

REHH 2.0: a reimplementation of the R package REHH to detect positive selection from haplotype structure

Mathieu Gautier, Alexander Klassmann, Renaud Vitalis

► **To cite this version:**

Mathieu Gautier, Alexander Klassmann, Renaud Vitalis. REHH 2.0: a reimplementation of the R package REHH to detect positive selection from haplotype structure. *Molecular Ecology Resources*, Wiley/Blackwell, 2017, 17 (1), pp.78-90. 10.1111/1755-0998.12634 . hal-01607599

HAL Id: hal-01607599

<https://hal.archives-ouvertes.fr/hal-01607599>

Submitted on 31 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



REHH 2.0: a reimplementation of the R package REHH to detect positive selection from haplotype structure.

Mathieu Gautier^{1,2}, Alexander Klassmann³ and Renaud Vitalis^{1,2}

¹ INRA, UMR CBGP, F-34988 Montferrier-sur-Lez, France

² Institut de Biologie Computationnelle, F-34095 Montpellier, France

³ Universität zu Köln, D-50674 Köln, Germany

keywords: footprints of selection, EHH, iHS, Rsb, XP-EHH

1 **Abstract**

2 Identifying genomic regions with unusually high local haplotype homozygosity represents a powerful
3 strategy to characterize candidate genes responding to natural or artificial positive selection. To that
4 end, statistics measuring the extent of haplotype homozygosity within (e.g., EHH, iHS) and between
5 (Rsb or XP-EHH) populations have been proposed in the literature. The REHH package for R was
6 previously developed to facilitate genome-wide scans of selection, based on the analysis of long-range
7 haplotypes. However, its performance wasn't sufficient to cope with the growing size of available data
8 sets. Here we propose a major upgrade of the REHH package, which includes an improved processing of
9 the input files, a faster algorithm to enumerate haplotypes, as well as multi-threading. As illustrated
10 with the analysis of large human haplotype data sets, these improvements decrease the computation time
11 by more than an order of magnitude. This new version of REHH will thus allow performing iHS-, Rsb- or
12 XP-EHH-based scans on large data sets. The package REHH 2.0 is available from the CRAN repository
13 (<http://cran.r-project.org/web/packages/rehh/index.html>) together with help files and a detailed
14 manual.

15 Introduction

16 Next-generation sequencing (NGS) technologies have deeply transformed the nature of polymorphism
17 data. Although population geneticists were, until recently, limited by the amount of available data in a
18 handful of presumably independent markers, they now have access to dense single nucleotide polymor-
19 phism (SNP) data in both model and non-model species (Davey *et al*, 2011). In those species where
20 genome assemblies are available, the analysis of haplotype structure in a population has proved useful
21 to detect recent positive selection (Sabeti *et al*, 2002). Consider neutral mutations appearing in a pop-
22 ulation: if, by chance, any of these increases in frequency after some time, then recombination should
23 tend to break down linkage disequilibrium (LD) around it, thereby decreasing the length of haplotypes
24 on which this mutation stands. Common variants are therefore expected to be old and standing on short
25 haplotypes. If a mutation is selected for, however, it should expand in the population before recombi-
26 nation has time to break down the haplotype on which it occurred. A powerful strategy to characterize
27 candidate genes responding to natural or artificial positive selection thus consists in identifying genomic
28 regions with unusually high local haplotype homozygosity, relatively to neutral expectation (Sabeti *et al*,
29 2002).

30 For that purpose, Sabeti *et al* (2002) introduced a new metric, referred to as the extended haplotype
31 homozygosity (EHH), which measures the decay of identity by descent, as function of distance, between
32 randomly sampled chromosomes carrying a focal SNP. Tests of departure of EHH from neutral expect-
33 ation were proposed, based on coalescent simulations of demographic history. Voight *et al* (2006) later
34 introduced a test statistic (iHS) based on the standardized log-ratio of the integrals of the observed decay
35 of EHH computed for the ancestral and the derived alleles at the focal SNP. Finally, cross-population
36 statistics were proposed, to contrast EHH profiles between populations: XP-EHH (Sabeti *et al*, 2007)
37 and Rsb (Tang *et al*, 2007). These haplotype-based methods of detecting selection have largely been
38 applied on human data (Vitti *et al*, 2013), a wide range of livestock (see, e.g. Flori *et al*, 2014; Bosse
39 *et al*, 2015; Barson *et al*, 2015) and plant species (see, e.g. Wang *et al*, 2014; Jin *et al*, 2016), and also
40 non-model species (see, e.g. Roesti *et al*, 2015; Mueller *et al*, 2016).

41 A few years ago, we developed REHH (Gautier & Vitalis, 2012), a package for the statistical software
42 package R (R Development Core Team, 2008), to detect recent positive selection from the analysis of long-
43 range haplotypes. Since then, two alternative programs were released: **selscan** (Szpiech & Hernandez,

44 2014), which introduces multithreading to improve computational efficiency and `hapbin` (Maclean *et al*,
45 2015), which in addition to multithreading offers considerable gain in computation time thanks to a new
46 computational approach based on a bitwise algorithm.

47 Here we propose a major upgrade of the REHH package (Gautier & Vitalis, 2012), which includes an
48 improved algorithm to enumerate haplotypes, as well as multi-threading. These improvements decrease
49 the computation time by more than an order of magnitude, as compared to the previous REHH version
50 (1.13), which eases the analysis of big datasets.

51 Below we provide a brief overview of the statistics and tests available in REHH 2.0, and give a detailed
52 worked example of the analysis of chromosome 2 in humans (HSA2), from HapMap samples CEU (Utah
53 residents with Northern and Western European ancestry from the CEPH collection) and JPT+CHB
54 (Japanese in Tokyo, Japan and Chinese from Beijing, China). We use this example as a guideline to use
55 REHH2.0. We further show how REHH was improved since the previous version, and how it compares to
56 the alternative programs `selscan` (Szpiech & Hernandez, 2014) and `hapbin` (Maclean *et al*, 2015).

57 Overview of the EHH-based tests

58 Within population tests

59 **The allele-specific extended haplotype homozygosity:** EHH (Sabeti *et al*, 2002)

60 At a focal SNP and for a given core allele (the ancestral or derived), the allele-specific extended haplotype
61 homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes (carrying the
62 core allele considered) are identical by descent (IBD). IBD is assayed by computing homozygosity at
63 all SNPs within an interval surrounding the core region (Sabeti *et al*, 2002). The EHH thus aims at
64 measuring to which extent an extended haplotype is transmitted without recombination. In practice,
65 the EHH ($EHH_{a_s,t}$) of a tested core allele a_s ($a_s = 1$ or $a_s = 2$) for a focal SNP s over the chromosome
66 interval comprised between the core allele a_s and the SNP t is computed as:

$$EHH_{a_s,t} = \frac{1}{n_{a_s}(n_{a_s} - 1)} \sum_{k=1}^{K_{a_s,t}} n_k(n_k - 1) \quad (1)$$

67 where $K_{a_s,t}$ represents the number of different extended haplotypes (from SNP s to SNP t) carrying
68 the core allele a_s , n_k is the number of the k th haplotype, and n_{a_s} represents the number of haplotypes
69 carrying the core allele a_s , i.e., $n_{a_s} = \sum_{k=1}^{K_{a_s,t}} n_k$.

70 **The integrated (allele-specific) EHH: iHH** (Voight *et al*, 2006)

71 By definition, irrespective of the allele considered, EHH starts at 1, and decays monotonically to 0 as one
72 moves away from the focal SNP. For a given core allele, the integrated EHH (iHH) (Voight *et al*, 2006) is
73 defined as the area under the EHH curve with respect to map position. In REHH (Gautier & Vitalis, 2012),
74 this definite integral is computed using the trapezoidal rule. In practice, the integral is only computed
75 for the regions of the curve above an arbitrarily small EHH value (e.g., $EHH > 0.05$). In their seminal
76 paper, Voight *et al* (2006) considered genetic distances and apply a penalty (proportional to physical
77 distances) for successive SNPs separated by more than 20 kb. In addition, they did not compute iHH
78 if any physical distance between a pair of neighboring SNPs was above 200 kb. We did not implement
79 such an approach in REHH although this might easily be done by modifying the positions of the markers
80 in SNP information input file. In addition, large gaps between successive SNPs (e.g., centromeres) might
81 also be treated by splitting the chromosomes. For instance, when analyzing metacentric chromosomes
82 (e.g., HSA2), each chromosome arm may be considered separately in the analyses by assigning a different
83 chromosome name (e.g., 2a and 2b) to the underlying SNPs.

84 **The standardized ratio of core alleles iHH: iHS** (Voight *et al*, 2006)

85 Let UniHS represent the log-ratio of the iHH for its ancestral (iHH_a) and derived (iHH_d) alleles:

$$\text{UniHS} = \log \left(\frac{iHH_a}{iHH_d} \right) \quad (2)$$

86 The iHS of a given focal SNP s ($iHS(s)$) is then defined following (Voight *et al*, 2006) as:

$$iHS(s) = \frac{\text{UniHS}(s) - \mu_{\text{UniHS}}^{p_s}}{\sigma_{\text{UniHS}}^{p_s}} \quad (3)$$

87 where $\mu_{\text{UniHS}}^{p_s}$ and $\sigma_{\text{UniHS}}^{p_s}$ represent, respectively, the average and the standard deviation of the UniHS
88 computed over all the SNPs with a derived allele frequency p_s similar to that of the core SNP s . In
89 practice, the derived allele frequencies are generally binned so that each bin is large enough (e.g., > 10

90 SNPs) to obtain reliable estimates of $\mu_{\text{UniHS}}^{p_s}$ and $\sigma_{\text{UniHS}}^{p_s}$. The iHS is constructed to have an approximately
 91 standard Gaussian distribution and to be comparable across SNPs regardless of their underlying allele
 92 frequencies. Hence, one may further transform iHS into p_{iHS} (Gautier & Naves, 2011):

$$p_{\text{iHS}} = -\log_{10}(1 - 2|\Phi(\text{iHS}) - 0.5|) \quad (4)$$

93 where $\Phi(x)$ represents the Gaussian cumulative distribution function. Assuming most of the genotyped
 94 SNPs behave neutrally (i.e., that the genome-wide empirical iHS distribution is a fair approximation
 95 of the neutral distribution), p_{iHS} may thus be interpreted as a two-sided p -value (in a $-\log_{10}$ scale)
 96 associated with the null hypothesis of selective neutrality.

97 Pairwise-population tests

98 **The site-specific extended haplotype homozygosity: EHHS** (Tang *et al*, 2007; Sabeti *et al*, 2007)

99 At a focal SNP, the site-specific extended haplotype homozygosity (EHHS) is defined as the probability
 100 that two randomly chosen chromosomes are IBD at all SNPs within an interval surrounding the core
 101 region. EHHS might roughly be viewed as linear combination of the EHH's for the two alternative alleles,
 102 with some weights depending on the corresponding allele frequencies. Two different EHHS estimators
 103 further referred to as $\text{EHHS}^{\text{Sabeti}}$ and $\text{EHHS}^{\text{Tang}}$ have been proposed by Sabeti *et al* (2007) and Tang
 104 *et al* (2007), respectively. For a focal SNP s over a chromosome interval extending to SNP t , these are
 105 computed as (using the same notation as above):

$$\text{EHHS}_{s,t}^{\text{Sabeti}} = \frac{1}{n_s(n_s - 1)} \sum_{a_s=1}^{a_s=2} \left(\sum_{k=1}^{K_{a_s,t}} n_k(n_k - 1) \right) \quad (5)$$

106 where $n_s = \sum_{a_s=1}^{a_s=2} n_{a_s}$ and

$$\text{EHHS}_{s,t}^{\text{Tang}} = \frac{1 - h_{\text{hap}}^{(s,t)}}{1 - h_{\text{all}}^{(s)}} \quad (6)$$

107 where:

108 • $h_{\text{all}}^{(s)} = \frac{n_s}{n_s - 1} \left(1 - \frac{1}{n_s^2} \sum_{a_s=1}^{a_s=2} n_{a_s}^2 \right)$ is an estimator of the focal SNP heterozygosity

109 • $h_{hap}^{(s,t)} = \frac{n_s}{n_s-1} \left(1 - \frac{1}{n_s} \sum_{a_s=1}^{a_s=2} \left(\sum_{k=1}^{K_{a_s,t}} n_k^2 \right) \right)$ is an estimator of haplotype heterozygosity over the chro-
 110 mosome region interval extending from SNP s to SNP t .

111 **The integrated EHHS: iES**

112 As for the EHH (see above), EHHS starts at 1 and decays monotonically to 0 with increasing distance
 113 from the focal SNP. At a focal SNP, and in a similar fashion as the iHH, iES is defined as the integrated
 114 EHHS (Tang *et al*, 2007). Depending on the EHHS estimator considered, $\text{EHHS}^{\text{Sabeti}}$ or $\text{EHHS}^{\text{Tang}}$, two
 115 different iES estimators, that we further refer to as $\text{iES}^{\text{Sabeti}}$ and iES^{Tang} can be computed.

116 **The standardized ratios of pairwise population iES: XP-EHH (Sabeti *et al*, 2007) and Rsb (Tang
 117 *et al*, 2007)**

118 For a given SNP s , let $\text{LRiES}^{\text{Sabeti}}(s)$ (respectively $\text{LRiES}^{\text{Tang}}(s)$) represent the (unstandardized) log-ratio
 119 of the $\text{iES}_{\text{pop1}}^{\text{Sabeti}}(s)$ and $\text{iES}_{\text{pop2}}^{\text{Sabeti}}(s)$ (respectively $\text{iES}_{\text{pop1}}^{\text{Tang}}(s)$ and $\text{iES}_{\text{pop2}}^{\text{Tang}}(s)$) computed in two different
 120 populations:

$$\text{LRiES}^{\text{Sabeti}}(s) = \log \left(\frac{\text{iES}_{\text{pop1}}^{\text{Sabeti}}(s)}{\text{iES}_{\text{pop2}}^{\text{Sabeti}}(s)} \right) \quad \text{and} \quad \text{LRiES}^{\text{Tang}}(s) = \log \left(\frac{\text{iES}_{\text{pop1}}^{\text{Tang}}(s)}{\text{iES}_{\text{pop2}}^{\text{Tang}}(s)} \right) \quad (7)$$

121 The XP-EHH (Sabeti *et al*, 2007) and the Rsb (Tang *et al*, 2007) for a given focal SNP are then
 122 standardized, as:

$$\text{xpEHH}(s) = \frac{\text{LRiES}^{\text{Sabeti}}(s) - \text{med}_{\text{LRiES}^{\text{Sabeti}}}}{\sigma_{\text{LRiES}^{\text{Sabeti}}}} \quad \text{and} \quad \text{rSB}(s) = \frac{\text{LRiES}^{\text{Tang}}(s) - \text{med}_{\text{LRiES}^{\text{Tang}}}}{\sigma_{\text{LRiES}^{\text{Tang}}}} \quad (8)$$

123 where $\text{med}_{\text{LRiES}^{\text{Sab}}}$ (respectively $\text{med}_{\text{LRiES}^{\text{Tang}}}$) and $\sigma_{\text{LRiES}^{\text{Sab}}}$ (respectively $\sigma_{\text{LRiES}^{\text{Tang}}}$) represent the me-
 124 dian and standard deviation of the $\text{LRiES}^{\text{Sabeti}}(s)$ (respectively $\text{LRiES}^{\text{Tang}}(s)$) computed over all the
 125 analyzed SNPs. As recommended by Tang *et al* (2007), the median is used instead of the mean because
 126 it is less sensitive to extreme data points. As for the iHS (see above), XP-EHH and Rsb are constructed
 127 to have an approximately standard Gaussian distribution. They may further be transformed into p_{xpEHH}

128 or p_{rSB} :

$$p_{\text{xpEHH}} = -\log_{10}(1 - 2|\Phi(\text{xpEHH}) - 0.5|) \quad \text{and} \quad p_{\text{rSB}} = -\log_{10}(1 - 2|\Phi(\text{rSB}) - 0.5|) \quad (9)$$

129 where $\Phi(x)$ represents the Gaussian cumulative distribution function. Assuming most of the genotyped
130 SNPs behave neutrally (i.e., the genome-wide empirical distributions of XP-EHH and Rsb are fair ap-
131 proximations of their corresponding neutral distributions), p_{xpEHH} and p_{rSB} may thus be interpreted as
132 a two-sided p -values (in a $-\log_{10}$ scale) associated with a null hypothesis of selective neutrality. Alter-
133 natively, one may also compute p'_{xpEHH} or p'_{rSB} as:

$$p'_{\text{xpEHH}} = -\log_{10}(\Phi(\text{xpEHH})) \quad \text{and} \quad p'_{\text{rSB}} = -\log_{10}(|\Phi(\text{rSB})|) \quad (10)$$

134 (see Gautier & Naves, 2011); p'_{xpEHH} and p'_{rSB} may then be interpreted as a one-sided p -values (in a
135 $-\log_{10}$ scale) allowing the identification of those sites displaying outstandingly high EHHs in population
136 $pop2$ (represented in the denominator of the corresponding LRiES) relatively to the reference population
137 ($pop1$).

138 Material and Methods

139 A new efficient algorithm to explore haplotype variability

140 In the previous version of REHH (1.13) the distribution of haplotype counts for the entire interval from
141 the core SNP to the distance x was computed for each x independently, entailing repeatedly the same
142 calculations. In the new version of REHH (2.0), the distribution of haplotype counts for the interval
143 from the core SNP to the distance x is updated consecutively from the distribution of haplotype counts
144 corresponding to the interval between the core SNP and $x - 1$. The new algorithm doesn't affect the
145 output, in particular, as in the previous version, all haplotypes carrying missing data are discarded from
146 the computation of long-range homozygosity.

147 Human haplotype data

148 Two HSA2 haplotype data sets were downloaded from the HAPMAP project (phase III) (The Interna-
149 tional HapMap3 Consortium, 2010) website (<ftp://ftp.ncbi.nlm.nih.gov/hapmap/>). They consisted
150 of 236 haplotypes of 116,430 SNPs from the CEU and 342 haplotypes from the JPT+CHB popula-
151 tions, respectively. Further details about these data (including the phasing procedure) can be found
152 on the HAPMAP website. For each SNP, the ancestral (and derived) allele was determined accord-
153 ing to the Chimpanzee genome reference (using the *dbsnp_chimp_B36.gff* annotation file available at
154 ftp://ftp.ncbi.nlm.nih.gov/hapmap/gbrowse/2010-08_phaseII+III/gff/). Such ancestral infor-
155 mation is indeed required to carry out iHS-based tests (see above). As a result, 6,230 SNPs (5.35%)
156 for which ancestral/derived states could not be unambiguously determined were discarded from further
157 analyses leading to a total of 110,200 SNPs per analyzed haplotype.

158 Computation

159 For comparison purposes, the different haplotype data sets were analyzed using the software packages
160 REHH (both the previous version 1.13 and the new version 2.0), **selscan** (version 1.1.0b) (Szpiech &
161 Hernandez, 2014) and **hapbin** (version 1.0.0) (Maclean *et al*, 2015). Default options were generally used
162 except for the minimal threshold on the minor allele frequency (MAF) that was set to 0.01 for all pro-
163 grams. In addition, for the **selscan** program, both the window size around the core SNPs (`--ehh-win`
164 option) and the maximum allowed gap in bp between two consecutive SNPs (`--max-gap` option) were set
165 to 10^9 (this was made to disallow these options that are not considered in other programs). Similarly,
166 the `--max-extent` option was inactivated by setting `--max-extent=-1`. For the **hapbin** programs (i.e.,
167 **ihsbin** and **xpehhbin**), the EHH and EHHS cut-off values (defined to stop the calculation of unstandard-
168 ized iHS and iES) were set to 0.05 (i.e., the default value in **selscan** and REHH). For all programs, the
169 standardization of iHS was performed with allele frequency bins of 0.01, as controlled by the **freqbins**
170 argument in the **ihh2ihs()** function of the REHH package, and the **bins** argument for the program **norm**
171 of the **selscan** package and the program **ihsbin** of the **hapbin** package. The command lines used for the
172 different programs, together with the corresponding input data files are provided in the Supplementary
173 Materials.

174 Finally, for each analysis and parameter set, the estimation of computation times was averaged over

175 ten independent runs. All analyses were run on a standard computer running under Linux Debian 8.5
176 and equipped with an Intel[®] Xeon[®] 6-core processor W3690 (3.46 GHz, 12M cache). Note that the Unix
177 command `taskset` was used to control the number of working threads for the analyses with the `hapbin`
178 programs (since neither the `ihsbin` nor the `xpehhbin` programs allow to chose the number of threads to
179 be used).

180 Results and Discussion

181 Analysis of the human chromosome 2 data sets

182 For illustration purpose, we used REHH 2.0 to analyze two human data sets consisting of 236 and 342
183 haplotypes of 110,200 SNPs mapping to HSA2 that were sampled in the CEU and JPT+CHB popu-
184 lations respectively. The chromosome-wide scans of iHS for the CEU and the JPT+CHB populations,
185 respectively, are plotted in Figure 1A. The most significant SNP map at position 136,503,121 bp for the
186 CEU population (iHS=-5.35) and at position 111,506,728 bp for the JPT+CHB population (iHS=-4.92).
187 The chromosome-wide scans of XP-EHH and Rsb, which contrast EHH profiles between the CEU and
188 the JPT+CHB populations, are plotted in Figure 1B. The most significant SNP mapped at position
189 136,533,558 bp for Rsb-based test (Rsb=6.13) and at position 136,523,244 bp for the XP-EHH-based test
190 (XP-EHH=5.59). For this latter SNP (mapping to region #7 as defined below), the haplotype bifurca-
191 tion diagrams for the ancestral and derived alleles within the CEU population are plotted in (Figure 1C)
192 and (Figure 1D), respectively, using the `bifurcation.diagram` function from the REHH package. Note
193 the extent of haplotype homozygosity associated with the derived allele (Figure 1C), relatively to that
194 associated with the ancestral allele (Figure 1D), which is consistent with the negative iHS measure at
195 this SNP (iHS=-3.24).

196 [Figure 1 about here.]

197 To further identify regions displaying strong footprints of selection, we split the HSA2 chromosome
198 into 950 consecutive 500 kb-windows (with a 250 kb overlap). Windows with at least 2 SNPs displaying
199 a statistic > 4 (in absolute value that roughly corresponds to a two-sided p -value $< 10^{-4}$, see above) for
200 at least one of the four test statistics were deemed significant. Significant overlapping windows were then
201 merged, leading to a total of 11 regions harboring strong signals of selection, which characteristics are

202 detailed in Table 1 (see also Figure 1). As expected, most of the regions identified in previously published
203 genome-scans for samples with the same origin (Sabeti *et al*, 2007; Tang *et al*, 2007; Voight *et al*, 2006)
204 overlap with the regions identified here (Table 1). For instance, regions #6 and #7 that are in the vicinity
205 of the EDAR gene (under selection in Asian populations) and the LCT gene (under selection in European
206 populations), respectively, have been extensively characterized in the literature (e.g., Peter *et al*, 2012).
207 Interestingly, we detected more regions than previously reported in the aforementioned studies, most
208 probably because our analyses are based on a larger dataset and different cut-off values. A more detailed
209 description of the newly identified regions is however beyond the scope of the present article.

210 [Table 1 about here.]

211 Note finally that XP-EHH- and Rsb-based scans gave consistent results, with the exception of the
212 region in the vicinity of the LCT gene (#7 in Table 1 and Figure 1) where a double peak was observed
213 with Rsb (consistent with the iHS profile within CEU) and a single peak with XP-EHH. The overall
214 correlation between these statistics was equal to 0.843, which illustrates the close similarity of these two
215 metrics.

216 **Comparing the performances of REHH 2.0 relatively to REHH 1.13, selscan and** 217 **hapbin packages**

218 The two CEU and JPT+CHB human data sets were further analyzed with REHH 1.13 to evaluate the
219 gain in computation time resulting from the modifications introduced in version 2.0. Note that extensive
220 tests were done during the development of version 2.0, to ensure that the same estimates were obtained
221 with both versions. Only very marginal differences were however sometimes observed in the estimates
222 of iES^{Tang} . For instance, the correlation between the resulting Rsb computed across the CEU and
223 JPT+CHB populations with version REHH 1.13 and REHH 2.0 was found equal to 0.999992 (instead of
224 1.0). This is actually due to the introduction of the computation of iES^{Sabeti} in version 2.0 to estimate
225 XP-EHH. Indeed, we chose to define the same cut-off value for both statistics during the computation of
226 the component variable EHHS (controlled with the option `limehhs`, set to 0.05 by default).

227 **An improved processing of the input file**

228 The first major modification introduced in REHH version 2.0 deals with the processing of input files
229 (haplotype and SNP information files) using the function `data2hap1ohh`. Indeed our own experience with
230 earlier versions of the package together with feedback from several users prompted us to optimize data
231 import and to improve allele recoding, which was inefficient in version 1.x. Considering standard input
232 haplotype file format (which is common to both versions), and with alleles encoded in the appropriate
233 format (`{0,1,2}` for missing data, ancestral and derived alleles respectively), the new `data2hap1ohh`
234 function is about 2.5 times faster than the previous one (see Table 2). In addition, the allele recoding
235 option results in slightly better processing performances, and is no more prone to errors as in version 1.x.
236 Finally, the new haplotype format (with haplotypes in columns), corresponding to the output file of the
237 SHAPEIT phasing program (O'Connell *et al*, 2014), was found to be the most efficient to process (see
238 Table 2).

239 [Table 2 about here.]

240 With datasets of increasing complexity and size, such improvement in the processing of input files is
241 critical to REHH users. Processing a data set as large as the JPT+CHB one (consisting of 342 haplotype
242 with 110,200 SNPs) now takes less than 12 seconds. Note however that for this file a maximum of about
243 1 Gb RAM was used, for a net memory size change of 240 Mb. For larger data sets, RAM requirements
244 may therefore be limiting for some computers.

245 **A faster and parallel algorithm to explore haplotype variability**

246 The second major modification introduced in REHH version 2.0 concerns the core algorithm that computes
247 the distribution of haplotype counts, which underlies the calculation of all the metrics of interest (iHS,
248 Rsb and XP-EHH). As shown in Table 3, this new algorithm allows to decrease the computation times
249 by more than one order of magnitude, as compared to the algorithm implemented in REHH version 1.13.
250 Hence, for the computation of iHS in the CEU population (respectively, the JPT+CHB population) on
251 a single thread, computation times were 13.7 (respectively 21.8) times smaller on average. Interestingly,
252 the computation time for the JPT+CHB dataset (which is approximately 1.34 times larger than the CEU
253 one in terms of number of SNPs \times number of haplotype) was only 1.09 times slower than for the latter.
254 Conversely, the computation time was 1.73 times slower for JPT+CHB relatively to CEU with REHH

255 version 1.13. Although a more detailed profiling of the algorithm would be required, these results suggest
256 that computational burden is approximately linearly related to the data set complexity.

257 To further improve computational speed, haplotype structure is now performed using OpenMP par-
258 allelization across SNPs in genome-wide scans. Using four threads then lead to an additional decrease of
259 about 3.5 times in computation times (see Table 3).

260 Overall, the whole analysis of the HSA2 haplotype files used in this study took about 1.5 minutes
261 (including the processing of input files) with REHH 2.0, and more than 1.3 hours with REHH 1.3. This
262 corresponds to the computation of iHS within the CEU and within the JPT+CHB populations, as well
263 as the computation of Rsb and XP-EHH.

264 [Table 3 about here.]

265 **Comparing REHH 2.0 to the selscan and hapbin programs**

266 Finally, we compared REHH 2.0 with **selscan** (Szpiech & Hernandez, 2014) and **hapbin** (Maclean *et al*,
267 2015), which were recently published. Both programs are written in C++ language and include paral-
268 lellisation. Computation times for the different analyses, either on a single or four threads, are provided
269 in Table 3. The new version of REHH outperforms **selscan** by about one order of magnitude. Moreover,
270 running REHH on a single thread is still more than twice as fast as running **selscan** on four threads. It
271 should also be noticed that running a full analysis consisting of the estimation of iHS within and XP-EHH
272 between the CEU and JPT+CHB populations result in a significant additional burden with **selscan** (Ta-
273 ble 3). Conversely, **hapbin** was found to be more than five times faster than REHH 2.0, most likely as a
274 result of its more efficient algorithm to explore haplotype variability. Yet, given the small computation
275 times achieved by both programs, REHH 2.0 remains competitive relative to **hapbin** for most practical
276 applications.

277 Correlation between the estimated iHS and XP-EHH obtained with the different programs are given in
278 Table 4. Estimates of XP-EHH were in almost perfect agreement among the different software packages.
279 Similarly, estimates for iHS were almost the same between REHH 2.0 and **selscan** but slightly depart from
280 those obtained with **hapbin**. Although we did not further investigate the origin of these discrepancies,
281 this might probably be related to a different definition of haplotype homozygosity in **hapbin**, as compared
282 to Sabeti *et al* (2007).

283

[Table 4 about here.]

284 Conclusion

285 Although the R package REHH (Gautier & Vitalis, 2012) has been widely used since its first release, the
286 increasing dimension of haplotype datasets typically available in most species led to serious limitations.
287 This stimulated the development of alternative R-free solutions (Szpiech & Hernandez, 2014; Maclean
288 *et al*, 2015). In this study, we introduced substantial changes in the REHH package to improve its
289 computational efficiency by one to several orders of magnitude. This was achieved by modifying the
290 processing of the input files and, most importantly, by improving and parallelizing the core algorithm
291 that computes the distribution of haplotype counts. As a result, REHH 2.0 clearly outperforms the
292 `selscan` (Szpiech & Hernandez, 2014) package and competes with `hapbin` (Maclean *et al*, 2015), the
293 fastest program to date. A decisive advantage of REHH 2.0 over these programs is that it allows working
294 within the multi-platform R environment. As such, it benefits from several graphical tools that facilitate
295 visual interpretation of the results.

296 REHH 2.0 is available from the CRAN repository ([http://cran.r-project.org/web/packages/
297 rehh/index.html](http://cran.r-project.org/web/packages/rehh/index.html)). A help file together with a detailed vignette manual (the current version is pro-
298 vided as a Supplementary File S2) are included in the package.

299 Acknowledgment

300 We wish to thank all users of the previous version for their feedback that helped to improve the package.
301 This work was supported in part by a grant of the German Science Foundation (DFG-SFB680) to AK.

References

- 302
- 303 Barson NJ, Aykanat T, Hindar K, *et al* (2015) Sex-dependent dominance at a single locus maintains
304 variation in age at maturity in salmon. *Nature*, **528**, 405–408.
- 305 Bosse M, Megens HJ, Madsen O, *et al* (2015) Using genome-wide measures of coancestry to maintain
306 diversity and fitness in endangered and domestic pig populations. *Genome research*, **25**, 970–981.
- 307 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic
308 marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, **12**, 499–510.
- 309 Flori L, Thevenon S, Dayo GK, *et al* (2014) Adaptive admixture in the west african bovine hybrid zone:
310 insight from the borgou population. *Mol Ecol*, **23**, 3241–3257.
- 311 Gautier M, Naves M (2011) Footprints of selection in the ancestral admixture of a new world creole cattle
312 breed. *Mol Ecol*, **20**, 3128–3143.
- 313 Gautier M, Vitalis R (2012) rehh: an r package to detect footprints of selection in genome-wide snp data
314 from haplotype structure. *Bioinformatics*, **28**, 1176–1177.
- 315 Jin J, Lee M, Bai B, *et al* (2016) Draft genome sequence of an elite dura palm and whole-genome patterns
316 of dna variation in oil palm. *DNA Research*, p. in press.
- 317 Maclean CA, Hong NPC, Prendergast JGD (2015) hapbin: An efficient program for performing haplotype-
318 based scans for positive selection in large genomic datasets. *Mol Biol Evol*, **32**, 3027–3029.
- 319 Mueller JC, Kuhl H, Timmermann B, Kempnaers B (2016) Characterization of the genome and transcrip-
320 tome of the blue tit *Cyanistes caeruleus*: polymorphisms, sex-biased expression and selection signals.
321 *Molecular ecology resources*, **16**, 549–561.
- 322 O’Connell J, Gurdasani D, Delaneau O, *et al* (2014) A general approach for haplotype phasing across
323 the full spectrum of relatedness. *PLoS Genet*, **10**, e1004234.
- 324 Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing
325 variation and from a de novo mutation. *PLoS Genet*, **8**, e1003011.
- 326 R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foun-
327 dation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

- 328 Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine
329 stickleback fish. *Nature communications*, **6**, 8767.
- 330 Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, *et al* (2002) Detecting recent positive selection
331 in the human genome from haplotype structure. *Nature*, **419**, 832–837.
- 332 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, *et al* (2007) Genome-wide detection and charac-
333 terization of positive selection in human populations. *Nature*, **449**, 913–918.
- 334 Szpiech ZA, Hernandez RD (2014) selscan: an efficient multithreaded program to perform ehh-based
335 scans for positive selection. *Mol Biol Evol*, **31**, 2824–2827.
- 336 Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent
337 positive selection in the human genome. *PLoS Biol*, **5**, e171.
- 338 The International HapMap3 Consortium (2010) Integrating common and rare genetic variation in diverse
339 human populations. *Nature*, **467**, 52–58.
- 340 Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annu Rev Genet*,
341 **47**, 97–120.
- 342 Voight BF, Kudravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human
343 genome. *PLoS Biol*, **4**, e72.
- 344 Wang M, Yu Y, Haberer G, *et al* (2014) The genome sequence of african rice (*oryza glaberrima*) and
345 evidence for independent domestication. *Nature genetics*, **46**, 982–988.

346 **Supplementary Material**

- 347 • File S1: compressed archive named `FileS1.tar.gz` containing example input haplotype data and
348 SNP information files in the REHH, `selscan` and `hapbin` format. The archive also contains command
349 lines that were used to run the different programs
- 350 • File S2: Detailed user manual (vignette) for the REHH 2.0.

351 **List of Figures**

352 1 **Analysis of the human chromosome 2 haplotype data sets (hg18 human genome**
353 **assembly) for the CEU and JPT+CHB populations with REHH 2.0.** A) Plot of iHS
354 against physical distance, in the CEU ($|iHS|$ in blue) and the JPT+CHB ($-|iHS|$ in red)
355 populations. B) Plot of XP-EHH ($|XP-EHH|$ in blue) and Rsb ($-|Rsb|$ in red) between the
356 CEU and JPT+CHB populations. In A) and B), the horizontal dotted lines indicate the
357 $|iHS|$ significance threshold of 4 that was used to identify significant regions (see Table 1)
358 and the arrows at the top of the graph indicate the mid-position of the significant regions
359 described in Table 1). C) and D) Haplotype bifurcation diagrams drawn for the ancestral
360 and derived allele, respectively, of the rs7377606 SNP in the CEU population (XP-EHH
361 peak position of region #7 described in Table 1 and containing the LCT gene). In C) and
362 D), the two grey vertical dotted lines delimit the LCT gene. 19

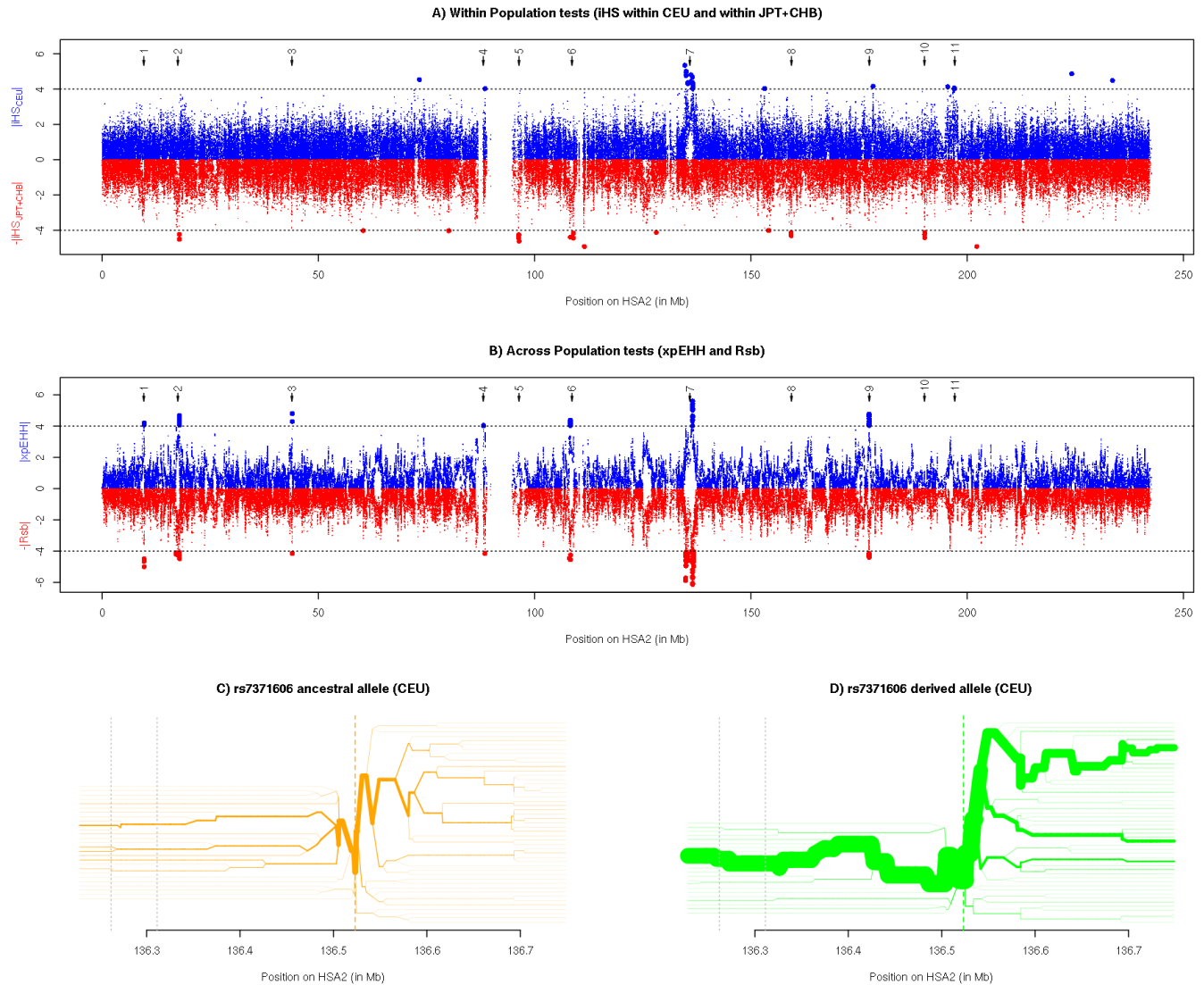


Figure 1. Analysis of the human chromosome 2 haplotype data sets (hg18 human genome assembly) for the CEU and JPT+CHB populations with REHH 2.0. A) Plot of iHS against physical distance, in the CEU ($|iHS|$ in blue) and the JPT+CHB ($-|iHS|$ in red) populations. B) Plot of XP-EHH ($|XP-EHH|$ in blue) and Rsb ($-|Rsb|$ in red) between the CEU and JPT+CHB populations. In A) and B), the horizontal dotted lines indicate the $|iHS|$ significance threshold of 4 that was used to identify significant regions (see Table 1) and the arrows at the top of the graph indicate the mid-position of the significant regions described in Table 1). C) and D) Haplotype bifurcation diagrams drawn for the ancestral and derived allele, respectively, of the rs7377606 SNP in the CEU population (XP-EHH peak position of region #7 described in Table 1 and containing the LCT gene). In C) and D), the two grey vertical dotted lines delimit the LCT gene.

363 **List of Tables**

364	1	Regions of HSA2 harboring strong signals of selection.	21
365	2	Comparison of the computation times (in seconds) required to process input	
366		data files with the <i>data2haplohh</i> function for the versions 1.13 and 2.0 of the	
367		REHH package. Two data sets consisting, respectively, of 236 and 342 haplotypes of	
368		110,200 SNPs for the CEU and JPT+CHB populations were considered (see the main	
369		text). For each of these datasets, the table gives the average computation times (\pm stan-	
370		dard deviation) across ten independent runs, either with or without (in parentheses) allele	
371		recoding (using the option <code>allele.recode</code>).	22
372	3	Comparison of the time (in seconds) required to compute the different EHH-	
373		based statistics for the versions 1.13 and 2.0 of the REHH, the <i>selscan</i> and the <i>hapbin</i>	
374		packages. For each analysis, the table gives the average computation time (\pm standard	
375		deviation) across ten independent runs. For each program, analyses were run either on a	
376		single thread or on four threads (except for REHH 1.13 version, which is not parallelised) .	23
377	4	Correlation between the estimated <i>iHS</i> and XP-EHH statistics across the programs	
378		REHH (version 2.0), <i>selscan</i> and <i>hapbin</i>. Pairwise correlation for the <i>iHS</i> computed in	
379		the CEU and the JPT+CHB (in parenthesis) populations are given in the upper diag-	
380		onal. Pairwise correlation for the XP-EHH computed across the CEU and JPT+CHB	
381		populations are given in the lower diagonal.	24

TABLES

ID	Position ^a (Size)	Candidate Gene (position)	Test	Peak Position ^b	Selected Population (Overlap with other studies ^c)
1	9.250-10.00 (0.75)	YWHAQ (9.641-9.688)	XP-EHH	9.700 (-4.21; 3)	JPT+CHB
			Rsb	9.701 (-5.00; 3)	
			iHS _{CEU}	9.700 (2.18; 0)	
			iHS _{JPT+CHB}	9.732 (3.52; 0)	
2	16.75-18.25 (1.50)	MSGN1 (17.861-17.862)	XP-EHH	17.871 (-4.68; 17)	JPT+CHB
			Rsb	17.890 (-4.49; 8)	
			iHS _{CEU}	18.150 (3.68; 0)	
			iHS _{JPT+CHB}	17.856 (4.51; 2)	
3	43.50-44.25 (1.75)	ABCG8 (43.919-43.959)	XP-EHH	43.955 (-4.80; 2)	JPT+CHB
			Rsb	43.957 (-4.15; 1)	
			iHS _{CEU}	44.177 (-3.13; 0)	
			iHS _{JPT+CHB}	43.783 (3.86; 0)	
4	87.75-88.50 (0.75)	SMYD1 (88.148-88.194)	XP-EHH	88.173 (-4.04; 2)	JPT+CHB
			Rsb	88.187 (-3.53; 0)	
			iHS _{CEU}	88.173 (-2.80; 0)	
			iHS _{JPT+CHB}	88.198 (-3.01; 0)	
5	96.00-96.75 (0.75)	NCAPH (96.365-96.405)	XP-EHH	96.281 (2.32; 0)	JPT+CHB
			Rsb	96.281 (3.07; 0)	
			iHS _{CEU}	96.609 (-3.88; 0)	
			iHS _{JPT+CHB}	96.403 (-4.61; 4)	
6	108.00-109.25 (1.25)	SULT1C2 (108.271-108.292) EDAR (108.877-108.972)	XP-EHH	108.273 (-4.38; 18)	JPT+CHB (Vo., Ta., Sa.)
			Rsb	108.253 (-4.55; 3)	
			iHS _{CEU}	109.016 (2.46; 0)	
			iHS _{JPT+CHB}	108.982 (4.44; 4)	
7	134.50-137.25 (2.75)	LCT (136.262-136.311) MCM6 (136.314-136.335)	XP-EHH	136.523 (5.60; 16)	CEU (Vo., Ta., Sa.)
			Rsb	136.533 (6.13; 71)	
			iHS _{CEU}	134.706 (-5.35; 19)	
			iHS _{JPT+CHB}	134.727 (-3.70; 0)	
8	159.00-159.75 (0.75)	PKP4 (159.021-159.246)	XP-EHH	159.381 (-2.96; 0)	JPT+CHB
			Rsb	159.380 (-2.86; 0)	
			iHS _{CEU}	159.745 (2.86; 0)	
			iHS _{JPT+CHB}	159.293 (4.31; 2)	
9	177.00-177.75 (0.75)	n.a.	XP-EHH	177.338 (-4.77; 16)	JPT+CHB (Sa.)
			Rsb	177.337 (-4.40; 7)	
			iHS _{CEU}	177.336 (-2.57; 0)	
			iHS _{JPT+CHB}	177.108 (3.43; 0)	
10	189.75-190.50 (0.75)	SLC40A1 (190.133-190.154)	XP-EHH	190.040 (0.58; 0)	JPT+CHB
			Rsb	189.919 (0.99; 0)	
			iHS _{CEU}	190.326 (2.92; 0)	
			iHS _{JPT+CHB}	190.177 (4.41; 3)	
11	196.75-197.50 (0.75)	HECW2 (196.772-197.166)	XP-EHH	196.794 (2.11; 0)	CEU (Ta.)
			Rsb	196.755 (2.08; 0)	
			iHS _{CEU}	197.030 (4.06; 2)	
			iHS _{JPT+CHB}	197.332 (2.32; 0)	

^aAll the position are given in Mb with respect to the hg18 human genome assembly

^bIn parentheses: the value of the test statistics at the peak position; the number of SNPs in the window that have a test statistic (in absolute value) above the threshold of 4

^cSignificant tests of selection found in other studies for the same regions are indicated: Vo. stands for Voight *et al* (2006); Ta. stands for Tang *et al* (2007) and Sa. stands for Sabeti *et al* (2007)

Table 1. Regions of HSA2 harboring strong signals of selection.

	haplotype format	CEU haplotypes	CHB+JPT haplotypes
REHH 1.13	standard	>36000 ^a (29.97 ± 0.29)	>36000 ^a (34.62 ± 0.60)
REHH 2.0	standard	9.858 ± 0.39 (10.73 ± 0.16)	14.56 ± 0.17 (15.61 ± 0.26)
REHH 2.0	transposed ^b	7.882 ± 0.10 (8.832 ± 0.50)	11.80 ± 0.20 (12.91 ± 0.14)

^aAs mentioned in the manual, REHH version 1.x is quite inefficient in allele recoding. Versions 1.x are also prone to error (e.g., if some alleles are coded as "T").

^busing the new option `haplotype.in.columns=T`

Table 2. Comparison of the computation times (in seconds) required to process input data files with the *data2haplohh* function for the versions 1.13 and 2.0 of the REHH package. Two data sets consisting, respectively, of 236 and 342 haplotypes of 110,200 SNPs for the CEU and JPT+CHB populations were considered (see the main text). For each of these datasets, the table gives the average computation times (± standard deviation) across ten independent runs, either with or without (in parentheses) allele recoding (using the option `allele.recode`).

program	#threads	iHS _{ceu}	iHS _{chb+jpt}	XP-EHH	Rsb	Total ^a
REHH 1.13	1	1759 ± 29	3045 ± 31	n.a.	4803 ± 58	4805 ± 58
REHH 2.0	1	128 ± 1.0	140 ± 2.1	268 ± 1.8	268 ± 1.8	269 ± 1.8
	4	37.8 ± 0.3	40.2 ± 0.3	77.1 ± 0.5	77.1 ± 0.5	78.5 ± 0.5
selscan	1	1237 ± 17	1503 ± 29	3833 ± 100	n.a.	6573 ± 86
	4	324 ± 6.5	391 ± 6.5	969 ± 5.6	n.a.	1684 ± 9.3
hapbin	1	17.6 ± 0.2	20.0 ± 0.1	47.4 ± 0.2	n.a.	85.0 ± 0.3
	4	5.68 ± 0.7	7.42 ± 0.1	13.2 ± 0.0	n.a.	26.2 ± 0.7

^aIn REHH the function `scanhh` computes iHH and iES simultaneously. It therefore needs to be run only once per haplotype data set. As a result, computing XP-EHH (and/or Rsb) requires almost no extra time, once iHS for the two populations has been computed.

Table 3. Comparison of the time (in seconds) required to compute the different EHH-based statistics for the versions 1.13 and 2.0 of the REHH, the selscan and the hapbin packages. For each analysis, the table gives the average computation time (\pm standard deviation) across ten independent runs. For each program, analyses were run either on a single thread or on four threads (except for REHH 1.13 version, which is not parallelised)

	REHH	selscan	hapbin
REHH	<i>na</i>	0.991 (0.993)	0.907 (0.945)
selscan	0.985	<i>na</i>	0.907 (0.945)
hapbin	0.986	0.994	<i>na</i>

Table 4. Correlation between the estimated iHS and XP-EHH statistics across the programs REHH (version 2.0), **selscan and **hapbin**.** Pairwise correlation for the iHS computed in the CEU and the JPT+CHB (in parenthesis) populations are given in the upper diagonal. Pairwise correlation for the XP-EHH computed across the CEU and JPT+CHB populations are given in the lower diagonal.