Package 'GenoScan'

December 21, 2018

Type Package

Title A Genome-Wide Scan Statistic Framework for Whole-Genome Sequence Data Analysis

Version 0.1

Date 2018-11-13

Author Zihuai He

Maintainer Zihuai He <zihuai@stanford.edu>

Description Functions for whole-genome sequencing studies, including genome-wide scan, candidate region scan and single window test.

License GPL-3

Depends R(>= 2.10.0), SKAT, Matrix, MASS, seqminer, data.table

NeedsCompilation no

Repository CRAN

Date/Publication 2018-12-21 15:20:16 UTC

R topics documented:

GenoScan.example *Data example for GenoScan (A Genome-Wide Scan Statistic Framework For Whole-Genome Sequence Data Analysis)*

Description

The dataset contains outcome variable Y, covariate X, genotype data G, positions of genetic variants pos, weight matrix for functional annotations Z.

Usage

```
data(GenoScan.example)
```

GenoScan.info *hg19 chromosome sizes*

Description

The dataset contains hg19 chromosome sizes from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes.

Usage

```
data(GenoScan.info)
```

GenoScan.prelim *The preliminary data management for GenoScan*

Description

This function does the preliminary data management and fit the model under null hypothesis. The output will be used in the other GenoScan functions.

Usage

```
GenoScan.prelim(Y, X=NULL, id=NULL, out_type="C", B=5000)
```

Arguments

Y	The outcome variable, an n*1 matrix where n is the total number of observations
X	An n*d covariates matrix where d is the total number of covariates.
id	The subject id. This is used to match phenotype with genotype. The default is NULL, where the matched phenotype and genotype matrices are assumed.
out_type	Type of outcome variable. Can be either "C" for continuous or "D" for dichotomous. The default is "C".
B	Number of resampling replicates. The default is 5000. A larger value leads to more accurate and stable p-value calculation, but requires more computing time.

Value

It returns a list used for function GenoScan.Region(), GenoScan.SingleWindow() and GenoScan.VCF.chr().

Examples

```
library(GenoScan)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by d matrix
# G: genotype matrix, n by p matrix where n is the total number of subjects
# Z: functional annotation matrix, p by q matrix

data(GenoScan.example)
Y<-GenoScan.example$Y;X<-GenoScan.example$X;G<-GenoScan.example$G;Z<-GenoScan.example$Z

# Preliminary data management
result.prelim<-GenoScan.prelim(Y,X=X,out_type="C",B=5000)
```

GenoScan.Region *Scan the association between an quantitative/dichotomous outcome variable and a region by score type statistics allowing for multiple functional annotation scores.*

Description

Once the preliminary work is done by "GenoScan.prelim()", this function scan a target region. This function is often used for candidate region analyses.

Usage

```
GenoScan.Region(result.prelim,G,pos,Gsub.id=NULL,Z=NULL,MAF.weights='beta',
test='combined',window.size=c(5000,10000,15000,20000,25000,50000),MAF.threshold=1,
impute.method='fixed')
```

Arguments

result.prelim	The output of function "GenoScan.prelim()"
G	Genetic variants in the target region, an n*p matrix where n is the subject ID and p is the total number of genetic variants.
pos	The positions of genetic variants, an p dimensional vector. Each position corresponds to a column in the genotype matrix.
Gsub.id	The subject id corresponding to the genotype matrix, an n dimensional vector. Each ID corresponds to a row in the genotype matrix. This is used to match phenotype with genotype. The default is NULL, where the matched phenotype and genotype matrices are assumed.
Z	Weight matrix for functional annotations, an p*q matrix where p is the total number of genetic variables and q is the number of weights. This is used to incorperate functional annotations. The default is NULL, where minor allele frequency weighted (see MAF.weights) dispersion and/or burden tests are applied.
MAF.weights	Minor allele frequency based weight. Can be 'beta' to up-weight rare variants or 'equal' for a flat weight. The default is 'beta'.
test	Can be 'dispersion', 'burden' or 'combined'. The test is 'combined', both dispersion and burden tests are applied. The default is 'combined'.
window.size	Candidate window sizes in base pairs. The default is c(5000,10000,15000,20000,25000,50000). Note that extemely small window size (e.g. 1) requires large sample size.
MAF.threshold	Threshold for minor allele frequency. Variants above MAF.threshold are ignored. The default is 1.
impute.method	Choose the imputation method when there is missing genotype. Can be "random", "fixed" or "bestguess". Given the estimated allele frequency, "random" simulates the genotype from binomial distribution; "fixed" uses the genotype expectation; "bestguess" uses the genotype with highest probability.

Value

n.marker	Number of tested variants in the window (heterozygous variants below MAF threshold).
window.summary	Results for all windows. Each row presents a window, including chromosome number, start position, end position,dispersion p-value(s), burden p-values(s).
M	Estimated number of effective tests.
threshold	Estimated threshold, 0.05/M. This threshold is for windows tested in this particular region.
p.value	P-value of entire region.

Examples

```
## GenoScan.prelim does the preliminary data management.
# Input: Y, X (covariates)
## GenoScan.Region scans a region.
# Input: G (genetic variants), pos (position) Z (weights) and result of GenoScan.prelim
```

```
library(GenoScan)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by d matrix
# G: genotype matrix, n by p matrix where n is the total number of subjects
# pos: positions of genetic variants, p dimention vector
# Z: functional annotation matrix, p by q matrix

data(GenoScan.example)
Y<-GenoScan.example$Y;X<-GenoScan.example$X
G<-GenoScan.example$G;pos<-GenoScan.example$pos
Z<-GenoScan.example$Z

# Preliminary data management
result.prelim<-GenoScan.prelim(Y,X=X,out_type="C",B=5000)

# Scan the region with functional annotations defined in Z
result<-GenoScan.Region(result.prelim,G,pos,Z=Z)
```

`GenoScan.SingleWindow` *Test the association between an quantitative/dichotomous outcome variable and a single window by dispersion or burden test allowing for multiple functional annotation scores.*

Description

Once the preliminary work is done by "GenoScan.prelim()", this function tests a single window. This is often used to double-check significant windows identified by GenoScan.Region or GenoScan.VCF.chr, with an increased number of resampling replicates in GenoScan.prelim.

Usage

```
GenoScan.SingleWindow(result.prelim,G,Gsub.id=NULL,Z=NULL,MAF.weights='beta',
test='combined',MAF.threshold=1,impute.method='fixed')
```

Arguments

`result.prelim`	The output of function "GenoScan.prelim()"
`G`	Genetic variants in the target region, an n*p matrix where n is the subject ID and p is the total number of genetic variants.
`Gsub.id`	The subject id corresponding to the genotype matrix, an n dimensional vector. Each ID corresponds to a row in the genotype matrix. This is used to match phenotype with genotype. The default is NULL, where the matched phenotype and genotype matrices are assumed.

Z	Weight matrix for functional annotations, an p*q matrix where p is the total number of genetic variables and q is the number of weights. This is used to incorperate functional annotations. The default is NULL, where minor allele frequency weighted (see MAF.weights) dispersion and/or burden tests are applied.
MAF.weights	Minor allele frequency based weight. Can be 'beta' to up-weight rare variants or 'equal' for a flat weight. The default is 'beta'.
test	Can be 'dispersion', 'burden' or 'combined'. The test is 'combined', both dispersion and burden tests are applied. The default is 'combined'.
MAF.threshold	Threshold for minor allele frequency. Variants above MAF.threshold are ignored. The default is 1.
impute.method	Choose the imputation method when there is missing genotype. Can be "random", "fixed" or "bestguess". Given the estimated allele frequency, "random" simulates the genotype from binomial distribution; "fixed" uses the genotype expectation; "bestguess" uses the genotype with highest probability.

Value

n.marker	Number of tested variants in the window (heterozygous variants below MAF threshold).
p.value	P-value(s) of the window (dispersion p-value(s), then burden p-values(s))

Examples

```
## GenoScan.prelim does the preliminary data management.
# Input: Y, X (covariates)
## GenoScan.Region scans a region.
# Input: G (genetic variants), pos (position) Z (weights) and result of GenoScan.prelim

library(GenoScan)

# Load data example
# Y: outcomes, n by 1 matrix where n is the total number of observations
# X: covariates, n by d matrix
# G: genotype matrix, n by p matrix where n is the total number of subjects
# pos: positions of genetic variants, p dimention vector
# Z: functional annotation matrix, p by q matrix

data(GenoScan.example)
Y<-GenoScan.example$Y;X<-GenoScan.example$X
G<-GenoScan.example$G;pos<-GenoScan.example$pos
Z<-GenoScan.example$Z

# Preliminary data management
result.prelim<-GenoScan.prelim(Y,X=X,out_type="C",B=5000)

# Scan the region with functional annotations defined in Z
result<-GenoScan.SingleWindow(result.prelim,G,Z=Z)
```

GenoScan.VCF.chr *Scan a VCF file to study the association between an quantitative/dichotomous outcome variable and a region or whole chromosome by score type statistics allowing for multiple functional annotation scores.*

Description

Once the preliminary work is done by "GenoScan.prelim()", this function scan a target region or chromosome, and output results for all windows as well as an estimated significance threshold. For genome-wide scan, users can scan each chromosome individually, then the genome-wide significance threshold can be obtained by combining chromosome-wise thresholds:

alpha=1/(1/alpha_1+1/alpha_2+...+1/alpha_22).

Usage

```
GenoScan.VCF.chr(result.prelim,vcf.filename,chr,pos.min=NULL,pos.max=NULL,
Gsub.id=NULL,annot.filename=NULL,cell.type=NULL,MAF.weights='beta',
test='combined',window.size=c(5000,10000,15000,20000,25000,50000),
MAF.threshold=1,impute.method='fixed')
```

Arguments

result.prelim	The output of function "GenoScan.prelim()"
vcf.filename	A character specifying the directory (including the file name) of the vcf file.
chr	Chromosome number.
pos.min	Minimum position of the scan. The default is NULL, where the scan starts at the first base pair.
pos.max	Maximum position of the scan. The default is NULL, where the scan ends at the last base pair, according to the chromosome sizes at: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes.
Gsub.id	The subject id corresponding to the genotype matrix, an n dimensional vector. This is used to match phenotype with genotype. The default is NULL, where the subject id in the vcf file is used.
annot.filename	A character specifying the directory (including the file name) of functional annotations. Currently GenoScan supports GenoNet scores across 127 tissues/cell types, which can be downloaded at: http://www.openbioinformatics.org/annovar/download/GenoNetScores/
cell.type	A character specifying the tissue/cell type integrated in the analysis, in addition to standard dispersion and/or burden tests. The default is NULL, where no functional annotation is included. If cell.type='all', GenoNet scores across all 127 tissues/cell types are incorperated.
MAF.weights	Minor allele frequency based weight. Can be 'beta' to up-weight rare variants or 'equal' for a flat weight. The default is 'beta'.

test	Can be 'dispersion', 'burden' or 'combined'. The test is 'combined', both dispersion and burden tests are applied. The default is 'combined'.
window.size	Candidate window sizes in base pairs. The default is c(5000,10000,15000,20000,25000,50000). Note that extemely small window size (e.g. 1) requires large sample size.
MAF.threshold	Threshold for minor allele frequency. Variants above MAF.threshold are ignored. The default is 1.
impute.method	Choose the imputation method when there is missing genotype. Can be "random", "fixed" or "bestguess". Given the estimated allele frequency, "random" simulates the genotype from binomial distribution; "fixed" uses the genotype expectation; "bestguess" uses the genotype with highest probability.

Value

window.summary	Results for all windows. Each row presents a window.
M	Estimated number of effective tests.
threshold	Estimated threshold, 0.05/M.

Examples

```
# load example vcf file from package "seqminer"
vcf.filename = system.file("vcf/all.anno.filtered.extract.vcf.gz", package = "seqminer")

# simulated outcomes, covariates and inidividual id.
Y<-as.matrix(rnorm(3,0,1))
X<-as.matrix(rnorm(3,0,1))
id<-c("NA12286", "NA12341", "NA12342")

# fit null model
result.prelim<-GenoScan.prelim(Y,X=X,id=id,out_type="C",B=5000)

# scan the vcf file
result<-GenoScan.VCF.chr(result.prelim,vcf.filename,chr=1,pos.min=196621007,pos.max=196716634)


## this is how the actual genotype matrix from package "seqminer" looks like
example.G <- t(readVCFToMatrixByRange(vcf.filename, "1:196621007-196716634",annoType='')[[1]])
```

Index