Article: Methods

Identifying and classifying shared selective sweeps from multilocus data

Alexandre M. Harris[1,2] and Michael DeGiorgio[1,3,4,*]

October 17, 2018

[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 Statistics, Pennsylvania State University, University Park, PA 16802, USA*

[4] *Institute for CyberScience, Pennsylvania State University, University Park, PA 16802,USA*

[*] *Corresponding author:* mxd60@psu.edu

Keywords: Expected haplotype homozygosity, multilocus genotype, ancestral sweep, convergent sweep

Running title: Detecting shared sweeps

Abstract

Positive selection causes beneficial alleles to rise to high frequency, resulting in a selective sweep of the diversity surrounding the selected sites. Accordingly, the signature of a selective sweep in an ancestral population may still remain in its descendants. Identifying genomic regions under selection in the ancestor is important to contextualize the timing of a sweep, but few methods exist for this purpose. To uncover genomic regions under shared positive selection across populations, we apply the theory of the expected haplotype homozygosity statistic H12, which detects recent hard and soft sweeps from the presence of high-frequency haplotypes. Our statistic, SS-H12, is distinct from other statistics that detect shared sweeps because it requires a minimum of only two populations, and properly identifies independent convergent sweeps and true ancestral sweeps, with high power. Furthermore, we can apply SS-H12 in conjunction with the ratio of a different set of expected haplotype homozygosity statistics to further classify identified shared sweeps as hard or soft. Finally, we identified both previously-reported and novel shared sweep candidates from whole-genome sequences of global human populations. Previously-reported candidates include the well-characterized ancestral sweeps at *LCT* and *SLC24A5* in Indo-European populations, as well as *GPHN* worldwide. Novel candidates include an ancestral sweep at *RGS18* in sub-Saharan African populations involved in regulating the platelet response and implicated in sudden cardiac death, and a convergent sweep at *C2CD5* between European and East Asian populations that may explain their different insulin responses.

Introduction

Alleles under positive selection increase in frequency in a population toward fixation, causing nearby linked neutral variants to also rise to high frequency. This process results in selective sweeps of the diversity surrounding the sites of selection, which can be hard or soft [Hermisson and Pennings, 2005, Pennings and Hermisson, 2006a,b, Hermisson and Pennings, 2017]. Under hard sweeps, beneficial alleles exist on a single haplotypic background at the time of selection, and a single haplotype rises to high frequency with the selected variants. In contrast, soft sweeps occur when beneficial alleles are present on multiple haplotypic backgrounds, each of which increases in frequency with the selected variants. Thus, individuals carrying the selected alleles do not all share a common haplotypic background. The signature of a selective sweep, hard or soft, is characterized by elevated levels of linkage disequilibrium (LD) and expected haplotype homozygosity [Maynard Smith and Haigh, 1974, Sabeti et al., 2002, Schweinsberg and Durrett, 2005]. Thus, the signature of a selective sweep decays with distance from the selected site as mutation and recombination erode tracts of sequence identity produced by the sweep, returning expected haplotype homozygosity and LD to their neutral levels [Messer and Petrov, 2013].

Various approaches exist to detect signatures of selective sweeps in single populations, but few methods can identify sweep regions shared across populations, and these methods primarily rely on allele frequency data as input. Identifying sweeps common to multiple populations provides an important layer of context that specifies the branch of a genealogy on which a sweep is likely to have occurred. Existing methods to identify shared sweeps [Bonhomme et al., 2010, Fariello et al., 2013, Racimo, 2016, Librado et al., 2017, Peyrégne et al., 2017, Cheng et al., 2017, Johnson and Voight, 2018] leverage the observation that study populations sharing similar patterns of genetic diversity at a putative site under selection descend from a common ancestor in which the sweep occurred. Such approaches therefore infer a sweep ancestral to the study populations from what may be coincidental (*i.e.*, independent) signals. Moreover, many of these methods require data from at least one reference population in addition to the study populations, and of these, most may be misled by sweeps in their set of reference populations. These constraints may therefore impede the application of these methods to study systems that do not fit their model assumptions or data requirements.

Here, we develop an expected haplotype homozygosity-based statistic, denoted SS-H12, that addresses the aforementioned constraints. SS-H12 detects shared selective sweeps from a minimum of two sampled populations using haplotype data (Figure 1), and classifies sweep candidates as either ancestral (shared through common ancestry) or convergent (occurring independently). We base our approach on the theory of the H12 statistic [Garud et al., 2015, Garud and Rosenberg, 2015], which applies a measure of expected homozygosity to haplotype data from a single population. This single-population statistic has high power to detect recent selective sweeps, and identifies hard and soft sweeps with similar power due to its unique formulation. For a genomic window in which there are I distinct haplotypes, H12 [Garud et al., 2015] is defined as

$$\mathrm{H12} = (p_1 + p_2)^2 + \sum_{i=3}^{I} p_i^2, \tag{1}$$

where p_i is the frequency of the ith most frequent haplotype, with $p_1 \geq p_2 \geq \cdots \geq p_I$. The two most common haplotype frequencies are pooled into a single value to reflect the presence of at least two high-frequency haplotypes in the population under a soft sweep. This pooling yields similar values of H12 for hard and soft sweeps. In addition, the framework of the single-population statistic distinguishes hard and soft sweeps using the H2/H1 ratio [Garud et al., 2015, Garud and Rosenberg, 2015], where $\mathrm{H1} = \sum_{i=1}^{I} p_i^2$ is the expected haplotype homozygosity, and where $\mathrm{H2} = \mathrm{H1} - p_1^2$ is the expected haplotype homozygosity omitting the most frequent haplotype. The value of H2/H1 is small under hard sweeps, because the second through Ith frequencies are small, as the beneficial alleles exist only on a single haplotypic background. Accordingly, H2/H1 is larger for soft sweeps [Garud et al., 2015], and can therefore be used to classify sweeps as hard or soft, conditioning on an elevated value of H12.

We now define SS-H12, which provides information about the location of a shared sweep on the sample population phylogeny, and describe its application to a sample consisting of individuals from two populations. The SS-H12 approach measures the diversity within each population, as well as within the pool of the populations, and is a modification of the single-population expected homozygosity-based approach. Consider a pooled sample consisting of haplotypes from two populations, in which a fraction γ of the haplotypes derives from population 1 and a fraction $1-\gamma$ derives from population 2. For a pooled sample consisting of individuals from two populations, we define the total-sample expected haplotype homozygosity statistic $\mathrm{H12}_{\mathrm{Tot}}$ within a genomic window containing I distinct haplotypes as

$$\mathrm{H12}_{\mathrm{Tot}} = (x_1 + x_2)^2 + \sum_{i=3}^{I} x_i^2, \tag{2}$$

where $x_i = \gamma p_{1i} + (1-\gamma) p_{2i}$, $x_1 \geq x_2 \geq \cdots \geq x_I$, is the frequency of the ith most frequent haplotype in the pooled population, and where p_{1i} and p_{2i} are the frequencies of this haplotype in populations 1 and 2, respectively. $\mathrm{H12}_{\mathrm{Tot}}$ therefore has elevated values at the sites of shared sweeps because the pooled genetic diversity at a site under selection in each sampled population remains small.

Next, we seek to define a statistic that classifies the putative shared sweep as ancestral or convergent between the pair of populations. To do this, we define a statistic $f_{\mathrm{Diff}} = \sum_{i=1}^{I} (p_{1i} - p_{2i})^2$, which measures the sum of the squared difference in the frequency of each haplotype between both populations. f_{Diff} takes on values between 0, for population pairs with identical haplotype frequencies, and 2, for populations that are each fixed for a different haplotype. The former case is consistent with an ancestral sweep scenario, whereas the latter is consistent with a convergent sweep.

From the summaries of H12_{Tot} and f_{Diff}, we now formulate SS-H12, which measures the extent to which an elevated H12_{Tot} is due to shared ancestry. First, we specify a statistic that quantifies the shared sweep, $\text{H12}_{\text{Anc}} = \text{H12}_{\text{Tot}} - f_{\text{Diff}}$. The value of H12_{Anc} lies between -1 for convergent sweeps, and 1 for ancestral sweeps, with a negative value near 0 in the absence of a sweep. H12_{Anc} is therefore easy to interpret because convergent sweeps on non-identical haplotypes cannot generate positive values, and ancestral sweep signals that have not eroded due to the effects of recombination and mutation cannot generate negative values. Because a sufficiently strong and complete sweep in one population (divergent sweep; Figure 1) may also yield negative values of H12_{Anc} with elevated magnitudes distinct from neutrality, we introduce a correction factor that yields SS-H12 by dividing the minimum value of H12 between a pair of populations by the maximum value. This modification allows SS-H12 to overlook spurious signals driven by strong selection in a single population by reducing their prominence relative to true shared sweep signals. Applying this correction factor yields SS-H12, which is computed as

$$\text{SS-H12} = \text{H12}_{\text{Anc}} \times \frac{\min[\text{H12}^{(1)}, \text{H12}^{(2)}]}{\max[\text{H12}^{(1)}, \text{H12}^{(2)}]}, \tag{3}$$

where $\text{H12}^{(1)}$ and $\text{H12}^{(2)}$ are the H12 values for populations 1 and 2, respectively. The correction factor has a value close to 1 for shared sweeps of either type, but a small value for divergent sweeps. Thus, the corrected SS-H12 is sensitive only to shared sweeps.

Using simulated genetic data, we show that SS-H12 has high power to detect recent shared selective sweeps in pairs of populations, displaying a similar range of detection to the single-population H12 statistic on which it is based. Additionally, we demonstrate that, in accordance with our expectations, SS-H12 correctly identifies recent ancestral sweeps from elevated positive values, and convergent sweeps from negative values of large magnitude, generally without confusing the two scenarios. Furthermore, we extended the application of SS-H12 to an arbitrary number of populations K (see *Materials and Methods*), finding once again that our approach classifies sweeps correctly and with high power. Moreover, the SS-H12 approach retains the ability to distinguish between ancestral and convergent hard and soft sweeps from the inferred number of distinct sweeping haplotypes, with each sweep type occupying a distinct subset of paired ($|\text{SS-H12}|$, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values, where $\text{H1}_{\text{Tot}} = \sum_{i=1}^{I} x_i^2$ and $\text{H2}_{\text{Tot}} = \text{H1}_{\text{Tot}} - x_1^2$. Finally, our analysis of whole-genome sequences from global human populations recovered previously-identified sweep candidates at the *LCT* and *SLC24A5* genes in Indo-European populations, corroborated recently-characterized sweeps that emerged from genomic scans with the single-population approach [Harris et al., 2018], such as *RGS18* in African and *P4HA1* in

Indo-European populations, and uncovered novel shared sweep candidates, such as the convergent sweeps *C2CD5* between Eurasian populations and *PAWR* between European and sub-Saharan African populations.

Results

We evaluated the ability of SS-H12 to differentiate among the simulated scenarios of shared selective sweeps, sweeps unique to only one sampled population, and neutrality, using the signature of expected haplotype homozygosity in samples consisting of individuals from two or more populations. We performed simulations using SLiM 2 [Haller and Messer, 2017] under human-inspired parameters [Takahata et al., 1995, Nachman and Crowell, 2000, Payseur and Nachman, 2000] for populations of constant diploid size (N) subject to changing selection start times (t) and strengths (s), across differing split times (τ) between sampled populations. Additionally, we evaluated the robustness of SS-H12 to confounding scenarios of population admixture and background selection. We then used an approximate Bayesian computation (ABC) approach in the same manner as Harris et al. [2018] to demonstrate our ability to distinguish between shared hard and soft sweeps in samples from multiple populations. Finally, we show that SS-H12 recovers previously-hypothesized signatures of shared sweeps in human whole-genome sequences [Auton et al., 2015], while also uncovering novel candidates. See *Materials and Methods* for further explanation of experiments.

Detection of ancestral and convergent sweeps with SS-H12

We conducted experiments to examine the ability of SS-H12 to not only identify shared sweep events among two or more sampled populations (K), but categorize them as shared due to common ancestry, or due to convergent evolution. Across all scenarios, we scanned 100 kb simulated chromosomes using a 40 kb sliding window with a step size of one kb. We selected this window size because it is over this interval that neutral pairwise LD, measured with r^2, decays to 1/3 of the value for loci one kb apart (Figure S1). For each sweep scenario, we studied the power at 1% and 5% false positive rates (FPRs) for detecting shared selective sweeps (Figures 2, 3, and S2-S7) as a function of time at which beneficial alleles arose, under scenarios of ancestral, convergent, and divergent sweeps.

First, we simulated scenarios in which an ancestral population split into $K = 2$ descendant populations $\tau = 1000$ generations prior to sampling, with strong ($s = 0.1$) hard sweeps occurring between 200 and 4000 generations prior to sampling (Figure 2). We chose a model of $\tau = 1000$ as an estimate of the approximate split time of Eurasian human populations [Gravel et al., 2011, Gronau et al., 2011, Schiffels and Durbin, 2014]. This series of experiments illustrates the range of sweep start times over which SS-H12 can detect sweeps. SS-H12 has high power for recent strong shared sweeps starting between 200 and 1500 generations prior to sampling, with power completely attenuated for shared sweeps older than 2000 generations (Figure 2).

Here, power is greater overall for convergent sweeps than for ancestral sweeps (Figure 2). As expected, the distributions of SS-H12 values for convergent sweeps center on negative values (Figure 2), whereas the SS-H12 distributions of ancestral sweeps center on positive values (Figure 2). The difference in power between these scenarios is primarily because the convergent sweeps are more recent, but also because we compute power from the distribution of maximum |SS-H12| values for each sweep scenario. This means that the magnitude of SS-H12 for replicates of shared sweeps must exceed the magnitude under neutrality, which for any combination of t and s is more likely for convergent than ancestral sweeps.

To further characterize the performance of SS-H12 for hard sweeps, we repeated experiments on simulated samples from $K = 2$ populations for more anciently-diverged populations (larger τ), and weaker sweeps (smaller s). Under strong sweeps but a more ancient $\tau = 2000$, approximately corresponding to lower estimates for the split time between African and non-African humans [Gravel et al., 2011, Gronau et al., 2011, Schiffels and Durbin, 2014], SS-H12 maintains power to distinguish convergent sweeps from neutrality (Figure 3), but the signal of ancestral sweeps is too attenuated to be detectable by the time of sampling (Figure 3). Reducing the selection coefficient to $s = 0.01$ for both the $\tau = 1000$ and 2000 scenarios had the effect of increasing the power of SS-H12 for more ancient selection times t, which allowed for the detection of ancestral sweeps starting between 1500 and 2100 generations before sampling (Figures S2 and S3), while still maintaining power for convergent sweeps more ancient than $t = 1000$ generations before sampling (Figures S2 and S3).

Because the single-population statistic H12 has power to detect both hard and soft sweeps, we next performed analogous experiments for simulated soft sweep scenarios. Maintaining values of t, τ, and s identical to those for hard sweep experiments, we simulated soft sweeps as selection on standing genetic variation for $\nu = 4$ and $\nu = 8$ distinct sweeping haplotypes (Figures S4-S7). We found that trends in the power of SS-H12 to detect shared soft sweeps remained consistent with those for hard sweeps. However, the power of SS-H12 for detecting soft sweeps was attenuated relative to hard sweeps, proportionally to the number of sweeping haplotypes, with a larger drop in power for older sweeps and little to no effect on power for more recent sweeps. Our observations therefore align with results for the single-population H12 statistic [Garud et al., 2015, Harris et al., 2018]. Thus, the ability to detect a sweep derives from the combination of s, t, and ν, with stronger, recent sweeps on fewer haplotypes being easiest to detect.

We contrast our results for shared sweeps across population pairs with those for divergent sweeps, which we present in parallel (Figures 2, 3, and S2-S7). Across identical values of t as for each convergent sweep experiment, we found that divergent sweeps, in which only one of the two simulated populations experiences a sweep ($t < \tau$), are not visible to SS-H12 for any combination of simulation parameters. To understand the properties of the divergent sweeps relative to the shared sweeps, we compared the distributions of their

SS-H12 values at peaks identified from the maximum values of |SS-H12| for each replicate. We observed that the distributions of the divergent sweeps remain broadly unchanged from one another under all parameter combinations, and closely resemble the distribution generated under neutrality, as all are centered on negative values with small magnitude, and have small variance. Thus, the use of a correction factor that incorporates the values of H12 from each component population in the sample (see Equation 3) provides an appropriate approach for preventing sweeps that are not shared from appearing as outlying signals.

Extending our analyses to $K = 3$ ($\tau_1 = 1000$ and $\tau_2 = 500$), 4 ($\tau_1 = 1000$, $\tau_2 = 750$, and $\tau_3 = 500$), and 5 ($\tau_1 = 1000$, $\tau_2 = 750$, $\tau_3 = 500$, and $\tau_4 = 250$) sampled populations and strong sweeps ($s = 0.1$) specifically, we found that SS-H12 maintains power to identify shared sweeps, but that the properties of the method change somewhat (Figures S8-S10). To assign values of SS-H12 to samples of $K \geq 3$ populations, we employed two types of approaches, and found that these generally yielded comparable power to detect shared sweeps. First, we measured SS-H12 for each possible population pair within the sample, and conservatively retained the value of smallest magnitude as the overall SS-H12. Second, we computed SS-H12 across the two branches directly subtending the root of the K-population phylogeny underlying the sample, grouping together all populations descending from the same internal branch (see *Materials and methods*).

For both the conservative and grouped approaches, the power of SS-H12 to detect strong shared sweeps is high for sweeps more recent than 1500 generations ago, and rapidly attenuates for more ancient sweeps, and power is once again greatest for convergent sweeps. However, we found that despite maintaining perfect or near-perfect power for convergent sweeps on samples from $K \geq 3$ populations, the distribution of SS-H12 includes many replicates with positive values, which are normally associated with ancestral sweeps. The shift toward positive values increases as the convergent sweep becomes more ancient, reflecting a greater fraction of ancestral sweeps between pairs of sampled populations within the overall convergent sweep. Although the conservative approach remains generally more robust to misclassifying shared sweeps within samples from $K \geq 3$ populations than does the grouped approach, both strategies may fail to identify a convergent sweep as convergent if the sweep time t is close enough to τ_1. Additionally, divergent sweeps yield a distribution of SS-H12 values for samples from $K \geq 3$ populations that may differ from neutrality as t approaches τ_1. Despite this observation, we emphasize that divergent sweeps once again do not produce values of SS-H12 that deviate appreciably from values generated under neutrality, leaving shared sweeps as the sole source of prominently outlying sweep signals in practice.

Addressing confounding scenarios

Because the SS-H12 approach relies on a signal of elevated expected haplotype homozygosity, it may be confounded by non-adaptive processes that alter levels of population-genetic diversity. For this reason, we first examined the effect of admixture on the power of SS-H12 in the context of ancestral, convergent, and divergent strong ($s = 0.1$) sweeps between population pairs separated by $\tau = 1000$ generations, wherein one population (the target) receives gene flow from a diverged, unsampled donor outgroup population (Figures 4 and S11). Admixture occurred as a single pulse 200 generations before sampling, and in the case of the divergent sweep, occurred specifically in the population experiencing the sweep. The donor split from the common ancestor of the two sampled populations (the target and its unadmixed sister) 2×10^4 generations before sampling—within a coalescent unit of the sampled populations, similar to the relationship between Neanderthals and modern humans [Juric et al., 2016, Harris and Nielsen, 2016])—and had an effective size either one-tenth, identical to, or tenfold the size of the target. Although the donor does not experience selection, extensive gene flow from a donor with low genetic diversity may resemble a sweep. Correspondingly, gene flow from a highly diverse donor may obscure sweeps.

As expected, gene flow into the target population distorted the SS-H12 distribution of the two-population sample relative to no admixture (compare Figure 4 to 2), and this distortion was proportional to the level of admixture from the donor, as well as the donor's effective size (Figure 4). Ancestral sweeps were the most likely to be misclassified following admixture from an unsampled donor of small effective size ($N = 10^3$; Figure 4, first row), increasingly resembling convergent sweeps as the rate of gene flow increased (though ultimately with little change in power to detect the shared sweep; Figure S11). The confounding effect of admixture on ancestral sweep inference emerges because low-diversity gene flow into one population yields a differing signal of elevated expected haplotype homozygosity in each population, spuriously resembling a convergent sweep. In contrast, the distributions of SS-H12 values and the power of SS-H12 for convergent and divergent sweeps remained broadly unchanged relative to no admixture (Figure 2) under low-diversity admixture scenarios (compare panels within the first rows of Figures 4 and S11 to Figure 2). Because two populations subject to convergent or divergent sweeps are already extensively differentiated, further differentiation due to admixture does not impact the accuracy of sweep timing inference using SS-H12.

For intermediate donor effective size ($N = 10^4$; Figures 4 and S11, second row), the magnitudes of both the ancestral and convergent sweep signals attenuates toward neutral levels, and the power of SS-H12 wanes as the admixture proportion increases. This is because the genetic diversity in the target population increases to levels resembling neutrality, overall yielding a pattern spuriously resembling a divergent sweep that SS-H12 cannot distinguish from neutrality. Accordingly, the magnitude and power of SS-H12 under a divergent sweep scenario following admixture scarcely change under the $N = 10^4$ scenario. As the effective size of the donor population grows large ($N = 10^5$; Figures 4 and S11, third row), SS-H12 becomes more robust to the effect of admixture for shared sweeps, accurately identifying ancestral and convergent sweeps with high power at greater admixture proportions relative to the $N = 10^4$ scenario. However, the power of SS-H12 spuriously rises to 1.0 for divergent sweeps under the $N = 10^5$ admixture scenario. Both the increased robustness to admixture for the ancestral and convergent sweeps, as well as the elevated power for divergent sweeps, result from a reduction in the magnitude of SS-H12 under neutrality for the $N = 10^5$ admixture scenario relative to $N = 10^4$, which does not occur for the sweep scenarios. That is, the magnitude of a sweep signature remains similar across the $N = 10^5$ and $N = 10^4$ admixture scenarios, while the magnitude of the neutral background is smaller, meaning that any sweep, even a divergent sweep, is more prominent for larger donor population sizes. We further address this observation in the *Discussion*.

We also observed the effect of long-term background selection on the neutral distribution of SS-H12 values (Figure S12). Background selection may yield signatures of genetic diversity resembling selective sweeps [Charlesworth et al., 1993, 1995, Seger et al., 2010, Nicolaisen and Desai, 2013, Cutter and Payseur, 2013, Huber et al., 2016], though previous work suggests that background selection does not drive particular haplotypes to high frequency [Enard et al., 2014, Harris et al., 2018]. Our two background selection scenarios for samples from $K = 2$ populations, with $\tau = 1000$ and 2000, were performed as described in the *Materials and Methods*, following the protocol of Cheng et al. [2017]. Briefly, we simulated a 100-kb sequence featuring a centrally-located 11-kb gene consisting of exons, introns, and untranslated regions, across which deleterious variants arose randomly throughout the entire simulation period. In agreement with our expectations, we found that background selection is unlikely to confound inferences from SS-H12, yielding only marginally larger values of |SS-H12| than does neutrality (Figure S12). Accordingly, SS-H12 does not classify background selection appreciably differently from neutrality.

Classifying shared sweeps as hard or soft from the number of sweeping haplotypes

Because the primary innovation of the single-population approach is its ability to classify sweeps as hard or soft from paired (H12, H2/H1) values, we evaluated the corresponding properties of our current approach for samples consisting of $K = 2$ populations (Figure 5). Here, we color a space of paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values, each bounded by $[0.005, 0.995]$, according to the inferred most probable number of sweeping haplotypes ν for each point in the space. Similarly to the approach of Harris et al. [2018], we inferred the most probable ν using an approximate Bayesian computation (ABC) approach in which we determined the posterior distribution of ν from 5×10^6 replicates of sweep scenarios with $\nu \in \{1, 2, \ldots, 16\}$ and $s \in [0.005, 0.5]$, both drawn uniformly at random for each replicate (the latter drawn from a log-scale). We were able to classify recent shared sweeps as hard or soft, but found our current approach to have somewhat different properties to the single-population approach (Figure 5).

For ancestral sweep scenarios and $\tau = 1000$ ($t \in [1020, 2000]$), the pattern of paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values generally followed that of single-population analyses [Harris et al., 2018], though with irregularities among larger values of |SS-H12| paired with intermediate values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ (Figure 5, top left). For a given value of |SS-H12|, smaller values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ were generally more probable for ancestral sweeps from smaller ν, and inferred ν increased with $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$. This fit our expectations, because as the number of ancestrally sweeping haplotypes in the pooled population increases, the value of H2_{Tot} increases relative to H1_{Tot}. Additionally, ancestral sweeps from larger ν (softer sweeps) are unlikely to generate large values of |SS-H12| or small values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$, and the most elevated values of |SS-H12| were rarely associated with more than four sweeping haplotypes. We note, however, the presence of paired values inferred to derive from $\nu = 1$ for some intermediate-to-large values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$, as well as the presence of points with inferred $\nu \geq 4$ at smaller $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$. This may indicate that among ancestral sweep replicates for $\tau = 1000$, weaker hard sweep signals may be difficult to resolve from stronger soft sweep signals, as both should yield intermediate levels of haplotypic diversity. Under simulated scenarios of $\tau = 2000$ ($t \in [2020, 3000]$; Figure 5, top right), the resolution of our approach greatly diminished. Here, most replicates were assigned to $\nu = 1$, with the largest values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$, as well as some larger values of |SS-H12|, inferred to be from soft sweeps.

The convergent sweep experiments yielded a distinctly different occupancy of possible paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values relative to ancestral sweeps, and provided a greater resolution among the tested values of ν, showing little irregularity in the assignment of ν (Figure 5, bottom row). In addition, trends in the occupancy of hard and soft sweeps were generally concordant between replicates for $\tau = 1000$ ($t \in [200, 980]$) and $\tau = 2000$ ($t \in [200, 1980]$). For these experiments, we simulated simultaneous independent sweeps, either both soft or both hard, allowing each population to follow a unique but comparable trajectory. Thus, there were always at least two sweeping haplotypes in the pooled population. Accordingly, convergent hard sweeps, unlike ancestral hard sweeps, are primarily associated with large values of |SS-H12| and intermediate values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$. Furthermore, strong convergent sweeps of any sort could not generate small $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ values unless |SS-H12| was also small. Even so, convergent sweeps from larger ν occupy a distinct set of paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values that is shifted either toward smaller |SS-H12|, larger $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$, or both, demonstrating that the accurate inference of ν is possible for convergent sweeps. Unlike for ancestral sweeps or single-population analyses, we observed a small subset of convergent soft sweeps, particularly from $\nu = 2$, that were able to consistently generate smaller values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ paired with smaller values of |SS-H12|. These observations represent convergent sweep replicates in which identical haplotypes were selected between subpopulations. Accordingly, convergent hard sweeps occupied the smallest paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values, which fits the expectation that hard sweeps yield smaller $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ than soft sweeps when all other parameters are identical.

Application of SS-H12 to human genetic data

We applied SS-H12 to whole-genome sequencing data from global human populations in phase 3 of the 1000 Genomes Project [Auton et al., 2015], which is ideal as input because it contains large sample sizes and no missing genotypes at polymorphic sites. We searched for shared sweep signals within the RNA- and protein-coding genes of geographically proximate and distant human population pairs, performing various comparisons of European, South Asian, East Asian, and Sub-Saharan African populations (Tables S2-S10). For the top 40 outlying candidate shared sweeps among population pairs, we assigned p-values from a neutral distribution of 10^6 replicates following human demographic models inferred from `smc++` (see *Materials and Methods*). Our Bonferroni-corrected significance threshold [Neyman and Pearson, 1928] was 2.10659×10^{-6} (see Table S1 for critical values associated with each population pair). We additionally inferred the maximum posterior estimates on $\nu \in \{1, 2, \ldots, 16\}$ for each candidate from a distribution of 5×10^6 simulated convergent or ancestral sweep replicates, depending on our classification of the candidate from the sign of SS-H12, following the same `smc++`-derived models. We categorized sweeps from $\nu = 1$ as hard, and sweeps from $\nu \geq 2$ as soft.

Across all comparisons, we found that ancestral hard sweeps comprised the majority of prominent candidates at RNA- and protein-coding genes, regardless of population pair. Many of these candidate ancestral sweeps were detected with H12 in single populations [Harris et al., 2018], including novel sweeps at *RGS18* in the sub-Saharan African pair of YRI and LWK ($p < 10^{-6}$, $\nu = 1$; previously identified in YRI; see Figure 6) and at *P4HA1* between the European CEU and South Asian GIH populations ($p < 10^{-6}$, $\nu = 1$; previously identified in GIH, though as a soft sweep). We also observed a dearth of high-magnitude negative values in Tables S2-S10, with prominent convergent sweep candidates only occurring between the most diverged population pairs. These consisted of *C2CD5* between CEU and the East Asian JPT population ($\nu = 1$), *PAWR* between Indo-European populations CEU and GIH with the sub-Saharan African YRI population (significant for the GIH-YRI comparison, $p = 2 \times 10^{-6}$, $\nu = 1$ for both comparisons; Table S7), and *MPHOSPH9* and *EXOC6B* between JPT and YRI (both significant with $p = 10^{-6}$ and $\nu = 1$). These observations reflect the broader pattern that negative SS-H12 values are rare between closely-related populations. Indeed, the majority of SS-H12 values at protein-coding genes between populations from the same geographic region are positive, and this distribution shifts toward negative values for more differentiated population pairs, consisting primarily of intermediate-magnitude negative values between the YRI and non-African populations (Figure S13). Our present results are also consistent with the H12-based observations of Harris et al. [2018] in single populations, in that we found a greater proportion of hard sweeps than soft sweeps among outlying sweep candidates in humans, though both were present between all population pairs.

Our top shared sweep candidates also comprised genes that have been described in greater detail in the literature [Bersaglieri et al., 2004, Sabeti et al., 2007, Gerbault et al., 2009, Liu et al., 2013], including *LCT* and the surrounding cluster of genes on chromosome 2 including *MCM6*, *DARS*, and *R3HDM1* in the European CEU-GBR pair ($\nu = 1$ for all; Table S2), reflecting selection for the lactase persistence phenotype. We also recovered the sweep on the light skin pigmentation phenotype in Indo-Europeans [Sabeti et al., 2007, Coop et al., 2009, Mallick et al., 2013, Liu et al., 2013] for comparisons between the CEU population with GBR (Table S2; near-significant with $p = 4 \times 10^{-6}$ and $\nu = 1$) and GIH (Table S3; $p < 10^{-6}$, $\nu = 1$). Although the selected allele for this sweep is thought to lie within the *SLC24A5* gene encoding a solute carrier [Lamason et al., 2005], the mappability and alignability filter that we applied to our data removed *SLC24A5*, but preserved the adjacent *SLC12A1*, which we use as a proxy for the expected signal. Finally, we find *KIAA0825* as a top candidate across comparisons between the CEU and GIH (Table S3; $p = 2 \times 10^{-6}$, $\nu = 1$), YRI and CEU (Table S5; $p = 3 \times 10^{-6}$, $\nu = 1$), YRI and LWK (Table S6; $p < 10^{-6}$, $\nu = 1$), JPT and YRI (Table S8; $p < 10^{-6}$, $\nu = 1$), and GIH and YRI (Table S7; $p < 10^{-6}$, $\nu = 1$) populations. Although the function of *KIAA0825* has not yet been characterized, it is a previously-reported sweep candidate ancestral to the split of African and non-African human populations [Racimo, 2016].

Across all population comparisons, the top shared sweep candidates at RNA- and protein-coding genes comprised both hard and soft sweeps, yielding a wide range of $\mathrm{H2}_{\mathrm{Tot}}/\mathrm{H1}_{\mathrm{Tot}}$ values. This emphasizes the multitude of sweep histories that have shaped shared variation among human populations. In Figure 6, we highlight four distinct results that capture the diversity of sweeps we encountered in our analysis. We first examine *GPHN*, which we found as an outlying candidate shared soft sweep in the East Asian JPT and KHV populations ($\nu = 2$; Table S10). *GPHN* encodes the scaffold protein gephyrin, which has been the subject of extensive study due to its central role in regulating the function of neurons, among the many other diverse functions of its splice variants [Ramming et al., 2000, Lencz et al., 2007, Tyagarajan and Fritschy, 2014]. *GPHN* has received attention as the candidate of a recent selective sweep ancestral to the human out-of-Africa migration event [Voight et al., 2006, Williamson et al., 2007, Park, 2012], which has resulted in the maintenance of two high-frequency haplotypes worldwide [Climer et al., 2015]. Although not meeting the genome-wide significance threshold, we see that a large signal peak is centered over *GPHN*, and the underlying haplotype structure shows two high-frequency haplotypes at similar frequency in the pooled population and in the individual populations (Figure 6, top row).

Next, we recovered *RGS18* as a significant novel outlying ancestral sweep signal in the sub-Saharan African LWK and YRI populations. *RGS18* occurs as a significant sweep in the YRI population [Harris et al., 2018] and correspondingly displays a single shared high-frequency haplotype between the LWK and YRI populations (Figure 6, second row), matching our assignment of this locus as a hard sweep ($\nu = 1$).

RGS18 has been implicated in the development of hypertrophic cardiomyopathy, a leading cause of sudden cardiac death in American athletes of African descent [Maron et al., 2003, Chang et al., 2007]. Between the CEU and YRI populations, we found another novel shared sweep at *SPRED3* (Figure 6, third row ; $p = 2 \times 10^{-6}$, $\nu = 1$), which encodes a protein that suppresses cell signaling in response to growth factors [Kato et al., 2003]. Although elevated levels of observed homozygosity at this gene have previously been reported in European and sub-Saharan African populations separately [Granka et al., 2012, Ayub et al., 2013], these observations have not previously been tied to one another. While both the CEU and YRI populations share their most frequent haplotype in an ancestral sweep, we found that this haplotype is at much lower frequency in the CEU than in the YRI, potentially indicating differences in the strength and duration of selection between these populations at *SPRED3* after their split in the out-of-Africa event.

Finally, we present the novel convergent hard sweep candidate that we uncovered at *C2CD5* (also known as *CDP138*) between the CEU and JPT populations. As expected of a convergent sweep, the signal peak here is large in magnitude but negative, corresponding to the presence of a different high-frequency haplotype in each population, each of which is also at high frequency in the pooled population (Figure 6, bottom row). The protein product of *C2CD5* is involved in insulin-stimulated glucose transport [Xie et al., 2011, Zhou et al., 2018], and the insulin response is known to differ between European and East Asian populations [Kodama et al., 2013]. Therefore, our discovery of *C2CD5* is in agreement with the results of Kodama et al. [2013], and illustrates the importance of differentiating ancestral and convergent sweeps in understanding the adaptive histories of diverse populations. We also highlight our discovery of *PAWR* as another outlying novel convergent hard sweep candidate with complementary clinical support, for comparisons between GIH and CEU with YRI. The protein product of *PAWR* is involved in promoting cancer cell apoptosis, and is implicated in the development of prostate cancer [Yang et al., 2013]. Because mutations within and adjacent to *PAWR* have been specifically implicated in the development of prostate cancer among individuals of African descent [Bonilla et al., 2011], our identification of a candidate convergent sweep at *PAWR* is consistent with the observation of elevated prostate cancer rates for populations with African ancestry [Kheirandish and Chinegwundoh, 2011, Shenoy et al., 2016].

Discussion

Characterizing the selective sweeps shared between geographically close and disparate populations can provide insights into the adaptive histories of these populations that may be unavailable when analyzing single populations separately. To this end, we extended the H12 framework of Garud et al. [2015] to identify genomic loci affected by selection in samples composed of individuals from two or more populations. Our approach, SS-H12, has high power to detect recent shared selective sweeps from phased haplotypes, and is sensitive to both hard and soft sweeps. SS-H12 can also distinguish hard and soft sweeps from one another in conjunction with the statistic $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$, thus retaining the most important feature of the single-population approach. Furthermore, SS-H12 has the unique ability to distinguish between sweeps that are shared due to common ancestry (ancestral sweeps), and shared due to independent selective events (convergent sweeps). Analysis with the SS-H12 framework therefore provides a thorough characterization of selection candidates, both previously-described and novel, that is unlike that of other methods for detecting shared sweeps.

Because the value of SS-H12 fundamentally derives from a measure of expected haplotype homozygosity, it is tailored toward the detection of recent shared selective sweeps. Accordingly, we found that in our simulated shared sweep experiments, the power of SS-H12 to distinguish sweeps from neutrality was greatest for selective events beginning between 200 and 2000 generations prior to the time of sampling, assuming demographic parameters based on human values (Takahata et al. [1995], Nachman and Crowell [2000], Payseur and Nachman [2000]; Figures 2, 3, and S2-S10). Due to their greater distortion of the haplotype frequency spectrum, stronger sweeps yield larger values of SS-H12 and larger sweep footprints [Gillespie, 2004, Garud et al., 2015, Hermisson and Pennings, 2017], and were indeed easier to detect than weaker sweeps (compare the maximum power of SS-H12 across Figures 2 and 3 to Figures S2 and S3). However, stronger sweeps reach fixation sooner than do weaker sweeps, and their signal therefore begins to erode sooner than that of weaker sweeps. This illustrates the inverse correlation between the strength of sweeps (s) that SS-H12 can detect, and the range of selection start times (t) for which SS-H12 can detect a sweep. Harris et al. [2018] also observed this effect for the single-population approach.

The timing and strength of a shared sweep were important not only for detecting the sweep, but also for classifying it as ancestral or convergent. Barring the rare occurrence of a convergent sweep on the same haplotype in two recently-diverged sister populations (which occurred at least once for each set of tested parameters across Figures 2, 3, and S2-S7), we found that simulated convergent sweeps could reliably be identified from the sign of SS-H12 under scenarios in which SS-H12 has power to detect sweeps shared between population pairs. For ancestral sweeps, however, it was possible for negative values of large magnitude to emerge for weaker sweeps, especially if the time of selection t was close to the split time τ (Figures S2, S3, S6, and S7). Due to the slower rate at which a more weakly selected allele rises to high frequency in a population relative to a strongly selected allele, it is likely that the beneficial allele, and the haplotypic background(s) on which it resides, have not risen to high frequency before the ancestral population splits into the modern sampled populations. Post-split, copies of the beneficial allele present in each of the two descendant populations may follow distinct trajectories, yielding differentiated haplotype frequency spectra between the two populations. Separate mutation and recombination events within each population also contribute to haplotype differentiation between populations. Using a smaller analysis window may therefore increase power to detect sweeps with less prominent footprints, but at the risk of misinterpreting elevated SS-H12 due to short-range LD as a sweep signal.

More generally, the ability of SS-H12 to identify a shared sweep as ancestral or convergent depends upon the underlying phylogeny of the sampled populations. For our simulated strong shared sweeps, the power of SS-H12 was greatest for sweeps occurring within 2000 generations of sampling. If the two sampled populations split 1000 generations before the time of sampling, then SS-H12 can detect a wide range of both ancestral and convergent sweeps (Figures 2 and S4). However, SS-H12 has little power to detect strong ancestral sweeps for $\tau = 2000$ under our simulated parameters because the signal of sweeps more ancient than $t = 2000$ erodes before the time of sampling, leaving convergent sweeps as the majority of outlying shared signals (Figures 3 and S5). For more moderate sweeps, the reverse may occur if the peak in power occurs after τ, with ancestral sweeps more likely to generate detectable outlying signals (Figures S2 and S6). In practice, we found that most outlying shared sweep candidates between pairs of human populations were ancestral (SS-H12 > 0; Tables S2-S10), indicating that despite the power of our approach to detect convergent sweeps, such events may simply be uncommon because beneficial mutations are rare [Orr, 2010], and so the independent establishment of a beneficial mutation at the same locus across multiple populations should be especially rare for all but the most strongly-selected mutations [Haldane, 1927, Kimura, 1962, Wilson et al., 2014].

Consistent with results from the single-population approach [Garud et al., 2015, Harris et al., 2018], SS-H12 has power to detect shared soft sweeps, and can assign these as ancestral or convergent similarly to shared hard sweeps (Figures S3 and S4). Our results for soft sweeps indicate, as expected, that increasing the number of sweeping haplotypes ν decreases the power of SS-H12 to detect shared sweeps. Sweeps from larger ν produce smaller distortions in the haplotype frequency spectrum relative to hard sweeps, yielding smaller values of |SS-H12| that are less likely to be distinct from neutrality. Our soft sweep simulations nonetheless consistently maintained similar distributions of SS-H12 to our hard sweep simulations (Figures 2, 3, S2, and S3), indicating that all haplotypes need not be shared between sampled populations, or at similar frequencies, in order to yield outlying SS-H12 signatures. That is, our simulated population split events represented a random sampling of haplotypes from the ancestor, which did not guarantee identical haplotype frequency spectra between descendant sister populations, or between the descendants and the ancestor, and SS-H12 could still identify the ancestral shared sweep with high power. This matches what we observed in the empirical data, wherein the frequencies of shared haplotypes at multiple top candidates differed considerably between populations. Our results for *SPRED3* between CEU and YRI provide a representative example of this (Figure 6, third row). While the H12 signal at *SPRED3* in the CEU population is not strong, there is enough overlap in its shared haplotype with YRI to identify it as a significant ancestral sweep. Thus, our simulated and empirical results suggest that SS-H12 is robust to deviations in haplotype frequency spectra between populations as long as sufficient haplotype sharing still persists.

Our application of SS-H12 to simulated samples composed of individuals from $K \in \{3, 4, 5\}$ populations (Figures S8-S10) demonstrated that our approach maintains excellent power to detect recent shared sweeps regardless of sample structure, with trends in power matching those we observed for samples deriving from $K = 2$ populations (Figure 2). Across all experiments, power curves were nearly identical to the equivalent $K = 2$ scenario (Figure 2, top row), with high power for strong, hard sweeps starting within $t \in [200, 1500]$ generations prior to sampling. Furthermore, the conservative and grouped approaches provide comparable power to one another, indicating the validity of either strategy. However, both the conservative and grouped approaches are unable to consistently classify convergent sweeps, frequently assigning a positive SS-H12 for convergent sweeps initiated before the most recent split time (*i.e.*, predating τ_{K-1}). In the case of the conservative approach, this is because we select the smallest-magnitude SS-H12 between pairs of populations as the SS-H12 for the whole sample, regardless of sign. Because the magnitude of SS-H12 is smaller for ancestral sweeps than for convergent sweeps occurring at the same t, an ancestral SS-H12 is often assigned if the convergent sweep predates at least one coalescence event between sampled populations, reflecting an internal ancestral sweep within the phylogeny. Accordingly, the distribution of SS-H12 values for $K = 3$ (Figure S8; $\tau_2 = 500$) and $K = 4$ (Figure S9; $\tau_3 = 500$) population samples and $t = 400$ is primarily centered on negative values, but we obtained many positive outliers for $K = 5$ (Figure S10; $\tau_4 = 250$). The grouped approach was overall more sensitive than the conservative to the presence of internal ancestral sweeps. This is likely the effect of the asymmetric topology of the population phylogeny we examined here, as older convergent sweeps result in a greater number of populations sharing the sweep ancestrally. For example, a convergent sweep at $t = 800$ in the $K = 5$ scenario (Figure S10) results in a sweep ancestral to the split of four sampled populations, which comprise 80% of the sample. Thus, a single high-frequency haplotype predominates, allowing for values of H12_{Tot} larger than f_{Diff}, and SS-H12 > 0. Because all ancestral sweeps are shared identically across all populations, ambiguity in their classification does not occur, and their distributions match those of $K = 2$ scenarios.

Across our tested simulation scenarios, SS-H12 assigned only negative values of small magnitude to divergent sweeps, eliminating them as potentially-outlying signals in our analyses. This feature is important because sweeps in one population lead to differentiated haplotype frequency spectra, providing values of f_{Diff} that may spuriously resemble convergent sweeps. We are able to dampen the signal of divergent sweeps through the application of a correction factor that prevents sweeps unshared among sampled populations from generating outlying values of SS-H12 (Equation 3), regardless of the number of sampled populations K. Indeed, the distribution of SS-H12 values generated under divergent sweeps often appears no different from neutrality, leaving no appreciable power to the method (Figures 2, 3, S2-S7). Although never prominent for samples composed of $K > 2$ populations, we note that divergent sweeps may spuriously show elevated power as more populations share the sweep ancestrally (Figures S8-S10), causing the distribution of SS-H12 values in the sample to become distinct from those under neutrality and shared sweeps alike.

SS-H12 also displayed an extensive robustness to admixture as a confounding factor in shared sweep detection and classification, allowing for the confident application of our approach to a wider set of complex demographic scenarios (Figures 4 and S11). We focused primarily on admixture as a confounding factor because it has been widely documented in the literature across multiple study systems, and is therefore an important consideration in many genome-wide analyses [Chun et al., 2010, Patterson et al., 2012, Pool et al., 2012, Nedić et al., 2014]. For our simulated study scenarios, we sought to model admixture events likely to produce non-negligible distortions in the haplotypic diversity of the sample. We chose as the admixing population a donor that was highly diverged from the sampled populations, splitting one coalescent unit (20,000 generations) before sampling, meaning that admixture would likely introduce new haplotypes into the sample. Admixture most impacted the ability of SS-H12 to detect and classify ancestral sweeps, whereas convergent sweeps remained broadly unobscured and distinct from neutrality for all but the most extreme scenarios. This fit our expectations because admixture into one population within a sampled pair sharing an ancestral sweep leads to differing haplotype frequency spectra between the pair, thus spuriously resembling a convergent sweep if donor genetic diversity is low, or neutrality otherwise (Figure 4, middle column). Accordingly, convergent sweeps may still appear convergent in many cases following admixture (Figure 4, left column). In most cases, the effect of admixture is likely to be an overall reduction in the prominence of SS-H12 at sweep loci, which may impact estimates of sweep age and intensity [Malaspinas et al., 2012, Mathieson and McVean, 2013, Smith et al., 2018], but not detection (Figure S11). Once again, the SS-H12 distribution of divergent sweeps showed little departure in prominence from neutrality (Figure 4, right column)—though with a spurious but not impactful rise in power (Figure S11, right column)—but this could be because only the sweeping population received gene flow from the donor population, resulting in little change to the correction factor (Equation 3) relative to no admixture. We caution that admixture from a donor of small size into the non-sweeping population may increase H12 in that population, potentially increasing the value of the correction factor, and SS-H12. Ultimately, only a narrow range of admixture scenarios is likely to affect inferences with SS-H12, and admixture was the only confounding factor we tested that affected SS-H12.

Beyond detecting and classifying recent shared sweeps with high power, accuracy, and specificity, the SS-H12 framework provides the only approach among comparable methods that can classify shared sweeps as hard or soft from the inferred number of sweeping haplotypes (ν). We based our shared sweep classification strategy on the approximate Bayesian computation (ABC) approach of Harris et al. [2018], leveraging the observation that shared sweeps yield differing paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) value profiles based the underlying ν parameter. Using an ABC approach, we found that the classification of recent ancestral sweeps broadly followed that of sweeps in single populations, with smaller $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ corresponding to harder sweeps, and the largest ν associated with the largest $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ (Figure 5, top-left). Because we must constrain ancestral sweeps such that $t > \tau$, the range of possible SS-H12 values is reduced relative to convergent sweeps, and the boundaries among ν classes are somewhat irregular within the posterior distribution. Resolving the most probable ν therefore becomes increasingly challenging for more ancient ancestral sweeps (Figure 5, top-right). In contrast, the classification of convergent sweeps as hard or soft is both easier and more regular than is the classification of ancestral sweeps, because they will always be more recent and therefore produce a stronger sweep signal on average (Figure 5, bottom). Hard convergent sweeps occupy the largest |SS-H12| values, and pair with intermediate $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ values, reflecting the presence of two independently-sweeping haplotypes in the pooled population. Sweeps on $\nu \geq 2$ haplotypes can only produce smaller |SS-H12| values, and sweeps from the largest ν once again occupy the largest $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ values. The posterior distribution of ν for convergent sweeps also yielded $\nu = 1$ for combinations of small paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values separate from the main aforementioned $\nu = 1$ band (Figure 5, bottom). This region of the plot reflects replicates in which the same haplotype was selected independently in both populations, yielding smaller values of $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ than are possible for sweeps on different haplotypes. This also explains the presence of adjacent $\nu = 2$ and $\nu = 3$ replicates within the posterior distribution, while the absence of larger ν indicates the small probability that soft sweeps could look identical by chance. Thus, the SS-H12 framework is powerful for classifying the softness of recent shared sweeps, provided that they are sufficiently recent, losing resolution proportionally to the power trends we observed across Figures 2, 3 and S2-S7.

Because phased haplotype data are often unavailable for non-model organisms, we briefly evaluated the power of our shared sweep approach applied to diploid unphased multilocus genotypes (MLGs), which we generated by manually merging pairs of haplotypes in existing simulated samples. Previously, Harris et al. [2018] showed that the single-population approach H12 [Garud et al., 2015] could be applied to MLGs as the analogous statistic G123, which pools the three most frequent MLG frequencies as $\text{G123} = (q_1 + q_2 + q_3)^2 + \sum_{j=4}^{J} q_j^2$, where there are J distinct unphased MLGs, with q_j denoting the frequency of the jth most frequent MLG, and with $q_1 \geq q_2 \geq \cdots \geq q_J$. We construct a MLG analogue of SS-H12, denoted SS-G123, as $\text{SS-G123} = \text{G123}_{\text{Anc}} \times \frac{\min[\text{G123}^{(1)}, \text{G123}^{(2)}]}{\max[\text{G123}^{(1)}, \text{G123}^{(2)}]}$, where $\text{G123}^{(1)}$ and $\text{G123}^{(2)}$ are G123 computed in populations 1 and 2, respectively, where $\text{G123}_{\text{Anc}} = \text{G123}_{\text{Tot}} - g_{\text{Diff}}$, where $\text{G123}_{\text{Tot}} = (y_1 + y_2 + y_3)^2 + \sum_{j=4}^{J} y_j^2$, and where $g_{\text{Diff}} = \sum_{j=1}^{J}(q_{1j} - q_{2j})^2$, assuming that $y_j = \gamma q_{1j} + (1 - \gamma)q_{2j}$, $y_1 \geq y_2 \geq \cdots \geq y_J$, is the frequency of the jth most frequent MLG in the pooled population, and q_{1j} and q_{2j} are the frequencies of this MLG in populations 1 and 2, respectively. SS-G123 performed best for recent, strong sweeps (Figure S14), with trends in power broadly mirroring those of SS-H12 for identical samples with phased data (Figure 2). Power was greatest once again for convergent sweeps, which generated distributions of SS-G123 with negative values and elevated magnitudes. For ancestral sweeps, power was considerably less than for SS-H12, and ancestral sweeps regularly generated negative SS-G123 of small magnitude, rendering classification unreliable. For moderate selection (Figure S15), SS-G123 rarely produced positive SS-G123 under ancestral sweeps, and the magnitudes of even the most prominent sweep replicates were scarcely different from neutrality, despite concordant power trends with SS-H12. Thus, our results indicate that power for MLG data is not comparable with power for haplotype data to the extent observed in the single-population statistics [Harris et al., 2018]. SS-G123 maintains power for recent strong sweeps, but this power attenuates more rapidly than for haplotype-based statistics, and so we expect that outlying shared sweep candidates discovered in MLG data using the SS-G123 approach will represent a narrower complement of observations than for SS-H12.

The high power, robustness, and flexibility of SS-H12 allowed us to discover outlying sweep candidates in humans that both corroborated the results of previous investigations, and uncovered previously-unknown shared sweep candidates. Most importantly, our SS-H12 framework provided inferences about the timing and softness of shared sweeps, yielding enhanced levels of detail about candidates that were until now not directly available. As SS-H12 is the only method that distinguishes between recent ancestral and convergent shared sweeps, our investigation was uniquely able to identify loci at which independent convergent sweeps, though rare, may have played a role in shaping modern patterns of genetic diversity. Among these loci was *EXOC6B*, which produces a protein component of the exocyst [Evers et al., 2014] and has been previously highlighted as a characteristic site of selection in East Asian populations [Baye et al., 2009, Durbin and Consortium, 2011, Pybus et al., 2014]. The shared hard sweep ($\nu = 1$) at *EXOC6B* appeared as convergent between the East Asian JPT and sub-Saharan African YRI populations (Table S8), but as ancestral between all other population pairs—pairs of non-African populations—in which it appeared (Tables S9, S3, S10, and S4). Thus, we have identified the sweep at *EXOC6B* as a global occurrence that affected African and non-African populations alike, and was not limited to a single region or a single event.

More generally, our investigation into sweep signals shared between disparate population pairs also updates existing notions about when during the peopling of the world particular selective events may have occurred. For example, a sweep at *NNT*, involved in the glucocorticoid response, has been previously reported in sub-Saharan African populations [Voight et al., 2006, Fagny et al., 2014]. As expected, we recovered *NNT* as a significant ancestral hard sweep ($p < 10^{-6}$, $\nu = 1$) in the comparison between LWK and YRI (Table S6), but additionally in all comparisons between YRI and non-African populations (Tables S5, S7, and S8;

significant for all but the CEU-YRI pair). This indicates that selection at *NNT* would have preceded the out-of-Africa event and would not have been exclusive to sub-Saharan African populations. Another candidate that appeared as a top outlier outside of previously-described populations was *SPIDR*, involved in double-stranded DNA break repair [Wan et al., 2013, Smirin-Yosef et al., 2017] and inferred to be a shared candidate among Eurasian populations [Racimo, 2016]. *SPIDR* previously appeared as an outlying H12 signal in the East Asian CHB population [Harris et al., 2018], but in our present analysis was shared ancestrally not only between the East Asian KHV and JPT populations (Table S10; $p = 10^{-6}$), but also between JPT and the European CEU (Table S4; $p < 10^{-6}$), and the sub-Saharan African LWK and YRI (Table S6; $p < 10^{-6}$) populations. Once again, we found evidence of a strong sweep candidate shared among a wider range of populations than previously expected, illustrating the role of shared sweep analysis in amending our understanding of the scope of sweeps in humans worldwide.

In addition to recovering expected and expanded sweep signatures across multiple loci, we also found top outlying ancestral shared sweep candidates whose signals are not especially prominent across single-population analyses, emphasizing that localizing an ancestral sweep depends not only on elevated expected homozygosity generating the signal, but highly on the presence of shared haplotypes between populations to maintain the signal. Foremost among such candidates was *CASC4*, which appeared as an ancestral hard sweep ($\nu = 1$) in all comparisons with YRI (Tables S5, S6, S7, and S8; significant for all but the CEU-YRI pair). Because this cancer-associated gene [Ly et al., 2014, Anczuków et al., 2015] had been previously described as a shared sweep ancestral to the out-of-Africa event [Racimo, 2016], we expected to see it represented across multiple comparisons. However, *CASC4* does not have a prominent H12 value outside of sub-Saharan African populations, and within YRI is a lower-end outlier [Harris et al., 2018]. Despite this, *CASC4* is within the top 12 outlying candidates across all comparisons with YRI, and appears as the eighth-most outlying signal in the CEU-JPT comparison (Table S4, $p = 10^{-6}$, $\nu = 2$), even though it does not appear as an outlier in either population individually. Similarly, we found *PHKB*, involved in glycogen storage [Hendrickx and Willems, 1996, Burwinkel et al., 1997, Burwinkel and Kilimann, 1998], as an ancestral hard sweep between the CEU and YRI populations (Table S5; $\nu = 1$) that was not prominent in either population alone, though it had been previously inferred to be a sweep candidate ancestral to Eurasian populations once again [Racimo, 2016]. We also identified *MRAP2*, which encodes a melanocortin receptor accessory protein that is implicated in glucocorticoid deficiency [Chan et al., 2009, Asai et al., 2013], similarly to *NNT*, as an ancestral hard sweep between the CEU and JPT populations (Table S4; $p < 10^{-6}$, $\nu = 1$), and is not prominent in either CEU or JPT. Thus, we expect that the analyses of shared sweeps should be sensitive to weaker sweeps than single-population analyses, given the larger sample size of the shared approach, and sufficient haplotype sharing.

The SS-H12 framework represents an important advancement in our ability to contextualize and classify shared sweep events using multilocus sequence data. Whereas previous experiments have identified shared sweeps and can do so with high power, or without the need for MLGs or phased haplotypes, the ability to distinguish hard and soft shared sweeps, as well as ancestral and convergent sweeps, is invaluable for understanding the manner in which an adaptive event has proceeded. Discerning whether a selective sweep has occurred multiple times or just once can provide novel and updated insights into the relatedness of study populations, and the selective pressures that they endured. Moreover, the sensitivity of our approach to both hard and soft sweeps, and our ability to separate one from the other, add an additional layer of clarity that is otherwise missing from previous analysis, and is especially relevant because uncertainty persists as to the relative contributions of hard and soft sweeps in human history [Jensen, 2014, Schrider and Kern, 2017, Mughal and DeGiorgio, 2018]. Ultimately, we expect inferences deriving from SS-H12 analysis to assist in formulating and guiding more informed questions about discovered candidates across diverse organisms for which sequence data—phased and unphased—exists. After establishing the timing and softness of a shared sweep, appropriate follow-up analyses can include inferring the age of a sweep [Smith et al., 2018], identifying the favored allele or alleles [Akbari et al., 2018], or identifying other populations connected to the shared sweep. We believe that our approach will serve to enhance investigations into a diverse variety of study systems, and facilitate the emergence of new perspectives and paradigms.

Finally, we provide open-source software (titled `SS-X12`) to perform scanning window analyses on haplotype input data using SS-H12 or multilocus genotype input data using SS-G123, as well as results from our empirical scans, at `http://personal.psu.edu/mxd60/SS-X12.html`. `SS-X12` provides flexible user control, allowing the input of samples drawn from arbitrary populations K, and the output of a variety of expected homozygosity summary statistics.

Materials and Methods

We first tested the power of SS-H12 to detect shared selective sweeps on simulated multilocus sequence data, including both phased haplotypes and unphased multilocus genotypes (MLGs; applied as SS-G123). We generated haplotype data using the forward-time simulator SLiM 2 [version 2.6; Haller and Messer, 2017], which follows a Wright-Fisher model [Fisher, 1930, Wright, 1931] and can reproduce complex demographic and selective scenarios. We generated MLGs from haplotypes by manually merging each simulated individual's pair of haplotypes into a single MLG. In this way, we were able to directly assess the effects of phasing on identical samples. This series of simulations followed human-inspired parameters, with mutation rate $\mu = 2.5 \times 10^{-8}$ per site per generation, recombination rate $r = 10^{-8}$ per site per generation, and a constant diploid population size of $N = 10^4$ [Takahata et al., 1995, Nachman and Crowell, 2000, Payseur and Nachman, 2000]. As is standard for forward-time simulations [Yuan et al., 2012, Ruths and Nakhleh, 2013], we scaled these parameters by a factor $\lambda = 20$ to reduce simulation runtime, dividing the population size and duration of the simulation by λ, and multiplying the mutation and recombination rates, as well as the selection coefficient (s) by λ. Thus, scaled simulations maintained the same expected levels of genetic variation as would unscaled simulations.

For each simulated replicate, we generated a sample consisting of data from $K = 2$ or more populations related by a rooted tree with K leaves. Simulations lasted for an unscaled duration of $12N$ generations, consisting of a burn-in period of $10N$ generations to produce equilibrium levels of variation [Messer, 2013], and $2N$ generations, the mean time to coalescence of a pair of lineages. All population split events occurred within the latter $2N$ generations of the simulation. Under this approach, we examined three broad classes of sweep scenarios, consisting of ancestral, convergent, and divergent sweeps. For ancestral sweeps, we introduced a selected allele to one or more randomly-drawn haplotypes in the ancestor of all sampled populations (*i.e.*, before any population splits), which ensured that the same selective event was shared in the histories of all populations. For convergent sweeps, we introduced the selected mutation independently in each extant population at the time of selection, after at least one split had occurred. Finally, divergent sweeps comprised any scenario in which the sweep event occurred in fewer than all sampled populations, meaning that at least one population did not experience a sweep by the time of sampling. Across all simulations, we conditioned on the maintenance of the selected allele in any affected population after its introduction.

We measured the ability of our approach to detect shared sweep events in simulated samples consisting of individuals from $K = 2$ to 5 populations by quantifying the abilities of SS-H12 and SS-G123 to distinguish between scenarios of selection and neutrality under equivalent demographic histories. SS-H12 and SS-G123 are compatible with an arbitrary number of K populations (see subsequent explanation). To generate distributions of the SS-H12 statistic, we scanned 100 kb of sequence data from simulated individuals using a sliding window approach. We computed SS-H12 and SS-G123 in 40 kb windows, advancing the window by increments of one kb across the simulated chromosome for a total of 61 windows. For each replicate, we retained the value of SS-H12 or SS-G123 from the window of absolute maximum value as the score. We selected a 40 kb window size to overcome the effect of short-range LD in the sample, which may produce a signature of expected haplotype homozygosity resembling a sweep. We measured the decay of LD for SNPs in neutral replicates separated by one to 100 kb at one kb intervals using mean r^2. For all parameter sets, we generated 10^3 sweep replicates and 10^3 neutral replicates between which population sizes, number of populations, and population split times were identical.

For simulated data consisting of individuals from $K \geq 3$ sampled populations (which we analyze only for haplotype data), we assigned values of SS-H12 in one of two ways. First, we employed a conservative approach in which we computed the SS-H12 score for each possible population pair in the aforementioned manner, but retained as the replicate score only the SS-H12 value with the smallest magnitude. That is, the replicate score had to satisfy the condition $|\text{SS-H12}_{K\geq 3}| = \min_{i\neq j}\{|\text{SS-H12}_{ij}|\}$, where SS-H12_{ij} is SS-H12 computed between populations i and j. Assigning SS-H12 in this manner ensured that only samples wherein all represented populations shared a sweep were likely to yield outlying values. Second, we explored a grouped approach in which we assigned the SS-H12 statistic between the two branches (denoted α and β) directly subtending the root of the phylogeny relating the set of K populations, treating the two subtrees respectively descending from these branches as single populations. Thus, $\text{SS-H12}_{\text{group}} = \text{H12}_{\text{Tot}} - f_{\text{Diff}}^{(\alpha,\beta)}$, where H12_{Tot} is the expected haplotype homozygosity of the pooled population, and $f_{\text{Diff}}^{(\alpha,\beta)} = \sum_{i=1}^{I}(p_{\alpha i} - p_{\beta i})^2$, where $p_{\alpha i}$ and $p_{\beta i}$ are the mean frequencies of haplotype i on branches α and β, respectively.

Across all tested population histories, we evaluated the ability of SS-H12 to identify hard selective sweeps from a *de novo* mutation and soft sweeps from selection on standing genetic variation, for both strong ($s = 0.1$) and moderate ($s = 0.01$) strengths of selection. This setting matched the experimental approach of Harris et al. [2018] for single-population statistics, and corresponds to scenarios for which those statistics have power under the specific mutation rate, recombination rate, and effective size we tested here. For all selection scenarios, we placed the beneficial allele at the center of the simulated chromosome, and introduced it at only one timepoint, constraining the selection start time, but not the selection end time. For hard and soft sweep scenarios, we allowed the selected allele to rise in frequency toward fixation (with no guarantee of reaching fixation). To specify soft sweep scenarios, we conditioned on the selected allele being present in the population on $\nu = 4$ or 8 distinct haplotypes at the start of selection (*i.e.*, 0.4 to 0.8% of haplotypes in the population would initially carry a beneficial allele), without defining the number of selected haplotypes remaining in the population at the time of sampling, as long as the selected allele was not lost.

For all method performance evaluation experiments, we observed the effect of varying the selection start time, and the time at which populations split from one another. As in Harris et al. [2018], we initiated selection at $t \in [200, 4000]$ generations prior to sampling. This range of selection times was chosen not to represent specific selective events in human history, but to cover the range of time over which hypothesized selective sweeps in recent human history have occurred [Przeworski, 2002, Sabeti et al., 2007, Beleza et al., 2012, Jones et al., 2013, Clemente et al., 2014, Fagny et al., 2014]. We varied split times (τ) across all experiments, with $\tau \in [250, 2000]$ generations before sampling. The combination of t and τ defined the simulation type. For two-population scenarios, simulations wherein $t > \tau$ produced ancestral sweeps, and simulations with $t < \tau$ yielded convergent or divergent sweeps depending on the number of populations under selection.

The number of population split events specified the number of populations in the simulated sample. For simulations in which the ancestral population split only once, at time τ, the sample consisted of $K = 2$ populations. To extend our notation for more than two populations, we index the divergence time as τ_k, $k = 1, 2, \ldots, K-1$, for K populations. Similarly, if we included two population splits, at times τ_1 and τ_2 (where $\tau_1 > \tau_2$), then the sample consisted of individuals from $K = 3$ populations. For experiments involving $K \geq 3$ sampled populations, we split populations at regularly repeating intervals, with each split generating a new population identical to its ancestor. We generated only asymmetric tree topologies, splitting each new population from the same ancestral branch. Furthermore, experiments with $K \geq 3$ populations allowed for the simulation of more complex selection scenarios featuring nested ancestral sweeps, part of larger convergent or divergent sweep events, wherein the time of selection occurs between τ_1 and τ_{K-1}, with $\tau_1 > \tau_2 > \cdots > \tau_{K-1}$ and $\tau_k = \frac{K-k}{K-1}\tau_1$ for $k = 1, 2, \ldots, K-1$.

We additionally observed the effects of two confounding factors on SS-H12 to establish the extent to which inferences of shared sweeps in the sampled populations could be misled. For these experiments, we studied only scenarios with $K = 2$ populations. First, we examined the effect of admixture on one of the two sampled populations. Second, we generated samples under long-term background selection, which is known to yield similar patterns of diversity to sweeps [Charlesworth et al., 1993, 1995, Seger et al., 2010, Nicolaisen and Desai, 2013, Cutter and Payseur, 2013, Huber et al., 2016]. For the admixture experiments, we simulated single pulses of admixture at fractions between 0.05 and 0.4, at intervals of 0.05, from a diverged unsampled donor, 200 generations prior to sampling (thus, following the start of selection in the sample). We simulated three different scenarios of admixture into the sampled target population from the donor population. These consisted of a highly-diverse donor population ($N = 10^5$, tenfold larger than the sampled population), which may obscure a sweep signature in the sampled target, and from a low-diversity donor population ($N = 10^3$, 1/10 the size of the sampled population), which may produce a sweep-like signature in the target, in addition to an intermediately-diverse donor population ($N = 10^4$, equal to the size of the sampled population). For divergent sweep experiments, only the population experiencing the sweep was the target.

Our background selection simulations followed the same protocol as in previous work [Cheng et al., 2017]. At the start of the simulation, we introduced a centrally-located 11-kb gene composed of UTRs (5' UTR of length 200 nucleotides [nt], 3' UTR of length 800 nt) flanking a total of 10 exons of length 100 nt separated by introns of length one kb. Strongly deleterious ($s = -0.1$) mutations arose throughout the course of the simulation across all three genomic elements under a gamma distribution of fitness effects with shape parameter 0.2 at rates of 50%, 75%, and 10% for UTRs, exons, and introns, respectively. The sizes of the genic elements follow human mean values [Mignone et al., 2002, Sakharkar et al., 2004]. We considered samples drawn from a scenario with $K = 2$ populations, and a split time of $\tau = 1000$ or 2000 between the populations.

As in Harris et al. [2018], we employed an approximate Bayesian computation (ABC) approach to demonstrate the ability of SS-H12, in conjunction with the $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$ statistics, to classify shared sweeps as hard or soft from the inferred number of sweeping haplotypes ν. Hard sweeps derive from a single sweeping haplotype, while soft sweeps consist of at least two sweeping haplotypes. Whereas the single-population approach [Garud et al., 2015, Garud and Rosenberg, 2015, Harris et al., 2018] identified hard and soft sweeps from their occupancy of paired (H12, H2/H1) values, we presently use paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values to classify shared sweeps. We defined a 100-by-100 grid corresponding to paired (|SS-H12|, $\text{H2}_{\text{Tot}}/\text{H1}_{\text{Tot}}$) values with each axis bounded by [0.005, 0.995] at increments of 0.01, and assigned the most probable value of ν to each test point in the grid.

We define the most probable ν for a test point as the most frequently-observed value of ν from the posterior distribution of 5×10^6 sweep replicates within a Euclidean distance of 0.1 from the test point. For each replicate, we drew $\nu \in \{1, 2, \dots, 16\}$ uniformly at random, as well as $s \in [0.005, 0.5]$ uniformly at random from a log-scale. Across ancestral and convergent sweep scenarios for $K = 2$ sampled sister populations, we generated replicates for $\tau = 1000$ and 2000 generations before sampling. For ancestral sweeps, we drew $t \in [1020, 2000]$ for $\tau = 1000$ scenarios, and $t \in [2020, 3000]$ for $\tau = 2000$ scenarios, both uniformly at random. Similarly, convergent sweeps were drawn from $t \in [200, 980]$ ($\tau = 1000$) and $t \in [200, 1980]$ ($\tau = 2000$). Simulated haplotypes were of length 40 kb, corresponding to the window size for method performance evaluations, because in practice a value of ν would be assigned to a candidate sweep based on its most prominent associated signal. Mutation and recombination rates, as well as sample size per population, were identical to previous experiments.

We applied SS-H12 to human empirical data from the 1000 Genomes Project Consortium [Auton et al., 2015]. We scanned all autosomes for signatures of shared sweeps in population pairs using 40 kb windows advancing by increments of four kb for samples of non-African populations, and 20 kb windows advancing by two kb for any samples containing individuals from any African populations. We based these window sizes on the interval over which LD, measured as r^2, decayed beyond less than half its original value relative to pairs of loci separated by one kb (Figure S16). As in Harris et al. [2018], we filtered our output data by removing analysis windows containing fewer than 40 SNPs, the expected number of SNPs corresponding to the extreme case in which a selected allele has swept across all haplotypes except for one, leaving two lineages [Watterson, 1975]. Following Huber et al. [2016], we also divided all chromosomes into non-overlapping bins of length 100 kb and assigned to each bin a mean CRG100 score [Derrien et al., 2012], which measures site mappability and alignability. We removed windows within bins whose mean CRG100 score was below 0.9, with no distinction between genic and non-genic regions. Thus, our overall filtering strategy was identical to that of Harris et al. [2018]. We then intersected remaining candidate selection peaks with the coordinates for protein- and RNA-coding genes from their hg19 coordinates.

Finally, we assigned p-values and the most probable inferred ν to the top 40 ancestral and convergent RNA- and protein-coding sweep candidates recovered in our genomic scans. We assigned p-values by generating 10^6 neutral replicates in *ms* [Hudson, 2002] under the appropriate two-population demographic history, inferred from `smc++` [Terhorst et al., 2017]. Here, we drew the sequence length for each replicate uniformly at random from the set of all hg19 gene lengths and appended the window size, providing as input at least one full-length analysis window. The p-value for a selection candidate is the proportion of |SS-H12| values generated under neutrality whose value exceeds the maximum |SS-H12| assigned to the candidate. Following Bonferroni correction for multiple testing [Neyman and Pearson, 1928], a significant p-value was $p < 0.05/23{,}735 \approx 2.10659 \times 10^{-6}$, where 23,735 is the number of protein- and RNA-coding genes for which we assigned a score.

We assigned the most probable ν for each sweep candidate following the same protocol as the constant-size demographic history simulation analyses, generating 5×10^6 replicates of sweep scenarios generated in SLiM 2 under `smc++`-inferred demographic histories for ancestral and convergent sweeps, drawing $t \in [200, 5000]$ uniformly at random, and $s \in [0.005, 0.5]$ uniformly at random on a log scale. Once again, $t > \tau$ for ancestral sweep scenarios and $t < \tau$ for convergent sweep scenarios, where τ is defined by the specific demographic history of the sample. Sequence length for each replicate was identical to analysis window length for equivalent empirical data (20 or 40 kb), because in practice we assign ν to windows of this size. For both p-value and most probable ν assignment, we used a per-site per-generation mutation rate of $\mu = 1.25 \times 10^{-8}$ and a per-site per-generation recombination rate of $r = 3.125 \times 10^{-9}$ [Narasimhan et al., 2017, Terhorst et al., 2017].

Acknowledgments

This work was funded by National Institutes of Health grant R35GM128590 and by the Alfred P. Sloan Foundation. Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by the Institute for CyberScience at Pennsylvania State University.

References

A Akbari, J J Vitti, A Iranmehr, M Bakhtiari, P C Sabeti, S Mirarab, and V Bafna. Identifying the favored mutation in a positive selective sweep. *Nat. Methods*, 15:279–282, 2018.

O Anczuków, M Akerman, A Cléry, J Wu, C Shen, N H Shirole, A Raimer, S Sun, M A Jensen, Y Hua, F H T Allain, and A R Krainer. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol. Cell*, 60: 105–117, 2015.

M Asai, S Ramachandrappa, M Joachim, Y Shen, R Zhang, N Nuthalapati, V Ramanathan, D E Strochlic, P Ferket, K Linhart, C Ho, T V Novoselova, S Garg, M Ridderstråle, C Marcus, J N Hirschhorn, J M Keogh, S O'Rahilly, L F Chan, A J Clark, I S Farooqi, and J A Majzoub. Loss of Function of the Melanocortin 2 Receptor Accessory Protein 2 Is Associated with Mammalian Obesity. *Science*, 341: 275–278, 2013.

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.

T M Baye, R A Wilke, and M Olivier. Genomic and geographic distribution of private SNPs and pathways in human populations. *Pers. Med.*, 6:623–641, 2009.

S Beleza, A M Santos, B McEvoy, I Alves, C Martinho, E Cameron, M D Shriver, E J Parra, and J Rocha. The Timing of Pigmentation Lightening in Europeans. *Mol. Biol. Evol.*, 30:24–35, 2012.

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.

M Bonhomme, C Chevalet, B Servin, S Boitard, J Abdallah, S Blott, and M SanCristobal. Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics*, 186:241–262, 2010.

C Bonilla, S Hooker, T Mason, C H Bock, and R A Kittles. Prostate Cancer Susceptibility Loci Identified on Chromosome 12 in African Americans. *PLoS ONE*, 6:e16044, 2011.

B Burwinkel and M W Kilimann. Unequal Homologous Recombination Between LINE-1 Elements as a Mutational Mechanism in Human Genetic Disease. *J. Mol. Biol.*, 277:513–517, 1998.

B Burwinkel, A J Maichele, Ø Aegenaes, H D Bakker, A Lerner, Y S Shin, J A Strachan, and M W Kilimann. Autosomal glycogenosis of liver and muscle due to phosphorylase kinase deficiency is caused by mutations in the phosphorylase kinase subunit β(*PHKB*). *Hum. Mol. Genet.*, 6:1109–1115, 1997.

L F Chan, T R Webb, T Chung, E Meimaridou, S N Cooray, L Guasti, J P Chapple, M Egertová, M R Elphick, M E Cheetham, L A Metherell, and A J L Clark. MRAP and MRAP2 are bidirectional regulators of the melanocortin receptor family. *Proc. Natl. Acad. Sci. U.S.A.*, 106:6146–6151, 2009.

Y C Chang, X Liu, J D O Kim, M A Ikeda, M R Layton, A B Weder, R S Cooper, S L R Kardia, D C Rao, S C Hunt, A Luke, E Boerwinkle, and A Chakravarti. Multiple Genes for Essential-Hypertension Susceptibility on Chromosome 1q. *Am. J. Hum. Genet.*, 80:253–264, 2007.

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.

X Cheng, C Xu, and M DeGiorgio. Fast and robust detection of ancestral selective sweeps. *Mol. Ecol.*, 2017. doi: 10.1111/mec.14416.

Y J Chun, B Fumanal, B Laitung, and F Bretagnolle. Gene flow and population admixture as the primary post-invasion processes in common ragweed (*Ambrosia artemisiifolia*) populations in France. *New Phytol.*, 185:1100–1107, 2010.

F J Clemente, A Cardona, C E Inchley, B M Peter, G Jacobs, L Pagani, D J Lawson, T Antão, M Vicente, M Mitt, M DeGiorgio, Z Faltyskova, Y Xue, Q Ayub, M Szpak, R Mägi, A Eriksson, A Manica, M Raghavan, M Rasmussen, S Rasmussen, E Willerslev, A Vidal-Puig, C Tyler-Smith, R Villems, R Nielsen, M Metspalu, B Malyarchuk, M Derenko, and T Kivisild. A Selective Sweep on a Deleterious Mutation in *CPT1A* in Arctic Populations. *Am. J. Hum. Genet.*, 95:584–589, 2014.

S Climer, A R Templeton, and W Zhang. Human *gephyrin* is encompassed within giant functional noncoding yinyang sequences. *Nat. Commun.*, 6, 2015. doi: 10.1038/ncomms7534.

G Coop, J K Pickrell, J Novembre, S Kudaravalli, J Li, D Absher, R M Myers, L L Cavalli-Sforza, M W Feldman, and J K Pritchard. The Role of Geography in Human Adaptation. *PLoS Genet.*, 5:e1000500, 2009.

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.

T Derrien, J Estellé, S M Sola, D G Knowles, E Rainieri, R Guigó, and P Ribeca. Fast Computation and Applications of Genome Mappability. *PLoS ONE*, 7:e30377, 2012.

R M Durbin and The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2011.

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

C Evers, B Maas, K A Koch, A Jauch, J W G Janssen, C Sutter, M J Parker, K Hinderhofer, and U Moog. Mosaic Deletion of EXOC6B: Further Evidence for An Important Role of the Exocyst Complex in the Pathogenesis of Intellectual Disability. *Am. J. Med. Genet. Part A*, 164:3088–3094, 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.

M I Fariello, S Boitard, H Naya, M SanCristobal, and B Servin. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics*, 193:929–941, 2013.

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

N R Garud and N A Rosenberg. Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theor. Popul. Biol.*, 102:94–101, 2015.

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.

P Gerbault, C Moret, M Currat, and A Sanchez-Mazas. Impact of Selection and Demography on the Diffusion of Lactase Persistence. *PLoS ONE*, 4:e6369, 2009.

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

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.

J B S Haldane. A mathematical theory of natural and artificial selection. V. selection and mutation. *Math. Proc. Camb. Philos. Soc.*, 23:838–844, 1927.

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

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*, In press, 2018.

K Harris and R Nielsen. The Genetic Cost of Neanderthal Introgression. *Genetics*, 203:881–891, 2016.

J Hendrickx and P J Willems. Genetic deficiencies of the glycogen phosphorylase system. *Hum. Genet.*, 97: 551–556, 1996.

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.

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.

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

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

K E Johnson and B F Voight. Patterns of shared signatures of recent positive selection across human populations. *Nat. Ecol. Evol.*, 2:713–720, 2018.

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.

I Juric, S Aeschbacher, and G Coop. The Strength of Selection against Neanderthal Introgression. *PLoS Genet.*, 12:e1006340, 2016.

R Kato, A Nonami, T Taketomi, T Wakioka, A Kuroiwa, Y Matsuda, and A Yoshimura. Molecular cloning of mammalian Spred-3 which suppresses tyrosine kinase-mediated Erk activation. *Biochem. Bioph. Res. Co.*, 302:767–772, 2003.

P Kheirandish and F Chinegwundoh. Ethnic differences in prostate cancer. *Brit. J. Cancer*, 105:481–485, 2011.

M Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

K Kodama, D Tojjar, S Yamada, K Toda, C J Patel, and A J Butte. Ethnic Differences in the Relationship Between Insulin Sensitivity and Insulin Response. *Diabetes Care*, 36:1789–1796, 2013.

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 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.

X Liu, R T Ong, E N Pillai, A M Elzein, K S Small, T G Clark, D P Kwiatowski, and Y Teo. Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. *Am. J. Hum. Genet.*, 92: 866–881, 2013.

T Ly, Y Ahmad, A Shlien, D Soroka, A Mills, M Emanuele, M R Stratton, and A I Lamond. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *eLife*, 3:e01630, 2014.

A Malaspinas, O Malaspinas, S N Evans, and M Slatkin. Estimating Allele Age and Selection Coefficient from Time-Serial Data. *Genetics*, 192:599–607, 2012.

C B Mallick, F M Iliescu, M Möls, S Hill, R Tamang, G Chaubey, R Goto, S Y W Ho, I G Romero, F Crivellaro, G Hudjashov, N Rai, M Metspalu, C G N Mascie-Taylor, R Pitchappan, L Singh, M Mirazon-Lahr, K Thangaraj, R Villems, and T Kivisild. The Light Skin Allele of SLC24A5 in South Asians and Europeans Shares Identity by Descent. *PLoS Genet.*, 9:e1003912, 2013.

B J Maron, K P Carney, H M Lever, J F Lewis, I Barac, S A Casey, and M V Sherrid. Relationship of Race to Sudden Cardiac Death in Competitive Athletes With Hypertrophic Cardiomyopathy. *J. Am. Coll. Cardiol.*, 41:974–980, 2003.

I Mathieson and G McVean. Estimating Selection Coefficients in Spatially Structured Populations from Time Series Data of Allele Frequencies. *Genetics*, 193:973–984, 2013.

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

P W Messer. SLiM: Simulating Evolution with Selection and Linkage. *Genetics*, 194:1037–1039, 2013.

P W Messer and D A Petrov. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.*, 28:659–669, 2013.

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. *BioRxiv*, 2018. doi: 10.1101/320523.

M W Nachman and S L Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156: 297–304, 2000.

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.

N Nedić, R M Francis, L Stanisavljević, I Pihler, N Kezić, C Bendixen, and P Kryger. Detecting population admixture in honey bees of Serbia. *J. Apicult. Res.*, 53:303–313, 2014.

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.

H A Orr. The population genetics of beneficial mutations. *Phil. Trans. R. Soc. B*, 365:1195–1201, 2010.

L Park. Linkage Disequilibrium Decay and Past Population History in the Human Genome. *PLoS ONE*, 7: e46603, 2012.

N Patterson, P Moorjani, Y Luo, S Mallick, N Rohland, Y Zhan, T Genschoreck, T Webster, and D Reich. Ancient Admixture in Human History. *Genetics*, 192:1065–1093, 2012.

B A Payseur and M W Nachman. Microsatellite Variation and Recombination Rate in the Human Genome. *Genetics*, 156:1285–1298, 2000.

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.

S Peyrégne, M J Boyle, M Dannemann, and K Prüfer. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.*, 27:1563–1572, 2017.

J E Pool, R B Corbett-Detig, R P Sugino, K A Stevens, C M Cardeno, M W Crepeau, P Duchen, J J Emerson, P Saelao, D J Begun, and C H Langley. Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genet.*, 8:e1003080, 2012.

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

M Pybus, G M Dall'Olio, P Luisi, M Uzkudun, A Carreño-Torres, P Pavlidis, H Laayouni, J Bertranpetit, and J Engelken. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.*, 42:D903–D909, 2014.

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

M Ramming, S Kins, N Werner, A Hermann, H Betz, and J Kirsch. Diversity and phylogeny of gephyrin: Tissue-specific splice variants, gene structure, and sequence similarities to molybdenum cofactor-synthesizing and cytoskeleton-associated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10266–10271, 2000.

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.

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

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.

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

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.

D Shenoy, S Packianathan, A M Chen, and S Vijayakumar. Do African-American men need separate prostate cancer screening guidelines? *BMC Urol.*, 16:19, 2016.

P Smirin-Yosef, N Zuckerman-Levin, S Tzur, Y Granot, L Cohen, J Sachsenweger, G Borck, I Lagovsky, M Salmon-Divon, L Wiesmüller, and L Basel-Vanagaite. A Biallelic Mutation in the Homologous Recombination Repair Gene SPIDR Is Associated With Human Gonadal Dysgenesis. *J. Clin. Endocrinol. Metab.*, 102:681–688, 2017.

J Smith, G Coop, M Stephens, and J Novembre. Estimating Time to the Common Ancestor for a Beneficial Allele. *Mol. Biol. Evol.*, 35:1003–1017, 2018.

N Takahata, Y Satta, and J Klein. Divergence Time and Population Size in the Lineage Leading to Modern Humans. *Theor. Popul. Biol.*, 48:198–221, 1995.

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.

S K Tyagarajan and J Fritschy. Gephyrin: a master regulator of neuronal function? *Nat. Rev. Neuro.*, 15: 141–156, 2014.

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.

L Wan, J Han, T Liu, S Dong, F Xie, H Chen, and J Huang. Scaffolding protein SPIDR/KIAA0146 connects the Bloom syndrome helicase with homologous recombination repair. *Proc. Natl. Acad. Sci. U.S.A.*, 110: 10646–10651, 2013.

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

S H Williamson, M J Hubisz, A G Clark, B A Payseur, C D Bustamante, and R Nielsen. Localizing Recent Adaptive Evolution in the Human Genome. *PLoS Genet.*, 3:e90, 2007.

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

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

X Xie, Z Gong, V Mansuy-Aubert, Q L Zhou, S A Tatulian, D Sehrt, F Gnad, L M Brill, K Motamedchaboki, Y Chen, M P Czech, M Mann, M Krüger, and Z Y Jiang. C2 Domain-Containing Phosphoprotein CDP138 Regulates GLUT4 Insertion into the Plasma Membrane. *Cell Metab.*, 14:378–389, 2011.

K Yang, J Shen, Y Xie, Y Lin, J Qin, Q Mao, X Zheng, and L Xie. Promoter-targeted double-stranded small RNAs activate *PAWR* gene expression in human cancer cells. *Int. J. Biochem. Cell B.*, 45:1338–1346, 2013.

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.

Q L Zhou, Y Song, C Huang, J Huang, Z Gong, Z Liao, A G Sharma, L Greene, J Z Deng, M C Rigor, X Xie, S Qi, J E Ayala, and Z Y Jiang. Membrane Trafficking Protein CDP138 Regulates Fat Browning and Insulin Sensitivity through Controlling Catecholamine Release. *Mol. Cell. Biol.*, 38:e00153–17, 2018.

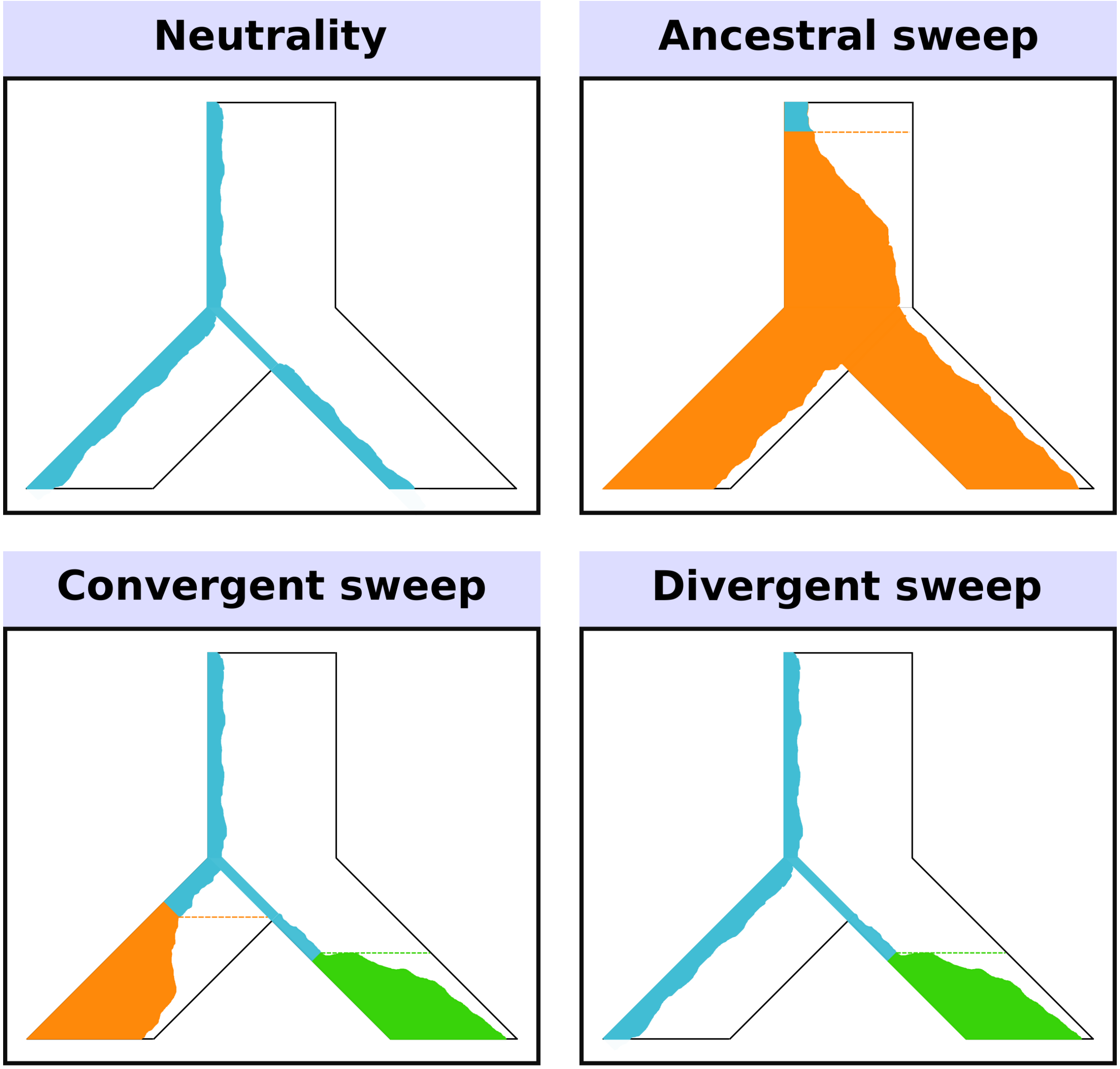

Figure 1: Model of a two-population phylogeny for which SS-H12 detects recent shared sweeps. Here, an ancestral population splits in the past into two modern lineages, which are sampled. Each panel displays the frequency trajectory of a haplotype in the population. Under neutrality, there is high haplotypic diversity such that many haplotypes, including the reference haplotype (blue), exist at low frequency. In the ancestral sweep, the reference haplotype becomes selectively advantageous (turning orange) and rises to high frequency prior to the split, such that both modern lineages carry the same selected haplotype at high frequency. The convergent sweep scenario involves different selected haplotypes independently rising to high frequency in each lineage after their split. Under a divergent sweep, only one sampled lineage experiences selection.

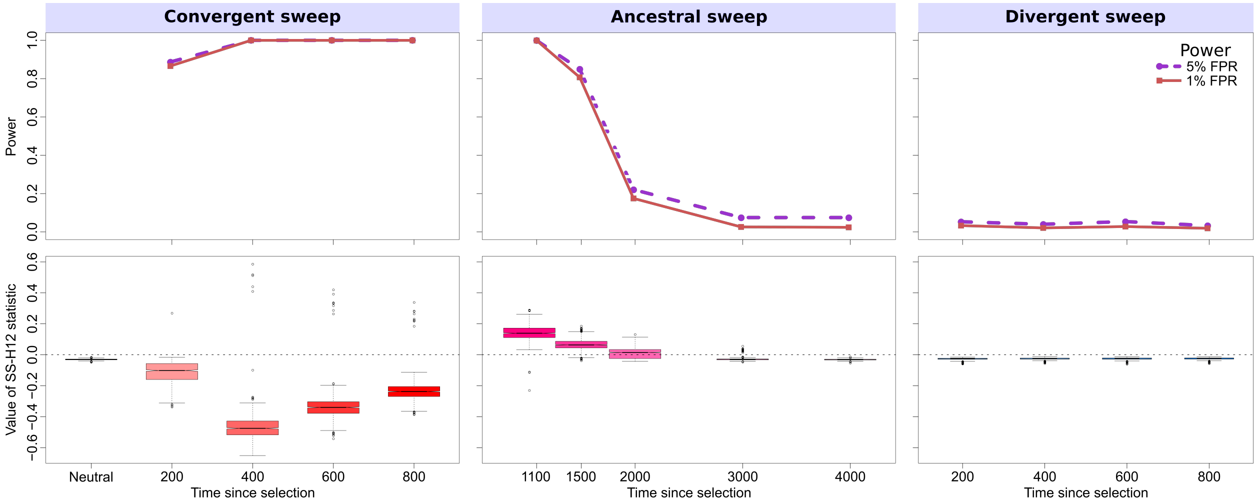

Figure 2: Properties of SS-H12 for simulated hard sweep scenarios in which $\tau = 1000$ generations before sampling. (Top row) Power at 1 and 5% false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps (see Figure 1) as a function of time at which selection initiated, with false positive rate based on the distribution of maximum |SS-H12| across simulated neutral replicates. (Bottom row) Box plots summarizing the distribution of SS-H12 values from windows of maximum |SS-H12| for each replicate, corresponding to each point in the power curve, with dashed lines in each panel representing SS-H12 = 0. Convergent and divergent sweeps occur after this time (200-800 generations before sampling), while ancestral sweeps occur before this time (1100-4000 generations before sampling). All sweeps are strong ($s = 0.1$) for a sample of $n = 100$ diploid individuals per population, with 1000 replicates performed for each scenario.

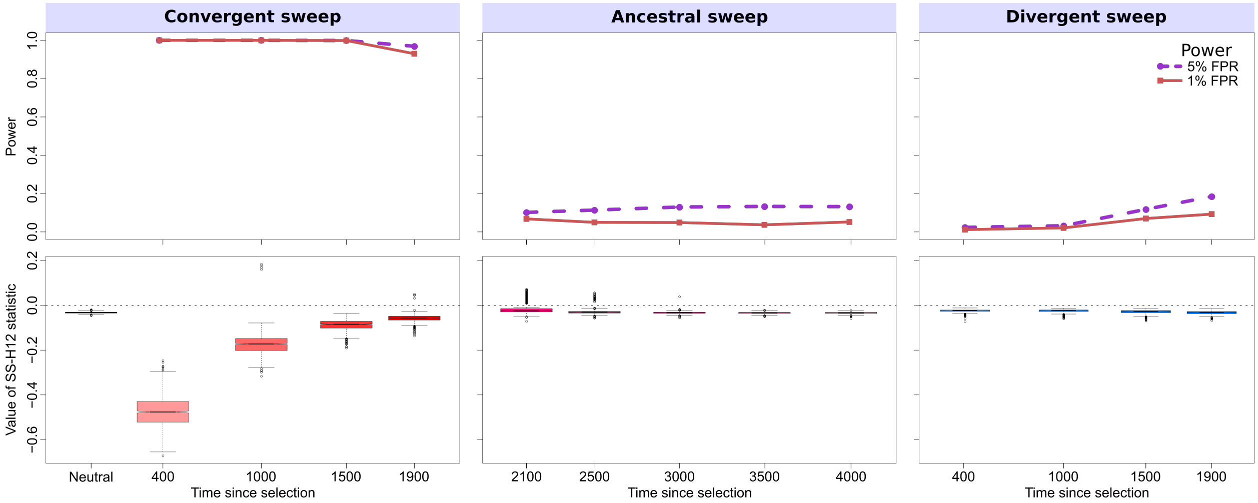

Figure 3: Properties of SS-H12 for simulated hard sweep scenarios in which $\tau = 2000$ generations before sampling. (Top row) Power at 1 and 5% false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps (see Figure 1) as a function of time at which selection initiated, with false positive rate based on the distribution of maximum |SS-H12| across simulated neutral replicates. (Bottom row) Box plots summarizing the distribution of SS-H12 values from windows of maximum |SS-H12| for each replicate, corresponding to each point in the power curve, with dashed lines in each panel representing SS-H12 = 0. Convergent and divergent sweeps occur after this time (400-1900 generations before sampling), while ancestral sweeps occur before this time (2100-4000 generations before sampling). All sweeps are strong ($s = 0.1$) for a sample of $n = 100$ diploid individuals per population, with 1000 replicates performed for each scenario.

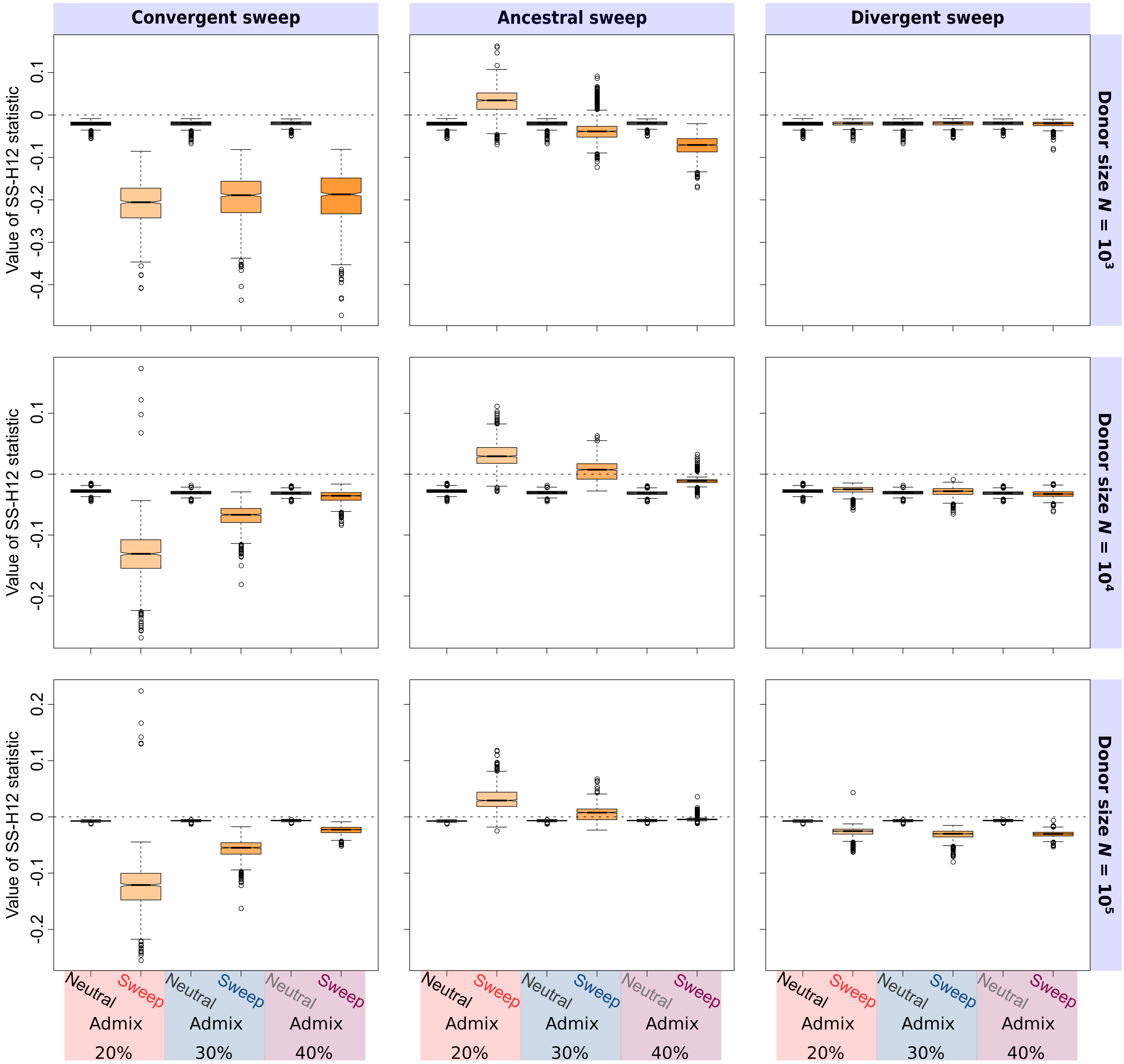

Figure 4: Effect of admixture from a diverged, unsampled donor lineage on distributions of SS-H12 values at peaks of maximum |SS-H12|, in samples consisting of individuals from $K = 2$ populations with $\tau = 1000$, under simulated recent ancestral, convergent, and divergent histories (see Figure 1). For ancestral sweeps, selection occurred 1400 generations before sampling. For convergent and divergent sweeps, selection occurred 600 generations before sampling. The effective size of the donor population varies from $N = 10^3$ (an order of magnitude less than the sampled populations), to $N = 10^5$ (an order of magnitude more), with admixture modeled as a single pulse occurring 200 generations before sampling at rates 0.2 to 0.4. The donor diverged from the sampled populations $2 \times 10^4 = 2N$ generations before sampling and in the case of divergent sweep scenarios, admixed specifically into the population experiencing a sweep. All sample sizes are of $n = 100$ diploid individuals, with 1000 replicates performed for each scenario.

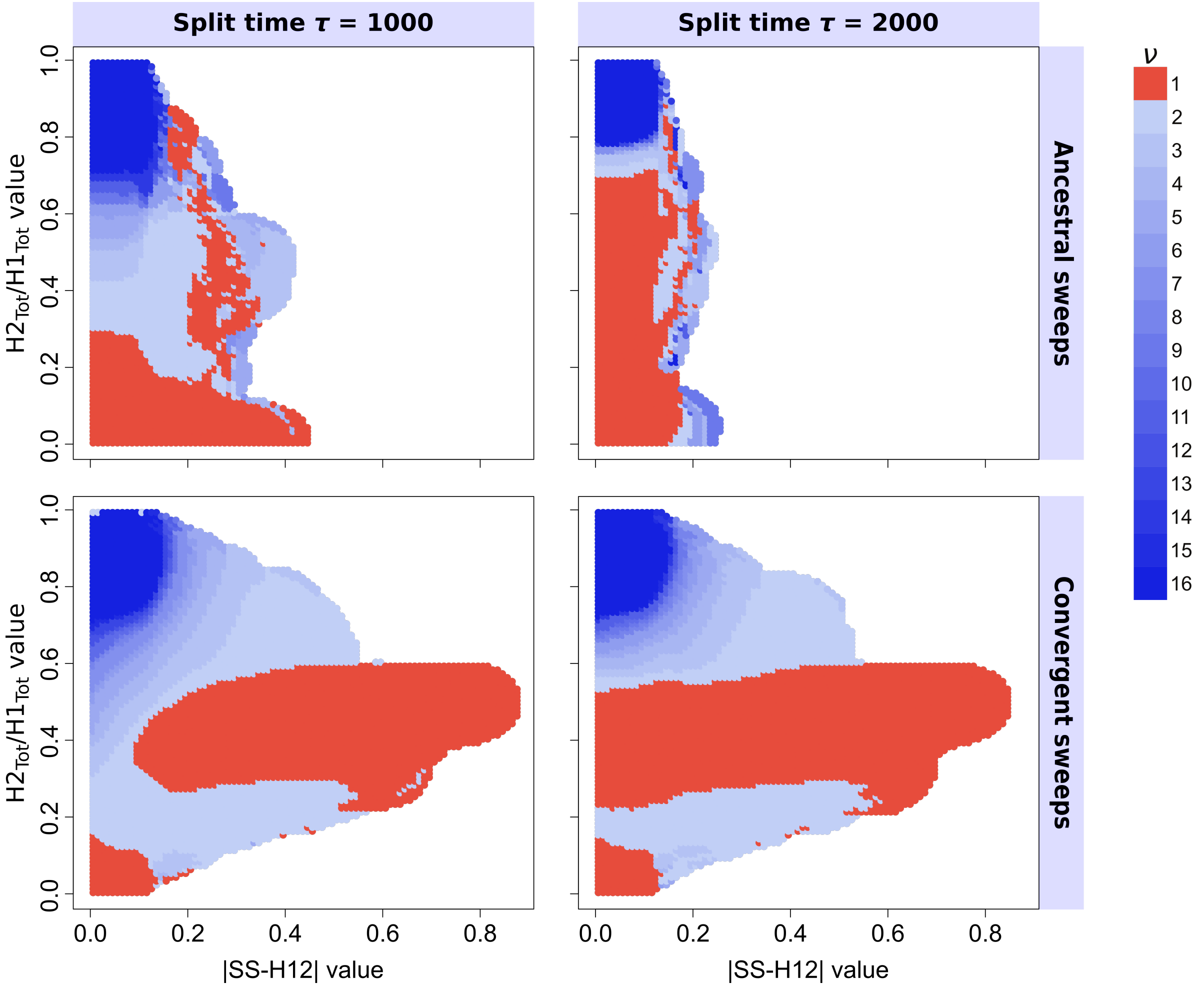

Figure 5: Ability of paired (|SS-H12|, $\mathrm{H2_{Tot}/H1_{Tot}}$) values to infer the most probable number of sweeping haplotypes ν in a shared sweep. Most probable ν for each test point was assigned from the posterior distribution of 5×10^6 sweep replicates with $\nu \in \{1, 2, \ldots, 16\}$, drawn uniformly at random. (Top row) Ancestral sweeps for $\tau = 1000$ (left) and $\tau = 2000$ (right), with $t \in [1020, 2000]$ (left) and $t \in [2020, 3000]$ (right). (Bottom row) Convergent sweeps for $\tau = 1000$ (left) and $\tau = 2000$ (right), with $t \in [200, 980]$ (left) and $t \in [200, 1980]$ (right). Colored in red are points whose paired (|SS-H12|, $\mathrm{H2_{Tot}/H1_{Tot}}$) values are more likely to result from hard sweeps, whereas those colored in shades of blue are points more likely to be generated from soft sweeps. Regions in white are those for which no observations of sweep replicates within a Euclidean distance of 0.1 were made.

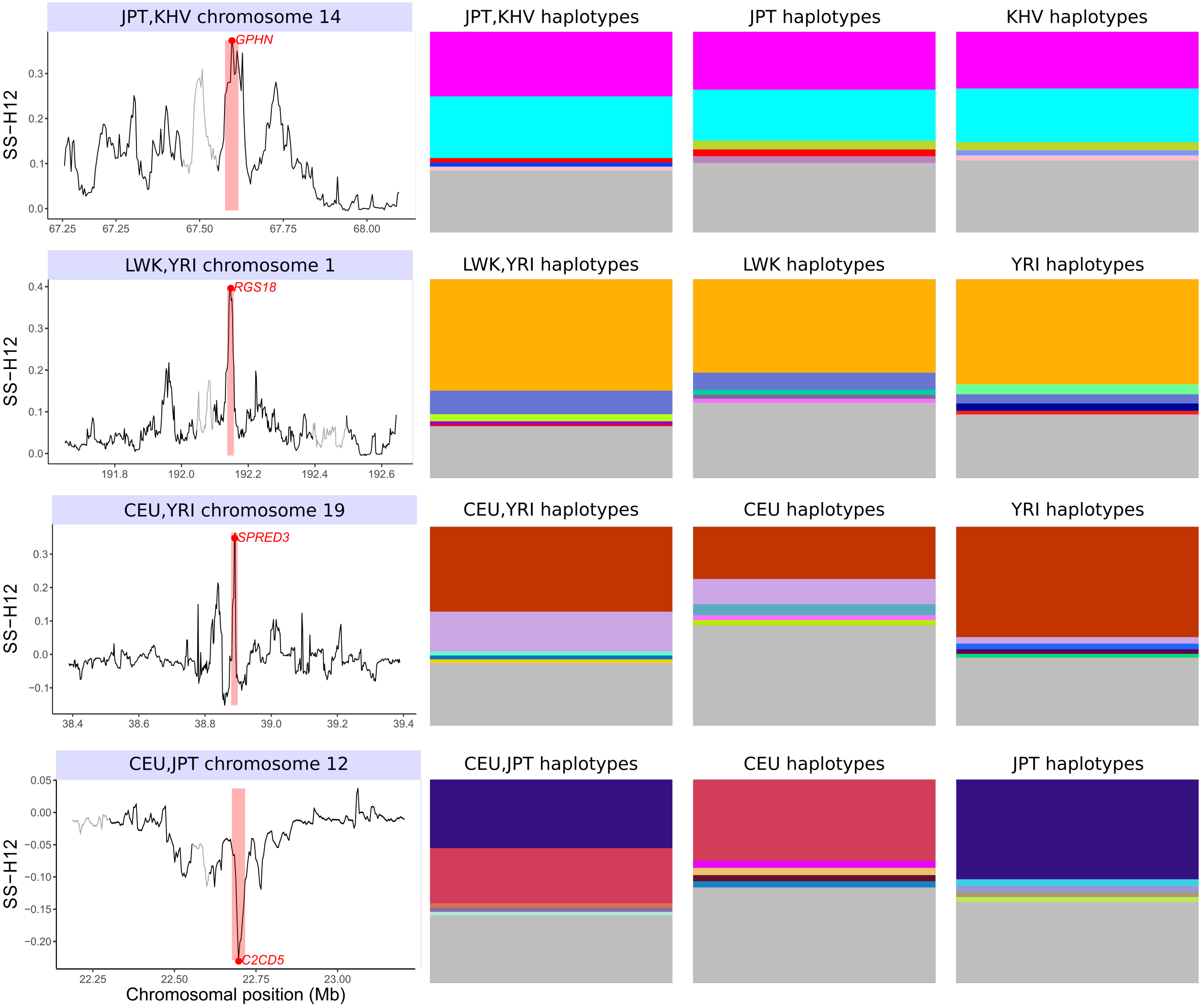

Figure 6: Top outlying shared sweep candidates at RNA- and protein-coding genes in global human populations. The signal peak, including chromosomal position, magnitude, and highlighted window of maximum SS-H12 (left column), as well as the haplotype frequency spectrum (right three columns) are displayed for each candidate. The East Asian JPT and KHV populations experience an ancestral soft sweep at *GPHN* (top row). The sub-Saharan African populations LWK and YRI share an ancestral hard sweep at *RGS18* (second row). The European CEU population experiences a shared sweep with YRI at *SPRED3* (third row). The European CEU and East Asian JPT have a convergent sweep at *C2CD5*, with a different, single high-frequency haplotype present in each population (bottom row). Coloration within the haplotype frequency spectrum plots indicates particular moderate to high-frequency haplotypes, while rows shaded in gray represent the remainder of haplotypes that are each at low frequency.