Detecting selection using *Extended Haplotype Homozygosity (EHH)*-based statistics on unphased or unpolarized data

A. Klassmann[a,*], R. Vitalis[b], M. Gautier[b]

[a] *Institut für Genetik, Universität zu Köln, 50674 Köln, Germany*
[b] *CBGP, Univ Montpellier, CIRAD, INRAE, IRD, Montpellier SupAgro, Montpellier, France*

Abstract

Analysis of population genetic data often includes the search for genomic regions with signs of recent positive selection. One of the approaches involves the concept of *Extended Haplotype Homozygosity* and its associated statistics. These statistics presume that haplotypes are phased and some of them that variants are polarized. Here we assess the consequences if one of the two conditions is not fulfilled. We find that phasing information is indispensable for the accurate estimation of within-population statistics and, if sample sizes are small, for cross-population statistics, too. Ancestry information, in contrast, is of lesser importance for both. We make use of a publicly available update of our R package REHH which, among other features, incorporates the adapted statistics presented here.

keywords: selective sweeps, neutrality tests, *EHH*, phasing, variant polarization

1. Introduction

The ease with which genomic sequences can be obtained constrasts sharply with the challenge to discern their functional elements. Finding molecular signatures of recent selection can help prioritizing regions for further investigation. Characterizing such signatures generally requires statistical tests which rely on

*Corresponding author

the null hypothesis of neutral evolution. Here we focus on the classical case of detecting recent strong positive selection in form of a hard "selective sweep". Differential selection across populations can be detected by means of classical statistics such as F_{ST} (Akey, 2002), but may be corroborated by more recent approaches which exploit other aspects of the selection signal. In contrast, the detection of selection within a population requires sophisticated methods and among the many methods proposed, the following three belong to the most commonly applied:

- *Tajima's D* (Tajima, 1989), *Fay & Wu's H* (Fay and Wu, 2000) and related metrics (Achaz, 2009) compare the observed site frequency spectrum of a genomic region with its expectation under neutrality. They were intended for regions short enough to neglect recombination. Although easy to apply and fast to compute, they are highly vulnerable to the confounding effects of demography and population structure. They are implemented in various software such as MEGA (Kumar *et al.*, 2018), dnasp (Rozas *et al.*, 2017) and the R package PopGenome (Pfeifer *et al.*, 2014).

- Sweepfinder (Nielsen *et al.*, 2005; DeGiorgio *et al.*, 2016) and SweeD (Pavlidis *et al.*, 2013) compare multiple frequency spectra in consecutive windows taking into account that recombination gradually erodes the signal of selection with increasing genetic distance. The likelihood of such a sequence of varying spectra is compared with either the background average or a model-derived "neutral" frequency spectrum.

- Sabeti *et al.* (2002) introduced the concept of *Extended Haplotype Homozygosity (EHH)* on top of which Voight *et al.* (2006) built a statistic called *iHS*, with later variations by Sabeti *et al.* (2007) and Tang *et al.* (2007). The quantity measures the decay of linkage around a specific site due to both recombination and mutations. *iHS* was first implemented in an eponymous program by the authors themselves (Voight *et al.*, 2006). Subsequently improved implementations have been Selscan (Szpiech

and HERNANDEZ, 2014), HAPBIN (MACLEAN *et al.*, 2015) and the R package REHH (GAUTIER and VITALIS, 2012; GAUTIER *et al.*, 2017).

In our view, there are two major points that distinguish *EHH*-based from frequency spectrum approaches:

- the latter are targeted at *completed* selective sweeps while *EHH* is more appropriate for the detection of *on-going* selective sweeps. At least in the human species, completed selective sweeps seem to be rare (HERNANDEZ *et al.*, 2011) and well-known examples of selection, such as variants near the *LCT* gene, extending the expression of the lactase enzyme into adulthood, are yet far from fixation (VITTI *et al.*, 2013).

- frequency spectrum based methods assign by definition a score to genomic intervals while *EHH*-based statistics assign a score to a precise site. Although selection is usually inferred only if sites with conspicuous scores cluster together, sites with a low score within such a cluster are unlikely to have driven selection.

The methods listed above, except *Tajima's D* in its original version, assume that alleles are *polarized*, i.e. that the ancestral vs. derived state of each allele is known. Polarization is typically achieved by using an outgroup: if a homologous site is monomorphic in the outgroup and coincides with one of the alleles in the investigated population, then that variant is called ancestral. However, an outgroup species needs to be chosen properly: if on the one hand the outgroup is phylogenetically too distant, then the probability of multiple mutations is high; if on the other hand the outgroup is too close then the probability of shared polymorphisms is high. Both situations lead to a mis-specified ancestry status (BAUDRY and DEPAULIS, 2003; HERNANDEZ *et al.*, 2007). Furthermore there must be a reference genome of that species available. Even so, the genomes of the outgroup and the focal species may not completely overlap, thereby leaving unpolarized chunks. For example, although considerable effort has been undertaken to infer the "ancestral sequence" of present-day humans, about 4% of the

SNPs found by the 1000 genome project cannot be polarized (see the Methods section below). Lack of ancestry information can be expected to entail a loss of statistical power for all statistics which fully or partially rely on it. Curiously, this has been overlooked so far.

In addition to polarization, the calculation of *EHH* as described by Sabeti *et al.* (2002) requires data to be *phased*, i.e. that all variants are assigned to each of the two homologous chromosomes of a diploid individual. While phased haplotypes are expensive to obtain experimentally, computational methods to infer them probabilistically often yield satisfactory results (Browning and Browning, 2011). In cases where phasing is not possible, the original method to calculate EHH cannot be applied. Attempts have been made to estimate extended haplotype homozygosity using unphased chromosomes of diploid individuals: Wang *et al.* (2006) for a within-population and Tang *et al.* (2007) for a cross-population test. Both assessed the power by simulations, however they did not compare directly phased and unphased estimators; the latter reported merely a correlation coefficient r of 80-85% for their statistic on a specific experimental data set.

The aim of this article is to investigate in more detail the consequences of calculating *EHH*-based statistics on unphased or unpolarized data. We first recapitulate the definition of three statistics which we want to investigate. Then, we present how the statistics must be adapted for unphased and/or unpolarized data. Finally we compare the regions found by original and modified statistics on simulated as well as experimental data. Along the way, we aim at providing potential users with some intuition for the various quantities involved.

2. Materials and Methods

2.1. The definition of the statistics iHS*,* XP-EHH *and* Rsb

The designation *haplotype* is sometimes used for both a specific sequence and the pattern of variants it contains. In the following, we use that term exclusively for the latter meaning and denote by *sequence* or *chromosome* an individual instance. Let s denote a site of interest within a genome. We call s

the *focal marker* (equivalent to *primary locus* in WANG *et al.* (2006)) and its variants *core alleles*. Let n_a refer to the number of sequences with core allele a and $n_s \equiv \sum_a n_a$ the total number of sequences. If there are no missing data at the focal marker, then n_s equals the sample size n. Sequences sharing a core allele are by definition homozygous at the focal marker. The *Extended Haplotype Homozygosity (EHH)* measures the decay of this homozygosity with increasing distance to the marker and is calculated independently in each direction (upstream/downstream) from the marker. More precisely, let t be another marker on the same chromosome and consider the genomic region between s and t. Let $K_{s,t}$ denote the number of distinct haplotypes in that region and $K^a_{s,t}$ the subset of haplotypes that carry allele a at the focal marker s. n_k refers to the number of sequences sharing haplotype k. The quantity EHH as defined by SABETI *et al.* (2002) is calculated for chromosomes carrying a specific core allele a and is denoted correspondingly by EHH^a. It is used to contrast two (or more) alleles within a population and estimated by:

$$\mathrm{EHH}^a_{s,t} = \frac{1}{n_a(n_a-1)} \sum_{k=1}^{K^a_{s,t}} n_k(n_k-1) \; . \tag{1}$$

In order to summarize $EHH^a_{s,t}$ to a single number assignable to the allele a at site s, VOIGHT *et al.* (2006) opted for an integration of EHH and named the resulting quantity *integrated Haplotype Homozygosity (iHH)*:

$$\mathrm{iHH}^a(s) = \int EHH^a_{s,t}\, dt \; . \tag{2}$$

The integration is performed numerically using the *trapezoidal rule*, a simple standard algorithm. However, its value is influenced by three technical parameters. First, in order to reduce statistical noise, integration is stopped when EHH, decreasing with distance to the focal marker, reaches a lower threshold or cut-off, usually chosen as 0.05. As is discussed subsequently, this parameter can be used to reduce statistical noise when dealing with unphased data. Second, the lower horizontal bound of the integrated area can either coincide with the cut-off (as by default in REHH) or with the x-axis (as by default in

hapbin). In fact, VOIGHT *et al.* (2006) adjusted additionally for sample size n by subtracting $\frac{1}{n}$ from the above cut-off value. SABETI *et al.* (2007) found this parameter to be of minor importance and we use the default value of REHH throughout our study. Third, integration can be downweighted, stopped or even discarded, if gaps between consecutive markers are encountered which most often arise from missing data at particular genomic structures like centromeres or highly repetitive regions. If not accounted for, gaps can lead to highly inflated estimates of Extended Haplotype Homozygosity. Although of great practical relevance, this issue is not a topic of the present study neither.

Given iHH for ancestral (A) and derived (D) alleles of a focal marker, VOIGHT *et al.* (2006) favoured a log-ratio for comparison of them, yielding the (as-yet unstandardized) *integrated Haplotype Homozygosity Score*

$$\text{uniHS}(s) = \ln\left(\frac{\text{iHH}^A(s)}{\text{iHH}^D(s)}\right) . \tag{3}$$

Finally, this quantity is standardized:

$$\text{iHS}(s) = \frac{\text{uniHS}(s) - \text{mean}(\text{uniHS}|p_s)}{\text{sd}(\text{uniHS}|p_s)} . \tag{4}$$

Since the expectated values under neutrality of *uniHS* depend strongly on the derived allele frequency p_s at the focal marker s, the standardization is ideally performed separately for each group of focal markers with the exact same frequency. In practice, the standardization is carried out over small frequency bins. VOIGHT *et al.* (2006) showed that this quantity follows approximately a standard normal distribution.

In order to detect selection using *iHS*, both alleles of a site must be present on enough sequences to reliably estimate their respective EHH^a. This limits the application to sites with a certain minor allele frequency, usually chosen as 5%, excluding particular variants near fixation. This limitation can be overcome if a second population with different allele frequencies can be used for comparison. SABETI *et al.* (2007) and TANG *et al.* (2007) independently created a new statistic by omitting the above distinction between core alleles. While SABETI *et al.* (2007) kept the designation *EHH*, we follow TANG *et al.* (2007) in denoting

the new quantity as site-specific *EHH*, abbreviated by *EHHS*. Its definition is given by

$$\mathrm{EHHS}_{s,t} = \frac{1}{n_s(n_s-1)} \sum_{k=1}^{K_{s,t}} n_k(n_k-1) \; . \tag{5}$$

Note that $EHHS_{s,s}$ is an estimate for the focal marker homozygosity. The subsequent statistics are build analogously to eq. (2-4). Sabeti *et al.* (2007) first integrated this quantity to yield *integrated EHHS (iES)*

$$\mathrm{iES}(s) = \int EHHS_{s,t}\, dt \; , \tag{6}$$

which is then compared between two populations to obtain the as-yet unstandardized *XP-EHH*

$$\text{unXP-EHH}(s) = \ln\left(\frac{\mathrm{iES}_{\mathrm{pop1}}(s)}{\mathrm{iES}_{\mathrm{pop2}}(s)}\right) \; , \tag{7}$$

which, in turn, is standardized to yield

$$\text{XP-EHH}(s) = \frac{\text{unXP-EHH}(s) - \text{mean(unXP-EHH)}}{\text{sd(unXP-EHH)}} \; . \tag{8}$$

Tang *et al.* (2007) defined a similar quantity that we call for distinction the *integrated normalized EHHS score*

$$\mathrm{inES}(s) = \frac{1}{EHHS_{s,s}} \int EHHS_{s,t}\, dt = \frac{iES(s)}{EHHS_{s,s}} \tag{9}$$

to obtain first the (unstandardized) *Ratio of integrated (normalized) EHHS between populations (Rsb)* [1]

$$\mathrm{unRsb}(s) = \ln\left(\frac{\mathrm{inES}_{\mathrm{pop1}}(s)}{\mathrm{inES}_{\mathrm{pop2}}(s)}\right) \; , \tag{10}$$

and, finally, standardizing by the median instead of the mean,

$$\mathrm{Rsb(s)} = \frac{\mathrm{unRsb}(s) - \mathrm{median(unRsb)}}{\mathrm{sd(unRsb)}} \; . \tag{11}$$

[1] Note that for the sake of uniformity our notation differs slightly from that given in Tang *et al.* (2007), where (11) was referred to as *ln(Rsb)* and the unlogarithmized value used only for plotting.

Importantly, for standardization of the cross-population statistics *XP-EHH* and *Rsb* no binning with respect to core allele frequencies is undertaken and thus no variant polarization is needed.

2.2. Adaptations for unphased sequences

The homozygosity of a population can not only be estimated by the pairwise comparison of all sequences in a sample (as formulated above), but also by the fraction of homozygous (diploid) individuals, assuming Hardy-Weinberg equilibrium. The latter does not require phase information and Wang *et al.* (2006) and Tang *et al.* (2007) applied it on EHH resp. $EHHS$: the crucial difference to eq. (1) resp. (5) is that only the two chromosomes of each individual are compared and can share a haplotype with one another. A shared haplotype is thus associated with a specific individual and for its determination no phase information is necessary. The quantities EHH and $EHHS$ are then estimated as above by the fraction of shared haplotypes among all sequence comparisons. In Figure 1 the two ways of estimating EHH are explained by an example.

Importantly, for any focal marker s, only the chromosomes of individuals homozygous at that position are taken into account. The extension step to either side is performed for each individual independently. When the first heterozygous marker is found, the shared haplotype ends. Let $I_{s,t}$ denote the number of individuals homozygous in the region between s and t and $I^a_{s,t}$ those among them that carry the core allele a. EHH^a resp. $EHHS$ at marker t is then estimated by the fraction of individuals homozygous until marker t:

$$\mathrm{EHH}^a_{s,t} = \frac{I^a_{s,t}}{I^a_{s,s}} \tag{12}$$

$$\mathrm{EHHS}_{s,t} = \frac{I_{s,t}}{I_{s,s}} \ . \tag{13}$$

[Figure 1 about here.]

All subsequent steps to obtain *iHS*, *XP-EHH* and *Rsb* remain the same as above. Since *EHHS* calculated by eq. (13) is normalized (equals to 1 at the focal

marker), for unphased data *XP-EHH* is essentially identical to *Rsb*; the only difference consists in the use of median resp. mean in the standardization step. Note that the integration of unphased *EHH* resp. *EHHS* yields essentially the average length of shared haplotypes; it would do so exactly, if the decay of *EHH* resp. *EHHS* would not be linearly interpolated between consecutive markers (see the right panel of Figure 1).

Discarding all heterozygous individuals entails that the number of sequences available for estimating EHH resp. $EHHS$, to which we will refer as *effective sample size*, may be small or even zero. In an attempt to reduce statistical noise, Wang *et al.* (2006) strengthened the above mentioned requirement of a minor (sample) allele frequency (MAF) of at least 5% for focal alleles by requiring that the sequences of homozygous individuals for each allele represent more than 5% of the sample. We modified this condition and require, in addition to a MAF of 5%, a fixed minimum number of 10 sequences (5 homozygous individuals) for each allele.

2.3. Adaptations for unpolarized variants

There is only one step where the information of allele ancestry status is exploited, namely the standardization of $uniHS$ in eq. (4), depending on the frequency of the derived core allele. In order to avoid an arbitrary assignment of ancestry status, we replace the ancestral resp. derived allele in eq. (3) by major (most frequent) and minor (second most frequent) allele, resp.

$$\mathrm{uniHS}(s) = \ln\left(\frac{\mathrm{iHH}^{MAJ}(s)}{\mathrm{iHH}^{MIN}(s)}\right) . \tag{14}$$

For unpolarized variants, the frequency-dependence of *EHH* under neutrality cannot be accounted for by a binning with respect to MAF, because such a binning would group derived alleles of frequency p_s together with those of frequency p_{1-s} whose respective expected values differ increasingly with increasing $|0.5 - s|$. Hence, we suggest standardization to be performed without consideration of allele frequencies:

$$\mathrm{iHS}(s) = \frac{\mathrm{uniHS}(s) - \mathrm{mean}(\mathrm{uniHS})}{\mathrm{sd}(\mathrm{uniHS})} . \tag{15}$$

2.4. Delineation of regions under selection

VOIGHT *et al.* (2006) showed that individual markers with outlier values are much less indicative of selective sweeps than a clustering of those (see Figure S2 of the Supporting Information in VOIGHT *et al.* (2006)). In effect, they delineated candidate regions of selection by requiring half of markers in such a region to fall into the 99% genome-wide percentile. We follow this approach; more precisely, we used overlapping sliding windows of width 250,000 base pairs (bp) with an offset of 50,000 bp. A window was called a candidate region, if at least 50% of markers exceeded a given threshold. Overlapping candidate windows were merged. For experimental data we required the number of markers in any window to exceed the (arbitrary) value of 150 in order to exclude regions with few genotyped markers; if phase information was ignored, this number was halved, corresponding to a similar decrease of exploitable markers. In order to evaluate our adapted statistics we did not use a fixed threshold but instead fixed the number of delineated regions by setting the thresholds accordingly. We assessed the power of the adapted statistics by the overlap with regions called by the original statistics.

2.5. Simulated data

We performed coalescent simulations using *msms* (EWING and HERMISSON, 2010). We assumed an effective population size of $N_e \approx 10,000$ for humans. In previous studies both population scaled mutation rate and recombination rate were set as $\theta = \rho = 0.001$ per base per generation (GUTENKUNST *et al.*, 2009; CRISCI *et al.*, 2013). Assuming a mutation-drift equilibrium this number translates to a mutation resp. recombination rate of $\mu = r = 10^{-3}/(4 \cdot 10,000) = 2.5 \cdot 10^{-8}$ per base per generation. However, experimental estimations of both rates yielded about half that value: $\mu \approx 1.3 \cdot 10^{-8}$ (JÓNSSON *et al.*, 2017), $r \approx 1.2 \cdot 10^{-8}$ (DUMONT and PAYSEUR, 2008). We employed hence population-scaled rates of $\theta = \rho \approx 5 \cdot 10^{-4}$ per base per generation. In our simulations we set $\theta = 25,000$ and $\rho = 25,000$ to mimic a genetic region of 50 Megabases. This large size proved necessary to reduce boundary effects because, as we show,

shared haplotypes can span 10 Megabases and more. It is known that most recombination events occur within hot spots (McVean *et al.*, 2004), however *msms* cannot handle varying recombination rates and other tools which can (e.g. *msHOT* (Hellenthal and Stephens, 2007)), are not able to simulate selection. The output format for site positions has been set to the highest precision in order to avoid different sites having the same (rounded) position. The commands have been (with appropriate modifications of sample sizes where indicated)

- for a sample from a single, neutrally evolving population:

```
msms 100 1 -N 10000 -t 25000 -r 25000 50000000
-oformat "#.################"
```

- for equal sized samples from two neutrally evolving populations that split symmetrically from an ancestral population $4N_e \cdot 0.05$ generations ago, without subsequent migration. The split time corresponds roughly to $50,000$ years in humans:

```
msms 200 1 -N 10000 -t 25000 -r 25000 50000000 -I 2 100 100 -ej 0.05 1 2
-oformat "#.################"
```

- for a sample from a population experiencing a single on-going selective sweep while otherwise evolving neutrally. The selected allele was set as dominant with a population-scaled selection coefficient of $2N_e s = 500$, having reached at sampling time a population frequency of 50% and located exactly at the center of the simulated chromosome (position 0.5). The selected site itself is included into the output:

```
msms 200 1 -N 10000 -t 25000 -r 25000 50000000 -SAA 500 -SaA 500
-SF 0 0.5 -Sp 0.5 -Smark -oformat "#.################"
```

Since the polymorphic sites simulated by *msms* have positions in the interval (0,1], they were multiplied by $5 \cdot 10^7$ in order to yield statistics comparable to human population genetic data.

2.6. Experimental data

We used data of Lowy-Gallego *et al.* (2019) who called variants on re-aligned reads from the *1000 Genomes Project* (The 1000 Genomes Project Consortium, 2015) to human reference genome assembly GRCh38. The data comprise only autosomes and contain fully phased bi-allelic SNPs with imputed

missing values. The ancestral alleles, inferred by an alignment of 12 primates, were obtained from ENSEMBL release 91 (Zerbino *et al.*, 2018). Almost 91% of the 73 million SNPSs are covered by ancestral states of "high confidence" and a further 6% of "low confidence"; using both, about 95.8% of SNPs can be polarized. We used samples of European origin (`CEU`, `GBR`), Asian origin (`CHB`, `CHS`, `JPT`) and African origin (`YRI`), see Table 1.

The calculation of the statistics for both simulated and experimental data has been performed using R-package rehh version 3.1 with options `maxgap = 100.000` and `discard_integration_at_border = FALSE`. The latter option means that integration is stopped at gaps and sequence borders, but the value not discarded. For unphased estimation, integration was stopped when *EHH* resp. *EHHS* reached 0.1 or less than 4 sequences remained homozygous.

[Table 1 about here.]

3. Results

3.1. Simulated neutral evolution

We first investigate the expected values of the original statistics *iHS*, *XP-EHH* and *Rsb* in neutrally evolving populations. Then, we calculate the correlation of these with their counterparts for unphased data. We present the distributions of the modified statistics and finally compare them on simulated selective sweeps.

3.1.1. Dependence on core allele frequencies

In this section we assess dependence of the three statistics on the frequency of the derived core allele p_s. We assume that the data are both phased and polarized. For $uniHS$ this was already reported by Voight *et al.* (2006) (see their Figure 4). However, we wanted to exclude that the observed dependence is an artifact due to the different amount of ancestral and derived sequences used for the calculation of iHH^A resp. iHH^D at each frequency p_s. We selected for each focal marker a random subsample with an equal number of ancestral and derived sequences and recomputed the statistics. Figure S1 shows the resulting

unstandardized *iHS* values from simulations of a neutrally evolving population: the values for the subsamples are virtually identical to those calculated from the complete sample, suggesting that *uniHS* indeed depends on the population frequency of the derived core allele and not its sample frequency.

Since the effect of p_s is greater than the range of values seen for each fixed p_s, disregarding this dependence in the standardization of unpolarized data can be expected to influence the final score considerably.

The cross-population statistics *XP-EHH* and *Rsb* are defined symmetrically with respect to the compared populations and consequently the expected values have to be zero for markers with the same derived allele frequency, assuming identical demographies for the two populations. We simulated two populations of equal size that recently split from a common ancestral population without subsequent migration. Figures 2 and 3 show the averaged unstandardized *XP-EHH* resp. *Rsb* values depending on the derived core allele frequency in each population. The values follow a varied pattern which differs between the two statistics. This dependence was neither reported by Sabeti *et al.* (2007) nor Tang *et al.* (2007) and consequently not accounted for in the standardization step. Fortunately, the effect is much smaller than for the *uniHS* statistics, making its consideration less stringent. Furthermore, a frequency-dependent standardization in the vein of *iHS* would need two-dimensional bins and is likely to be too conservative because high allele frequency differences between populations often are indicative of differential selection. In other words, contrary to *iHS*, the implicit assumption that each bin is dominated by neutral variants almost certainly does not hold. Hence in lack of a better solution we continue to use these statistics as they are. Note, however, that any such hypothetical bin-wise standardization would make *XP-EHH* and *Rsb* essentially identical, except for the respective usage of the mean and the median in eqs. (8) and (11).

[Figure 2 about here.]

[Figure 3 about here.]

3.1.2. Correlation between phased and unphased data

The estimators of *EHH* and *EHHS* on unphased data by eq. (12) evaluate only sequences belonging to individuals that are homozygous for the focal marker. Furthermore, exclusively the two chromosomes from each individual are compared to define shared haplotypes, forming a subset of all possible cross-individual comparisons in phased data. Consequently the number of comparisons used for the estimation of *EHH* and *EHHS* is considerably reduced, thereby increasing the variance of the estimates (right panel of Figure S1. The problem is exacerbated by the fact that the distribution of the lengths of shared haplotypes is very uneven, as shown by Figure 4 for phased data: a few very long shared haplotypes (extending over 10 Megabases in our simulations) increase disproportionately the maximum range of haplotypes, hence the boundaries for the integration step (eqs. 2, 6, and 9).

[Figure 4 about here.]

[Figure 5 about here.]

The increased variance is most problematic for the estimation of *EHH* on the minor allele, relying on possibly very few sequences. In order to mitigate the problem we tuned two parameters: first, in addition to demand a minor core allele frequency of 5% we require a minimum number of 10 evaluated sequences (5 homozygous individuals). Although this might seem a mild criterion, assuming Hardy-Weinberg genotype proportions, the condition on average entails the discard of markers with a MAF below $\sqrt{0.1} \approx 31\%$, resp. $\sqrt{0.05} \approx 22\%$ in samples of size $n = 100$ resp. 200.

Second, we analyzed the effect of changing the cut-off value at which integration of *EHH* or *EHHS* stops. For unphased estimation this value corresponds to the fraction of sequences which remain homozygous. We increased this value from its default of 0.05 to 0.1. Additionally, we required an absolute number of at least 4 homozygous sequences (2 homozygous individuals). Both conditions cap the contribution of the longest extended haplotypes, yielding a considerably

stronger correlation with *iHH* values obtained by phased estimation (Figure 5). The latter requirement, though, introduces a bias in the sense that it entails a relatively stronger decrease of *iHH* for alleles with low frequency (compare right panel of Figure S1).

The situation is somewhat different for the cross-population statistics. A requirement of a minimum number of evaluated chromosomes (now comprising all core alleles) appears to be of lesser importance. In fact, in our simulations with sample sizes of $n = 100$ for each population, even a requirement of at least 50 chromosomes (25 homozygous individuals) excluded less than 1% of focal markers, so that we did not apply any such condition. For smaller sample sizes, though, a minimum requirement of 20-30 evaluable sequences may still be reasonable.

Analogously to Figure 5, increasing the cut-off for integration from the default 0.05 to 0.1 increases the correlation between phased and unphased *inES* from 69% to 85% in two-population simulations with sample sizes of 100 sequences in each; an additional cut-off requirement of 4 homozygous sequences has no effect here.

3.1.3. Distributions

The statistics *iHS*, *XP-EHH* and *Rsb* have been constructed to be approximately Gaussian distributed under neutrality. Figures S2 and S4 show that using simulated data, this approximation is indeed very good, while our modified quantities show notable deviations. Neglecting ancestry information leads to a skew in *iHS* values and omitting phase results in "heavier tails" of the distributions of all three statistics.

3.2. Simulated selective sweeps

3.2.1. The signal of a single sweep

We simulated a single strong selective sweep in an otherwise neutrally evolving chromosome. The selected variant is located in the middle of the chromosome, dominant and has reached a population frequency of 50%. Figure 6 shows for a large sample as well as a small subsample of it the *iHS* values using its

original definition together with the adjustments defined above. It is clearly visible that omission of ancestry status results mainly in a reduction of the signal caused by the selected site. A lack of phase, by contrast, rather increases the statistical "noise" from the neutral part of the chromosome. In the subsample the signal is severely reduced, but not entirely absent. Note that the relative lack of low values is a more robust feature of a sweep than the attainment of extreme values. Our requirement of at least 10 sequences per allele in unphased data is of little importance if the sample size is large, but reduces drastically the number of evaluated markers in small samples. Finally, note that the selected variant neither has the most extreme value nor does it lie in the exact center of the region with elevated values.

[Figure 6 about here.]

3.2.2. Comparison of the statistics

We simulated a genome consisting of 100 chromosomes, each experiencing a selective sweep at the central position (see methods). We performed a sliding window scan and adapted the threshold such that 100 different candidate regions of selection were called. A candidate region was valued as correctly identified, if it contained the simulated site of selection. A perfect method would find exactly those, while a random delineation yields a "correct" interval with probability 0.01. Figure 7 shows the *False Discovery Rate*, the fraction of erroneously called regions among all called regions. Although an increase of the integration cut-off to 0.1 improves accuracy, the usage of unphased data still yields unreliable statistics for sample sizes below 200 sequences. The additional cut-off criterion of at least 4 homozygous sequences leads only to an improvement for sample sizes below $n = 100$. In contrast, discarding ancestry information yields only a slight loss of accuracy, largely independent of sample size. The Figure also shows the size of the threshold needed to yield a fixed amount of selected regions. The thresholds for unphased and phased data converge to the same value of roughly the 0.95% percentile of the genome-wide distribution. The corresponding values for unpolarized date are lower due to the left-skewness of the distribution.

[Figure 7 about here.]

3.3. Experimental data

We applied our modified statistics to six population samples of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). First, we probed the signal for two classical examples for strong recent selection in humans. Then, we examined the genome-wide distributions and last, we compared the delineated regions. The calculated values are available on Dryad ([dataset] Klassmann *et al.*, 2020).

3.3.1. Two examples of an on-going selective sweep

Several SNPs in the enhancer of the gene *LCT* confer *Lactase persistence* which enables adult humans to digest fresh milk (Enattah *et al.*, 2002; Tishkoff *et al.*, 2007; Enattah *et al.*, 2008). We are here concerned with the SNP *rs4988235* whose derived variant attains its highest frequency of 74% in population `CEU`, while it is virtually absent in all East Asian and non-admixed African populations investigated by the 1000 Genomes Project. Figure 8 depicts *EHH* around this SNP for its two alleles. It can be seen that *EHH* extends far further for the derived variant than for the ancestral one, a sign that the allele reached its current population frequency faster than under neutrality. The curves for *EHH* using the estimator for unphased data are more coarse-grained, but still quite similar in shape and scale. Figure 9 shows the genome-wide standardized *iHS* values around the *LCT* gene. Omission of polarization leads to a reduction of high values, but leaves the overall pattern intact. Omission of phasing, instead, leads to a notable increase of "noise" in the sense that many low values get inflated. Most conspicuous, however, is the massive lack of values in the putative center of the sweep. This gap owes to discarded sites, having the minor allele on less than 10 sequences. In fact, only 7 individuals are homozygous for the minor, i.e. ancestral, allele of the SNP *rs4988235* itself. Figures S6-S9 illustrate that the situation is similar in other candidate regions.

The SNP *rs1426654* influences skin pigmentation (Lamason *et al.*, 2005). The derived variant has low frequency in the African populations, is almost fixed

in the European populations and all but absent in the East Asian populations of the 1000 Genomes Project. Because population sample `CEU` is monomorphic for the derived variant, only cross-population statistics are applicable. Figure 10 shows that *EHHS* in population `CEU` extends far further than in the populations `CHB` and `YRI`. Again, ignoring phase information, we obtain a coarser, but otherwise similar picture. Figure 11 compares the genome-wide standardized *XP-EHH* and *Rsb* values obtained by phased/unphased estimation around the gene *SLC24A5*. All panels look quite similar, suggesting that the respective quantities contain equivalent information.

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

3.3.2. Distributions

Figures S3 and S5 show that the statistics from experimental data have more extreme values, or with other words, their distributions have heavier "tails" than those from simulated neutral evolution. This holds particularly if the estimators for unphased data are employed.

3.3.3. Comparison of the statistics

Ultimately, it is of most interest, whether delineated putative regions under selection are robust with respect to the amendments we made to the original statistics. As discussed in section 2, we largely employ the settings of Voight *et al.* (2006), but adjust the threshold value in order to yield exactly 20 candidate regions for each statistic. We applied the within- and cross-population statistics to the populations `CEU`, `CHB`, `JPT` and `YRI`. Furthermore, in order to obtain larger samples, we merged very closely related populations (see Supplementary Table 5 of The 1000 Genomes Project Consortium (2015)), namely `CEU` and `GBR` as well as `CHB` and `CHS`. Table 2 shows the number of overlapping regions using iHS.

It can be seen that overlap between the regions called from the original statistics with those neglecting ancestral information is considerable, while neglecting phase information yields scarce overlap, even for large sample sizes. Table 3 compares the standard statistics *Rsb* and *XP-EHH* with one another and each of the two with its equivalent for unphased data. Here, all measures lead to overlapping regions with the exception for the comparison between populations `CHB` and `JPT`. However, because these two populations are rather similar, the signal of differential selection is likely to be small and masked by the increased variance arising from neglecting phase. The precise chromosomal locations of all ascertained regions as well as the strengths of the signal are listed in the supplement.

[Table 2 about here.]

[Table 3 about here.]

4. Discussion

While in the last decade the focus of research shifted away from detection of classical hard sweeps to more subtle modes of selection (Stephan, 2016), methods to detect the former will continue to be applied as a first-pass analysis to population genetic data. The aim of our study was to broaden the applicability of the established statistics *iHS*, *XP-EHH* and *Rsb* by relaxing their requirement on sequences to be phased and variants polarized. Although the issue of phasing can often be solved computationally and its importance is likely to wane further given the rapid improvement of sequencing technologies, in the meantime methods that can cope with unphased data might find their niche. In contrast, the polarization of alleles will always remain imperfect and incomplete, notwithstanding rare cases of available ancient DNA. This holds even more so for cases of "reticulate" evolution such as hybridization/admixture where the very concept of an ancestral allele gets blurred. We hence expect any method apt to handle unpolarized variants to remain a useful complement to methods that cannot.

We showed that although omission of ancestry information entails a substantial decrease in peak values, another feature of selective sweeps, namely the relative absence of low scores within a certain region, remains intact and can be exploited to delineate candidate regions.

In contrast, omission of phasing information leads not only to a reduction of the signal but also to a massive increase of statistical noise. This is mainly due to the increased variance of the corresponding estimators of *EHH* resp. *EHHS* arising from a decreased effective sample size. The latter is reduced in two ways: only sequences from homozygous individuals are exploited and only the two sequences of an individual can share a haplotype, discarding all those which are potentially shared across individuals.

This is less of a problem for *EHHS* which is calculated using all alleles of a focal marker. Under Hardy-Weinberg proportions the number of homozygous individuals is on average greater than 50% in a population. Hence in a sample of 100 chromosomes, typically at least 50 chromosomes can be used to calculate *EHHS* and its derivates *XP-EHH* and *Rsb*. This seems enough to yield substantial correlation with their homologues for phased data.

In contrast, for within-population scans, *EHH* has to be estimated for each allele independently which often renders estimation for the minor allele unreliable, because few sequences can be exploited. In order to increase robustness of estimation, we deployed two strategies. First, similarly to Wang *et al.* (2006) who chose to disregard sites where less than 5% of the sampled sequences can be used for estimation, we opted for requiring a fixed minimum *number* of 10 sequences. This condition suffices to detect selective sweeps, if the selected variant is of middle frequency and the sample size exceeds 200 sequences (Figure 7). However, the depletion of variants with intermediate frequency is a major hallmark of a selective sweep (near completion) (Tajima, 1989; Fay and Wu, 2000). Hence, this seemingly mild condition leads to the exclusion of many informative markers not only in the example of the *LCT* locus (Figure 9), but in other candidate regions, too (Figures S6-S9). Figure 5 suggests that generally a minimum number of 20-30 sequences might be needed for accurate estimates.

If we want to detect alleles with a minor frequency of 10% and expect 1% of individuals homozygous for it, we need samples of 2000-3000 sequences; for most purposes a prohibitively high number. The poor overlap of inferred regions in experimental data shown in Table 2 confirms this conclusion.

We intended to reduce the variance in shared haplotype length by increasing the cut-off level for the integration boundaries from 0.05 to 0.1, thus in essence capping the longest observed shared haplotypes. Although this does increase the correlation with the non-phased statistics, the improvement is rather moderate. Both Wang *et al.* (2006) and Tang *et al.* (2007) invented more sophisticated measures to counteract the increased variance of the statistics on unphased sequences. In fact, the former did not integrate *EHH*, but chose to fit a logistic function describing its decay with increasing distance to the focal marker (more precisely, they fitted the increase of $\frac{1}{2}(1-EHH)$). The latter repeated the whole genome scan 50 times on a bootstrapped sample to eliminate the most volatile 50% of significant markers. We doubt, however, that any such noise reduction can overcome the general problem of very small sample sizes for minor alleles. To summarize, without phasing information, the *iHS* statistic yields low power and can detect only strong signals in a range of intermediate allele frequencies.

We aimed at investigating the robustness of delineated candidate regions of selection under a loss of input information, not the power of the statistics as such, having been shown already in the original publications (Voight *et al.*, 2006; Tang *et al.*, 2007). For this reason, we fixed the number of candidate regions and measured the robustness by the amount of intersections. Obviously this approach is too rigid for a real-word whole-genome scan on selection. Furthermore, the fine-scale plots of our candidate regions in Figures S6-S9 remind that their delineation depends on various, to some degree arbitrarily chosen parameters such as the window size, handling of gaps and boundary regions, a minimum number of markers, the condition imposed to yield a "cluster" of significant scores and not least the significance threshold itself, which is notoriously uncertain given that null-models can be specified only approximately.

References

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics **183**: 249–258.

Akey, J. M., 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Research **12**: 1805–1814.

Baudry, E., and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics **165**: 1619–1622.

Browning, S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**: 703–714.

Crisci, J. L., Y. P. Poh, S. Mahajan, and J. D. Jensen, 2013 The impact of equilibrium assumptions on tests of selection. Frontiers in Genetics **4**: 1–7.

[dataset] Klassmann, A., R. Vitalis, and M. Gautier, 2020 Detecting selection using *Extended Haplotype Homozygosity (EHH)*-based statistics on unphased or unpolarized data. Dataset. Dryad 10.5061/dryad.rfj6q5775.

DeGiorgio, M., C. D. Huber, M. J. Hubisz, I. Hellmann, and R. Nielsen, 2016 SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics **32**: 1895–1897.

Dumont, B. L., and B. A. Payseur, 2008 Evolution of the genomic rate of recombination in mammals. Evolution **62**: 276–294.

Enattah, N. S., T. G. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, *et al.*, 2008 Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. American Journal of Human Genetics **82**: 57–72.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, *et al.*, 2002 Identification of a variant associated with adult-type hypolactasia. Nature Genetics **30**: 233–237.

Ewing, G., and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–5.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155**: 1405–13.

Gautier, M., A. Klassmann, and R. Vitalis, 2017 rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Molecular Ecology Resources **17**: 78–90.

Gautier, M., and R. Vitalis, 2012 rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics **28**: 1176–7.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics **5**: 1–11.

Hellenthal, G., and M. Stephens, 2007 msHOT: Modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23**: 520–521.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. Science **331**: 920–924.

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Molecular Biology and Evolution **24**: 1792–800.

Jónsson, H., P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, *et al.*, 2017 Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature **549**: 519–522.

Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura, 2018 MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution **35**: 1547–1549.

Lamason, R. L., M.-a. A. P. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, *et al.*, 2005 Genetics: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science **310**: 1782–1786.

Lowy-Gallego, E., S. Fairley, X. Zheng-Bradley, M. Ruffier, L. Clarke, *et al.*, 2019 Variant calling on the grch38 assembly with the data from phase three of the 1000 genomes project [version 2; peer review: 2 approved]. Wellcome Open Research **4**: 1–41.

Maclean, C. A., N. P. Chue Hong, and J. G. D. Prendergast, 2015 Hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. Molecular Biology and Evolution **32**: 3027–3029.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

Nielsen, R., S. Williamson, Y. Kim, R. Nielsen, S. Williamson, *et al.*, 2005 Genomic scans for selective sweeps using SNP data Genomic scans for selective sweeps using SNP data. Genome research : 1566–1575.

Pavlidis, P., D. Živković, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Molecular Biology and Evolution **30**: 2224–34.

Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, 2014 PopGenome: An efficient swiss army knife for population genomic analyses in R. Molecular Biology and Evolution **31**: 1929–1936.

649 Rozas, J., A. Ferrer-Mata, J. C. Sanchez-DelBarrio, S. Guirao-Rico,
650 P. Librado, *et al.*, 2017 DnaSP 6: DNA sequence polymorphism analysis of
651 large data sets. Molecular Biology and Evolution **34**: 3299–3302.

652 Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J.
653 Richter, *et al.*, 2002 Detecting recent positive selection in the human
654 genomes from haplotype structure. Nature **419**: 832–7.

655 Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. B. Hostetter,
656 *et al.*, 2007 Genome-wide detection and characterization of positive selection
657 in human populations. Nature **449**: 913–8.

658 Stephan, W., 2016 Signatures of positive selection: From selective sweeps
659 at individual loci to subtle allele frequency changes in polygenic adaptation.
660 Molecular Ecology **25**: 79–88.

661 Szpiech, Z. A., and R. D. Hernandez, 2014 Selscan: An efficient multi-
662 threaded program to perform EHH-based scans for positive selection. Molec-
663 ular Biology and Evolution **31**: 2824–2827.

664 Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis
665 by DNA polymorphism. Genetics **123**: 585–95.

666 Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for
667 using genome scans to detect recent positive selection in the human genome.
668 PLoS Biology **5**: 1587–1602.

669 The 1000 Genomes Project Consortium, 2015 A global reference for hu-
670 man genetic variation. Nature **526**: 68–74.

671 Tishkoff, S. A., F. a. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt,
672 *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and
673 Europe. Nature Genetics **39**: 31–40.

674 Vitti, J. J., S. R. Grossman, and P. C. Sabeti, 2013 Detecting natural
675 selection in genomic data. Annual Review of Genetics **47**: 97–120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biology .

Wang, E. T., G. Kodama, P. Baldi, and R. K. Moyzis, 2006 Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences **103**: 135–140.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, *et al.*, 2018 Ensembl 2018. Nucleic Acids Research **46**: D754–D761.

Data accessibility

Released packages of rehh are available from the CRAN repository at https://cran.r-project.org/package=rehh. The statistics calculated for the samples of the 1000 Genomes Project are available in the Dryad Digital Repository by identifier https://doi.org/10.5061/dryad.rfj6q5775.

Author contributions

A.K. designed and performed the study. A.K., R.V. and M.G. wrote, reviewed and revised the manuscript.

Supporting Information

Additional Supporting Information may be found on the online version of this article:

Figure S1 Unstandardized *iHS* values in dependence of the derived allele frequency.

Figures S2-S5 Distribution and qq-plots of *iHS*, *Rsb* and *XP-EHH*.

Figures S6-S9 Manhattan plots of candidate regions of selection delineated by *iHS*.

Tables S1-S39 Coordinates and statistics of candidate regions delineated by *iHS*, *Rsb* and *XP-EHH*.

List of Figures

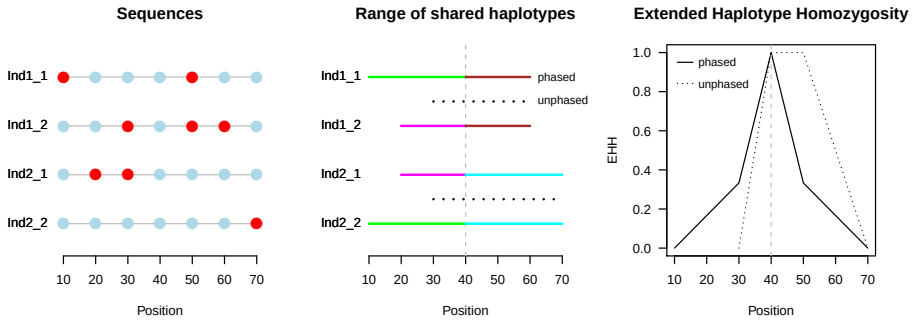

Figure 1: An example for the calculation of EHH using the estimator for phased (eq. (1)) resp. unphased sequences (eq. (5)). The left panel depicts the variants seen in four aligned sequences belonging to two diploid individuals. At the central marker (position 40) all sequences share the same allele and this marker is taken as focal in the other two panels. The middle panel shows the range of shared extended haplotypes around the focal marker. In the drawing, the boundaries of shared haplotypes are defined by the position of the marker that introduces a difference between the hitherto identical haplotypes. Assuming unphased sequences, only the two sequences of each individual are compared. Those of individual 1 become different at the first marker to the left of the focal marker. Consequently, the dashed line for individual 1 has its left end at position 30. In contrast, if their phase is known, all sequences can be compared with each other, yielding 6 comparisons. For each sequence, the range of its longest shared haplotype is indicated and the haplotype identified by color. Shorter haplotypes are not shown (e.g. all 4 sequences share a single extended haplotype between positions 30 and 50). The right panel shows the EHH values calculated at each marker position as the fraction of sequences sharing a haplotype among all sequence comparisons. Note that the EHH curve has been defined as linearly interpolating values between consecutive markers.

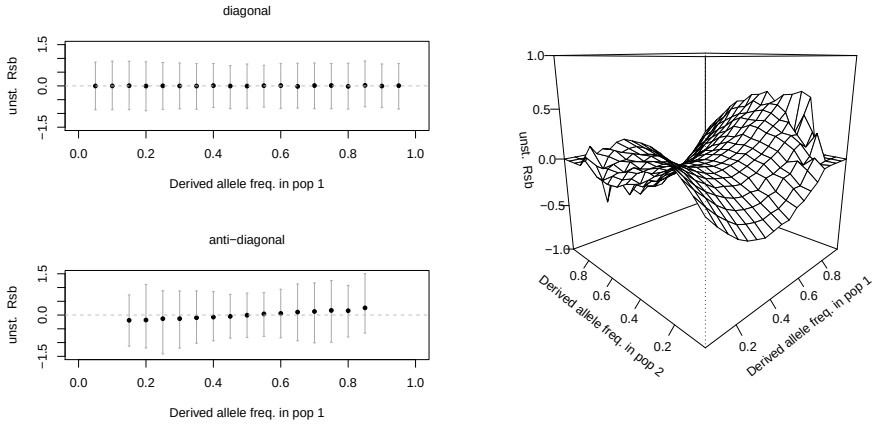

Figure 2: Unstandardized *XP-EHH* in dependence of the derived allele frequency in both populations. Simulated was a region of 50 Megabases evolving neutrally in two recently split populations. The sample size was $n = 100$ in each population. Results are averaged over 200 runs. The grid points represent values at discrete frequencies, not averages over frequency intervals. The values for mutual frequency differences greater than 0.7 relied on less than 25 observed markers and were set to zero. The left panels show sections of the right panel along the diagonal ($p_s^{pop1} = p_s^{pop2}$) and antidiagonal ($p_s^{pop2} = 1 - p_s^{pop1}$) together with the 0.05 and 0.95 quantiles.

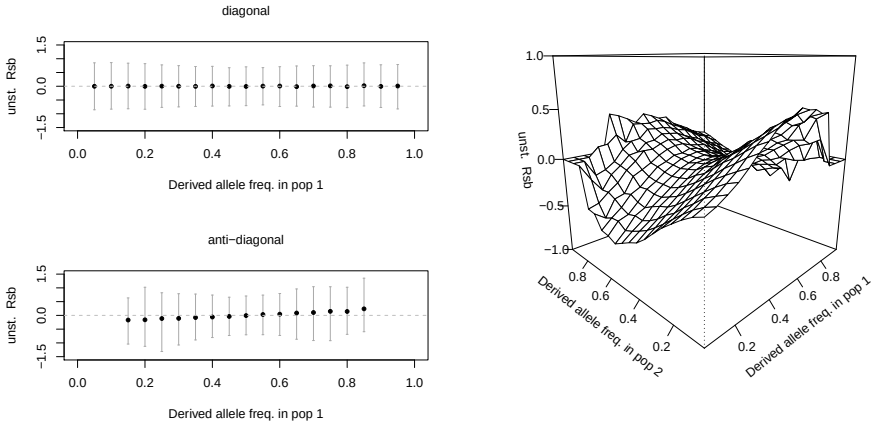

Figure 3: Same as Figure 2, but for unstandardized *Rsb*.

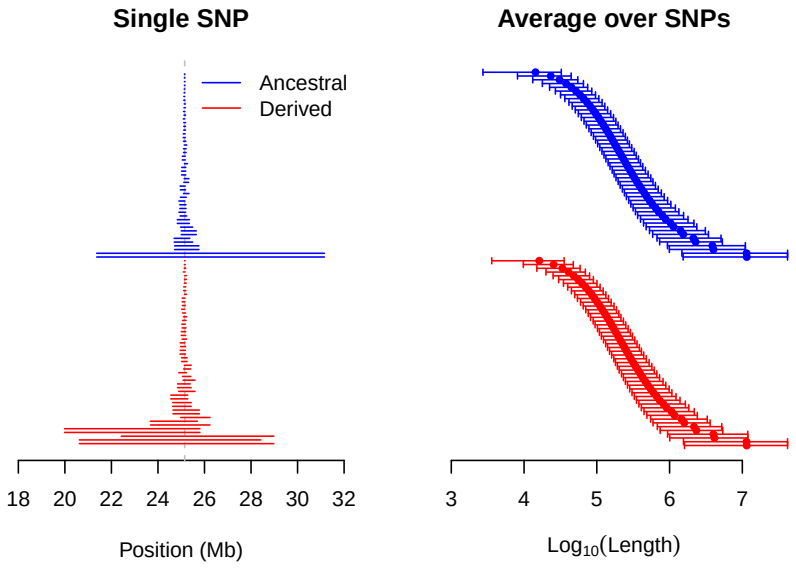

Figure 4: Length of shared haplotypes. Simulated was a region of 50 Megabases in a neutrally evolving population and a sample size of $n = 100$. Like in the middle panel of Figure 1, the lines in the left panel symbolize the range of shared extended haplotypes, here ordered by their length. These are shown for a single SNP near the center of the chromosome. Both core alleles have a sample frequency of 50%. Clearly visible is the very long range of a few haplotypes. The right panel is intended to show that this is a typical feature: here the lengths (left+right to the focal marker) of ordered shared haplotypes is averaged over SNPs with 50% core allele frequencies. For this, 100 sample replicates were generated and only SNPs less than 5 Mb away from the center were taken into acccount in order to minimize boundary effects. Note that although not visible on this scale, the distribution is actually capped to the right by the length of the chromosome which in a few replicates was attained by a shared haplotype.

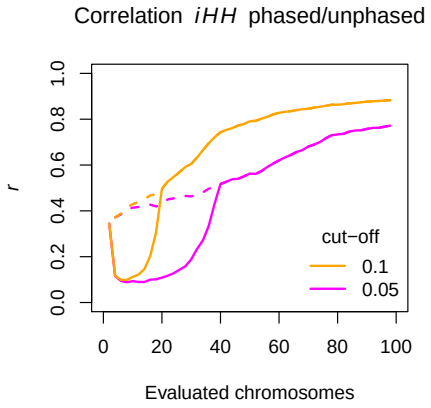

Figure 5: The correlation between *iHH* values obtained by phased vs unphased estimators on the same simulated data as for Figure 4. Both ancestral and derived allele were used. The x axis indicates the number of sequences that were evaluable for unphased estimation, hence belonging to homozygous individuals. The "cut-off" yields the threshold at which integration over *EHH* is stopped. For phased estimation, the standard cut-off of 0.05 was used, for unphased estimation an additional value of 0.1. Note that in the unphased case the cut-off value defines the minimum fraction of evaluated sequences. The dashed lines show the correlation, if an absolute cut-off condition is imposed on top: a minimum number of 4 sequences (2 homozygous indiviudals). Beyond a certain number of evaluated sequences (in the example 20 resp. 40) the first condition implies the second. For very small allele frequencies all approaches converge; in the extreme case of a core allele present in a single homozygous individual, phased and unphased estimators are identical (but too unreliable to be useful).

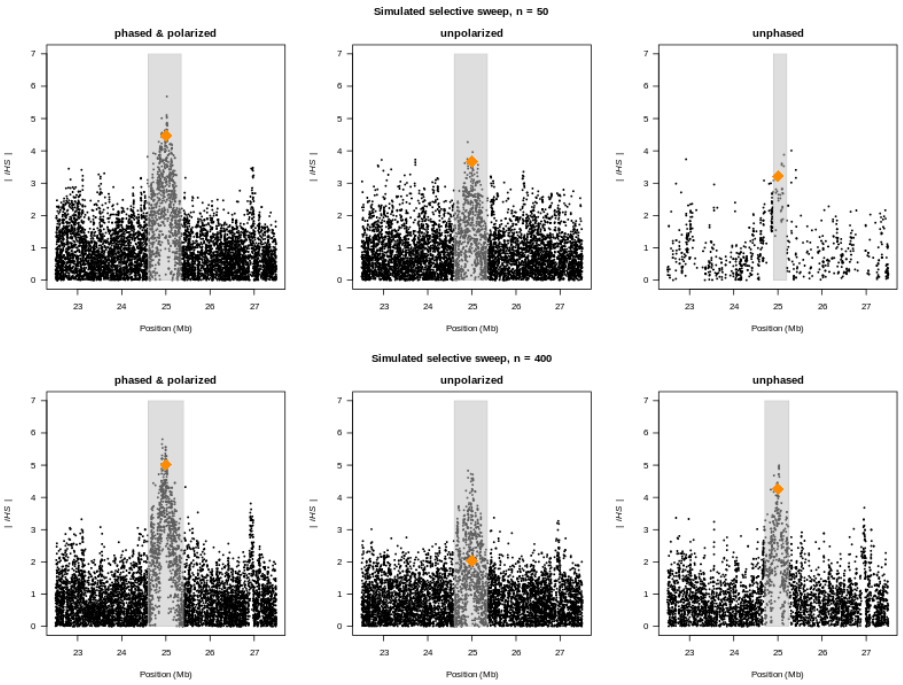

Figure 6: Standardized *iHS* values for a single simulated on-going selective sweep. The selected variant (marked in orange) is at the center of the simulated region and has reached a population frequency of 50%. The simulated sample was of size $n = 400$; the top panels show values obtained from a random sub sample of size $n = 50$. Delineated candidate regions for selection are marked in gray.

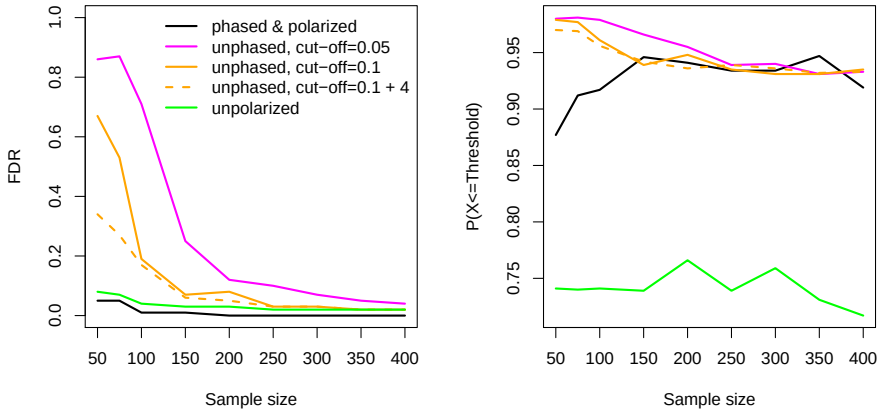

Figure 7: Comparison of the adapted statistics for different sample sizes. We simulated a genome of 50 chromosomes, each of length 50 Mb and with a single selected site at the center. A threshold was fitted such that the number of delineated regions equaled the number of selected sites. As a measure of the power of the adapted statistics, the left panel shows the *False Discovery Rate*, the fraction of erroneously called regions among all delineated regions. The right panel shows the fitted thresholds as quantiles of the respective genome-wide distributions.

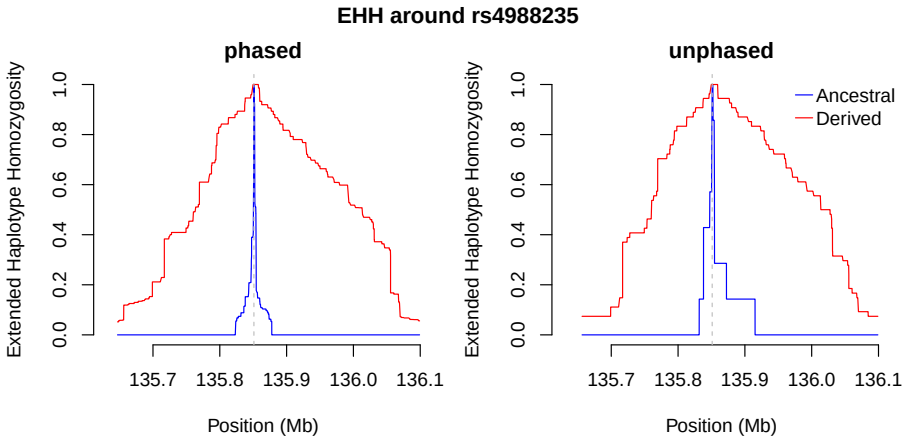

Figure 8: *EHH* for ancestral and derived alleles of SNP *rs4988235* in population `CEU` of the 1000 genomes project. The SNP is located on chromosome 2, about 13kb upstream (in 3'-direction) of the gene *LCT*.

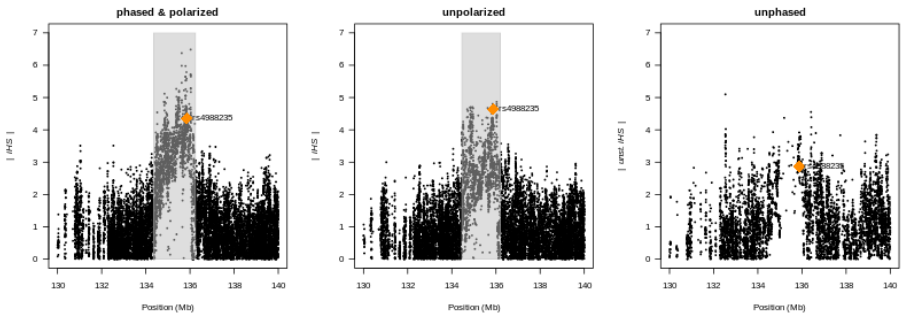

Figure 9: Standardized *iHS* values in a region around the gene *LCT* in population `CEU`. Candidate regions for selection are marked in gray. That the putatively causal site has a more prominent score using unpolarized estimation is, in our opinion, entirely accidental.

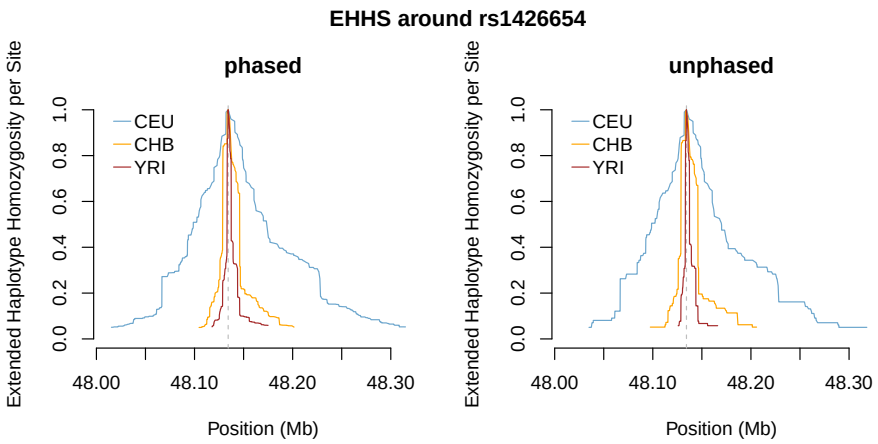

Figure 10: Normalized *EHHS* around SNP *rs1426654* in populations `CEU`, `CHB` and `YRI`. The SNP is located within gene *SLC24A5*.

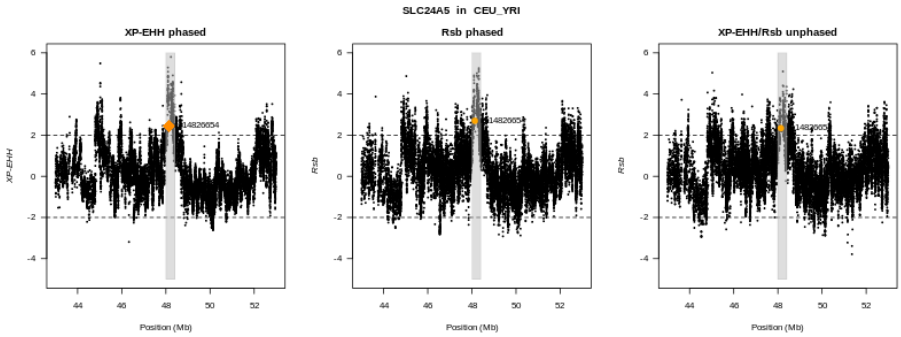

Figure 11: Standardized *XP-EHH* and *Rsb* values in a region around the gene *SLC24A5* for population `CEU` versus `YRI`. Delineated candidate regions for selection are marked in gray.

List of Tables

Sample	Population	# Individuals
CEU	Central Europeans in Utah (CEPH individuals)	99
CHB	Han Chinese in Beijing, China	106
CHS	Han Chinese South, China	105
GBR	British from England and Scotland	100
JPT	Japanese in Tokyo, Japan	105
YRI	Yoruba in Ibadan, Nigeria	107

Table 1: The population samples of the 1000 genomes project used in this study.

	iHS phased polarized/unpolarized	*iHS* polarized phased/unphased
CEU	10	2
CHB	12	1
JPT	9	2
YRI	14	3
CEU+GBR	11	4
CHB+CHS	12	3

Table 2: The number of overlapping delineated candidate regions for selection using iHS. For each statistic a threshold was fitted in order to yield exactly 20 candidate regions.

	Rsb/*XP-EHH* phased	*XP-EHH* phased/unphased	*Rsb* phased/unphased
CEU vs CHB	12	11	11
CEU vs JPT	11	10	13
CEU vs YRI	11	6	10
CHB vs JPT	13	4	3
CHB vs YRI	12	6	10
JPT vs YRI	11	5	8
CEU+GBR vs CHB+CHS	13	12	12

Table 3: The number of overlapping delineated candidate regions for differential selection using *XP-EHH* and *Rsb*. For each statistic the 20 most conspicous regions were considered. Note that unphased *XP-EHH* and *Rsb* are virtually identical and so are the respective candidate regions.