ORIGINAL ARTICLE

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

Alexander Klassmann[1] | **Mauthieu Gautier**[2]

[1]Institute for Genetics, University of Cologne, Cologne, Germany

[2]CBGP, Univ Montpellier, CIRAD, INRAE, IRD, Montpellier SupAgro, Montpellier, France

Correspondence
Alexander Klassmann, Institute for Genetics, University of Cologne, Cologne, Germany
Email: alex.klassmann@koeln.de

Funding information
[No fundings to declare]

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 (EHH)* and its associated statistics. These statistics typically need phased haplotypes and, some of them, polarized variants. Here, we unify and extend previously proposed modifications to loosen these requirements. We compare the modified versions with the original ones by measuring the False Discovery Rate on simulated whole-genome scans and quantifying the overlap of inferred candidate regions on empirical data. We find that phasing information is indispensable for the accurate estimation of within-population statistics for all but very large samples and of cross-population statistics for small samples. Ancestry information, in contrast, is of lesser importance for both. Our publicly available R package rehh incorporates the modified statistics presented here.

KEYWORDS
selective sweep, neutrality tests, *EHH*, phasing, variant polarization

1 | INTRODUCTION

The ease with which genomic sequences can be obtained contrasts sharply with the challenge to discern their functional elements. Finding molecular signatures of recent selection can help prioritizing regions for further investigation. The search for them is often performed by means of statistical tests refuting 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 conceptually simple statistics such as F_{ST} (Akey, 2002), but may be corroborated by more sophisticated approaches, including those presented here, which exploit other aspects of the selection signal. In contrast, the detection of selection within a single population has proven more challenging, with various methods trying to capture an aspect of reduction in genetic variation (Oleksyk, Smith, & O'Brien, 2010; Sabeti, 2006). Measures for the average sample homozygosity and length of "runs of homozygosity" in individuals can be regarded, in our opinion, as pre-stages of the frequency spectrum- and *EHH*-based statistics presented here, respectively. Hands-on overviews are provided by (Cadzow et al., 2014; Utsunomiya, Pérez O'Brien, Sonstegard, Sölkner, & Garcia, 2015; Weigand & Leese, 2018). We focus on the following three approaches which have been widely applied for more than a decade (Haasl & Payseur, 2016):

- *Tajima's D* (Tajima, 1989), *Fay & Wu's H* (Fay & 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 dnasp (Rozas et al., 2017) and the R package PopGenome (Pfeifer, Wittelsbürger, Ramos-Onsins, & Lercher, 2014).
- SweepFinder (DeGiorgio, Huber, Hubisz, Hellmann, & Nielsen, 2016; Nielsen et al., 2005) and SweeD (Pavlidis, Živković, Stamatakis, & Alachiotis, 2013) are two programs implementing the same method. They calculate the frequency spectrum around specific chromosomal positions and compare the likelihood of a fitted sweep model (assuming gradual erosion of the signal of selection with increasing genetic distance) with the likelihood of a

position-independent null spectrum. The latter is taken either from the empirical genome-wide "background" or derived from an explicit demographic model.

- Sabeti et al. (2002) introduced the concept of *Extended Haplotype Homozygosity (EHH)* on top of which Voight, Kudaravalli, Wen, and Pritchard (2006) built a statistic called *iHS*, with later variations by Sabeti et al. (2007) and Tang, Thornton, and Stoneking (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 & Hernandez, 2014), hapbin (Maclean, Chue Hong, & Prendergast, 2015) and the R package rehh (Gautier, Klassmann, & Vitalis, 2017; Gautier & Vitalis, 2012).

In our view, there are two major points that distinguish *EHH*-based from frequency spectrum approaches (see our Supplementary Information on Site Frequency Spectrum-based methods for a short review):

- the latter are constructed to detect *completed* selective sweeps while the former are focussed on *on-going* selective sweeps. At least in the human species, completed selective sweeps seem to be rare (Hernandez et al., 2011) and prime examples of selection, such as variants influencing the expression of the *LCT* gene (discussed below), are yet far from fixation (Vitti, Grossman, & Sabeti, 2013).
- Tajima's D and similar quantities refer to genomic intervals and although SweepFinder/SweeD compute scores for exact genomic positions, these are not directly associated with any particular polymorphism. In contrast, *EHH*-based statistics are tied to specific sites.

The methods listed above, except Tajima's D in its original version, exploit 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 & Depaulis, 2003; Hernandez, Williamson, & Bustamante, 2007). Furthermore, a reference genome of that species has to be 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 Genomes Project cannot be polarized (see below).

In addition to polarization, the calculation of *EHH* as described by Sabeti et al. (2002) requires genotype data to be *phased*, i.e. that is known for di- or polyploid individuals which variant of a heterozygous locus belongs to which chromosome. While phased haplotypes are expensive to obtain experimentally, computational methods to infer them probabilistically often yield satisfactory results (Browning & Browning, 2011). Nevertheless, two studies with the same basic approach stated that phasing can be omitted in case of diploid individuals: Wang, Kodama, Baldi, and Moyzis (2006) for a within-population and Tang et al. (2007) for a cross-population test. Both assessed the statistical power by simulations, yet they did not directly compare phased and unphased estimators; the latter merely reported a coefficient of correlation r^2 of 65-73% on empirical data.

The aim of this article is to assess the robustness of *EHH*-based statistics against loss of information about phase or variant ancestry status. We first recapitulate and unify the definition of the three statistics we want to investigate. Then, we present how the statistics can be adapted to account for unphased and/or unpolarized data. For the within-population test we compare the False Discovery Rate for original and modified statistics on simulated whole-genome scans and set them in relation to the above-mentioned frequency spectrum-based methods. For all three statistics we calculate the overlap of candidate regions found by original and modified versions on empirical 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*

At start, we want to clarify that the term *homozygosity* as part of the name *EHH* refers to the probability that two randomly chosen chromosomes from a population are identical (at a certain locus or region) which does not imply any statement about individuals.

Let s denote a site of interest within a chromosome. 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 $n_s = \sum_a n_a$ the total number of sequences. If there are no missing data at the focal marker, then n_s equals the sample size n. All chromosomes 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 region between s and t. Any two (or more) chromosomes identical in that region constitute a *shared haplotype*. Let $K_{s,t}$ denote the number of all distinct shared haplotypes in the sample and $K^a_{s,t}$ the subset with allele a at the focal marker s. n_k refers to the number of sequences sharing haplotype k. The quantity EHH^a as defined by Sabeti et al. (2002) is calculated for chromosomes carrying the core allele a:

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

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 and stopped when EHH, monotonically decreasing with increasing distance to the focal marker, reaches a lower threshold or cut-off, usually chosen as 0.05.

Note that, although iHS was historically defined in this two-step way, it is equivalent, but conceptually simpler to perceive it as the average of the lengths $l_{ij}(s)$ of shared haplotypes over all chromosomes $i \neq j$ carrying core allele a:

$$\mathrm{iHH}^a(s) = \frac{1}{n_a(n_a-1)} \sum_{i \neq j}^{n_a} l_{ij}(s) \ . \tag{3}$$

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

$$\mathrm{uniHS}(s) = \ln\left(\frac{\mathrm{iHH}^A(s)}{\mathrm{iHH}^D(s)}\right) \ . \tag{4}$$

Finally, this quantity is standardized:

$$\mathrm{iHS}(s) = \frac{\mathrm{uniHS}(s) - \mathrm{mean}(\mathrm{uniHS}|p_s)}{\mathrm{sd}(\mathrm{uniHS}|p_s)} \ . \tag{5}$$

Since the expected 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 all markers with the same frequency. In practice, the standardization is carried out over small frequency bins. Voight et al. (2006) stated that *iHS* 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. Typically, a Minor Allele Frequency (MAF) of at least 5% is required which excludes variants near fixation.

To overcome this limitation Sabeti et al. (2007) and Tang et al. (2007) independently modified the above statistic in order to compare two populations instead of two alleles. While Sabeti et al. (2007) kept the designation *EHH*, we follow Tang et al. (2007) in distinguishing *site-specific EHH* by *EHHS*:

$$\mathrm{EHHS}_{s,t} = \frac{1}{n_s(n_s-1)} \sum_{k=1}^{K_{s,t}} n_k(n_k-1) \; . \tag{6}$$

Note that $EHHS_{s,s}$ is an estimate for the focal marker homozygosity. The subsequent statistics are build analogously to Equations (2-5). Sabeti et al. (2007) first integrated this quantity to yield *integrated EHHS (iES)*

$$\mathrm{iES}(s) = \int EHHS_{s,t} \, dt \; , \tag{7}$$

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

$$\text{unXP-EHH}(s) = \ln\left(\frac{\mathrm{iES}_{\mathrm{pop1}}(s)}{\mathrm{iES}_{\mathrm{pop2}}(s)}\right) \; , \tag{8}$$

127 which, in turn, is standardized to yield

$$\text{XP-EHH}(s) = \frac{\text{unXP-EHH}(s) - \text{mean}(\text{unXP-EHH})}{\text{sd}(\text{unXP-EHH})} \ . \tag{9}$$

128 The approach of Tang et al. (2007) differs in so far as $EHHS_{s,t}$ is normalized by its value at marker $t = s$. We thus

129 refer to the integral as *integrated normalized EHHS score*:

$$\text{inES}(s) = \frac{1}{EHHS_{s,s}} \int EHHS_{s,t}\, dt = \frac{iES(s)}{EHHS_{s,s}} \tag{10}$$

130 to obtain first the (unstandardized) *Ratio between populations (Rsb)* [1]

$$\text{unRsb}(s) = \ln\left(\frac{\text{inES}_{\text{pop1}}(s)}{\text{inES}_{\text{pop2}}(s)}\right) , \tag{11}$$

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

$$\text{Rsb(s)} = \frac{\text{unRsb}(s) - \text{median}(\text{unRsb})}{\text{sd}(\text{unRsb})} \ . \tag{12}$$

132 Importantly, for standardization of the cross-population statistics *XP-EHH* and *Rsb* no binning with respect to core

[1] Note that for the sake of uniformity our notation differs slightly from that given in Tang et al. (2007), where (12) was referred to as *ln(Rsb)* and the unlogarithmized value used only for plotting.

allele frequencies is undertaken and thus no variant polarization is presupposed.

2.2 | Modifications for unphased sequences

The probability that two sequences of a population are identical 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) used the idea to estimate EHH (under a different name) and $EHHS$, respectively: the crucial difference to Equations (1) and (6), respectively, is that only the two chromosomes of each individual are compared. The quantities EHH and $EHHS$ are then estimated as above by the fraction of shared haplotypes among all sequence comparisons. 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. At marker t the quantities EHH^a and $EHHS$, respectively, are estimated by

$$\mathrm{EHH}^a_{s,t} = \frac{I^a_{s,t}}{I^a_{s,s}} \tag{13}$$

$$\mathrm{EHHS}_{s,t} = \frac{I_{s,t}}{I_{s,s}} \,. \tag{14}$$

[Figure 1 about here.]

Figure 1 illustrates the original and modified way to estimate EHH (and iHH). All subsequent steps to obtain *iHS*, *XP-EHH* and *Rsb* remain the same as above. Since *EHHS* calculated by Equation (14) is normalized (to yield 1 at the focal marker), for unphased data *XP-EHH* is essentially identical to *Rsb*; the only difference consists in the use of

median and mean, respectively, in the standardization step.

Importantly, only the chromosomes of individuals homozygous at that focal marker can share a haplotype. The resulting set of mutual chromosome comparisons is hence a (typically much smaller and possibly even empty) subset of those made by the original approach.

[Figure 2 about here.]

Figure 2 shows why this entails a major problem: the length of shared haplotypes is distributed very unevenly among the chromosomes of a sample and even in the absence of selection a few shared haplotypes of extreme length can occur. In small samples, these can easily yield "outlier" values of the final statistics, confounding the signal arising from selection. As an attempt to reduce this statistical noise, we imposed the following restrictions:

- only focal markers with at least 10 homozygous sequences (5 individuals) are considered: sample-wise for *XP-EHH* and *Rsb*, and independently for each core allele in case of *iHS* (the latter on top of the original requirement of a Minor Allele Frequency of at least 0.05),
- the cut-off value which stops integration of EHH/EHHS is increased from its original value of 0.05 to 0.10,
- another integration cut-off is added, leading to a stop when less than four chromosomes (two individuals) remain homozygous (for the original statistics this condition follows from the former two).

2.3 | Modifications for unpolarized variants

There is only one step where the information of allele ancestry status is exploited, namely the standardization of $uniHS$ in Equation (5), depending on the frequency of the derived core allele. In order to avoid an arbitrary assignment of ancestry status, we replace the ancestral and derived allele in Equation (4) by major (most frequent) and minor (second

most frequent) allele, respectively.

$$\text{uniHS}(s) = \ln\left(\frac{\text{iHH}^{MAJ}(s)}{\text{iHH}^{MIN}(s)}\right) . \tag{15}$$

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 respectively expected values differ increasingly with increasing distance $|0.5 - s|$. Hence, in lack of a better solution we suggest standardization to be performed without consideration of allele frequencies:

$$\text{iHS}(s) = \frac{\text{uniHS}(s) - \text{mean}(\text{uniHS})}{\text{sd}(\text{uniHS})} . \tag{16}$$

2.4 | Delineation of regions under selection

Voight et al. (2006) showed that single markers with extreme values of *iHS* are less indicative of selective sweeps than a cluster of high values (see figure S2 of their Supporting Information). In effect, they determined candidate intervals of selection by requiring half of markers to have values above the 99% genome-wide percentile. We followed this approach with the modification that we adapted the threshold value in order to obtain a fixed number of candidate regions. We used overlapping sliding windows of width 250 kb with an offset of 50 kb and overlapping candidate windows were merged. For empirical 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 was ignored, this number was halved for *iHS*, corresponding to a similar decrease of markers for which a score could be obtained.

For ease of comparison, we applied sliding windows of same size and overlap to the values of the frequency-spectrum-based tests, although here we required single markers to exceed a given threshold. Because the values of

Tajima's D and Fay & Wu's H are calculated for intervals, we took the interval centers as corresponding positions.

2.5 | Whole genome scans on simulated data

We performed coalescent simulations using msms (Ewing & Hermisson, 2010). We assumed an effective population size of $N_e \approx 10,000$ for humans. In previous simulation studies both population scaled mutation rate and recombination rate were set as $\theta = \rho = 0.001$ per base per generation (Crisci, Poh, Mahajan, & Jensen, 2013; Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) and we followed them for simplicity, although we acknowledge that, depending on the estimation method, rates of half that size can be inferred for both quantities (Dumont & Payseur, 2008; Jónsson et al., 2017; Scally, 2016; Spence & Song, 2019).

For our simulation we set the population-scaled rates of θ and ρ both to $50,000$ to correspond very roughly to a physical length of 50 Mb in humans. This large size proved necessary to reduce boundary effects because, as shown in Figure 2, shared haplotypes can span several Mb even under neutrality. We ignored that recombination events in reality occur within hot spots (McVean et al., 2004), because msms cannot handle varying recombination rates and other tools which can (e.g. msHOT (Hellenthal & Stephens, 2007)), are not able to simulate selection. In order to investigate distributional properties under neutrality, for *iHS* we simulated chromosomes evolving in a single constant-size population and for *XP-EHH*/*Rsb* two neutrally evolving populations which split symmetrically from an ancestral population $4N_e \cdot 0.05$ generations ago ($\approx$ 50,000 years in humans), without subsequent migration.

In order to study selection signals, we created a "genome" consisting of 100 independently simulated samples of chromosomes, each experiencing a single on-going selective sweep located at its center 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% (70%, 90%, respectively). The simulated sample size was $n = 400$ from which we took subsamples down to sample size $n = 50$. We applied on this genome the original as well as the modified *iHS* statistics. For the estimator on unphased data, we tried two cut-off values: the standard one of $EHH = 0.05$ and the more stringent of $EHH = 0.10$. Furthermore, we computationally re-established phase information from randomized genotypes using fastPHASE (Scheet & Stephens, 2006) with subsequent application of

the original statistics. Additionally, we computed values for Tajima's D (Tajima, 1989) and Fay & Wu's H (Fay & Wu, 2000) as well as the Composite Likelihood Score (CLS) as implemented by SweepFinder (DeGiorgio et al., 2016) and SweeD (Pavlidis et al., 2013). The latter was calculated with and without allowance for variant ancestry status.

To evaluate the performance of the tests, we estimated the False Discovery Rate (FDR) of delineated candidate regions for selection. A region was regarded as a "true positive" when it overlapped a true selected site. The FDR hence measures the proportion of mislocated regions among regions deemed significant. For each scan the significance threshold was adjusted so as to call exactly 100 candidate regions. With these settings, the lower the FDR, the more optimal the test. The FDR is equal to zero when each of the 100 simulated sites under selection is identified by a distinct candidate region. If, on the contrary, candidate regions are assigned to random places within the genome, then the probability of a "true positive" equals the combined length of all candidate regions divided by the length of the genome; in this case the expected FDR is 1 minus this probability. Note that the length of candidate regions is not fixed, because they may comprise several merged windows.

See also the Supplementary Information on software and technical details.

2.6 | Whole genome scans on empirical 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 SNPs are covered by ancestral states classified as "high confidence" and a further 6% as "low confidence"; using both, we could polarize 95.8% of SNPs. We calculated the statistics for samples of European origin (`CEU`, `GBR`), Asian origin (`CHB`, `JPT`) and African origin (`YRI`), see Table 1. Additionally, we combined the samples of two closely related populations (see Supplementary Table 5 of The 1000 Genomes Project Consortium (2015)), namely the two European samples mentioned and the Chinese samples `CHB` and `CHS`, respectively.

We assessed the robustness of the statistics against loss of phase or ancestry information by the number of

overlapping candidate regions.

[Table 1 about here.]

3 | RESULTS

3.1 | General properties of the statistics under neutrality

3.1.1 | Dependence on core allele frequencies under neutrality

We examined the dependence (under neutrality) of the three original statistics on the frequency of the derived core allele p_s. For $uniHS$ this was already reported by Voight et al. (2006) (see their Figure 4). We present a similar, but more detailed plot in our Figure S1, left Panel. In order to exclude the possibility that the dependence is an artifact of the number of chromosomes carrying the derived allele in the sample, we re-calculated the *uniHS* values using at each marker a subsample containing an equal number of chromosomes with the derived and the ancestral variant, respectively. The middle panel shows clearly that *uniHS* indeed depends on the population frequency of the derived core allele and not its sample frequency.

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 from populations of identical demography. Figures S2 and S3 show for two recently diverged populations of equal size the averaged unstandardized *XP-EHH* and *Rsb* values in dependence 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 smaller than for the *uniHS* statistics, making a correction less necessary. Furthermore, a frequency-dependent standardization in the vein of *iHS* would need two-dimensional bins and, contrary to *iHS*, the implicit assumption that each bin is dominated by neutral variants does not hold because large frequency differences are indicative of differential selection. 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 mean and median in Equations (9) and (12).

3.1.2 | Distributions of the statistics under neutrality

The statistics *iHS*, *XP-EHH* and *Rsb* have been constructed to be approximately standard-normal distributed under neutrality. Figures S4 and S6 show that for simulated data the distributions of the original statistics are approximated quite well by a Gaussian curve, while our modified scores show notable deviations: neglecting ancestry information leads to a skew in *iHS* values and using the estimator for unphased variants results in "heavier tails" in all three statistics.

3.2 | Whole genome scans on simulated data

3.2.1 | A single selective sweep in detail

In Figure 3 we present an example of the *iHS* values obtained in the vicinity of a strongly selected variant, located at the middle of a chromosome which otherwise evolves neutrally. The variant has reached a population frequency of 70%. It is evident that omission of ancestry status entails a decrease of values around the selected site. A lack of phase, by contrast, primarily increases the statistical "noise" from the neutral part of the chromosome. This can be observed, too, for unstandardized *iHS* in the right panel of Figure S1. The relative lack of low values around the selected site is in each case a more prominent feature of the sweep than the attainment of extreme values, hence the reason to search for such "clusters". Further examples, including those of the frequency spectrum-based tests and calculated for different sample sizes, are given in Figures S8-S13 in Supplementary Figures and Tables. These plots show that our requirement of at least 10 sequences per allele in unphased data is of lesser importance when the sample size is large, but reduces drastically the number of suitable markers in small samples. Note that the selected variant neither necessarily has the most extreme value nor lies in the exact center of the region with elevated values.

[Figure 3 about here.]

3.2.2 | Comparison of the statistics by the False Discovery Rate

Figure 4 summarizes the results of our whole-genome scan on simulated data. It shows the *False Discovery Rate*, the fraction of the 100 delineated candidate regions that did not overlap one of the 100 true selected sites. For each test and sample size the average area covered by the candidate regions comprised about 500 kb per chromosome, hence about 1% of its length, representing the probability to yield a "true positive" by chance alone. First, we can observe that on-going sweeps in early stages can be better recognized by *iHS* than by frequency spectrum-based tests. Second, a lack of polarization yields in every case, and almost independently of sample size, an increase of "false positives", with the effect being smaller for *iHS* than for the other statistics. Third, lack of phase drastically increases the FDR for *iHS* for all but the largest sample sizes and an increased cut-off achieves only partial compensation. Lastly, computational phasing of genotypes is, at least in our high-density simulated data, much more effective than using the modified estimator for unphased sequences; to our surprise the FDR for the reconstructed phase turned out to be partially even lower than for the "true" data. We do not know the reason for this and can only speculate that fastPHASE does not recognize all recombination events, increasing thereby the length of shared haplotypes and hence the signal of selection.

[Figure 4 about here.]

3.3 | Whole genome scans on empirical data

3.3.1 | Two selective sweeps in detail

Several variants in the enhancer of the human gene *LCT* confer *Lactase persistence* which enables adults to digest fresh milk (Enattah et al., 2008, 2002; Tishkoff et al., 2007). Although undisputedly under strong selection, the precise advantage of this capability is still debated (Segurel et al., 2020). 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 of the 1000 Genomes Project. Figure 5 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 6 shows the genome-wide standardized *iHS* values around the *LCT* gene. As with simulated data, 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. Again, most conspicuous is the massive lack of values in the putative center of the sweep owing to the discard of sites with the minor allele present on less than 10 sequences (or 5 individuals). In fact, only 7 individuals are homozygous for the minor, i.e. ancestral, allele of the SNP *rs4988235* itself. Figures S14-S17 illustrate that the situation is similar in other candidate regions. Out of interest, we computed the standard *iHS* values for further populations as well (Figure S18): almost all European populations of the 1000 Genomes Project show a similarly strong signal, while none of the African populations do. However, a further African population investigated within the HapMap3 project (The international HapMap Consortium, 2010) shows a signal like Europeans, as observed already by (Ferrer-Admetlla, Liang, Korneliussen, & Nielsen, 2014).

The SNP *rs1426654* within gene *SLC24A5* translates to a `Ala111Thr` polymorhism in the corresponding protein and influences skin pigmentation (Lamason et al., 2005). The level of pigmentation needs to balance the opposing requirements of protecting from UV radiation as well as ensuring sufficient vitamin D production (Quillen et al., 2019). 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 the population sample `CEU` is monomorphic for the derived variant, only cross-population statistics are applicable. Figure 7 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 8 compares the original *XP-EHH* and *Rsb* statistics with their counterpart for unphased data (where both statistics are essentially identical) around the gene *SLC24A5*. The panels look quite similar, suggesting that the statistics are largely equivalent.

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

3.3.2 | Distribution of the statistics on empirical data

Figures S5 and S7 show that the statistics from empirical data have more extreme values, or with other words, their distributions have heavier "tails" than those from simulated neutral evolution. This holds particularly when the estimators for unphased data are applied.

3.3.3 | Overlap of delineated candidate regions

We are interested in 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 employed the settings of Voight et al. (2006), but adjusted the threshold value in order to yield exactly 20 candidate regions for each statistic. 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 the version for unphased data. Here, the overlap between the modified statistics with the original ones is not much less than between the two original statistics, except for the comparison of populations `CHB` and `JPT`. Because these two populations are rather similar, the signal of differential selection might be too small to be detected without phasing.

The precise chromosomal locations of all ascertained candidate regions as well as the strengths of the signals are listed in Supplementary Figures and Tables.

[Table 2 about here.]

[Table 3 about here.]

The *iHS*, *XP-EHH* and *Rsb* values calculated for the 1000 Genomes Project populations are available on Dryad ([dataset] Klassmann & Gautier, 2020).

4 | DISCUSSION

While ever more sophisticated methods to detect selective sweeps are being developed (Alachiotis & Pavlidis, 2018; Harris & DeGiorgio, 2020; Stern, Wilton, & Nielsen, 2019) and other, more subtle modes of selection (Stephan, 2016) are under increasing scrutiny, the comparatively simple summary statistics presented here will continue to be applied as a first-pass analysis to population genetic data. The aim of our study was to examine whether the established scores *iHS*, *XP-EHH* and *Rsb* can be used without the requirement of sequences to be phased and variants to be 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 compared the different approaches to detect selective sweeps by the False Discovery Rate, because typically in whole-genome scans only a handful of most extreme "outlier" regions can be investigated in detail further on and it is more important that these are correctly identified than to know the overall level of selection as would be described by Statistical Power. We even want to caution that reporting large numbers of putative selective sweeps may inadvertently suggest a precision that cannot be warranted. The fine-scale plots of our candidate regions in Figures S14-S17 should remind that their delineation depends on various, often overlooked parameters such as the handling of gaps and boundary regions, the clustering of significant scores and not least the thresholds applied, which are notoriously uncertain given that in many cases null-models can be specified only roughly.

The results presented in Figure 4 show that frequency spectrum-based methods, constructed for the detection

of sweeps near completion, are unable to detect on-going sweeps when the selected variant is still in intermediate frequency. For the same reason, polarization is more important for those approaches than for *EHH*-based ones. Surprisingly, sample size (at least in the range investigated) plays almost no role for the former.

Concerning *EHH*-based statistics, we showed that although omission of ancestry information entails a substantial decrease in peak values, the conspicuous absence of low scores can still be exploited to delineate candidate regions.

In contrast, the claims of (Wang et al., 2006) and (Tang et al., 2007) that phase can be neglected without major loss of information, must be regarded as too optimistic. The main reason is that the estimation of the statistics in this case relies solely on individuals which are homozygous at the respective focal markers. This is less of a problem for *EHHS* because under Hardy-Weinberg proportions more than half of individuals in a population can be expected to be homozygous for a given marker. Hence in a sample of 100 chromosomes, typically around 50 chromosomes are available to calculate *EHHS* and the derived *XP-EHH* and *Rsb*. This seems enough to yield a substantial similarity with their homologues for phased data as Table 3 shows for empirical data. For *iHS*, however, *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 the robustness of estimation on unphased data, we required a minimum number of 10 sequences to be available for estimation at the focal site. However, the depletion of variants with intermediate frequency is a major hallmark of a selective sweep near completion (Fay & Wu, 2000; Tajima, 1989) and hence, this seemingly mild condition can entail for *iHS* the exclusion of many markers around the selected site. This phenomenon can be seen most clearly at the *LCT* locus (Figure 6), but seems to be general (Figures S14-S17).

Furthermore, we increased the cut-off level for *EHH*/*EHHS* integration from 0.05 to 0.1 and stopped integration as well, when only a single homozygous individual (a single shared haplotype) remained. These added restrictions aim at capping shared haplotype(s) with extreme length. However, as Figure 4 shows, the improvement is rather moderate. Both Wang et al. (2006) and Tang et al. (2007) invented more sophisticated measures: 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 few exploitable sequences.

To summarize, without phasing information, selective sweeps can be located by *iHS* only on very large samples; consequently, phasing should be performed whenever possible. The poor overlap of inferred regions using *iHS* with and without phase on empirical data (Table 2) confirms this conclusion.

On a more fundamental level, Figure 2 reveals the limits of any *EHH*-based approach: the extremely uneven length of shared haplotypes under neutrality produces a difficult to handle background noise. Were this length log-normal distributed as suggested by the right panels of the figure, a remedy could lie in replacing the arithmetic average in Equation 3 by a geometric one. We shortly probed such a replacement, but recognized that the cut-off parameters are more important than the type of averaging. Indeed, (Ferrer-Admetlla et al., 2014) concluded from a coalescent-based reasoning that this problem cannot have an "optimal" solution because the expected length of shared haplotypes is infinite. Hence we do not expect that our ad hoc cut-off rules can be substantially improved or even motivated by theory.

References

Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, *183*(1), 249–258. https://doi.org/10.1534/genetics.109.104042

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

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, *1*(79), 1–11. https://doi.org/10.1038/s42003-018-0085-8

Baudry, E., & Depaulis, F. (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, *165*(3), 1619–1622.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Cadzow, M., Boocock, J., Nguyen, H. T., Wilcox, P., Merriman, T. R., & Black, M. A. (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, *5*(293), 1–8. https://doi.org/10.3389/fgene.2014.00293

Crisci, J. L., Poh, Y. P., Mahajan, S., & Jensen, J. D. (2013). The impact of equilibrium assumptions on tests of selection. *Frontiers in Genetics*, *4*(NOV), 1–7. https://doi.org/10.3389/fgene.2013.00235

[dataset] Klassmann, A., & Gautier, M. (2020). *Detecting selection using* Extended Haplotype Homozygosity-*based statistics on unphased or unpolarized data. Dataset.* https://doi.org/doi:10.5061/dryad.8cz8w9gns

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, *32*(12), 1895–1897. https://doi.org/10.1093/bioinformatics/btw051

Dumont, B. L., & Payseur, B. A. (2008). Evolution of the genomic rate of recombination in mammals. *Evolution*, *62*(2), 276–294. https://doi.org/10.1111/j.1558-5646.2007.00278.x

Enattah, N. S., Jensen, T. G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., … Peltonen, L. (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*(1), 57–72. https://doi.org/10.1016/j.ajhg.2007.09.012

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, *30*(2), 233–237. https://doi.org/10.1038/ng826

Ewing, G., & Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, *26*(16), 2064–5. https://doi.org/10.1093/bioinformatics/btq322

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–13.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*, 78–90. https://doi.org/10.1101/067629

Gautier, M., & Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176–7. https://doi.org/10.1093/bioinformatics/bts115

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), 1–11. https://doi.org/10.1371/journal.pgen.1000695

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5–23. https://doi.org/10.1111/mec.13339

Harris, A. M., & DeGiorgio, M. (2020). A likelihood approach for uncovering selective sweep signatures from haplotype data. *Molecular Biology and Evolution*, *37*(10), 3023–46. https://doi.org/10.1093/molbev/msaa115

Hellenthal, G., & Stephens, M. (2007). msHOT: Modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, *23*(4), 520–521. https://doi.org/10.1093/bioinformatics/btl622

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G. A. T., … Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, *331*(6019), 920–924. https://doi.org/10.1126/science.1198878

Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2007). Context dependence, ancestral misidentification, and

spurious signatures of natural selection. *Molecular Biology and Evolution*, *24*(8), 1792–800. https://doi.org/10.1093/molbev/msm108

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., … Stefansson, K. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, *549*(7673), 519–522. https://doi.org/10.1038/nature24018

Lamason, R. L., Mohideen, M.-a. A. P., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., … Cheng, K. C. (2005). Genetics: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, *310*(5755), 1782–1786. https://doi.org/10.1126/science.1116238

Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P., & The 1000 Genomes Project Consortium. (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. https://doi.org/10.12688/wellcomeopenres.15126.1

Maclean, C. A., Chue Hong, N. P., & Prendergast, J. G. D. (2015). Hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Molecular Biology and Evolution*, *32*(11), 3027–3029. https://doi.org/10.1093/molbev/msv172

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*(5670), 581–584. https://doi.org/10.1126/science.

Nielsen, R., Williamson, S., Kim, Y., Nielsen, R., Williamson, S., Kim, Y., … Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome research*, *15*, 1566–1575. https://doi.org/10.1101/gr.4252305

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 185–205. https://doi.org/10.1098/rstb.2009.0219

Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, *30*(9), 2224–34. https://doi.org/10.1093/molbev/mst112

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Quillen, E. E., Norton, H. L., Parra, E. J., Lona-Durazo, F., Ang, K. C., Illiescu, F. M., … Jablonski, N. G. (2019). Shades of complexity: New perspectives on the evolution and genetic architecture of human skin. *American Journal of Physical Anthropology*, *168*(September 2018), 4–26. https://doi.org/10.1002/ajpa.23737

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

Sabeti, P. C. (2006). Positive natural selection in the human lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/

10.1126/science.1124309

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., … Lander, E. S. (2002). Detecting recent positive selection in the human genomes from haplotype structure. *Nature*, *419*(6909), 832–7. https://doi.org/10.1038/nature01027.1.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E. B., Cotsapas, C., … The international HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–8. https://doi.org/10.1038/nature06250

Scally, A. (2016). The mutation rate in human evolution and demographic inference. *Current Opinion in Genetics and Development*, *41*, 36–43. https://doi.org/10.1016/j.gde.2016.07.008

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Segurel, L., Guarino-Vignon, P., Marchi, N., Lafosse, S., Laurent, R., Bon, C., … Heyer, E. (2020). Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS biology*, *18*(6), 1–11. https://doi.org/10.1371/journal.pbio.3000742

Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, *5*(10), 1–14. https://doi.org/10.1126/sciadv.aaw9206

Stephan, W. (2016). Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, *25*(1), 79–88. https://doi.org/10.1111/mec.13288

Stern, A. J., Wilton, P. R., & Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, *15*(9), 1–32. https://doi.org/10.1371/journal.pgen.1008384

Szpiech, Z. A., & Hernandez, R. D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. https://doi.org/10.1093/molbev/msu211

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–95.

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, *5*(7), 1587–1602. https://doi.org/10.1371/journal.pbio.0050171

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The international HapMap Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–8. https://doi.org/10.1038/nature09298

Tishkoff, S. A., Reed, F. a., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., … Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, *39*(1), 31–40. https://doi.org/10.1038/ng1946

Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Sölkner, J., & Garcia, J. F. (2015). Genomic data as the "hitchhiker's guide" to cattle adaptation: Tracking the milestones of past selection in the bovine genome. *Frontiers in Genetics*, *5*(FEB), 1–13. https://doi.org/10.3389/fgene.2015.00036

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), 446–58. https://doi.org/10.1371/journal.pbio.0040072

Wang, E. T., Kodama, G., Baldi, P., & Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences*, *103*(1), 135–140. https://doi.org/10.1073/pnas.0509691102

Weigand, H., & Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society*, *184*(2), 528–583. https://doi.org/10.1093/zoolinnean/zly007

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

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:10.5061/dryad.rfj6q5775.

Author contributions

A.K. designed and performed the study. A.K. and M.G. wrote and revised the manuscript.

Acknowledgements

We thank Renaud Vitalis for helpful comments.

Supporting Information

Three files with supporting information can be found on the online version of this article:

- Supplementary Information on Site Frequency Spectrum-based methods
- Supplementary Information on software and technical details
- Supplementary Figures and Tables comprising

Figures S1-S3 Unstandardized *iHS*, *XP-EHH* and *Rsb* values in dependence of the derived allele frequency.

Figures S4-S7 Distribution and qq-plots of simulated and empirical *iHS*, *Rsb* and *XP-EHH*.

Figures S8-S13 Manhattan plots for the simulated whole-genome scans around the selected site for different sample sizes and population frequencies of the selected variant.

Figures S14-S17 Manhattan plots for the empirical whole-genome scans around candidate regions of selection.

Figure S18 *iHS* values in the *LCT* region for European and African populations.

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

List of Figures

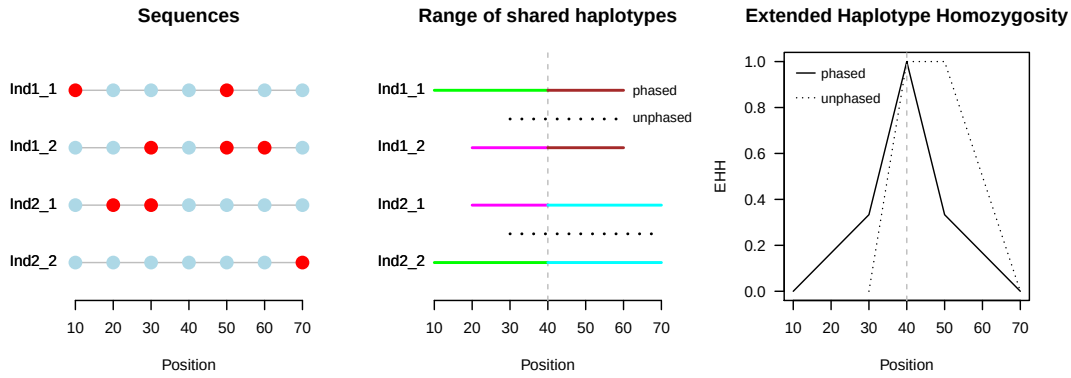

FIGURE 1 An example for the calculation of EHH using the estimator for phased (Equation (1)) and unphased sequences (Equation (6)). The left panel depicts the variants seen in four aligned sequences belonging to two diploid individuals. At the central marker (position 40), all sequences share the same 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. The boundaries of shared haplotypes are defined by the position of the marker that introduces a difference between the hitherto identical sequences. Without phase information, only the two sequences of each individual can be compared and the resulting shared haplotypes are visualized by dashed lines. For instance, the two sequences of individual 1 become different at the first marker to the left of the focal marker and consequently, their shared haplotype ends at position 30. In contrast, when variants are phased, all sequences can be compared with each other, yielding at most 6 different shared haplotypes. The panel depicts for each sequence only its longest shared haplotype, indicated by a solid line, with the constituting sequences in the same color. The remaining comparisons yield (trivial) shared haplotypes, extending to position 30 and 50, respectively. The right panel shows the EHH values, calculated at each marker position as the fraction of sequences sharing a haplotype among all comparisons. Note that the EHH curve is typically defined as linearly interpolating between consecutive markers (as depicted), although for completely sequenced data a stepwise constant function would be more appropriate. With the latter definition, the integral over the EHH-curve, iHH, becomes identical to the average length of shared haplotypes: $\frac{30+40}{2} = 35$ and $\frac{180}{6} = 30$ for unphased and phased sequences, respectively

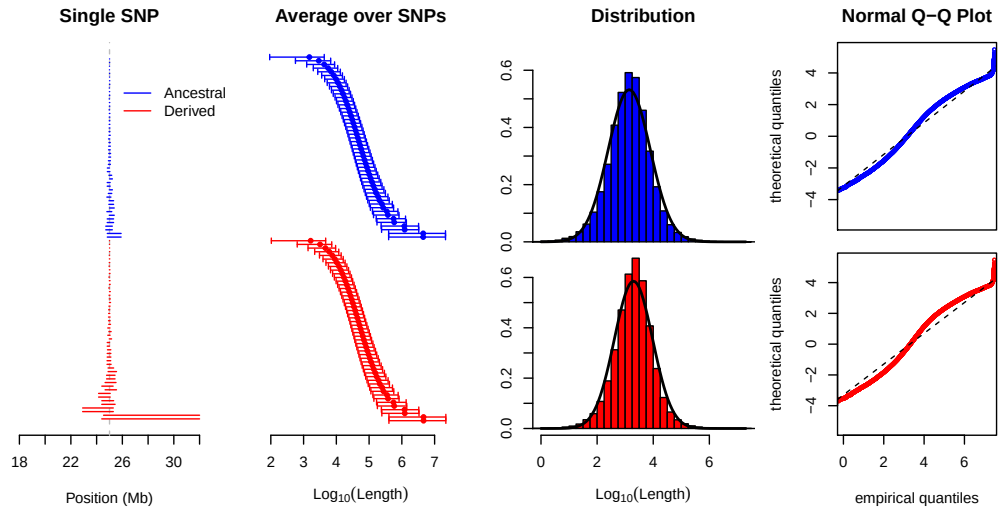

FIGURE 2 Length of shared haplotypes. Simulated was a region of 50 Mb in a neutrally evolving population with a sample size of $n = 100$. We considered only SNPs where both core alleles have a sample frequency of 50% and we assumed that phase is known. Like in the middle panel of Figure 1, the lines in the left panel symbolize the range of the longest shared extended haplotypes, here ordered by their length, for a single, "arbitrarily" chosen SNP (the most central one on the first simulation "run"). Outstanding is the extreme length of a single shared haplotype. The center left panel shows that this is not an exceptional feature: here, the shared haplotype lengths (restricted to those to the "right" of the focal marker) are averaged over SNPs from 100 independent simulation runs, restricted to those less than 5 Mb away from the center in order to minimize boundary effects. The ends of the bars represent the 5% and 95% quantiles, respectively. The center right panel shows for the same SNPs the length distributions of all pairwise shared haplotypes ($\frac{50 \cdot 49}{2}$ per SNP and allele). The distributions are overlayed with a fitted Gaussian curve. The right panel shows Q-Q-plots of the distributions. Note that the largest lengths are actually capped, because in 11 simulation runs, shared haplotypes reached the chromosomal boundary

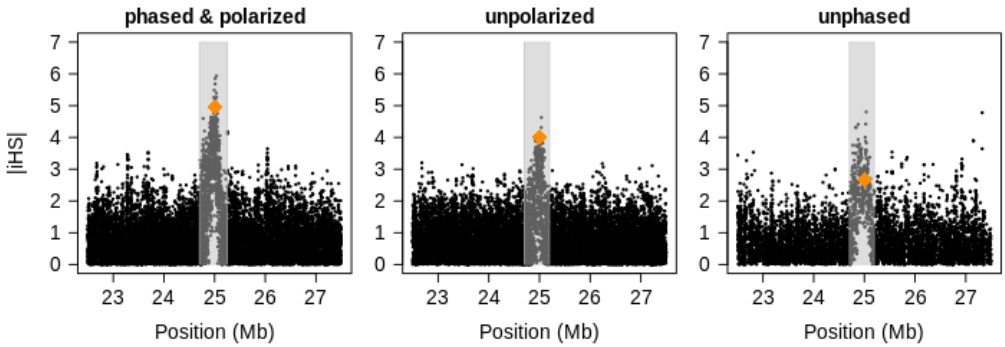

FIGURE 3 *iHS* values of a single simulation "run" (arbitrarily chosen as the first of the 100 runs) around a site with a selected variant of population frequency 70% using a sample of size $n = 200$. The value for the site with the selected variant is marked in dark orange and delineated candidate regions for selection are marked in gray. See also Figures S8-S13

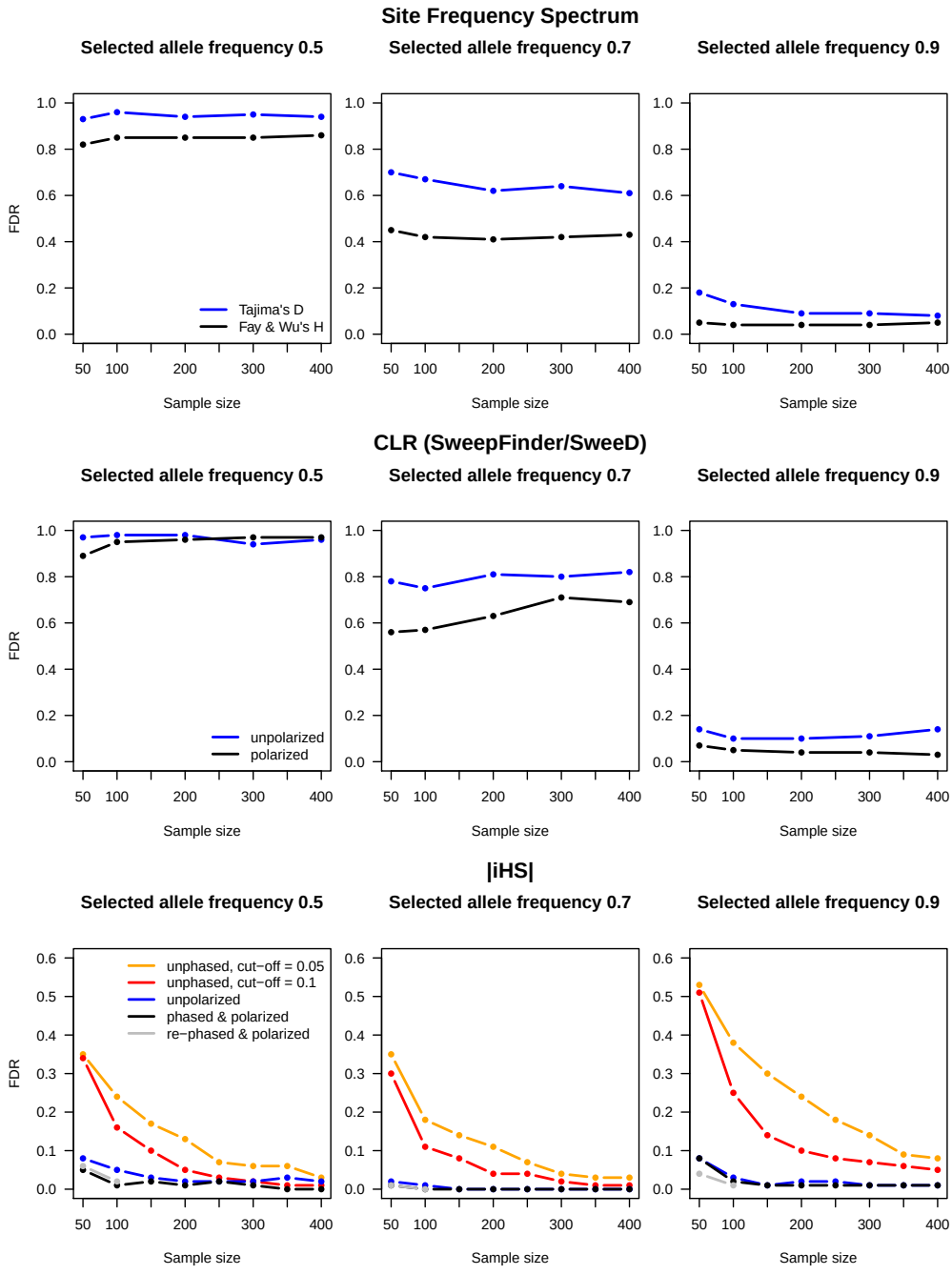

FIGURE 4 Comparison of the False Discovery Rate for different statistics, different sample sizes and different frequencies of the selected allele. 100 candidate regions for selection were delineated on a simulated genome containing 100 sites under selection. The FDR represents the fraction of incorrectly located regions, i.e. regions that do not overlap any "true" site under selection. An ideal test should yield a FDR of zero. Re-phasing was performed only for sample sizes 50 and 100

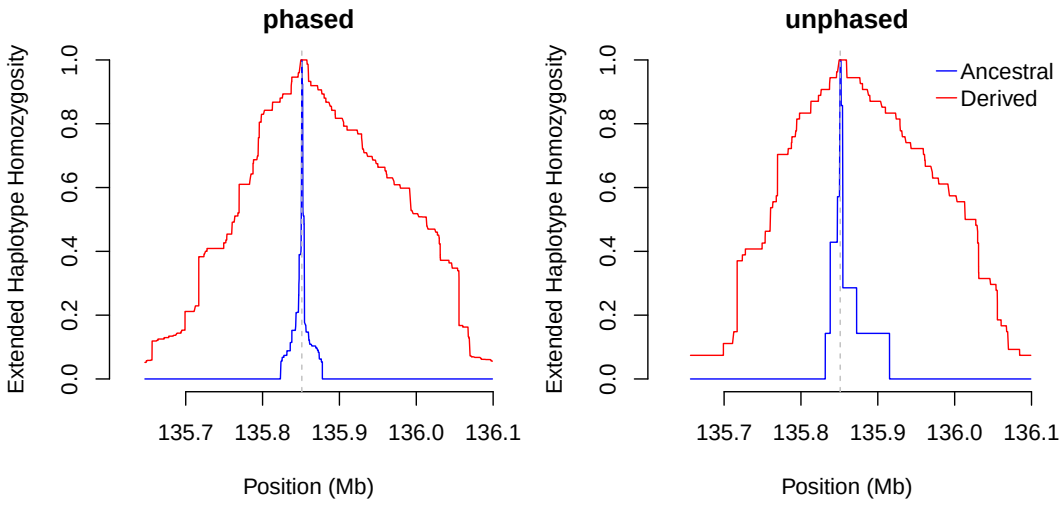

FIGURE 5 *EHH* for ancestral and derived alleles of SNP *rs4988235* in population `CEU` of the 1000 Genomes Project. The SNP is located on chromosome 2, about 13kb upstream (in 3'-direction) of the gene *LCT*

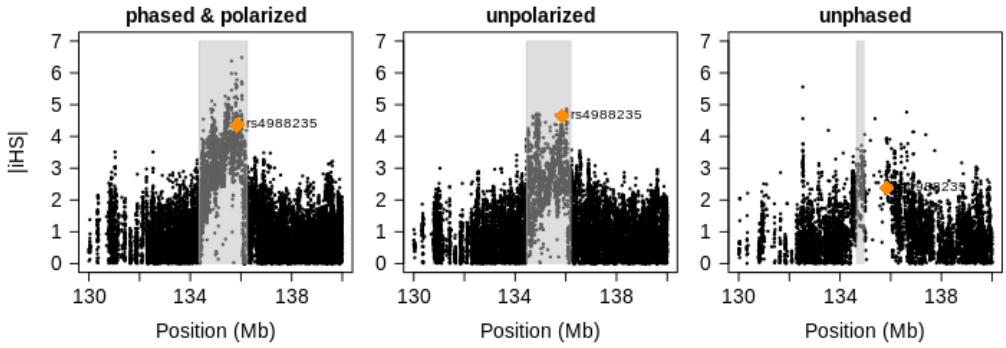

FIGURE 6 *iHS* values in a region around the gene *LCT* in population `CEU`. The value of the putatively selected site is marked in dark orange and delineated candidate regions for selection are marked in gray. That the putatively causal site has a more prominent score using unpolarized estimation is, in our opinion, entirely accidental

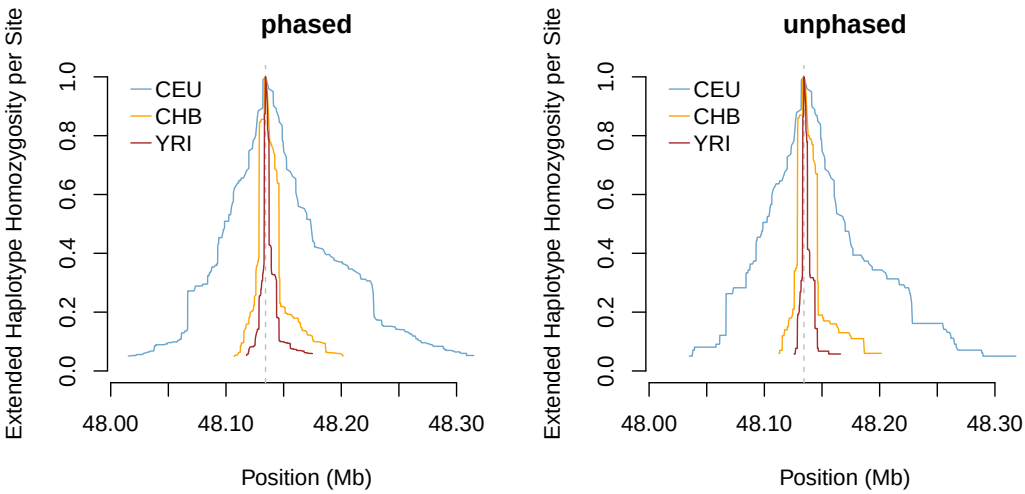

FIGURE 7 Normalized *EHHS* around SNP *rs1426654* in populations `CEU`, `CHB` and `YRI`. The SNP is located within gene *SLC24A5*

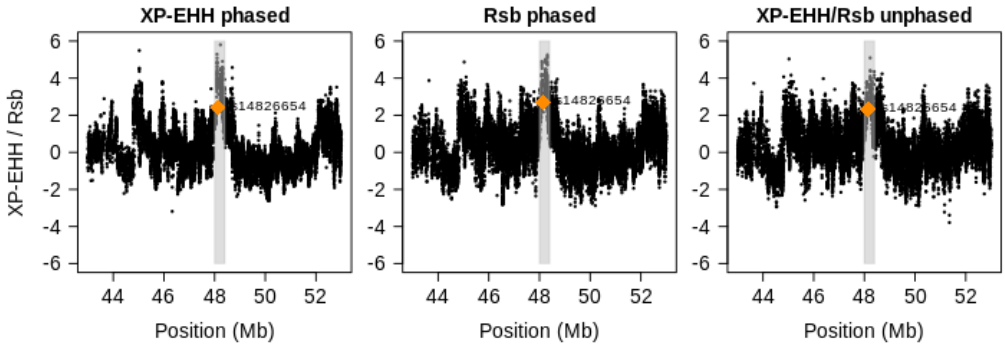

FIGURE 8 *XP-EHH* and *Rsb* values in a region around the gene *SLC24A5* for a comparison of population `CEU` with `YRI`. The value of the putatively selected site is marked in dark orange and delineated candidate regions for selection are marked in gray

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	5
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	*Rsb*/*XP-EHH* unphased
CEU vs CHB	12	11	11	20
CEU vs JPT	11	9	14	18
CEU vs YRI	11	7	10	20
CHB vs JPT	13	4	3	20
CHB vs YRI	12	6	10	18
JPT vs YRI	11	8	11	20
CEU+GBR vs CHB+CHS	13	12	12	20

TABLE 3 The number of overlapping delineated candidate regions for differential selection using *XP-EHH* and *Rsb*. For each statistic a threshold was fitted in order to yield exactly 20 candidate regions. Note that unphased *XP-EHH* and *Rsb* are almost identical and so are the respective candidate regions