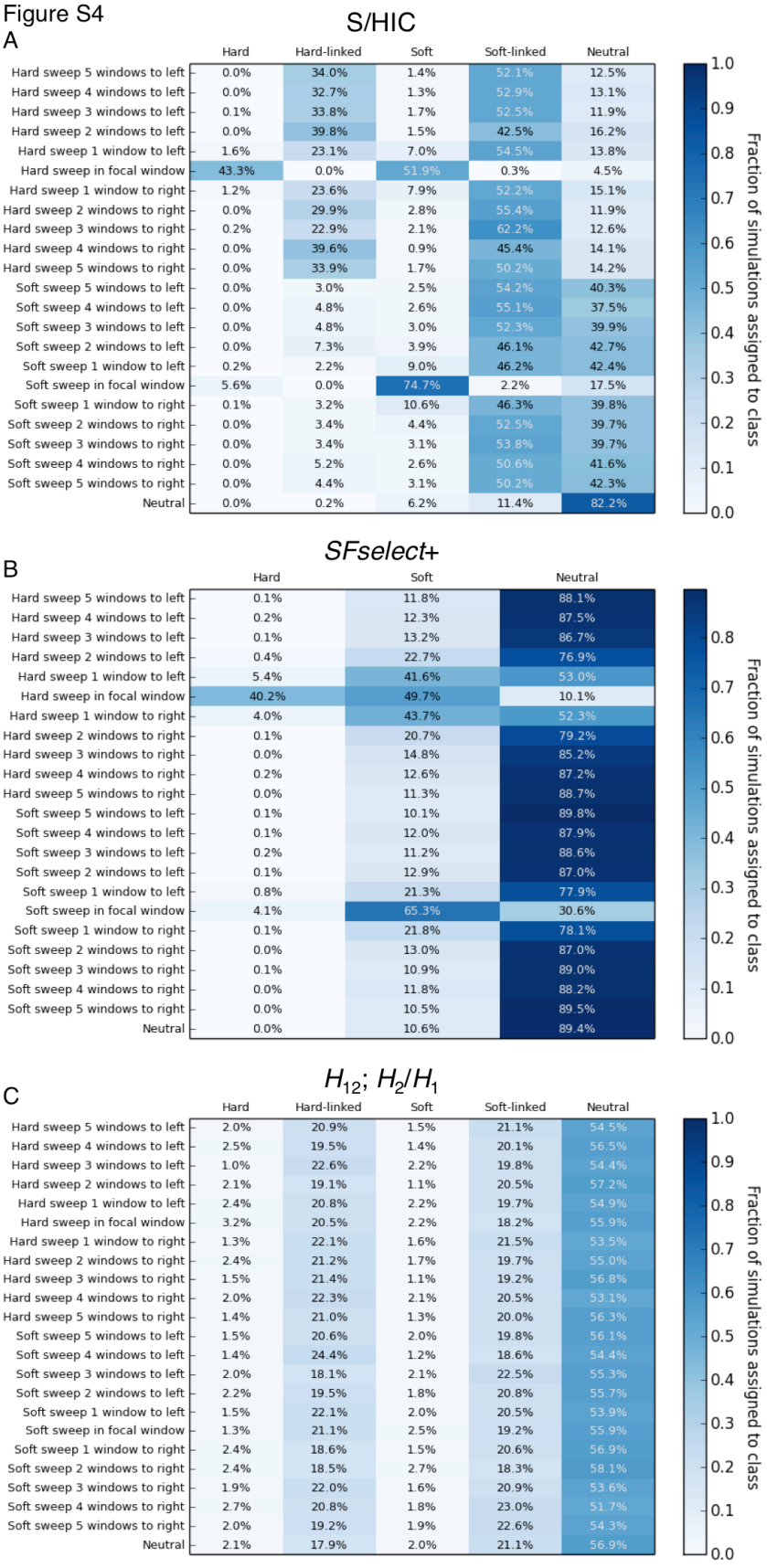

Figure S4: Heatmaps showing the fraction of regions at varying distances from weak sweeps inferred to belong to each class by S/HIC, *SFselect+*, and Garud et al.'s method using H_{12} and H_2/H_1. The location of any sweep relative to the classified window (or "Neutral" if there is no sweep) is shown on the *y*-axis, while the inferred class on the *x*-axis. Here, $\alpha \sim U(2.5\times10^1, 2.5\times10^2)$.

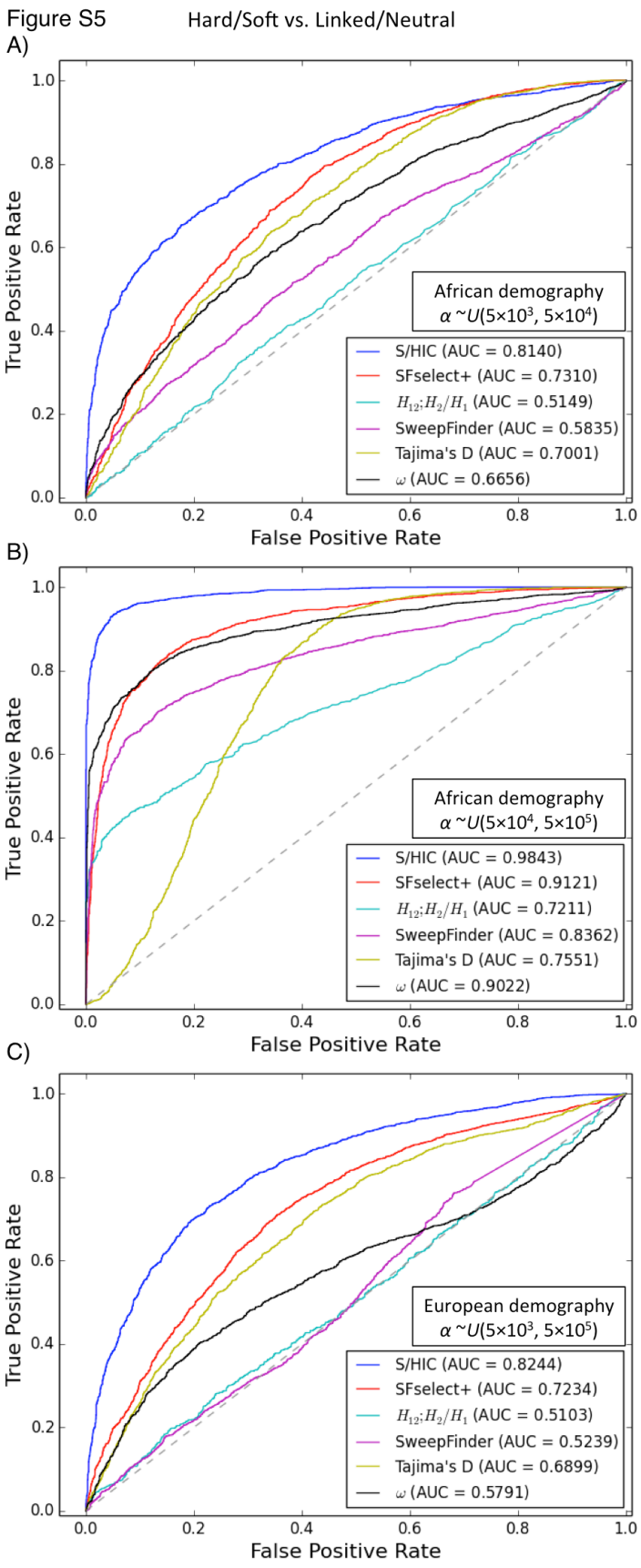

Figure S5: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a sweep (either hard or soft) and unselected regions (either neutral or linked to sweeps) when testing on non-equilibrium demography. Here, the methods that require training from simulated sweeps were trained from the same demographic model used for testing. A) Testing on the African demographic model, with $\alpha \sim U(5\times10^3, 5\times10^4)$. B) The African demographic model, with $\alpha \sim U(5\times10^4, 5\times10^5)$. C) The European demographic model, with $\alpha \sim U(5\times10^3, 5\times10^5)$.

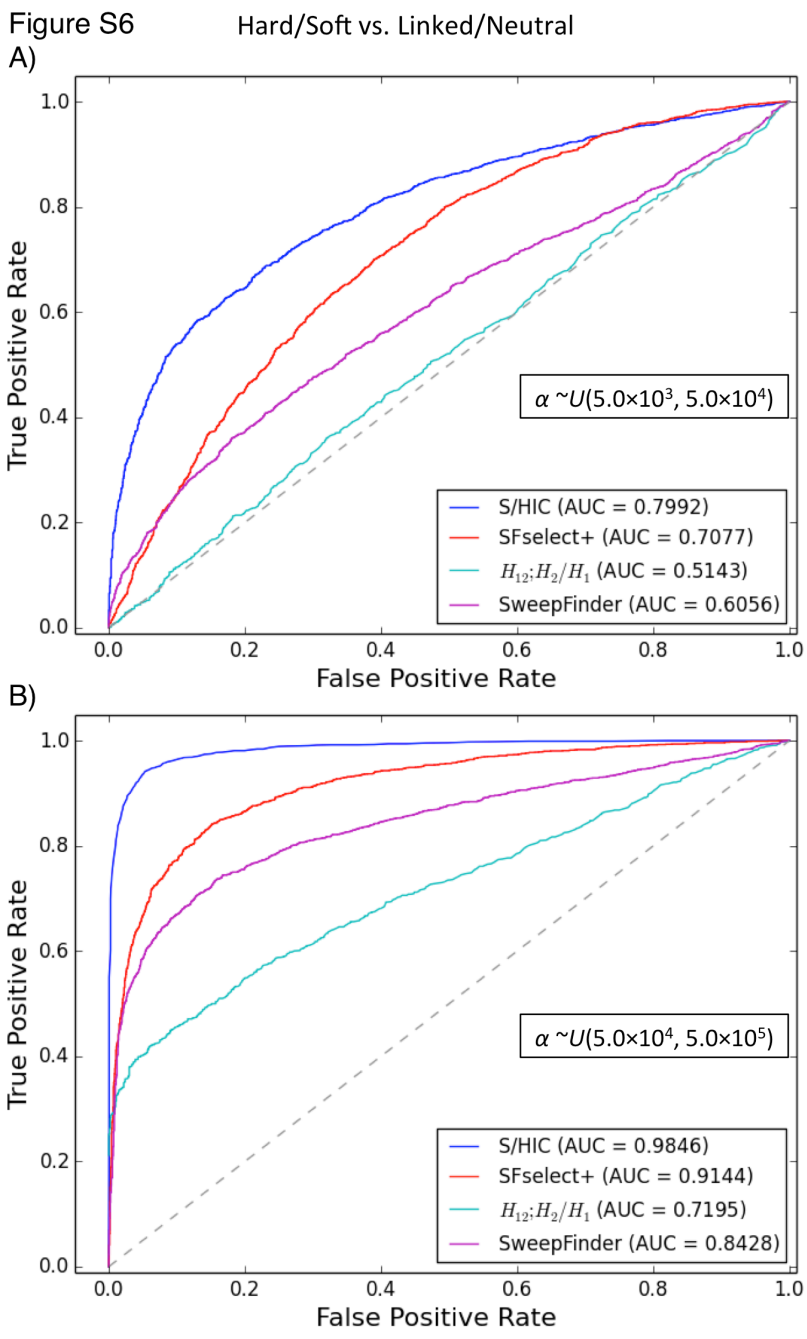

Figure S6: ROC curves showing the true and false positive rates of various methods/statistics when tasked with discriminating between regions containing a sweep (either hard or soft) and unselected regions (either neutral or linked to sweeps) when training with equilibrium demography but testing on non-equilibrium demography. Here, the methods that require training from simulated sweeps were trained from the same simulations with equilibrium demography as used for Figures 2-7. A) Testing on the African demographic model, with $\alpha \sim U(5\times10^3, 5\times10^4)$. B) The African demographic model, with $\alpha \sim U(5\times10^4, 2.5\times10^5)$. C) The European demographic model, with $\alpha \sim U(5\times10^3, 5\times10^5)$. Note that Tajima's D and Kim and Nielsen's ω were omitted from this figure, as we simply used the values of these statistics to generate ROC curves without respect to any demographic model.

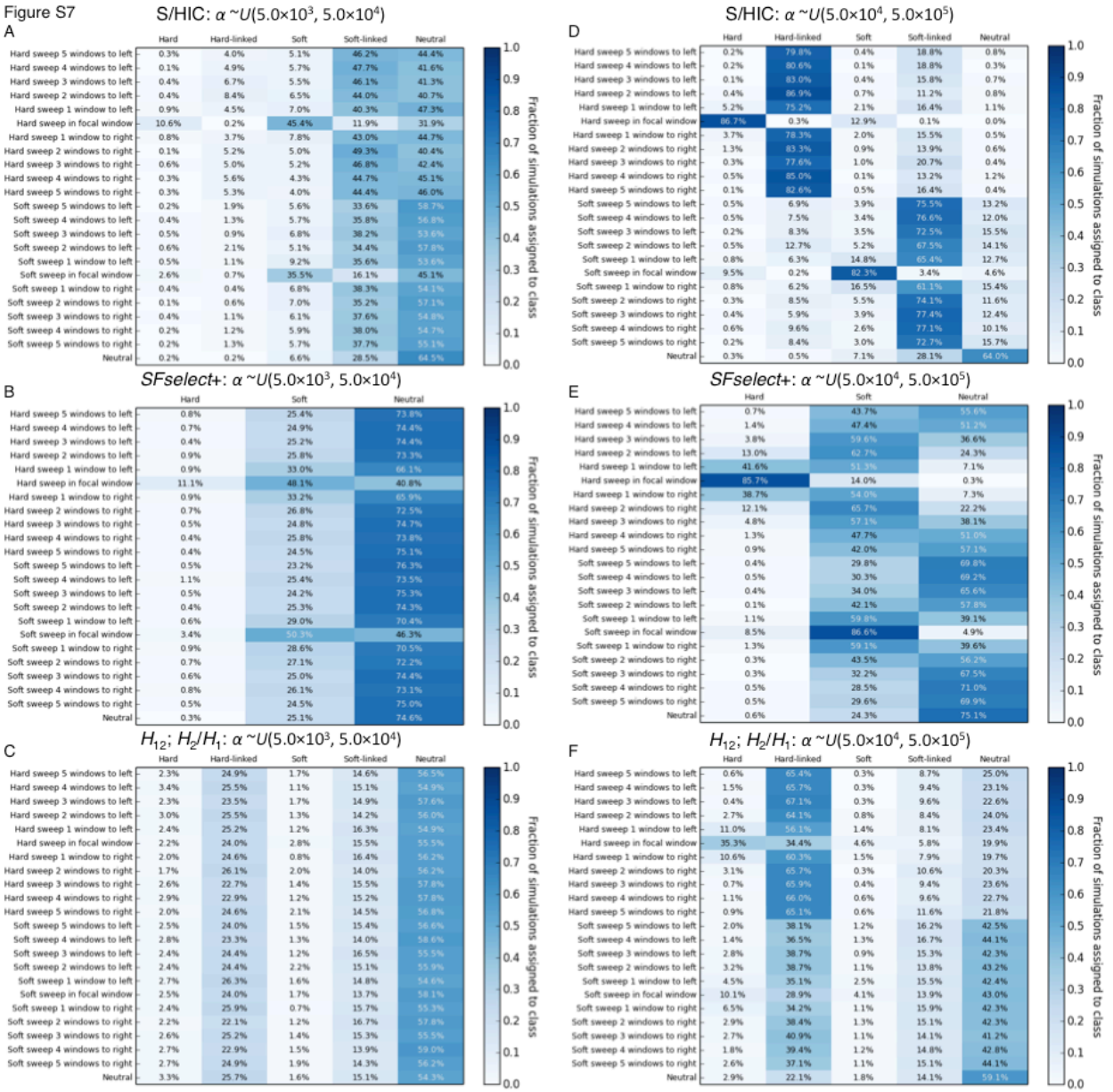

Figure S7: Heatmaps showing the fraction of regions simulated under Tennessen et al.'s African demographic model located at varying distances from sweeps inferred to belong to each class by S/HIC, *SFselect+*, and Garud et al.'s method using H_{12} and H_2/H_1. The location of any sweep relative to the classified window (or "Neutral" if there is no sweep) is shown on the *y*-axis, while the inferred class on the *x*-axis. For panels A–C, $\alpha \sim U(5\times10^3, 5\times10^4)$, and for D–F $\alpha \sim U(5\times10^4, 5\times10^5)$. These three classifiers were trained from simulations with equilibrium demography.

Figure S8

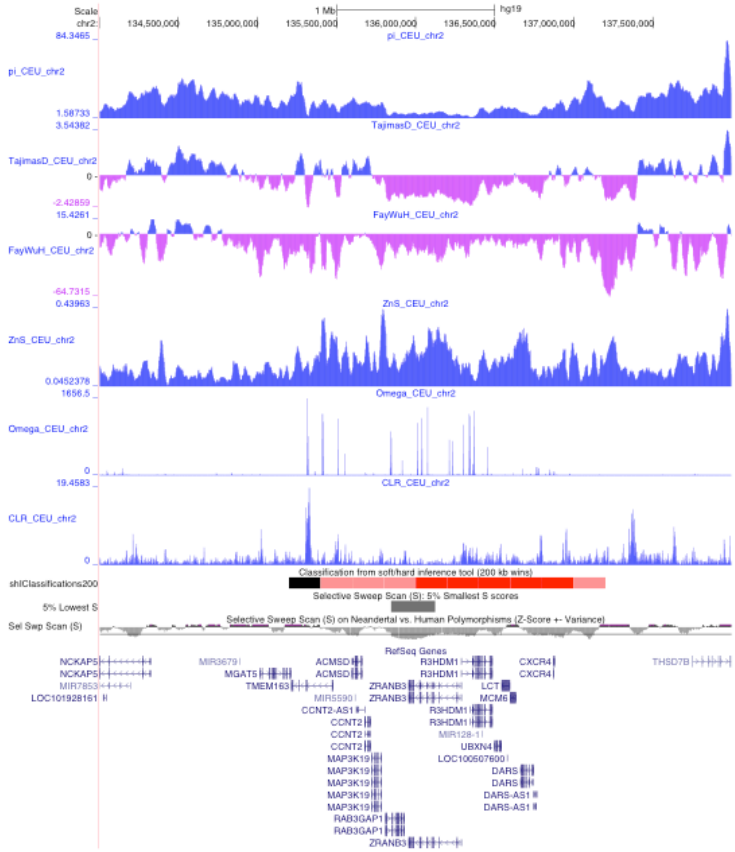

Figure S8: Patterns of variation around the *LCT* locus in the CEU population. Values of π, Tajima's D, Fay and Wu's H, Kelley's Z_{nS}, Kim and Nielsen's ω, and Nielsen et al's composite likelihood ratio, all from Pybus et al. (2013), are shown. Beneath these statistics we show the classifications from S/HIC (red: hard sweep; faded red: hard-linked; blue: soft sweep; faded blue: soft-linked; black: neutral). This image was generated using the UCSC Genome Browser (http://genome.ucsc.edu).

SUPPLEMENTAL TABLE LEGENDS

Table S1: Parameters used for simulating training and test datasets of large chromosomal regions.

Table S2: Predicted sweeps on chromosome 18 from the CEU sample from Phase I of the 1000 Genomes Project.