Package 'IFP'

November 26, 2020

Version 0.2.4

Date 2020-10-22

Title Identifying Functional Polymorphisms

Author Park L

Maintainer Leeyoung Park <lypark@yonsei.ac.kr>

Description
A suite for identifying causal models using relative concordances and identifying causal polymorphisms in case-control genetic association data, especially with large controls re-sequenced data.

License GPL (>= 2)

Depends R (>= 2.11.1)

Imports haplo.stats,coda

Suggests

URL https://www.r-project.org

NeedsCompilation yes

Repository CRAN

Date/Publication 2020-11-26 00:50:05 UTC

R topics documented:

`allele.freq` *Allele Frequency Computation from Genotype Data*

Description

Computes allele frequencies from genotype data.

Usage

```
allele.freq(geno)
```

Arguments

`geno`	matrix of alleles, such that each locus has a pair of adjacent columns of alleles, and the order of columns corresponds to the order of loci on a chromosome. If there are K loci, then ncol(geno) = 2*K. Rows represent the alleles for each subject. Each allele shoud be represented as numbers (A=1,C=2,G=3,T=4).

Value

array of allele frequencies of each SNP. The computed allele is targeted as an order of alleles, "A", "C", "G", and "T".

Examples

```
data(apoe)
allele.freq(apoe7)
allele.freq(apoe)
```

allele.freq.G — *Allele Frequency Computation from the sequencing data with a vcf type of the 1000 Genomes Project*

Description

Computes allele frequencies from the sequencing data with a vcf type of the 1000 Genomes Project.

Usage

```
allele.freq.G(genoG)
```

Arguments

genoG	matrix of haplotypes. Each row indicates a variant, and each column ind icates a haplotype of an individual. Two alleles of 0 and 1 are available.

Value

array of allele frequencies of each variant.

Examples

```
data(apoeG)
allele.freq.G(apoeG)
```

apoe — *Genetic data of APOE gene region*

Description

This data set came from a re-sequenced data of APOE gene region in the Molecular Diversity and Epidemiology of Common Disease (MDECODE) database. Sixteen polymorphic sites were included. "apoe7" data contains the genetic data of seven single nucleotide polymorphisms with allele frequencies higher than 0.1 from the apoe data.

Usage

```
data(apoe)
```

Format

A matrix with 48 rows and 32 columns

Source

http://droog.gs.washington.edu/mdecode/

References

Nickerson, D. A., S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark et al. (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res* 10: 1532-1545.

`apoeG` *Sequencing data of APOE gene region from the 1000 Genomes Project*

Description

This data set came from a re-sequenced data of APOE gene region from the 1000 Genomes Project. Thirty three polymorphic sites with allele frequencies higher than 0.001 were included for the original data set, apoeG. The test data sets, apoeT and apoeC, indicate the data of 100 controls and 100 cases respectively when the dominant variant is 15th variant with the odds ratio of 3.

Usage

```
data(apoeG)
```

Format

A matrix with 33 rows and 2184 columns

Source

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/

References

Abecasis, G. R. et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073.

`drgegggne` *causal models with all possible causal factors: G, G*G, G*E and E*

Description

provides concordance probabilities of relative pairs for a causal model with G, G*G, G*E and E components

Usage

```
drgegggne(fdg,frg,fdgg,frgg,fdge,frge,eg,e)
```

Arguments

fdg	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G component
frg	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G component
fdgg	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G*G component
frgg	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G*G component
fdge	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G*E component
frge	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G*E component
eg	a proportion of population who are exposed to environmental cause of G*E interactiong the genetic cause of G*E during their entire life
e	a proportion of population who are exposed to environmental cause during their entire life

Value

matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drggn drgegne

Examples

```
### PLI=0.01.
ppt<-0.01



### for a model without one or more missing causal factors,
### set the relevant parameters as zero.

pg<-0.002  # the proportion of G component in total populations
pgg<-0.002  # the proportion of G*G component in total populations
pge<-0.003  # the proportion of G*E component in total populations
e<-1-(1-ppt)/(1-pg)/(1-pgg)/(1-pge)
   # the proportion of E component in total populations

fd<-0.001  # one dominant gene
tt<-3      # the number of recessive genes

temp<-sqrt(1-((1-pg)/(1-fd)^2)^(1/tt))
```

```
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

ppd<-sqrt(pgg)
fdg<-array(1-sqrt(1-ppd^(1/2)),2)
ttg<-1
temp<-(pgg/ppd)^(1/2/ttg)
frg<-c(array(0,length(fdg)),array(temp,ttg))
fdg<-c(fdg,array(0,ttg))

ppe<-0.5
ppg<-pge/ppe

fdge<-0.002
ttge<-2      # the number of recessive genes

temp<-sqrt(1-((1-ppg)/(1-fdge)^2)^(1/ttge))
frge<-c(array(0,length(fdge)),array(temp,ttge))
fdge<-c(fdge,array(0,ttge))


drgegggne(fd,fr,fdg,frg,fdge,frge,ppe,e)
```

---

drgegne *causal models with three possible causal factors: G, G*E and E*

Description

provides concordance probabilities of relative pairs for a causal model with G, G*E and E components

Usage

```
drgegne(fdg,frg,fdge,frge,eg,e)
```

Arguments

fdg	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G component
frg	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G component
fdge	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G*E component
frge	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G*E component

eg	a proportion of population who are exposed to environmental cause of G*E interactiong the genetic cause of G*E during their entire life
e	a proportion of population who are exposed to environmental cause during their entire life

Value

matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drgn drgene

Examples

```
### PLI=0.01.
ppt<-0.01



pg<-0.002  # the proportion of G component in total populations
pge<-0.005  # the proportion of G*E component in total populations
e<-1-(1-ppt)/(1-pg)/(1-pge)
  # the proportion of E component in total populations

fd<-0.001  # one dominant gene
tt<-2      # the number of recessive genes

temp<-sqrt(1-((1-pg)/(1-fd)^2)^(1/tt))
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

ppe<-0.5
ppg<-pge/ppe

fdge<-0.002
ttge<-2      # the number of recessive genes

temp<-sqrt(1-((1-ppg)/(1-fdge)^2)^(1/ttge))
frge<-c(array(0,length(fdge)),array(temp,ttge))
fdge<-c(fdge,array(0,ttge))


drgegne(fd,fr,fdge,frge,ppe,e)
```

drgen *causal models with G*E*

Description

provides concordance probabilities of relative pairs for a causal model with G*E component

Usage

```
drgen(fd,fr,e)
```

Arguments

fd	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G component of G*E interacting with E of G*E
fr	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G component of G*E interacting with E of G*E
e	a proportion of population who are exposed to environmental cause of G*E interacting with genetic cause of G*E during their entire life

Value

a list of the g*e proportion in population and a matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drgene.gm

Examples

```
### PLI=0.01.
ppt<-0.01



### g*e model

pge<-ppt  # the proportion of G*E component in total populations

ppe<-0.5
ppg<-pge/ppe

fd<-0.0005  # one dominant gene
```

```
tt<-3      # the number of recessive genes

temp<-sqrt(1-((1-ppg)/(1-fd)^2)^(1/tt))
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

drgen(fd,fr,ppe)
```

---

drgene *causal models with G*E and E*

Description

provides concordance probabilities of relative pairs for a causal model with G*E and E components

Usage

```
drgene(fdg,frg,eg,e)
```

Arguments

fdg	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G component of G*E interacting with E of G*E
frg	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G component of G*E interacting with E of G*E
eg	a proportion of population who are exposed to environmental cause of G*E interacting with genetic cause of G*E during their entire life
e	a proportion of population who are exposed to environmental cause during their entire life

Value

matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th(causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drgen.gm

Examples

```
### PLI=0.01.
ppt<-0.01



### g*e+e model

pge<-0.007  # the proportion of G*E component in total populations
e<-1-(1-ppt)/(1-pge)   # the proportion of E component in total populations

ppe<-0.5
ppg<-pge/ppe

fd<-0.0005  # one dominant gene
tt<-3      # the number of recessive genes

temp<-sqrt(1-((1-ppg)/(1-fd)^2)^(1/tt))
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

drgene(fd,fr,ppe,e)
```

---

drggn *causal models with G*G*

Description

provides concordance probabilities of relative pairs for a causal model with G*G component

Usage

```
drggn(fd,fr)
```

Arguments

fd	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G*G component
fr	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G*G component

Value

a list of PLI and a matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drgegggne

Examples

```
### PLI=0.01.
ppt<-0.01



### g*g model

pp<-ppt  # the proportion of G*G component in total populations

gd<-sqrt(pp) # dominant gene proportion = recessive gene proportion
fd<-array(1-sqrt(1-gd^(1/2)),2)  # two dominant genes
tt<-2      # the number of recessive genes: 2

temp<-(pp/gd)^(1/2/tt)
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

drggn(fd,fr)
```

---

drgn *causal models with G*

Description

provides concordance probabilities of relative pairs for a causal model with G component

Usage

```
drgn(fd,fr)
```

Arguments

fd	an array (size=number of dominant genes+recessive genes) of dominant gene frequencies including 0 values of recessive genes of G component
fr	an array (size=number of dominant genes+recessive genes) of recessive gene frequencies including 0 values of dominant genes of G component

Value

list of the value of PLI and the matrix of NN, ND, and DD probabilities of 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

See Also

drgegne.gm

Examples

```
### PLI=0.01.
ppt<-0.01



### g model

pp<-ppt  # the proportion of G component in total populations

fdt<-0.001 # one dominant gene with frequency of 0.001
tt<-5      # the number of recessive genes: 5

fd<-c(fdt,array(0,tt))
temp<-sqrt(1-((1-pp)/(1-fdt)^2)^(1/tt))
fr<-c(0,array(temp,tt))

drgn(fd,fr)
```

`error.rates` *Error Rates Estimation for Likelihood Ratio Tests Designed for Identifying Number of Functional Polymorphisms*

Description

Compute error rates for a given model.

Usage

```
error.rates(H0,Z, pMc, geno, no.ca, no.con=nrow(geno), sim.no = 1000)
```

Arguments

`H0`	the index number for a given model for functional SNPs
`Z`	number of functional SNPs for the given model
`pMc`	array of allele frequencies of case samples
`geno`	matrix of alleles, such that each locus has a pair of adjacent columns of alleles, and the order of columns corresponds to the order of loci on a chromosome. If there are K loci, then ncol(geno) = 2*K. Rows represent the alleles for each subject. Each allele shoud be represented as numbers (A=1,C=2,G=3,T=4).
`no.ca`	number of case chromosomes
`no.con`	number of control chromosomes
`sim.no`	number of simulations for error rates estimation

Value

array of results consisted of Type I error rate (alpha=0.05), Type I error rate (alpha=0.01), Type II error rate (beta=0.05), Type II error rate (beta=0.01), percent when the target model has the lowest corrected -2 log likelihood ratio.

See Also

allele.freq hap.freq lrtB

Examples

```
## LRT tests when SNP1 & SNP6 are the functional polymorphisms.


data(apoe)

n<-c(2000, 2000, 2000, 2000, 2000, 2000, 2000) #case sample size = 1000
x<-c(1707, 281,1341, 435, 772, 416, 1797) #allele numbers in case samples

Z<-2  #number of functional SNPs for tests
n.poly<-ncol(apoe7)/2  #total number of SNPs

#index number for the model in this case is 5 for SNP1 and 6.
#apoe7 is considered to represent the true control allele and haplotype frequencies.
#Control sample size = 1000.

error.rates(5, 2, x/n, apoe7, 2000, 2000, sim.no=2)

# to obtain valid rates, use sim.no=1000.
```

---

geno.freq *Genotype Frequency Computation from the sequencing data with a vcf type of the 1000 Genomes Project*

Description

Computes genotype frequencies from the sequencing data with a vcf type of the 1000 Genomes Project.

Usage

```
geno.freq(genoG)
```

Arguments

genoG	matrix of haplotypes. Each row indicates a variant, and each column ind icates a haplotype of an individual. Two alleles of 0 and 1 are available.

Value

matrix of genotype frequencies of each variant.

Examples

```
data(apoeG)
geno.freq(apoeG)
```

`genotype` *Conversion to Genotypes from Alleles using the sequencing data with a vcf type of the 1000 Genomes Project*

Description

Convert sequencing data to genotypes.

Usage

```
genotype(genoG)
```

Arguments

`genoG`	matrix of haplotypes. Each row indicates a variant, and each column ind icates a haplotype of an individual. Two alleles of 0 and 1 are available.

Value

matrix of genotypes with rows of variants and with columns of individuals.

Examples

```
data(apoeG)
genotype(apoeG)
```

`hap.freq` *Estimation of Haplotype Frequencies with Two SNPs*

Description

EM computation of haplotype frequencies with two SNPs. The computation is relied on the package"haplo.stats".

Usage

```
hap.freq(geno)
```

Arguments

geno	matrix of alleles, such that each locus has a pair of adjacent columns of alleles, and the order of columns corresponds to the order of loci on a chromosome. If there are K loci, then ncol(geno) = 2*K. Rows represent the alleles for each subject. Each allele shoud be represented as numbers (A=1,C=2,G=3,T=4).

Value

matrix of haplotype frequencies consisted of two alleles from each SNP. These alleles are the same ones computed for frequency using the function "allele.freq".

See Also

allele.freq

Examples

```
data(apoe)
hap.freq(apoe7)
hap.freq(apoe)
```

---

iter.mcmc *mcmc inference of causal models with all possible causal factors: G, G*G, G*E and E*

Description

provides proportions of each causal factor of G, G*G, G*E and E based on relative concordance data

Usage

```
iter.mcmc(ppt,aj=2,n.iter,n.chains,thinning=5,init.cut,darray,x,n,model,mcmcrg=0.01)
```

Arguments

ppt	population lifetime incidence
aj	a constant for the stage of data collection
n.iter	number of mcmc iterations
n.chains	number of mcmc chain
thinning	mcmc thinning parameter (default=5)
init.cut	mcmc data cut
darray	indicating the array positions of available data among 9 relative pairs: 1:mzt,2:parent-offspring,3:dzt,4:sibling,5:2-direct(grandparent-grandchild),6:3rd(uncle-niece),7:3-direct(great-grandparent-great-grandchild),8:4th (causin),9:4d(great-great-grandparent-great-great-grandchild)

x	number of disease concordance of relative pairs
n	total number of relative pairs
model	an array, size of 4 (1: E component; 2: G component; 3: G*E component; 4: G*G component), indicating the existance of the causal component: 0: excluded; 1: included.
mcmcrg	parameter of the data collection stage (default=0.01)

Value

a list of rejectionRate, result summary, Gelman-Rubin diagnostics (point est. & upper C.I.) for output variables: e[1]: proportion of environmental factor (E) g[2]: proportion of genetic factor (G) ge[3]: proportion of gene-environment interaction (G*E) gg[4]: proportion of gene interactions (G*G) gn[5]: number of recessive genes in G ppe[6]: population proportion of interacting environment in G*E ppg[7]: population proportion of interacting genetic factor in G*E fd[8]: frequency of dominant genes in G fdge[9]: frequency of dominant genes in G*E gnge[10]: number of recessive genes in G*E ppd[11]: population proportion of dominant genes in G*G ppr[12]: population proportion of recessive genes in G*G kd[13]: number of dominant genes in G*G kr[14]: number of recessive genes in G*G

References

L. Park, J. Kim, A novel approach for identifying causal models of complex disease from family data, Genetics, 2015 Apr; 199, 1007-1016.

Examples

```
### PLI=0.01.
ppt<-0.01

### a simple causal model with G and E components

pg<-0.007  # the proportion of G component in total populations
pgg<-0  # the proportion of G*G component in total populations
pge<-0  # the proportion of G*E component in total populations
e<-1-(1-ppt)/(1-pg)   # the proportion of E component in total populations

fd<-0.001  # one dominant gene
tt<-3      # the number of recessive genes

temp<-sqrt(1-((1-pg)/(1-fd)^2)^(1/tt))
fr<-c(array(0,length(fd)),array(temp,tt))
fd<-c(fd,array(0,tt))

rp<-drgegggne(fd,fr,c(0,0),c(0,0),c(0,0),c(0,0),0,e)

sdata<-rp[,3]/(rp[,2]+rp[,3])
#sdata<-round(sdata*500)

darray<-c(1:2,4:6)
  ## available data= MZT, P-O, sibs, grandparent-grandchild, avuncular pair
```

```
n<-array(1000,length(darray))
x<-array()
for(i in 1:length(darray)){
x[i]<-rbinom(1,n[i],sdata[darray[i]])
}
model<-c(1,1,0,0)

## remove # from the following lines to test examples.
#iter.mcmc(ppt,2,15,2,1,1,darray,x,n,model) # provide a running test
#iter.mcmc(ppt,2,2000,2,10,500,darray,x,n,model) # provide a proper result
```

`lrt` *Likelihood Ratio Tests for Identifying Number of Functional Polymorphisms*

Description

Compute p-values and likelihoods of all possible models for a given number of functional SNP(s).

Usage

```
lrt(n.fp, n, x, geno, no.con=nrow(geno))
```

Arguments

`n.fp`	number of functional SNPs for tests.
`n`	array of each total number of case sample chromosomes for SNPs
`x`	array of each total allele number in case samples
`geno`	matrix of alleles, such that each locus has a pair of adjacent columns of alleles, and the order of columns corresponds to the order of loci on a chromosome. If there are K loci, then ncol(geno) = 2*K. Rows represent the alleles for each subject. Each allele shoud be represented as numbers (A=1,C=2,G=3,T=4).
`no.con`	number of control chromosomes.

Value

matrix of likelihood ratio test results. First n.fp rows indicate the model for each set of disease polymorphisms, and followed by p-values, -2 log(likelihood ratio) with corrections for variances, maximum likelihood ratio estimates, and likelihood.

References

L. Park, Identifying disease polymorphisms from case-control genetic association data, Genetica, 2010 138 (11-12), 1147-1159.

See Also

allele.freq hap.freq

Examples

```
## LRT tests when SNP1 & SNP6 are the functional polymorphisms.

data(apoe)

n<-c(2000, 2000, 2000, 2000, 2000, 2000, 2000) #case sample size = 1000
x<-c(1707, 281,1341, 435, 772, 416, 1797) #allele numbers in case samples


Z<-2  #number of functional SNPs for tests
n.poly<-ncol(apoe7)/2  #total number of SNPs

#control sample generation( sample size = 1000 )
con.samp<-sample(nrow(apoe7),1000,replace=TRUE)
con.data<-array()
for (i in con.samp){
con.data<-rbind(con.data,apoe7[i,])
}
con.data<-con.data[2:1001,]

lrt(1,n,x,con.data)
lrt(2,n,x,con.data)
```

`lrtG` *Likelihood Ratio Tests for Identifying Disease Polymorphisms with Same Effects*

Description

Compute p-values and likelihoods of all possible models for a given number of disease SNP(s).

Usage

```
lrtG(n.fp, genoT, genoC)
```

Arguments

`n.fp`	number of disease SNPs for tests.
`genoT`	matrix of control genotypes. Each row indicates a variant, and each column indicates a haplotype of an individual. Two alleles of 0 and 1 are allowed.
`genoC`	matrix of case genotypes. Each row indicates a variant, and each column indicates a haplotype of an individual. Two alleles of 0 and 1 are allowed.

Value

matrix of likelihood ratio test results. First row indicates the index, and following n.fp rows indicate the model for each set of disease polymorphisms, and followed by p-values, -2 log(likelihood ratio) with corrections for variances, and the degree of freedom.

References

L. Park, J. Kim, Rare high-impact disease variants: properties and identification, Genetics Research, 2016 Mar; 98, e6.

See Also

allele.freq.G

Examples

```
## LRT tests for a dominant variant (15th variant)
## the odds ratio: 3, control: 100, case: 100.

data(apoeG)
lrtG(1,genoT[,1:20],genoC[,1:20])

# use "lrtG(1,genoT,genoC)" for the actual test.
```

Index