Title  Haplotype Data Simulation

Version  0.2

Date  2005-09-22

Author  Giovanni Montana

Maintainer  Giovanni Montana <g.montana@imperial.ac.uk>

Description  Package for haplotype data simulation. Haplotypes are generated such that their allele frequencies and linkage disequilibrium coefficients match those estimated from an input data set.

Depends  MASS

License  GPL version 2 or later

R topics documented:

ACEdata  ..........................................................  2
allelefreqs  ..........................................................  2
checkpd  ..........................................................  3
cor2dprime  ..........................................................  3
cortocov  ..........................................................  4
covtocov  ..........................................................  4
divlocus  ..........................................................  5
haplodata  ..........................................................  6
haplofreqs  ..........................................................  7
haplosim  ..........................................................  7
ldplot  ..........................................................  9
makepd  .......................................................... 10
mergemats  .......................................................... 10
mse  .......................................................... 11
sumsqscale  .......................................................... 12

Index  13
ACEdata  

**Description**

ACE (angiotensin I converting enzyme) data set

**Usage**

data(ACEdata)

**Format**

A data set with 22 haplotypes and 52 SNPs.

**Source**


**References**


---

allelefreqs  

**Description**

Estimates allele frequencies from a binary matrix

**Usage**

allelefreqs(dat)

**Arguments**

- **dat**  
  A binary matrix, rows are haplotypes and columns are binary markers

**Value**

A list containing:

- **freqs**  
  Vector of allele "0" frequencies

- **all.polym**  
  If TRUE, all loci are polymorphic

- **non.polym**  
  Vector of non-polymorphic loci, if any

**Author(s)**

Giovanni Montana
**checkpd**

*Checks that a covariance matrix is positive definite*

**Description**

Checks that a matrix is positive definite

**Usage**

```r
ccheckpd(sigma, tol = 1e-06)
```

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sigma</td>
<td>Covariance matrix</td>
</tr>
<tr>
<td>tol</td>
<td>Tolerance value</td>
</tr>
</tbody>
</table>

**Value**

If TRUE, the input matrix is positive definite.

---

**cor2dprime**

*Correlations to D’ coefficients conversion*

**Description**

Converts a matrix of correlation coefficients into a D’ matrix

**Usage**

```r
cor2dprime(mat, probs)
```

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mat</td>
<td>A correlation matrix</td>
</tr>
<tr>
<td>probs</td>
<td>A vector of allele frequencies</td>
</tr>
</tbody>
</table>
The resulting matrix of $D'$ coefficients

Giovanni Montana


cortocov

Correlation to covariance conversion

Converts a correlation matrix into a covariance matrix

cortocov(cor.mat, sd)

Correlation matrix

Vector of standard deviations

The covariance matrix

covtocor

Covariance to correlation conversion

Converts a covariance matrix into a correlation matrix

covtocor(cov.mat)

Covariance matrix
divlocus

Value

The correlation matrix

Examples

divlocus

Description

Compute a measure of genetic diversity at each locus

Usage

divlocus(dat)

Arguments

dat A binary matrix, rows are haplotypes and columns are binary markers

Details

This function implements a measure of diversity for a locus \( j \) as in Clayton (2002). If \( z_{ij} \) represents the allele \( j \) of haplotype \( i \), for \( i = 1, \ldots, N \) and assuming that alleles are coded as 0 and 1, the diversity measure can be written as

\[
D_j = 2 * N \left( \sum_{i=1}^{N} z_{ij}^2 - \left( \sum_{i=1}^{N} z_{ij} \right)^2 \right)
\]

Value

A vector containing the diversity measure for all markers

Author(s)

Giovanni Montana

References


Examples

data(ACEdata)
divlocus(ACEdata)
haplodata  

Haplotype object creator

Description

Creates an haplotype data object needed for simulating haplotypes with haplosim. This object also contains some summary statistics about the real data.

Usage

haplodata(dat)

Arguments

dat  A binary matrix, rows are haplotypes and columns are binary markers

Value

A list containing:

freqs  Allele frequencies
cor  Correlation matrix (LD coefficients)
div  Locus-specific diversity measure
cov  Covariance matrix for the normal distribution

Author(s)

Giovanni Montana

References


See Also

See also haplosim

Examples

data(ACEdata)

# creates the haplotype object
x <- haplodata(ACEdata)

# simulates 100 random haplotypes
y <- haplosim(100, x)
haplofreqs

Haplotype frequencies

Description

Compute haplotype frequencies

Usage

haplofreqs(dat, firstl, lastl)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dat</td>
<td>A binary matrix, rows are haplotypes and columns are binary markers</td>
</tr>
<tr>
<td>firstl</td>
<td>Position of the first locus</td>
</tr>
<tr>
<td>lastl</td>
<td>Position of the last locus</td>
</tr>
</tbody>
</table>

Value

A vector of haplotype frequencies

Author(s)

Giovanni Montana

References


Examples

```r
data(ACEdata)
freqs <- haplofreqs(ACEdata, 17, 22)
```

haplosim

Haplotype data simulator

Description

Generates a random sample of haplotypes, given an haplotype object created from a data set

Usage

haplosim(n, hap, which.snp = NULL, seed = NULL, force.polym = TRUE, summary = TRUE)
Arguments

n  Number of haplotypes to generate
hap  Haplotype object created with haplodata
which.snp  A vector specifying which SNPs to include
seed  Seed for the random number generator
force.polym  if TRUE, all loci are polymorphic
summary  if TRUE, additional summary statistics are returned

Value

A list containing:

data  Simulated sample
freqs  Allele frequency vector
cor  Correlation matrix
div  Locus-specific diversity scores
mse.freqs  MSE of allele frequencies
mse.cor  MSE of correlations

Author(s)

Giovanni Montana

References


See Also

See also haplodata

Examples

#  # Example 1  
#
data(ACEdata)

# create the haplotype object
x <- haplodata(ACEdata)

# simulates a first sample of 100 haplotypes using all markers
y1 <- haplosim(100, x)

# compares allele frequencies in real and simulated samples
plot(x$freqs, y1$freqs, title=paste("MSE:",y1$mse.freqs)); abline(a=0, b=1)

# compares LD coefficients in real and simulated samples
ldplot(mergemats(x$cor, y1$cor), id.type='r')
```r
# simulates a second sample of 1000 haplotypes using the first 20 markers only
y2 <- haplosim(1000, which.snp=seq(20), x)
#
# Example 2
#
# simulate a sample of 500 haplotypes based on the ACE data set
set.seed(100)
data(ACEdata)
n <- 500
x <- haplodata(ACEdata)
y <- haplosim(n, x)

# compute the haplotype frequencies
# an haplotype starts at markers 17 and ends at marker 22
freq1 <- haplofreqs(ACEdata, 17, 22)
freq2 <- haplofreqs(y$data, 17, 22)

# extract the set of haplotypic configurations that are shared
# by real and simulated data and their frequencies
commonhapls <- intersect(names(freq1), names(freq2))
cfreq1 <- freq1[commonhapls]
cfreq2 <- freq2[commonhapls]

# compare real vs simulated haplotype frequencies
par(mar=c(10.1, 4.1, 4.1, 2.1), xpd=TRUE)
legend.text <- names(cfreq1)
bp <- barplot(cbind(cfreq1, cfreq2), main="Haplotype Frequencies",
             names.arg=c("Real","Simulated"), col=heat.colors(length(legend.text)))
legend(mean(range(bp)), -0.3, legend.text, xjust = 0.5,
       fill=heat.colors(length(legend.text)), horiz = TRUE)
chisq.test(x=n*cfreq2, p=cfreq1, simulate.p.value = TRUE, rescale.p = TRUE)
```

---

**ldplot**

**LD plot**

**Description**

Creates a linkage disequilibrium plot from a matrix of pair-wise LD coefficients

**Usage**

```r
ldplot(ld.mat, ld.type, color = heat.colors(50), title = NULL)
```

**Arguments**

- `ld.mat`: A square matrix of LD coefficients
- `ld.type`: A character value specifying what coefficients are used as input: either 'r' for correlation coefficients or 'd' for D/Dprime scores
- `color`: A range of colors to be used for drawing. Default is `heat.colors`
- `title`: Character string for the title of the plot
Author(s)

Giovanni Montana

References


Examples

data(ACEdata)

# LD plot of ACEdata using r^2 coefficients
ldplot(cor(ACEdata), ld.type='r')

makepd

Makes a matrix positive definite

Description

Modifies a matrix so that it is positive definite

Usage

makepd(mat, eig.tol = 1e-06)

Arguments

mat Matrix
eig.tol Tolerance value

Value

A positive-definite matrix

mergemats

Merges two LD matrices

Description

Merges two LD matrices. It can be used to compare the LD coefficients estimated in the real and simulated data sets

Usage

mergemats(mat1, mat2)
mse

Arguments
mat.1  First square matrix
mat.2  Second square matrix of same dimensions

Details

Value
The resulting matrix has upper triangular matrix from mat.1 and lower triangular matrix from mat.2

Author(s)
Giovanni Montana

References

Description
Computes the Mean Square Error between two vectors

Usage
mse(x, y)

Arguments
x  First vector
y  Second vector, must have same length

Value
The MSE score
sumsqscale  Scales a matrix

Description
Scales the columns of a matrix so that the squared elements sum to unity

Usage

sumsqscale(mat)

Arguments

mat  A matrix

Value
The scaled matrix.

Examples
Index

*Topic** datasets
   ACEdata, 1
*Topic** internal
   checkpd, 3
   cor2dprime, 3
   cortocov, 4
   covtocor, 4
   makepd, 10
   mse, 11
   sumsqscale, 12
*Topic** utilities
   allelefreqs, 2
   divlocus, 5
   haplodata, 6
   haplofreqs, 7
   haplosim, 7
   ldplot, 9
   mergemats, 10

ACEdata, 1
allelefreqs, 2
checkpd, 3
cor2dprime, 3
cortocov, 4
covtocor, 4
divlocus, 5
haplodata, 6, 8
haplofreqs, 7
haplosim, 6, 7
ldplot, 9
makepd, 10
mergemats, 10
mse, 11
sumsqscale, 12