Package ‘rehh’

May 13, 2015

Maintainer Mathieu Gautier <mathieu.gautier@supagro.inra.fr>
Author Mathieu Gautier and Renaud Vitalis
Version 1.13
License GPL (>= 2)
Title Searching for Footprints of Selection using Haplotype
  Homozygosity Based Tests
Description Functions for the detection of footprints of selection on
dense SNP data using Extended Homozygosity Haplotype (EHH)
based tests. The package includes computation of EHH, iHS
(within population) and Rsb (across pairs of populations)
statistics. Various plotting functions are also included to
facilitate visualization and interpretation of the results.
Depends R (>= 2.10), gplots, methods
NeedsCompilation yes
Repository CRAN
Date/Publication 2015-05-13 19:54:58

R topics documented:

  rehh-package .................................................. 2
  bifurcation.diagram .......................................... 3
  calc_ehh .......................................................... 4
  calc_ehhs ......................................................... 6
  data2haploh  .................................................. 7
  distribplot ..................................................... 9
  haplohh-class .................................................. 10
  haplohh_cgu_bta12 .............................................. 10
  ies2rsb .......................................................... 11
  ihh2ihs ........................................................ 12
  ihsplot ........................................................ 13
  make.example.files .......................................... 14
  rsbplot ........................................................ 15
  scan_hh ......................................................... 16
Searching for footprints of selection using Haplotype Homozygosity based tests

Description

Functions for the detection of footprints of selection on dense SNP data using Extended Homozygosity Haplotype (EHH) based tests. The package includes computation of EHH, iHS (within population) and Rsb (across pairs of populations) statistics. Various plotting functions are also included to facilitate visualization and interpretation of the results.

Details

Package:  rehh
Version:  1.1
License:  GPL(>=2)
Depends:  gplots, methods

Index:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bifurcation.diagram</td>
<td>plot of an haplotype bifurcation diagram</td>
</tr>
<tr>
<td>calc_ehh</td>
<td>EHH and iHH computations at a given core SNP</td>
</tr>
<tr>
<td>calc_ehhs</td>
<td>EHHS and iES computations at a given core SNP</td>
</tr>
<tr>
<td>data2haplohh</td>
<td>Converting data into an object of class haplohh</td>
</tr>
<tr>
<td>distribplot</td>
<td>Distribution of standardized iHS or Rsb values</td>
</tr>
<tr>
<td>haplohh_cgu_bta12</td>
<td>Example of an 'haplohh' object</td>
</tr>
<tr>
<td>ies2rsb</td>
<td>Compute Rsb (standardized ratio of iES from two populations)</td>
</tr>
<tr>
<td>ihh2ihs</td>
<td>Compute iHS (standardized iHH)</td>
</tr>
<tr>
<td>ihplot</td>
<td>Plot iHS over a genome</td>
</tr>
<tr>
<td>make_example.files</td>
<td>Generate example input files</td>
</tr>
<tr>
<td>rsbplot</td>
<td>Plot Rsb over a genome</td>
</tr>
<tr>
<td>scan_hh</td>
<td>Computing EHH based statistics over a whole chromosome</td>
</tr>
<tr>
<td>wgscan.cgu</td>
<td>Whole genome scan results for the CGU (Creole from Guadeloupe island)</td>
</tr>
</tbody>
</table>
wgscan.eut Whole genome scan results for a pool of European taurine cattle

References


bifurcation.diagram

Description

Haplotype Bifurcation diagram visualizes the breakdown of LD at increasing distances from the core allele at the selected focal SNPs.

Usage

```
bifurcation.diagram(haplohh, mrk_foc, all_foc=1, nmrk_l=10, nmrk_r=10, limhapcount=10, refsize=0.1, linecol="blue", main_leg=NA, xlab_leg="Position")
```

Arguments

- `haplohh` An object of class haplohh (see data2haplohh).
- `mrk_foc` Integer representing the number of the focal marker
- `all_foc` either 1 or 2 depending on the chosen core allele (resp. ancestral or derived)
- `nmrk_l` Number of markers to be considered upstream of the focal SNP
- `nmrk_r` Number of markers to be considered downstream of the focal SNP
- `limhapcount` Minimal number of haplotypes containing the core allele at the focal SNP
- `refsize` Controls the relative width of the diagram lines on the plot
- `linecol` Color of the lines on the diagram
- `main_leg` Main legend of the diagram. By default, the name of the SNP together with the allele considered
- `xlab_leg` Legend on the xaxis of the diagram

Details

Haplotype Bifurcation diagram visualizes the breakdown of LD at increasing distances from the core allele at the selected focal SNPs. The root (focal SNP) of each diagram is the core allele, identified by a vertical dashed line. The diagram is bi-directional, portraying both centromere-proximal and centromere-distal LD. Moving in one direction, each marker is an opportunity for a node; the diagram either divides or not based on whether both or only one allele is present. Thus the breakdown of LD on the core haplotype background is portrayed at progressively longer distances. The thickness of the lines corresponds to the number of samples with the indicated long-distance haplotype.
Value

The function returns a plot.

References


Examples

```r
#example haplohh object (280 haplotypes, 1424 SNPs)
#see ?haplohh_cgu_bta12 for details
data(haplohh_cgu_bta12)
#plotting bifurcation diagram for both ancestral and derived allele
#from the focal SNP at position 456
#which display a strong signal of selection
layout(matrix(c(1:2,2,1)))
#ancestral allele
bifurcation.diag(haplohh_cgu_bta12,mrk_foc=456,all_foc=1,
nmrk_l=20,rmrk_r=20)
#derived allele
bifurcation.diag(haplohh_cgu_bta12,mrk_foc=456,all_foc=2,
nmrk_l=20,rmrk_r=20)
##
dev.off()
```

calc_ehh

**EHH and iHH computations at a given core SNP**

Description

Compute Extended Haplotype Homozygosity (EHH) and integrated EHH (iHH) for a given focal SNPs.

Usage

```r
calc_ehh(haplohh,mrk,limhaplo=2,limehh=0.05,
plotehh=TRUE,main_leg="EHH plot")
```

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>haplohh</td>
<td>An object of class haplohh (see data2haplohh).</td>
</tr>
<tr>
<td>mrk</td>
<td>Integer representing the number of the focal marker</td>
</tr>
</tbody>
</table>
calc_ehh

limhaplo Minimal number of haplotypes to continue computing EHH away from the core SNP. Useless, if no missing data. However, when some data are missing, haplotypes with missing data are removed from the computation. Hence as we compute EHH further from the core SNP, less haplotypes are expected.

limehh Limit below which EHH stops to be evaluated

plotehh If TRUE, EHH estimates for both the ancestral and derived allele are plotted for each position

main_leg Legend of the EHH plot.

Details

EHH are computed at each position upstream and downstream the focal SNP for both the derived and ancestral allele. This allows in turn the computation of the integrated EHH relative to map distances (iHH).

Value

The returned value is a list containing the following components:

ehh A matrix of two rows and nsnps columns containing EHH estimates at each chromosome position relative to the focal SNP for the ancestral (first row) and derived (second row) alleles.

nhaplo_eval A matrix of two rows and nsnps columns containing the number of evaluated haplotypes at each chromosome position relative to the focal SNP for the ancestral (first row) and derived (second row) alleles.

ihh A vector of two elements corresponding respectively to the iHH (integrated EHH) for the ancestral and derived allele.

References


See Also

calc_ehhs, data2haplohh, scan_hh
Examples

#example haplohh object (280 haplotypes, 1424 SNPs)
#see ?haplohh_cgu_bta12 for details
data(haplohh_cgu_bta12)

#computing EHH statistics for the focal SNP at position 456
# which displays a strong signal of selection
res.ehh<-calc_ehh(haplohh_cgu_bta12, mrk=456)

calc_ehhs

EHHS and iES computations at a given core SNP

Description

Compute site Extended Haplotype Homozygosity (EHHS) and integrated EHH (iES) for a given focal SNPs.

Usage

calc_ehhs(haplohh, mrk, limhaplo=2, limehhs=0.05, plotehhs=TRUE, main_leg="EHHS plot")

Arguments

haplohh An object of class haplohh (see data2haplohh).

mrk Integer representing the number of the focal marker

limhaplo Minimal number of haplotypes to continue computing EHHS away from the core SNP. Useless, if no missing data. However, when some data are missing, haplotypes with missing data are removed from the computation. Hence as we compute EHH further from the core SNP, less haplotypes are expected

limehhs Limit below which EHHS stops to be evaluated

plotehhs If TRUE, EHHS estimates are plotted for each position

main_leg Legend of the EHHS plot

Details

EHHS are computed at each position upstream and downstream the focal SNP. This allows in turn the computation of the integrated EHHS relative to map distances (iES).

Value

The returned value is a list containing the following components:

ehhs A vector of nsnps columns containing EHHS estimates at each chromosome position relative to the focal SNP.
A matrix of two rows and nsnps columns containing the number of evaluated haplotypes at each chromosome position relative to the focal SNP for the ancestral (first row) and derived (second row) alleles.

freqani Frequency of the ancestral allele matrix for the focal SNP.

ies Integrated EHHS (IES) over the chromosome.

References


See Also
calc_ehh, data2haplohh, scan_hh

Examples

```r
#example haplohh object (280 haplotypes, 1424 SNPs)
#see ?haplohh_cgu_bta12 for details
data(haplohh_cgu_bta12)
#computing EHH statistics for the focal SNP at position 456
#which displays a strong signal of selection
res.ehhs<-calc_ehhs(haplohh_cgu_bta12, mrk=456)
```

data2haplohh  Converting data into an object of class haplohh

Description

Converts input file data into an object of class haplohh.

Usage

```r
data2haplohh(hap_file, map_file, min_maf=0, min_perc_genophap=100,
min_perc_genon.snp=100, chr.name=NA, popsel=NA, recode.allele=FALSE)
```

Arguments

- **hap_file** Path to the file containing haplotype data (see details section below for information about input file format)
- **map_file** Path to the file containing map information (see details section below for information about input file format)
- **min_maf** Threshold on Minor Allele Frequency (SNPs displaying a MAF<min_maf are discarded)
data2haplohh

`min_perc_genohap`  
Threshold on percentage of missing data for haplotypes (Haplotypes with less than `min_perc_genohap` percent SNPs genotyped are discarded)

`min_perc_genosnp`  
Threshold on percentage of missing data for SNPs (SNPs genotyped on less than `min_perc_genosnp` percent haplotypes are discarded)

`chr.name`  
Name of the chromosome considered (relevant if several chromosomes are represented in the map file)

`popsel`  
Code of the population considered in the fastPHASE output haplotype file (relevant if `hap_file` is a fastPHASE output and haplotypes originate from different population)

`recode.allele`  
If TRUE, allele in the haplotypes are recoded according to the map file information. If FALSE a rough verification is performed to check only 0 (code for missing data), 1 (code for ancestral allele) or 2 (code for derived allele) are present in the haplotype file

**Details**

Two haplotype input formats are supported: i) a standard format with haplotype in row and snps in column (with no header and a haplotype id) and ii) output files from fastPHASE program (Sheet and Stephens, 2006). The function automatically checks if the file is in fastPHASE output format. In this latter case, if haplotypes originate from several different population were phased simultaneously (-u fastPHASE option was used), the function ask interactively which population should be considered (a list of population number are proposed) unless specified with the `popsel` argument. Map file contains SNPs information in five columns SNP names, chromosome, position, ancestral and derived allele. SNPs must be in the same order as in the haplotype for the chromosome considered. If several chromosomes are represented in the map file, one can provide the name of the chromosome of interest (corresponding to the haplotype under study) with `chr.name` argument. Haplotype are recoded (if `recode.allele` option is activated) according to the ancestral and derived allele definition available in the map file (fourth and fifth columns) as : 0=missing data, 1=ancestral allele, 2=derived allele. If such a coding is detected, no recoding is performed. Note that Rsb statistics does not consider ancestral and derived allele status information. Finally, the arguments `min_perc_genohap`, `min_perc_genosnp` and `min_maf` are evaluated in this order.

**Value**

The returned value is an object of class `haplohh`

**See Also**

`calc_ehh,calc_ehhs,scan_hh,make.example.files`

**Examples**

```r
#Copy example files in the current working directory.
make.example.files()
#using the fastPHASE output haplotype example file
hap<-data2haplohh(hap_file="bta12_hapguess_switch.out",map_file="map.inp", min_maf=0.05,popsel=7,chr.name=12)
```
# Using the standard output haplotype example file

```r
hap <- data2haplohh(hap_file="bta12_cgu.hap", map_file="map.inp",
min_maf=0.05, chr.name=12)
```

## distribplot

### Distribution of standardized iHS or Rsb values

**Description**

Plot the observed distribution of standardized iHS or Rsb values together with the expected standard Gaussian distribution

**Usage**

```r
distribplot(data, col=c("blue","red"),
main="iHS distribution", xlab="iHS")
```

**Arguments**

- **data**: A vector of iHS or Rsb values.
- **col**: A vector describing color of the Observed and expected Gaussian distribution
- **main**: Character string for the plot legend
- **xlab**: Character string for the X-axis legend

**Value**

The function returns a plot.

**See Also**

- `scan_hh`, `ihs2ihs`, `ies2rsb`, `ihsplot`, `rsbplot`

**Examples**

```r
data(wgscan.cgu)
## results from a genome scan (44,057 SNPs) see ?wgscan.eut for details
val.ihs <- ihh2ihs(wgscan.cgu)$res.ihs[,3]
##standardize
distribplot(val.ihs, main="iHS (CGU population)")
dev.off()
```
Description
An object of class haplohh contains all relevant haplotype information (see below).

Objects from the Class
Objects can be created by calls of the form `new("haplohh", ...)`. 

Slots
- `haplo` Object of class "matrix": haplotypes with alleles coded as 0 (missing data), 1 (ancestral allele) or 2 (derived allele)
- `position` Object of class "numeric": position of the SNPs in the chromosome
- `snp.name` Object of class "character": names of the SNP
- `chr.name` Object of class "numeric": name of the chromosome SNPs are mapping to
- `nhap` Object of class "numeric": number of haplotypes
- `nsnp` Object of class "numeric": number of SNPs in the haplotypes

See Also
data2haplohh

Examples
`showClass("haplohh")`

Example of an haplohh object

Description
The object contains haplotype data for 140 cattle individuals (280 haplotypes) belonging to the Creole breed from Guadeloupe (CGU) and 1424 SNPs (mapping to chromosome BTA12).

Usage
data(haplohh_cgu_bta12)

References
ies2rsb

See Also
data2haplohh

ies2rsb  Compute Rsb (standardized ratio of iES from two populations)

Description

Compute Rsb (standardized ratio of iES from two populations).

Usage

ies2rsb(hh_pop1, hh_pop2, popname1=NA, popname2=NA, method="bilateral")

Arguments

hh_pop1  A matrix with nsnps rows and six columns (Chromosome name, position of the SNP, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES) obtained after performing a scan on the first population.

hh_pop2  A matrix with nsnps rows and six columns (Chromosome name, position of the SNP, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES) obtained after performing a scan on the second population.

popname1  Name of the first population compared (character string).

popname2  Name of the second population compared (character string).

method  Either "bilateral" or "unilateral". If bilateral (resp. unilateral), pvalue (assuming Rsb follows a standard Gaussian distribution under neutrality) corresponds to a bilateral (resp. unilateral) test.

Details

Ratio of iES (population 1 over population 2) standardized as described in Tang et al. (2007)

Value

The returned value is a matrix with nsnps rows and four columns (Chromosome name, position of the SNP, Rsb and Pvalue)

References


ihh2ihs

Compute iHS (standardized iHH)

Description

Compute iHS (standardized iHH).

Usage

ihh2ihs(res_ihh,freqbin=0.025,minmaf=0.05)

Arguments

res_ihh A matrix with nsnps rows and six columns (Chromosome name, position of the SNP, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES) obtained after performing a scan on the population of interest.

freqbin Size of the bin to standardize log(iHH1/iHH2) according to the underlying Derived Allele frequency. Allele frequency bins vary from minmaf to 1-minmaf per step of size freqbin. If freqbin is set to 0 (e.g. in the case of a large number of SNPs and few haplotypes), standardization is performed considering each observed frequency as a frequency class.

minmaf SNPs with a MAF (Minor Allele Frequency) lower than minmaf are discarded from the analysis.

Details

iHS (standardized iHH) are standardized as described in Voight et al. (2006)

Value

The returned value is a list containing two elements

res.ihs matrix with nsnps rows and four columns (Chromosome name, position of the SNP, Rsb and Pvalue)

summary.class matrix with nclasses rows and four columns (Class, Number of SNPs belonging to this class, position of the SNP, mean iHH in this class, standard deviation of iHH in this class)
References


See Also
calc_ehh, scan_hh, distribplot, ihsplot

Examples

data(wgscan.cgu)
## results from a genome scan (44,057 SNPs)
## see ?wgscan.eut and ?wgscan.cgu for details
res.ihs<-ihh2ihs(wgscan.cgu)

ihsplot(data, plot.pval="TRUE", ylim.scan=2, pch=16, main="iHS")

Description

Plot iHS over a genome.

Usage

ihsplot(data, plot.pval="TRUE", ylim.scan=2, pch=16, main="iHS")

Arguments

data A matrix with nsnps rows and 4 columns (Chromosome name, position of the SNP, iHS, Pvalue) (e.g. obtained using ihh2ihs function.
plot.pval Either TRUE or FALSE if Pvalue should not be plotted
ylim.scan An horizontal line is added at the corresponding coordinate, for instance to represent a significance threshold
pch Type of the points representing SNPs in the plot(s)
main Main Legend of the plot

Value

The function returns a plot

References

Description

This function allows to copy in the working directory three different example files: bta12_cgu.hap (an haplotype input file in standard format), bta12_hapguess_switch.out (an haplotype input file in fastphase output format) and map.inp (a SNP information input file). These files contains data for 280 haplotypes (originating from 140 individuals belonging to the Creole cattle breed from Guadeloupe) of 1,424 SNPs mapping to bovine chromosome 12 (BTA12) (see reference below).

Usage

make.example.files()

References


See Also

data2haplohh

Examples

make.example.files()
rsbplot

Plot Rsb over a genome

Description

Plot Rsb over a genome.

Usage

rsbplot(data, plot.pval="TRUE", ylim.scan, pch=16, main=NA)

Arguments

data A matrix with nsnps rows and 4 columns (Chromosome name, position of the
SNP, Rsb, Pvalue (either unilateral or bilateral)) (e.g. obtained using ies2rsb
function.
plot.pval Either TRUE or FALSE if Pvalue should not be plotted
ylim.scan An horizontal line is added at the corresponding coordinate, for instance to repre-
sent a significance threshold
pch Type of the points representing SNPs in the plot(s)
main Main Legend of the plot

Value

The function returns a plot

References


See Also
calc_ehhs, scan_hh, distribplot, ihsplot

Examples

data(wgscan.cgu); data(wgscan.eut)
## results from a genome scan (44,057 SNPs)
# see ?wgscan.eut and ?wgscan.cgu for details
res.rsb<ies2rsb(wgscan.cgu, wgscan.eut,"CGU","EUT")
rsbplot(res.rsb$res.rsb)
Computing EHH based statistics over a whole chromosome

Description

Compute Extended Haplotype Homozygosity (EHH), site-specific EHH (EHHS), integrated EHH (iHH) and integrated EHHS (iES) for all SNPs of the chromosome (or linkage group).

Usage

scan_hh(haplohh,limhaplo=2,limehh=0.05,limehhs=0.05)

Arguments

- **haplohh**: An object of class haplohh (see data2haplohh).
- **limhaplo**: Minimal number of haplotypes to continue computing EHH away from the core SNP. Useless, if no missing data. However, when some data are missing, haplotypes with missing data are removed from the computation. Hence as we compute EHH further from the core SNP, less haplotypes are expected.
- **limehh**: Limit below which EHH stops to be evaluated.
- **limehhs**: Limit below which EHHS stops to be evaluated.

Details

Extended Haplotype Homozygosity (EHH), site-specific EHH (EHHS), integrated EHH (iHH) and integrated EHHS (iES) are computed for all SNPs of the chromosome (or linkage group). This function is approximately twice as faster as a procedure calling in turn calc_ehh and calc_ehhs for all the SNP. To perform a whole genome-scan this function needs to be called for each chromosome and results concatenated.

Value

The returned value is a matrix with nsnps rows and six columns (Chromosome name, position of the SNP, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES).

References


**See Also**

calc_ehh, calc_ehhs, data2haplohh, ihh2lhs, ies2rsb

**Examples**

```r
# example haplohh object (280 haplotypes, 1424 SNPs)
# see ?haplohh_cgu_bta12 for details
data(haplohh_cgu_bta12)
res.scan<-scan_hh(haplohh_cgu_bta12)
```

---

**wgscan.cgu**

*Whole genome scan results for the CGU (Creole from Guadeloupe island)*

**Description**

A matrix of 44,057 rows (SNPs) and 6 columns (Chromosome name, position of the SNP in bp, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES).

**Usage**

data(wgscan.cgu)

**References**


---

**wgscan.eut**

*Whole genome scan results for a pooled of European taurine cattle*

**Description**

A matrix of 44,057 rows (SNPs) and 6 columns (Chromosome name, position of the SNP in bp, Frequency of the ancestral allele, iHH for the ancestral allele, iHH for the derived allele and iES).

**Usage**

data(wgscan.eut)
References

Index

*Topic **classes**
  *Topic **haplohh-class, 10**

*Topic **package**
  **rehh-package, 2**

bifurcation_diagram, 3
calc_ehh, 4
calc_ehhs, 6
data2haplohh, 7
distribplot, 9

haplohh-class, 10
haplohh_cgu_bta12, 10

ies2rsb, 11
ihh2ihs, 12
ihsplot, 13

make_example_files, 14

rehh (rehh-package), 2
rehh-package, 2
rsbplot, 15

scan_hh, 16

wgscan_cgu, 17
wgscan_eut, 17