Type: Application Note

selscan 2.0: scanning for sweeps in unphased data

Zachary A. Szpiech[1,2,*]

[1] *Department of Biology, Pennsylvania State University, University Park, PA 16802, USA*
[2] *Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA 16802, USA*
[*] *Correspondence: szpiech@psu.edu*

Abstract

Summary: Several popular haplotype-based statistics for identifying recent or ongoing positive selection in genomes require knowledge of haplotype phase. Here we provide an update to selscan which implements a re-definition of these statistics for use in unphased data.

Availability and Implementation: Source code and binaries freely available at https://github.com/szpiech/selscan, implemented in C/C++ and supported on Linux, Windows, and MacOS.

Contact: *szpiech@psu.edu*

Supplemental Information: Online supplemental information available

1 Introduction

Haplotype-based summary statistics—such as iHS (Voight, et al., 2006), nSL (Ferrer-Admetlla, et al., 2014), XP-EHH (Sabeti, et al., 2007), and XP-nSL (Szpiech, et al., 2021)—have become commonplace in evolutionary genomics studies to identify recent and ongoing positive selection in populations (e.g.,Colonna, et al., 2014; Crawford, et al., 2017; Lu, et al., 2019; Meier, et al., 2018; Nedelec, et al., 2016; Salmon, et al., 2021; Zhang, et al., 2020; Zoledziewska, et al., 2015). When an adaptive allele sweeps through a population, it leaves a characteristic pattern of long high-frequency haplotypes and low genetic diversity in the vicinity of the allele.

These statistics aim to capture these signals by summarizing the decay of haplotype homozygosity as a function of distance from a putatively selected region, either within a single population (iHS and nSL) or between two populations (XP-EHH and XP-nSL).

These haplotype-based statistics are powerful for detecting recent positive selection (Colonna, et al., 2014; Crawford, et al., 2017; Lu, et al., 2019; Meier, et al., 2018; Nedelec, et al., 2016; Salmon, et al., 2021; Zhang, et al., 2020; Zoledziewska, et al., 2015), and the two-population versions can even out-perform pairwise Fst scans on a large swath of the parameter space (Szpiech, et al., 2021). Furthermore, haplotype-based methods have also been shown to be robust to background selection (Fagny, et al., 2014; Schrider, 2020). However, each of these statistics presumes that haplotype phase is known or well-estimated.

As the generation of genomic sequencing data for non-model organisms is becoming routine (Ellegren, 2014), there are many great opportunities for studying recent adaptation across the tree of life (e.g., Campagna and Toews (2022)). However, often these organisms/populations do not have a well-characterized demographic history or recombination rate map, two pieces of information which are important inputs for statistical phasing methods (Browning, et al., 2021; Delaneau, et al., 2013).

Recent work has shown that haplotype-based statistics can be adapted for use on unphased data (Klassmann and Gautier, 2022) and that converting haplotype data into "multi-locus genotype" data is an effective approach for using haplotype-based selection statistics such as G12, LASSI, and saltiLASSI (DeGiorgio and Szpiech, 2022; Harris and DeGiorgio, 2020; Harris, et al., 2018) in unphased data. Recognizing this, we have reformulated the iHS, nSL, XP-EHH, and XP-nSL statistics to use multi-locus genotypes and provided an easy-to-use implementation in selscan 2.0 (Szpiech and Hernandez, 2014). We evaluate the performance of

these unphased statistics under various generic demographic models and compare against the original statistics applied to simulated datasets when phase is either known or unknown.

2 Methods

When the --unphased flag is set in selscan v2.0+, biallelic genotype data is collapsed into multi-locus genotype data by representing the genotype as either 0, 1, or 2—the number of derived alleles observed. In this case, selscan v2.0+ will then compute iHS, nSL, XP-EHH, and XP-nSL as described below. We follow the notation conventions of Szpiech and Hernandez (2014).

2.1 Extended Haplotype Homozygosity

In a sample of n diploid individuals, let $\mathcal{C}$ denote the set of all possible genotypes at locus x_0. For multi-locus genotypes, $\mathcal{C} \coloneqq \{0,1,2\}$, representing the total counts of a derived allele. Let $\mathcal{C}(x_i)$ be the set of all unique haplotypes extending from site x_0 to site x_i either upstream or downstream of x_0. If x_1 is a site immediately adjacent to x_0, then $\mathcal{C}(x_1) \coloneqq \{00,01,02,10,11,12,20,21,22\}$, representing all possible two-site multi-locus genotypes. We can then compute the extended haplotype homozygosity (EHH) of a set of multi-locus genotypes as

$$\mathrm{EHH}(x_i) \sum_{h \in \mathcal{C}(x_i)} \frac{\binom{\mathrm{n_h}}{2}}{\binom{\mathrm{n}}{2}},$$

where n_h is the number of observed haplotypes of type h.

If we wish to compute the EHH of a subset of observed haplotypes that all contain the same 'core' multi-locus genotype, let $\mathcal{H}_c(x_i)$ be the partition of $\mathcal{C}(x_i)$ containing genotype $\mathrm{c} \in \mathcal{C}$ at x_0. For example, choosing a homozygous derived genotype ($\mathrm{c} = 2$) as the core, $\mathcal{H}_2 \coloneqq \{20,21,22\}$. Thus, we can compute the EHH of all individuals carrying a given genotype at site x_0 extending out to site x_i as

$$\mathrm{EH}H_c(x_i) = \sum_{h \in \mathcal{H}_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}},$$

where n_h is the number of observed haplotypes of type h and n_c is the number of observed multi-locus genotypes with core genotype of c. Finally, we can compute the complement EHH of a sample of multi-locus genotypes as

$$\mathrm{cEH}H_c(x_i) = \sum_{h \in \mathcal{C}(x_i) \setminus \mathcal{H}_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_{c'}}{2}},$$

where $n_{c'}$ is the number of observed multi-locus genotypes with a core genotype of not c.

2.2 iHS and nSL

Unphased iHS and nSL are calculated using the equations above. First, we compute the integrated haplotype homozygosity (iHH) for the homozygous ancestral ($c = 0$) and derived ($c = 2$) core genotypes as

$$iHH_c = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\big(EHH_c(x_{i-1}) + EHH_c(x_i)\big)d(x_{i-1}, x_i)$$

$$+ \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\big(EHH_c(x_{i-1}) + EHH_c(x_i)\big)d(x_{i-1}, x_i)\,,$$

where $\mathcal{D}$ is the set of downstream sites from the core locus and $\mathcal{U}$ is the set of upstream sites. $d(x_{i-1}, x_i)$ is a measure of genomic distance between to markers and is the genetic distance in centimorgans or physical distance in basepairs for iHS (Voight, et al., 2006) or the number of sites observed for nSL (Ferrer-Admetlla, et al., 2014). We similarly compute the complement integrated haplotype homozygosity (ciHH) for both homozygous core genotypes as

$$ciHH_c = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\left(cEHH_c(x_{i-1}) + cEHH_c(x_i)\right) d(x_{i-1}, x_i)$$

$$+ \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\left(cEHH_c(x_{i-1}) + cEHH_c(x_i)\right) d(x_{i-1}, x_i).$$

The (unstandardized) unphased iHS is then calculated as

$$iHS = \begin{cases} iHS_2, & if\ iHS_2 > iHS_0 \\ -iHS_0, & otherwise \end{cases},$$

where $iHS_2 = \log_{10}\left(\frac{iHH_2}{ciHH_2}\right)$ and $iHS_0 = \log_{10}\left(\frac{iHH_0}{ciHH_0}\right)$. Conceptually, this is nearly identical to the phased version of iHS, where the log ratio of the integrated haplotype homozygosity is computed between all haplotypes carrying the ancestral allele at the core locus versus all haplotypes carrying the derived allele at the core locus. In this case, however, we compare the iHH of the haplotypes containing homozygous genotypes of one allele at the core locus to the iHH of the haplotypes containing all other genotypes at the core locus. Doing this for both homozygous derived and homozygous ancestral haplotypes separately, we then choose the most extreme value. We assign a positive sign for long low-diversity haplotypes containing the derived homozygous genotype at the core locus, and we assign a negative sign for long low-diversity haplotypes containing the ancestral homozygous genotype at the core locus. Unstandardized iHS scores are then normalized in frequency bins, as previously described (Ferrer-Admetlla, et al., 2014; Voight, et al., 2006). Unstandardized unphased nSL is computed similarly with the appropriate distance measure (see Ferrer-Admetlla, et al. (2014) where they show that nSL can be reformulated as iHS with a different distance measure). Large positive scores indicate long high-frequency haplotypes with a homozygous derived core genotype, and large negative scores indicate long high-frequency haplotypes with a homozygous ancestral core genotype. Clusters of extreme scores in both directions indicate evidence for a sweep.

2.3 XP-EHH and XP-nSL

Unphased XP-EHH and XP-nSL are calculated by comparing the iHH between populations A and B, using the entire sample in each population. iHH in a population P is computed as

$$iHH_P = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\big(EHH(x_{i-1}) + EHH(x_i)\big)d(x_{i-1}, x_i)$$

$$+ \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\big(EHH(x_{i-1}) + EHH(x_i)\big)d(x_{i-1}, x_i),$$

where the distance measure is given as centimorgans or basepairs for XP-EHH (Sabeti, et al., 2007) and number of sites observed for XP-nSL (Szpiech, et al., 2021). The XP statistics between population A and B are then computed as $XP = \log_{10}\left(\frac{iHH_A}{iHH_B}\right)$ and are normalized genome wide. Large positive scores indicate long high-frequency haplotypes in population A, and large negative scores indicate long high-frequency haplotypes in population B. Clusters of extreme scores in one direction indicate evidence for a sweep in that population.

3 Results

We find that the unphased versions of iHS and nSL generally have good power at large sample sizes (Figures 1A, 1B, S1, S7, and S8) to detect selection prior to fixation of the allele, with nSL generally outperforming iHS. In smaller populations (Figures S1C and S1D), power does suffer relative to larger populations (Figures S1A, S1B, S1E, S1F). We note that these statistics struggle to identify soft sweeps when the population is undergoing exponential growth (Figures S1E and S1F). Each of these statistics also have low false positive rates hovering around 1% (Tables S2-S5). These single-population statistics only perform well for relatively large samples (Figures 1A, 1B, and Figures S19, S25, S26, S31, S32, S37, S43, S44, S55, S61, S62).

Similarly, we find that the unphased versions of XP-EHH and XP-nSL have good power as well even for relatively low sample sizes (Figures 1C, 1D, 1G, 1H and Figures S2, S3, S9-S12, S20, S21, S27-30, S38, S39, S45-48, S56, S57, S63-S66). When the sweep takes place in the smaller of the two populations (Figures S2C, S2D, S20C, S20D, S38D, S38C, S56C, and S56D), we see a similar decrease in power, likely related to the lower efficiency of selection in small populations. When one population is undergoing exponential growth (Figures S3, S21, S39, S57) performance is generally quite good, likely the result of a larger effective selection coefficient in large populations. These two-population statistics generally outperform their single-population counterparts, especially at small diploid sample sizes and for sweeps that have reached fixation recently. Each of these statistics also have low false positive rates hovering around 1% (Tables S2-S5).

Next, we turn to comparing the performance of these unphased statistics to their phased counterparts when they are used to analyze either phased data or unphased data. In Figures 1E, 1F, 1G, 1H and Figures S4-S6, S13-S18, S22-S24, S31-S36, S40-S42, S49-S54, S58-S60, S67-S72) we plot the difference in power between the unphased statistics and the phased counterpart applied to data with phase known (red lines) or phase scrambled (blue lines). Where these lines are greater than or equal to 0 indicates that the unphased statistic performed as well as or better than the phased counterpart.

We find that iHS tends to underperform the traditional phased implementations, but nSL tends to perform as well as the phased versions (Figures 1E, 1F and Figures S4, S13, S14, S22, S31, S32, S40, S49, S50, S58, S67, S68). Although we note noticeable drops in unphased nSL power for softer sweeps in exponential growth scenarios (Figures S4F, S13F, S14F, S22F, S31F, S32F, S40F, S49F, S50F, S58F, S67F, S68F) and for sweeps near completion in small

population sizes (Figures S4E, S13E, S14E, S22E, S31E, S32E, S40E, S49E, S50E, S58E, S67E, S68E).

When comparing the unphased versions of XP-EHH and XP-nSL, we find that they consistently perform as well or better than their phased counterparts (Figures 1G, 1H and Figures S5, S6, S17, S18, S23, S24, S35, S36, S41, S42, S53, S54, S59, S60, S71, S72), except in limited circumstances where phase is known, and the sweep is fairly young (sweeping allele at 0.7 frequency) or the divergence time is further in the past.

4 Discussion

We introduce multi-locus genotype versions of four popular haplotype-based selection statistics—iHS (Voight, et al., 2006), nSL (Ferrer-Admetlla, et al., 2014), XP-EHH (Sabeti, et al., 2007), and XP-nSL (Szpiech, et al., 2021)—that can be used when the phase of genotypes is unknown. Although phase would seem to be a critically important component of any haplotype-based method for detecting selection, here we show that, by collapsing haplotypes into derived allele counts (thus erasing phase information), we can achieve similar power to using this information. We observed that single-population statistics such as iHS and nSL require relatively large diploid sample sizes ($n >= 100$ for iHS, $n >= 50$ for nSL), but the two-population statistics XP-EHH and XP-nSL perform well even for diploid sample sizes down to $n = 10$ per population. This follows other work that has shown similar patterns with other haplotype-based statistics for detecting selection (DeGiorgio and Szpiech, 2022; Harris and DeGiorgio, 2020; Harris, et al., 2018; Klassmann and Gautier, 2022). Importantly, this approach now opens up the application of several popular haplotype-based selection statistics (based on extended haplotype homozygosity) to more species where phase information is challenging to know or infer.

For ease of use of these new unphased versions of iHS, nSL, XP-EHH, and XP-nSL, we implement these updates in the latest v2.0 update of the program selscan (Szpiech and Hernandez, 2014), with source code and pre-compiled binaries available at https://www.github.com/szpiech/selscan.

5 Methods

5.1 Simulations

We evaluate the performance of the phased and unphased versions of iHS, nSL, XP-EHH, and XP-nSL under a generic two-population divergence model using the coalescent simulation program discoal (Kern and Schrider, 2016). We explore five versions of this generic model and name them Demo 1 through Demo 5 (Table S1). Let N_0 and N_1 be the effective population sizes of Population 0 and Population 1 after the split from their ancestral population (of size N_A). For Demo 1, we keep a constant population size post-split and let $N_0 = N_1 = 10{,}000$. For Demo 2, we keep a constant population size post-split and let $N_0 = 2N_1 = 10{,}000$. For Demo 3, we keep a constant population size post-split and let $2N_0 = N_1 = 10{,}000$. For Demo 4, we initially set $N_0 = N_1 = 10{,}000$ and let N_0 grow stepwise exponentially every 50 generations starting at 2,000 generations ago until $N_0 = 5N_1 = 50{,}000$. For Demo 5, we initially set $N_0 = N_1 = 10{,}000$ and let N_1 grow stepwise exponentially every 50 generations starting at 2,000 generations ago until $5N_0 = N_1 = 50{,}000$.

For each demographic history we vary the population divergence time $t_d \in \{2000,\ 4000,\ 8000\}$ generations ago. For non-neutral simulations, we simulate a sweep in Population 0 in the middle of the simulated region across a range of selection coefficients $\mathrm{s} \in \{0.005,\ 0.01,\ 0.02\}$. We vary the frequency at which the adaptive allele starts sweeping as $\mathrm{e} \in \{0,\ 0.01,\ 0.02,\ 0.05,\ 0.10\}$, where $\mathrm{e} = 0$ indicates a hard sweep and $\mathrm{e} > 0$ indicates a soft

sweep, and we also vary the frequency of the selected allele at time of sampling $f \in \{0.7,\ 0.8,\ 0.9,\ 1.0\}$ as well as $g \in \{50,\ 100\}$ representing fixation of the sweeping allele g generations ago. For all simulations we set the genome length to be $L = 500{,}000$ basepairs, the ancestral effective population size to be $N_A = 10{,}000$, the per site per generation mutation rate at $\mu = 2.35 \times 10^{-8}$, and the per site per generation recombination rate at $r = 1.2 \times 10^{-8}$. For neutral simulations, we simulate 1,000 replicates for each parameter set, and for non-neutral simulations we simulate 100 replicates for each parameter set. We sample $2n \in \{200, 100, 40, 20\}$ haplotypes, randomly paired together to form $n \in \{100, 50, 20, 10\}$ diploid individuals, from each population for analysis. These data sets represent the case where phase is known perfectly. We also create a set of "unphased" data sets from these phased data sets by swapping the alleles of each heterozygote to the opposing haplotype with probability 0.5.

As iHS and nSL are single population statistics, we only analyze Demo 1, Demo 3, and Demo 4 with these statistics, as Demo 2 and Demo 5 have a constant size history identical to Demo 1 for Population 0, where the sweeps are simulated. For XP-EHH and XP-nSL we analyze all five demographic histories.

For all simulations, we compute the relevant statistics (--ihs, --nsl, --xpehh, or --xpnsl) with selscan v2.0 using the --trunc-ok flag. We set --unphased when computing the unphased versions of these statistics, and we do not set it when computing the original phased versions. For iHS and XP-EHH, we also use the --pmap flag to use physical distance instead of a recombination map.

5.2 Power and False Positive Rate

Here we evaluate the power and false positive rate for the unphased version of iHS, nSL, XP-EHH, and XP-nSL. For comparison, we also compute the power for the original phased

versions of these statistics in two different ways. We compute the phased statistics for a set of simulated datasets where perfect phase is known, and we compute them again for a set of simulated datasets where we destroy phase information (see section 5.1). As the unphased statistics collapse genotypes into derived allele counts, there is no functional difference between these two datasets for these statistics. We compute power in the same way for each statistic regardless of underlying dataset analyzed as described below.

To compute power for iHS and nSL, we follow the approach of Voight, et al. (2006). For these statistics, each non-neutral replicate is individually normalized jointly with all neutral replicates with matching demographic history in 1% allele frequency bins. Because extreme values of the statistic are likely to be clustered along the genome (Voight, et al., 2006), we then compute the proportion of extreme scores ($|iHS| > 2$ or $|nSL| > 2$) within 100kbp non-overlapping windows. We then bin these windows into 10 quantile bins based on the number of scores observed in each window and call the top 1% of these windows as putatively under selection. We calculate the proportion of non-neutral replicates that fall in this top 1% as the power. To compute the false positive rate, we compute the proportion of neutral simulations that fall within the top 1%.

To compute power for XP-EHH and XP-nSL, we follow the approach of (Szpiech, et al., 2021). For these statistics, each non-neutral replicate is individually normalized jointly with all matching neutral replicates. Because extreme values of the statistic are likely to be clustered along the genome (Szpiech, et al., 2021), we then compute the proportion of extreme scores (XP-EHH > 2 or XP-nSL > 2) within 100kbp non-overlapping windows. We then bin these windows into 10 quantile bins based on the number of scores observed in each window and call the top 1% of these windows as putatively under selection. We calculate the proportion of non-

neutral replicates that fall in this top 1% as the power. To compute the false positive rate, we compute the proportion of neutral simulations that fall within the top 1%.

6 Acknowledgements

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM146926 and by start-up funds from the Pennsylvania State University's Department of Biology. Computations for this research were performed using the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer.

References

Browning, B.L., *et al.* Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* 2021;108(10):1880-1890.

Campagna, L. and Toews, D.P.L. The genomics of adaptation in birds. *Curr Biol* 2022;32(20):R1173-R1186.

Colonna, V., *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol* 2014;15(6):R88.

Crawford, N.G., *et al.* Loci associated with skin pigmentation identified in African populations. *Science* 2017;358(6365).

DeGiorgio, M. and Szpiech, Z.A. A spatially aware likelihood test to detect sweeps from haplotype distributions. *PLoS Genet* 2022;18(4):e1010134.

Delaneau, O., Zagury, J.F. and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013;10(1):5-6.

Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol* 2014;29(1):51-63.

Fagny, M., *et al.* Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Mol Biol Evol* 2014;31(7):1850-1868.

Ferrer-Admetlla, A., *et al.* On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* 2014;31(5):1275-1291.

Harris, A.M. and DeGiorgio, M. A likelihood approach for uncovering selective sweep signatures from haplotype data. *Mol Biol Evol* 2020.

Harris, A.M., Garud, N.R. and DeGiorgio, M. Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. *Genetics* 2018;210(4):1429-1452.

Kern, A.D. and Schrider, D.R. Discoal: flexible coalescent simulations with selection. *Bioinformatics* 2016;32(24):3839-3841.

Klassmann, A. and Gautier, M. Detecting selection using extended haplotype homozygosity (EHH)-based statistics in unphased or unpolarized data. *PLoS One* 2022;17(1):e0262024.
Lu, K.*, et al.* Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat Commun* 2019;10(1):1154.
Meier, J.I.*, et al.* Genomics of Parallel Ecological Speciation in Lake Victoria Cichlids. *Mol Biol Evol* 2018;35(6):1489-1506.
Nedelec, Y.*, et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* 2016;167(3):657-669 e621.
Sabeti, P.C.*, et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007;449(7164):913-918.
Salmon, P.*, et al.* Continent-wide genomic signatures of adaptation to urbanisation in a songbird across Europe. *Nat Commun* 2021;12(1):2983.
Schrider, D.R. Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics* 2020;216(2):499-519.
Szpiech, Z.A. and Hernandez, R.D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* 2014;31(10):2824-2827.
Szpiech, Z.A.*, et al.* Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *Evol Lett* 2021;5(4):408-421.
Voight, B.F.*, et al.* A map of recent positive selection in the human genome. *Plos Biol* 2006;4(3):e72.
Zhang, S.J.*, et al.* Genomic regions under selection in the feralization of the dingoes. *Nat Commun* 2020;11(1):671.
Zoledziewska, M.*, et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat Genet* 2015;47(11):1352-1356.

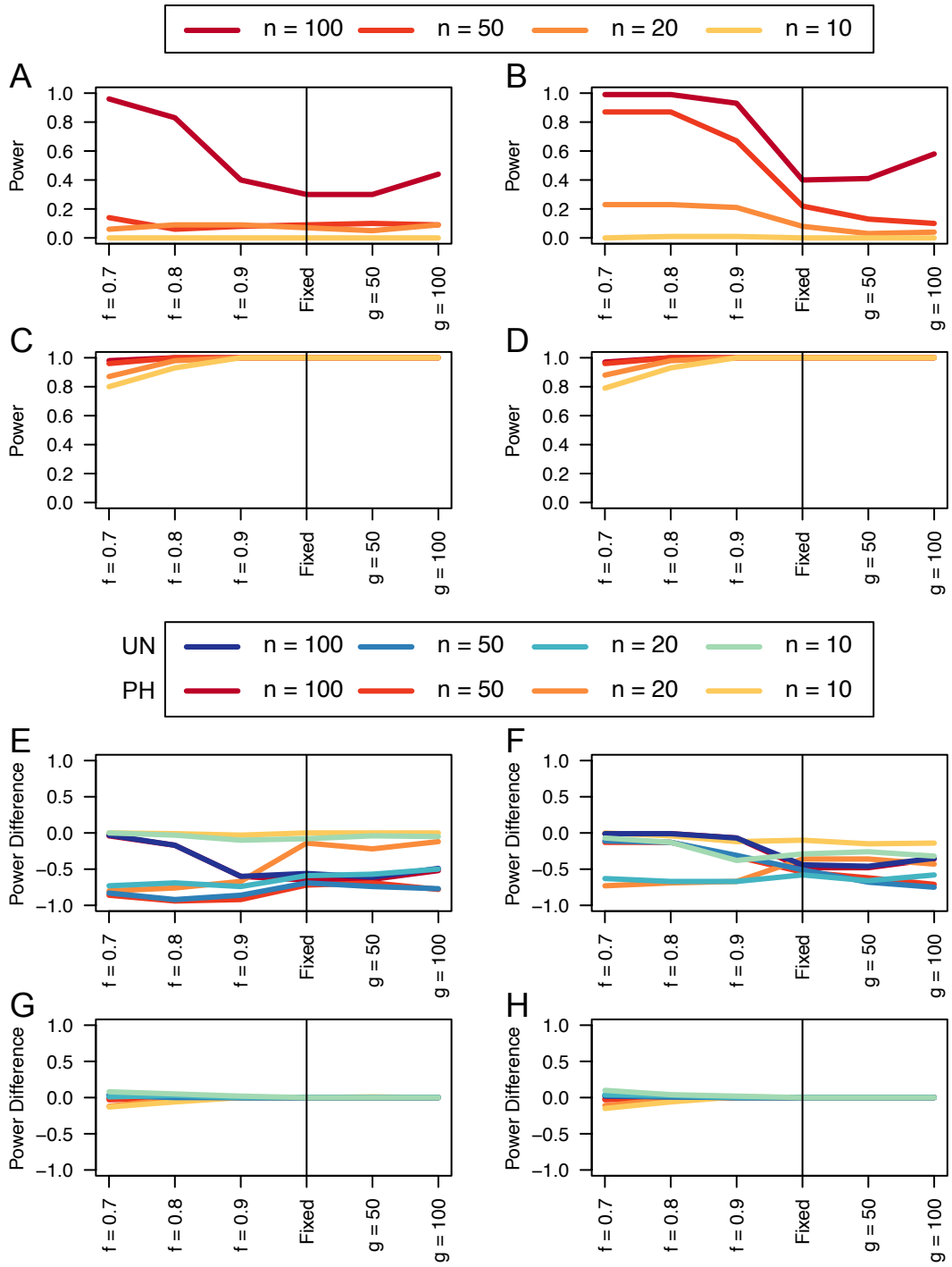

Figure 1. Power curves for unphased implementations of iHS (A), nSL (B), XP-EHH (C), and XP-nSL (D), and power difference between unphased implementations of iHS (E), nSL (F), XP-EHH (G), and XP-nSL (H) and phased implementations. Blue curves represent the power difference between the unphased and phased statistics when applied to unphased data (UN). Red curves represent the power difference between the unphased and phased statistics when applied to perfectly phased data (PH). Values greater than 0 indicate the unphased statistic had higher power. All panels represent analyses with demographic history Demo 1 and n = 100, 50, 20, or 10 diploid samples. For these plots the selection coefficient is set at $s = 0.01$, the frequency at which selection began is set at $e = 0$ (i.e., a hard sweep),

and the divergence time in generations is set at $t_d = 2000$. f is the frequency of the adaptive allele at time of sampling, g is the number of generations at time of sampling since fixation.