Package 'pwLD'

June 2, 2017

Type Package

Title Estimation of pairwise haplotype frequencies and linkage disequilibrium measures

Version 2.0

Date 2008-07-23

Author Karsten Krug <karsten_krug@gmx.net>

Maintainer Karsten Krug <karsten_krug@gmx.net>

Depends gtools, corrplot, R (>= 3.0.0)

Description Estimation of pairwise haplotype frequencies and linkage disequilibrium from unphased genotyped SNP data.

License GPL (>= 2)

NeedsCompilation yes

RoxygenNote 6.0.1

Archs x64

R topics documented:

cardano.complex *Analytic solution of a cubic polynomial*

Description

Analytic solution of a cubic polynomial by Cardano's formula.

Usage

```
cardano.complex(coeff)
```

Arguments

coeff	numeric vector of length four indicating the coefficients of the polynomial in descending order. See _Details_ below.

Details

The function calculates the roots of the polynomial within the set of complex numbers and hence returns complex roots. The coefficients are determined by parameter `coeff` and correspond to the standard form of cubic polynomial

y = a x^3 + b x^2 + c x + d

in the following order: coeff = c(a,b,c,d).

Value

D	discriminant of Cardano's formula
x	complex vector containing the roots

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>

Examples

```
cardano.complex(c(4,6,3,1))
```

HapMap *Sample HapMap data set (r21)*

Description

This data set consists of the first 500 SNPs from chromosome 22 genotyped in the HapMap JPT+CHB population (N=90).

Usage

```
HapMap_geno
```

Format

HapMap_geno is a matrix with 500 rows and 90 columns. The genotypes are coded as follows: homozygote: 0 or 2, heterozygote: 1, missing data: 3. HapMap_anno is a matrix with 500 rows and 10 columns containing some annotations of the SNPs.

Source

www.hapmap.org

LD.all *Estimation of pairwise linkage disequilibrium measures*

Description

Estimation of pairwise haplotype frequencies and linkage disequilibrium measures from genotyping SNP data. LD.allestimates LD measures of all pairwise SNP combinations contained in data.

Usage

```
LD.all(data, code = c(0, 1, 2, 3),
LD = c("D", "Dprime", "Q", "r", "OR", "MI", "chi2", "Y", "HS"), MAF = 0,
paradigm = c("freq", "bayes", "fullbayes"),
strategy = c("bootstrap", "jackknife"), dirich = rep(1, 4),
verbose = T, tol = .Machine$double.eps^0.6, digits = 12,
CI = F, HSweight = 4, alpha = 0.2, nSim = 1000, seed = F, intervall = c(0, 1), mc=1000)
```

Arguments

data	P x N matrix representing P loci with N observed genotypes. The entries are items of the parameter code.
code	vector of length four that indicates the code used in snps to mark allele combinations. The ordering is NOT arbitrary and is stated as follows: allele 1 homozygote, heterozygote, allele 2 homozygote, missing data. For instance, if there are two alleles, A and B, and missing values are denoted as NN, the parameter has to be code=c("AA", "AB", "BB", "NN").
LD	character vector indicating the LD measures to estimate. See _Details_ below.
MAF	numeric specifying the minor allele frequency. All loci with allele frequency $<=$ MAF are excluded from data. If MAF = 0 and paradigm = "bayes" no loci are excluded.
paradigm	character indicating the statistical paradigm to use in order to estimate haplotype frequencies. If freq, haplotype frequencies are estimated by dividing absolute counts by N. If bayes, a Dirichlet prior is assumed and the mean of posterior distribution serves as an estimate.If fullbayes, a Dirichlet prior is assumed and the mean of posterior distribution serves as an estimate using a monte carlo simulation.
strategy	character, if bootstrap a bootstrap confidence interval is constructed, if jackknife a jackknife confidence interval is constructed
dirich	numeric vector of length 9 indicating the shape parameters of the Dirichlet prior.

verbose	logical, if TRUE some informations about the progress of function LD.all will be displayed.
tol	numeric, accuracy value for numerical comparison of floating point numbers. See _Details_ below.
digits	numeric, specifies the number of digits for rounding of the cubic polynomial's coefficients. See _Details_ below.
CI	logical, if TRUE either a frequentistic confidence interval or a Bayesian credible interval is estimated, depending on parameter paradigm.
HSweight	numeric, parameter n of the HS measure for up-and downweighting
alpha	numerical, confidence level of confidence or credible interval, respectively.
nSim	numerical, indicating the number of bootstrap or Dirichlet replicates to use in order to determine confidence or credible intervals, respectively. Default value is set to 1000.
seed	logical, if TRUE random seed of the CI construction was set
intervall	integer, if strategy=="bootstrap"&&intervall=="0" bootstrap quantile method is used, if strategy=="bootstrap"&&intervall=="1" bootstrap standard error method is used, if strategy=="jackknife"&&intervall=="1" jackknife leave one out method is used, if strategy=="jackknife"&&intervall=="0" jackknife pseudo value mehtod is used.
mc	integer, number of Monte Carlo iterations for full bayesian estimator.

Value

The output of LD.all is a list of correlation matrizes. If the paramter: CIis FALSE, a list of one entry containing the correlation matrix of all defined LD measures between all genotyped SNPs. CIis TRUE, a list of three entries containing the correlation matrizies of all defined LD measures between all genotyped SNPs, the lower bounds of confidence interval for all defined LD measures between all genotyped SNPs, the upper bounds of confidence interval for all defined LD measures between all genotyped SNPs.

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>, Jakub Fusiak <<fusiakjakub@gmx.net>>

References

Further informations and documentation can be found on github

See Also

LD.cardano

Examples

```
## load sample HapMap data
data(HapMap)

## estimate LD measures Dprime, Q, r, OR, Y and HS between all SNP pairs:
#res.all <- LD.all(data = HapMap_geno, code = c(0, 1, 2, 3),
#LD = c("Dprime", "Q", "r", "OR", "Y", "HS"), MAF = 0.1,
#paradigm = c("freq"), strategy = c("jackknife"), verbose = TRUE,
#tol = .Machine$double.eps^0.6, digits = 12, CI = TRUE, HSweight = 4,
#alpha = 0.05, intervall = 0)
```

LD.cardano *Estimation of pairwise linkage disequilibrium measures*

Description

Estimation of pairwise haplotype frequencies and linkage disequilibrium measures from genotyping SNP data. LD.cardano determines the fix-point of the EM algorithm analytically in order to infer double heterozygote haplotype frequencies.

Usage

```
LD.cardano(snps, code = c(0, 1, 2, 3), LD = c("Dprime", "Q", "r", "Y", "HS"),
HSweight = 4, CI = T, strategy = c("jackknife", "bootstrap"),
alpha = 0.05, n.sim = 1000, returnLDdist = F, paradigm = c("freq", "bayes", "fullbayes"),
dirich = rep(1, 9), all.solutions = F, tol = .Machine$double.eps^0.6,
digits = 12, seed = F, mc = 1000, intervalll = c(0, 1))
```

Arguments

snps	2 x N matrix representing two loci with N observed genotypes. The entries are items of the parameter code.
code	vector of length four that indicates the code used in snps to mark allele combinations. The ordering is NOT arbitrary and is stated as follows: allele 1 homozygote, heterozygote, allele 2 homozygote, missing data. For instance, if there are two alleles, A and B, and missing values are denoted as NN, the parameter has to be code=c("AA", "AB", "BB", "NN").
LD	character vector indicating the LD measures to estimate. See _Details_ below.
HSweight	numeric, parameter n of the HS measure for up-and downweighting
CI	logical, if TRUE either a frequentistic confidence intervall or a Bayesian credible intervall is estimated, depending on parameter paradigm.
strategy	haracter, if bootstrap a bootstrap confidence intervall is constructed, if jackknife a jackknife confidence intervall is constructed
alpha	numerical, confidence level of confidence or credible intervall, respectively.
n.sim	numerical indicating the number of bootstrap or Dirichlet replicates to use in order to determine confidence or credible intervalls, respectively. Default value is set to 5000.
returnLDdist	logical, if TRUE the distribution of each LD measure over n.sim simulations is returned.
paradigm	character indicating the statistical paradigm to use in order to estimate genotypic frequencies. If freq, genotype frequencies are estimated by dividing absolute counts by N. If bayes, a Dirichlet prior is assumed and the mean of posterior distribution serves as an estimate.
dirich	numeric vector of length 9 indicating the shape parameters of the Dirichlet prior.
all.solutions	logical, if TRUE all solutions of the cubic polynomial will be returned.
tol	numeric, accuracy value for numerical comparison of floating point numbers. See _Details_ below

digits	numeric, specifies the number of digits for rounding of the cubic polynomial's coefficients. See _Details_ below.
seed	logical, if TRUE random seed of the CI construction was set
mc	numberic, numbers of iterations of the monte carlo simulation for the full bayesian haplotype frequency estimator
intervalll	integer, if strategy=="bootstrap"&&intervall=="0" bootstrap quantile method is used, if strategy=="bootstrap"&&intervall=="1" bootstrap standard error method is used, if strategy=="jackknife"&&intervall=="1" jackknife leave one out method is used, if strategy=="jackknife"&&intervall=="0" jackknife pseudo value mehtod is used

Value

Pairwise haplotype frequencies are represented by a 2×2 contingency table with following notation:

alleles	0	1	Sum
0	p_{00}	p_{01}	$p_{0.}$
1	p_{10}	p_{11}	$p_{1.}$
Sum	$p_{.0}$	$p_{.1}$	1

The rows correspond to the first, the columns to the second SNP. The marginal distributions of this table are the allele frequencies that are directly estimated from genotypic frequencies. Haplotype frequencies are estimated via maximum-likelihood by determining the fix-point of the EM algorithm analytically, i.e. the resulting cubic polynomial is solved by the use of Cardano's formulae and the solution that is maximum-likelihood serves as an estimate. The measures of linkage disequilibrium specified by parameter LD are:

The parameters tol and digits are used to avoid numerical problems during the computation. Parameter tol specifies an epsilon for a numerically robust comparison of floating point numbers, parameter digits guarantees a robust case discrimination within the Cardano formulae. The chosen default values should be unaffected by the user.

The output of LD.cardano is a list of haplotype frequencies, chi-. If the paramter: If CI ist FALSE a list containing haplotype frequencies, chi square value, number of genotype counts, number of double heterozygotes, names of the SNPs LD is measured between and the defined LD measrues If CI a list containing haplotype frequencies, chi square value, number of genotype counts, number of double heterozygotes, names of the SNPs LD is measured between and the defined LD measrues, CIs for the LD measures with the wanted CI method If all.solutions is TRUE the solutions of maximizing the likelihood of the data is printed out too. If LDdist a list of tables of used by the Bootstrap/Jackknife confidence intervall estimation and the LD measrues for these tables are printed out too.

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>, Jakub Fusiak <<fusiakjakub@gmx.net>>

References

Further informations and documentation can be found on github

See Also

LD.all

Examples

```
## load sample HapMap data
data(HapMap)

##Calculation of pairwise LD measures with default values, no conifidence intervall
res.ci <- LD.cardano(snps=HapMap_geno[c(4,5), ], LD=c("Dprime"),
paradigm="bayes", CI=FALSE,strategy = "bootstrap",
intervalll = 1,alpha = 0.05,HSweight = 4,returnLDdist = FALSE,
mc = 1000,seed = FALSE,all.solutions = FALSE ,n.sim = 10)

##Calculation of pairwise LD measures D', Q, r, Y and HS with the bayesian plug in estimator
res.ci <- LD.cardano(snps=HapMap_geno[c(4,5), ], LD=c("Dprime"),
paradigm="bayes", CI=TRUE,strategy = "bootstrap",
intervalll = 1,alpha = 0.05,HSweight = 4,returnLDdist = FALSE,
mc = 1000,seed = FALSE,all.solutions = FALSE ,n.sim = 10)
```

LD.corrplot *Estimation of pairwise linkage disequilibrium measures*

Description

Graphical output of a correlation matrix consists of the LD measures between all genotyped SNPs. This data is the result of the LD.all command.

Usage

```
LD.corrplot(valuematrix, values = c("values", "upperCI", "lowerCI"),
LD = c("D", "Dprime", "Q", "r", "OR", "MI", "chi2", "Y", "HS"),
snp.zoom = FALSE, snp.target, snp.region, corr.labels=FALSE)
```

Arguments

valuematrix	object, output of the LD.all command - symmatric matrix of the LD measure between all genotyped SNPs.
values	object, parameter to define wich values shoul be plotted. If vales is "values" the LD values are taken for the plotting. If values is "upperCI" the upper bounds of the confidence interval estimation are taken for the plotting. If values is "lowerCI" the lower bounds of the confidence interval estimation are taken for the plotting.
LD	character, defines which LD measure will be plotted
snp.zoom	logical, if FALSE all data will be plotted if TRUE only part of the data will be plotted
snp.target	character, definition of the SNP, that is wanted to be investigated, e.g. "rs16984366"
snp.region	integer, definition of the region of the SNP, that is wanted to be investigated, e.g. 5 SNPs before and after the "rs16984366" SNP
corr.labels	logical, show labels? TRUE Yes. FALSE No.

Value

Output is a correlation matrix plot.

Note

This command uses the package `corrplot` to give out the graphical output.

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>, Jakub Fusiak <<fusiakjakub@gmx.net>>

References

Further informations and documentation can be found on github

See Also

`LD.snpplot`

Examples

```
## load sample HapMap data
data(HapMap)

##Creating correlation matrix
res.all <- LD.all(data = HapMap_geno, code = c(0, 1, 2, 3),
LD = c("Dprime", "Q", "r", "OR", "Y", "HS"), MAF = 0.1,
paradigm = c("freq"), strategy = c("jackknife"), verbose = TRUE,
tol = .Machine$double.eps^0.6, digits = 12, CI = TRUE, HSweight = 4,
alpha = 0.05, intervall = 0)

##Plotting correlation matrix of D' without zoom
LD.corrplot(res.all, values = c("values"), LD = c("Dprime"), snp.zoom = FALSE)

##Plotting correlation matrix of D' with zoom on rs16984366 around a region of 5 SNPs
LD.corrplot(res.all, values = c("values"), LD = c("Dprime"),
snp.zoom = TRUE, snp.target="rs16984366", snp.region=5)
```

`LD.pattern` *Visualisation of LD patterns along chromosomes or chromosomal regions.*

Description

The function visualises patterns of LD along chromosomes or chromosomal regions.

Usage

```
LD.pattern(LDs, map = NULL, window = 1700000, minSNPs = 10, scale = c("Mb", "Kb", "bp"),
plot = T)
```

Arguments

LDs	object returned by function LD.all.
map	numeric vector containing the genomic positions of SNPs.
window	numeric indicating the size of the sliding window in bp. See _Details_ below.
minSNPs	minimum number of SNPs that have to be within a window.
scale	character indicating the scale used for the plot. One of the following values is possible: Mb, Kb, bp.
plot	logical, if TRUE the plot will be produced.

Details

The absolute LD values of SNPs within a sliding window of size window are averaged and plotted against the middle position of the corresponding interval. Parameter minSNPs indicates the minimal number of SNPs that have to be within the window in order to average the LD values.

Value

LD.mean	vector containing the averaged LD values of each sliding window, i.e. the y-axis of the plot.
BP.mean	vector containing the middle position of each sliding window, i.e. the x-axis of the plot.
#SNPs	vector containing the number of SNPs within each sliding window.
info	some general informations.

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>

References

Further informations and documentation can be found on github

See Also

LD.snpplot

Examples

```
## load sample HapMap data
data(HapMap)

## estimate pairwise LD between all SNPs with MAF > 0
res.LD <- LD.all(HapMap_geno, LD=c("Dprime","Q"), MAF=0, CI=FALSE)


## make the plot
res.pattern <- LD.pattern(res.LD[1], window=3e5,
map=HapMap_anno[rownames(res.LD[[1]]), "pos"], scale="Kb" )
```

LD.snpplot *Estimation of pairwise linkage disequilibrium measures*

Description

Graphical output of a scatterplot consists of the LD measures between all genotyped SNPs. This data is the result of the LD.all command.

Usage

```
LD.snpplot(valuematrix, annotation, values = c("values", "upperCI", "lowerCI"),
LD = c("D", "Dprime", "Q", "r", "OR", "MI", "chi2", "Y", "HS"),
snp.zoom = FALSE, snp.target, snp.region, unit = c("Mb", "Kb", "bp"), legends=TRUE)
```

Arguments

valuematrix	object, output of the LD.all command - symmatric matrix of the LD measure between all genotyped SNPs.
annotation	object, numeric vector containing the genomic positions of SNPs.
values	character, Parameter to define wich values shoul be plotted. If values is "values" the LD values are taken for the plotting. If values is "upperCI" the upper bounds of the confidence interval estimation are taken for the plotting. If values is "lowerCI" the lower bounds of the confidence interval estimation are taken for the plotting.
LD	character, Defines which LD measure will be plotted. One of the following values is possible: Dprime, r, Q, Y, HS
snp.zoom	logical, if FALSE all data will be plotted if TRUE only part of the data will be plotted
snp.target	character, definition of the SNP, that is wanted to be investigated, e.g. "rs16984366"
snp.region	integer, definition of the region of the SNP, that is wanted to be investigated, e.g. 5 SNPs before and after the "rs16984366" SNP
unit	character, indicating the scale used for the plot. One of the following values is possible: Mb, Kb, bp.
legends	logical, show legend? TRUE Yes. FALSE No.

Value

Output is a scatterplot of LD values of a region of SNPs against the position of the Chromosome in base pairs

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>, Jakub Fusiak <<fusiakjakub@gmx.net>>

References

Further informations and documentation can be found on github

See Also

LD.corrplot

Examples

```
## load sample HapMap data
data(HapMap)

## Creating correlation matrix
res.all <- LD.all(data = HapMap_geno, code = c(0, 1, 2, 3),
LD = c("Dprime", "Q", "r", "OR", "Y", "HS"), MAF = 0.1, paradigm = c("freq"),
strategy = c("jackknife"), verbose = TRUE, tol = .Machine$double.eps^0.6,
digits = 12, CI = TRUE, HSweight = 4, alpha = 0.05, intervall = 0)

## Plotting scatterplot for D' without zoom
LD.snpplot(valuematrix =  res.all, annotation = HapMap_anno,
values = "values",LD ="Dprime",snp.zoom = FALSE, unit = "Mb",
snp.target = "rs16984366", snp.region = 5)

## Plotting scatterplot for D' with zoom on rs16984366 around a region of 5 SNPs
LD.snpplot(valuematrix =  res.all, annotation = HapMap_anno,
values = "values",LD ="Dprime",snp.zoom = TRUE, snp.target = "rs16984366",
snp.region = 5,unit = "Mb")
```

---

plotLikelihood *Plot of the log-likelihood function of pairwise haplotype frequencies*

Description

The function visualises the log-likelihood function of haplotype frequencies of a given SNP pair.

Usage

```
plotLikelihood(snps, code = c(0, 1, 2, 3),
paradigm = c("freq", "bayes"), dirich = rep(1,9), main = "",
ylim = NULL, legend = F, tol = .Machine$double.eps^0.6,
plot.tol=.Machine$double.eps^.25, digits=12)
```

Arguments

snps	2 x N matrix representing two loci with N observed genotypes. The entries are items of parameter code.
code	vector of length four that indicates the code used in snps to mark allele combinations. The ordering is NOT arbitrary and is stated as follows: allele 1 homozygote, heterozygote, allele 2 homozygote, missing data. For instance, if there are two alleles, A and B, and missing values are denoted as NN, the parameter has to be code=c("AA", "AB", "BB", "NN").
paradigm	character indicating the statistical paradigm to use in order to estimate genotypic frequencies. If freq, genotype frequencies are estimated by dividing absolute counts by N. If bayes, a Dirichlet prior is assumed and the mean of posterior distribution serves as estimate.

dirich	numeric vector of length 9 indicating the shape parameters of Dirichlet prior.
main	character, main title for the plot.
ylim	numeric vector of length two indicating the y limits of the plot.
legend	logical, if TRUE a legend will be plotted.
plot.tol	numeric, tolerance at the margins in order to avoid NaN values. (can be ignored by the user)
tol	numeric, accuracy value for numerical comparison of floating point numbers. See the help pages of LD.cardano for further details.
digits	numeric, specifies the number of digits for rounding of the cubic polynomial's coefficients. See the help pages of LD.cardano for further details.

Author(s)

Karsten Krug <<karsten_krug@gmx.net>>

References

Further informations and documentation can be found on github

See Also

LD.ml, LD.cardano

Examples

```
## load sample HapMap data
data(HapMap)

## plot the log likelihood and add a legend
##plotLikelihood(HapMap_geno[c(45,46),], legend=TRUE)
```

Index