Package ‘eSCAN’

June 29, 2020

Type Package

Title Scan the Enhancers: Scan Regulatory Regions for Rare Variants Aggregate Association Testing using Whole Genome Sequencing Data

Version 0.1.1

Date 2020-06-23

Author Yingxi Yang

Maintainer Yingxi Yang <yangyx117@gmail.com>

Description

eSCAN (or “Scan the Enhancers”, with “enhancers” as a shorthand for any potential regulatory regions in the genome) is an R package for performing genome-wide assessment of rare variants residing in enhancer regions that are significantly associated with a phenotype, combining the advantages of dynamic window selection with the advantages of increasing genomic annotation information, including chromatin accessibility, histone markers, and 3D chromatin conformation. eSCAN defines test windows across the genome by genomic annotation either specified by users or provided by eSCAN package such that each window marks putative regulatory region(s). eSCAN then searches the defined windows using fastSKAT and detects significant regions by an empirical/analytic threshold which adjusts for multiple testing of all the searching windows across the genome, of different sizes, including some overlapping windows.

License GPL-3

Encoding UTF-8

Imports RcppArmadillo, Rcpp (>= 1.0.4.6), Matrix, GENESIS, methods, survey, CompQuadForm

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.1.0

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

eGenerator 2
eSCAN 2
fitNull 3
preprocess 4

Index **5**

eGenerator *Generate locations of the enhancers*

Description

The eGenerator function generates locations of the regulatory regions if not specified by users.

Usage

```
eGenerator(geno, maxgap = 10^4)
```

Arguments

geno	an n*p genotype matrix, where n is the sample size and p is the number of rare variants included.
maxgap	threshold to split independent loci (default=10^4).

Value

The function returns a data frame containing the generated locations of candidate enhancers. The first column is index of the enhancers. The next two columns are start and end positions of the enhancers. The next two columns are start and end index of the enhancers (the index in the geno matrix sorted by genomic positions). The last column is length of the enhancers.

eSCAN *Scan the enhancers*

Description

The eSCAN function is the main function in the package. It takes in genotype matrix, null model object and annotation information which could be specified by users, and then detects rare variants association between a quantitative/dichotomous phenotype and regulatory regions in whole-genome sequencing data by using eSCAN procedure.

Usage

```
eSCAN(
  genotype,
  nullmod,
  new_enloc = NULL,
  gap = 10^4,
  times = 1000,
  alpha = 0.05,
  analy = FALSE
)
```

Arguments

`genotype`	an n*p genotype matrix, where n is the sample size and p is the number of rare variants included.
`nullmod`	a null model object returned from `fitNull`. Note that `fitNull` is a wrapper of `fitNullModel` function from the `GENESIS` package.
`new_enloc`	a data frame of annotation information with dimension q*6, where q is the number of candidate regulatory regions. The six columns indicate index, start position, end position, start index, end index in the `genotype` matrix sorted by genomic positions and length of the enhancers, respectively. If annotation information is not specified by users (default=NULL), a data frame of locations of the enhancers will be automatically created by `eGenerator`.
`gap`	if `new_enloc` is not specified by users, this parameter will be used to generate locations of the enhancers in the function `eGenerator`, where gap is the threshold to split independent loci (default=10^4).
`times`	the number of MC simulations (default=1000).
`alpha`	significance level (default=0.05).
`analy`	TRUE indicates analytic threshold, FALSE indicates empirical threshold (default=FALSE).

Value

The function returns a list containing the following elements:

`res`: a matrix of significant regions detected by eSCAN. The first column is the p-value of the detected region(s). The next two columns are start and end positions of the detected region(s).

`res0`: a matrix to summarize all the regions included in the analysis. The first column is the p-value of the regulatory regions. The next two columns are start and end positions of the regulatory regions.

`thres`: threshold of eSCAN to control the family-wise/genome-wide error at `alpha` level.

`fitNull` *Fit a generalized linear model under the null hypothesis for unrelated samples*

Description

The `fitNull` function is a wrapper of `fitNullModel` from the `GENESIS` package. It fits a regression model under the null hypothesis for unrelated samples which is a preparation for subsequent analysis.

Usage

```
fitNull(x, outcome = NULL, covars = NULL, fam = "gaussian")
```

Arguments

`x`	a data frame containing outcome variable and covariates.
`outcome`	a character string specifying the name of outcome variable in `x`.
`covars`	a vector of character strings specifying the names of covariates in `x`.
`fam`	can be either "gaussian" for a continuous phenotype or "binomial" for a binary phenotype.

Value

The function returns an object of model fitted from `fitNullModel`. See `fitNullModel` in the `GENESIS` package for more details.

References

Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics.

`preprocess` *Data preprocessing*

Description

This function is data preparation for subsequent analysis using eSCAN.

Usage

```
preprocess(geno, enhancer, gap)
```

Arguments

`geno` an n*p genotype matrix, where n is the sample size and p is the number of rare variants included.

`enhancer` a data frame of annotation information with dimension q*6, where q is the number of candidate regulatory regions. The six columns indicate index, start position, end position, start index, end index the `geno` matrix sorted by genomic positions and length of the enhancers, respectively. If annotation information is not specified by users (default=NULL), a data frame of locations of the enhancers will be automatically created by `eGenerator`.

`gap` threshold to split independent loci (default=10^4).

Value

The function returns a list with the following elements:

`genotype`: an n*p genotype matrix sorted by genomic positions.

`MAF`: a vector (length=p) of minor allele frequencies.

`new_enloc`: a data frame of locations of the enhancers of dimension q*6, where q is the number of candidate regulatory regions. The six columns indicate index, start position, end position, start index, end index in the `genotype` matrix sorted by genomic postions and length of the enhancers, respectively.

Index

eGenerator, 2, *3*, *4*
eSCAN, 2

fitNull, *3*, 3
fitNullModel, *3*, *4*

GENESIS, *3*, *4*

preprocess, 4