HaploSweep v1.0 User Manual

July 11, 2024

Contents

1 Introduction

HaploSweep is a method or detecting and categorizing soft and hard selective sweeps based on haplotype structure. HaploSweep builds on and advances the extended haplotype homozygosity (EHH) methods (Sabeti et al., 2002), adapting to the complexity of soft sweeps. Taking into account the multi-founder clustering characteristic of haplotypes carrying the beneficial allele during soft selective sweeps, we propose novel statistics, integrated haplotype homozygosity for local clusters (iHHL) and a logarithmic ratio variant (iHSL). In addition, we introduce RiHS, which represents the logarithmic ratio between iHHL and iHH, to facilitate the classification of sweep types. We also introduce the composite statistic RiHSL = $\text{iHSL}^2 + \text{RiHS}^2$, which follows a chi-square distribution with 2 degrees of freedom. RiHSL exhibits higher power than iHSL in simulations with a low beneficial allele frequency, but shows lower power in simulations with a high beneficial allele frequency. Users of HaploSweep have the option to access both iHSL and RiHSL values for their analysis.

If you have any issues or suggestions with the software, please get in touch with Shilei Zhao at zhaoshilei2018d@big.ac.cn.

2 Installation

HaploSweep can run on Linux platforms. The HaploSweep software is freely available at http://github.com/ChenHuaLab/HaploSweep. Download the file "HaploSweep-1.0.tar.gz" from the release HaploSweep-v1.0 and execute the following commands:

```
tar -zxvf HaploSweep-1.0.tar.gz
cd HaploSweep-1.0
make
```

After completing the setup, you can test the software by running it with the toy example:

```
mkdir out_neutral
./HaploSweep calc -i testdata/neutral.vcf.gz -m testdata/neutral.map -t 8 -o1
out_neutral/result.out
./HaploSweep norm -dir out_neutral -t 1 -out result_norm.out
```

The program takes a few minutes to complete on an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, and generates output files "result.out" and "result_norm.out".

3 Command line arguments

3.1 HaploSweep calc

To obtain the value of uiHS, uiHSL, and uRiHS, enter the following command:

```
./HaploSweep calc [arguments]
```

Inputs:

-i input genetic data file (.vcf .vcf.gz .hap .tped)

The detailed description of the input genetic data format is provided in sections 4.1-4.3.

-m input genetic map file (not required for .tped)

The detailed description of the genetic map file is provided in section 4.4.

-o1 output file

Specify the path and name for the output file. A detailed description of the output file is provided in section 5.1.

Options:

-w (int) window size (default: 400)

Specify the window size for the determination of clusters (default: 400 SNPs upstream and downstream from the core SNP).

-e (int) max extend distance (default: 1000000)

The maximum allowable distance for the EHH decay curve to extend from the core SNP.

-t (int) thread number(default: 8)

Specify the number of threads. Test results for runtime across different sample sizes and speedup across different threads are shown in Tables 1 and 2.

-region (int) start (include)-end (include), 1-based (default: full region)

Specify the region for running HaploSweep; by default, it runs on the entire region of the input data.

-trunc-ok (int) 1: retain the locus which not extend to maximum (default: 0)

If you specify "-trunc-ok 1", the locus will be retained regardless of the -trunc setting.

-trunc (double) truncation proportion (default: 0.5)

Discard the locus if the proportion of EHH decay for clusters does not extend to the maximum distance smaller than this value.

-maf (double) minor allele frequency (default: 0.05)

If the minor allele frequency (MAF) at a site falls below this threshold, the program will exclude it as a core SNP.

-cutoff (double) min EHH value (default: 0.05)

The EHH decay cutoff.

-gap-scale (int) gap scale(default: 20000)

Gap scale parameter in base pairs (bp). If a gap between two SNPs is encountered that is greater than GAP_SCALE and less than MAX_GAP, then the genetic distance is scaled by GAP_SCALE/GAP.

-max-gap (int) max extend gap (default: 200000)

Maximum permissible gap size between two SNPs, measured in base pairs (bp).

-o2 Output details of the EHH curves

A detailed description of the output file is provided in section xx. Warning: Avoid using -o2 for large genome regions, as this will result in a very large output file.

3.2 HaploSweep norm

To normalize the values of iHS, iHSL, and RiHS, use the following command:

```
./HaploSweep norm [arguments]
```

Inputs:

-dir input directory

Specify the directory containing the output files of the HaploSweep calc.

-out output file

Specify the path and name for the output file. The detailed description of the output file is provided in section xx.

Options:

-t (int) thread number (default: 8)

Specify the thread number.

-bin (double) bin size (default: 0.01)

Specify the allele frequency bin size for normalization. For a core locus with a derived allele frequency f_c, all loci with derived allele frequencies f such that f_c-bin/2 $\leq f \leq$ f_c+bin/2 are used for normalization.

Note: We recommend placing the results of all chromosomes into a single directory and normalizing them together.

4 Input Files

4.1 vcf format

A .vcf file is structured as follows (see details on https://github.com/samtools/hts-specs): it begins with meta-information lines (lines starting with "##"), followed by a header line (beginning with "#CHROM"), and then data lines that include marker and genotype data (one variant per line). The first nine columns contain locus information. For example:

```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  sample1 sample2  sample3
chr1  108 .  .  .  .  .  .  .    0|0   0|0   0|0
chr1  167 .  .  .  .  .  .  .    1|1   0|0   1|1
chr1  306 .  .  .  .  .  .  .    1|1   1|0   0|1
chr1  336 .  .  .  .  .  .  .    0|0   0|0   0|0
```

represents variant information across four loci for three diploid samples. Columns within the red box are essential for HaploSweep. **The genetic data is required to be phased**.

HaploSweep accepts various chromosome symbols, such as chr1, 1, chromosome1, etc. It is important to ensure consistent formatting across all input files.

4.2 hap format

A .hap file organizes variant data where each row represents a single haploid copy from an individual, and columns represent consecutive loci separated by whitespace. For example,

```
0 1 1 0
0 1 1 0
0 1 1 0
0 1 0 0
0 1 0 0
0 1 1 0
```

represents six haploid samples with variant information at four loci.

4.3 tped format

A tped file (transposed PLINK file, details available at http://pngu.mgh.harvard.edu/) is organized as follows:

<chr#> <id > <genetic position > <physical position > <haploid copy 1> ... <haploid copy N>

For example,

```
chr1 rs108    0.000108 108  0    0    0    0    0    0
chr1 rs167    0.000167 167  1    1    0    0    1    1
chr1 rs306    0.000306 306  1    1    1    0    0    1
chr1 rs336    0.000336 336  0    0    0    0    0    0
```

represents six haploid samples with variant information at four loci.

4.4 map format

Note that for .vcf and .hap genetic variant files, HaploSweep requires an additional file providing genetic position information. This file should have four columns: chromosome, rs, genetic position (cM), and physical position (bp), delimited by whitespace. Each row in the .vcf file and each column in the .hap file should correspond exactly to a row in the genetic variant file.

For example,

```
chr1 rs108 0.000108 108
chr1 rs167 0.000167 167
chr1 rs306 0.000306 306
chr1 rs336 0.000336 336
```

represents the positions of four SNPs in the example .vcf and .hap files mentioned earlier.

5 Output File

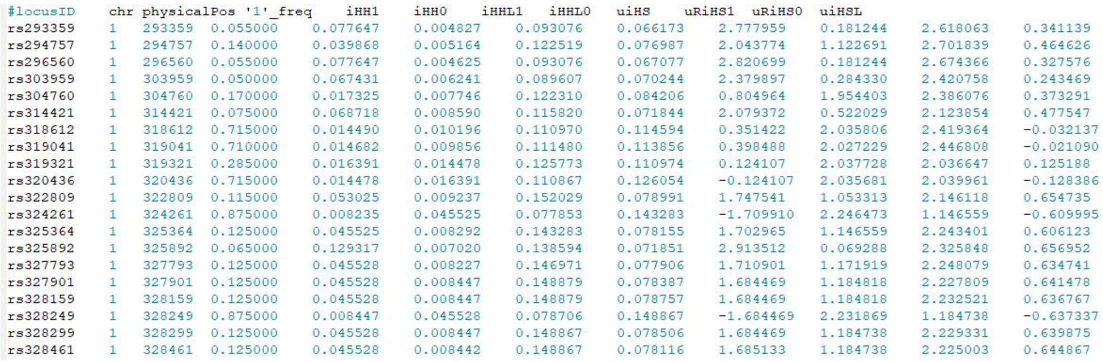
5.1 Output file -o1 for HaploSweep calc

HaploSweep calc produces a single output file for -o1. The file contains 12 columns organized as follows:

<#locusID> <chr> <physicalPos> <'1'_freq> <iHH1> <iHH0> <iHHL1> <iHHL0> <uiHS> <uRiHS1> <uRiHS0> <uiHSL>,

where '1'_freq denotes the derived allele frequency. iHH represents integrated haplotype homozygosity for haplotypes with the derived (1) or ancestral (0) allele. iHHL is the mean iHH for sub-clusters of haplotypes with the derived (1) or ancestral (0) allele. uiHS stands for unstandardized iHS. uRiHS represents the unstandardized RiHS for haplotypes with the derived (1) or ancestral (0) allele. uiHSL denotes the unstandardized iHSL.

For example,



```
#locusID    chr physicalPos '1'_freq    iHH1     iHH0     iHHL1    iHHL0    uiHS     uRiHS1   uRiHS0   uiHSL
rs293359    1   293359  0.055000    0.077647    0.004827    0.093076    0.066173    2.777959    0.181244    2.618063    0.341139
rs294757    1   294757  0.140000    0.039868    0.005164    0.122519    0.076987    2.043774    1.122691    2.701839    0.464626
rs296560    1   296560  0.055000    0.077647    0.004625    0.093076    0.067077    2.820699    0.181244    2.674366    0.327576
rs303959    1   303959  0.050000    0.067431    0.006241    0.089607    0.070244    2.379897    0.284330    2.420758    0.243469
rs304760    1   304760  0.170000    0.017325    0.007746    0.122310    0.084206    0.804964    1.954403    2.386076    0.373291
rs314421    1   314421  0.075000    0.068718    0.008590    0.115820    0.071844    2.079372    0.522029    2.123854    0.477547
rs318612    1   318612  0.715000    0.014490    0.010196    0.110970    0.114594    0.351422    2.035806    2.419364    -0.032137
rs319041    1   319041  0.710000    0.014682    0.009856    0.111480    0.113856    0.398488    2.027229    2.446808    -0.021090
rs319321    1   319321  0.285000    0.016391    0.014478    0.125773    0.110974    0.124107    2.037728    2.036647    0.125188
rs320436    1   320436  0.715000    0.014478    0.016391    0.110867    0.126054    -0.124107   2.035681    2.039961    -0.128386
rs322809    1   322809  0.115000    0.053025    0.009237    0.152029    0.078991    1.747541    1.053313    2.146118    0.654735
rs324261    1   324261  0.875000    0.008235    0.045525    0.077853    0.143283    -1.709910   2.246473    1.146559    -0.609995
rs325364    1   325364  0.125000    0.045525    0.008292    0.143283    0.078155    1.702965    1.146559    2.243401    0.606123
rs325892    1   325892  0.065000    0.129317    0.007020    0.138594    0.071851    2.913512    0.069288    2.325848    0.656952
rs327793    1   327793  0.125000    0.045528    0.008227    0.146971    0.077906    1.710901    1.171919    2.248079    0.634741
rs327901    1   327901  0.125000    0.045528    0.008447    0.148879    0.078387    1.684469    1.184818    2.227809    0.641478
rs328159    1   328159  0.125000    0.045528    0.008447    0.148879    0.078757    1.684469    1.184818    2.232521    0.636767
rs328249    1   328249  0.875000    0.008447    0.045528    0.078706    0.148867    -1.684469   2.231869    1.184738    -0.637337
rs328299    1   328299  0.125000    0.045528    0.008447    0.148867    0.078506    1.684469    1.184738    2.229331    0.639875
rs328461    1   328461  0.125000    0.045528    0.008442    0.148867    0.078116    1.685133    1.184738    2.225003    0.644867
```

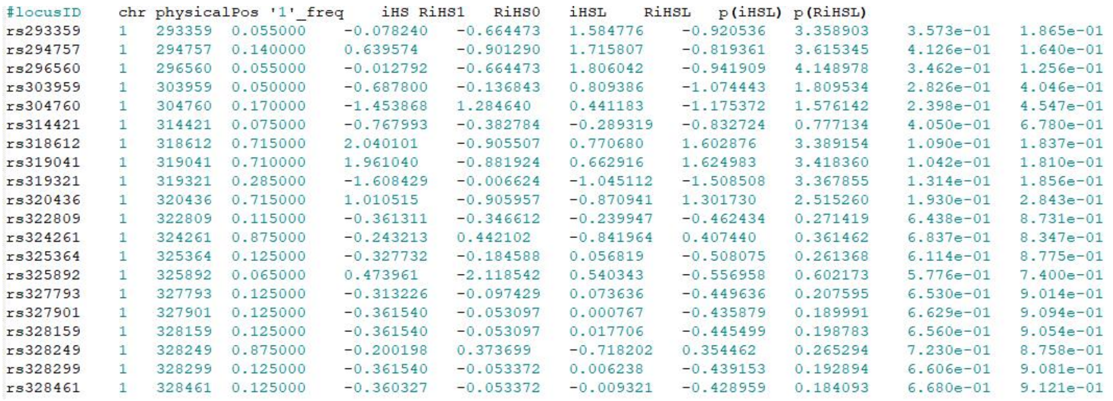
5.2 Output file -out for HaploSweep norm

The output file for HaploSweep norm contains 11 columns organized as follows:

<#locusID> <chr> <physicalPos> <'1'_freq> <iHS> <RiHS1> <RiHS0> <iHSL> <RiHSL> <p(iHSL)> <p(RiHSL)>,

where iHS, RiHS1, RiHS0, and iHSL denote the normalized values of uiHS, uRiHS1, uRiHS0, and uiHSL, respectively. If iHSL>0, then RiHSL=iHSL^2+RiHS1^2; otherwise, RiHSL=iHSL^2+RiHS0^2. iHSL follows a standard normal distribution, and RiHSL follows a chi-square distribution with 2 degrees of freedom (Figure 1). The p-value for iHSL is calculated as p(iHSL) = 2*Φ*(-|niHSL|), where *Φ* denotes the cumulative distribution function of the standard normal distribution. The p-value for RiHSL is calculated as p(RiHSL)=F(RiHSL|v=2), where F(x|v=2) denotes the cumulative distribution function of the chi-square distribution with 2 degrees of freedom.

For example,



```
#locusID    chr physicalPos '1'_freq    iHS RiHS1   RiHS0   iHSL    RiHSL   p(iHSL) p(RiHSL)
rs293359    1   293359  0.055000    -0.078240   -0.664473   1.584776    -0.920536   3.358903    3.573e-01   1.865e-01
rs294757    1   294757  0.140000    0.639574    -0.901290   1.715807    -0.819361   3.615345    4.126e-01   1.640e-01
rs296560    1   296560  0.055000    -0.012792   -0.664473   1.806042    -0.941909   4.148978    3.462e-01   1.256e-01
rs303959    1   303959  0.050000    -0.687800   -0.136843   0.809386    -1.074443   1.809534    2.826e-01   4.046e-01
rs304760    1   304760  0.170000    -1.453868   1.284640    0.441183    -1.175372   1.576142    2.398e-01   4.547e-01
rs314421    1   314421  0.075000    -0.767993   -0.382784   -0.289319   -0.832724   0.777134    4.050e-01   6.780e-01
rs318612    1   318612  0.715000    2.040101    -0.905507   0.770680    1.602876    3.389154    1.090e-01   1.837e-01
rs319041    1   319041  0.710000    1.961040    -0.881924   0.662916    1.624983    3.418360    1.042e-01   1.810e-01
rs319321    1   319321  0.285000    -1.608429   -0.006624   -1.045112   -1.508508   3.367855    1.314e-01   1.856e-01
rs320436    1   320436  0.715000    1.010515    -0.905957   -0.870941   1.301730    2.515260    1.930e-01   2.843e-01
rs322809    1   322809  0.115000    -0.361311   -0.346612   -0.239947   -0.462434   0.271419    6.438e-01   8.731e-01
rs324261    1   324261  0.875000    -0.243213   0.442102    -0.841964   0.407440    0.361462    6.837e-01   8.347e-01
rs325364    1   325364  0.125000    -0.327732   -0.184588   0.056819    -0.508075   0.261368    6.114e-01   8.775e-01
rs325892    1   325892  0.065000    0.473961    -2.118542   0.540343    -0.556958   0.602173    5.776e-01   7.400e-01
rs327793    1   327793  0.125000    -0.313226   -0.097429   0.073636    -0.449636   0.207595    6.530e-01   9.014e-01
rs327901    1   327901  0.125000    -0.361540   -0.053097   0.000767    -0.435879   0.189991    6.629e-01   9.094e-01
rs328159    1   328159  0.125000    -0.361540   -0.053097   0.017706    -0.445499   0.198783    6.560e-01   9.054e-01
rs328249    1   328249  0.875000    -0.200198   0.373699    -0.718202   0.354462    0.265294    7.230e-01   8.758e-01
rs328299    1   328299  0.125000    -0.361540   -0.053372   0.006238    -0.439153   0.192894    6.606e-01   9.081e-01
rs328461    1   328461  0.125000    -0.360327   -0.053372   -0.009321   -0.428959   0.184093    6.680e-01   9.121e-01
```

For the simulated neutral data, the Pearson correlation coefficient between iHSL and RiHS is 0.0063, with a p-value of 0.4162.

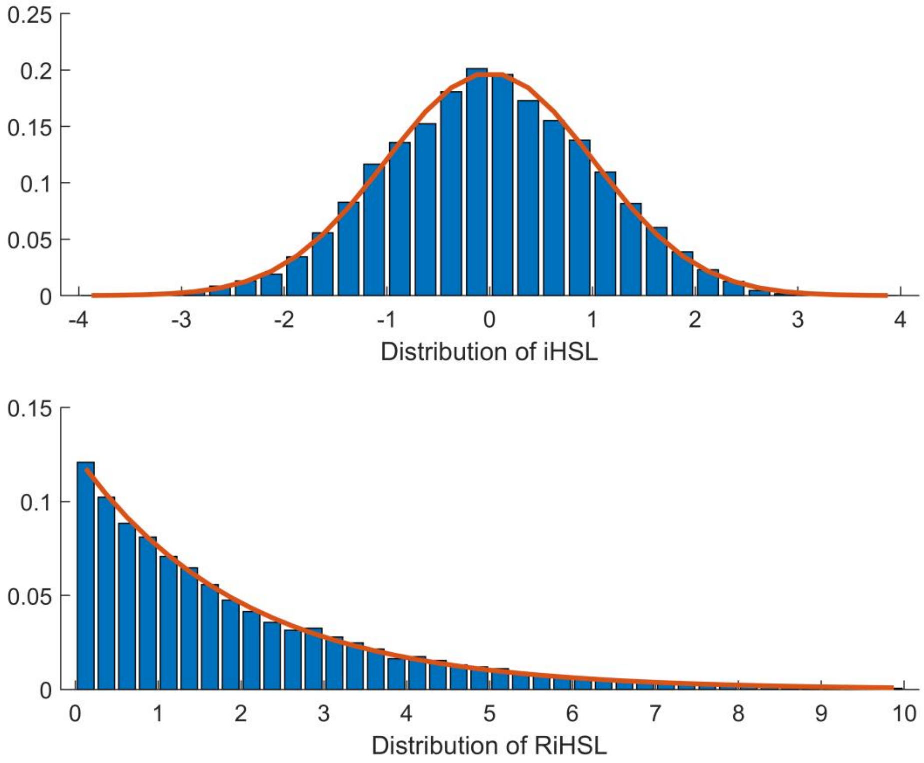


Figure 1. Distribution of iHSL and RiHSL for simulated neutral data. Top: Distribution of iHSL values with bars representing simulated data and a red line indicating the standard normal distribution. Below: Distribution of RiHSL values with bars representing simulated data and a red line indicating the chi-square distribution with 2 degrees of freedom.

5.3 Output file -o2 for HaploSweep calc

If you specify -o2, HaploSweep will produce data of traditional EHH curves and average EHH curve for clusters. The output file of -o2 contains 8 additional columns compared with -o1:

<#locusID> <chr> <physicalPos> <'1'_freq> <iHH1> <iHH0> <iHHL1> <iHHL0> <uiHS> <uRiHS1> <uRiHS0> <uiHSL> <pos1> <EHH1> <pos0> <EHH0> <pos1_cluster> <EHH1_cluster> <pos0_cluster> <EHH0_cluster>

where EHH1 denotes traditional EHH curve for haplotypes carrying derived allele, pos1 denotes the physical position for EHH1, EHH0 denotes traditional EHH curve for haplotypes carrying ancestral allele, and pos0 denotes the physical position for EHH0. EHH1_cluster, pos1_cluster, EHH0_cluster, and pos0_cluster denote the average EHH curve for clusters and the corresponding physical position.

```
1.  mkdir out_sweep
2.  ./HaploSweep calc -i ./testdata/sweep.vcf.gz -m ./testdata/sweep.map -t 8 -o1 ./out_sweep
    /result.out
3.  ./HaploSweep norm -dir out_sweep -t 1 -out result_sweep_norm.out
4.  ./HaploSweep calc -i ./testdata/sweep.vcf.gz -m ./testdata/sweep.map -region 10000000-10000000
    -t 1 -o2 result_sweep.detail.out
```

The file "sweep.vcf.gz" contains the simulated data under a soft sweep scenario. The simulated sweep data is 20 Mb long, with the adaptive allele located at the center of the genome (10 Mb). After running the commands in lines 1-3, HaploSweep produces the output file "result_sweep_norm.out". Figure 2 shows the p-values of iHSL and RiHSL from the file "result_sweep_norm.out". To view the EHH curve of the adaptive locus, you can run the command in line 4. By specifying "-region 10000000-10000000" and "-o2 result_sweep.detail.out", you can obtain the EHH curve for only the target locus. If you do not specify "-region", the EHH curve for all sites in the genetic data will be recorded, resulting in a very large output file. The EHH curves are shown in Figure 3.

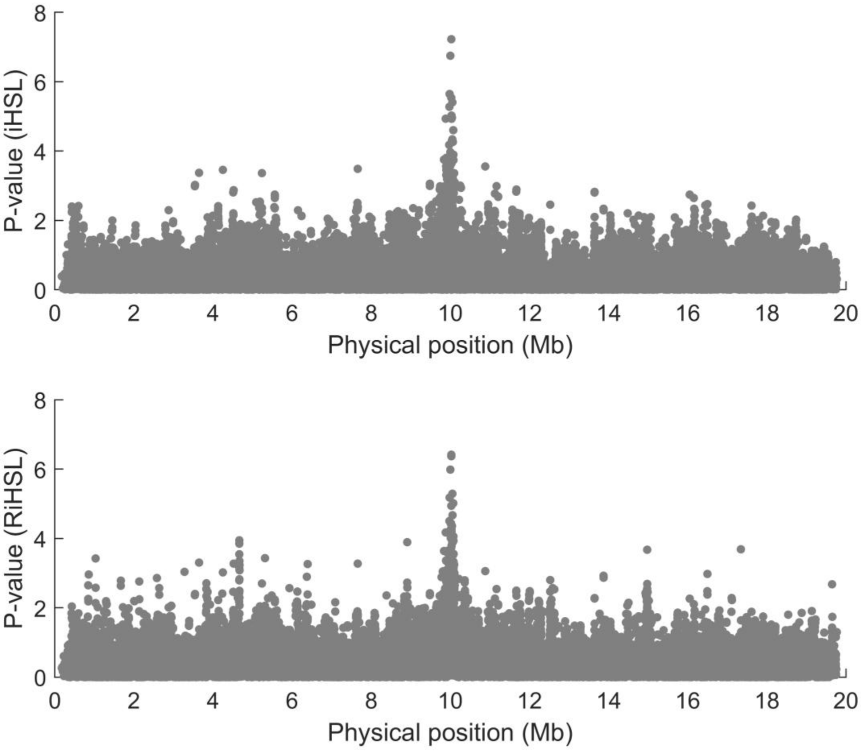


Figure 2. iHSL and RiHSL p-values

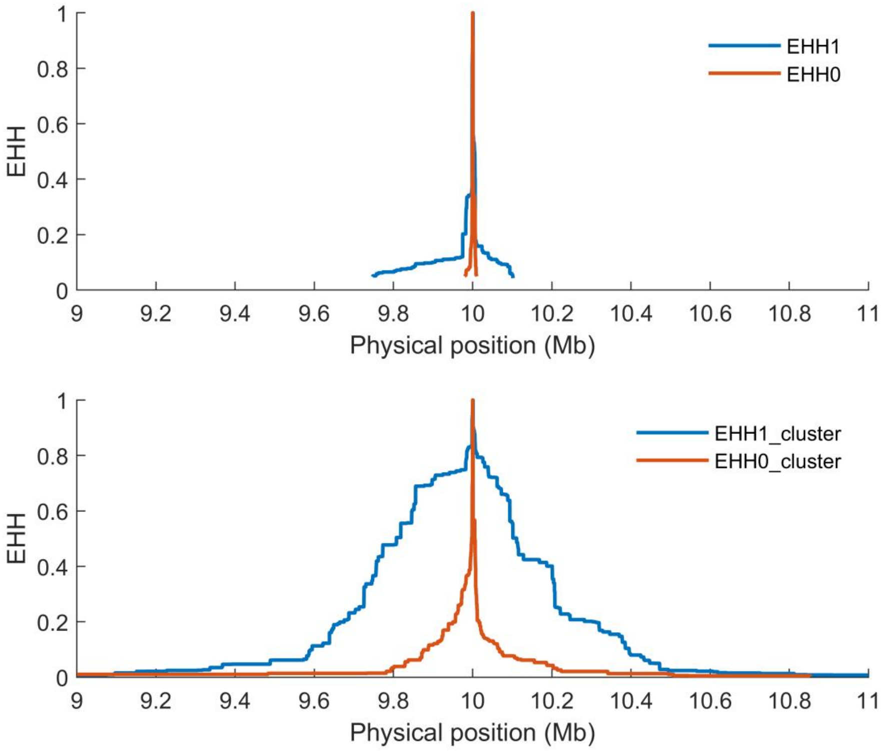


Figure 3. EHH curves

6 Runtime and speedup

The mean runtime of ten replicates under various sample sizes is presented in Table 1. Table 2 shows the speedup achieved using 2 to 10 threads (Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz).

Table 1. Runtime per 1,000 snps under different sample sizes

#haplotypes	50	100	200	400	600	800	1000
Run time (seconds)	6.84	18.98	46.47	120.08	276.01	410.60	635.27

Table 2. Runtime per 1,000 snps and speedup under different threads (200 haplotypes)

#threads	2	3	4	5	6	7	8	9	10
Run time (seconds)	22.69	15.42	11.69	9.61	8.26	7.11	6.38	6.00	5.44
Speed up	2.05	3.01	3.97	4.84	5.63	6.53	7.28	7.75	8.54

Reference

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature, 419(6909): 832–7.