Package 'TriadSim'

October 12, 2022

Title Simulating Triad Genomewide Genotypes

Version 0.3.0

Description

Simulate genotypes for case-parent triads, case-control, and quantitative trait samples with realistic linkage diequilibrium structure and allele frequency distribution. For studies of epistasis one can simulate models that involve specific SNPs at specific sets of loci, which we will refer to as ``pathways". TriadSim generates genotype data by resampling triad genotypes from existing data. The details of the method is described in the manuscript under preparation ``Simulating Autosomal Genotypes with Realistic Linkage Disequilibrium and a Spiked in Genetic Effect" Shi, M., Umbach, D.M., Wise A.S., Weinberg, C.R.

Depends R (>= 3.2.2)

License GPL-3

biocViews snpStats

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Imports methods, parallel, snpStats, foreach, doParallel

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Min Shi [aut, cre]

Maintainer Min Shi <shi2@niehs.nih.gov>

Repository CRAN

Date/Publication 2021-09-08 17:20:05 UTC

R topics documented:

`fit.risk.model.par` *Resample families based on the risk model*

Description

This function selects families based on the prespecified risk model. It can simulate a homogenous scenario or a stratified scenario with two subpopulations. When e.fr is given rather than the default NA the risk model can involve exposure main effects as well as gene by exposure interation. This function is parallelized and the default number of cores for parallelization is set as the ceiling of half of the total number of CPU cores.

Usage

```
fit.risk.model.par(
  n.ped,
  brks,
  target.snp,
  fam.pos,
  mom.tar,
  dad.tar,
  kid.tar,
  pathways,
  betas.e0,
  e.fr = NA,
  betas.e,
  pop1.frac = NA,
  rate.beta = NA,
  is.case = TRUE,
  qtl = FALSE,
  out.put.file = NA,
  no_cores = NA
)
```

Arguments

`n.ped`	is an integer giving the number of trios to be simulated
`brks`	a matrix of integers showing where the chromosomal breaks is to take place for each individual in the simulated trios.
`target.snp`	is a vector of integers showing the row number of the target SNPs in the .bim file.

`fam.pos` is a matrix showing the chromosomal segments out of which is each target SNP selected for each simulated trio.

`mom.tar` is a matrix containing genotypes of the target SNPs in the mothers of the original data for simulations of a homogenous population. For simulations under population stratification it is a list of two matrices each containing genotypes of the mothers' target SNP genotypes in one of the two subpopulations.

`dad.tar` is a matrix containing the genotypes of the target SNPs in the fathers of the original data for simulations of a homogenous population. For simulations under population stratification it is a list of two matrices each containing fathers' target SNP genotypes in one of the two subpopulations.

`kid.tar` is a matrix with containing genotypes of the target SNP in the children stacking on top of the complements of the original data for simulations of a homogenous population. For simulations under population stratification it is a list of two matrices each containing children's and complements' target SNP genotypes in one of the two subpopulations.

`pathways` is a list of vectors of integers. Each vector of integers denotes the SNPs involved in a particular pathway. E.g. list(1:4,5:8) denote that there are two pathways. SNPs 1-4 are in the first pathway and SNPs 5-8 are in the second.

`betas.e0` is a vector of doubles giving the beta coefficients of the logit risk model for the unexposed individuals. The length of the vector should be 1+ number_of_risk_pathway. The first number is a function of the disease prevalence in the unexposed individual who does not carry any copies of the risk pathway. The numbers after that gives the odds ratios for carrying one/two copies of the risk pathways comparing to those who do not carry any copies of the pathways in the unxposed group. e.g., c(-6.4, 0.5,1) means the baseline disease prevalence is exp(-6.4)/(1+exp(-6.4)) and the log OR for carrying at least one copy of the first pathway is 0.5 and that for carrying at least one copy of the second pathway is 1.

`e.fr` is a double number between 0 and 1 which gives the exposure prevalence.

`betas.e` is a vector of doubles giving the beta coefficients of the logit risk model for the exposed individuals. The length of the vector should be 1+ number_of_risk_pathway. The first number is a function the disease prevalence in the exposed individual who does not carry any copies of the risk pathway. The numbers after that gives the odds ratios for carrying one/two copies of the risk pathways comparing to those who do not carry any copies of the pathways in the exposed group.

`pop1.frac` is a double number between 0 and 1 which gives the fraction of subpopulation 1 out of the two subpopulations for a population stratification scenario.

`rate.beta` is a double number giving the log OR of disease prevalence in population 2 over that in population 1.

`is.case` is a boolean variable. When is.case = TRUE case-parents trios will be simulated.Otherwise, control-parents trios will be simulated.

`qtl` is a boolean variable denoting whether a quantitative trait (qtl=TRUE) or a binary trait (qtl=FALSE) is to be simulated. For a binary trait only affected families will be kept. The default value is qtl=FALSE.

`out.put.file` is a character string giving the base file name for the output file. When a non-default value is given the fucntion will write the following files to the designated

directory: a file with name ending with "exp.txt" containing the exposure data when exposure is involved in the risk model. a file with name ending with "pop.txt" containing information on subpopulation membership when the simulation involves a stratified scenario. a file with name ending with "pheno.tx" containing quantitative trait phenotype when a quantitative trait is involved. When out.put.file is the default value NA the file names for the above three files are: exposure.txt, population.txt, phenotype.txt.

`no_cores` is an integer which specifies the number of CPU cores to be parallelized.

Value

The function returns a list of five elements. The first one is a matrix of integers giving the families (in terms of row number) selected for each simulated trio and each chromosomal segment. The second one is a matrix giving the genotypes on the target SNPs in the simulated trio. The third one is relevant only when exposure is involved. It is a vector of 0's and 1's giving the exposure status of each simulated trio when the risk model involves exposure. The fourth element is relevant only in simulations of stratified scenarios. It is a vector of 1's and 2's giving the memebership of the subpopulation groups of each simulated trio. The fifth element is relevant only in simulations of a quantitative trait. It is a vector of doubles giving the phenotype values for simulations of a quantitative trait.

Examples

```
tar.snp <- c(21, 118, 121, 140, 155, 168, 218, 383)
found.brks <- get.brks(N.brk=3,n.ped=1000, snp.all2, tar.snp,rcmb.rate=NA)
breaks <- found.brks[[1]]
family.position <- found.brks[[2]]
betas <- c(-6.4, 3.2, 5.8)
pwy <- list(1:4,5:8)
m.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_mom')
f.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_dad')
k.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_kid')
# the preloaded data frame snp.all2 contains the data frame read from the corresponding .bim file.
target.geno <- get.target.geno(c(m.file,f.file,k.file), tar.snp,snp.all2)
mom.target <- target.geno[[1]]
dad.target <- target.geno[[2]]
kid.target <- target.geno[[3]]
## Not run:
fitted.model <- fit.risk.model.par(n.ped=1000,brks=breaks,target.snp=tar.snp,
fam.pos=family.position,mom.tar=mom.target,dad.tar=dad.target, kid.tar=kid.target,
pathways=pwy,betas, e.fr=NA, betas,pop1.frac= NA,rate.beta=NA,no_cores=2)

## End(Not run)
```

`get.brks` *Picking chromosomal breaking points*

Description

The breaking points at each chromosome can be picked manually or use this function. When a data frame containing the recombination rates (rcmb.rate) is provided the function tends to pick the breaking points at recombination hotspots.

Usage

```
get.brks(N.brk, n.ped, snp.all2, target.snp, rcmb.rate = NA, same.brk = FALSE)
```

Arguments

N.brk	is an integer giving the number of breaks to be picked for each chromosome.
n.ped	is an integer giving the number of trios to be simulated
snp.all2	is a dataframe containing the list of SNPs in PLINK .bim format. Two columns of the dataframe is used: column 1 with column name "V1" containing the chromosome number and column 4 with column name "V4" containing the chromosomal postion of the SNPs.
target.snp	is a vector of integers showing the row number of the target SNPs in the .bim file.
rcmb.rate	the default value is NA. rcmb.rate is a dataframe containing the recombination rates at each SNP. The ordering of the SNPs should be identical to that of snp.all2. It contains 4 columns with column names 'CHR','RS','POS',and 'RATE with the corresponding values for "the chromosomal number", "SNP rs number", "chromosomal position", and "recombination rate". The recombination rate represents the maximum recombination rate in the chromosomal region between the current SNP and the SNP above (or the first basepair of the chromosome for the first SNP on a chromosome). When no rcmb.rate is provided the function will pick the breaking points randomly where keeping the breaking points in between target SNPs. An example recombination rate data frame "rcmb.rate" is already loaded with the package.
same.brk	is an indicator variable to denote whethere the same set of breaking points will be used for all simulated triads

Value

A list of two elements is returned. The first one is a matrix of integers showing where the chromosomal breaks is to take place for each individuals in the simulated trios. The second one is a matrix showing the chromosomal segments out of which each target SNP is selected for each simulated trio.

Examples

```
tar.snp <- c(21, 118, 121, 140, 155, 168, 218, 383)
found.brks <- get.brks(N.brk=3,n.ped=1000, snp.all2, tar.snp,rcmb.rate=NA)
breaks <- found.brks[[1]]
family.pos <- found.brks[[2]]
```

`get.target.geno` *Getting genotypes of the target SNPs*

Description

This function read out the genotypes of the selected target SNPs from the orignal data set (the data set on which simulation is based).

Usage

```
get.target.geno(input.plink.file, target.snp, snp.all2)
```

Arguments

`input.plink.file`	is a vector of three character strings for the file names of the mother's father's and child's plink base filenames with the necessary path to the directory. The plink files are in bed format and three files with extensions .bed .bim and .fam are expected for each individual's genotypes. The mothers, fathers, and chil-dredn must be from the same set of trio families even though the ordering of the families can be different for the three sets of data.
`target.snp`	is a vector of integers showing the row number of the target SNPs in the .bim file
`snp.all2`	is a dataframe containing list of SNPs in PLINK .bim format. Only the second column is used which contains the rs number of the SNPs. The colname name of the second column has to be "V2".

Value

A list of three matrices is returned. The three matrices are the observed genotypes of the mothers from family 1 to family n repeated twice, genotypes of the fathers from family 1 to family n repeated twice and genotypes of children from family 1 to n followed by (stacking on top of) genotypes of the complements at the target SNPs.

Examples

```
tar.snp <- c(21, 118, 121, 140, 155, 168, 218, 383)
m.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_mom')
f.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_dad')
k.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_kid')
# the preloaded data frame snp.all2 contains the data frame read from the corresponding .bim file.
## Not run:
target.geno <- get.target.geno(c(m.file,f.file,k.file), tar.snp,snp.all2)

## End(Not run)
```

`glue.chr.segment.par` *Splicing chromosomal segments*

Description

This function splices the triad chromosomal segments into "complete" trios. The spliced trio sets are written into separate plink files chromosome by chromosome. It is parallelized and if no no_cores value is given the ceiling of half of the total number of CPUs available will be used in the parallelization.

Usage

```
glue.chr.segment.par(
  input.plink.file,
  out.put.file,
  brks,
  sel.fam.all,
  snp.all2,
  pathway.all,
  target.snp,
  pop.vec = NA,
  no_cores = NA,
  flip = TRUE
)
```

Arguments

`input.plink.file`	for simulations of homogenous population, it is a vector of three character strings for the base filenames of the mother's father's and child's plink base filenames. The plink files are in bed format and in the same folder three files with extensions .bed .bim and .fam are expected for each individual's genotypes. The mothers, fathers, and childredn must be from the same set of trio families even though the ordering of the families can be different for the three sets of data. For simulations under population stratification it is a list of two vectors. Each vector is a vector of three character strings for the base filenames as described above.The two vectors correspond to the two subpopulations.
`out.put.file`	is a character string giving the base file name for the output file. Genotypes on different chromosomes are output to different files. The final file name also contains information on chromosome number. E.g., for a base filename "trio" and for chromosome 1 the final file name is "trio1sim".
`brks`	is a matrix of integers showing where the chromosomal breaks is to take place for each individual in the simulated trios.
`sel.fam.all`	is a matrix of integer giving the families (in terms of row number) selected for each chromosomal segment and each simulated trio.

snp.all2	is a dataframe containing the list of SNPs in PLINK .bim format. Two columns of the dataframe is used: column 1 with column name "V1" containing the chromosome number and column 2 with column name "V2" containing the rs number of the SNPs.
pathway.all	is a matrix giving the genotypes on the pathway SNPs in the simulated trio.
target.snp	is a vector of integers showing the row number of the target SNPs in the .bim file.
pop.vec	is a vector of 1's and 2's giving the subpopulation group of each simulated trio. This parameter is relevant only for stratified scenarios.
no_cores	is an integer which specifies the number of CPU cores to be parallelized.
flip	is a boolean indicating whether the mother's and the father's genotypes will be swapped to wipe out potential maternal effects in the orignal data.

Value

This function does not return values. Instead it writes PLINK files into the designated directory. Each set of PLINK files contains genotype data for one chromosome for all trios. The first one third of the rows are genotypes of the mothers'. The second one third are those of the fathers' and the last one third are the children's.

Examples

```
tar.snp <- c(21, 118, 121, 140, 155, 168, 218, 383)
found.brks <- get.brks(N.brk=3,n.ped=1000, snp.all2, tar.snp,rcmb.rate=NA)
breaks <- found.brks[[1]]
family.position <- found.brks[[2]]
betas <- c(-6.4, 3.2, 5.8)
pwy <- list(1:4,5:8)
m.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_mom')
f.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_dad')
k.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_kid')
# the preloaded data frame snp.all2 contains the data frame read from the corresponding .bim file.
target.geno <- get.target.geno(c(m.file,f.file,k.file), tar.snp,snp.all2)
mom.target <- target.geno[[1]]
dad.target <- target.geno[[2]]
kid.target <- target.geno[[3]]
 ## Not run:
fitted.model <- fit.risk.model.par(n.ped=1000,brks=breaks,target.snp=tar.snp,
fam.pos=family.position,mom.tar=mom.target,dad.tar=dad.target, kid.tar=kid.target,
pathways=pwy,betas, e.fr=NA, betas,pop1.frac= NA,rate.beta=NA,no_cores=2)
sel.fam <- fitted.model[[1]]
sim.pathway.geno <-  fitted.model[[2]]
glue.chr.segment.par(c(m.file,f.file,k.file),file.path(tempdir(),'trio'), breaks,sel.fam,
                  snp.all2,sim.pathway.geno,target.snp,pop.vec=NA,no_cores=1,flip=TRUE)

## End(Not run)
```

pick_target.snp *Pick target SNPs in the pathway*

Description

The target SNPs in the pathway can be picked by users manually or use this facility function. It helps pick the set of target SNPs in the pathway(s) based on a desired allele frequency. If picked manually, the target SNPs should be in the order from the smallest to the largest.

Usage

```
pick_target.snp(input.plink.file, fr.desire = "double", n.snp = "integer")
```

Arguments

input.plink.file	is a vector of two character strings for the file names of the mother's and father's plink base filenames with the necessary path to the directory. The plink files are in bed format and three files with extensions .bed .bim and .fam are expected for each parent's genotypes. In addition the allele frequnecy files generated by PLINK (base filename with .frq extension) are expected to be in the same directory as the .bed file.
fr.desire	is a double number giving the desired frequency of the target SNPs.
n.snp	is an integer giving the number of target SNPs to be picked.

Value

The function returns a list of two: first element is the SNPs read from the .bim file now with allele frequncies merged and the second is the row numbers of the target SNPs selected among all SNPs in the .bim file.

Examples

```
m.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_mom')
f.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_dad')
picked.target <- pick_target.snp(c(m.file,f.file),0.05, 8)
cat('Target SNPs picked:',picked.target[[2]],'\n')
```

rcmb.rate *An example recombination rate dataset*

Description

This dataset is an example dataset of recombination rates between SNPs The variables are as follows:

Usage

```
rcmb.rate
```

Format

A data frame with 478 rows and 6 variables

CHR Chromosome

RS SNP rs number

POS Chromosomal position

RATE recombination rate between the SNP and the SNP above

snp.all2 *SNPs in the PLINK files*

Description

A dataset containing the list of variables in the PLINK files The variables are as follows:

Usage

```
snp.all2
```

Format

A data frame with 10279 rows and 12 variables:

ord Ordering of the SNPs

RS SNP rs number

CHR Chromosome

POS Chromosomal position

A1 A1 allele

A2 A2 allele

MAF Minor allele frequency

NCHROBS Number of observed chromosomes used in MAF calculation

`TriadSim` *Simulation main function*

Description

TriadSim can simulate genotypes for case-parent triads, case-control, and quantitative trait samples with realistic linkage diequilibrium structure and allele frequency distribution. For studies of epistasis one can simulate models that involve specific SNPs at specific sets of loci, which we will refer to as "pathways". TriadSim generates genotype data by resampling triad genotypes from existing data. It takes genotypes in PLINK format as the input files.

Usage

```
TriadSim(
  input.plink.file,
  out.put.file,
  fr.desire,
  pathways,
  n.ped,
  N.brk,
  target.snp = NA,
  P0,
  is.OR,
  risk.exposure,
  risk.pathway.unexposed,
  risk.pathway.exposed,
  is.case = TRUE,
  e.fr = NA,
  pop1.frac = NA,
  P0.ratio = 1,
  rcmb.rate = NA,
  no_cores = NA,
  qtl = FALSE,
  same.brk = FALSE,
  flip = TRUE
)
```

Arguments

`input.plink.file`

gives the filenames (as well as the path) of the source data used for resampling. The input files are in PLINK format. For simulations of a homogenous population, it is a vector of three character strings for the base filenames of the mother's father's and child's PLINK files. The PLINK files are in bed format and three files with extensions .bed .bim and .fam are expected for each individual's genotypes. The mothers, fathers, and children must be from the same set of triad families even though the ordering of the families can be different for the three

sets of data. For simulations under population stratification it is a list of two vectors. Each vector is a vector of three character strings giving the base filenames for the PLINK files as described above.The two vectors correspond to the two subpopulations.

`out.put.file` is a character string giving the pathway to and the base filename of the output file. The names of the final output files also contain information on chromosome number. E.g., for a base filename "trio" and for chromosome 1 the final filenames for the PLILK files are "trio1.bim","trio1.bed" and "trio1.fam".

`fr.desire` is a double number giving the desired frequency of the target SNPs.

`pathways` is a list of vectors of integers. Each vector of integers denotes the SNPs involved in a particular pathway. E.g. list(1:4,5:8)

`n.ped` is an integer giving the number of trios to be simulated

`N.brk` is an integer giving the number of breaks to be picked for each chromosome.

`target.snp` is a vector of integers showing the row number of the target SNPs in the .bim file.

`P0` gives the baseline disease prevalence in the unexposed individuals with 0 copies of the risk pathways.

`is.OR` is a boolean varialbe denoting wether the input risk parameters are odds ratios. It is TRUE when the input risks are odds ratios.

`risk.exposure` is a double giving the relative risk (or odds ratio, if is.OR=TRUE) of the exposure main effect.

`risk.pathway.unexposed`
is a vector of doubles giving the relative risk (or odds ratio, if is.OR=TRUE) of each risk pathways in the unexposed individuals with the risk of unexposed individuals who carry no copies of the pathways as a reference.For scenarios that do not involve exposure the value of this vector is for all individuals.

`risk.pathway.exposed`
is a vector of doubles giving the relative risk (or odds ratio, if is.OR=TRUE) of each risk pathways in the exposed individuals. with the risk of exposed individuals who carry no copies of the pathways as a reference. For scenarios that do not involve exposure the value of this vector is not used.

`is.case` is a boolean variable. When is.case = TRUE case-parents trios will be simulated.Otherwise, control-parents trios will be simulated.

`e.fr` is a double number between 0 and 1 which gives the exposure prevalence.

`pop1.frac` is a double number between 0 and 1 which gives the fraction of population 1 for a population stratification scenario.

`P0.ratio` gives the ratio of the baseline disease prevalence in the second subpopulation to that of the first subpopulation.

`rcmb.rate` the default value is NA. rcmb.rate is a dataframe containing the recombination rates at each SNP. The ordering of the SNPs (in rows) should be identical to that of snp.all2. It has 4 columns with the column names 'CHR','RS','POS', and 'RATE' representing "the chromosomal number", "SNP rs number", "chromosomal position", and "recombination rate", respectively. The recombination rate represents the maximum recombination rate in the chromosomal region between

	the current SNP and the SNP above (or the first basepair of the chromosome for the first SNP on a chromosome).When no rcmb.rate is provided the function will pick the breaking points randomly.
no_cores	is an integer which specifies the number of CPU cores to parallelized.contain values
qtl	is a boolean variable denoting whether a quantitative trait (qtl=TRUE) or a binary trait (qtl=FALSE) is to be simulated. For a binary trait only affected families will be kept. The default value is qtl=FALSE.
same.brk	is an indicator variable to denote whether the same set of breaking points will be used for all simulated triads. The default value is FALSE.
flip	is an indicator variable denoting whether the mother's and the father's genotypes will be swapped to wipe out potential maternal effects in the orignal data. The default value is TRUE.

Value

this function simulates genotypes of parent-offspring triads and writes PLINK files into the designated directory. Genotypes on each chromosome will be written into a separate set of PLINK files. In each set of PLINK files genotypes of the mothers, fathers, and children are stacked on top of each other. The first third of the rows are genotypes of the mothers'.The seond third are those of the fathers' and the last third are those of the children's. The following files are also generated under specific scenarios: a file with name ending with "exp.txt" containing the exposure data when exposure is involved in the risk model. a file with name ending with "pop.txt" containing information on subpopulation membership when the simulation involves a stratified scenario. a file with name ending with "pheno.tx" containing quantitative trait phenotype when a quantitative trait is involved.

Examples

```
m.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_mom')
f.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_dad')
k.file <- file.path(system.file(package = "TriadSim"),'extdata/pop1_4chr_kid')
input.plink.file <- c(m.file, f.file, k.file)
## Not run: TriadSim(input.plink.file, file.path(tempdir(),'triad'), fr.desire=0.05,pathways=list(1:4,5:8),
       n.ped=1000, N.brk=3, target.snp=NA,P0=0.001,is.OR=FALSE,risk.exposure= 1,
     risk.pathway.unexposed=c(1.5, 2), risk.pathway.exposed=c(1.5, 2), is.case=TRUE, e.fr=NA,
       pop1.frac=NA, P0.ratio=1, rcmb.rate, no_cores=1)
## End(Not run)
```

Index