A spatially aware likelihood test to detect sweeps from haplotype distributions

Michael DeGiorgio[1,*], Zachary A. Szpiech[2,3,*]

[1]*Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA*
[2]*Department of Biology, Pennsylvania State University, University Park, PA 16801, USA*
[3]*Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA 16801, USA*
[*]*Corresponding authors:* `mdegiorg@fau.edu` (M.D.), `szpiech@psu.edu` (Z.A.S.)

Keywords: haplotypes, hard sweeps, soft sweeps, maximum likelihood

Abstract

The inference of positive selection in genomes is a problem of great interest in evolutionary genomics. By identifying putative regions of the genome that contain adaptive mutations, we are able to learn about the biology of organisms and their evolutionary history. Here we introduce a composite likelihood method that identifies recently completed or ongoing positive selection by searching for extreme distortions in the spatial distribution of the haplotype frequency spectrum relative to the genome-wide expectation taken as neutrality. Furthermore, the method simultaneously infers two parameters of the sweep: the number of sweeping haplotypes and the "width" of the sweep, which is related to the strength and timing of selection. We demonstrate that this method outperforms the leading haplotype-based selection statistics. Then, as a positive control, we apply it to two well-studied human populations from the 1000 Genomes Project and examine haplotype frequency spectrum patterns at the *LCT* and MHC loci. To facilitate use of this method, we have implemented it in user-friendly open source software.

Introduction

The identification and classification of genomic regions undergoing positive selection in populations has been of long standing interest for studying organisms across the tree of life. By investigating regions containing putative adaptive variation, one can begin to shed light on a population's evolutionary history and the biological changes well-suited to cope with various selection pressures.

The genomic footprint of positive selection is generally characterized by long high-frequency haplotypes and low nucleotide diversity in the vicinity of the adaptive locus, the result of linked genetic material "sweeping" to high frequency faster than mutation and recombination can introduce novel variation. These selective sweeps are often described by two paradigms—"hard sweeps" and "soft sweeps". Whereas a hard sweep is the result of a beneficial mutation that brings a single haplotype to high frequency [Przeworski, 2002], soft sweeps are the result of selection on multiple haplotype backgrounds, often the result of selection on standing variation or a high adaptive mutation rate. Soft sweeps are thus characterized by multiple sweeping haplotypes rising to high frequency [Hermisson and Pennings, 2005, Pennings and Hermisson, 2006a].

Many statistics have been proposed to capture these haplotype patterns to make inferences about recent or ongoing positive selection [Sabeti et al., 2002, Voight et al., 2006, Sabeti et al., 2007, Ferrer-Admetlla et al., 2014, Garud et al., 2015, Harris et al., 2018, Torres et al., 2018, Harris and DeGiorgio, 2020, Szpiech et al., 2020], most of which focus on summarizing patterns of haplotype homozygosity in a local genomic region. A particularly novel approach, the T statistic implemented in `LASSI` [Harris and DeGiorgio, 2020], employs a likelihood model based on distortions of the haplotype frequency spectrum (HFS). In this framework, Harris and DeGiorgio [2020] model a shift in the HFS toward one or several high-frequency haplotypes as the result of a hard or soft sweep in a local region of the genome. In addition to the likelihood test statistic T, for which larger values suggest more support for a sweep, `LASSI` also infers the parameter $\widehat{m}$. This parameter estimates the number of sweeping haplotypes in a genomic region, and $\widehat{m} > 1$ indicates support for a soft sweep.

A drawback of the original formulation of the T statistic implemented in `LASSI` is that it does not account for or make use of the genomic spatial distribution of haplotypic variation expected from a sweep. Specifically, Harris and DeGiorgio [2020] demonstrated that if the spatial distribution of T was directly accounted for in the machine learning approach (*Trendsetter*) of Mughal and DeGiorgio [2019], the power for detecting sweeps was greatly enhanced. Indeed, modern statistical learning machinery to detect sweeps has been greatly enhanced by incorporating spatial distributions of summary statistics [Lin et al., 2011, Schrider and Kern, 2016, Sheehan and Song, 2016, Kern and Schrider, 2018, Mughal and DeGiorgio, 2019, Mughal et al., 2020]. However, these machine learning methods need extensive simulations under an accurate and explicit demographic model to train the classifier. An alternative approach is to directly integrate this spatial distribution into the likelihood model, as has been performed for site frequency spectrum (SFS) composite likelihood methods to detect sweeps [Kim and Stephan, 2002, Nielsen et al., 2005, Chen et al., 2010, Huber et al., 2015, Vy and Kim, 2015, DeGiorgio et al., 2016, Racimo, 2016, Lee and Coop, 2017, Setter et al., 2020]. Here we incorporate the spatial distribution of HFS variation into the `LASSI` framework and introduce the Spatially Aware Likelihood Test for Improving `LASSI`, or `saltiLASSI`. For easy application to genomic datasets, we implement `saltiLASSI` in the open source program `lassip` along with `LASSI` [Harris and DeGiorgio, 2020], and other HFS-based statistics H12, H2/H1, G123, and G2/G1 [Garud et al., 2015, Harris et al., 2018]. `lassip` is available at `https://www.github.com/szpiech/lassip`.

After validating `saltiLASSI` through simulations and comparing it to other popular haplotype-based selection scans, we apply `saltiLASSI` to two well-studied populations from the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2015] as a positive control in an empirical data set. We reproduce several well-known signals of selection in the European CEU population and the African YRI population, including the *LCT* (CEU), MHC (CEU and YRI), and *APOL1* (YRI) loci, demonstrating that this method works well in real data.

Results

In this section we begin by developing a new likelihood ratio test statistic, termed Λ, that evaluates spatial patterns in the distortion of the HFS as evidence for sweeps. We then demonstrate that Λ has substantially higher power than competing single-population haplotype-based approaches, across a number of model parameters related to the underlying demographic and adaptive processes. Similar to the T statistic implemented in the `LASSI` framework of Harris and DeGiorgio [2020], we also show that Λ is capable of approximating the softness of a sweep by estimating the current number of high-frequency haplotypes $\widehat{m}$. We then apply the Λ statistic to whole-genome sequencing data from two human populations from the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2015].

Definition of the statistic

Here we extend the `LASSI` maximum likelihood framework for detecting sweeps based on haplotype data [Harris and DeGiorgio, 2020], by incorporating the spatial pattern of haplotype frequency distortion in a statistical model of a sweep. Recall that Harris and DeGiorgio [2020] defined a genome-wide background K-haplotype truncated frequency spectrum vector

$$\mathbf{p} = (p_1, p_2, \ldots, p_K),$$

which they assume represents the neutral distribution of the K most-frequent haplotypes, with $p_1 \geq p_2 \geq \cdots \geq p_K \geq 0$ and normalization such that $\sum_{k=1}^{K} p_k = 1$. Harris and DeGiorgio [2020] then define the vector

$$\mathbf{q}^{(m)} = \big(q_1^{(m)}, q_2^{(m)}, \ldots, q_K^{(m)}\big),$$

with $q_1^{(m)} \geq q_2^{(m)} \geq \cdots \geq q_K^{(m)}$ and $\sum_{k=1}^{K} q_k^{(m)} = 1$. This represents a distorted K-haplotype truncated frequency spectrum vector in a particular genomic region with a distortion consistent with m sweeping haplotypes. To create the these distorted haplotype spectra, Harris and DeGiorgio [2020] used the equation

$$q_k^{(m)} = \begin{cases} p_k + f_k \sum_{j=m+1}^{K} (p_j - q_j^{(m)}) & k = 1, 2, \ldots, m \\ U - \frac{k-m-1}{K-m-1}(U - \varepsilon) & k = m+1, m+2, \ldots, K \end{cases}$$

where $f_k \geq 0$ for $k \in \{1, 2, \ldots, m\}$ and $\sum_{k=1}^{m} f_k = 1$, defines the way at which mass is distributed to the m "sweeping" haplotypes from the $K - m$ non-sweeping haplotypes with frequencies $p_{m+1}, p_{m+2}, \ldots, p_K$. Harris and DeGiorgio [2020] propose several reasonable choices of f_k, and for all computations here we use $f_k = e^k / \sum_{j=1}^{m} e^{-j}$. The variables U and ε are associated with the amount of mass from non-sweeping haplotypes that are converted to the m sweeping haplotypes [see Harris and DeGiorgio, 2020]. The schematic in Figure 1A illustrates the `LASSI` framework of generating the distorted haplotype spectra.

To incorporate the spatial distribution haplotypic variation into the `LASSI` framework, consider an index set $\mathcal{W} = \{1, 2, \ldots, I\}$ of $I \in \mathbb{Z}^+$ contiguous (potentially overlapping) windows such that window $i \in \mathcal{W}$ has position along a chromosome denoted z_i. This position could be in physical units (such as bases), in genetic map units (such as centiMorgans), in number of polymorphic sites [such as employed by nS_L in Ferrer-Admetlla et al., 2014], or in window number. We model the relative contribution of a sweep with m sweeping haplotypes at target window with index $i^\star \in \mathcal{W}$ by a parameter $\alpha_i \in [0, 1]$ on window $i \in \mathcal{W}$ and the relative contribution of neutrality by $1 - \alpha_i$.

Following a similar powerful framework introduced by Cheng and DeGiorgio [2020] for modeling balancing selection, we employ a mixture model to model the K-haplotype truncated frequency spectrum in window i, with a proportion

$$\alpha_i(A) = \exp\big(-A|z_i - z_{i^\star}|\big)$$

deriving from a sweep model and a proportion $1 - \alpha_i(A)$ deriving from the genome-wide background haplotype spectrum to represent neutrality. Here, A is a parameter that we optimize over, describing the

rate of decay of the effect of the sweep at target window $i^\star$ on the flanking windows a certain distance away. Specifically, we model the K-truncated haplotype spectrum in window i as the vector

$$\mathbf{g}_i^{(m,A)} = \big(g_{i1}^{(m,A)}, g_{i2}^{(m,A)}, \ldots, g_{iK}^{(m,A)}\big),$$

where

$$g_{ik}^{(m,A)} = \alpha_i(A) q_k^{(m)} + [1 - \alpha_i(A)] p_k$$

for $k = 1, 2, \ldots, K$ and $i \in \mathcal{W}$. Note here that for target window $i^\star$, $\alpha_{i^\star}(A) = 1$, and hence $\mathbf{g}_{i^\star}^{(m,A)} = \mathbf{q}_{i^\star}^{(m)}$—*i.e.*, the target window is on top of the sweep, and so it is entirely determined by the distorted m-sweeping haplotype spectrum. However given a fixed A value, for windows i far enough away from the central window $i^\star$, we have the $\alpha_i(A) = 0$, and therefore $\mathbf{g}_i^{(m,A)} = \mathbf{p}$—*i.e.*, the expectation of a neutral window. Based on these trends, windows far from the putatively selected target window are modeled as neutral, and windows close to the target window are heavily distorted due to the sweep. Moreover, because $\alpha_i(A)$ tends to zero for windows far enough away for the central window, the model of neutrality is nested within our proposed sweep model. The schematic in Figure 1B illustrates the `saltiLASSI` framework of generating the spatially-distorted haplotype spectra.

Assume that in window $i \in \mathcal{W}$, there is a K-truncated vector of counts

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK}),$$

which are the observed counts of the K most-frequent haplotypes, with $x_{i1} \geq x_{i2} \geq \cdots \geq x_{iK} \geq 0$ and normalized such that $\sum_{k=1}^{K} x_{ik} = n_i$, where n_i is the total number of sampled haplotypes in window i. Following Cheng and DeGiorgio [2020] and Harris and DeGiorgio [2020], we then compute the log composite likelihood ratios for null hypothesis of neutrality at target window $i^\star$ as

$$\log \mathcal{L}_0\big(\mathbf{p}, K; i^\star, \{\mathbf{x}_i\}_{i\in\mathcal{W}}\big) = \sum_{i\in\mathcal{W}} \sum_{k=1}^{K} x_{ik} \log(p_k)$$

and for the alternative hypothesis of m sweeping haplotypes at target window $i^\star$ as

$$\log \mathcal{L}_1\big(\mathbf{p}, K, \varepsilon, m, A; i^\star, \{\mathbf{x}_i, z_i\}_{i\in\mathcal{W}}\big) = \sum_{i\in\mathcal{W}} \sum_{k=1}^{K} x_{ik} \log\big(g_{ik}^{(m,A)}\big).$$

Using these log likelihoods, we follow Harris and DeGiorgio [2020] and construct a log likelihood ratio test statistic of a sweep at target window $i^\star$ as

$$\Lambda = 2\big[\log \mathcal{L}_1\big(\mathbf{p}, K, \widehat{\varepsilon}, \widehat{m}, \widehat{A}; i^\star, \{\mathbf{x}_i, z_i\}_{i\in\mathcal{W}}\big) - \log \mathcal{L}_0\big(\mathbf{p}, K; i^\star, \{\mathbf{x}_i\}_{i\in\mathcal{W}}\big)\big],$$

where

$$\big(\widehat{m}, \widehat{A}, \widehat{\varepsilon}\big) = \underset{(m, A, \varepsilon)}{\operatorname{argmax}} \log \mathcal{L}_1\big(\mathbf{p}, K, \varepsilon, m, A; i^\star, \{\mathbf{x}_i, z_i\}_{i\in\mathcal{W}}\big).$$

Power to detect sweeps

The power to detect sweeps will depend on a number of factors, including window size used to compute a statistic, whether phasing information for genotypes is used, the selection strength of the beneficial mutation s, the age of the sweep t, the number of selected haplotypes ν, and the underlying demographic history. To explore the power of Λ, we evaluate its power to detect sweeps of varying strengths, softness, and ages. Under each setting, we interrogated its robustness to demographic history, both through idealized constant-size histories and histories with recent severe bottlenecks. Moreover we gauged whether Λ yields false sweep signals under settings of background selection. Furthermore, for each setting described, we investigated the power and robustness of using unphased multilocus genotypes as input to Λ instead of

phased haplotypes. Finally, we compared Λ to competing contemporary methods that use the same type of input data, using the T statistic of Harris and DeGiorgio [2020] for phased and unphased input data, and also considered the H12 [Garud et al., 2015], nS_L [Ferrer-Admetlla et al., 2014], and iHS [Voight et al., 2006] statistics for phased data and the G123 statistic [Harris et al., 2018] for unphased data. The simulation protocol for all settings is described in the *Methods* section.

To begin, we compare the performance of Λ to T, H12, nS_L, and iHS under a constant-size demographic history with diploid effective size of $N = 10^4$ diploid individuals. The Λ, T, and H12 statistics were computed for different window sizes, consisting of 51, 101, or 201 SNPs per window. Figures 2A and S1 show that across sweeps of varying degrees of softness (beneficial mutation on $\nu \in \{1, 2, 4, 8, 16\}$ distinct haplotypes) and for both moderate (per-generation selection coefficient $s = 0.01$) and strong ($s = 0.1$) sweeps, the method with highest power regardless of time of selection ($t \in \{500, 100, 1500, 2000, 2500, 3000\}$ generations prior to sampling) is Λ, thereby outperforming the competing methods. Interestingly, Λ applied to 51 SNP windows has generally higher power than with 101 and 201 SNP windows. Furthermore, smaller window sizes enable Λ to achieve high power even for old sweeps—with this elevated power often substantially higher than the closest competing method. This result recapitulates a finding of Harris and DeGiorgio [2020], where they observed that if the spatial distribution of the T statistic was used within a machine learning framework, computing the T statistic in a greater number of small windows yielded higher power for ancient sweeps than when a smaller number of large windows was used. This is an intriguing result, because smaller windows have poorer estimates of the distortion of the HFS, yet it appears that for detecting ancient sweeps what matters is capturing the overall spatial trend of the distortion of the HFS. That is, when using too large of windows, Λ is averaging the HFS across too large of a region, which has likely been broken up over time due to recombination for ancient sweeps. Instead, smaller windows focus on genomic segments with less shuffling of haplotype variation due to recombination events, such that distortions in the HFS are due to the effect of a sweep at a nearby selected site.

Figure S1 also highlights a key distinction between moderate and strong sweeps. Specifically, regardless of method considered, each achieves its highest power when strong sweeps are recent, whereas for moderate sweeps, highest power for each method is shifted farther in the past toward more ancient sweep. This pattern was also found previously for H12 [Harris et al., 2018] and T [Harris and DeGiorgio, 2020]. The likely reason for this result is that moderate sweeps require more time for the beneficial allele to reach high frequency and leave a conspicuous genomic footprint. In contrast, strong sweeps create an immediate selection signature to appear in the genome due to the rapid rise in frequency of a beneficial mutation, but traces of this sweep pattern erode over time due to recombination, mutation, and drift. However, regardless, the Λ statistic paired with a small window size yields uniformly better or comparable sweep detection ability than the other approaches we examined.

During a scan with Λ, the composite likelihood ratio is optimized over the number of high frequency (sweeping) haplotypes m and the footprint size of the sweep A, leading to respective estimates $\widehat{m}$ and $\widehat{A}$. Therefore, at a genomic location with evidence for a sweep (high Λ value), we may better understand properties of the putative sweep by evaluating its softness through $\widehat{m}$ and its strength or age through $\widehat{A}$. Figure S2 shows that for moderate sweeps, the estimated number of sweeping haplotypes $\widehat{m}$ is considerably different from the actual number of initially-selected haplotypes ν, regardless of window size used or age of the sweep. In contrast, Figures 2A and S2 reveal that for strong hard sweeps ($\nu = 1$), the estimate of the number of sweeping haplotypes when using 51 SNP windows is often consistent with hard sweeps ($\widehat{m} = 1$) provided that the sweep is recent enough (within the last 500 generations). Similarly, under these same settings but with soft sweeps of $\nu \in \{2, 4\}$ selected haplotypes (Figures 2A and S2), the estimated number of sweeping haplotypes tends to be consistent with sweeps of the same level of softness ($\widehat{m} \approx \nu$). Moreover, for softer sweeps ($\nu \in \{8, 16\}$) the number of sweeping haplotypes tends to be underestimated ($\widehat{m} < \nu$) but is still consistent with a soft sweep ($\widehat{m} > 1$). Therefore, provided that a sweep is recent enough, when using 51 SNP windows the value of the estimated number of sweeping haplotypes can be used to lend evidence of a hard ($\widehat{m} = 1$) or a soft ($\widehat{m} > 1$) sweep.

Similarly, the other parameter estimate $\widehat{A}$ may also help characterize identified sweeps. Specifically,

Figures 2A and S3 show that the footprint size of the sweep (measured as $\log_{10}\widehat{A}$) is substantially elevated compared to expectation for neutral simulations for sweep times at which there is high power to detect sweeps (Figures 2A and S1). Interestingly, the shape of the curves relating the mean sweep footprint size over time mirror the power of the Λ statistic with corresponding window size as a function of sweep initiation time (t), sweep softness (ν), and sweep strength (s). These results suggest that the estimate of the sweep footprint size ($\log_{10}\widehat{A}$) can be used to learn about the age or strength of a candidate sweep (the signatures of which appear to be confounded between the two parameters). Coupled with an estimate of the sweep softness ($\widehat{m}$), our `saltiLASSI` framework provides a means to not only detect sweeps with high power, but to also learn the underlying parameters that may have shaped the adaptive evolution of candidate sweep regions.

Obtaining phased haplotypes for input to Λ represents an error-prone step that, without sufficient reference panels or high-enough quality genotypes, may make identification of sweeps difficult or potentially impossible for a number of diverse study systems. It is therefore beneficial if the favorable performance of Λ transfers to datasets that have not been phased. Similar to prior studies [*e.g.*, Harris et al., 2018, Kern and Schrider, 2018, Harris and DeGiorgio, 2020, Harpak et al., 2021], we sought to evaluate the power of Λ when applied to unphased multilocus genotype data, and to compare its performance with the T statistic and G123 [analogue of H12 for use with unphased data; Harris et al., 2018], both of which are also applied to unphased multilocus genotypes. Figure S4 shows that Λ maintains high power to detect sweeps of differing ages, strengths, and softness. Consistent with the results on haplotype data (Figures 2A and S1), Λ generally displays higher power than, or comparable power to, T and G123, with the best performance deriving from Λ with a small window size of 51 SNPs, and with substantially higher power for old sweeps compared to other approaches. An exception is that for recent ($t \leq 1000$ generations) and highly soft ($\nu = 16$) sweeps, using a window size of 101 SNPs for Λ had substantially higher power than using the smaller 51 SNP window. Moreover, for strong ($s = 0.1$) highly soft ($\nu = 16$) and ancient ($t \geq 2000$) sweeps, the power of Λ is much lower with unphased multilocus genotypes compared to phased haplotypes (compare Figures S1 and S4). Interpretation of $\widehat{m}$ is more difficult for multilocus genotypes compared to haplotypes. However, consistent with the results for haplotypes (Figure S2), Figure S5 shows that when using 51 SNP windows, Λ tends to estimate a small number of sweeping multilocus genotypes (smaller $\widehat{m}$) for harder sweeps (smaller ν) than for softer sweeps (larger ν).

While adaptive processes generally affect variation locally in the genome, neutral processes such as demographic history influence overall levels of genome diversity. Specifically, it is common to consider that demographic processes impact the mean value of genetic diversity, and numerous likelihood approaches for detecting sweeps [Kim and Stephan, 2002, Nielsen et al., 2005, Chen et al., 2010, Huber et al., 2015, Vy and Kim, 2015, DeGiorgio et al., 2016, Racimo, 2016, Lee and Coop, 2017, Harris and DeGiorgio, 2020, Setter et al., 2020] and other forms of natural selection [DeGiorgio et al., 2014, Cheng and DeGiorgio, 2019, 2020] have been created to specifically account for this average effect of demographic history on genome diversity. However, demographic processes, such as recent severe bottlenecks, not only alter mean diversity but also influence higher-order moments of diversity, potentially making it insufficient to account solely for the mean effect of diversity [Barton, 1998, Jensen et al., 2005, Pavlidis et al., 2008]. Given that Λ does not account for higher moments than the mean effect of demographic history on the HFS, we sought to evaluate its properties under recent strong bottlenecks—a setting that has proven challenging for other sweep statistics in the past.

The Λ statistic generally exhibits superior power to T, H12, nS_L, and iHS when applied to haplotype data (Figures 2B and S7) or to T and G123 when applied to unphased multilocus genotype data (Figure S10). Moreover, the general trends in method power as a function sweep strength, softness, and age observed for the constant-size history (Figures 2A, S1, and S4) hold for this complex demographic setting (Figures 2B, S7, and S10), with the caveat that, as expected, power for all methods is generally lower under the bottleneck compared to the constant-size history. A clear difference between these two demography settings is that, whereas Λ had exhibited uniformly superior or comparable power with smaller 51 SNP windows compared to larger 101 or 201 SNP windows (Figures 2A and S1), under the bottleneck model the

best window size depends on age of the sweep (Figures 2B and S7). In particular, recent sweeps often had highest power with 201 SNP windows, sweeps of intermediate age with 101 SNPs, and ancient sweeps with 51 SNPs. Therefore, under complex demographic histories, choice of window size for Λ is more nuanced than with constant-size histories. This result is consistent with those of Harris and DeGiorgio [2020] who demonstrated that, when accounting for the spatial distribution of the T statistic in a machine learning framework (referred to as *T-Trendsetter*), power to detect recent sweeps is higher for larger windows and power to detect ancient sweeps is higher for smaller windows under the bottleneck history considered here.

In addition to demographic history, a pervasive force acting to reduce variation across the genome is background selection [McVicker et al., 2009, Lohmueller et al., 2011, Comeron, 2014, Wilson Sayres et al., 2014], which is the loss of genetic diversity at neutral sites due to negative selection at nearby loci [Charlesworth et al., 1993, Hudson and Kaplan, 1995a, Charlesworth, 2012]. Background selection has been demonstrated to alter the neutral SFS [Charlesworth et al., 1993, 1995, Seger et al., 2010, Nicolaisen and Desai, 2013], and masquerade as false signals of positive selection [Charlesworth et al., 1993, 1995, Hudson and Kaplan, 1995a,b, Nordborg et al., 1996, McVean and Charlesworth, 2000, Boyko et al., 2008, Akashi et al., 2012, Charlesworth, 2012, Huber et al., 2015]. However, because this process does not generally lead to haplotypic variation consistent with sweeps [Enard et al., 2014, Fagny et al., 2014, Schrider, 2020], like prior studies developing haplotype approaches for detecting sweeps [Harris et al., 2018, Harris and DeGiorgio, 2020] we sought to evaluate the robustness of Λ to background selection. We find that under both simple and complex demographic histories, using either phased haplotype or unphased multilocus genotype data, all methods considered here demonstrate robustness to background selection by not falsely attributing genomic regions evolving under background selection as sweeps (Figure S13).

Application to empirical data

We next apply the Λ statistic to empirical data representing a European (CEU) and an African (YRI) population from the 1000 Genomes Project Phase 3 [The 1000 Genomes Project Consortium, 2015] to demonstrate its use and check that it identifies sensible results.

We plot the genome-wide Λ statistics for the CEU population in Figure 3A and the YRI population in Figure 3B. We find several conspicuous peaks of notably large Λ values, which indicates strong support for a highly distorted HFS in these regions compared to the genome-wide mean HFS. Using a conservative threshold for determining significance (maximum observed Λ from genome-wide neutral simulations; see *Methods*), we identify several regions in both populations with scores consistently above this threshold, including five regions in the CEU population (Table 1) and 14 in the YRI population (Table 2). Among these regions, we find several well-studied genes that are known to have been under selection in these populations. These include the lactase gene [*LCT*; Tishkoff et al., 2007, Field et al., 2016, Ségurel and Bon, 2017, Taliun et al., 2021], the major histocompatibility complex [MHC; Field et al., 2016, Pierini and Lenz, 2018, Taliun et al., 2021], and the apolipoprotein L1 [*APOL1*; Ko et al., 2013].

We next explore two peaks in detail, the *LCT* and MHC loci in each population (Figure 4). The *LCT* locus has been previously identified as under selection in some northern European populations and eastern African populations [Tishkoff et al., 2007]. As the CEU population has largely northern European ancestry and the YRI population is from western Africa, we expect to find a peak near *LCT* in CEU but not in YRI. Indeed, this is what we see in Figure 4A, which plots Λ statistics in the vicinity of the *LCT* locus on Chromosome 2. Furthermore, we examine the truncated HFS among eleven windows spanning *LCT* in both YRI (Figure 4B) and CEU (Figure 4C). We see in Figure 4B that YRI has haplotype frequencies similar to the genome-wide mean (plotted and highlighted on the left), whereas Figure 4C shows that the CEU population is dominated largely by a single haplotype near 80% frequency. Indeed, the `saltiLASSI` method also infers a $\widehat{m} = 1$ in this region (Table 1), indicating a single sweeping haplotype (*i.e.*, a hard sweep). Furthermore, we can see the HFS in this region trending toward the genome-wide mean as the windows move farther from the sweep's focal point, illustrating the pattern that the `saltiLASSI` method was designed to capture.

Figures 4D-F illustrate the Λ statistics and HFS patterns in the vicinity of the MHC locus. This locus contains a large cluster of immune system genes, and selection at this locus is distinguished from *LCT* in that high diversity is preferred in order for the body to be able to mount a robust response to unknown pathogen exposure. As expected, both populations have extreme Λ values (Figure 4D) and a greatly distorted HFS in this region (Figures 4E and F). However, we note that the HFS is clearly distorted in favor of multiple haplotypes, in contrast to *LCT*, which we expect at a locus that favors diversity. Indeed, the `saltiLASSI` method infers $\widehat{m}$ to be between seven and nine in the CEU population and between eight and 11 for YRI (variance due to multiple regions within the MHC being separately identified; Tables 1 and 2).

Finally, we repeated our analyses of these two populations and two loci using the unphased multilocus-genotype approach (Figures S14 and S15; Tables S2 and S3), and we find good concordance with the phased haplotype approach.

Discussion

In this study, we developed a new likelihood ratio test statistic Λ that examines the spatial distribution of the HFS for evidence of sweeps. We demonstrated that this statistic has high power to detect both hard and soft sweeps, with performance substantially better than competing haplotype-based approaches for the same task. Moreover, while optimizing the model parameters of Λ we obtain estimates of sweep softness m and footprint size A, which is correlated with age and strength of the sweep. These additional parameters have the potential to further characterize well-supported sweep signals from large Λ values.

In addition to lending exceptional performance on simulated data, application of Λ to whole-genome variant calls from central European and sub-Saharan African individuals recapitulated the well-established signal at the *LCT* gene in Europeans due to lactase persistence [Bersaglieri et al., 2004], as well as sweep footprints at the MHC locus in both populations related to immunity, which have previously been detected with other sweep statistics [Albrechtsen et al., 2010, Goeury et al., 2018, Harris and DeGiorgio, 2020]. Though not novel findings, the clear (Figure 4) and strong (Figure 3) signals at these two loci serve as positive controls to highlight the efficacy of Λ. Furthermore, these findings were similarly recapitulated with unphased multilocus genotype data (Figures S14 and S15), lending support for the utility of Λ when applied to study systems for which obtaining phased haplotypes data is challenging.

A key parameter that must be chosen when applying Λ is the number of SNPs per window. Specifically, we found that larger windows had greatest power for more recent sweeps, and smaller windows for more ancient sweeps (Figures 2, S1, and S7), mirroring the window size results observed in Figures S8 and S9 of Harris and DeGiorgio [2020] for the spatial distribution of the T statistic using a different modeling approach. Therefore, choice of window size may be informed by the time frame of selective events that is being investigated. As highlighted in Figures 2B and S7, the Λ statistic computed within windows of 201 SNPs had highest power of all other tested window sizes within the past 2000 generations under the central European demographic history. Because selective events within this time frame are consistent with adaptive events in recent evolution of modern humans [Gravel et al., 2011, Gronau et al., 2011, Schiffels and Durbin, 2014], we selected this size so that we could recapitulate expected well-established sweeps—*e.g.*, Figures 3 and 4 highlighting the sweep signal at *LCT*. In addition to using simulation results to aid in selecting appropriate window sizes, an alternate method such as choosing sizes based on the expected decay of linkage disequilibrium in the genome has been demonstrated to also work well in practice [*e.g.*, Garud et al., 2015, Harris and DeGiorgio, 2020].

The T statistic of Harris and DeGiorgio [2020] presented the first likelihood approach that evaluated distortions in the HFS to detect selective sweeps, importantly because neutrality and soft sweeps leave similar signatures in the SFS but different within the HFS [Pennings and Hermisson, 2006b]. As demonstrated by Harris and DeGiorgio [2020], using the spatial distribution of the T statistic within a machine learning framework enhanced it's detection ability, specifically for ancient sweeps. However, machine learning frameworks require extensive simulations to train [*e.g.*, Schrider and Kern, 2016, Sheehan and Song,

2016, Mughal and DeGiorgio, 2019], and these simulations must be based on a set of critical assumptions, such as demographic, mutation rate, and recombination rate parameters. Yet, accurate inferences of these parameters is not always possible, or can be highly error prone, and prior studies have found that these machine learning methods can make highly incorrect predictions if the distribution of training data is different from that of the test or empirical data [Mughal and DeGiorgio, 2019, Mughal et al., 2020]. Furthermore, generation of these training datasets and training the models on them often requires substantial computational time and resources. Instead, our Λ statistic is the first likelihood method to model the spatial distribution of the HFS, providing the power of modeling the spatial distribution of T afforded by current machine learning frameworks (*e.g.*, compare Figures S1 and S7 with Figures S8 and S9 of Harris and DeGiorgio [2020]), yet with massive savings in computational speed and with predictions not hinging on accurate estimates of genetic and evolutionary model parameters to generate training sets.

While optimizing the Λ statistic, we also obtain estimates of the number of presently-sweeping haplotypes m and the footprint size A. For recent strong sweeps, estimates of m correlate well with the number of initially-selected haplotypes ν. For older and less strong sweeps, mutation and recombination events accumulate leading to more distinct haplotypes, thereby inflating m estimates. Moreover, estimates of the footprint size A correlate with power of Λ, suggesting that the estimated footprint size will be large under scenarios in which sweeps are highly supported. The relationship between A and power of Λ is related to prominence of the distortions in the HFS, which also erode due mutation and recombination rates. Therefore, though we found that estimates of m and A were not highly accurate under non-ideal sweeps settings, they may still be useful. Specifically, though not directly associated with population-genetic parameters such as ν or the strength s and time t of a sweep, estimated Λ, μ, and A values can be used as input features to machine learning regression algorithms to predict underlying evolutionary model parameters of ν, s, and t [Hastie et al., 2009]. Such strategies are typically computationally expensive, but may be required for accurate characterization of sweep footprints, even though they are unnecessary for detecting sweeps due to the already high power of Λ.

The Λ statistic developed here represents an important step in advancing methodology for sweep detection by interrogating the spatial distribution of distortions in the HFS. Prior studies focused either on spatial distributions of the SFS, which cannot distinguish between hard and soft sweeps, or only local distortions in the HFS. Specifically, methods that explore the skews in the SFS typically do so with an explicit analytical population-genetic model [Kim and Stephan, 2002, Nielsen et al., 2005, Huber et al., 2015, Vy and Kim, 2015, DeGiorgio et al., 2016], which can be underpowered if the assumed model is incorrect and are unable to detect soft sweeps [Pennings and Hermisson, 2006b]. In contrast, analytical population-genetic modeling of distortions in the HFS is difficult, and alternative statistical models that capture relevant features of sweeps are often used, focusing either on local distortions in the HFS [Harris and DeGiorgio, 2020] or haplotype length distributions [Voight et al., 2006, Ferrer-Admetlla et al., 2014]. Instead, our Λ statistic represents a compromise of these two extremes, permitting simultaneous interrogations of haplotype frequency distributions and correlates of their length distributions in a computationally efficient framework that leads to expected patterns that are informed by theoretical results. Our methodological framework therefore provides a foundation for developing tools that can identify other evolutionary processes that may act locally in the genome, enhancing future investigations of sweeps and other forces across a variety of study systems.

Methods

In this section we outline the methods used to assess the power of a diversity of sweep statistics using simulations. These simulations examine an array of model parameters, including sweep strength, age, and softness as well as the confounding effects of demographic history, background selection, and haplotype phasing. We also describe pre- and post-analysis processing for the application of the Λ statistic to phased haplotype and unphased multiloucs genotype data from a pair of human populations.

Power Analysis

To assess the ability of Λ to detect sweeps, we conducted forward-time simulations using `SLiMv3.2` [Haller and Messer, 2019] for sweeps of varying strength, age, and softness under a constant-size demographic history as well as under a realistic non-equilibrium demographic history inspired by human studies. Specifically, for each simulation scenario, we generated 1000 independent replicates of length 500 kb, so that Λ was able to interrogate the spatial distribution of variation across a large genomic segment. We employed a mutation rate of $\mu = 2.5 \times 10^{-8}$ per site per generation [Nachman and Crowell, 2000] and a recombination rate of $r = 10^{-8}$ per site per generation [Payseur and Nachman, 2000]. For the constant-size demographic history, we considered a population size of $N = 10^4$ diploid individuals [Takahata, 1993], and to investigate complex non-equilibrium demographic histories, we also employed the model inferred in Terhorst et al. [2017] of central European humans (CEU), which incorporates a recent bottleneck with a severe population collapse followed by rapid population expansion. In particular, we used this non-equilibrium model as it was inferred by the contemporary method `SMC++` [Terhorst et al., 2017], which attempts to fit model parameters that can both recapitulate haplotype diversity and allele frequency distributions [Beichman et al., 2017] observed in genomic data from the CEU population of the 1000 Genomes Project dataset [The 1000 Genomes Project Consortium, 2015].

In addition to these genetic and demographic parameters, for selection simulations, we modeled sweeps on $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes, where each of these haplotypes harbored a beneficial allele in the center of the simulated genomic segment with strength $s \in \{0.01, 0.1\}$ per generation that immediately appeared and became beneficial at time $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generation prior to sampling. To ensure that a sweep signature had the potential to be uncovered (especially under settings with $s = 0.01$), we required that the beneficial allele established in the population by reaching a frequency of 0.1 in the population. Simulation replicates for which the beneficial allele did not reach a frequency of 0.1 in the population were repeated until the beneficial allele established in the population. All neutral and selection simulations were run for $11N$ generations, where the first $10N$ generations were used as burn-in and $n = 50$ diploid individuals were sampled from the population after $11N$ generations (*i.e.*, the present). Because forward-time simulations are computationally intensive, as is commonly-practiced [Yuan et al., 2012, Ruths and Nakhleh, 2013] we scaled all constant-size demographic history simulations by a factor $\lambda = 10$ and the European human history history by $\lambda = 20$, such that the selection coefficient, mutation rate, and recombination rate were multiplied by λ and the population size at each generation and the total number of simulated generations was divided by λ. This scaling leads to a speedup of approximately λ^2 in computing time, such that the constant-size simulations run roughly 100 times faster than without scaling and the CEU model simulations run approximately 400 times faster, making a large-scale simulation study feasible.

When analyzing each simulated replicate, we examined the performance of Λ with the likelihood T statistic [Harris and DeGiorgio, 2020] that does not account for the spatial distribution of genomic variation, the summary statistic H12 [Garud et al., 2015] that was developed to detect hard and soft sweeps with similar power, and the standardized iHS [Voight et al., 2006] and nS_L Ferrer-Admetlla et al. [2014] methods that summarize the lengths of haplotypes centered on core SNPs. To investigate the effect of window size on the relative powers of Λ, T, and H12, we considered their applications in central windows of 51, 101, and 201 SNPs, and analyzed windows every 25 SNPs across a simulated sequence. We chose SNP-delimited windows rather than windows based on physical length as they should be more robust to variation in recombination and mutation rate across the genome, as well as random missing genomic segments due to poor mappability, alignability, or sequence quality. That is, we expect SNP-delimited to be more conservative than windows based on the physical length of an analyzed genomic segment. We also examined the application of Λ, T, and G123 [Harris et al., 2018, analogue of H12] to unphased multilocus genotype input data to evaluate the relative powers of these three approaches when applied on study systems for which obtaining phased haplotypes is difficult, unreliable, or impossible [Mallick et al., 2009]. We applied the `lassip` software released with this article for application of the `saltiLASSI` Λ statistic, the `LASSI` T statistic, and H12 (and G123), and the `selscan` software [Szpiech and Hernandez, 2014] to compute standardized iHS and

nS_L.

Analysis of 1000 Genomes Data

We extracted the phased genomes of CEU (99 diploids) and YRI (108 diploids) populations, separately, from the full 1000 Genomes Project Phase 3 dataset (2504 diploids) [The 1000 Genomes Project Consortium, 2015]. For each population, we retained only autosomal biallelic SNPs that were polymorphic in the sample. In order to avoid potentially spurious signals, we also filtered any regions with poor mapability as indicated by mean CRG100 < 0.9 [Derrien et al., 2012, Huber et al., 2015]. This left 12,400,078 SNPs in CEU and 20,417,698 SNPs in YRI.

We compute `saltiLASSI` Λ statistics for both phased (haplotype-based) and unphased (multilocus-genotype-based) analyses with `lassip`. We set `--winsize 201` and `--winstep 100`, and we choose `--k 20` to use the ranked HFS for the top 20 most frequent haplotypes. By default `lassip` assumes phased data and computes haplotype-based statistics, when the `--unphased` flag is set, all statistics are computed using multilocus genotypes.

To determine significance thresholds, we simulated neutral whole genomes with a realistic recombination map and demographic history using `stdpopsim` [Adrion et al., 2020] and `msprime` [Kelleher et al., 2016]. Using the `OutOfAfrica_2T12` demographic history [Tennessen et al., 2012] and the `HapMapII_GRCh37` genetic map [Consortium, 2007] in `stdpopsim`, we simulate 100 replicates of all 22 autosomes for each population separately, sampling 99 diploid individuals for CEU simulations and 108 diploid individuals for YRI simulations. For each replicate, we then compute `saltiLASSI` Λ statistics for both phased and unphased analyses with `lassip`, setting `--winsize 201`, `--winstep 100`, and `--k 20`. We then compute the max Λ, the top-0.1% Λ, and the top-1% Λ across all replicates for each population and each analysis (phased/unphased), which are given in Table S1. Putatively selected regions were identified by concatenating consecutive windows with Λ greater than the max observed across all simulations for a given population and analysis (phased/unphased).

Acknowledgments

This work was supported by National Institutes of Health grant R35GM128590, by National Science Foundation grants DEB-1949268 and BCS-2001063, and by Pennsylvania State University startup funds. Computations for this research were performed using the services provided by Research Computing at the Florida Atlantic University and using the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer.

References

JR Adrion, CB Cole, N Dukler, JG Galloway, AL Gladstein, G Gower, CC Kyriazis, AP Ragsdale, G Tsambos, F Baumdicker, J Carlson, RA Cartwright, A Durvasula, I Gronau, BY Kim, P McKenzie, PW Messer, E Noskova, D Ortega-Del Vecchyo, F Racimo, TJ Struck, S Gravel, RN Gutenkunst, KE Lohmueller, PL Ralph, DR Schrider, A Siepel, J Kelleher, and AD Kern. A community-maintained standard library of population genetic models. *eLife*, 9:e54967, 2020.

H Akashi, N Osada, and T Ohta. Weak selection and protein evolution. *Genetics*, 192:15–31, 2012.

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.

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

AC Beichman, TN Phung, and KE Lohmueller. Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3 (Bethesda)*, 7:3605–3620, 2017.

T Bersaglieri, PC Sabeti, N Patterson, T Vanderploeg, T Schaffner, JA Drake, M Rhodes, DE Reich, and JN Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*, 74:1111–1120, 2004.

AR Boyko, SH Williamson, AR Indap, JD Degenhardt, RD Hernandez, KE Lohmueller, MD Adams, S Schmidt, JJ Sninsky, SR Sunyaev, TJ White, R Nielsen, AG Clark, and CD Bustamante. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 30:e1000083, 2008.

B Charlesworth. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. *Genetics*, 191:233–2463, 2012.

B Charlesworth, MT Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303, 1993.

D Charlesworth, B Charlesworth, and MT Morgan. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141:1619–1632, 1995.

H Chen, N Patterson, and D Reich. Population differentiation as a test for selective sweeps. *Genome Res*, 20:393–402, 2010.

X Cheng and M DeGiorgio. Detection of shared balancing selection in the absence of trans-species polymorphism. *Mol Biol Evol*, 36:177–199, 2019.

X Cheng and M DeGiorgio. Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. *Mol Biol Evol*, 37:3267–3291, 2020.

JM Comeron. Background selection as a baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genet*, 10:e1004434, 2014.

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:841, 2007.

M DeGiorgio, KE 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, CD Huber, MJ Hubisz, I Hellmann, and R Nielsen. *SweepFinder2*: Increased sensitivity, robustness, and flexibility. *Bioinformatics*, 32:1895–1897, 2016.

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

D Enard, PW Messer, and DA Petrov. Genome-wide signals of positive selection in human evolution. *Genome Res*, 24:884–895, 2014.

M Fagny, E Patin, D Enard, 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.

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, EA Boyle, N Telis, Z Gao, KJ Gaulton, D Golan, L Yengo, G Rocheleau, P Froguel, MI McCarthy, and JK Pritchard. Detection of human adaptation during the past 2000 years. *Science*, 354:760–764, 2016.

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

T Goeury, LE Creary, L Brunet, M Galan, M Pasquier, B Kervaire, A Langaney, J-M Tiercey, MA Fernández-Viña, JM Nunes, and A Sachez-Mazas. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*, 91: 36–51, 2018.

S Gravel, BM Henn, RN Gutenkunst, AR Indap, GT Marth, AG Clark, F Yu, R Gibbs, The 1000 Genomes Project, and D Bustamante. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA*, 108:11983–11988, 2011.

I Gronau, MJ Hubisz, B Gulko, CG Danko, and A Siepel. Bayesian inference of ancient human demography from individuals genomes. *Nat Genet*, 43:1031–1034, 2011.

BC Haller and PW Messer. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol*, 36:632–637, 2019.

A Harpak, N Garud, NA Rosenberg, DA Petrov, M Combs, PS Pennings, and J Munshi-South. Genetic adaptation in New York City rats. *Genome Biol Evol*, 13:evaa247, 2021.

AM Harris and M DeGiorgio. A likelihood approach for uncovering selective sweep signatures from haplotype data. *Mol Biol Evol*, 215:143–171, 2020.

AM Harris, NR Garud, and M DeGiorgio. Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. *Genetics*, 210:1429–1452, 2018.

T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, NY, 2nd edition, 2009.

J Hermisson and PS Pennings. Soft sweeps. *Genetics*, 4:2335–2352, 2005.

CD 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, 2015.

RR Hudson and NL Kaplan. Deleterious background selection with recombination. *Genetics*, 141:1605–1617, 1995a.

RR Hudson and NL Kaplan. The coalescent process and background selection. *Philos Trans R Soc B*, 349: 19–23, 1995b.

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

J Kelleher, AM Etheridge, and G McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12:1–12, 2016.

AD Kern and DR Schrider. diploS/HIC: an updated approach to classifying selective sweeps. *G3 (Bethesda)*, 8:1959–1970, 2018.

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

W-Y Ko, P Rajan, F Gomez, L Scheinfeldt, P An, CA Winkler, A Froment, TB Nyambo, SA Omar, C Wambebe, A Ranciaro, JB Hirbo, and SA Tishkoff. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. *Am J Hum Genet*, 93:54–66, 2013.

KM Lee and G Coop. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics*, 207:1591–1619, 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.

KE Lohmueller, A Albrechtsen, Y Li, SY Kim, T Koneliussen, N Vinckenbosch, G Tian, E Huerta-Sanchez, AF Feder, N Garup, T Jørgensen, T Jiang, DR Witte, A Sandbæk, I Hellmann, T Lauritzen, T Hansen, O Pedersen, J Wang, and R Nielsen. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet*, 7:e1002326, 2011.

S Mallick, S Gnerre, and D Reich. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*, 19:922–933, 2009.

GA McVean and B Charlesworth. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, 155:929–944, 2000.

G McVicker, D Gordon, C Davis, and P Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5:e1000471, 2009.

MR Mughal and M DeGiorgio. Localizing and classifying selective sweeps with trend filtered regression. *Mol Biol Evol*, 36:252–270, 2019.

MR Mughal, H Koch, J Huang, F Chiaromonte, and M DeGiorgio. Learning the properties of adaptive regions with functional data analysis. *PLoS Genet*, in press, 2020.

MW Nachman and SL Crowell. Estimation of the mutation rate per nucleotide in humans. *Genetics*, 156: 297–304, 2000.

LE Nicolaisen and MM Desai. Distortions in genealogies due to purifying selection and recombination. *Genetics*, 194:221–230, 2013.

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

M Nordborg, B Charlesworth, and D Charlesworth. The effect of recombination of background selection. *Genet Res*, 67:159–174, 1996.

P Pavlidis, S Hutter, and W Stephan. A population genomic approach to map recent positive selection in model species. *Mol Ecol*, 17:3585–2598, 2008.

BA Payseur and MW Nachman. Micorsatelllite variation and recombination rate in the human genome. *Genetics*, 156:1285–1298, 2000.

PS Pennings and J Hermisson. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol*, 23:1076–1084, 2006a.

PS Pennings and J Hermisson. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet*, 2:1–15, 2006b.

F Pierini and TL Lenz. Divergent allele advantage at human MHC genes: Signatures of past and ongoing selection. *Mol Biol Evol*, 35:2145–2158, 2018.

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*, 202:733–750, 2016.

T Ruths and L Nakhleh. Boosting forward-time population genetic simulators through genotype compression. *BMC Bioinformatics*, 14, 2013. doi: 10.1186/1471-2105-14-192.

PC Sabeti, DE Reich, JM Higgins, HZP Levine, DJ Richter, SF Schaffner, SB Gabriel, JV Platko, NJ Patterson, GJ McDonald, HC Ackerman, SJ Campbell, D Altshuler, R Cooper, D Kwiatkowski, R Ward, and ES Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, 2002.

PC Sabeti, P Varilly, B Fry, J Lohmueller, E Hostetter, C Cotsapas, X Xie, EH Byrne, SA McCarroll, R Gaudet, SF Schaffner, ES Lander, and The International HapMap Consortium. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918, 2007.

S Schiffels and R Durbin. Inferring human popualtion size and separation history from multiple genome sequences. *Nat Genet*, 46:919–925, 2014.

DR Schrider. Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics*, 216:499–519, 2020.

DR Schrider and AD Kern. S/HIC: robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, 12:1–31, 2016.

J Seger, WA Smith, JJ Prry, J Hunn, ZA Kaliszewska, L La Sala, L Pozzi, VJ Rowntree, and FR Adler. Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics*, 184:529–545, 2010.

L Ségurel and C Bon. On the evolution of lactase persistence in humans. *Ann Rev Genomics Hum Genet*, 18:297–319, 2017.

D Setter, S Mousset, X Cheng, R Nielsen, M DeGiorgio, and J Hermisson. VolcanoFinder: genomic scans of adaptive introgression. *PLoS Genet*, 16:e1008867, 2020.

S Sheehan and YS Song. Deep learning for population genetic inference. *PLoS Comput Biol*, 12:1–28, 2016.

ZA Szpiech and RD Hernandez. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*, 31:2824–2827, 2014.

ZA Szpiech, TE Novak, NP Bailey, and LS Stevison. Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *bioRxiv*, doi.org/10.1101/2020.05.19.104380, 2020.

N Takahata. Allelic genealogy and human evolution. *Mol Biol Evol*, 10:2–22, 1993.

D Taliun, DN Harris, MD Kessler, J Carlson, ZA Szpiech, R Torres, SA Gagliano Taliun, A Corvelo, SM Gogarten, HM Kang, AN Pitsillides, J LeFaive, S-b Lee, X Tian, BL Browning, S Das, A-K Emde, WE Clarke, DP Loesch, AC Shetty, TW Blackwell, AV Smith, Q Wong, X Liu, MP Conomos, DM Bobo, F Aguet, C Albert, A Alonso, KG Ardlie, DE Arking, S Aslibekyan, PL Auer, J Barnard, RG Barr, L Barwick, LC Becker, RL Beer, EJ Benjamin, LF Bialek, J Blangero, M Boehnke, DW Bowden, JA Brody, EG Buchard, BE Cade, JF Casella, B Chalazan, DI Chasman, IY-D Chen, MH Cho, SH Choi, MK Chung, CB Clish, A Correa, JE Curran, B Custer, D Darbar, M Daya, M de Andrade, DL DeMeo,

SK Dutcher, PT Ellinor, LS Emery, C Eng, D Fatkin, T Fingerlin, L Forer, M Fornage, N Franceschini, C Fuchsberger, SM Fullerton, S Germer, MT Gladwin, DJ Gottlieb, X Guo, ME Hall, J He, NL Heard-Costa, SR Heckbert, MR Irvin, JM Johnsen, AD Johnson, R Kaplan, SLR Kardia, T Kelly, S Kelly, EE Kenny, DP Kiel, R Klemmer, BA Konkle, C Kooperberg, A Köttgen, LA Lange, J Lasky-Su, D Levy, X Lin, K-H Lin, C Liu, RJF Loos, L Garman, R Gerszten, SA Lubitz, KL Lunetta, ACY Mak, A Manichaikul, AK Manning, RA Mathias, DD McManus, ST McGarvey, JB Meigs, DA Meyers, JL Mikulla, MA Minear, BD Mitchell, S Mohanty, ME Montasser, C Montgomery, AC Morrison, JM Murabito, A Natale, P Natarajan, SC Nelson, KE North, JR O'Connell, ND Palmer, N Pankratz, GM Peloso, PA Peyser, J Pleiness, WS Post, BM Psaty, DC Rao, S Redline, AP Reiner, D Rode, JI Rotter, I Ruczinski, C Sarnowski, S Schoenherr, DA Schwartz, J-S Seo, S Seshadri, VA Sheehan, WH Sheu, BM Shoemaker, NL Smith, JA Smith, N Sotoodehnia, AM Stilp, W Tang, KD Taylor, M Telen, TA Thornton, RP Tracy, DJ Van Den Berg, RS Vasan, KA Viaud-Martinez, S Vrieze, DE Weeks, BS Weir, ST Weiss, L-C Weng, CJ Willer, Y Zhang, X Zhao, DK Arnett, AE Ashley-Koch, KC Barnes, E Boerwinkle, S Gabriel, R Gibbs, KM Rice, SS Rich, EK Silverman, P Qasba, W Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, GJ Papanicolaou, DA Nickerson, SR Browning, MC Zody, S Zöllner, JG Wilson, LA Cupples, CC Laurie, CE Jaquish, RD Hernandez, TD O'Connor, and GR Abecasis. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590:290–299, 2021.

J Tennessen, AW Bigham, TD O'Connor, W Fu, EE Kenny, S Gravel, S McGee, R Do, X Liu, G Jun, HM Kang, D Jordan, SM Leal, S Gabriel, MJ Rieder, G Abecasis, D Altshuler, DA Nickerson, E Boerwinkle, S Sunyaev, CD Bustamante, MJ Bamshad, JM Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337:64–69, 2012.

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

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

SA Tishkoff, FA Reed, A Ranciaro, BF Voight, CC Babbitt, JS Silverman, K Powell, HM Mortensen, JB Hirbo, M Osman, M Ibrahim, SA Omar, G Lema, TB Nyambo, J Ghori, S Numpstead, JK Pritchard, GA Wray, and P Deloukas. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*, 39:31–40, 2007.

R Torres, ZA Szpiech, and RD Hernandez. Human demographic history has amplified the effects of background selection across the genome. *PLoS genetics*, 14(6):e1007387, 2018.

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

HMT Vy and Y Kim. A composite-likelihood method for detecting incomplete selective sweep from population genomic data. *Genetics*, 200:633–649, 2015.

MA Wilson Sayres, KE Lohmueller, and R Nielsen. Natural selection reduced diversity on human Y chromosomes. *PLoS Genet*, 10:e1004064, 2014.

X Yuan, D J Miller, J Zhang, D Herrington, and Y Wang. An overview of population genetic data simulation. *J Comput Biol*, 19:42–54, 2012.

Chromosome	Start (bp)	Stop (bp)	$\widehat{m}$	$\log_{10}(\widehat{A})$	Max Λ	Genes
2	135,517,106	136,318,189	1	7.817	1889.370	*ACMSD*, *MIR5590*, *CCNT2-AS1*, *CCNT2*, *MAP3K19*, *RAB3GAP1*, *ZRANB3*, *R3HDM1*
2	136,524,766	136,816,336	1	7.817	2250.050	*UBXN4*, *LCT*, *LOC100507600*, *MCM6*, *DARS*, *DARS-AS1*
4	34,296,435	34,400,578	1	8.252	1296.390	–
6	29,782,470	29,984,591	9	7.817	1366.550	*HLA-G*, *HLA-H*, *HCG4B*, *HLA-A*, *HCG9*, *ZNRD1-AS1*, *HLA-J*, *HCG8*
6	32,384,933	32,723,916	7	7.817	4565.340	*HLA-DRA*, *HLA-DRB5*, *HLA-DRB6*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DQB1-AS1*, *HLA-DQA2*, *MIR3135B*, *HLA-DQB2*

Table 1: Regions of extreme Λ values in the CEU population and the genes contained therein. $\widehat{m}$ is the inferred number of sweeping haplotypes, and $\log_{10}(\widehat{A})$ is the estimated sweep width.

Chromosome	Start (bp)	Stop (bp)	$\widehat{m}$	$\log_{10}(\widehat{A})$	Max Λ	Genes
3	46,125,575	46,360,655	6	8.252	839.195	*CCR1*, *CCR3*
3	87,273,225	87,311,124	6	8.252	665.802	*MIR4795*, *CHMP2B*, *POU1F1*
3	162,552,546	162,631,343	6	8.252	711.747	–
3	162,640,076	162,659,093	8	8.252	619.686	–
6	31,210,058	31,382,135	11	7.817	918.650	*HLA-C*, *HLA-B*, *MIR6891*, *MICA*
6	32,374,191	32,743,860	8	7.817	5385.050	*BTNL2*, *HLA-DRA*, *HLA-DRB5*, *HLA-DRB6*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DQB1-AS1*, *HLA-DQA2*, *MIR3135B*, *HLA-DQB2*
6	32,983,818	33,185,400	8	7.817	1196.890	*HLA-DPA1*, *HLA-DPB1*, *HLA-DPB2*, *COL11A2*, *RXRB*, *SLC39A7*, *HSD17B8*, *MIR219A1*, *RING1*
8	50,077,099	50,205,908	6	7.817	672.218	–
10	102,184,232	102,282,125	5	8.252	717.760	*WNT8B*, *SEC31B*
12	79,578,861	79,785,498	6	8.252	917.585	*SYT1*
15	55,148,217	55,267,961	7	8.252	730.824	–
17	3,529,529	3,672,429	6	8.252	911.490	*SHPK*, *CTNS*, *TAX1BP3*, *P2RX5-TAX1BP3*, *EMC6*, *P2RX5*, *ITGAE*, *GSG2*
20	37,432,116	37,470,174	5	8.686	613.034	*PPP1R16B*
22	36,584,162	36,727,622	6	8.252	796.478	*APOL4*, *APOL2*, *APOL1*, *MYH9*, *MIR6819*

Table 2: Regions of extreme Λ values in the YRI population and the genes contained therein. $\widehat{m}$ is the inferred number of sweeping haplotypes, and $\log_{10}(\widehat{A})$ is the estimated sweep width.

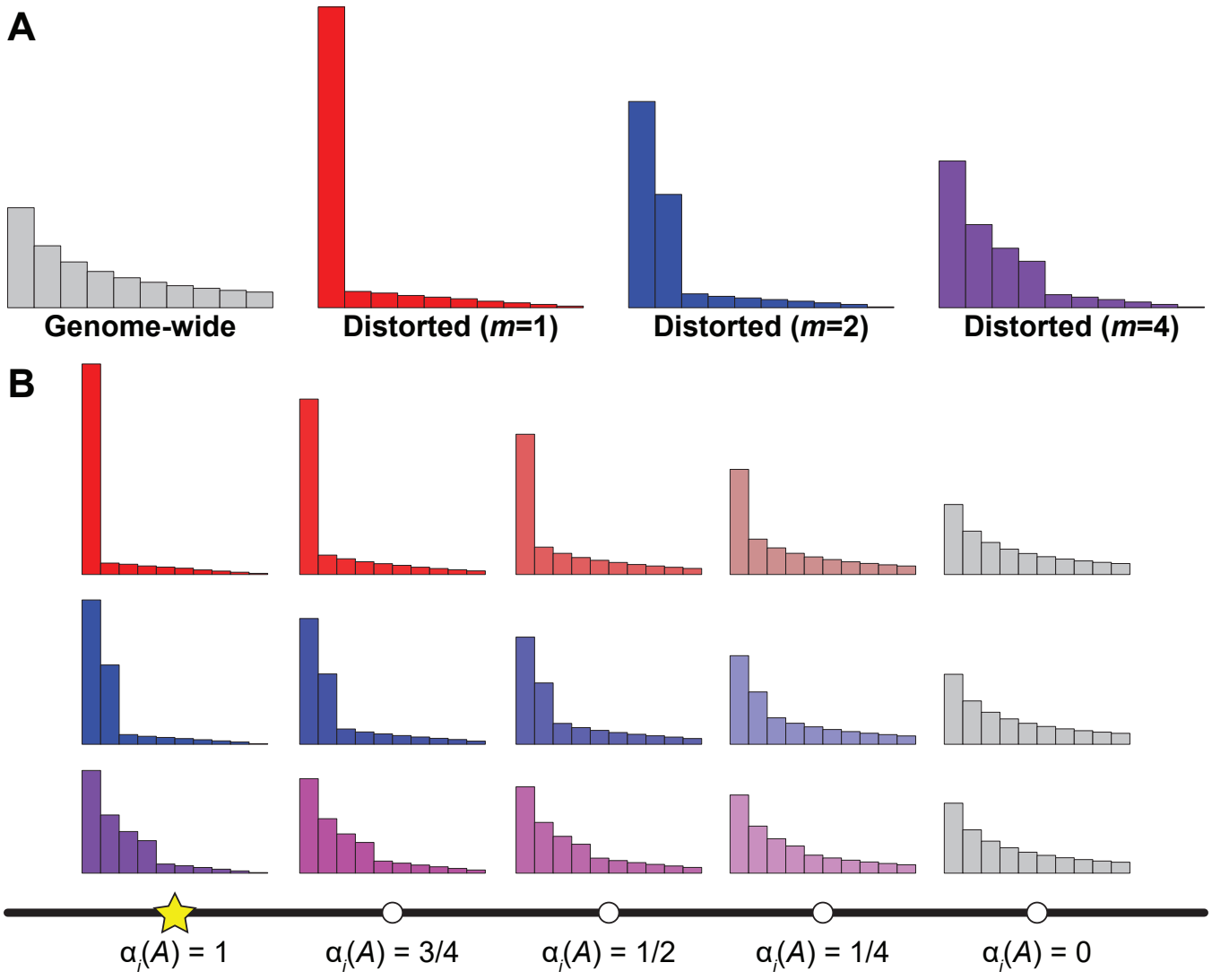

Figure 1: Schematic of the `saltiLASSI` mixture model framework. (A) Generation of distorted haplotype frequency spectra (HFS) for $m = 1$ (red), 2 (blue), and 4 (purple) sweeping haplotypes from a genome-wide (gray) neutral HFS under the `LASSI` framework of Harris and DeGiorgio [2020]. (B) Generation of spatially-distorted HFS under the `saltiLASSI` framework for a window i (white circles) with increasing distance from the sweep location (yellow star). When the window is on top of the sweep location, the HFS is identical to the distorted `LASSI` HFS, and $\alpha_i(A) = 1$. When a window is far from the sweep location, the HFS is identical to the genome-wide (neutral) HFS, and $\alpha_i(A) = 0$. For windows at intermediate distances from the sweep location, the HFS is a mixture of the distorted and genome-wide HFS, with the distorted HFS contributing $\alpha_i(A)$ and the genome-wide HFS contributing $1 - \alpha_i(A)$.

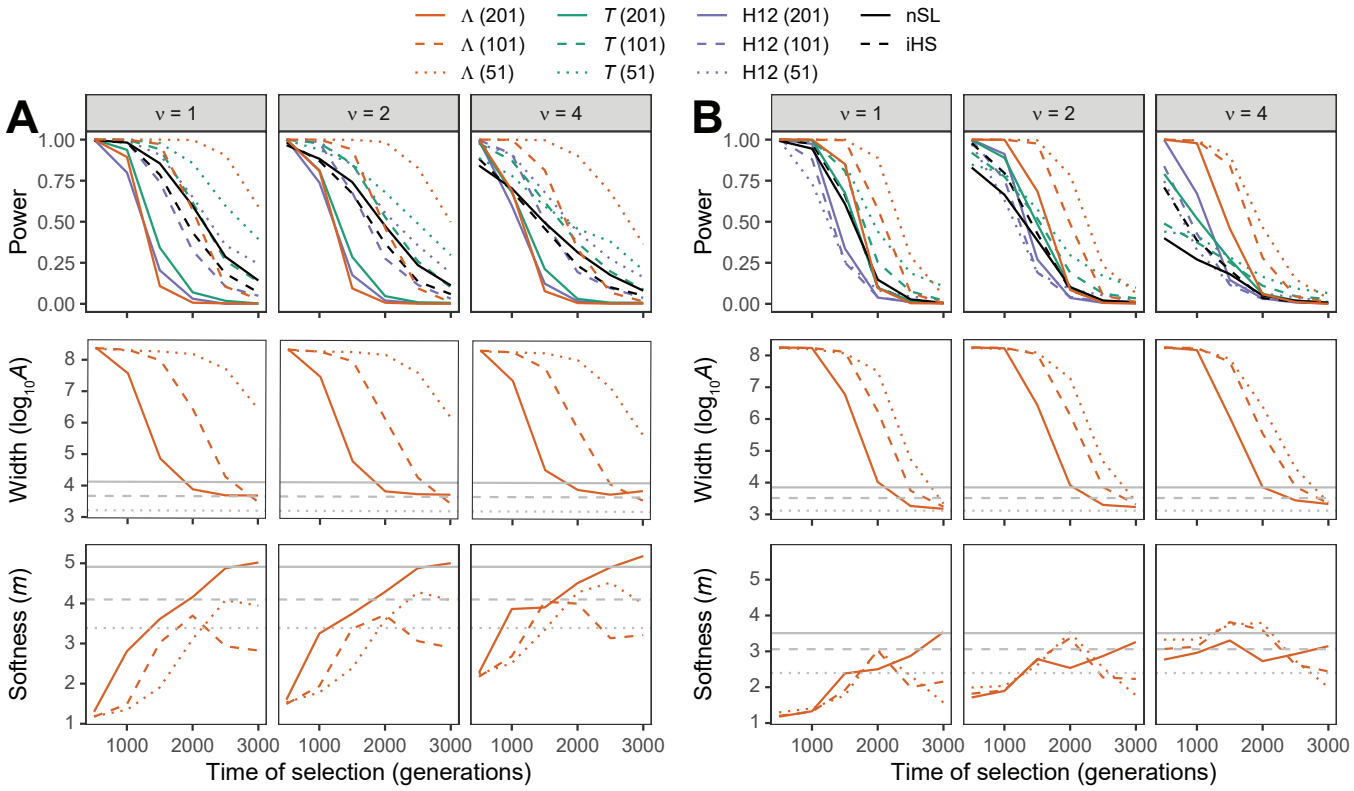

Figure 2: Performance of detecting and characterizing sweeps for applications of Λ, T, and H12 with windows of size 51, 101, and 201 SNPs, as well nS_L and iHS under simulations of (A) a constant-size demographic history or (B) the human central European (CEU) demographic history of Terhorst et al. [2017]. (Top row) Power at a 1% false positive rate as a function of selection start time. (Middle row) Estimated sweep width illustrated by mean estimated genomic size influenced by the sweep ($\log_{10} \widehat{A}$) as a function of selection start time. (Bottom row) Estimated sweep softness illustrated by mean estimated number of sweeping haplotypes ($\widehat{m}$) as a function of selection start time. Sweep scenarios consist of hard ($\nu = 1$) and soft ($\nu \in \{2, 4\}$) sweeps with per-generation selection coefficient of $s = 0.1$ that started at $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling. Results expanded across wider range of simulation settings can be found in Figures S1-S3 and S7-S9.

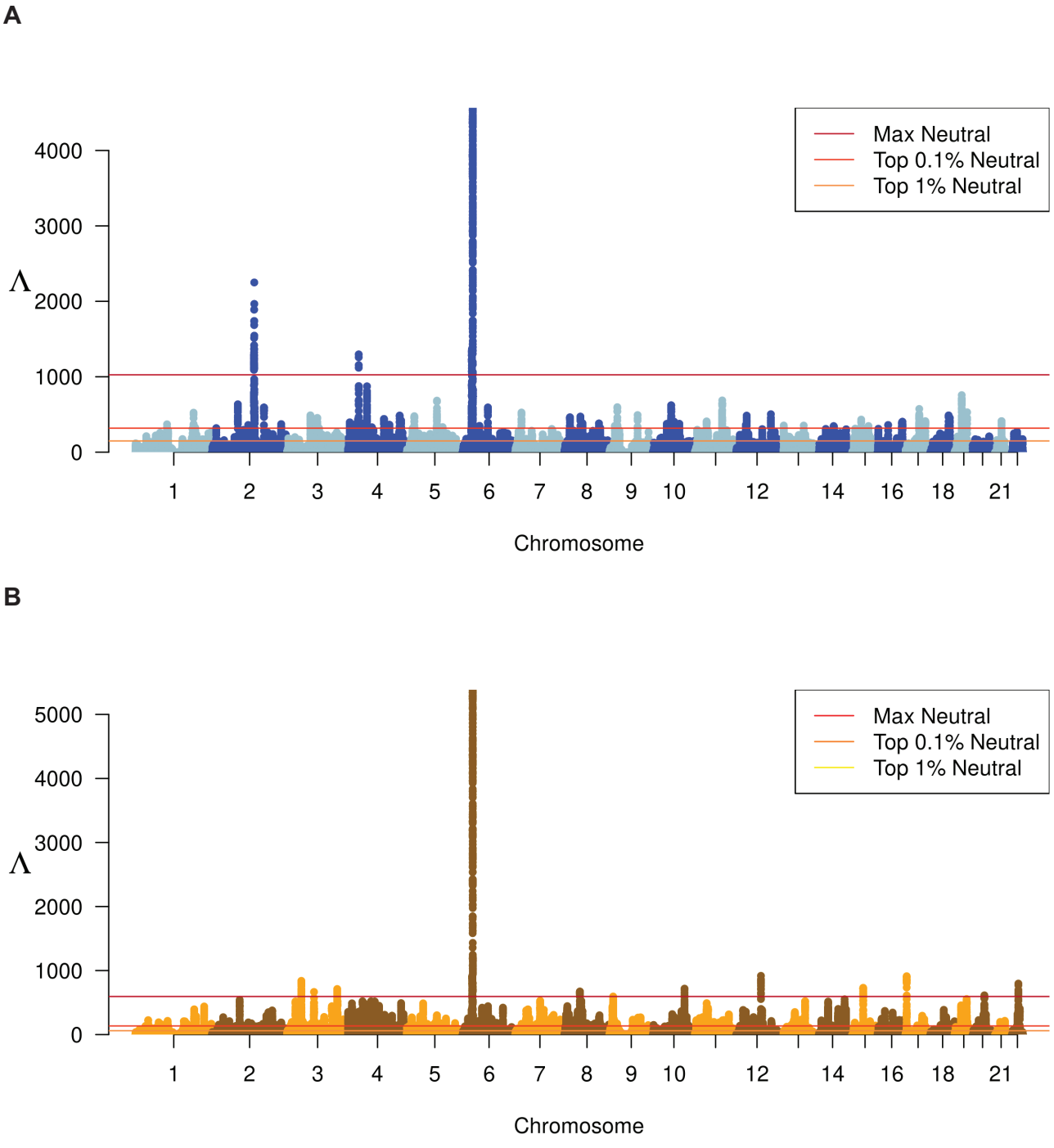

Figure 3: Manhattan plot of Λ-statistics for the (A) CEU and (B) YRI populations from the 1000 Genomes Project. Each point represents a single 201-SNP window along the genome. Horizontal lines represent the top 1%, top 0.1%, and maximum observed Λ statistic in demography-matched neutral simulations.

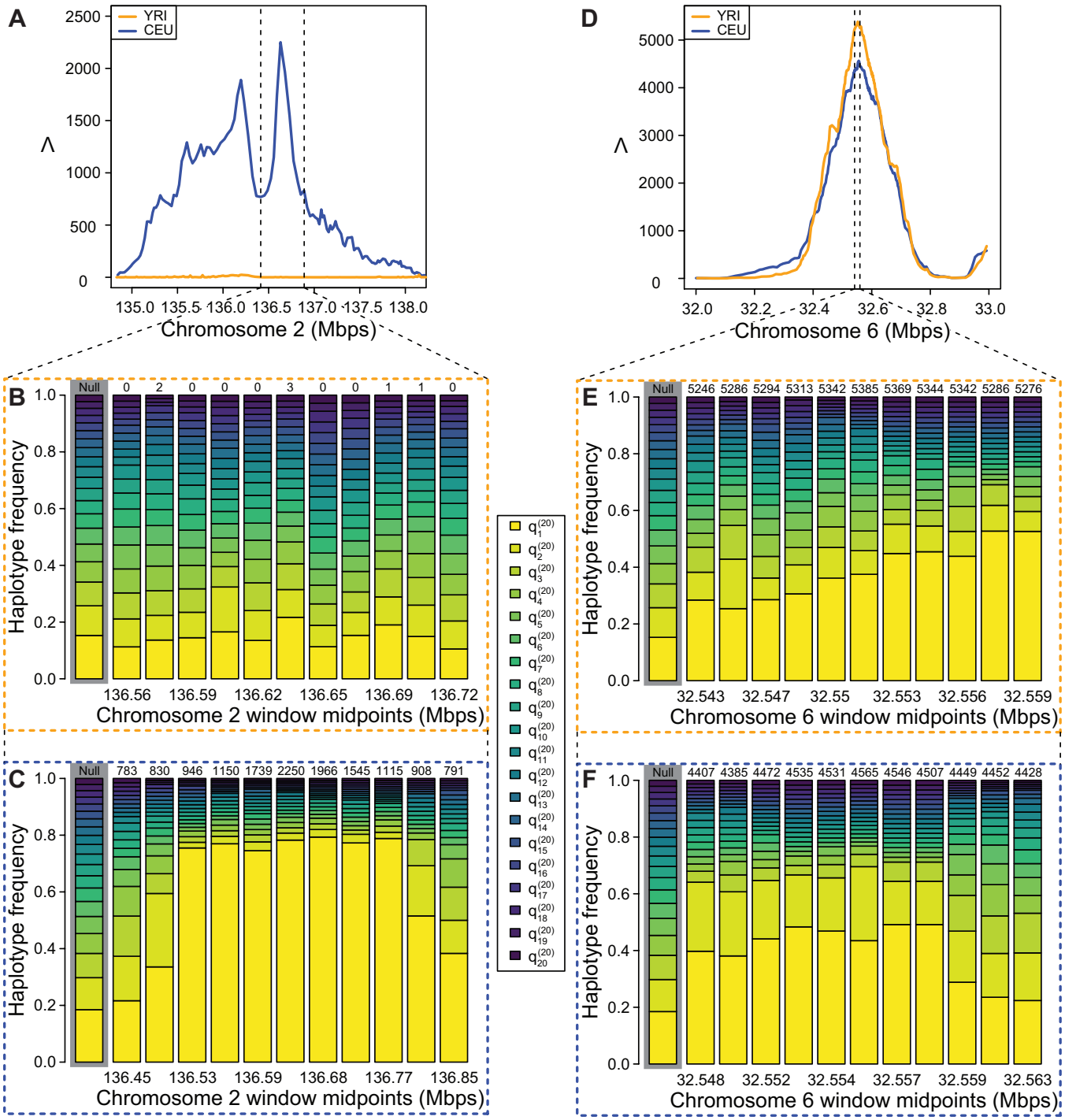

Figure 4: Detailed illustration of Λ statistics and haplotype frequency spectra in CEU and YRI. (A) Λ plotted in the *LCT* region, vertical dotted lines indicate zoomed region shown in (B) and (C). (B) YRI empirical HFS for 11 windows in the *LCT* region. (C) CEU empirical HFS for 11 windows in the *LCT* region. (D) Λ plotted in the MHC region, vertical dotted lines indicate zoomed region shown in (E) and (F). (E) YRI empirical HFS for 11 windows in the MHC region. (F) CEU empirical HFS for 11 windows in the MHC region. In (B), (C), (E), and (F), numbers above HFS are Λ values for the window rounded to the nearest whole number, and the genome-wide average HFS is highlighted in grey. q_i^{20} is the frequency of the ith most common haplotype truncated to $K = 20$.

Supplementary material

Population	Phased	Top-1% Λ	Top-0.1% Λ	Max Λ
CEU	Yes	148.817	318.419	1025.690
CEU	No	66.933	142.547	488.995
YRI	Yes	60.457	134.969	594.606
YRI	No	26.925	54.072	220.665

Table S1: Λ statistic thresholds as calculated from demography-matched neutral simulations.

Chromosome	Start (bp)	Stop (bp)	$\widehat{m}$	$\log_{10}(\widehat{A})$	Max Λ	Genes
2	136,052,012	136,271,543	1	8.686	677.751	*ZRANB3*
2	136,494,985	136,788,904	1	8.686	890.887	*UBXN4*, *LCT*, *LOC100507600*, *MCM6*, *DARS*, *DARS-AS1*
6	32,500,928	32,666,396	7	7.817	699.735	*HLA-DRB6*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DQB1-AS1*

Table S2: Regions of extreme Λ values (unphased analysis) in the CEU population and the genes contained therein. $\widehat{m}$ is the inferred number of sweeping haplotypes, and $\log_{10}(\widehat{A})$ is the estimated sweep width.

Chromosome	Start (bp)	Stop (bp)	$\widehat{m}$	$\log_{10}(\widehat{A})$	Max Λ	Genes
3	46,105,118	46,308,665	6	8.252	348.033	*CCR1*, *CCR3*
3	162,501,324	162,524,925	8	8.252	221.017	–
3	162,524,926	162,661,235	8	8.252	296.913	–
6	31,191,709	31,385,508	9	7.817	352.419	*HLA-C*, *HLA-B*, *MIR6891*, *MICA*
6	32,545,459	32,571,159	9	7.817	248.927	*HLA-DRB1*
6	32,593,356	32,605,370	10	7.817	240.087	*HLA-DQA1*
6	32,995,608	33,173,606	7	7.817	324.768	*HLA-DPA1*, *HLA-DPB1*, *HLA-DPB2*, *COL11A2*, *RXRB*, *SLC39A7*, *HSD17B8*
6	130,583,306	130,635,803	7	7.817	244.991	*SAMD3*
8	50,064,801	50,172,664	8	7.817	275.713	–
10	102,081,939	102,106,184	6	8.252	232.238	*PKD2L1*
10	102,173,525	102,287,281	6	8.252	327.657	*WNT8B*, *SEC31B*, *NDUFB8*
12	79,589,655	79,779,306	6	8.252	334.766	*SYT1*
15	55,139,290	55,335,782	7	8.252	302.699	–
17	3,545,315	3,651,534	6	8.252	294.719	*CTNS*, *TAX1BP3*, *P2RX5-TAX1BP3*, *EMC6*, *P2RX5*, *ITGAE*, *GSG2*

Table S3: Regions of extreme Λ values (unphased analysis) in the YRI population and the genes contained therein. $\widehat{m}$ is the inferred number of sweeping haplotypes, and $\log_{10}(\widehat{A})$ is the estimated sweep width.

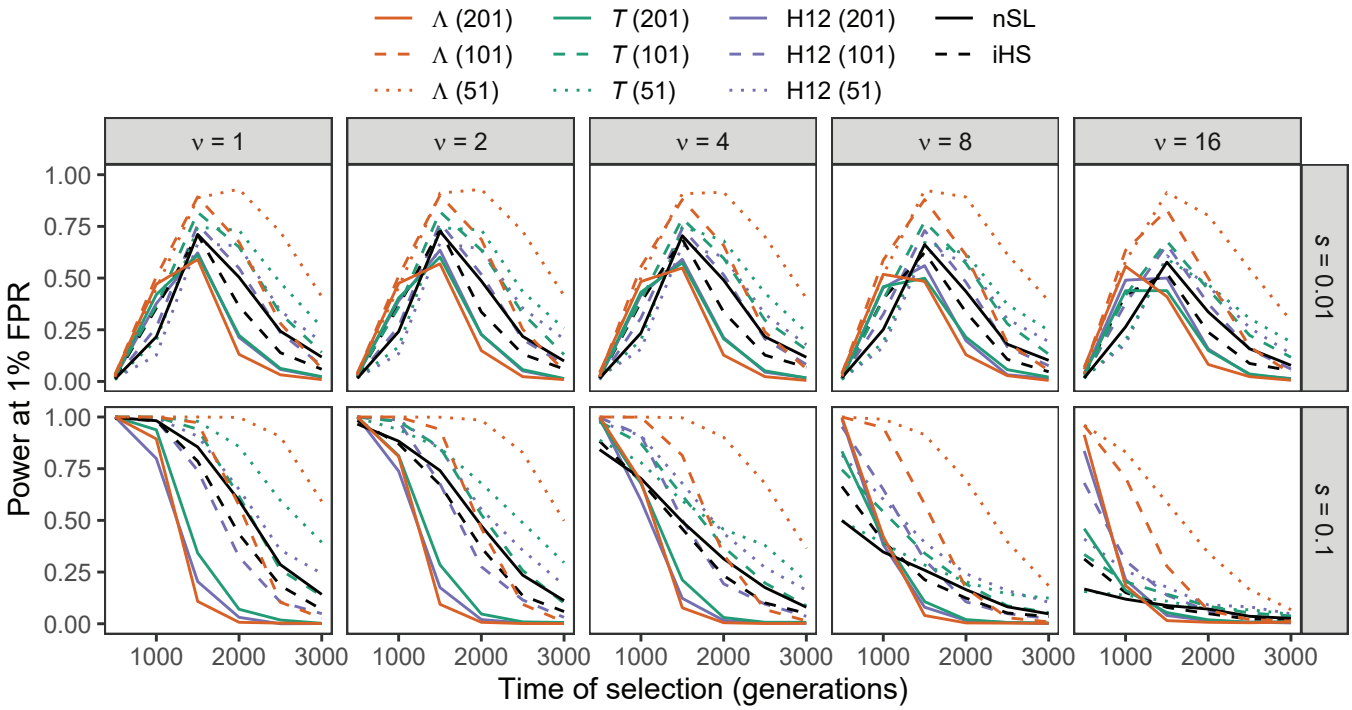

Figure S1: Power at a 1% false positive rate (FPR) as a function of selection start time for applications of Λ, T, and H12 with windows of size 51, 101, and 201 SNPs, as well nS_L and iHS under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Classification ability demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns).

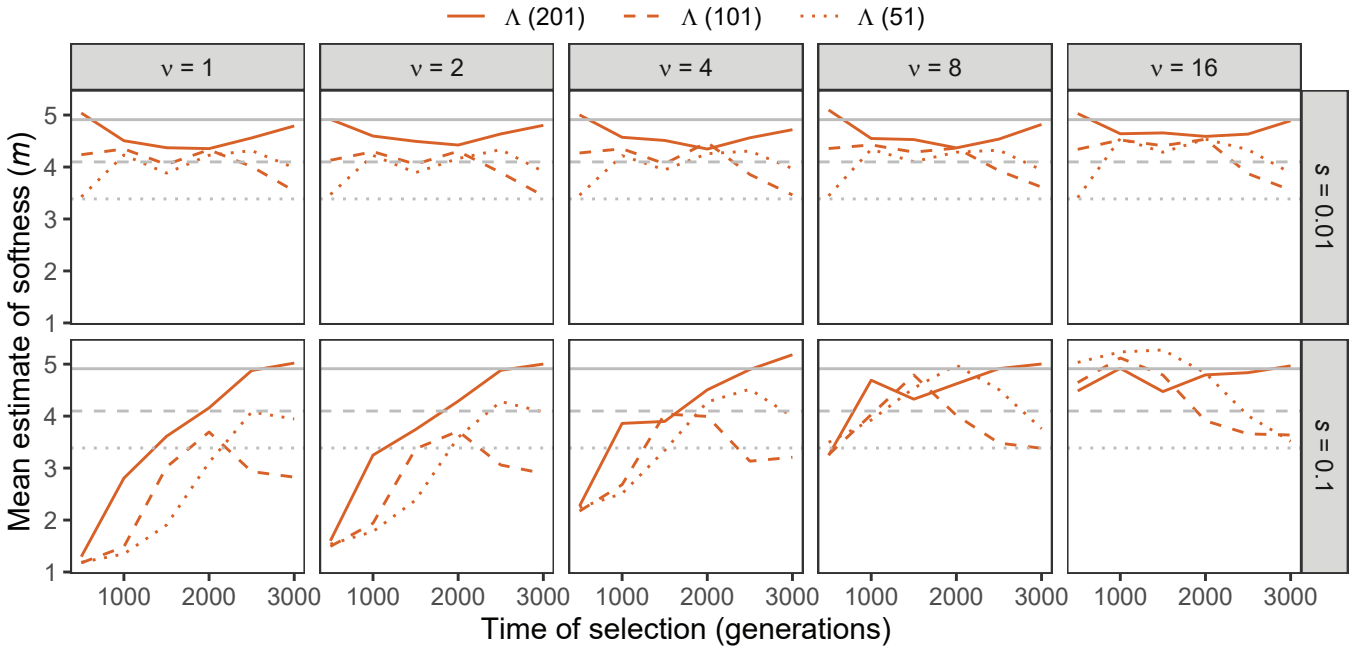

Figure S2: Estimated sweep softness illustrated by mean estimated number of sweeping haplotypes ($\widehat{m}$) in Λ with windows of size 51, 101, and 201 SNPs under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated softness demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\widehat{m}$ values for Λ applied to neutral simulations.

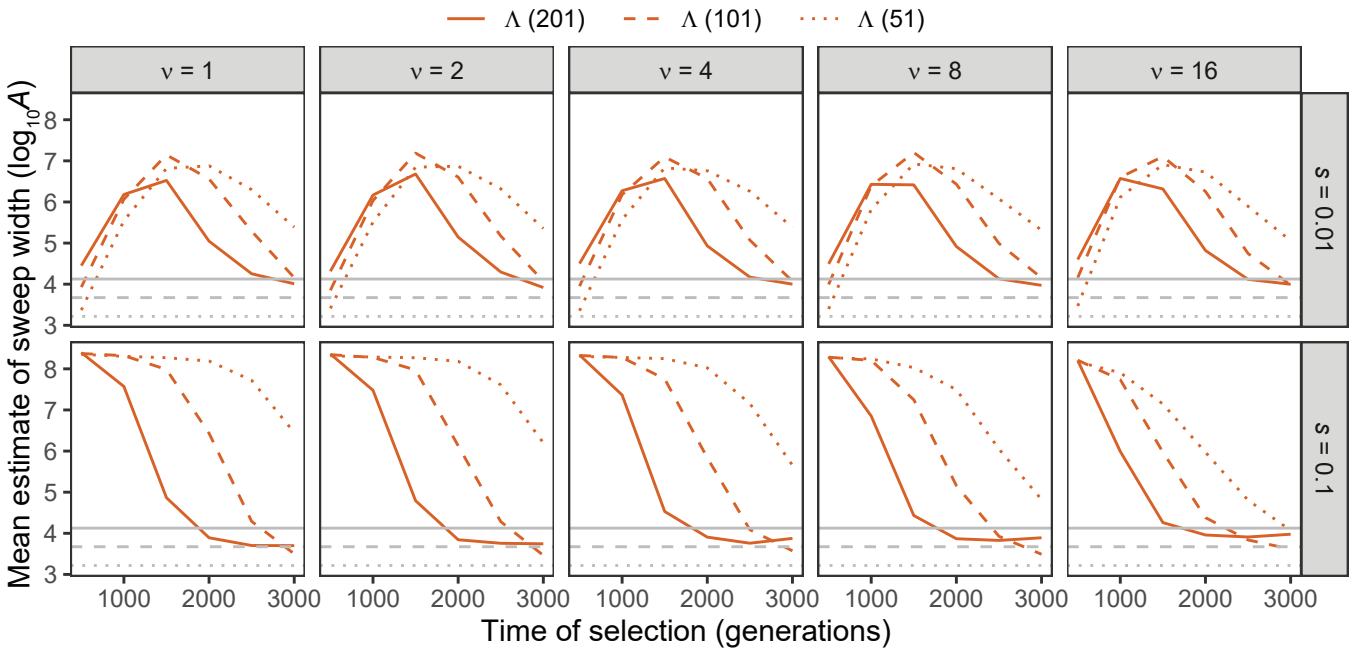

Figure S3: Estimated sweep width illustrated by mean estimated genomic size influenced by the sweep ($\log_{10}\widehat{A}$) in Λ with windows of size 51, 101, and 201 SNPs under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated genomic size influenced by sweeps demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\log_{10}\widehat{A}$ values for Λ applied to neutral simulations.

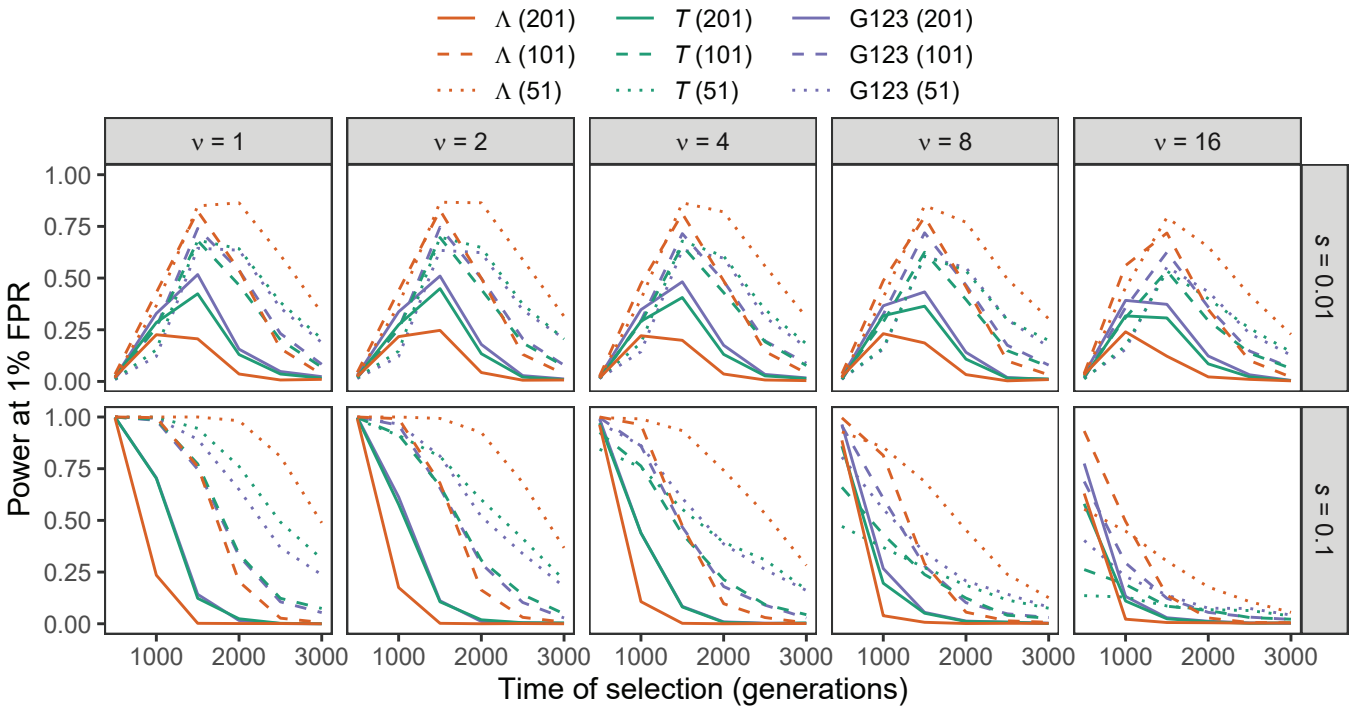

Figure S4: Power at a 1% false positive rate (FPR) as a function of selection start time for applications of Λ, T, and G123 with windows of size 51, 101, and 201 SNPs to unphased multilocus genotype input data under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Classification ability demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns).

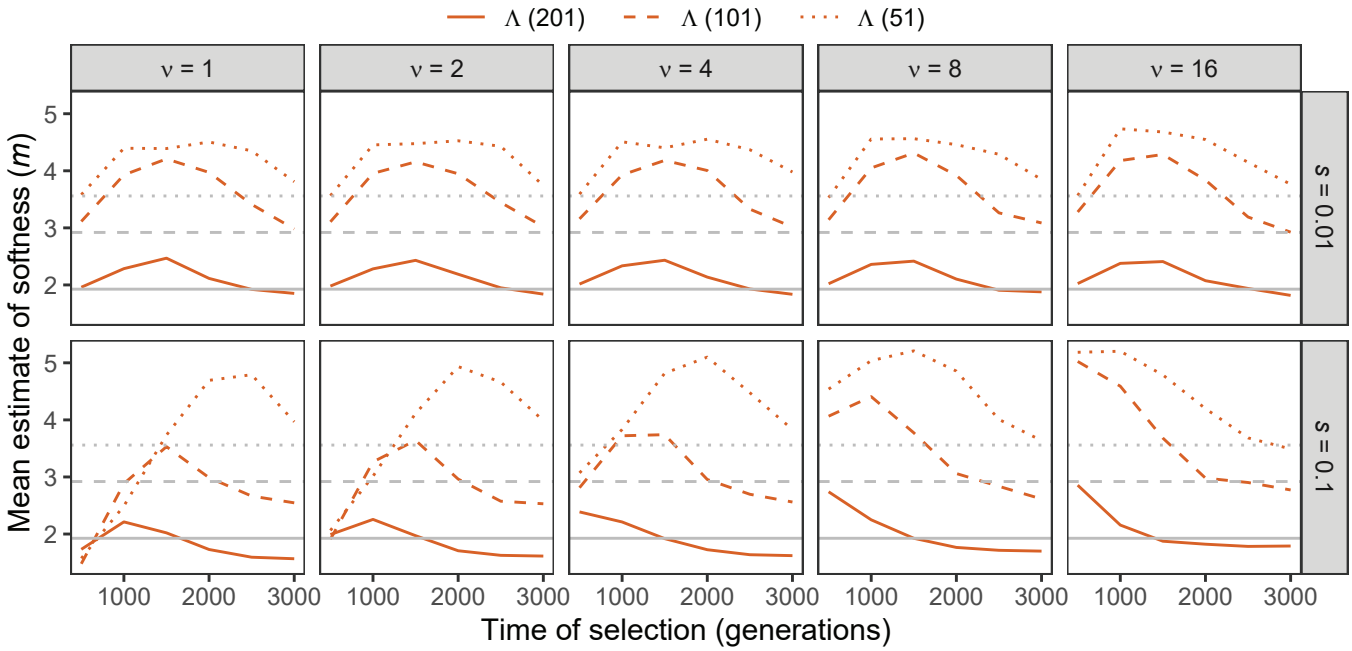

Figure S5: Estimated sweep softness illustrated by mean estimated number of sweeping haplotypes ($\widehat{m}$) in Λ with windows of size 51, 101, and 201 SNPs applied to unphased multilocus input data under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated softness demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\widehat{m}$ values for Λ applied to neutral simulations.

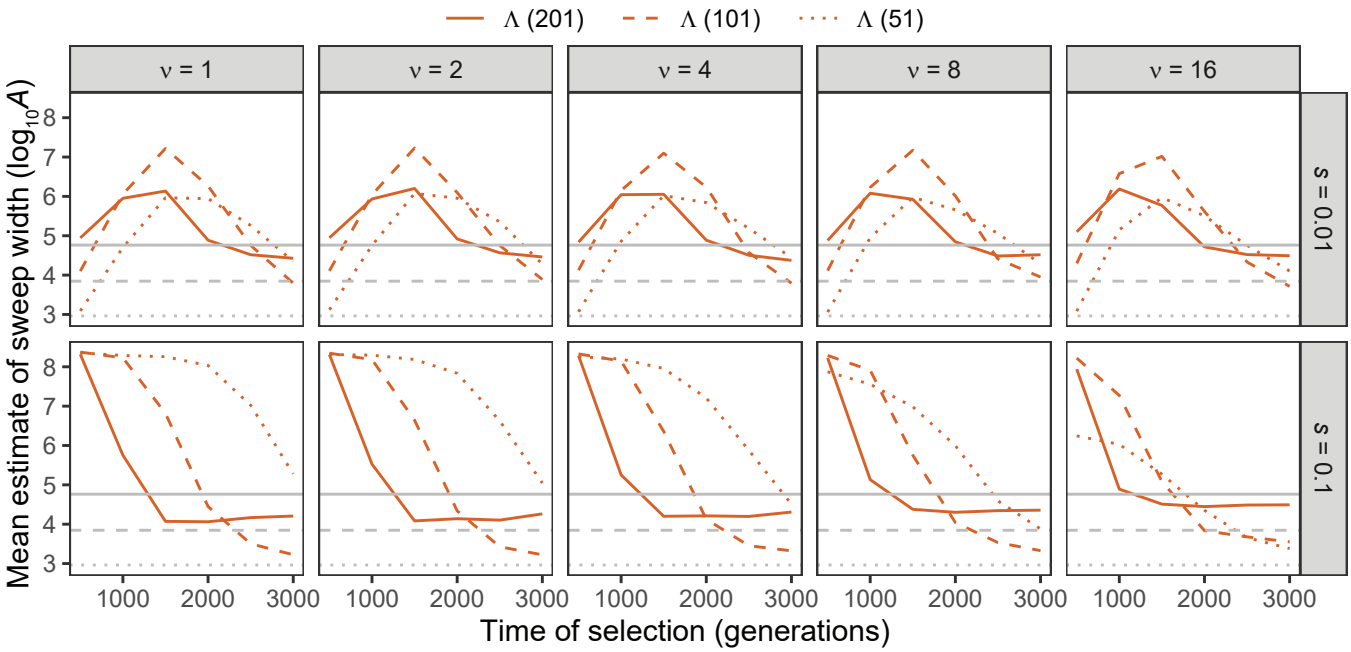

Figure S6: Estimated sweep width illustrated by mean estimated genomic size influenced by the sweep ($\log_{10}\widehat{A}$) in Λ with windows of size 51, 101, and 201 SNPs applied to unphased multilocus input data under simulations of a constant-size demographic history for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated genomic size influenced by sweeps demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\log_{10}\widehat{A}$ values for Λ applied to neutral simulations.

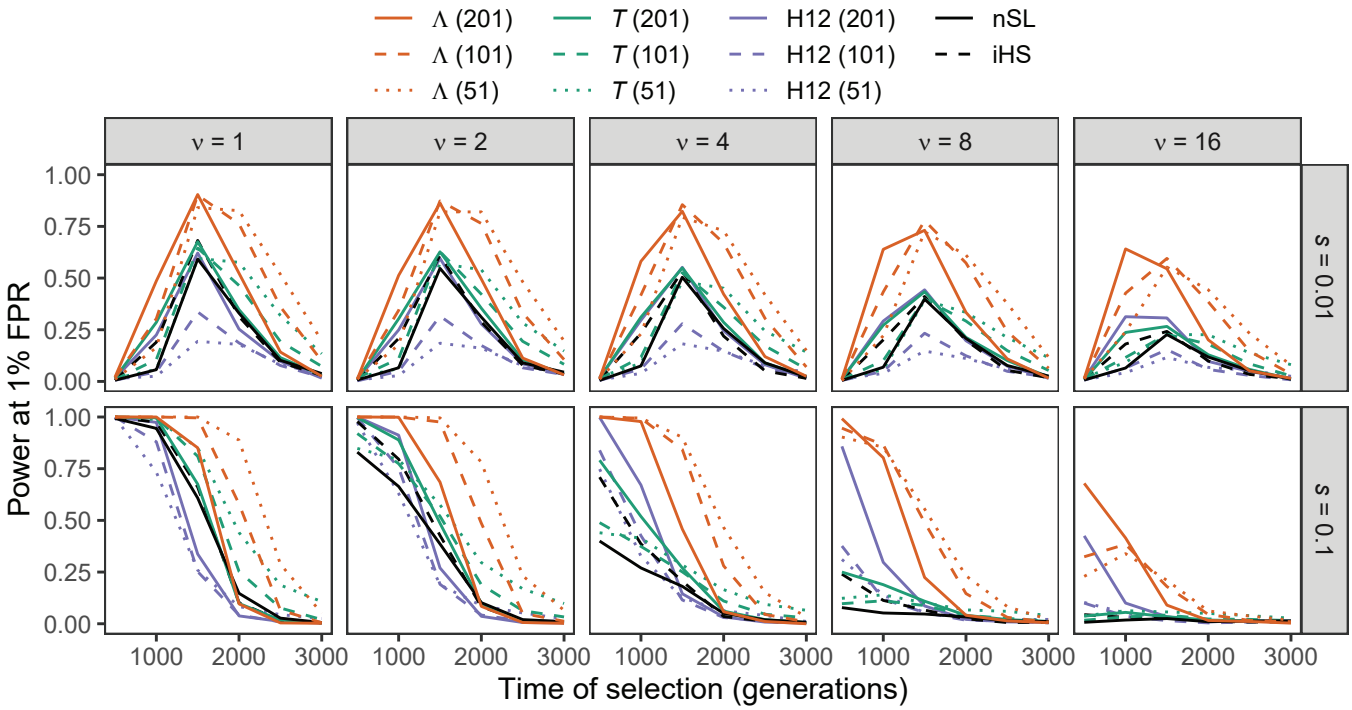

Figure S7: Power at a 1% false positive rate (FPR) as a function of selection start time for applications of Λ, T, and H12 with windows of size 51, 101, and 201 SNPs, as well nS_L and iHS under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Classification ability demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns).

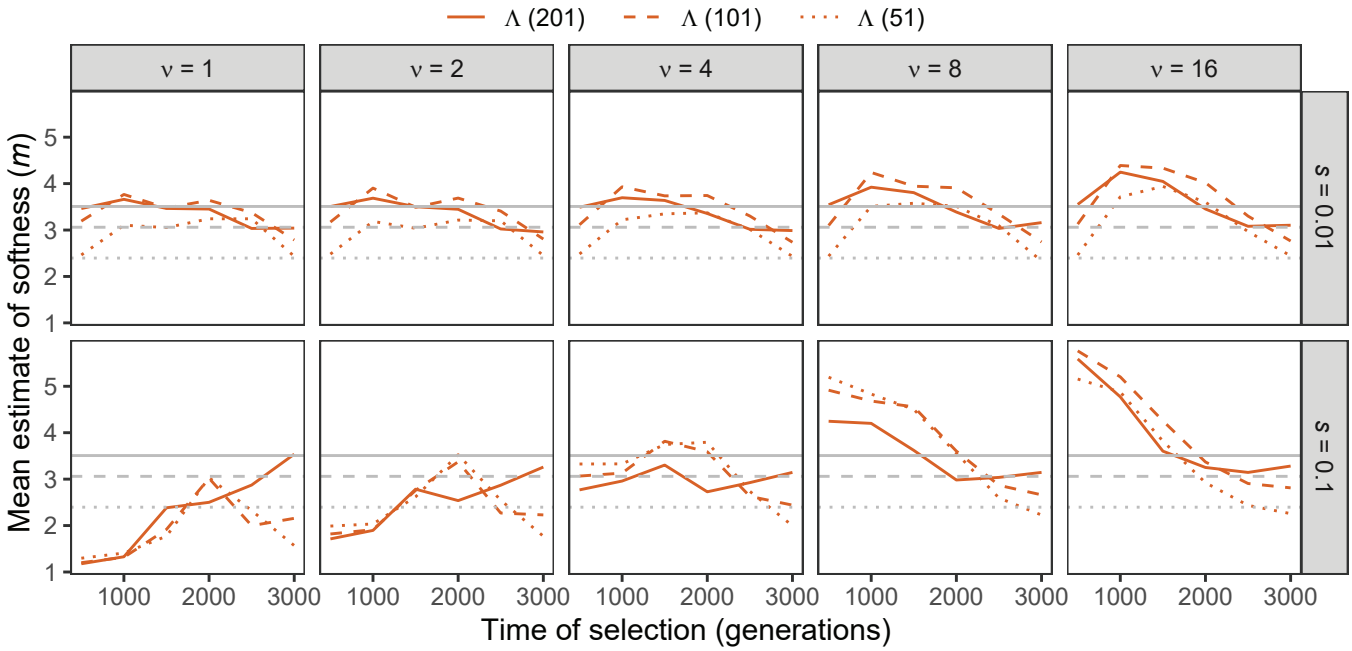

Figure S8: Estimated sweep softness illustrated by mean estimated number of sweeping haplotypes ($\widehat{m}$) in Λ with windows of size 51, 101, and 201 SNPs under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated softness demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\widehat{m}$ values for Λ applied to neutral simulations.

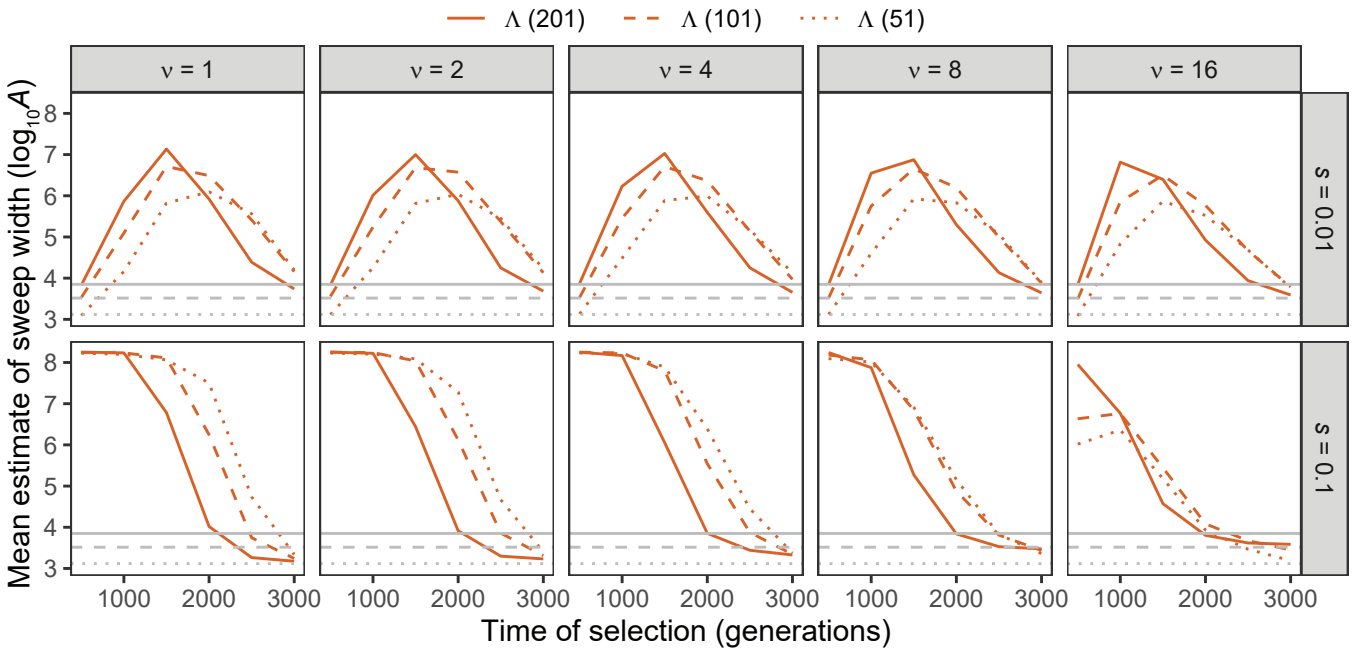

Figure S9: Estimated sweep width illustrated by mean estimated genomic size influenced by the sweep ($\log_{10}\widehat{A}$) in Λ with windows of size 51, 101, and 201 SNPs under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated genomic size influenced by sweeps demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\log_{10}\widehat{A}$ values for Λ applied to neutral simulations.

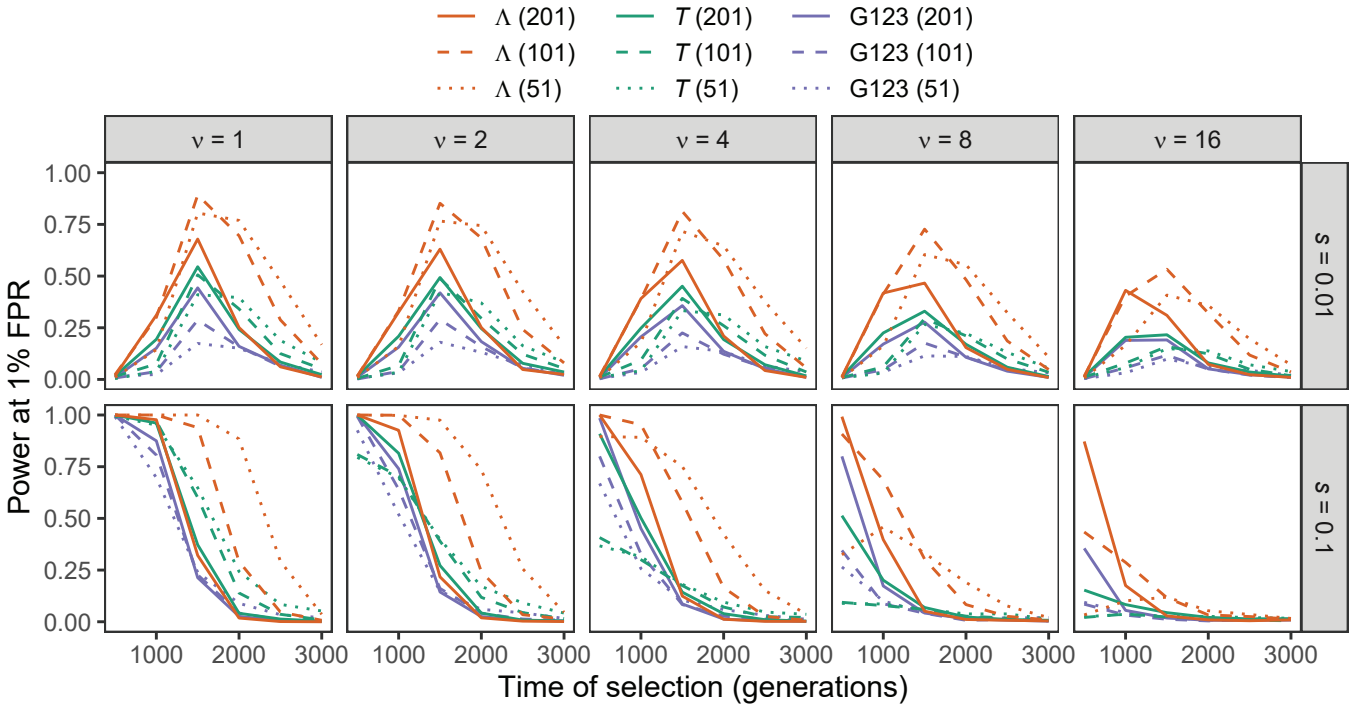

Figure S10: Power at a 1% false positive rate (FPR) as a function of selection start time for applications of Λ, T, and G123 with windows of size 51, 101, and 201 SNPs to unphased multilocus genotype input data under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Classification ability demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns).

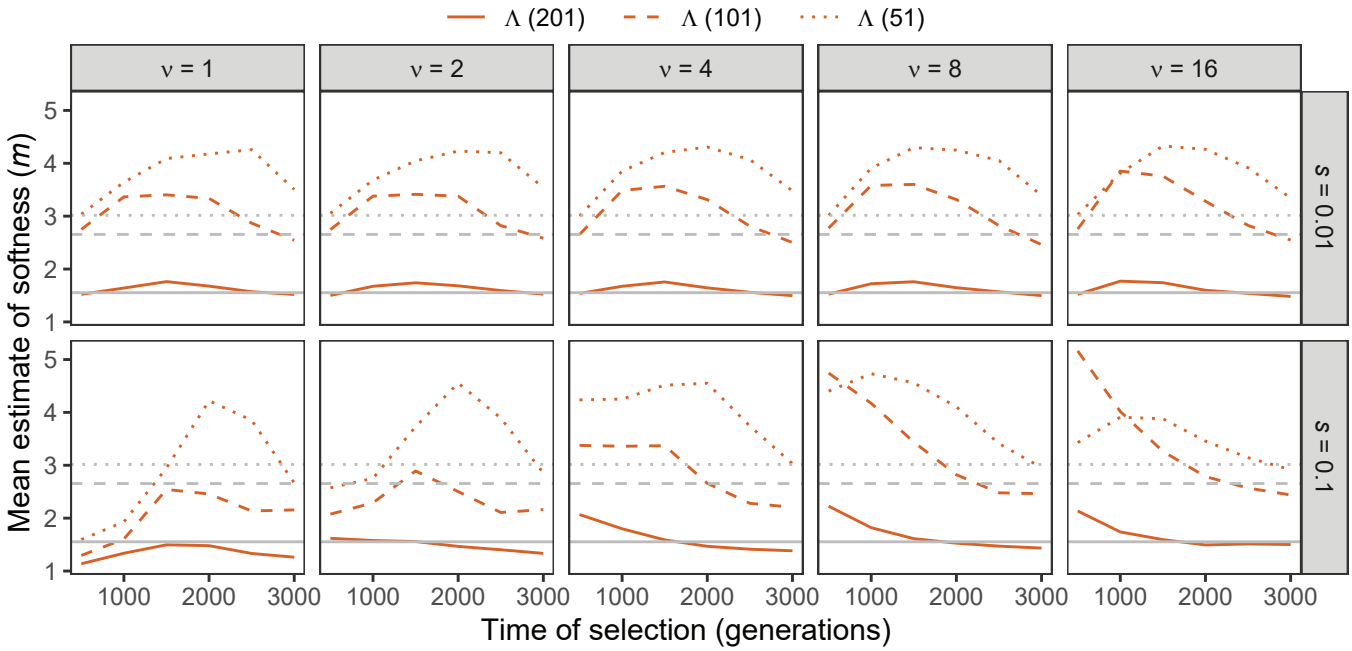

Figure S11: Estimated sweep softness illustrated by mean estimated number of sweeping haplotypes ($\widehat{m}$) in Λ with windows of size 51, 101, and 201 SNPs applied to unphased multilocus input data under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated softness demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\widehat{m}$ values for Λ applied to neutral simulations.

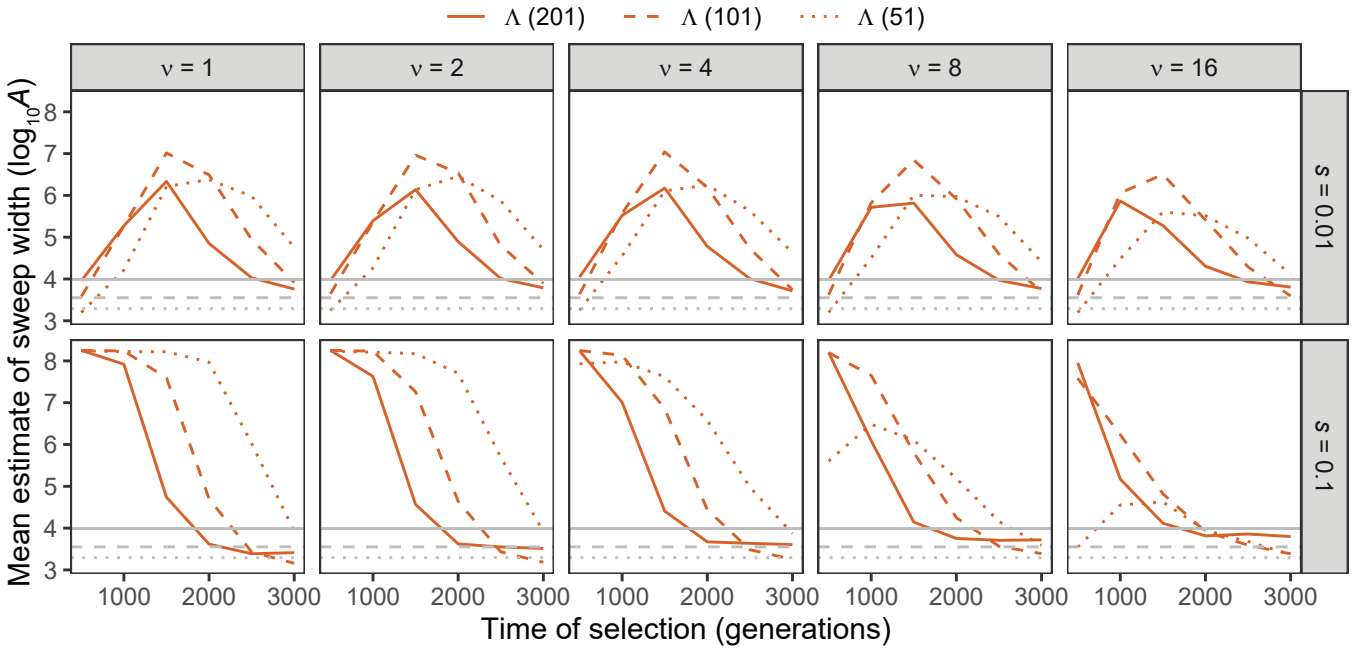

Figure S12: Estimated sweep width illustrated by mean estimated genomic size influenced by the sweep ($\log_{10}\widehat{A}$) in Λ with windows of size 51, 101, and 201 SNPs applied to unphased multilocus input data under simulations of the human central European (CEU) demographic history of Terhorst et al. [2017] for per-generation selection coefficients of $s \in \{0.01, 0.1\}$ on the rows. Mean estimated genomic size influenced by sweeps demonstrated for selection start times of $t \in \{500, 1000, 1500, 2000, 2500, 3000\}$ generations prior to sampling for $\nu \in \{1, 2, 4, 8, 16\}$ initially-selected haplotypes (columns). Gray solid, dashed, and dotted horizontal lines are the corresponding mean $\log_{10}\widehat{A}$ values for Λ applied to neutral simulations.

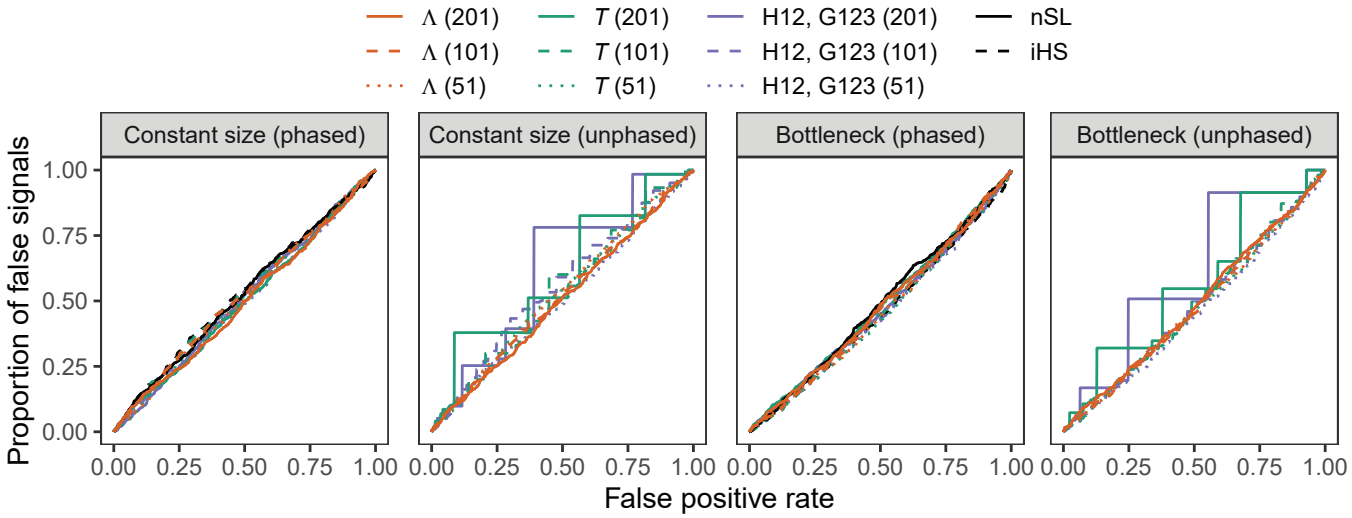

Figure S13: Proportion of false signals as a function of false positive rate for applications of Λ, T, H12, and G123 with windows of size 51, 101, and 201 SNPs, as well nS_L and iHS under simulations of a constant-size demographic history and the human central European (CEU) demographic history of Terhorst et al. [2017] (bottleneck scenario) under background selection using either phased haplotype input data (Λ, T, H12, nS_L, and iHS) or unphased multilocus genotype input data (Λ, T, and G123). Proportion of false signals is computed as the fraction of background selection simulations in which the score computed for Λ, T, H12, G123, nS_L, or iHS exceeded the corresponding score threshold defined by a particular false positive rate.

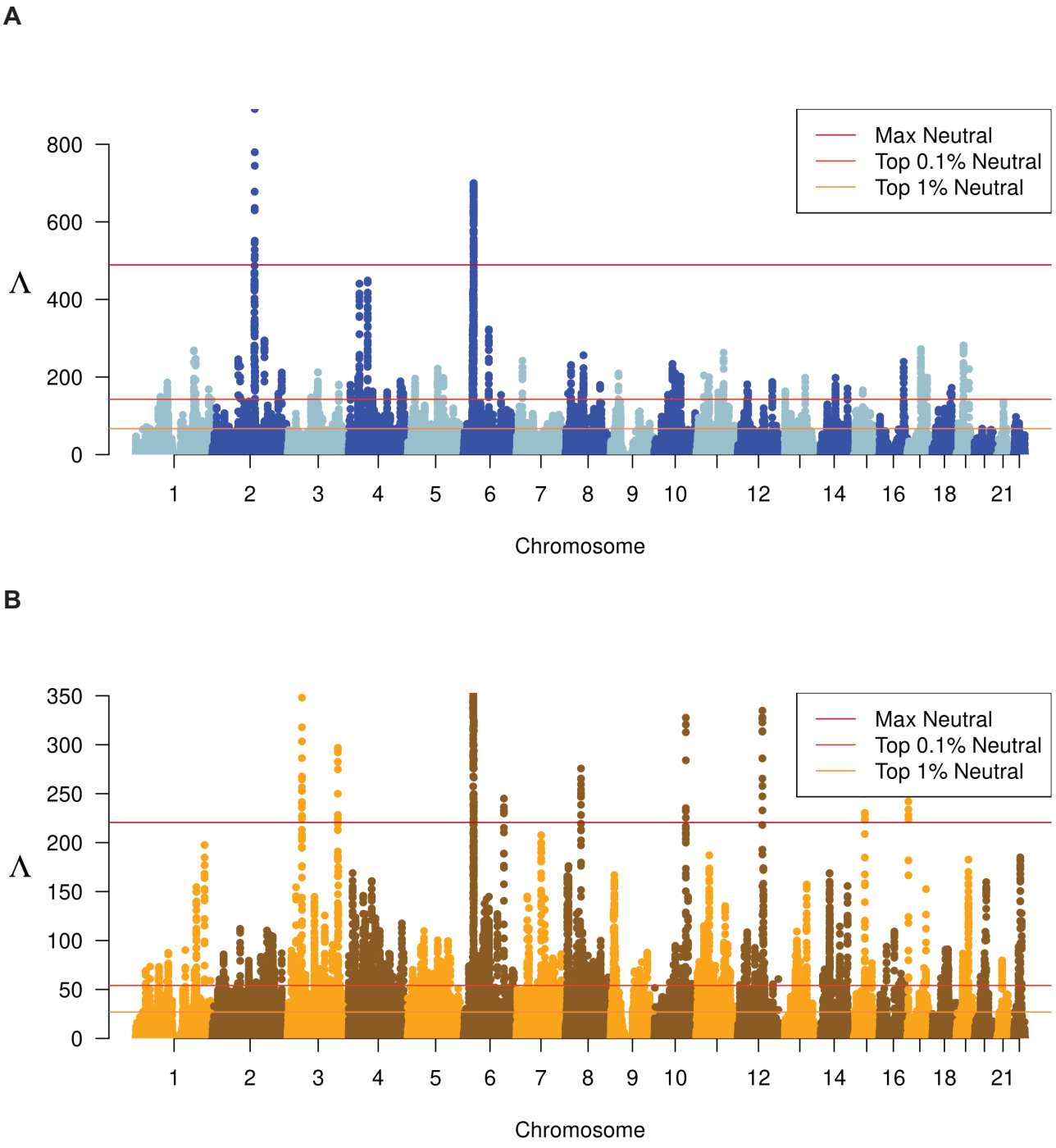

Figure S14: Manhattan plot of unphased multi-locus genotype Λ-statistics for the (A) CEU and (B) YRI populations from the 1000 Genomes Project. Each point represents a single 201-SNP window along the genome. Horizontal lines represent the top 1%, top 0.1%, and maximum observed Λ statistic in demography-matched neutral simulations.

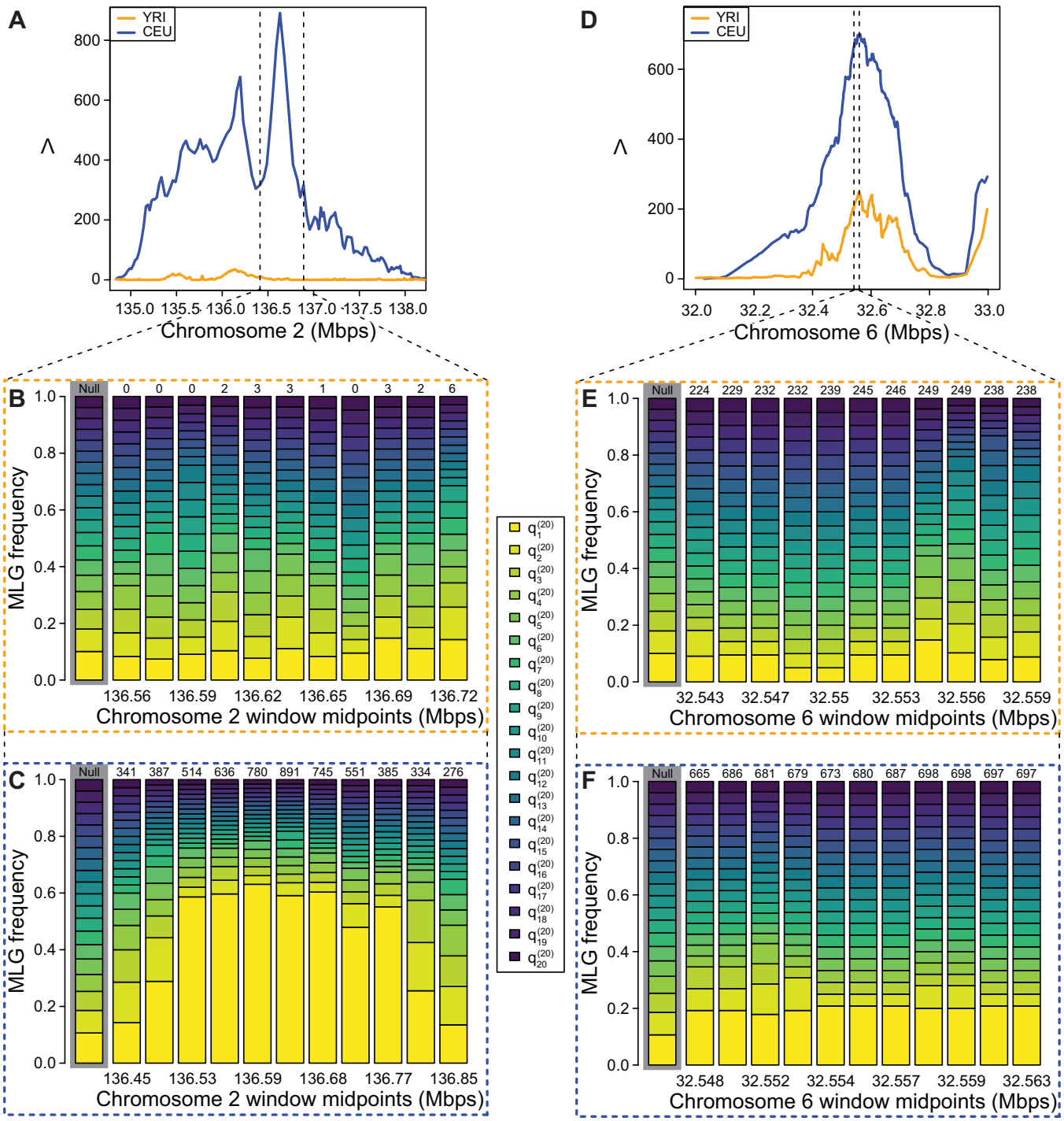

Figure S15: Detailed illustration of Λ statistics and multi-locus genotype frequency spectra in CEU and YRI. (A) Λ plotted in the *LCT* region, vertical dotted lines indicate zoomed region shown in (B) and (C). (B) YRI empirical HFS for 11 windows in the *LCT* region. (C) CEU empirical HFS for 11 windows in the *LCT* region. (D) Λ plotted in the MHC region, vertical dotted lines indicate zoomed region shown in (E) and (F). (E) YRI empirical HFS for 11 windows in the MHC region. (F) CEU empirical HFS for 11 windows in the MHC region. In (B), (C), (E), and (F), numbers above HFS are Λ values for the window rounded to the nearest whole number, and the genome-wide average HFS is highlighted in grey. q_i^{20} is the frequency of the ith most common MLG truncated to $K = 20$.