A likelihood approach for uncovering selective sweep signatures from haplotype data

Alexandre M. Harris[1,2] and Michael DeGiorgio[3,*]

March 26, 2020

[1]*Department of Biology, Pennsylvania State University, University Park, PA 16802, USA*

[2]*Molecular, Cellular, and Integrative Biosciences at the Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA*

[3]*Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA*

[*]*Corresponding author:* mdegiorg@fau.edu

Keywords: Maximum likelihood, selective sweep, haplotype

Running title: Maximum likelihood sweeps

Abstract

Selective sweeps are frequent and varied signatures in the genomes of natural populations, and detecting them is consequently important in understanding mechanisms of adaptation by natural selection. Following a selective sweep, haplotypic diversity surrounding the site under selection decreases, and this deviation from the background pattern of variation can be applied to identify sweeps. Multiple methods exist to locate selective sweeps in the genome from haplotype data, but none leverage the power of a model-based approach to make their inference. Here, we propose a likelihood ratio test statistic T to probe whole genome polymorphism datasets for selective sweep signatures. Our framework uses a simple but powerful model of haplotype frequency spectrum distortion to find sweeps and additionally make an inference on the number of presently sweeping haplotypes in a population. We found that the T statistic is suitable for detecting both hard and soft sweeps across a variety of demographic models, selection strengths, and ages of the beneficial allele. Accordingly, we applied the T statistic to variant calls from European and sub-Saharan African human populations, yielding primarily literature-supported candidates, including *LCT*, *RSPH3*, and *ZNF211* in CEU, *SYT1*, *RGS18*, and *NNT* in YRI, and *HLA* genes in both populations. We also searched for sweep signatures in *Drosophila melanogaster*, finding expected candidates at *Ace*, *Uhg1*, and *Pimet*. Finally, we provide open-source software to compute the T statistic and the inferred number of presently sweeping haplotypes from whole-genome data.

Introduction

A selective sweep is a genomic signature resulting from positive selection in which the linked variants surrounding the site under selection rise to high frequency together in a population, thereby yielding a footprint of reduced diversity that can span across megabases [Przeworski, 2002, Gillespie, 2004, Kim and Nielsen, 2004, Garud et al., 2015, Hermisson and Pennings, 2017]. Thus, a recent selective event is identifiable in polymorphism data from a region of extended haplotype homozygosity, and the signal of a selective sweep accordingly decays over time as mutation and recombination break up long haplotypes [Sabeti et al., 2002, Schweinsberg and Durrett, 2005, Voight et al., 2006]. Selective sweeps can arise from multiple processes, including the *de novo* emergence of a selectively advantageous allele, selection on standing population haplotypic variation, and recurrent mutation to a selectively advantageous allele [Hermisson and Pennings, 2005, Pennings and Hermisson, 2006a,b]. The former scenario is a hard sweep, in which a single haplotype rises to high population frequency, gradually replacing all other haplotypes as the sweep proceeds to fixation. The latter two scenarios are soft sweeps, in which multiple haplotypes simultaneously rise to high population frequency, and a greater haplotypic diversity underlies the sweep.

Identifying selective sweeps is important because sweeps serve as indicators of recent rapid adaptation in a population, providing insight into the pressures that shaped its present-day levels of genetic diversity [Vatsiou et al., 2016, Librado et al., 2017]. These pressures can vary considerably in their intensity and duration, resulting in selection signals of varying magnitude ranging from prominent, such as *LCT* in Europeans [Bersaglieri et al., 2004], to the more subtle *ASPM*, implicated in the development of human brain size [Kouprina et al., 2004, Peter et al., 2012]. Whereas strong sweeps are typically easy to detect, weaker sweeps typically require a large sample size for detection [Jensen et al., 2007, Pavlidis et al., 2013], and may only be identifiable through sophisticated approaches [Chen et al., 2010]. Selective sweeps, while not the only signature of adaptation in natural populations, are likely to occur at loci where mutations have a large effect size, little negative pleiotropic effects, and contribute to phenotypes that are either monogenic or controlled by few genes [Pritchard and DiRienzo, 2010]. In addition, identifying selective sweeps is important to make inferences about the relative contributions of hard and soft sweeps to adaptive events in study organisms [Garud et al., 2015, Schrider and Kern, 2016, Harris et al., 2018a], which is a topic of continued debate [Hernandez et al., 2011, Jensen, 2014, Schrider and Kern, 2017, Harris et al., 2018b, Mughal and DeGiorgio, 2019].

Multiple powerful methods have been proposed to characterize selective sweeps, and well-established among these are composite likelihood ratio (CLR) methods [Kim and Stephan, 2002, Kim and Nielsen, 2004, Nielsen et al., 2005, Chen et al., 2010, Pavlidis et al., 2013, Vy and Kim, 2015, Racimo, 2016, Huber et al., 2016, DeGiorgio et al., 2016], and haplotype homozygosity-based methods [Voight et al., 2006, Ferrer-

Admetlla et al., 2014, Garud et al., 2015, Harris et al., 2018a]. The former category of methods represents approaches in which the probability of neutrality in a genomic region under analysis is compared to the probability of a selective sweep in that region, based on a model of distortion in the site frequency spectrum expected under a sweep. A CLR statistic quantifies support for the alternative hypothesis of selection, with larger values indicating greater support. Although CLR methods make simplifying assumptions in their models [Beaumont et al., 2010, Pavlidis and Alachiotis, 2017], they have demonstrated a powerful capacity for identifying multiple different signatures of selection without the need for computationally intense calculations of full likelihood functions [Kim and Stephan, 2002, DeGiorgio et al., 2014, Huber et al., 2016]. However, because they are typically allele frequency-based approaches, the CLR methods may lack in power to detect soft sweeps in comparison to haplotype-based methods, which can generally detect both [Pennings and Hermisson, 2006b, Ferrer-Admetlla et al., 2014]. Accordingly, the need exists for methods that leverage the power and efficiency of CLR approaches, while providing the sensitivity of haplotype-based approaches.

We introduce an approach for identifying selective sweep signatures using a likelihood ratio framework T that is the first haplotype-based method of its kind, intended to address the limitations of previous methods. Our T statistic (see *Definition of statistic*) has high power to detect recent sweeps from genome-wide polymorphism data and additionally infers the number of presently sweeping haplotypes as a model parameter, providing an additional layer of insight not shared with other CLR methods. This attribute is especially important because it eliminates the need for time- and computation-heavy alternatives, such as training a machine-learning classifier [Lin et al., 2011, Kern and Schrider, 2018, Mughal and DeGiorgio, 2019], or drawing inferences from a posterior distribution by approximate Bayesian computation [Garud et al., 2015, Harris et al., 2018a, Harris and DeGiorgio, 2019]. We demonstrate with simulated data that the T-statistic identifies recent hard and soft sweeps, and performs especially well for population size expansion models. As such, our application of the T statistic to human and *Drosophila melanogaster* datasets recovered multiple previously-characterized candidate sweeps in both organisms, allowing us to corroborate and enhance our understanding of adaptation in each of their histories. We implement the T statistic in an open-source software package termed `LASSI` (**L**ikelihood-based **A**pproach for **S**elective **S**weep **I**nference), which can be downloaded at `http://degiorgiogroup.fau.edu/LASSI.html`.

Definition of statistic

The goal of our approach is to identify genomic signatures of selective sweeps. We achieve this by assigning a T statistic to each SNP-delimited window of analysis in the genome. The T statistic is a measure of the likelihood that an analysis window is consistent with a selective sweep rather than neutrality. We base this inference on the sample haplotype frequency spectrum, reasoning that a spectrum with few high-

frequency haplotypes indicates a sweep, and a spectrum with no moderate- or high-frequency haplotypes indicates neutrality. For this reason, genomic regions with low mutation and recombination rates can resemble selective sweep regions owing to their reduced nucleotide and haplotypic diversity [Pollinger et al., 2005, Wiehe et al., 2007, O'Reilly et al., 2008, Pavlidis and Alachiotis, 2017], and so caution is warranted as with any approach. The T statistic is a likelihood ratio test in which the model of neutrality, based on the genome-wide haplotype frequency spectrum, is nested within the model of selection, based on a distortion of the genome-wide haplotype frequency spectrum toward few moderate- or high-frequency haplotypes. We illustrate examples of haplotype frequency spectra for neutrality and sweeps in Figure 1, and also provide a schematic showing how key model parameters relate to distortions in the haplotype frequency spectrum.

To begin, we must first define the haplotype spectrum on which we will base our neutral expectation. That is, the spectrum that we will assign as representative for a genomic window under neutrality. For all genomic windows in the sample, we extract the haplotype frequency spectrum, arrange frequencies in descending order, and truncate the spectrum at an arbitrary value K most frequent haplotypes (compare top and middle panels of Figure 1, first column). Thus, for each window ℓ, $\ell = 1, 2, \ldots, L$ for L windows, we have a truncated spectrum $\mathbf{p}^{(\ell)} = (p_1^{(\ell)}, p_2^{(\ell)}, \ldots, p_K^{(\ell)})$, where $p_1^{(\ell)} \geq p_2^{(\ell)} \geq \ldots \geq p_K^{(\ell)}$, and normalized such that $\sum_{i=1}^{K} p_i^{(\ell)} = 1$. Next, we define the vector $\mathbf{p} = (p_1, p_2, \ldots, p_K)$, such that $p_i = \frac{1}{L} \sum_{\ell=1}^{L} p_i^{(\ell)}$ for $i = 1, 2, \ldots, K$. We now use $\mathbf{p}$ as our neutral expectation for likelihood computations.

From the vector $\mathbf{p}$, we define the vector $\mathbf{q}^{(m)} = (q_1^{(m)}, q_2^{(m)}, \ldots, q_K^{(m)})$, which represents a hypothetical distorted frequency spectrum consistent with a model of m sweeping haplotypes in an analysis window, with $q_1^{(m)} \geq q_2^{(m)} \geq \ldots \geq q_K^{(m)}$. Accordingly, the choice of K, while flexible for any analysis, most directly affects the resolution with which we can classify soft sweeps. For example, identifying a sweep on nine distinct, presently-sweeping haplotypes requires at minimum a $K = 10$ truncation, while a sweep on 14 haplotypes requires $K = 15$ and would likely appear as neutral under smaller truncations. We generate $\mathbf{q}^{(m)}$ by increasing the frequency of sweeping haplotype classes $\{q_1^{(m)}, q_2^{(m)}, \ldots, q_m^{(m)}\}$ at the expense of non-sweeping haplotype classes $\{q_{m+1}^{(m)}, q_{m+2}^{(m)}, \ldots, q_K^{(m)}\}$ in a heuristic manner. We note therefore that our approach is purely statistical and does not feature an underlying population genetic model, but is an attempt to capture the features in the haplotype frequency spectrum consistent with those expected from a sweep. The vector $\mathbf{q}^{(m)}$ is related to $\mathbf{p}$ by

$$q_i^{(m)} = \begin{cases} p_i + f_i \sum_{j=m+1}^{K} (p_j - q_j^{(m)}) & i = 1, 2, \ldots, m \\ U - \frac{i-m-1}{K-m-1}(U - \varepsilon) & i = m+1, m+2, \ldots, K \end{cases} \quad (1)$$

where f_i, with $\sum_{i=1}^{m} f_i = 1$ and $f_i \geq 0$ for $i = 1, 2, \ldots, m$, is a term defining the manner in which the mass associated with haplotype frequencies $\{p_{m+1}, p_{m+2}, \ldots, p_K\}$ in the neutral frequency spectrum is distributed

among $\{p_1, p_2, \ldots, p_m\}$ to generate the sweep frequency spectrum of the alternative model; U and ε are respectively the frequencies of the most and least frequent non-sweeping haplotype classes, $q_{m+1}^{(m)}$ and $q_K^{(m)}$.

We can define f_i in multiple ways. Choosing $f_i = 1/m$ (model A) generates an alternative model in which value is uniformly added to each of $p_1, \ldots, p_m$. We can also specify a distortion in which value is added proportionally to each sweeping haplotype frequency, where $f_1 > f_2 > \ldots > f_m$, such as $f_i = (1/i)/\sum_{j=1}^{m} 1/j$ (model B), $f_i = (1/i^2)/\sum_{j=1}^{m} 1/j^2$ (C), $f_i = e^{-i}/\sum_{j=1}^{m} e^{-j}$ (D), or $f_i = e^{-i^2}/\sum_{j=1}^{m} e^{-j^2}$ (E). The latter non-uniform models may provide a more accurate representation of the haplotype frequency spectrum following a sweep, as sweeping haplotypes, in contrast to neutral haplotypes, may not exist at similar frequencies to one another (see Figure 1, right column). As such, we primarily use model (D) for inferences in the *Results* The choices of U and ε determine the frequency of the non-sweeping haplotype classes in the alternative model. For $U > \varepsilon$, we make the simplifying assumption that the value of $q_i^{(m)}$ decreases linearly for $i = m+1, m+2, \ldots, K$, whereas $U = \varepsilon$ constrains all $q_i^{(m)}$ to equal ε for $i = m+1, m+2, \ldots, K$. Regardless of the choice of U and ε, their relationship with each other and p_{m+1} is necessarily $p_{m+1} \geq U \geq \varepsilon$. We also note that $\mathbf{q}^{(K)} = \mathbf{p}$ by definition, illustrating that the null (neutral) model is nested within the alternative (sweep distortion) model.

For each analysis window, we must finally obtain a vector of counts $\mathbf{x}$, observed for the most frequent K haplotypes. We define $\mathbf{x} = (x_1, x_2, \ldots, x_K)$, where elements are once again arranged in descending order, with $x_1 \geq x_2 \geq \ldots \geq x_K$. We normalize each x_i to satisfy the constraint $\sum_{i=1}^{K} x_i = n$, where n is the number of haplotypes in the sample.

Using the model haplotype frequency spectra $\mathbf{p}$ and $\mathbf{q}^{(m)}$ in conjunction with the observed vector of counts $\mathbf{x}$ for the most-frequent K haplotypes in a particular genomic window, we define likelihood functions, which are based on the multinomial distribution. Our use of the multinomial distribution is reasonable as it describes the probability of observing the vector of haplotype counts ($\mathbf{x}$) across K haplotype categories given the vector of respective haplotype frequencies ($\mathbf{p}$ or $\mathbf{q}^{(m)}$). The likelihood of the model parameters under the null hypothesis (neutrality) given the haplotype counts in an analysis window, equivalent to the probability of obtaining the observed haplotype counts $\mathbf{x}$ given $\mathbf{p}$ and K, is

$$\mathcal{L}_0(\mathbf{p}, K; \mathbf{x}) = \prod_{i=1}^{K} p_i^{x_i}, \tag{2}$$

whereas the likelihood under the alternative hypothesis (sweep distortion) is

$$\mathcal{L}_1(\mathbf{p}, K, \varepsilon, m; \mathbf{x}) = \prod_{i=1}^{K} \left[q_i^{(m)}\right]^{x_i}. \tag{3}$$

Therefore, the log-likelihoods are

$$\ell_0(\mathbf{p}, K; \mathbf{x}) = \sum_{i=1}^{K} x_i \log(p_i) \tag{4}$$

and

$$\ell_1(\mathbf{p}, K, \varepsilon, m; \mathbf{x}) = \sum_{i=1}^{K} x_i \log(q_i^{(m)}). \tag{5}$$

We optimize $\ell_1(\mathbf{p}, K, \varepsilon, m; \mathbf{x})$ over $m \in \{1, 2, \ldots, K\}$ and $\varepsilon \in [1/(100K), U]$, keeping U fixed, to find

$$(\widehat{m}, \widehat{\varepsilon}) = \underset{(m,\varepsilon)}{\operatorname{argmax}}\, \ell_1(\mathbf{p}, K, \varepsilon, m; \mathbf{x}).$$

Thus, our test statistic is defined as

$$T = 2\{\ell_1(\mathbf{p}, K, \widehat{\varepsilon}, \widehat{m}; \mathbf{x}) - \ell_0(\mathbf{p}, K; \mathbf{x})\}. \tag{6}$$

Each analysis window in the genome is assigned a test statistic in this manner, and larger test statistics indicate greater support for a sweep in the window (*i.e.*, greater distortion toward few moderate- or high-frequency haplotypes). Because in the process we also infer the most likely number of presently sweeping haplotypes $\widehat{m}$ to yield the underlying distorted haplotype spectrum, our approach can also be used to quantify the softness of an identified sweep.

Results

We first performed experiments with simulated data in which we generated populations based on non-equilibrium human demographic models inferred with `smc++` [Terhorst et al., 2017], covering a variety of neutral and selection scenarios. These demographic models consisted of a history based on that of the CEU European population, featuring a prominent bottleneck about 2000 generations prior to sampling, and a sub-Saharan African history resembling that of the YRI population, characterized by large population size with a recent expansion; these attributes of both models are consistent with previous estimates [Gravel et al., 2011, Gronau et al., 2011]. Using these simulations, we measured the power of the T statistic, and contextualized our results by comparing T to other popular methods, comprising (in order of increasing sophistication) H12 [Garud et al., 2015], nS_L [Ferrer-Admetlla et al., 2014], *SweepFinder2* [Nielsen et al., 2005, Huber et al., 2016, DeGiorgio et al., 2016], and *Trendsetter* [Mughal and DeGiorgio, 2019], across hard and soft sweep scenarios. We applied *Trendsetter*, a machine learning method that uses information on the spatial autocorrelation of statistics, using both its standard six-statistic approach—incorporating pairwise sequence difference $\widehat{\pi}$, squared correlation coefficient r^2, number of distinct haplotypes N_{haps}, and

the expected haplotype homozygosity statistics H1, H12, and H2/H1—and using contiguous T statistic analysis windows as input ("*T-Trendsetter*").

To test the versatility of the T statistic, we probed the effects of various confounding factors on T statistic inferences. Foremost among these was admixture, which can mimic the signature of a sweep when a donor population of small effective size contributes ancestry to the sampled population [Lohmueller et al., 2009, Harris et al., 2018a]. We also computed the value of the T statistic in regions of low mutation and recombination rates to evaluate whether their associated reductions in haplotypic diversity could be mistaken for sweeps. Due to its pervasiveness in genomes, we then generated models of background selection to determine whether it can affect the value of the T statistic, as background selection has been implicated as a confounding factor when searching for selective sweeps [Charlesworth et al., 1993, 1995, Seger et al., 2010, Nicolaisen and Desai, 2013, Cutter and Payseur, 2013, Huber et al., 2016]. Additionally, we evaluated the effects of dataset confounding factors, exploring the impact of missing data and small sample size on power. Complementing the power analyses, we evaluated the performance of our method in terms of its ability to infer the number of sweeping haplotypes at the time of sampling ($\widehat{m}$). We use $\widehat{m}$ as a proxy for the number of distinct presently sweeping haplotypes in the population (model parameter m), which itself is a proxy for the true number of initially sweeping haplotypes (ν), an unknown parameter. Finally, we applied our method to data from the 1000 Genomes Project [Auton et al., 2015] and the *Drosophila* Genetic Reference Panel [DGRP; Mackay et al., 2012] to measure our ability to properly identify and classify selective sweep candidates.

Detection and characterization of selective sweeps

We measured the power of our likelihood ratio test statistic (T) to differentiate selective sweeps from neutrality across diverse simulated scenarios using a sliding analysis window approach. For hard sweeps, as well as soft sweeps on $\nu = 4$ initially sweeping haplotypes, we compare the power of T to that of the four alternate methods. Larger values of the T statistic for an analysis window indicate a greater departure from the neutral haplotype frequency spectrum, and therefore a greater probability of a sweep within that genomic region. To measure power, we first simulated 1000 neutral replicates of one Mb chromosomes under the CEU and YRI demographic models. From these simulations, we obtained each model's expected truncated neutral haplotype frequency spectrum $\mathbf{p} = (p_1, p_2, \ldots, p_K)$, which was the basis of our likelihood computations (see *Definition of statistic*). The spectrum $\mathbf{p}$ for a model represents the mean across all genomic windows of all replicates, truncated at a particular value of K. Thus, $K = 10$ indicates the spectrum of the most frequent 10 haplotypes in a genomic window, whose frequencies are labeled p_1 through p_{10}. To assess power, we computed the T statistic for each 117-SNP (see *Materials and methods*) genomic window of each simulated

neutral replicate. We solely retained the maximum value of the T statistic across all windows for each neutral replicate, and similarly retained the maximum T statistic across each replicate of each selection scenario we tested. In our experiments, we assessed power at the 1 and 5% false positive rates (FPRs), meaning that we measured the proportion of selection replicates respectively exceeding the top 1 or 5% of T statistics within the neutral distribution.

The T statistic has high power to detect a recent hard sweep ($\nu = 1$ sweeping haplotype) affecting the CEU-based demographic history, provided that the selection coefficient is at least $s = 0.005$ (Figure 2, top). At both the 1 and 5% (Figure S1) FPRs, the T statistic reliably detects hard sweeps beginning between 1000 and 1500 generations before sampling, with the strongest sweeps extending the lower bound of this range to 200 generations (Figure 2, rightmost column). The power of the T statistic attenuates for more ancient sweep events because haplotype identity surrounding the selected site decays over time in the population as mutation and recombination generate new haplotypes. Additionally, power to detect the most recent and weakest sweeps is low because sufficient time has not elapsed for the selected haplotype to reach high frequency. Sweeps on $s < 0.005$ are specifically difficult to detect due to their smaller footprint and shorter time over which elevated haplotype homozygosity persists as selection proceeds. For these reasons, there is no point in time at which the T statistic can detect these sweeps (Figure S3).

Across simulated CEU selection scenarios, each of the alternate methods we examined was subject to the same power limitations as the T statistic, which outperformed all except for the more sophisticated *Trendsetter*. The relative performance of all methods indicates that pairwise sequence identity tract length, which nS_L measures, is the most volatile sweep signal, decaying more rapidly than others. nS_L consistently had the lowest power of all tested methods, and reached high power only for the most recent strong hard sweeps (Figures 2 and S1, top-right), quickly losing power as the sweep footprint eroded. Likewise, H12 never matched the power of the T statistic except in detecting the strongest sweeps, but in using a fixed window size retained somewhat more power than did nS_L. *SweepFinder2* displayed greater power than H12, with higher maxima and longer signal duration. Despite not using haplotype information, *SweepFinder2* incorporates a population-genetic model of a recent hard sweep, which results in more power than methods which do not. Finally, *Trendsetter* had easily the greatest power to detect hard sweeps under the CEU model, losing little resolution for sweeps up to 4000 generations old. Using evidence from multiple signals may therefore be necessary to maximize power, as the strengths of each component statistic can complement the others' weaknesses across different parameter configurations.

The T statistic achieves greater power for hard sweeps on simulated YRI demographic models than for CEU models across all tested scenarios (Figures 2 and S1, bottom). This increased power is due to the greater effective size of African relative to European human populations, which results in greater background

haplotype diversity and therefore increased prominence of selective sweeps. Furthermore, because the size of the YRI population is an order of magnitude greater than that of the CEU for hundreds of generations, the population-scaled selection coefficient $\sigma = 4Ns$ for YRI remains much larger, resulting in a stronger sweep. Accordingly, power declines more slowly for older sweeps, and remains for sweeps as old as 4000 generations before sampling. All methods therefore show greater power when applied to simulated YRI data. Notably, while *Trendsetter* and the T statistic display excellent performance once again, *SweepFinder2* demonstrates consistently superior power for older sweeps under the YRI model, and this power scarcely decays. This suggests that *SweepFinder2* may be more susceptible to demographic history than other methods, losing considerable power under the CEU bottleneck [Jensen et al., 2005, Huber et al., 2016]. Meanwhile, the choice of K truncation—15, 20, or 25 (Figure S5)—yielded little change in power to detect simulated sweeps from $s \in [0.01, 0.1]$ relative to our highlighted value of $K = 10$ for either population model (Figures 2 and S1, fourth column). However, we note that power is slightly greater as K decreases, and so we recommend the use of smaller K truncations where possible.

For soft sweeps from selection on standing genetic variation (SSV, $\nu \in \{2, 4, 8, 16, 32\}$; Figure S4), the power of the T statistic attenuates more rapidly than for hard sweeps, and T rarely reaches values as large, especially for weaker sweeps. Under both CEU (Figure S4, top) and YRI (Figure S4, bottom) demographic histories, trends in power remain consistent regardless of the number of sweeping haplotypes, with maximum power of T achieved once again for sweeps between 1000 and 1500 generations old (or up to 3000 for YRI); however, power declines as the number of sweeping haplotypes increases. Assessing power at the 5% FPR indicates that we nonetheless maintain sufficient differentiation between sweeps and neutrality for up to $\nu = 4$ distinct initially sweeping haplotypes for CEU models, or up to $\nu = 8$ for YRI models. To better understand the relationship between power and ν, we tracked the mean number of distinct sweeping haplotypes through time for simulated soft sweeps across each selection strength range and choice of ν using an *in-silico* barcoding approach (see *Materials and methods*). We found that weak soft sweeps frequently lose most of their sweeping haplotypes by the time of sampling, undergoing a hardening [Wilson et al., 2014] during the early stages of the sweep when the beneficial allele's frequency is low and still subject to genetic drift (Figures S13 and S14). After the beneficial allele becomes established in the population, the number of sweeping haplotypes remains generally stable. Because weaker sweeps require more time to establish, this provides more time for haplotypes to be lost, and for fewer sweeping haplotypes to be sampled, relative to stronger sweeps. Thus, the T statistic can have greater power for our weaker simulated soft sweeps from larger ν than for stronger soft sweeps because the former case ultimately yields a more distinct sweep signal with fewer high-frequency haplotypes, while the latter often results in scenarios of high haplotypic diversity that are difficult to distinguish from neutrality.

The power of each alternative method responded to soft sweep ($\nu = 4$) scenarios in the same manner as that of T. Methods generally had poor to middling performance at the 1% FPR for the CEU history (Figure 3, top), but decent power at the 5% FPR, especially for sweep strengths between 0.005 and 0.1 (Figure S2, top), while the power of all methods was improved for the YRI model (Figures 3 and S2, bottom). However, *SweepFinder2* retains little power to detect soft sweeps, and lost power proportionally to the number of sweeping haplotypes at the time of sampling, as it is specifically formulated to detect hard sweeps through distortions in the site frequency spectrum, and soft sweeps do not dramatically alter the site frequency spectrum [Pennings and Hermisson, 2006b]. Expectedly, *Trendsetter* was still the most powerful method, with T and H12 following closely behind for recent sweeps, and nS_L lagging once again. Thus, the demographic and selective histories of the sampled population play an important role in the power of the T statistic and other sweep detection methods. Our results nonetheless indicate that the T statistic is flexible as to the selection scenarios it can distinguish from neutrality, and detects recent sweeps especially well for the relatively little computation time it requires.

Selective sweeps produce elevated values of the T statistic along the simulated chromosome that on average peaks in the region surrounding the site under selection (Figures S6 and S7, first and third rows). Furthermore, T remains elevated beyond the 900 kb bounds that we examined, indicating that on average, the shape of its distribution in a genomic region, as well as its overall elevated value, may be used to distinguish selection from neutrality. A signal peak often exists even for scenarios in which we do not have high power, though its maximum associated value remains small on average. Because neutral regions are likely to feature plateaus rather than peaks in the value of the T statistic, our observations illustrate the potential importance of considering the correlation in signal between windows to identify more subtle selection signatures. This is especially important for soft sweeps, which lose prominence proportionally to the number of sweeping haplotypes, but still produce a peak-like distortion of local T statistic values.

To motivate this point, we reevaluated the power of our T statistic for sweeps on $\nu = 1$ and $\nu = 4$ haplotypes with the penalized multinomial regression method, *Trendsetter* [Mughal and DeGiorgio, 2019], as it directly incorporates the genomic spatial distribution of sweep (summary) statistics in its inferences (Figures S8 and S9). Using T as the sole statistic in the statistical learning protocol across 201 11-SNP windows (*Trendsetter*'s standard approach; see *Materials and methods*), we considerably improved our ability to detect more ancient sweep signatures than in our standard application of T, especially under the CEU model, yielding power comparable to the full six-statistic approach. Moreover, we saw little change in power to detect recent sweeps, and power was improved overall for soft sweeps. When we instead used our typical 117-SNP windows for T, and 75 windows total, the T-*Trendsetter* approach yielded a uniformly enhanced power profile whose trends matched our original results with the standard T statistic, but with greater overall

power, especially for soft sweeps (Figure S9). Thus, the spatial distribution of the T statistic provides an informative sweep signature that can be used to isolate sweep regions from neutrality. By learning this spatial distribution, we can enhance the power of T to detect sweeps that may be overlooked using an isolated per-window approach.

In addition to evaluating the power of the T statistic, we measured the ability of our approach to infer the number of presently sweeping haplotypes ($\widehat{m}$) at the site under selection. The ability to infer $\widehat{m}$ is a result of optimizing the likelihood function ℓ_1 over all possible m distortion models for the chosen truncation K (see *Definition of statistic*). In Figure 4, we show the distribution of T statistics with their associated haplotype frequency spectra and $\widehat{m}$, for each of 1000 neutral, mixed hard sweep ($s \in [0.001, 0.5]$, $\nu = 1$), and mixed soft sweep ($\nu = 4$) replicates, under both the CEU and YRI models (same data as Figures 2 and 3; $t = 1500$ for CEU, and $t = 2500$ for YRI, representing points of maximum power). Relative to neutrality (Figure 4, left), we more often assign smaller $\widehat{m}$ to sweep simulations (Figure 4, center and right). This result fits with the expectation that under a sweep, the first few haplotype classes exist at elevated frequency relative to the remaining classes, and this also translates to larger values of T for those replicates. Accordingly, sweeps that have weaker signatures due to their age or selection coefficient are not only difficult to distinguish from neutrality, but also difficult to accurately classify with $\widehat{m}$, yielding patterns that fit within the neutral distribution. We found that trends were highly congruent between the CEU and YRI sweep models, but the large neutral background diversity for YRI made it less likely that we would infer a small $\widehat{m}$ in the absence of a sweep.

To further understand the sweep classification properties of the T statistic, we generated box plots summarizing the distribution of $\widehat{m}$ across the more prominent strong sweep scenarios ($s \in [0.01, 0.1]$) we previously analyzed (Figures 2, 3, and S4). In this way, we were able to better understand our ability to correctly classify sweeps as hard or soft, especially because our trajectory results (Figures S13 and S14) provided us with an expectation of the number of remaining sweeping haplotypes at the time of sampling for prominent sweep scenarios. We found that sweeps initiated from larger ν were more likely to be classified as soft using our $K = 10$ truncated spectra, but frequently we found that the median inferred $\widehat{m}$ for prominent soft sweeps was one, consistent with a hard sweep (Figures S10 and S11). Regardless of ν, the spatial signature of $\widehat{m}$ along the chromosome forms a valley surrounding the site of selection that mirrors the signal peak when a sweep is detectable (Figures S6 and S7). These results suggest that our present approach may therefore be more accurate as a binary classifier (hard versus soft), though we still assign soft sweeps on a continuum.

Because phasing haplotypes may not be possible in all cases, such as in the study of non-model organisms, we sought to expand our application of the T statistic to unphased multilocus genotype (MLG) data. To

evaluate power for MLGs, we used the previous simulated human demographic model replicates of prior experiments (represented in Figures 2 and 3), merging each individual's two haplotypes. Whereas haplotypes are character strings indicating the state of a biallelic SNP as either reference or alternate along a region of one copy of an individual's genome, MLGs have three possible states for each biallelic SNP—homozygous reference, homozygous alternate, or heterozygous—and half the sample size of phased haplotypes. We found that, as with the transition between phased and unphased data for haplotype homozygosity statistics [Harris et al., 2018a, Harris and DeGiorgio, 2019], trends in power for the unphased application of the T statistic were wholly consistent with those of the phased application, for both hard and soft sweeps (Figure S15). The smaller size of the MLG samples resulted in slight decreases in power for each sweep scenario, as well as smaller values of the T statistic relative to the phased application, but our results indicate that selective sweeps may be reliably identified nonetheless without the need to phase haplotypes. Likewise, we found that the T statistic applied to MLGs could generate inferences of sweep softness from $\widehat{m}$ that matched those of haplotype data, further underscoring the parallel performance of our approach on unphased data (Figure S16).

Effects of confounding factors of the T statistic

There are multiple genetic and non-genetic factors that may affect the ability of sweep statistics to properly localize a selection signature. Though these factors are varied in origin, they may each reduce genetic diversity locally, and spuriously generate patterns similar to those of selective sweeps. Accordingly, we examined the effects of introducing these confounding factors to our simulated data, allowing us to understand the scenarios for which T is robust and susceptible to misclassifying sweeps. Among genomic factors, we observed the impact of admixture into the sampled population, reductions in mutation and recombination rates, and background selection. Common non-genomic factors that can change the interpretation of sweep statistics are missing data, small sample sizes, and reliance on a misspecified demographic model.

We begin with admixture, which can be pervasive in natural populations [Hudjashov et al., 2017, Kopatz et al., 2017, Browning et al., 2018, Barría et al., 2019]. Previous work has shown that under certain scenarios, admixture from an unsampled donor population can lead to reductions in haplotypic diversity across the genome of a study population [Harris et al., 2018a]. Specifically, admixture from a diverged donor population of small effective size into the sampled study population may introduce large tracts of homozygous sequences that methods may interpret as a selection signal. To assess the extent to which the T statistic—which makes use of the study genome's background haplotypic patterns—is affected by admixture, we simulated neutral replicates under the CEU and YRI models receiving pulse admixture from a highly diverged donor ($\tau = 2N = 2 \times 10^4$ generations prior to sampling). Our tested admixture proportions were $\alpha \in \{0.05, 0.1, \ldots, 0.4\}$

at $t_{\text{adm}} = 200$ generations prior to sampling, and we examined donor population effective sizes of $N_{\text{adm}} \in \{10^3, 10^4, 10^5\}$ diploids (roughly 1/10, equal to, and tenfold the effective size of the sampled populations).

We found that admixture had a considerable effect on the haplotype frequency spectrum of the target population, and assessed this effect in two ways. First, we computed the T statistic for each admixture scenario, but using an unadmixed background $\mathbf{p}$ spectrum (Figures S17 and S19). Our results demonstrate that admixture from a donor with small effective size yields the expected haplotypic diversity reduction in the sampled population, producing inflated T statistics. In contrast, gene flow from medium- and large-sized donors rarely produced large values of T, except for large-donor admixture at $\alpha \leq 0.1$. We attempted to address the confounding effects of admixture by computing T from an appropriately matched admixed background $\mathbf{p}^\alpha$ spectrum for each tested value of α (Figures S18 and S20). Using a matched $\mathbf{p}^\alpha$ resulted in T statistic distributions under admixture that more closely resembled unadmixed distributions. Regardless of scenario, the T distribution deriving from a $\mathbf{p}^\alpha$ spectrum informed by admixture resulted in median values of T for admixed scenarios closer to the median for unadmixed replicates, especially for small-donor admixture. While this was uniformly beneficial for the CEU model, $\mathbf{p}^\alpha$ overcorrected for large-donor scenarios under the YRI model. The susceptibility of our T statistic to confounding admixture is a consequence of using a model that does not account for mixed ancestry in the target population. Instead, the reliance of our approach on an average neutral background spectrum, means that we are not capturing the higher moment effects of admixture on the admixed T statistic distribution, such as its variance. Even so, we expect that in human populations, admixture is unlikely to be as extreme as in our simulated examples, and is likely to feature populations that are less diverged from one another, and of closer effective size to one another, reducing its overall detrimental impact when searching for sweeps.

Because natural genomes may feature wide variation in recombination and mutation rates, we sought to determine the effect of such variation on the value of T in the absence of selection in order to quantify its potential misleading effect on T. In comparison to standard simulations featuring $\mu = 1.25 \times 10^{-8}$ and mean $r = 10^{-8}$ drawn from an exponential distribution, we generated simulated replicates with $\mu = 1.25 \times 10^{-9}$ and mean $r = 10^{-9}$, reduced by one order of magnitude (see *Materials and methods*). Reducing the mean recombination rate had the anticipated effect of slightly inflating the distribution of T relative to normal values of r, a result of the reduction in genetic diversity in regions where new haplotypes rarely arise (Figure S21). In contrast, reducing the mutation rate by an order of magnitude resulted in a deflation of T statistic values relative to the original rate. This is because our SNP-delimited windows become physically wider when SNP density is reduced, leading to the incorporation of SNPs in lower mean linkage disequilibrium, and therefore more haplotypic diversity. Reducing both μ and mean r led to an intermediately deflated T statistic distribution. This result may suggest that mutation rate variation is more important than

recombination rate variation in determining the T statistic value under SNP-delimited windows (Figure S21). However, because we already draw recombination rates from a distribution in our default protocol, it is possible that we observed a greater effect by changing μ because our reduced-μ scenarios represent a greater departure from standard simulations. Thus, we caution that recombination rate variation should not be ignored as a source of false signals in analyses with any sweep statistic.

We next examined background selection scenarios for both haplotype and MLG data to determine whether the loss of genetic diversity associated with linked purifying selection could spuriously yield elevated values of the T statistic. Simulating one Mb chromosomes as previously under both human demographic models, we found that background selection had no discernible effect on the distribution of T relative to neutrality, even as we reduced mean r by two orders of magnitude to 10^{-10} across the central gene. We determined this by observing the receiver operating characteristic curves comparing neutral scenarios to those in which a central gene experiences strong ($s = -0.1$) background selection for the duration of the simulation (see *Materials and methods*). For both the CEU (Figure S22, top) and YRI (Figure S22, bottom) populations, across central genes of size 11 kb (Figure S22, left), 55 kb (Figure S22, center), and 110 kb (Figure S22, right), we see that all curves fit tightly along the diagonal, indicating no difference between compared replicate sets. Therefore, we expect that the presence of background selection, for which we do not explicitly account in our model, should not affect inferences with the T statistic.

Among the most common non-genomic confounders that may be encountered in analyses of natural populations is the presence of missing data. That is, sites for which the allelic state is indeterminate. We evaluated the performance of the T statistic for sequences with missing data at polymorphic sites by randomly removing alleles from our existing neutral replicates (see *Materials and methods*). To address missingness in our data, we modified our scan using two corrective approaches. First, we removed polymorphic sites with greater than 5% missing data, and second, we incorporated all remaining missing alleles as a new character "N", thereby conservatively diversifying the remaining set of haplotypes relative to no missing data. Following this approach, we compared the distributions of T with and without missing data. We found that introducing missing data had little effect on the distribution of T statistic values under neutrality (Figure S23). We expect that as long as sufficient polymorphism remains in the sample, that missing data is unlikely to yield false sweep inferences, and in extreme cases, is likely to resemble the effect of decreased mutation rate due to the depletion of allelic information.

A limitation of haplotype-based approaches is that their power derives from sample size (n). That is, sufficient diversity must be captured in a sample in order to distinguish between the unique signals of neutrality and selection. As sample size decreases, subtle signatures of selection recede into the genomic background and become imperceptible. To determine the minimum sample size to which the T statistic

should be confidently applied, we resampled our existing hard sweep replicates (see Figure 2) for reduced sample sizes $n' \in \{20, 50, 100\}$ haplotypes, corresponding to 1/10, 1/4, and 1/2 the original sample size of $n = 200$. Reducing sample sizes led to a reduction in the minimum detectable range of selection coefficients s, and in the range of sweep ages t over which the T statistic had power (Figure S24). While power scarcely changed for larger samples of size $n' = 100$ relative to $n = 200$ (as in our application to MLG data), we are unable to reliably detect even strong sweeps older than $t = 500$ generations for the CEU model when only $n' = 20$ haplotypes are sampled (Figure S24, top-right). Analysis of the YRI model was more accommodating to sample size reduction owing to the preexisting greater ease of detecting sweeps for populations with large effective sizes, but power quickly drops when small selection coefficients ($s < 0.005$) are included (Figure S24).

A practical consideration when applying sweep statistics is inferring an accurate demographic model. A proper model can be used to generate simulated replicates from which p-value cutoffs and false discovery rate thresholds may be assigned. We subsequently demonstrate this in our own *Application to empirical datasets*. To motivate the selection of an appropriate model, we show the effect on the neutral T statistic distribution of using non-ideal demographic models. First, we generated T statistic distributions for CEU and YRI under constant-size models. Here, the constant sizes were equal to the effective size of the populations under each model (see *Materials and methods*). Because these models included no population size fluctuations, they provided uniformly deflated T statistic distributions relative to `smc++` models, with the constant-size CEU model especially underestimating values relative to its more accurate counterpart (Figure S25). We also examined T statistic distributions resulting from the popular Gravel et al. [2011] model, which is based on the site frequency spectrum. Relative to the `smc++` model, the T statistic distribution for the Gravel et al. [2011] CEU model model was comparable, but the YRI model, consisting of only two phases (constant size followed by twofold expansion), resulted in much smaller values of T (Figure S26). Choosing a demographic model that captures both recent and ancient history is therefore important [Beichman et al., 2017], and model choice should be approached with caution.

Application to empirical datasets

We searched for candidate selective sweeps in human and *D. melanogaster* datasets using the T statistic, choosing these datasets because of their high quality, size, and availability of phased haplotypes. Specifically, the 1000 Genomes Project [Auton et al., 2015] dataset contains no missing data, as all allelic states have been imputed. Meanwhile, the DGRP dataset [Mackay et al., 2012] provides a classic invertebrate model whose properties deviate considerably in history and genomic architecture from the mammalian model of humans. For each protein- and RNA-coding gene in each study population, we obtained values of T using inferences from a $K = 10$ truncation, and assigned a p-value to each of the top 40 candidate genes based

on the window of maximum T overlapping that gene (Tables S1-S3). For windows to be associated with a gene, their central SNP must lie between the transcription start and stop sites of the gene. Additionally, we assigned an $\widehat{m}$ value to each gene using both $K = 10$ and $K = 20$ truncations. Analysis windows for scans of human data were of size 117 SNPs, advancing by 12-SNP increments, while windows for *D. melanogaster* analysis were 400 SNPs in size (as in the analyses of Garud et al. [2015] and Harris et al. [2018b]) with a step size of 40 SNPs. To eliminate the effect of background LD on inferences, window sizes were based upon the minimum physical interval across which LD decayed beyond one-third of its value between SNPs separated by one kb (see *Materials and methods*). Following our successful application of T to simulated unphased MLG data, we analyzed the human 1000 Genomes Project data as MLGs by manually merging individuals' two haplotypes. This was unnecessary for the DGRP data, as individuals are inbred. For haplotype data, we also determined false discovery rate (FDR) thresholds for each population based on simulated replicates, inferring T statistic cutoffs at a 5% FDR for all study populations (Table S4).

For human data, we examined the CEU and YRI populations (Tables S1 and S2), matching the demographic models used in our simulations. Though few of our top 40 sweep candidates produced a significant p-value, all easily exceeded their population's 5% FDR (Table S4). Among these candidates, hard sweeps predominated within either population, and we found that this depended somewhat on our K truncation. For $K = 20$, hard sweeps comprised all but two top candidates among the CEU, and 67.5% of top candidates among the YRI, while for $K = 10$, we did not classify any top CEU or YRI candidates as soft from phased haplotypes. Additionally, $K = 20$ candidate soft sweeps, except for *BTNL2* in YRI ($\widehat{m} = 6$) featured only three or fewer sweeping haplotypes. These results indicate that the T statistic is more sensitive to harder sweeps than to softer ones, which is a consequence of the greater distortion in the haplotype frequency spectrum of hard sweeps relative to soft sweeps. This finding matches our simulated results, in which the value of T was proportional to the number of sweeping haplotypes in the population. Moreover, we find that our choice of K truncation impacts our ability to classify sweeps as soft, with a greater distortion of haplotype classes two through m required for a sweep in a $K = 10$ truncated spectrum to be classified as soft relative to $K = 20$. Regardless of truncation, the increased presence of candidate soft sweeps in YRI relative to CEU mirrors our observation from simulated data that the T statistic has greater power to detect softer sweeps for populations that have not experienced a bottleneck in their history. Furthermore, these patterns corroborate results from the H12 analysis of this dataset [Harris et al., 2018a], which found more hard sweeps than soft in the CEU population, and among top candidates generally.

Across both the CEU and YRI populations, we were able to recover most of the top 40 candidates from the haplotype data within the MLG data, indicating the reliability of using MLGs for inference with the T statistic in natural populations when phased data are unavailable. The MLG results primarily deviated from

the haplotype results when classifying candidates as hard or soft. Multiple candidates that were inferred to be hard sweeps from the haplotype data were classified as soft from their MLG spectra, particularly for the $K = 20$ truncation. These candidates include *XIRP2* and *BCAS3* in the CEU population, as well as *ITGAE*, *SUGCT*, *NNT*, and *HLA-DPB2* in the YRI population. We examine the latter candidate more closely in Figures 5 and S30. These differing inferences may arise from the slightly different interpretation of $\widehat{m}$ between phased and unphased data. For phased data, $\widehat{m}$ refers to the number of sweeping haplotypes, whereas for unphased, it measures the number of MLGs involved in the sweep, which may be different for the same genomic window between the different data types if MLG frequencies are at or near their expected proportions under Hardy-Weinberg equilibrium. We also note that multiple top candidates in the MLG data inferred as soft are simply not present among top haplotype candidates, indicated by the absence of a turquoise-colored background in Tables S1 and S2. We consider the application of the T statistic to MLGs further in the *Discussion*.

Among top sweep candidates in human data were expected results, including a hard sweep ($\widehat{m} = 1$) at the cluster of genes on CEU chromosome 2 comprising *LCT*, *MCM6*, *DARS*, and *ZRANB3* (minimum p-value 3×10^{-6}), related to a well-documented adaptation to milk-based diets in European populations [Bersaglieri et al., 2004]. Additionally, we found two noteworthy top candidates in CEU that have not been explicitly described as sweeps previously, *RSPH3* and *ZNF211* (both $\widehat{m} = 1$). *RSPH3* encodes a radial spoke protein that is integral in the structure of 9 + 2 motile cilia across diverse cell types, including flagellated cells [Teves et al., 2016], and so we speculate that selection here could be related to ancient sperm competition in humans [Leivers et al., 2014]. *ZNF211* is among a diverse set of zinc-finger genes whose products are believed to participate in the inactivation of endogenous retroviruses, parasitic mobile DNA whose effects can be deleterious to their hosts [Lukic et al., 2014]. We recovered *SYT1* ($\widehat{m} = 2$ for $K = 20$, $\widehat{m} = 1$ for $K = 10$; $p = 3 \times 10^{-6}$), *NNT* ($\widehat{m} = 1$), *HEMGN* ($\widehat{m} = 1$), and *RGS18* ($\widehat{m} = 2$ for $K = 20$, $\widehat{m} = 1$ for $K = 10$) in YRI, which have all received attention as potential adaptive targets [Voight et al., 2006, Pickrell et al., 2009, Fagny et al., 2014, Harris et al., 2018a]. Our significant candidates, *SPRED3* and *ITGAE* have also been previously identified [Granka et al., 2012, Ayub et al., 2013, Grossman et al., 2013], though the effect of selection at these genes has not yet been elucidated. Both populations yielded *HLA* genes as top sweep candidates, overlapping at *HLA-DRB5* ($\widehat{m} = 1$), whereas *HLA-DPB1* ($\widehat{m} = 1$) was exclusive to CEU and *HLA-DPB2* ($\widehat{m} = 1$) was exclusive to YRI. This shared signal supports the recent evidence [Albrechtsen et al., 2010, Goeury et al., 2017] that sweeps at HLA loci, including those which we describe here, were important in the development of modern genetic diversity in human immune-related genes.

In Figures 5 and S30, we take a closer look at top candidate sweeps uncovered in our scan of the 1000 Genomes dataset [Auton et al., 2015] for both $K = 10$ and $K = 20$ truncations, across both haplotypes

and MLGs. Each top candidate fell within a well-defined T statistic peak region surrounded by regions of low signal, and T-statistic spatial signatures were consistent between all data types. First, we found *SYT1* as a near-significant ($p = 3 \times 10^{-6}$) top sweep candidate in the YRI population, featuring both $\widehat{m} = 2$ sweeping haplotypes and $\widehat{m} = 2$ MLGs involved in the sweep at the window of maximum signal for $K = 20$ (Figures 5 and S30, first row of first column). For the $K = 10$ truncation, the window of maximum signal contains only a single sweeping haplotype and is located within an adjacent, upstream sub-peak. *SYT1* is the cell surface receptor through which the type B botulinum neurotoxin of *Clostridium botulinum* bacteria enters human neurons [Connan et al., 2017], and so a sweep here may be involved in resistance to this infection [Harris et al., 2018a]. Next, we identified *HLA-DPB2* as an outlying hard sweep in YRI ($\widehat{m} = 1$) based on haplotypes and $K = 10$ data, but featuring three elevated MLGs within the window of maximum signal for $K = 20$ (Figures 5 and S30, second row). Looking at the $K = 20$ haplotype frequency spectrum, it is clear that one haplotype predominates, and equivalently, only one MLG predominates, but individuals heterozygous for the first haplotype and either the second or third comprise just under 20% of the population, leading to an inference of $\widehat{m} = 3$. *COL5A2* was the most outlying soft sweep candidate we identified in CEU using the $K = 20$ truncation, harboring $\widehat{m} = 2$ inferred sweeping haplotypes, but with a sevenfold disparity between their frequencies, which could occur due to a recombination event during the sweep, or a recurrent selected mutation [Hermisson and Pennings, 2017]. This gene has received little attention, but is located within a significantly overrepresented run of homozygosity associated with schizophrenia [Lencz et al., 2007]. Additionally, selection on collagen-related genes has been implicated in cold adaptations in European populations [Yudin et al., 2017]. Finally, we propose the spermatogenesis-associated protein *SPATA6L* as a hard sweep candidate in CEU. Our finding here of an isolated T peak fits with existing evidence of selection at other spermatogenesis proteins [Schrider and Kern, 2017], and with the result that European and sub-Saharan African populations are diverged at this locus, with selection in the hunter-gatherer Batwa population inferred here [Bergey et al., 2018].

We contextualize our results for outlying sweep candidates by illustrating the background haplotype frequency spectrum patterns we observed in regions of low T. In Figure S31, we highlight four regions each in the CEU and YRI populations with $T \approx 0$, 10, 20, or 30. In accordance with the expectation that classic selective sweep patterns are rare in the human genome [Hernandez et al., 2011], we observe that the majority of analysis windows had a small associated T, and accordingly resembled our example windows. We see from these examples that small peaks in the T statistic are common, and associated with haplotype frequency spectra that are distinct from those under selection, containing no high-frequency classes and an abundance of low-frequency classes of similar size. As we increase from $T \approx 0$ to $T \approx 30$, we see the spectra begin to distort and contain higher frequency classes, but this distortion is far from what we expect under a sweep.

Our scan of the North American DGRP population of *D. melanogaster* also identified expected sweep candidates among the top genic T statistic peaks. We note that while we were unable to establish statistical significance against a neutral model based on the DGRP demographic history of Duchen et al. [2013] (see *Materials and methods*), and only our top candidate, *CG11902*, exceeded the 5% FDR threshold (Table S4), our top candidates have literature support as potential adaptive targets. Foremost among functionally-characterized candidates was *Ace*, which encodes the acetylcholinesterase enzyme and has long been implicated in the development of resistance to organophosphate and carbamate insecticides within *D. melanogaster* [Menozzi et al., 2004, Karasov et al., 2010, Garud et al., 2015]. Contrary to previous studies alleging a soft sweep at *Ace* [Karasov et al., 2010, Garud et al., 2015], we found the greatest support for a model of only one sweeping haplotype. We identified another candidate hard sweep of similar magnitude at *Uhg1*, which also contributes to insecticide resistance, but to the organochlorine DDT [Pedra et al., 2004]. The methyltransferase-encoding gene *Pimet* emerged as the most prominent candidate soft sweep ($\widehat{m} = 3$) in our search using $K = 20$ ($\widehat{m} = 1$ for $K = 10$), and is central to the viral RNA degradation pathway that is subject to ongoing coevolution against pathogen incursion and deleterious transposable element activity [Kolaczkowski et al., 2011, Lee and Langley, 2012]. We finally highlight *ana3* as a candidate for adaptation in *D. melanogaster*. This prospective hard sweep affects a highly-conserved gene encoding a centriole protein fundamental to the structural integrity of basal bodies within cells [Stevens et al., 2009]. A sweep here may contribute to enhanced success in sperm competition, and fits with the expectation that sperm gene evolution is an ongoing and central part of positive selection in *D. melanogaster* [Nurminsky et al., 1998, Dorus et al., 2008, Wong et al., 2008, Yeh et al., 2012].

Discussion

We have proposed a likelihood-based approach to detect selective sweeps in whole-genome polymorphism data that is applicable to a variety of different demographic scenarios, classifies detected sweeps as hard or soft without relying on additional analyses or statistics, and is the first likelihood-based method to leverage distortions in the haplotype frequency spectrum to make these inferences. Each of these attributes is important because selective sweeps are multifaceted genomic signatures that are not always characterized by the presence of a single high-frequency haplotype [Jones et al., 2013, Wilson et al., 2017], may be ongoing or incomplete at the time of sampling [Vy and Kim, 2015, Vy et al., 2017], and may range in strength across multiple orders of magnitude [Messer and Neher, 2012, Nam et al., 2017]. Thus, our simulation experiments probed a realistically diverse complement of sweep scenarios likely to be relevant in a variety of study systems. Most importantly, the T statistic demonstrated high and consistent power and classification ability across examined parameters, highlighting its suitability to make inferences within variable contexts.

Expectedly, the T statistic achieved its maximum power for recent selective sweeps on fewer haplotypes, and lost power proportional to the extent of departure from these ideal conditions (Figures 2, 3, and S4). Because it is haplotype-based, the T statistic captures distortions in the haplotype frequency spectrum relative to neutral expectations. These distortions require time to establish, and decay over time as well. Thus, we found that for human demographic models, the T statistic could reliably identify sweeps that initiated between 500 and 2000 generations before sampling. For stronger sweeps ($s \geq 0.05$), power was consistently elevated across this range, but because weaker sweeps require more time to establish, this range narrows and power peaks for older sweeps as s decreases. Additionally, we uniformly had more power to detect sweeps under the YRI demographic model than the CEU. This is due to the severe bottleneck underlying the history of the CEU, as well as all non-African human populations. Bottlenecks may reduce the diversity of haplotypes within a population, reducing the distinctiveness of sweeps relative to neutrality, whereas population expansions have the opposite effect [Jensen et al., 2005, Campbell and Tishkoff, 2008]. Nonetheless, the T statistic could generally detect sweep strengths across all but our weakest selection coefficient range for sweeps aged between 1000 and 2000 generations under either demographic model, comprising events that in humans cover the period from 25,000 to 58,000 years ago, between the out-of-Africa event and the spread of agriculture [Lukić and Hey, 2012, Nakagome et al., 2015, Haber et al., 2019].

Importantly, ours is a powerful single-statistic approach that provides an ideal balance of detection capability and computational efficiency. Compared to popular recent methods, we found that the T statistic is generally more powerful than other single-statistic approaches, and is also sensitive to the same range of sweep times as they are (Figures 2, 3, S1, and S2). This highlights the usefulness of T as a substitute for other single-statistic approaches, which may miss sweeps that the T statistic can detect. Although T is likely to be underpowered relative to machine-learning methods, such as *Trendsetter*, analyses with T have no need to train a classifier, which may be computationally intensive when training a composite of multiple signals, and must be undertaken for each study scenario. However, using the T statistic within a machine learning framework can greatly enhance its performance. By learning the spatial distribution of T statistic signals within our T-*Trendsetter* construct, we were able to enhance the performance of our method considerably. Using a large number of small windows, we could extend our range of sweep detection to identify simulated selective events up to 4000 generations in age, reflecting ancient sweeps far older than the out-of-Africa expansion. In contrast, training a classifier from a smaller amount of standard 117-SNP windows greatly improved performance for recent soft sweeps relative to the unassisted T statistic (Figures S8 and S9). Our results suggest that optimizing the power of a machine learning approach relies not only on the choice of input statistics, but also on the manner in which those statistics are applied to make inferences.

The choice of simulated human demographic history did not impact our inferences on the number of currently sweeping haplotypes ($\widehat{m}$) in a population. Under equivalent sweep scenarios (s, t, and ν), our analyses yielded similar distributions of $\widehat{m}$ for both the CEU and YRI models, and could accurately identify soft sweeps, provided that at least two sweeping haplotypes remained in the population at the time of sampling (Figures 4, S10, and S11). We found that simulated soft sweeps were frequently assigned an inferred $\widehat{m}$ that was smaller than the ν with which we initiated the sweep. Furthermore, we observed that T statistic signal peaks were on average associated with valleys in $\widehat{m}$ regardless of ν (Figures S6 and S7). To better understand these results, we devised an *in-silico* barcoding approach to complement our existing simulations. We found that soft sweeps from selection on standing variation frequently lose the majority of their selected haplotypes within generations of selection start time t, meaning that many soft sweeps, especially for smaller s, appear hard at the time of sampling (Figures S13 and S14). This hardening effect [Wilson et al., 2014], due to genetic drift at the early stage of selection, affected both simulated CEU and YRI populations equally, corroborating our consistently similar $\widehat{m}$ observations between the two populations. This means that even if soft sweeps are the dominant mode of adaptation in human history [Schrider and Kern, 2017], there may be considerably more that can never be identified as soft.

As an attempt to improve the performance of the T statistic, We sought to examine whether the choice of sweep distortion model, based on the choice of f_i (see *Definition of statistic*), would affect our inferences. Ultimately, using our YRI simulations as a basis, we found that all of our tested models yielded little difference in the power of T to identify sweeps (Figure S28). The five models we examined, consisting of (A) $f_i = 1/m$, (B) $f_i = (1/i)/\sum_{j=1}^{m} 1/j$, (C) $f_i = (1/i^2)/\sum_{j=1}^{m} 1/j^2$, (D, our primary model for analyses) $f_i = e^{-i}/\sum_{j=1}^{m} e^{-j}$, and (E) $f_i = e^{-i^2}/\sum_{j=1}^{m} e^{-j^2}$, differ in the amount of weight allocated to the secondary sweeping haplotypes $q_i^{(m)}$ for $i \in \{2, 3, \ldots, m\}$ relative to $q_1^{(m)}$ when distorting $\mathbf{p}$. In model A, each sweeping haplotype gains the same amount of weight after distortion, ensuring that each is prominent within spectrum $\mathbf{q}^{(m)}$. Models B through E represent increasingly uneven weight distributions that favor frequency $q_1^{(m)}$ at the expense of $q_2^{(m)}, q_3^{(m)}, \ldots,$ and $q_m^{(m)}$. We believe this is reasonable based on the observation in simulated data that soft sweeps do not affect each sweeping haplotype evenly, and one sweeping haplotype may still be considerably more prominent than the rest (Figure 1; see also, Figure 3 of Garud et al. [2015]). Furthermore, the different T statistic variants demonstrated little difference in sweep classification with $\widehat{m}$ (Figure S29), suggesting that the most important consideration in constructing our sweep models lies in simply distinguishing sweeping from non-sweeping haplotype classes, and not the manner in which they sweep.

An important feature of the T statistic is its ability to detect sweeps from unphased MLG data. Our ability to extend the power of our approach to MLGs is meaningful because it provides the ability to

interrogate polymorphism data from non-model organisms for which phased haplotypes are unavailable, difficult to obtain, or unreliable [Browning and Browning, 2011, O'Connell et al., 2014, Castel et al., 2016, Laver et al., 2016, Zhang et al., 2017, Harris et al., 2018a]. Overall, we found no difference in power trends between the two data types, such that scenarios under which we have high power with phased data are scenarios of high power with unphased data (compare Figures 2 and 3 to Figure S15). However, we find that the T statistic applied to haplotypes always matched or exceeded power for MLG data. This is to be expected because MLGs are a more diverse data type drawn from a smaller sample size. Under a random mating assumption, the presence of a single high-frequency haplotype implies that only one MLG should exist at high frequency, but in the case of two high frequency haplotypes, both homozygotes, as well as their heterozygote, will be prominent in the population. In this way, a sweep on two haplotypes can appear as a sweep on three MLGs, and sweeps on larger numbers of haplotypes will result in even larger numbers of elevated MLGs, which may be more difficult to separate from neutrality proportional to their $\widehat{m}$. Likewise, one high frequency haplotype and one medium frequency haplotype can yield two high frequency MLGs, meaning that an inferred $\widehat{m} = 2$ in MLG data could underlie a true hard sweep. Our results indicate that this may be a common occurrence (compare Figures 4 and S16), and so we recommend scrutinizing $\widehat{m}$ results obtained from MLG data more carefully.

Our extensive testing revealed that T is overall robust to the most common confounding scenarios that affect sweep statistics. Making use of the sample average background haplotype frequency spectrum for inference allows T to account for the effects of mutation rate, recombination rate, and sample size on inferences by creating an expectation specific to the study data (Figures S21 and S24). Using haplotype information provides complete resistance to the effect of background selection on nucleotide diversity (Figure S22), as background selection does not cause haplotypes to rise to high frequency [Enard et al., 2014]. The choice of a SNP-delimited window, meanwhile, is ideal for analyzing datasets with missing sites (Figure S23) or low polymorphism density because by fixing the number of SNPs included in a window, we avoid generating windows that have low diversity simply because they contain few SNPs. SNP-delimited windows may also be more robust to the effect of population bottlenecks on inferences [Harris et al., 2018a]. Despite these strengths, we found that T may be misled by certain admixture scenarios (Figures S17-S20). Although the admixture we simulated was extreme, our results are still informative as to the limits of our model. We found that, as expected, admixture from a donor of small size could inflate the neutral T statistic distribution because small-sized donors have a smaller haplotypic diversity, but we also found that low-level admixture from a large-sized donor also had this effect. We attribute this to the lack of admixture parameter in our model, and expect that models directly incorporating admixture could overcome the confounding effects that we have observed.

The results from our empirical analyses with T served as a validation of our method, yielding expected sweep candidate genes in agreement with previous investigations (Tables S1-S3). More specifically, our top candidates matched extensively with those inferred with H12 and G123 [Garud et al., 2015, Harris et al., 2018a]. Our top candidates in the European-descent CEU population were centered on the *LCT* locus, associated with adaptation to dairy consumption [Bersaglieri et al., 2004]. In the sub-Saharan African YRI population, we saw commonly-recurring candidates at *SYT1*, *NNT*, *LONP2*, and *HEMGN* [Voight et al., 2006, Pickrell et al., 2009, Fagny et al., 2014, Pierron et al., 2014]. The largest discrepancy between the H12 and T statistic analyses of the 1000 Genomes Project data we observed was the absence of *SLC12A1* from the top 40 candidates in CEU haplotype data. *SLC12A1* is a proxy for *SLC24A5* (which was filtered out), a solute transporter has been implicated in the shift toward lighter skin pigmentation among Indo-Europeans [Lamason et al., 2005], and still yields an outlying value of $T = 163.35$, but there are dozens of genes with larger T, whereas only ten genes produced a larger H12 [Harris et al., 2018a]. Even so, *SLC12A1* easily passes our 5% FDR threshold, and even our more stringent 1% threshold (Table S4; all top YRI candidates also pass a 1% threshold). Though our simulation experiments show that T has greater power than H12 for the same data, each method prioritizes sweep signatures differently. H12 is most sensitive to the sum of the two most frequent haplotypes in the frequency spectrum, while T places high emphasis on the relative values of all haplotypes in a truncated spectrum. Similarly, *SweepFinder2* is unlikely to find any soft sweeps, but will find older hard sweeps than what H12 and T could find, particularly for the YRI population (Figure 2).

All of our top candidates for the DGRP scan overlapped with one of the top 10 H12 peaks that Garud et al. [2015] identified, except for *Uhg1*, *CG8552*, *Skeletor/CG14681* (which were in lower-ranked peaks), and *corn*. However, none of our top candidates was statistically significant when compared against the neutral distribution we generated under the Duchen et al. [2013] model of North American *D. melanogaster* population history, and only our top candidate, *CG11902* (located within the second-largest signal peak of Garud et al. [2015]), passed the 5% FDR threshold. This result matches that of Harris et al. [2018b], who also found that they could not reject neutrality when using a model that incorporates uncertainty in key parameters (Figure S27). In contrast, the original interpretation of Garud et al. [2015] found the top signal peaks to be significant, but used fixed model parameters rather than drawing them from a posterior distribution. Even so, the top *D. melanogaster* candidates uncovered in recent analyses show literature support for selection, especially in response to pressure from pesticide application [Menozzi et al., 2004, Pedra et al., 2004, Karasov et al., 2010]. Furthermore, future analyses with more certain demographic parameter estimates may aid in ultimately rejecting neutrality for these genes.

Finally, empirical analysis allowed us to understand the practical effect of using different K truncations to generate inferences. The most apparent difference between scans with $K = 10$ and $K = 20$ truncations

was each configuration's inference on $\widehat{m}$. Using a larger K truncation had the effect of classifying a greater number of candidate sweeps as soft ($\widehat{m} \geq 2$), and we observed this for both phased and unphased data, and all study populations. Without exception, $\widehat{m}$ for $K = 10$ truncations was less than or identical to $\widehat{m}$ for $K = 20$. We find that when $\widehat{m}$ changes, it is for borderline cases, such as that of *COL5A2* in CEU. For this candidate, one high-frequency haplotype predominates, but there is clearly a second haplotype at an elevated frequency as well. In the $K = 20$ spectrum, the contrast in size between the second haplotype frequency and the others is sufficient to assign an inference of $\widehat{m} = 2$, while this is not the case for the $K = 10$ truncation, where haplotype frequencies three through ten are relatively large enough to the second that a hard sweep serves as a better explanation of the data (Figure 5). The MLG spectrum underlying *COL5A2*, in contrast, reflects $\widehat{m} = 2$ regardless of truncation, and that is because we no longer have a borderline case, and two high-frequency MLGs are evident. Ultimately, T is more sensitive to hard sweeps, and any scan with T is likely to yield a greater number of hard sweeps, especially when choosing a smaller K. This is not to say that soft sweeps are uncommon or rarer than hard sweeps [Hernandez et al., 2011, Jensen, 2014, Schrider and Kern, 2017], but that we may simply be overlooking these more often due to the nature of our approach.

We believe that our T statistic will serve as an important contribution to the field of selective sweep detection methods, providing the first maximum-likelihood approach that exploits a haplotype and MLG frequency spectrum distortion model. As such, the T statistic offers high power for recent selective sweeps with little computation time, and can additionally assign an $\widehat{m}$ value to candidates with no additional analysis required. This makes it an appropriate complement to methods such as the singleton density score [Field et al., 2016], which detects sweeps occurring within the past 100 generations of human history (outside the range of detection of T). The T statistic also complements machine learning methods [Lin et al., 2011, Sheehan and Song, 2016, Schrider and Kern, 2016, Sugden et al., 2018, Kern and Schrider, 2018, Mughal and DeGiorgio, 2019, Mughal et al., 2019], which are more powerful in exchange for more computation time (and can also incorporate T into their algorithms). Our lack of dependence on phased data provides the opportunity to search for sweep signatures in any non-model organism for which whole-genome polymorphism data exist. We expect that our simple yet powerful statistical model of selective sweeps will yield novel insights into the adaptive histories of diverse populations. Even in well-studied species such as humans, there are yet understudied populations for which future analyses of selection signatures will provide important insights about population history that is missing from the current literature. To motivate this point, we highlight that insights into local adaptation within human populations continue to emerge [Hu et al., 2017, Buckley et al., 2017, Fan et al., 2019], more than a decade after the first investigations began [Ronald and Akey, 2005, Bustamante et al., 2005, Sabeti et al., 2006]. Finally, we make available the open-source software package `LASSI`, which implements the T statistic protocol in a single efficient pipeline.

Materials and methods

General simulation protocol

We applied the T statistic to simulated data based on demographic models consistent with recent estimates of human [Terhorst et al., 2017] and *D. melanogaster* [Duchen et al., 2013] population history. For some experiments evaluating power under human models, we also applied the T statistic to unphased multilocus genotype (MLG) data, which we produced by manually merging each simulated individual's two haplotypes. We generated these data using the population-genetic simulation software SLiM [Haller and Messer, 2017], as well as with the coalescent simulator *ms* [Hudson, 2002]. For power experiments based on human models, we exclusively performed simulations forward in time using a Wright-Fisher model implemented in SLiM [Fisher, 1930, Wright, 1931, Hartl and Clark, 2007]. These simulations lasted for a total of 200,000 generations, of which the former 100,000 (equivalent to $10N$, where $N = 10^4$ is the diploid effective population size of the simulated population) was a burn-in period to achieve equilibrium values of neutral variation, and the latter 100,000 was the period over which population size variation occurred. To speed up run time, human-modeled simulations were scaled by a factor of $\lambda = 20$, while *D. melanogaster* simulations, which featured larger population sizes, were scaled by $\lambda = 100$. In order to scale, we multiplied mutation rates, recombination rates, and selection coefficients by λ, while the size of the simulated population and the duration of the simulation in generations were divided by λ. Thus, simulation duration was reduced by a factor of 400 ($\lambda = 20$) or 10^4 ($\lambda = 100$).

For all simulations other than for the above human model power experiments, we generated data for each replicate population using *ms*, either for use as input into SLiM (burn-in), or to simulate neutrality. We used the former approach to generate false discovery rate thresholds under the human and *D. melanogaster* models (see *Results* and *Selection protocols* below). Here, we used *ms* to run the majority of each simulation. For the *D. melanogaster* model, we ran *ms* up to the earliest point in time that selection could occur, which was the time of admixture between the African and European populations (Figure S27; see also below). For human models, we simulated the first $10N$ generations prior to sampling. Simulations then proceeded forward in time with SLiM, and selection was allowed to take place. Meanwhile, we outputted neutral simulations to compute p-values (see below) for both human and *D. melanogaster models* directly from *ms*. For human simulations, we chose a mutation rate of 1.25×10^{-8} per site per generation [Narasimhan et al., 2017], and an exponentially-distributed recombination rate with mean 10^{-8} per site per generation, with maximum value truncated at 3×10^{-8}, as in Schrider and Kern [2017] and Mughal and DeGiorgio [2019]. For *D. melanogaster*, our recombination rate was a uniform 5×10^{-9} per site per generation (equivalent to

5×10^{-7} cM per base), and our mutation rate was 10^{-9} per site per generation, specifically chosen to match Garud et al. [2015] and Harris et al. [2018b], but smaller than the value inferred by [Keightley et al., 2009].

Our simulated human demographic histories consisted of a European-descended CEU model (individuals of northern and western European descent sampled in Utah, USA) and a sub-Saharan African YRI model (Yoruba individuals from Ibadan, Nigeria). Both models were inferred by [Terhorst et al., 2017] using `smc++`. The CEU model features a severe bottleneck reducing population effective size by an order of magnitude approximately 2000 generations before sampling, followed by a population expansion over two orders of magnitude leading to present day. The YRI model contains population size fluctuations, but with no severe bottlenecks, and also includes an expansion similarly to the CEU model (see Figure 5 of Terhorst et al. [2017]). Thus, the simulated CEU population has an approximately twofold reduction in its level of background genetic diversity relative to the YRI. For every replicate within each simulated human-model selection scenario (see below), we outputted a simulated chromosome in SLiM of length one megabase (Mb) and scanned it with a sliding analysis window of size 117 SNPs, advancing by 12 SNPs per iteration. A window of 117 SNPs roughly corresponds to the number of SNPs expected in a physical window of size 40 kb for our sample size of 100 European diploid individuals, or 20 kb for 100 sub-Saharan African diploid individuals [Watterson, 1975]. We selected this window size because it is over this interval that pairwise linkage disequilibrium (LD) between SNPs decays by more than one-third on average in the human genome [Jakobsson et al., 2008]. This makes it unlikely that elevated values of the T statistic are due to background LD.

We simulated the *Drosophila* Genetic Reference Panel [DGRP; Mackay et al., 2012] *D. melanogaster* demographic history following the protocol of Harris et al. [2018b], adapting the model of Duchen et al. [2013] (Figure S27). Here, an ancestral African population (effective size N_1) experiences a bottleneck at time T_{B}, contracting to size N_{B} for 1000 generations before expanding to size $N_1^{'}$. After the bottleneck, the ancestral European population diverges from the ancestral African population at time τ_1, and begins with an effective size N_2. The European population grows exponentially to its modern size, $N_2^{'}$. At time τ_2, the North American population ancestral to the modern DGRP sample is generated with initial size N_3 from the admixture of the European and African populations, modeled as a single event, with a proportion α of North American genomes deriving from African ancestors, and a proportion $1 - \alpha$ deriving from European ancestors, where $\alpha < 1/2$. The North American population grows exponentially to its final size, $N_3^{'}$. We draw each of the aforementioned model parameters according to their posterior probability density (Harris et al. [2018b], Table S1), thereby incorporating uncertainty into the demographic model. We only output data from the North American population, and selection only occurs in that population. We used analysis windows of size 400 SNPs and a step size of 40 SNPs for *D. melanogaster* simulations. This represents the

expected number of SNPs in a 10 kb window, over which a pairwise LD decay of greater than one-third occurs for the DGRP dataset [Garud et al., 2015].

Across all experiments, we primarily used a haplotype frequency spectrum truncation of $K = 10$, but for human model analyses, we also examined $K \in \{15, 20, 25\}$. As described in the *Definition of statistic* section, we generate an averaged truncated haplotype frequency spectrum from each neutral replicate analysis window, and use this as an estimate of the baseline variation in the absence of a selective sweep. Except where otherwise mentioned, we used weight allocation scheme (D) from the *Definition of statistic* section, with $f_i = e^{-i} / \sum_{j=1}^{m} e^{-j}$, favoring a greater increase in the frequency of the first sweeping haplotype class relative to the neutral baseline. In practice, all schemes showed similar power, however (Figure S28). Additionally, we only optimized models with $U = p_K$, and did so over a grid of $\varepsilon \in [1/(100K), U]$ (with intervals of $(U - \varepsilon)/100$), providing us with a range of different U and ε combinations. Fixing $U = p_K$ ensured that we would under all circumstances have frequencies $\{q_{m+1}^{(m)}, q_{m+2}^{(m)}, \ldots, q_K^{(m)}\}$ smaller than their equivalents in the $\mathbf{p}$ spectrum, though in doing so we often underestimate the empirically observed putatively non-sweeping classes. We emphasize, however, that due to the structure of our likelihood equations, the value of the T statistic depends more on the fit of the model's sweeping classes than its non-sweeping classes, and thus did not include an optimization of U because it would be unlikely to provide additional power to T, but could potentially require considerably more computational burden.

Selection protocols

To assess the power of the T statistic to differentiate between neutrality and sweeps, we simulated a variety of human selective sweep scenarios, defined primarily by their combination of selection time t, selection strength s, and simulated number of initially sweeping haplotypes ν. We simulated selection on *de novo* mutations arising at times $t \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000\}$ generations prior to sampling. We chose this range of t because the T statistic is suited to detecting recent sweeps that have established in a population, and so this range spans sweeps that are too recent to detect (as they are unlikely to have established), too ancient to detect (as their footprints have likely eroded), and optimal to detect. We simulated sweeps on $\nu \in \{1, 2, 4, 8, 16, 32\}$ distinct sweeping haplotypes, drawn uniformly at random across the set of haplotypes in the population at time t, and conditioned that at least one copy of the selected allele would remain in the population for the duration of the simulation. We chose selection coefficients uniformly at random following five schemes to illustrate the effect of the selection coefficient on sweep detection. Weak selection protocols covered $s \in [0.001, 0.01]$ and $s \in [0.005, 0.05]$. Strong selection comprised $s \in [0.01, 0.1]$ and $s \in [0.05, 0.5]$. We drew mixed selection coefficients as a combination of the weak and strong ranges, with $s \in [0.001, 0.5]$ drawn uniformly at random specifically from a log-scale. We computed the T statistic

for each simulated genomic analysis window using its truncated counts spectrum as input for the likelihood functions ℓ_0 (Equation 4) and ℓ_1 (Equation 5), and retained the largest T for a replicate as its score. We also computed the score in this manner for each existing neutral replicate. We assessed the power of our approach for each parameter set at 1% and 5% false positive rates as the proportion of sweep replicates whose scores exceeded the top 1% or 5% of scores under neutrality, respectively.

For experiments to identify false discovery rate thresholds, selection parameters were drawn at random from distributions of all parameters described above. Sweeps under the human model were initiated uniformly at random across $t \in [200, 4000]$ generations prior to sampling, and selection coefficients uniformly at random from $s \in [0.005, 0.5]$ on a log scale. We raised the lower bound on s relative to power experiments because we found that the T statistic has little ability to identify sweeps on $s < 0.005$. Likewise, we drew $\nu \in \{1, 2, \ldots, 8\}$ to account for the range over which T has the most power. For the *D. melanogaster* model, we simulated hard and soft sweeps once again from $\nu \in \{1, 2, \ldots, 8\}$, but chose $t \in [500, 7000]$ and $s \in [0.001, 0.5]$. The constraint on selection time here derives from the requirement that sweeps must occur in the North American *D. melanogaster* population, which we have fixed to arise 7943 unscaled generations before sampling (this is the sole parameter that we do not draw from the posterior distribution of Duchen et al. [2013]), while we expect that the larger size of the *D. melanogaster* population should allow us to detect sweeps from smaller s due to the larger population-scaled selection coefficient $\sigma = 4N_e s$.

Because we were interested in tracking the number of haplotypes carrying the selected allele in a soft sweep over the course of a sweep, we implemented an *in-silico* barcoding protocol in SLiM. This approach allowed us to observe the effect of hardening on soft sweeps due to genetic drift [Wilson et al., 2014], as well as the relationship among selection coefficient, $\widehat{m}$, and method power. We augmented the simulation procedure of previous power experiments by additionally introducing a unique neutral mutation adjacent to the site of selection for each selected haplotype at time t. Thus, no two selected haplotypes could be identical at time t. Then, at the end of each generation, we measured the frequency of both the selected allele and the count of each unique mutation, which served as a proxy for each unique haplotype's count in the population. We quantified selected allele frequency and selected haplotype count trajectories by taking the mean of 1000 replicates for each scenario. Figure S12 summarizes our *in-silico* barcoding protocol for $\nu = 4$ initially-selected haplotypes. For haplotype tracking experiments, we focused specifically on human scenarios across our five selection strength schemes and $\nu \in \{2, 4, 8, 16, 32\}$, but only studied the selection times for which we had the greatest power. For CEU, these were $t = 2000$ ($s \in [0.001, 0.01]$), $t = 1500$ ($s \in [0.005, 0.05]$ and $s \in [0.001, 0.05]$), $t = 1000$ ($s \in [0.01, 0.1]$), and $t = 500$ ($s \in [0.05, 0.5]$). Across the same s ranges for YRI, we respectively used $t = 2500$, 2000, 2500, 1500, and 500.

Scans of simulated data with multiple methods

In addition to the T statistic, we also applied other popular and powerful methods to our simulated data in order to thoroughly contextualize the power of our approach. Here, we reused the data generated for power experiments across all ranges of s for hard ($\nu = 1$) and soft ($\nu = 4$) sweeps. We elected to compare H12 [Garud et al., 2015], nS_L [Ferrer-Admetlla et al., 2014], *SweepFinder2* [Nielsen et al., 2005, Huber et al., 2016, DeGiorgio et al., 2016], and *Trendsetter* [Mughal and DeGiorgio, 2019], as each presents a unique approach in identifying signatures of selective sweeps. H12 and nS_L represent perhaps the most similar alternatives to T, as they are summary statistics that also leverage measures of haplotype frequency to make inferences. H12 is a haplotype homozygosity-based method that detects sweeps based on their reduced haplotypic diversity, while nS_L identifies sweeps based on their large tracts of sequence identity. Both approaches also have power to detect soft sweeps in addition to hard sweeps (this being the primary purpose of H12). *SweepFinder* is a likelihood method that detects sweeps from distortions in the site frequency spectrum, and has high power to detect hard sweeps, but is underpowered for soft sweeps because it does not use haplotype information. Finally, *Trendsetter* is a sophisticated machine-learning approach that leverages the spatial autocorrelation of summary statistic signals along the chromosome to identify and classify sweeps, and is therefore likely to outperform any single-statistic approach.

We applied these statistics in ways that allowed us to most directly compare their performance with that of T. For H12, we used the same SNP-based window and step sizes as with the T statistic, meaning that we computed H12 using the exact data as for T. Next, we implemented unstandardized nS_L with `selscan` [Szpiech and Hernandez, 2014] running default options, and obtained a value of the statistic for each SNP within a replicate. For *SweepFinder2*, we scanned with a step size of 1000 nucleotides to ensure a dense grid of analysis windows. We generated a helper file from our neutral replicates, which served a similar purpose to our neutral haplotype frequency spectrum $\mathbf{p}$, and only included polymorphic sites in computations [Huber et al., 2016]. Finally, we implemented two versions of *Trendsetter*, each trained on simulated data across three classes—neutrality, hard sweeps, and soft sweeps ($\nu \in \{2, 3, \ldots, 8\}$ drawn uniformly at random)—and 5000 replicates per training class, with $s \in [0.001, 0.5]$ and $t \in [200, 4000]$ as previously. First, we used the standard approach, which studies the spatial signatures of six statistics—mean pairwise sequence difference $\widehat{\pi}$, squared correlation coefficient of linkage disequilibrium r^2, the number of distinct haplotypes N_{haps}, expected haplotype homozygosity H1, H12, and H2/H1 [Garud et al., 2015]. Second, we applied *Trendsetter* to employ the spatial signature of only the T statistic ("T-*Trendsetter*"). This experiment served to highlight the power gain from incorporating spatial autocorrelation of T signals across genomic sequence tracts. In both applications, we used the default behavior of *Trendsetter*, drawing inferences across 201 windows of size 11 SNPs spaced apart by five SNPs for a total of 1010 SNPs. For T-*Trendsetter*, we also trained a classifier on

75 117-SNP windows to better recapitulate our approach with the unaided solitary T statistic. To measure power, we retained the maximum value of each summary statistic (or for *Trendsetter*, the probability of assigning the replicate as a sweep, which is the sum of the hard and soft sweep class probabilities) as the score.

Modeling confounding scenarios

We considered admixture as a confounding scenario because its effect can mimic that of a selective sweep under certain circumstances [Harris et al., 2018a]. To determine the impact of admixture on the T statistic, we implemented contrived models overlaid onto our existing CEU and YRI models in which a distantly-related donor population (diverged $\tau = 2N = 2 \times 10^4$ generations before sampling) unidirectionally admixes into the sampled CEU or YRI population as a single pulse 200 generations before sampling. We varied the size of the admixture pulse from 0.05 to 0.4, meaning between five and 40% of the subsequent generation derive their ancestry from the admixing donor population, at increments of 0.05, and tested admixing population sizes of $N = 10^3$, 10^4, and 10^5. For all scenarios, we generated neutral background haplotype frequency spectra $\mathbf{p}^\alpha$ matching the admixture scenario, but additionally computed the T statistic using the unadmixed $\mathbf{p}$ spectra to demonstrate the effect of not accounting for admixture. For these experiments, all scenarios included no selection in order to highlight the range of T statistics emerging from admixture in the absence of a sweep.

To evaluate background selection as a potential confounding factor that may also produce spurious sweep signals, we performed simulations in which we allowed for deleterious mutations to arise within the simulated chromosome throughout the simulation while maintaining all other parameters identical to neutrality. Our protocol was similar to that of Harris et al. [2018a], and covered the human CEU and YRI models. As with our previous simulations, we generated a genomic region of length 500 kb with identical mutation rate and population sizes as previously, evolving once again for a duration of $20N$ generations ($N = 10^4$ diploids, the effective size during the burn-in period). At the center of the simulated sequence, we introduced a gene of length either 11 kb (small), 55 kb (medium), or 110 kb (large) consisting of a 5' untranslated region (UTR, length 200 bases), either 10 (small), 50 (medium), or 100 (large) exons (100 bases each) and nine (small), 49 (medium), or 99 (large) introns (one kb each) alternating for 10 (small), 54 (medium), or 109 (large) kb, and a 3' UTR (800 bases). We based the sizes of genetic elements on human genome-wide mean values [Mignone et al., 2002, Sakharkar et al., 2004]. Within the gene, strongly deleterious mutations ($s = -0.1$; gamma distribution of fitness effects with shape parameter 0.2) arose at rates of 50% within the UTRs, 75% within the exons, and 10% within the introns, while all other mutations within the gene and across the rest of the chromosome were selectively neutral. To enhance the effect of background selection under this scenario, we

reduced the mean recombination rate from $r = 10^{-8}$ to $r = 10^{-10}$ per site per generation within the central gene.

Within natural genomes, and across different study systems, there can be considerable variability in mutation and recombination rates, which ultimately affects the density of SNPs and number of distinct haplotypes within a genomic analysis window. To understand this effect with respect to the value of the T statistic, we performed neutral simulations under both of our human models in which we reduced the mutation and mean recombination rates by an order of magnitude, both separately and simultaneously. Thus, altered mutation rates were lowered to $\mu = 1.25 \times 10^{-9}$, and recombination rates were drawn as previously, but centered on a mean of $r = 10^{-9}$. We performed 1000 simulations of one-Mb sequences once again, recorded the value of the T statistic to generate a distribution, and measured the proportion of false signals deriving from the above scenarios as a function of the true positive rate.

Similarly, analyses of datasets with missing sites can also reduce the number of SNPs and therefore haplotypes within an analysis window relative to ideal data, and so we performed a similar analysis measuring the distribution of the T statistic under neutrality after removing polymorphic sites with $> 5\%$ missing data. To generate a dataset with missing sites, we followed two types of approaches. First, we selected a number of SNPs, drawn uniformly at random for each existing replicate, and removed data from these SNPs in five to twenty diploid individuals (both haplotypes were treated identically). Second, we performed five iterations of data removal, with SNPs from between one and four individuals removed per iteration. For each approach, we performed three different intensities of data removal. The lowest intensity involved the removal of between 200 and 500 SNPs for the single-iteration approach, or between 40 and 100 SNPs per iteration for the five-iteration approach. The middle intensity approach removed $[400, 1000]$ or $[80, 200]$ SNPs, and the most intense approach removed $[600, 1500]$ or $[120, 300]$. The single-iteration approach is more likely to force the removal of sites with missing data, since the possibility is greater that enough missing alleles will be present at an affected site. The five-iteration approach results in a scattering of missing alleles, but a lower chance of any SNP having $> 5\%$ missing data. Sites with $\leq 5\%$ missing data were incorporated into haplotypes, encoded as a third character state "N" (in contrast to the binary 0/1 for non-missing sites). Doing so allowed us to conservatively lower the value of the T statistic in such cases by introducing greater haplotypic diversity to samples with missing data.

An important component of analysis with the T statistic is the computation of statistical thresholds, such as p-value and FDR cutoffs for determining candidate significance. In order to compute these thresholds, it is necessary to perform simulations under an appropriate demographic history and compare the T statistics of sweep candidates to the distribution of T statistics under simulated replicates. If the demographic history is improperly inferred, then p-values and FDR cutoffs will also be incorrect, and may result in unwarranted

emphasis on nonadaptive regions, or spurious disregard for true selective events. To demonstrate the effect of misspecifying the demographic history of a study population, we show the impact of generating neutral replicates from CEU and YRI demographic histories wherein population size remains constant and equal to the effective size of `smc++`-derived models, computed as the harmonic mean of the population size through the simulation. We also compared distributions of the T statistic generated under the Gravel et al. [2011] CEU and YRI models, which were inferred solely from site frequency spectrum (SFS) information in contrast to the potentially more accurate hybrid approach of `smc++`, which uses the SFS and whole-genome sequence information [Beichman et al., 2017].

Finally, we also examined the power of the T statistic for variably small sample sizes. Haplotype-based methods are sensitive to the number of sampled individuals because sufficient variation needs to be captured in a sample for the difference between neutral and sweep haplotypic diversity to become apparent [Harris et al., 2018a]. That is, reductions in diversity following a sweep become more apparent as more individuals are sampled, and more subtle signatures of sweeps can be elucidated. Accordingly, we resampled all of our existing sweep replicates used in the previous power analysis by drawing $n = 100$, 50, or 20 haplotypes uniformly at random and applying the T statistic as previously. We note that we computed a size-adjusted neutral background haplotype frequency spectrum $\mathbf{p}$ for each reduced-size sample as well.

Application of T statistic to empirical data

We applied the T statistic to human empirical data from the 1000 Genomes Project [Auton et al., 2015], as well as to the DGRP inbred *D. melanogaster* dataset [Mackay et al., 2012]. The former application served primarily as a validation of our approach, as positive selection in the human genome has been widely explored. The latter application represented a typical insect model system that has also been well studied and diverges in size, genome architecture, and population history from humans. The complete outputs for our scans of these datasets are available at `http://degiorgiogroup.fau.edu/LASSI.html`. Our protocols for analyzing either dataset were identical in approach. For each, we searched for candidate peaks by applying a sliding window to all autosomes in the subject genome, basing window size on the interval over which LD, measured as r^2, decayed below one-third of its original value relative to pairs of loci separated by one kb. This matched the prior approaches of [Garud et al., 2015, Harris et al., 2018a]. A candidate peak simply refers to an elevated instance of T within an RNA- or protein-coding region, and we retained the most prominent peak for each gene. For humans, our window was size 117 SNPs, and for *D. melanogaster*, it was 400 SNPs, both matching our values for simulation experiments.

After performing each scan of the CEU and YRI datasets, we filtered windows overlapping chromosomal regions of low alignability and mappability, removing windows overlapping with chromosomal regions of

mean CRG100 score less than 0.9. For *D. melanogaster*, we removed strains 49, 85, 101, 109, 136, 153, 237, 309, 317, 325, 338, 352, 377, 386, 426, 563, and 802 from our analysis due to their high number of heterozygous sites, and treated remaining heterozygous sites as missing data, as in Garud et al. [2015]. We also only used SNPs that had a quality score (reported in the DGRP data) between 1 and 30. We computed T statistic for the $K = 10$ truncation, and assigned $\widehat{m}$ values based on both $K = 10$ and $K = 20$ to examine the practical effect of truncation on candidate classifications.

We intersected the locations of computed T statistic values with the coordinates for protein- and RNA-coding genes based on hg19 and Dmel 5.13 annotations for humans and *D. melanogaster*, respectively. We assigned p-values based on $K = 10$ truncations to the 40 genes with the largest associated values of T by generating 10^6 neutral replicates simulated in *ms* [Hudson, 2002]. For humans, we generated neutral replicates under demographic models inferred with `smc++` [Terhorst et al., 2017], and for *D. melanogaster*, neutral replicates were based on the Duchen et al. [2013] model, drawing parameters as previously. For each replicate, we simulated a sequence of length drawn uniformly at random from the set of all gene lengths, appended with the minimum number of nucleotides necessary to allow the application of full analysis windows centered across the entire length of the simulated gene. As an example, for a simulated human gene of length L nucleotides, we appended additional sequence length guaranteeing that 117-SNP windows centered at the first SNP and the last SNP of the simulated gene could be constructed. This allowed us to obtain a T statistic for at least one whole analysis window centered on the simulated gene during each replicate.

The p-value for a selection candidate is the proportion of T statistics across all neutral replicates (using the maximum value for a replicate if there was more than one analysis window) that exceeded the maximum T associated with the candidate. All p-values were Bonferroni corrected for multiple testing [Neyman and Pearson, 1928], where a significant p-value was $p < 0.05/G$ and where G is the number of genes for which we assigned a score in the organism. Accordingly, we have for humans $G_{\text{human,CEU}} = 18,785$, $G_{\text{human,YRI}} = 19,379$, $p_{\text{human,CEU}} = 2.6617 \times 10^{-6}$, and $p_{\text{human,YRI}} = 2.5801 \times 10^{-6}$, whereas $G_{\text{Dm}} = 10,000$ and $p_{\text{Dm}} = 5 \times 10^{-6}$ for *D. melanogaster*. We ultimately examined fewer than the total number of genes for each study system as a consequence of our filtering and sliding window step size. Filtering had the effect of removing SNPs, which can lead to genes losing representation in the final dataset, while the choice of window step size may result in genes being skipped over. The total number of protein- and RNA-coding autosomal genes in hg19 is 23,735, of which approximately 20% were uncounted, while the total number of autosomal genes in Dmel 5.13 is 12,215, meaning that once again approximately 20% were omitted. Finally, we computed FDR cutoffs for each population by generating simulated 10^6 neutral and sweep replicates as described in *Selection protocols*, generating a sample of size 2×10^6. The 5% FDR cutoff, which we assigned to all populations, was the T statistic value for which 5% of the replicates exceeding that value were neutral,

and 95% were sweeps. There was no value of T which served as a 1% cutoff for the *D. melanogaster* model (that is, there was no T value for which only 1% of replicates were neutral), but we did assign a 1% FDR cutoff to human populations (Table S4).

Acknowledgments

We thank two anonymous reviewers for their insightful comments that helped improve this manuscript. We also thank Mehreen Mughal for her help with *Trendsetter*. This work was funded by National Institutes of Health grant R35-GM128590, by National Science Foundation grants DEB-1753489, DEB-1949268, and BCS-2001063, and by the Alfred P. Sloan Foundation. Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences Advanced CyberInfrastructure (ICDS-ACI).

References

A Albrechtsen, I Moltke, and R Nielsen. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *GENETICS*, 186:295–308, 2010.

A Auton, G R Abecasis, and The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

Q Ayub, B Yngvadottir, Y Chen, Y Xue, M Hu, S C Vernes, S E Fisher, and C Tyler-Smith. FOXP2 Targets Show Evidence of Positive Selection in European Populations. *Am. J. Hum. Genet.*, 92:696–706, 2013.

A Barría, K A Christensen, G Yoshida, A Jedlicki, J S Leong, E B Rondeau, J P Lhorente, B F Koop, W S Davidson, and J M Yáñez. Whole Genome Linkage Disequilibrium and Effective Population Size in a Coho Salmon (*Oncorhynchus kisutch*) Breeding Population Using a High-Density SNP Array. *Front. Genet.*, 10, 2019. doi: 10.3389/fgene.2019.00498.

M A Beaumont, R Nielsen, C Robert, J Hey, O Gaggiotti, L Knowles, A Estoup, M Panchal, J Corander, M Hickerson, S A Sisson, N Fagundes, L Chikhi, P Beerli, R Vitalis, J Cornuet, J Huelsenbeck, M Foll, Z Yang, F Rousset, D Balding, and L Excoffier. In defence of model-based inference in phylogeography. *Mol. Ecol.*, 19:436–446, 2010.

A C Beichman, T N Phung, and K E Lohmueller. Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. *G3—Genes Genom. Genet.*, 7:3605–3620, 2017.

C M Bergey, M Lopez, G F Harrison, E Patin, J A Cohen, L Quintana-Murci, L B Barreiro, and G H Perry. Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. *Proc. Natl. Acad. Sci. U.S.A.*, 115:E11256E11263, 2018.

T Bersaglieri, P C Sabeti, N Patterson, T Vanderploeg, S F Schaffner, J A Drake, M Rhodes, D E Reich, and J N Hirschhorn. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am. J. Hum. Genet.*, 74:1111–1120, 2004.

S R Browning and B L Browning. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12:703–714, 2011.

S R Browning, B L Browning, M L Daviglus, R A Durazo-Arvizu, N Schneiderman, R C Kaplan, and C C Laurie. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.*, 14:e1007385, 2018.

M T Buckley, F Racimo, M E Allentoft, M K Jensen, A Jonsson, H Huang, F Hormozdiari, M Sikora, D Marnetto, E Eskin, et al. Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol. Biol. Evol.*, 34:1307–1318, 2017.

C D Bustamante, A Fledel-Alon, S Williamson, R Nielsen, M T Hubisz, S Glanowski, D M Tanenbaum, T J White, J J Sninsky, R D Hernandez, D Civello, M D Adams, M Cargill, and A G Clark. Natural selection on protein-coding genes in the human genome. *Nature*, 437:1153–1157, 2005.

M C Campbell and S A Tishkoff. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genom. Hum. G.*, 9:403–433, 2008.

S E Castel, P Mohammadi, W K Chung, Y Shen, and T Lappalainen. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.*, 7:12817, 2016.

B Charlesworth, M T Morgan, and D Charlesworth. The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics*, 134:1289–1303, 1993.

B Charlesworth, D Charlesworth, and M T Morgan. The Pattern of Neutral Molecular Variation Under the Background Selection Model. *Genetics*, 141:1619–1632, 1995.

H Chen, N J Patterson, and D E Reich. Population differentiation as a test for selective sweeps. *Genome Res.*, 20:393402, 2010.

C Connan, M Voillequin, C V Chavez, C Mazuet, C Levesque, S Vitry, A Vandewalle, and M R Popoff. Botulinum neurotoxin type B uses a distinct entry pathwaymediated by CDC42 into intestinal cells versus neuronal cells. *Cell. Microbiol.*, 19:e12738, 2017.

A D Cutter and B A Payseur. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.*, 14:262–274, 2013.

M DeGiorgio, K E Lohmueller, and R Nielsen. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genet.*, 10:e1004561, 2014.

M DeGiorgio, C D Huber, M J Hubisz, I Hellmann, and R Nielsen. SWEEPFINDER2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32:1895–1897, 2016.

S Dorus, Z N Freeman, E R Parker, B D Heath, and T L Karr. Recent Origins of Sperm Genes in *Drosophila*. *Mol. Biol. Evol.*, 25:2157–2166, 2008.

P Duchen, S Živković, Hutter, W Stephan, and S Laurent. Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population. *Genetics*, 193:291301, 2013.

D Enard, P W Messer, and D A Petrov. Genome-wide signals of positive selection in human evolution. *Genome Res.*, 24:885–895, 2014.

M Fagny, E Patin, D Enard, L B Barreiro, L Quintana-Murci, and G Laval. Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. *Mol. Biol. Evol.*, 31: 1850–1868, 2014.

S Fan, D E Kelly, M H Beltrame, M E B Hansen, S Mallick, A Ranciaro, J Hirbo, S Thompson, W Beggs, T Nyambo, et al. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.*, 20:82, 2019.

A Ferrer-Admetlla, M Liang, T Korneliussen, and R Nielsen. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Mol. Biol. Evol.*, 31:1275–1291, 2014.

Y Field, E A Boyle, N Telis, Z Gao, K J Gaulton, D Golan, L Yengo, G Rocheleau, P Froguel, M I McCarthy, and J K Pritchard. Detection of human adaptation during the past 2000 years. *Science*, 354:760–764, 2016.

R A Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, Inc., Clarendon, Oxford, 1st edition, 1930.

N R Garud, P W Messer, E O Buzbas, and D A Petrov. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLoS Genet.*, 11:e1005004, 2015.

J H Gillespie. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, MD, 2nd edition, 2004.

T Goeury, L E Creary, L Brunet, M Galan, M Pasquier, B Kervaire, A Langaney, J-M Tiercy, M A Fernández-Viña, J M Nunes, and A Sanchez-Mazas. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a welldocumented population from subSaharan Africa. *HLA*, 91:36–51, 2017.

J M Granka, B M Henn, C R Gignoux, J M Kidd, C D Bustamante, and M W Feldman. Limited Evidence for Classic Selective Sweeps in African Populations. *Genetics*, 192:1049–1064, 2012.

S Gravel, B M Henn, R N Gutenkunst, A R Indap, G T Marth, A G Clark, F Yu, R A Gibbs, The 1000 Genomes Project, and C D Bustamante. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.*, 108:11983–11988, 2011.

I Gronau, M J Hubisz, B Gulko, C G Danko, and A Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.*, 43:1031–1034, 2011.

S R Grossman, K G Andersen, I Shylakhter, S Tabrizi, S Winnicki, A Yen, D J Park, D Griesemer, E K Karlsson, S H Wong, M Cabili, R A Adegbola, R N K Bamezai, A V S Hill, F O Vannberg, J L Rinn, 1000 Genomes Project, E S Lander, S F Schaffner, and P C Sabeti. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, 152:703–713, 2013.

M Haber, A L Jones, B A Connell, Asan, E Arciero, H Yang, M G Thomas, Y Xue, and C Tyler-Smith. A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of Modern Humans Out of Africa. *Genetics*, 212:1421–1428, 2019.

B C Haller and P W Messer. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol. Biol. Evol.*, 34:230–240, 2017.

A M Harris and M DeGiorgio. Identifying and classifying shared selective sweeps from multilocus data. *bioRxiv*, 2019. doi: 10.1101/446005.

A M Harris, N R Garud, and M DeGiorgio. Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. *Genetics*, 210:1429–1452, 2018a.

R B Harris, A Sackman, and J D Jensen. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLoS Genet.*, 14:e1007859, 2018b.

D L Hartl and A G Clark. *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland MA, 4th edition, 2007.

J Hermisson and P S Pennings. Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*, 169:2335–2352, 2005.

J Hermisson and P S Pennings. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.*, 8:700–716, 2017.

R D Hernandez, J L Kelley, E Elyashiv, S C Melton, A Auton, G McVean, 1000 Genomes Project, G Sella, and M Przeworski. Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science*, 331:920–924, 2011.

H Hu, N Petousi, G Glusman, Y Yu, R Bohlender, T Tashi, J M Downie, J C Roach, A M Cole, F R Lorenzo, et al. Evolutionary history of Tibetans inferred from whole-genome sequencing. *PLoS Genet.*, 13:e1006675, 2017.

C D Huber, M DeGiorgio, I Hellmann, and R Nielsen. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol. Ecol.*, 25:142–156, 2016.

G Hudjashov, T M Karafet, D J Lawson, S Downey, O Savina, H Sudoyo, J S Lansing, M F Hammer, and M P Cox. Complex Patterns of Admixture across the Indonesian Archipelago. *Mol. Biol. Evol.*, 34: 24392452, 2017.

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

M Jakobsson, S W Scholz, P Scheet, J R Gibbs, J M VanLiere, H Fung, Z A Szpiech, J H Degnan, K Wang, R Guerreiro, J M Bras, J C Schymick, D G Hernandez, B J Traynor, J Simon-Sanchez, M Matarin, A Britton, J van de Leemput, I Rafferty, M Bucan, H M Cann, J A Hardy, N A Rosenberg, and A B Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451:998–1003, 2008.

J D Jensen. On the unfounded enthusiasm for soft selective sweeps. *Nat. Commun.*, 5:5281, 2014.

J D Jensen, Y Kim, V B DuMont, C F Aquadro, and C D Bustamante. Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics*, 170:1401–1410, 2005.

J D Jensen, K D Thornton, C D Bustamante, and C F Aquadro. On the Utility of Linkage Disequilibrium as a Statistic for Identifying Targets of Positive Selection in Nonequilibrium Populations. *Genetics*, 176: 23712379, 2007.

B L Jones, T O Raga, A Liebert, P Zmarz, E Bekele, E T Danielson, A K Olsen, N Bradman, J T Troelsen, and D M Swallow. Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. *Am. J. Hum. Genet.*, 93:538–544, 2013.

T Karasov, P W Messer, and D A Petrov. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet.*, 6:e1000924, 2010.

P D Keightley, U Trivedi, M Thomson, F Oliver, S Kumar, and M L Blaxter. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.*, 19: 11951201, 2009.

A D Kern and D R Schrider. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3-Genes Genom. Genet.*, 8:1959–1970, 2018.

Y Kim and R Nielsen. Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, 167:1513–1524, 2004.

Y Kim and W Stephan. Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. *Genetics*, 160:765–777, 2002.

B Kolaczkowski, D N Hupalo, and A D Kern. Recurrent Adaptation in RNA Interference Genes Across the *Drosophila* Phylogeny. *Mol. Biol. Evol.*, 28:1033–1042, 2011.

A Kopatz, H G Eiken, J Schregel, J Aspi, I Kojola, and S B Hagen. Genetic substructure and admixture as important factors in linkage disequilibriumbased estimation of effective number of breeders in recovering wildlife populations. *Ecol. Evol.*, 7:10721–10732, 2017.

N Kouprina, A Pavlicek, G H Mochida, G Solomon, W Gersch, Y Yoon, R Collura, M Ruvolo, J C Barrett, C G Woods, C A Walsh, J Jurka, and V Larionov. Accelerated Evolution of the *ASPM* Gene Controlling Brain Size Begins Prior to Human Brain Expansion. *PLoS Biol.*, 2:e126, 2004.

R L Lamason, M P K Mohideen, J R Mest, A C Wong, H L Norton, M C Aros, M J Jurynec, X Mao, V R Humphreville, J E Humbert, S Sinha, J L Moore, P Jagadeeswaran, W Zhao, G Ning, I Makalowska, P M McKeigue, D O'Donnell, R Kittles, E J Parra, N J Mangini, D J Grunwald, M D Shriver, V A Canfield, and K C Cheng. *SLC24A5*, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science*, 310:1782–1786, 2005.

T W Laver, R C Caswell, K A Moore, J Poschmann, M B Johnson, M M Owens, S Ellard, K H Paszkiewicz, and M N Weedon. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.-U.K.*, 6:21746, 2016.

Y C G Lee and C H Langley. Long-Term and Short-Term Evolutionary Impacts of Transposable Elements on *Drosophila*. *Genetics*, 192:1411–1432, 2012.

S Leivers, G Rhodes, and L W Simmons. Sperm Competition in Humans: Mate Guarding Behavior Negatively Correlates with Ejaculate Quality. *PLoS ONE*, 9:e108099, 2014.

T Lencz, C Lambert, P DeRosse, K E Burdick, T V Morgan, J M Kane, R Kucherlapati, and A K Malhotra. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.*, 104:19942–19947, 2007.

P Librado, C Gamba, C Gaunitz, C D Sarkissian, M Pruvost, A Albrechtsen, A Fages, N Khan, M Schubert, V Jagannathan, et al. Ancient genomic changes associated with domestication of the horse. *Science*, 356: 442–445, 2017.

K Lin, H Li, C Schlötterer, and A Futschik. Distinguishing Positive Selection From Neutral Evolution: Boosting the Performance of Summary Statistics. *Genetics*, 187:229–244, 2011.

K E Lohmueller, C D Bustamante, and A G Clark. Methods for Human Demographic Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data. *Genetics*, 182:217–231, 2009.

S Lukić and J Hey. Demographic Inference Using Spectral Methods on SNP Data, with an Analysis of the Human Out-of-Africa Expansion. *Genetics*, 192:619–639, 2012.

S Lukic, J-C Nicolas, and A J Levine. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death Differ.*, 21: 381–387, 2014.

T F C Mackay, S Richards, E A Stone, A Barbadilla, J F Ayroles, D Zhu, S Casillas, Y Han, M M Magwire, J M Cridland, et al. The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, 482:173–178, 2012.

P Menozzi, M A Shi, A Lougarre, Z H Tang, and D Fournier. Mutations of acetylcholinesterase which confer insecticide resistance in *Drosophila melanogaster* populations. *BMC Evol. Biol.*, 4:4, 2004.

P W Messer and R A Neher. Estimating the Strength of Selective Sweeps from Deep Population Diversity Data. *Genetics*, 191:593–605, 2012.

F Mignone, C Gissi, S Liuni, and G Pesole. Untranslated regions of mRNAs. *Genome Biol.*, 3:reviews0004–1, 2002.

M R Mughal and M DeGiorgio. Localizing and Classifying Adaptive Targets with Trend Filtered Regression. *Mol. Biol. Evol.*, 36:252–270, 2019.

M R Mughal, H Koch, J Huang, F Chiaromonte, and M DeGiorgio. Learning the properties of adaptive regions with functional data analysis. *bioRxiv*, 2019. doi: 10.1101/834010.

S Nakagome, G Alkorta-Aranburu, R Amato, B Howie, Peter B M, R R Hudson, and A Di Rienzo. Estimating the Ages of Selection Signals from Different Epochs in Human History. *Mol. Biol. Evol.*, 33:657–669, 2015.

K Nam, K Munch, T Mailund, A Nater, M P Greminger, M Krützen, T Marquès-Bonet, and M H Schierup. Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proc. Natl. Acad. Sci. U.S.A.*, 114:1613–1618, 2017.

V M Narasimhan, R Rahbari, A Scally, A Wuster, D Mason, Y Xue, J Wright, R C Trembath, E R Maher, D A van Heel, A Auton, M E Hurles, C Tyler-Smith, and R Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.*, 8, 2017. doi: 10.1038/s41467-017-00323-y.

J Neyman and E S Pearson. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A:175–240, 1928.

L E Nicolaisen and M M Desai. Distortions in Genealogies due to Purifying Selection and Recombination. *Genetics*, 195:221–230, 2013.

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

D I Nurminsky, M V Nurminskaya, D De Aguiar, and D L Hartl. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature*, 396:572–575, 1998.

J O'Connell, D Gurdasani, O Delaneau, N Pirastu, S Ulivi, M Cocca, M Traglia, J Huang, J E Huffman, I Rudan, R McQuillan, R M Fraser, H Campbell, O Polasek, G Asiki, K Ekoru, C Hayward, A F Wright, V Vitart, P Navarro, J Zagury, J F Wilson, D Toniolo, P Gasparini, N Soranzo, M S Sandhu, and J Marchini. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.*, 10:e1004234, 2014.

P F O'Reilly, E Birney, and D J Balding. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.*, 18:1304–1313, 2008.

P Pavlidis and N Alachiotis. A survey of methods and tools to detect recent and strong positive selection. *J. Biol. Res.-Thessalon*, 24, 2017. doi: 10.1186/s40709-017-0064-0.

P Pavlidis, D Živković, A Stamatakis, and N Alachiotis. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Mol. Biol. Evol.*, 30:2224–2234, 2013.

J H F Pedra, L M McIntyre, M E Scharf, and B R Pittendrigh. Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, 101:7034–7039, 2004.

P S Pennings and J Hermisson. Soft Sweeps II: Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Mol. Biol. Evol.*, 23:1076–1084, 2006a.

P S Pennings and J Hermisson. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLoS Genet.*, 2:e186, 2006b.

B M Peter, E Huerta-Sánchez, and R Nielsen. Distinguishing between Selective Sweeps from Standing Variation and from a *De Novo* Mutation. *PLoS Genet.*, 8:e1003011, 2012.

J K Pickrell, G Coop, J Novembre, S Kudaravalli, J Z Li, D Absher, B S Srinivasan, G S Barsh, R M Myers, M W Feldman, and J K Pritchard. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.*, 19:826–837, 2009.

D Pierron, H Razafindrazaka, L Pagani, F Ricaut, T Antao, M Capredon, C Sambo, C Radimilahy, J Rakotoarisoa, R M Blench, T Letellier, and T Kivisild. Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc. Natl. Acad. Sci. U.S.A.*, 111: 936–941, 2014.

J P Pollinger, C D Bustamante, A Fledel-Alon, S Schmutz, M M Gray, and R K Wayne. Selective sweep mapping of genes with large phenotypic effects. *Genome Res.*, 15:1809–1819, 2005.

J K Pritchard and A DiRienzo. Adaptation not by sweeps alone. *Nat. Rev. Genet.*, 11:665–667, 2010.

M Przeworski. The Signature of Positive Selection at Randomly Chosen Loci. *Genetics*, 160:1179–1189, 2002.

F Racimo. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics*, 202:733750, 2016.

J Ronald and J M Akey. Genome-wide scans for loci under selection in humans. *Hum. Genomics*, 2:113–125, 2005.

P C Sabeti, D E Reich, J M Higgins, H Z P Levine, D J Richter, S F Schaffner, S B Gabriel, J V Platko, N J Patterson, G J McDonald, H C Ackerman, S J Campbell, D Altshuler, R Cooper, D Kwiatkowski, R Ward, and E S Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, 2002.

P C Sabeti, S F Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, T S Mikkelsen, D Altshuler, and E S Lander. Positive Natural Selection in the Human Lineage. *Science*, 312:1614–1620, 2006.

M K Sakharkar, V T K Chow, and P Kangueane. Distributions of exons and introns in the human genome. *In Silico Biol.*, 4:387–393, 2004.

D R Schrider and A D Kern. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genet.*, 12:e1005928, 2016.

D R Schrider and A D Kern. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Mol. Biol. Evol.*, 34:1863–1877, 2017.

J Schweinsberg and R Durrett. Random Partitions Approximating the Coalescence of Lineages During a Selective Sweep. *Ann. Appl. Probab.*, 15:1591–1651, 2005.

J Seger, W A Smith, J J Perry, J Hunn, Z A Kaliszewska, L La Sala, L Pozzi, V J Rowntree, and F R Adler. Gene Genealogies Strongly Distorted by Weakly Interfering Mutations in Constant Environments. *Genetics*, 184:529–545, 2010.

S Sheehan and Y S Song. Deep Learning for Population Genetic Inference. *PLoS Comput. Biol.*, 12:e1004845, 2016.

N R Stevens, J Dobbelaere, A Wainman, F Gergely, and J W Raff. Ana3 is a conserved protein required for the structural integrity of centrioles and basal bodies. *J. Cell Biol.*, 187:355–363, 2009.

L A Sugden, E G Atkinson, A P Fischer, S Rong, B M Henn, and S Ramachandran. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat. Commun.*, 9:703, 2018.

Z A Szpiech and R D Hernandez. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol. Biol. Evol.*, 31:2824–2827, 2014.

J Terhorst, J A Kamm, and Y S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.*, 49:303–309, 2017.

M E Teves, D R Nagarkatti-Gude, Z Zhang, and J F Strauss III. Mammalian axoneme central pair complex proteins: Broader roles revealed by gene knockout phenotypes. *Cytoskeleton.*, 73:3–22, 2016.

A I Vatsiou, E Bazin, and O E Gaggiotti. Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.*, 25:89–103, 2016.

B F Voight, S Kudaravalli, X Wen, and J K Pritchard. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol.*, 4:e72, 2006.

H M T Vy and Y Kim. A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data. *Genetics*, 200:633–649, 2015.

H M T Vy, Y Won, and Y Kim. Multiple modes of positive selection shaping the patterns of incomplete selective sweeps over African populations of *Drosophila melanogaster*. *Mol. Biol. Evol.*, 34:2792–2807, 2017.

G A Watterson. On the Number of Segregating Sites in Genetical Models without Recombination. *Theor. Popul. Biol.*, 7:256–276, 1975.

T Wiehe, V Nolte, D Zivkovic, and C Schlötterer. Identification of Selective Sweeps Using a Dynamically Adjusted Number of Linked Microsatellites. *Genetics*, 175:207–218, 2007.

B A Wilson, D A Petrov, and P W Messer. Soft Selective Sweeps in Complex Demographic Scenarios. *Genetics*, 198:669–684, 2014.

B A Wilson, P S Pennings, and D A Petrov. Soft Selective Sweeps in Evolutionary Rescue. *Genetics*, 205: 1573–1586, 2017.

A Wong, M C Turchin, M F Wolfner, and C F Aquadro. Evidence for Positive Selection on *Drosophila melanogaster* Seminal Fluid Protease Homologs. *Mol. Biol. Evol.*, 25:497–506, 2008.

S Wright. Evolution in Mendelian Populations. *Genetics*, 16:97–159, 1931.

S Yeh, T Do, C Chan, A Cordova, F Carranza, E A Yamamoto, M Abbassi, K A Gandasetiawan, P Librado, E Damia, P Dimitri, J Rozas, D L Hartl, J Roote, and J M Ranz. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc. Natl. Acad. Sci. U.S.A.*, 109: 2043–2048, 2012.

N S Yudin, D M Larkin, and E V Ignatieva. A compendium and functional characterization of mammalian genes involved in adaptation to Arctic or Antarctic environments. *BMC Genet.*, 18:111, 2017.

F Zhang, L Christiansen, J Thomas, D Pokholok, R Jackson, N Morrell, Y Zhao, M Wiley, E Welch, E Jaeger, A Granat, S J Norberg, A Halpern, M C Rogert, M Ronaghi, J Shendure, N Gormley, K L Gunderson, and F J Steemers. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol.*, 35, 2017.

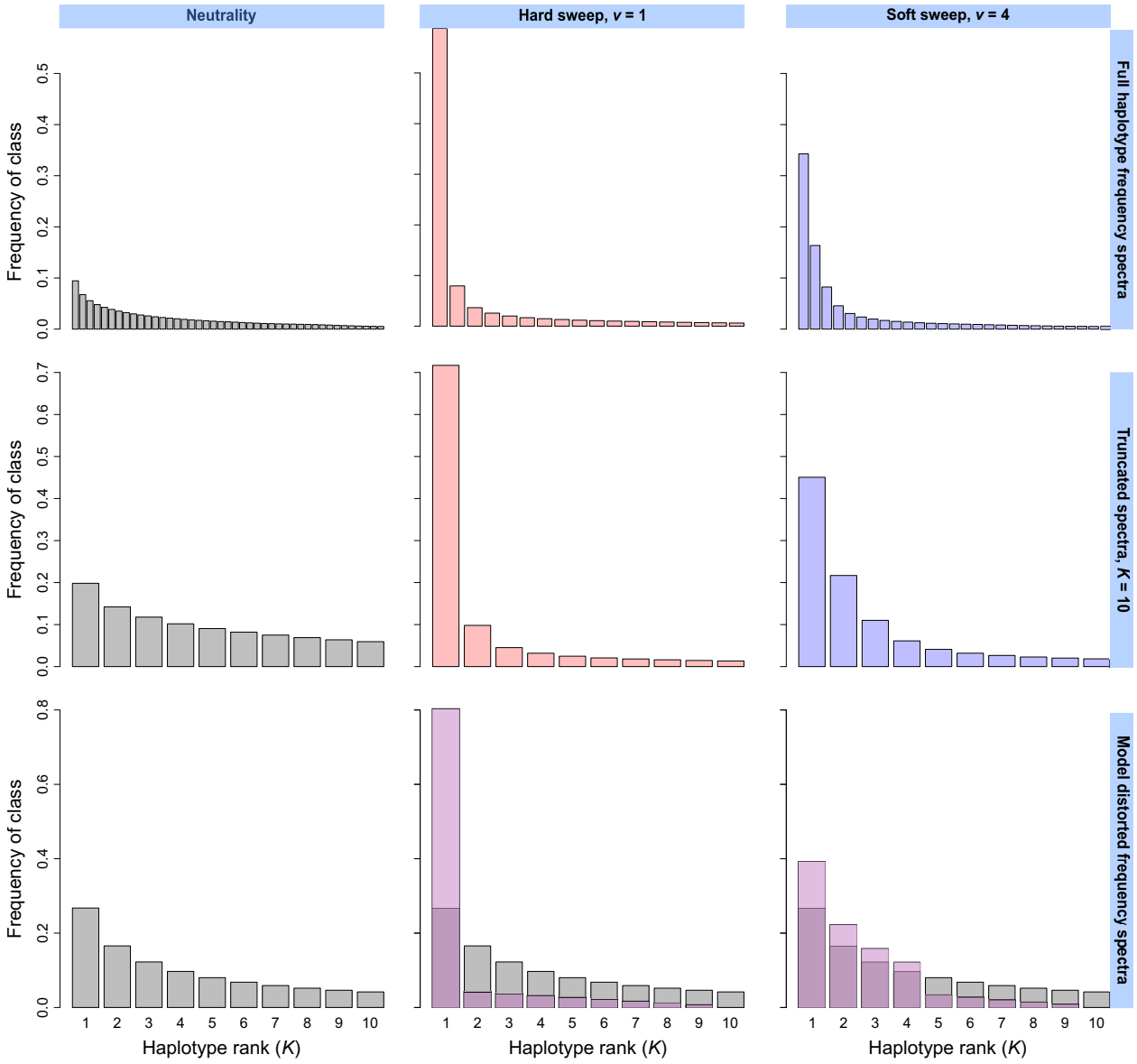

Figure 1: Example simulated haplotype frequency spectra for neutrality, hard sweeps ($\nu = 1$), and soft sweeps ($\nu = 4$). Under neutrality, all sampled haplotypes in an analysis window exist at low frequency, and there are many haplotypes. In contrast, selective sweeps yield high-frequency haplotypes, and fewer total haplotypes (top). Truncated spectra ($K = 10$) preserve their overall shape relative to untruncated spectra above (middle). We distort the truncated neutral spectrum computed from sampled haplotypes to yield spectra corresponding to alternative models (purple), in which the mass of non-sweeping classes is transferred to sweeping classes, resembling the expected pattern under a true selection event (bottom). Spectra represent the mean frequencies of each distinct haplotype across 10^3 simulated replicates in a sample of $n = 100$ diploids under a constant-size simulated demographic history. Selective sweeps were simulated as one or more strongly-selected ($s = 0.1$) haplotypes rising to high frequency starting at the time of selection $t = 400$ generations before sampling.

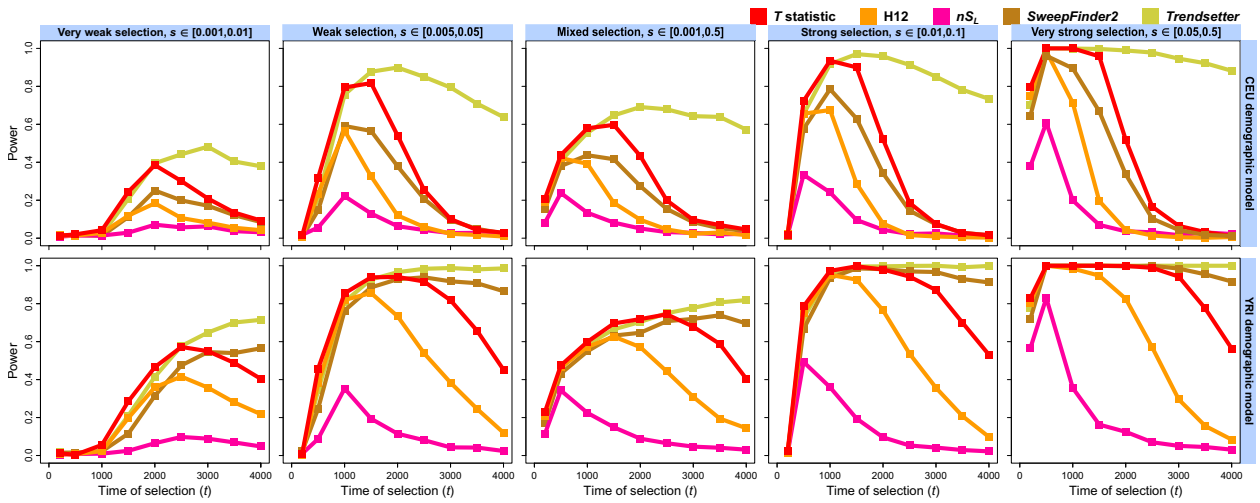

Figure 2: Powers of the T statistic and sweep detection methods—H12, nS_L, *SweepFinder2*, and *Trendsetter*—at the 1% false positive rate (FPR) to detect hard selective sweeps originating from a single beneficial *de novo* mutation arising at times $t \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000\}$ generations prior to sampling, for the European CEU (top) and sub-Saharan African YRI (bottom) human demographic models inferred with `smc++`. Analysis data consisted of phased haplotypes of length one megabase, with 1000 replicates for each distinct scenario. Selective sweeps were simulated for five ranges of selection coefficients (s) spanning very weak to very strong, with s for each replicate drawn uniformly at random (from a log-scale for s drawn across orders of magnitude). All inferences used a spectrum of $K = 10$ for likelihood computations.

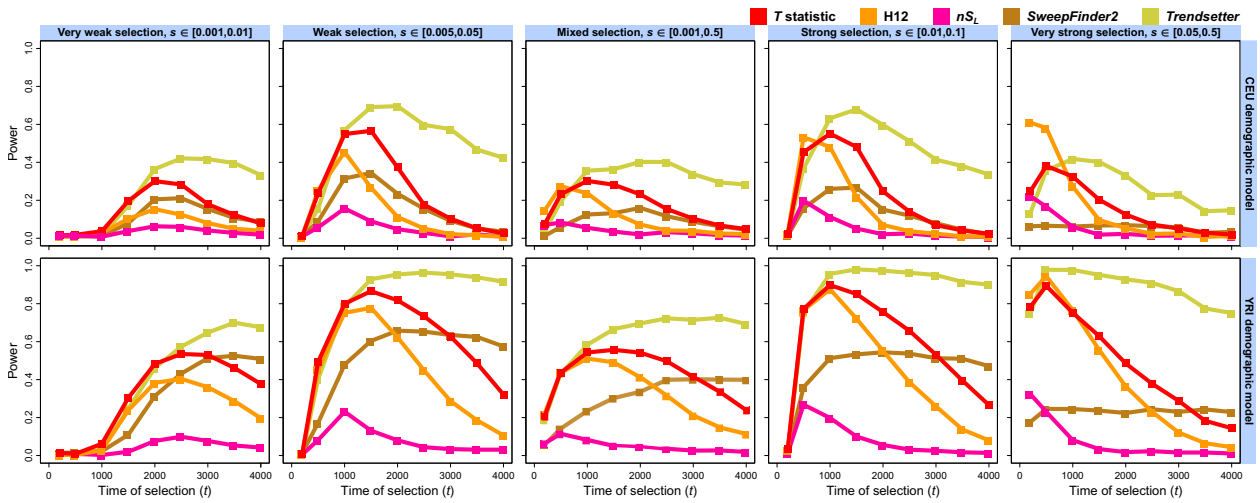

Figure 3: Powers of the T statistic and other sweep detection methods—H12, nS_L, *SweepFinder2*, and *Trendsetter*—at the 1% false positive rate (FPR) to detect soft selective sweeps from selection on standing variation on $\nu = 4$ distinct sweeping haplotypes beginning at times $t \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000\}$ generations prior to sampling, for the European CEU (top) and sub-Saharan African YRI (bottom) human demographic models inferred with `smc++`. Analysis data consisted of phased haplotypes of length one megabase, with 1000 replicates for each distinct scenario. Selective sweeps were simulated for five ranges of selection coefficients (s) spanning very weak to very strong, with s for each replicate drawn uniformly at random (from a log-scale for s drawn across orders of magnitude). All inferences used a spectrum of $K = 10$ for likelihood computations.

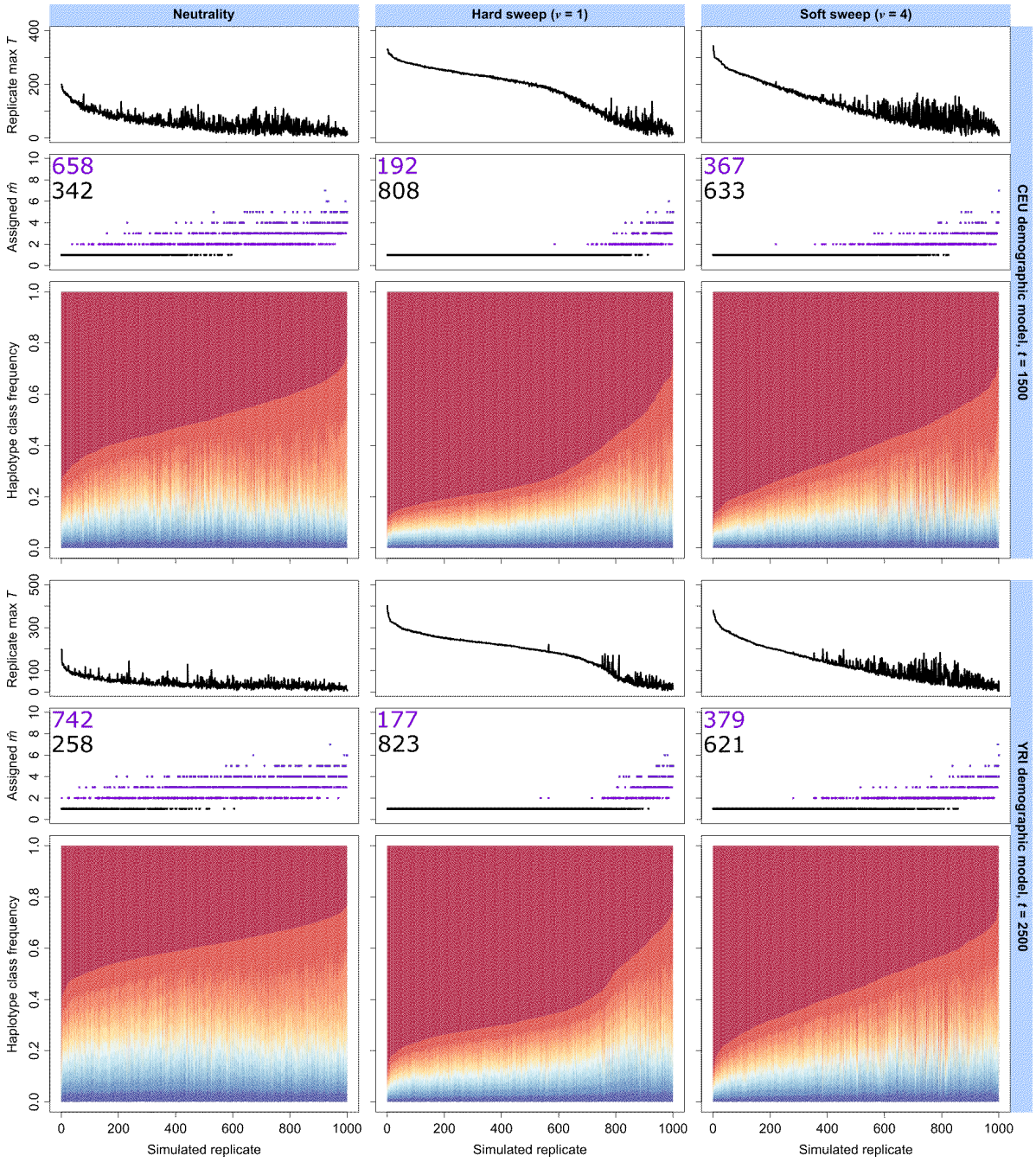

Figure 4: Truncated haplotype frequency spectra ($K = 10$) across 10^3 simulated replicates for analysis window of maximum replicate-wide T statistic under neutral (left), hard sweep (center), and soft sweep (right) scenarios, for European CEU (top) and sub-Saharan African YRI (bottom) human demographic models. Each simulated replicate is one vertical slice within the lower panel, and replicate spectra are arranged in decreasing order of most-frequent haplotype frequency. Replicates are associated with their T statistic (upper panel) and their inferred $\widehat{m}$ (middle panel). Inferred hard sweeps ($\widehat{m} = 1$) are indicated with black dots, whereas inferred soft sweeps ($\widehat{m} \geq 2$) are indicated in purple. We indicate within the $\widehat{m}$ panel the number of replicates classified as hard (black text) or soft (purple text). Sweep replicates were drawn from mixed selection coefficients $s \in [0.001, 0.5]$ uniformly at random on a log-scale, and are identical to those in Figures 2 and 3.

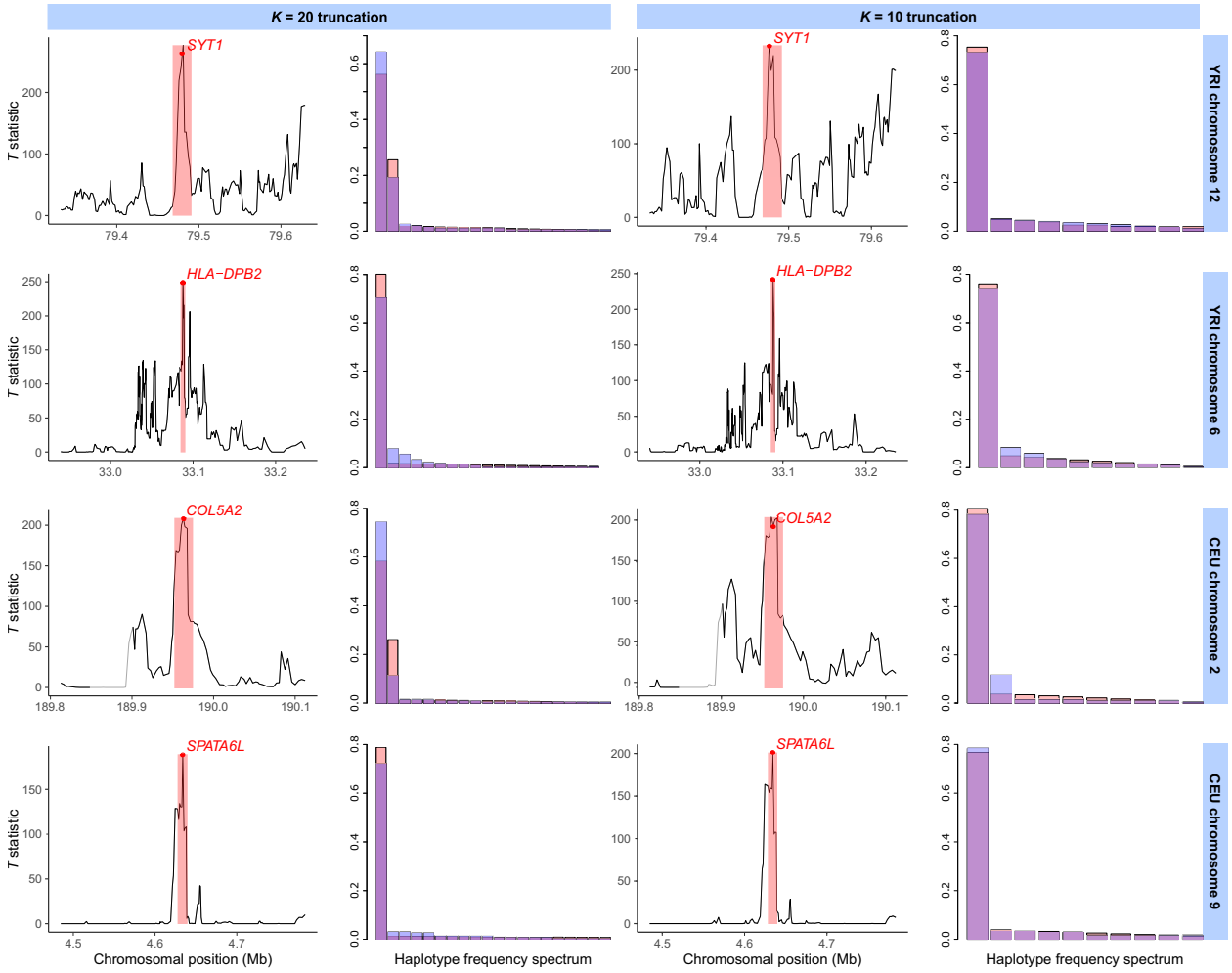

Figure 5: Selective sweep candidates detected with the T statistic from the 1000 Genomes Project dataset [Auton et al., 2015] as for scans with a $K = 20$ (left) and $K = 10$ (right) truncation. For each of four sweep candidates in the human YRI (top two rows) and CEU (bottom two rows) populations, we show the T statistic across the 300 kb interval surrounding the candidate peak, as well as the frequency spectra for the most likely sweep model corresponding to the candidate at the 117-SNP analysis window of maximum T. The window of maximum T is shaded in red, with the position of the window center (median SNP) as a red dot. The frequency spectrum of the most likely model is also shown in red, whereas the observed frequency spectrum at the point of maximum T is overlaid in blue. The displayed candidates are a putative soft sweep ($\widehat{m} = 2$) at *SYT1* in YRI (top row), hard sweep ($\widehat{m} = 1$) at *HLA-DPB2* in YRI (second row), soft sweep ($\widehat{m} = 2$ when analyzed with $K = 20$; hard for $K = 10$) at *COL5A2* in CEU (third row), and hard sweep at *SPATA6L* in CEU (bottom row). We note that the window of maximum signal for $K = 10$ and $K = 20$ differed for *SYT1* (top row). The gray segment upstream of *COL5A2* (third row) indicates a portion of the genome that was filtered out (see *Materials and methods*).