Supplementary Note

Identifying and classifying shared selective sweeps from multilocus data

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

March 6, 2020

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

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

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

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

Contents

1 Detecting selective sweeps in a simplified demographic model

We measured the power of our approach to detect shared sweep events in simulated samples consisting of individuals from $K \in \{2, 3, 4, 5\}$ populations, related by a rooted tree with K leaves and under a simpler demographic model, in order to illustrate the effect of sampling multiple diverged populations. A classic example of this is the colonization of the Galápagos Islands, which for many founding populations resulted in their fragmentation due to the limitations on gene flow between islands [Ciofi et al., 2006, Steinfartz et al., 2009, Hedrick, 2019]. For these experiments, which we intended as a representation of a more general mammalian model (Table 2, first row), population sizes remained at a constant $N = 10^4$ diploids throughout the simulation, rather than fluctuating as in the human model experiments. 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. 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. Doing so allowed us to fully control the number of populations k affected by a simulated sweep by simply changing the time of selection t. Thus, we could explore sweep scenarios affecting any number of simulated populations $k \in \{1, 2, \ldots, K\}$. Our chosen split times for experiments were $\tau = 1000$ ($K = 2$ sampled populations); $\tau_1 = 1000$ and $\tau_2 = 750$ ($K = 3$); $\tau_1 = 1000$, $\tau_2 = 750$, and $\tau_3 = 500$ ($K = 4$), and $\tau_1 = 1000$, $\tau_2 = 750$, $\tau_3 = 500$, and $\tau_4 = 250$ ($K = 5$), and we simulated strong hard sweeps specifically ($s = 0.1$; see Figure S1 for model).

Because SS-H12 is compatible with an arbitrary number of populations, we used our simulations under the generalized mammalian model to evaluate our ability to detect sweeps on $K > 2$ sampled populations. This is important because experiments with $K > 2$ populations could feature complex convergent and divergent sweep events harboring nested ancestral sweeps when the time of selection t occurred between τ_1 and τ_{K-1}, with $\tau_1 > \tau_2 > \cdots > \tau_{K-1}$ and $\tau_k = \tau_1 - 250(k-1)$ for $k = 1, 2, \ldots, K-1$ (Figure S1). We define SS-H12 for $K > 2$ samples in one of two ways. First, we can employ a conservative approach in which we compute SS-H12 for each possible population pair, but retain as the sample value only the smallest-magnitude SS-H12 value. Meaning, we constrain that $|\text{SS-H12}_{K\geq 3}| = \min_{i \neq j}\{|\text{SS-H12}_{ij}|\}$, where SS-H12_{ij} is SS-H12 computed between populations i and j. Assigning SS-H12 in this manner ensures that only samples wherein all represented populations share a sweep are likely to yield outlying values. Second, we can follow a grouped

approach in which we assign 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 individual populations. Thus, $\text{H12}_{\text{Anc,group}} = \text{H12}_{\text{Tot}} - f_{\text{Diff}}^{(\alpha,\beta)}$, 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, and where α and β are weighted proportionally to their sample sizes. Accordingly, we have that $\text{SS-H12}_{\text{group}} = \text{H12}_{\text{Anc,group}} \times \frac{\min[\text{H12}^{(\alpha)},\text{H12}^{(\beta)}]}{\max[\text{H12}^{(\alpha)},\text{H12}^{(\beta)}]}$.

The conservative and grouped approaches yielded comparable power to one another, and could readily detect strong shared sweeps more recent than 1500 generations old, with power rapidly attenuating for more ancient sweeps, and once again greatest for convergent sweeps (Figures SN1-SN3, top). Furthermore, trends in power for detecting shared sweeps remained consistent between $K = 2$ (Figure S9) and $K > 2$ (Figures SN1-SN3, top) scenarios, regardless of the choice of approach. 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 (Figures SN1-SN3, middle and bottom). 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 (Figures SN1-SN3, middle) remains generally more robust to misclassifying shared sweeps within samples from $K \geq 3$ populations than does the grouped approach (Figures SN1-SN3, bottom), 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.

2 Effect of inter-sister admixture on SS-H12

To complement experiments based upon the diverse-donor admixture model (main text Figures 4 and S19), we examined a second possible scenario in which gene flow occurs between sampled sister populations. Here, a selective sweep occurs either ancestrally, convergently, or divergently as previously, but unidirectional admixture occurs as a single pulse from one (donor) population into its (target) sister. We maintained times of selection and admixture, as well as the split time between sisters identical to the diverse-donor model (see

main text Table 2, "admixture, inter-sister"). Additionally, in the case of a divergent sweep, we initiated selection in the donor population and examined both a scenario in which the selected allele was adaptive in the target, and one in which it became neutral in the target. All sweep scenarios yielded SS-H12 distributions consistent with ancestral sweeps, characterized by positive values of high magnitude reflecting the shared haplotypes between populations (Figure SN4, bottom). Accordingly, SS-H12 distinguishes these scenarios from neutrality with high power (Figure SN4, top), but the selective history of the sample is misidentified. The effect of admixture following a divergent sweep is somewhat reduced for the neutral scenario relative to the adaptive, but because SS-H12 is highly sensitive to the presence of shared haplotypes, a misleading inference nonetheless emerges. Thus, we caution that it is helpful to search for shared sweeps in relatively unadmixed samples in order to minimize the possibility of both overlooking shared sweeps, or misinterpreting introgressed sweeps as true shared sweeps.

3 SS-H12 for non-simultaneous convergent sweeps

We tested the power and classification ability of SS-H12 for convergent sweeps initiating at different time-points, maintaining all other parameters identical to the $K = 2$ generalized mammalian model (Figure S9). Previously, we made the simplifying assumption in simulations that convergent sweeps would start simultaneously at time t in all affected populations. This is unlikely to be the case in natural populations, particularly under allopatry. To demonstrate the effect of varying relative selection times, we modeled a history in which one of a pair of sister populations related by the generalized mammalian model experiences a strong hard sweep at $t_1 = 800$ generations before sampling, and the other experiences a sweep at $t_2 \in \{200, 400, 600, 800\}$ generations before sampling. SS-H12 has excellent power to resolve each convergent sweep scenario from neutrality, and properly identify them as convergent from negative values (Figure SN5) The magnitude of SS-H12 responds to the time at which the variable sweep occurred, showing the greatest value for $t_2 = 400$, as with the identically-timed scenario (Figure S9, left column). We therefore expect that as long as all sweeps occur within the detectability window for SS-H12, meaning that the haplotypic signature of selection has not degraded due to the age of the selective event, the relative timing of convergent sweeps does not impact the performance of our approach.

4 Performance for a star phylogeny

As another deviation from the generalized mammalian model, we also simulated sweeps on a simple star phylogeny with $K = 4$ descendants radiating from the root node, each of which were part of the sample (Figures SN6 and SN7). We simulated shared sweeps as previously, choosing strong and hard sweeps ($\nu = 1$, $s = 0.1$) for maximum signal, specifying ancestral sweeps as those with $t > \tau$ ($\tau = 1000$), and convergent sweeps occurring simultaneously to one another at a time more recent than τ ($t < \tau$). Unlike for our prior asymmetric tree topology (Figure SN2), we simulated divergent sweeps as one to three independent events (rather than a single event) depending on the scenario. Because grouping (see *Materials and Methods*) has no meaning for equally-related populations, we report only the conservative SS-H12 for the sample. Under the $K = 4$ star topology, SS-H12 performs nearly identically to the $K = 2$ scenario (Figure S9), but with a somewhat faster decay in power for ancestral sweeps, while it outperforms the asymmetric $K = 4$ topology for convergent sweeps (Figure SN2, left column) by rarely yielding positive values of SS-H12 (Figure SN6, left column). The former observation reflects the greater likelihood that between more sampled populations there will be a greater haplotypic diversity and therefore weaker SS-H12 signal over time, while the latter observation derives from the lack of internal ancestral sweeps to distort inferences of population history. Meanwhile, SS-H12 is robust to divergent sweeps occurring independently across any number of branches less than four, though with a slight increase in spurious power for three independent divergent sweeps (Figure SN7, right column). Our results indicate that our formulation of SS-H12 is likely to be indifferent to tree topology, and identify only prominent shared sweeps.

References

C Ciofi, G A Wilson, L B Beheregaray, C Marquez, J P Gibbs, W Tapia, H L Snell, A Caccone, and J R Powell. Phylogeographic History and Gene Flow Among Giant Galápagos Tortoises on Southern Isabela Island. *GENETICS*, 172:1727–1744, 2006.

P W Hedrick. Galapagos Islands Endemic Vertebrates: A Population Genetics Perspective. *J. Hered.*, 110: 137–157, 2019.

S Steinfartz, S Glaberman, D Lanterbecq, M A Russello, S Rosa, T C Hanley, C Marquez, H L Snell, H M Snell, G Gentile, G Dell'Olmo, A M Powell, and A Caccone. Progressive colonization and restricted gene flow shape island-dependent population structure in Galápagos marine iguanas (*Amblyrhynchus cristatus*). *BMC Evol. Biol.*, 9:297, 2009.

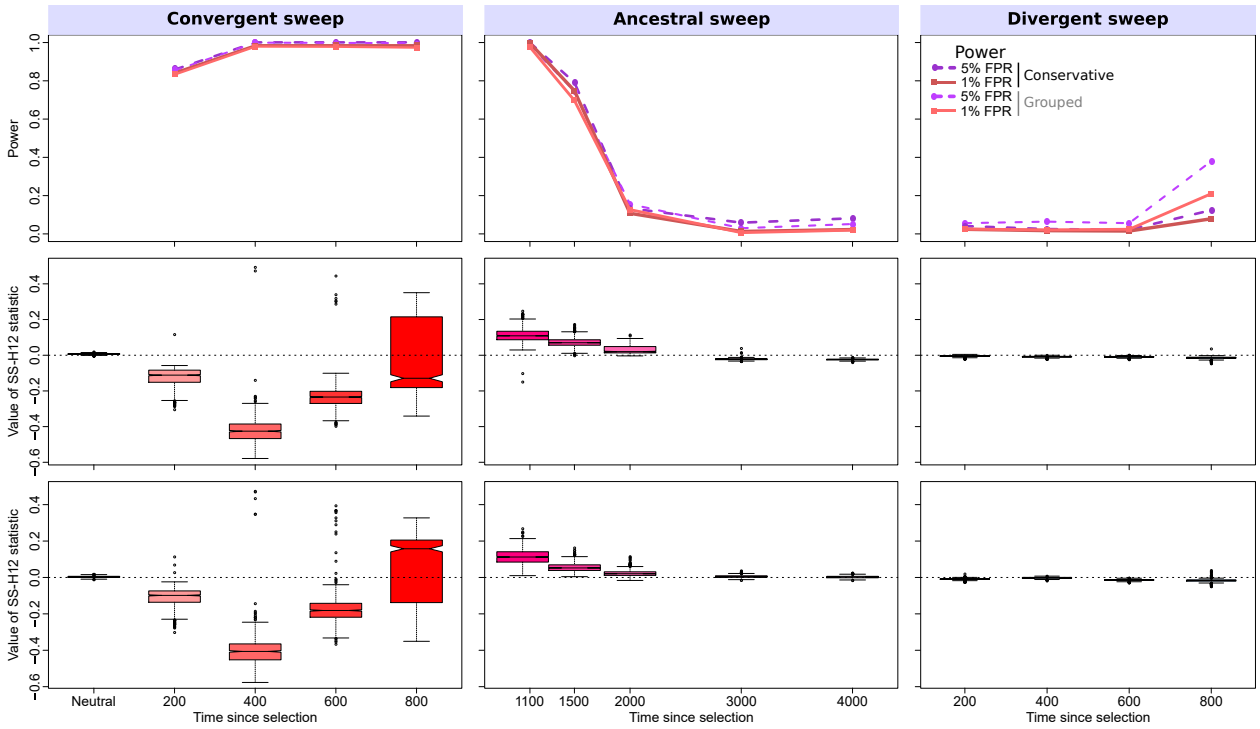

Figure SN1: Properties of SS-H12 for simulated hard sweep scenarios for samples drawn from $K = 3$ equally-sized populations in which $\tau_1 = 1000$ (0.05 coalescent units) and $\tau_2 = 750$ (0.0375) generations before sampling. (Top) Power at 1 and 5% false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps (see main text 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. (Middle and bottom rows) 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 for conservative (middle row) and grouped (bottom row) approaches, with dashed lines in each panel representing SS-H12 = 0. Convergent and divergent sweeps occur more recently than this time (200-800 generations, or 0.01-0.04 coalescent units, before sampling), while ancestral sweeps occur more anciently than this time (1100-4000 generations, or 0.055-0.2 coalescent units, before sampling). All sweeps are strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) for a sample of $n = 100$ diploid individuals per population, with 1000 replicates performed for each scenario.

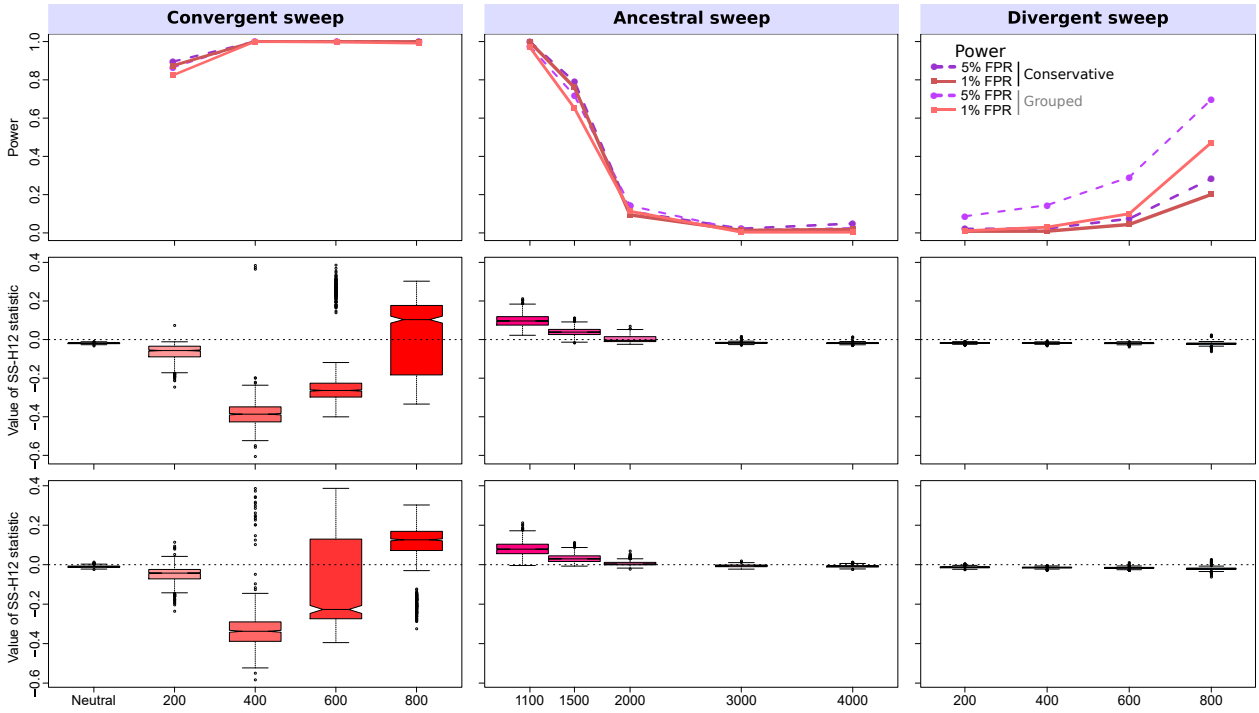

Figure SN2: Properties of SS-H12 for simulated hard sweep scenarios for samples drawn from $K = 4$ equally-sized populations in which $\tau_1 = 1000$ (0.05 coalescent units), $\tau_2 = 750$ (0.0375), and $\tau_3 = 500$ (0.025) generations before sampling. (Top row) Power at 1 and 5% false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps (see main text 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. (Middle and bottom rows) 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 for conservative (middle row) and grouped (bottom row) approaches, with dashed lines in each panel representing SS-H12 = 0. Convergent and divergent sweeps occur more recently than this time (200-800 generations, or 0.01-0.04 coalescent units, before sampling), while ancestral sweeps occur more anciently than this time (1100-4000 generations, or 0.055-0.2 coalescent units, before sampling). All sweeps are strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) for a sample of $n = 100$ diploid individuals per population, with 1000 replicates performed for each scenario.

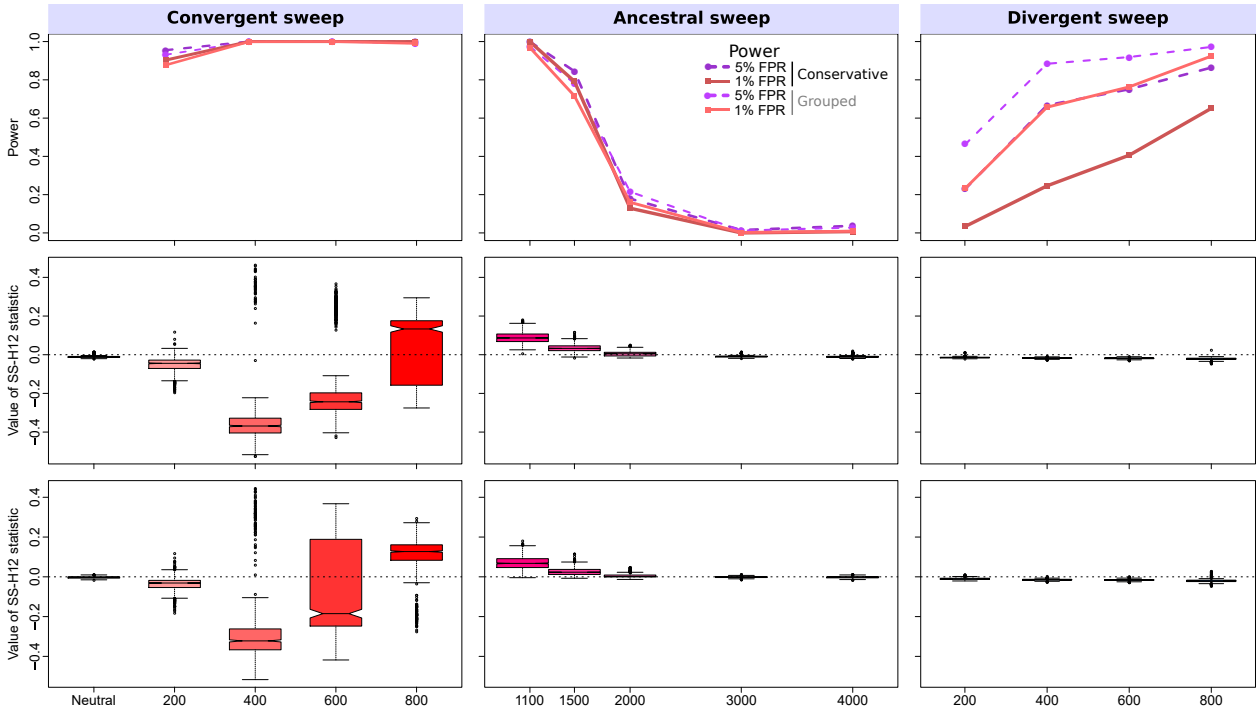

Figure SN3: Properties of SS-H12 for simulated hard sweep scenarios for samples drawn from $K = 5$ equally-sized populations in which $\tau_1 = 1000$ (0.05 coalescent units), $\tau_2 = 750$ (0.0375), $\tau_3 = 500$ (0.025), and $\tau_4 = 250$ (0.0125) generations before sampling. (Top row) Power at 1 and 5% false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps (see main text 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. (Middle and bottom rows) 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 for conservative (middle row) and grouped (bottom row) approaches, with dashed lines in each panel representing SS-H12 = 0. Convergent and divergent sweeps occur more recently than this time (200-800 generations, or 0.01-0.04 coalescent units, before sampling), while ancestral sweeps occur more anciently than this time (1100-4000 generations, or 0.055-0.2 coalescent units, before sampling). All sweeps are strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) for a sample of $n = 100$ diploid individuals per population, with 1000 replicates performed for each scenario.

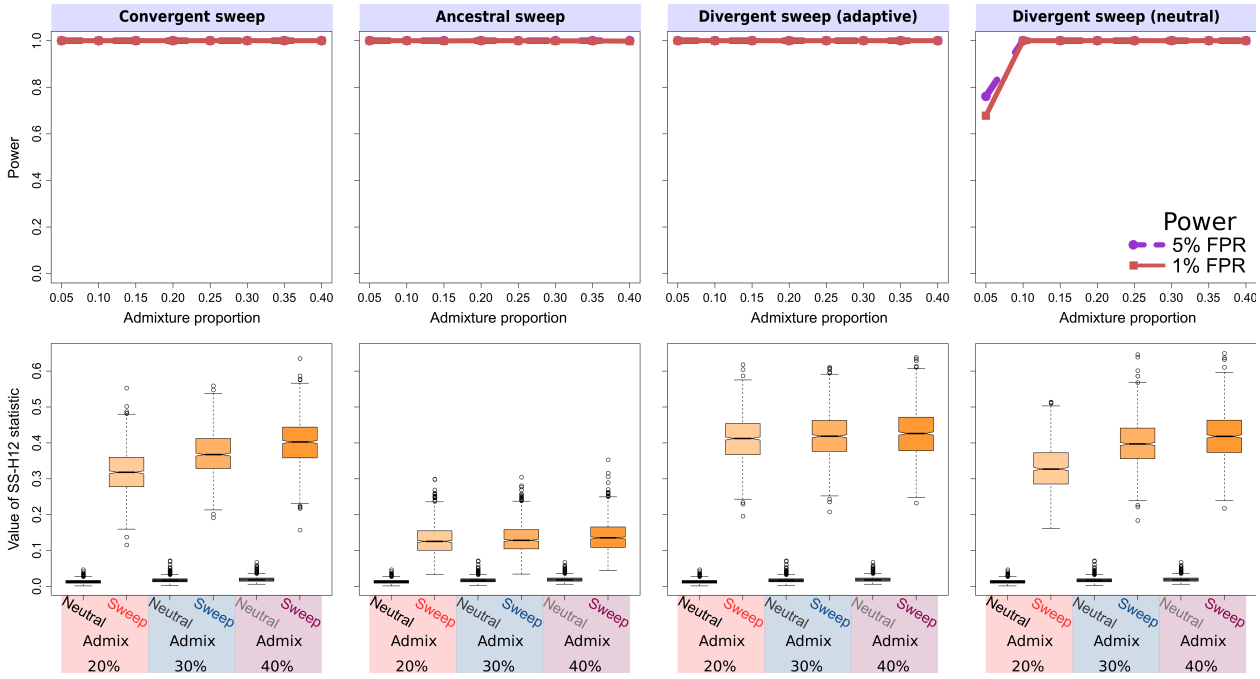

Figure SN4: Power (top) at 1 and 5% false positive rates (FPRs) and distribution (bottom) of SS-H12 for a $K = 2$ population history under the simplified mammalian model in which the populations split $\tau = 1000$ generations, or 0.05 coalescent units, prior to sampling, and one population (donor) admixed into its sister (target) 200 (0.01 coalescent units) generations prior to sampling as a single pulse, at proportions 0.2, 0.3, or 0.4. Ancestral sweeps occurred at $t = 1400$ (0.07) generations before sampling, while convergent and divergent sweeps occurred at $t = 600$ (0.03), as in main text Figure 4. Two divergent sweep scenarios were modeled, one in which the sweep was adaptive in the target, and another in which it was adaptive only in the donor and neutral in the target. All sweeps were strong ($s = 0.1$; $\sigma = 4N_e s = 4000$).

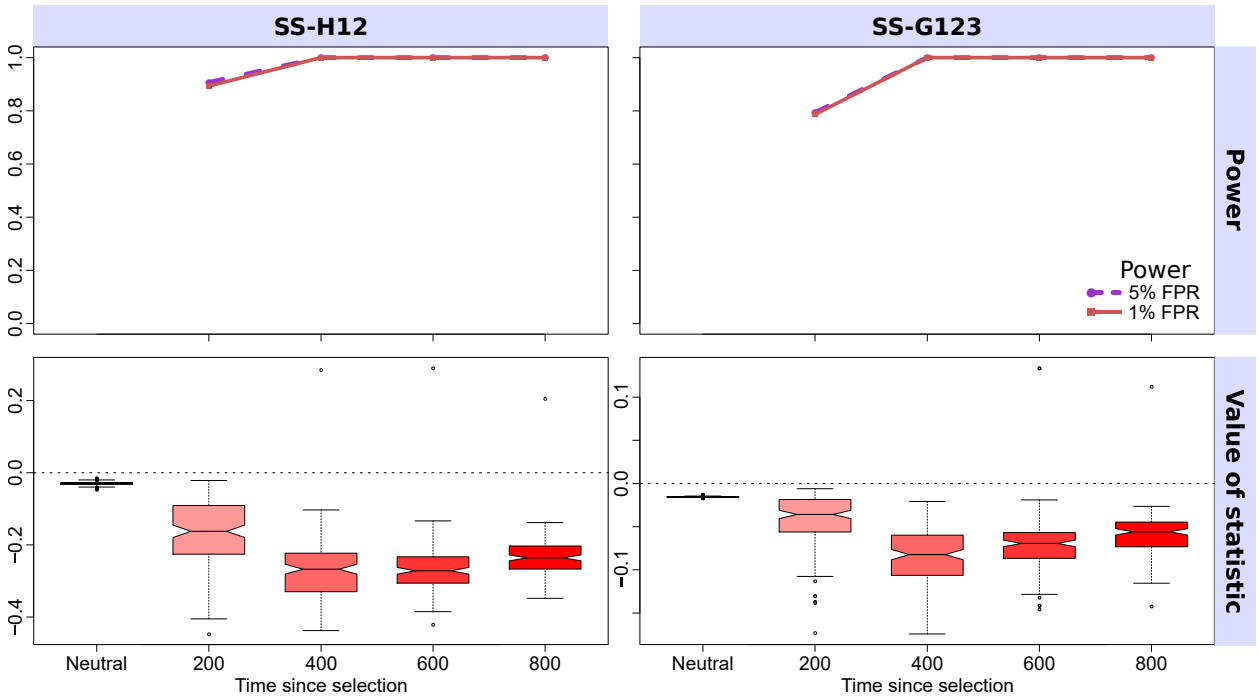

Figure SN5: Properties of SS-H12 (left) and SS-G123 (right) for simulated strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) hard convergent sweep scenarios under the simplified mammalian model ($K = 2$, $\tau = 1000$ generations, or 0.05 coalescent units, before sampling). Sweep time was fixed in one population at $t_1 = 800$ generations, or 0.04 coalescent units, before sampling, while in the other population it varied across $t_2 \in \{200, 400, 600, 800\}$. (Top row) Power at 1% (red) and 5% (purple) false positive rates (FPRs) to detect recent convergent sweeps), with FPR based on the distribution of maximum |SS-H12| (left) or |SS-G123| (right) across simulated neutral replicates. (Bottom row) Box plots summarizing the distribution of SS-H12 (left) or SS-G123 (right) values from windows of maximum |SS-H12| (left) or |SS-G123| (right) across replicates, with dashed lines in each panel representing SS-H12 = 0. Other than differing sweep times, protocol was identical to that of experiments in Figure S9.

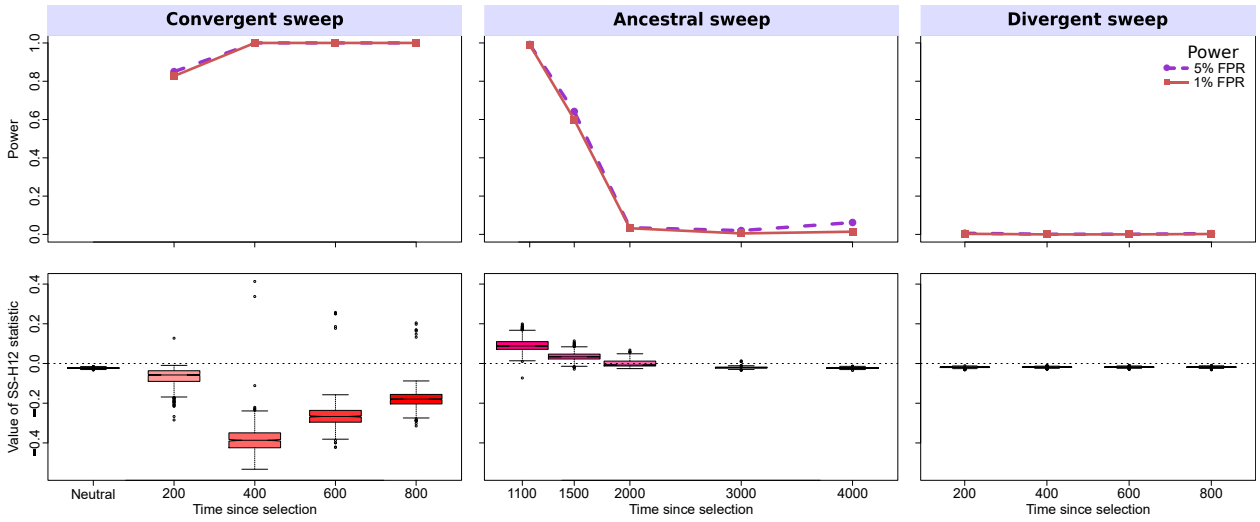

Figure SN6: Properties of SS-H12 for simulated strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) hard sweep scenarios under the simplified mammalian model and star tree topology wherein all descendant lineages split from their common ancestor simultaneously ($K = 4$, $\tau = 1000$ generations, or 0.05 coalescent units, before sampling). (Top row) Power at 1% (red) and 5% (purple) false positive rates (FPRs) to detect recent ancestral, convergent, and divergent hard sweeps, with FPR 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| across strong sweep replicates, corresponding to each time point in the power curves, with dashed lines in each panel representing SS-H12 = 0. Note that the divergent sweep occurs in only one population. Other than tree topology, protocol was identical to that of experiments in Figure SN2.

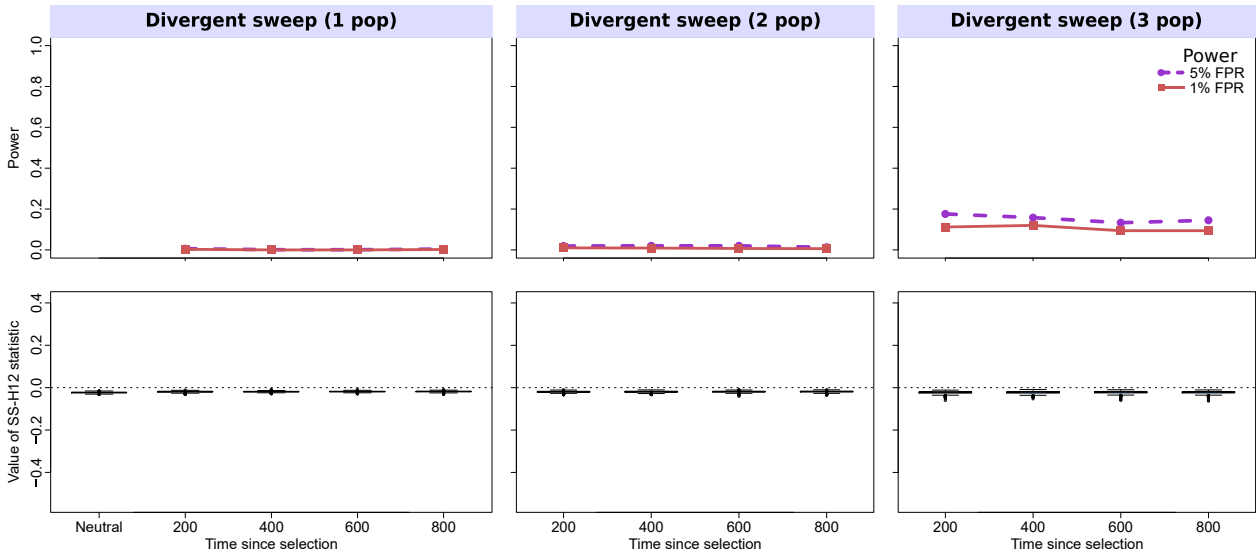

Figure SN7: Properties of SS-H12 for simulated strong ($s = 0.1$; $\sigma = 4N_e s = 4000$) hard divergent sweep scenarios under the simplified mammalian model and star tree topology wherein all descendant lineages split from their common ancestor simultaneously ($K = 4$, $\tau = 1000$ generations, or 0.05 coalescent units, before sampling). Expanding upon Figure SN6, we compare power for divergent sweeps on one (left; identical to Figure SN6), two (center), or three (right) populations out of the total four. For the latter two scenarios, divergent sweeps initiated simultaneously.