Package ‘SCANG’

November 13, 2018

Type Package
Title Scan the Genome
Version 0.1.0
Date 2018-11-12
Authors Zilin Li <li@hsph.harvard.edu>, Xihao Li <xihaoli@g.harvard.edu>
Maintainer Zilin Li <li@hsph.harvard.edu>
Description An R package for performing SCANG procedure in whole genome sequencing studies
License GPL-3
Copyright See COPYRIGHTS for details.
Encoding UTF-8
LazyData true
Depends R (>= 3.0.0)
Imports Matrix
LinkingTo Rcpp, RcppArmadillo
RoxygenNote 6.1.0
Suggests knitr,
rmarkdown
VignetteBuilder knitr

R topics documented:

SCANG_B .......................................................... 2
SCANG_O .......................................................... 3
SCANG_S .......................................................... 4

Index 6
SCANG_B

SCANG procedure using burden test

Description

Take in genotype, phenotype and covariates, and analyzes the target sequence by using SCANG-B procedure. For each region, the scan statistic is the set-based p-value of Burden(1,1) or Burden(1,25).

Usage

SCANG_B(genotype, phenotype, X, family, weightstype, Lmax, Lmin, maf, 
steplength = 5, rnt = FALSE, times = 2000, alpha = 0.05, 
f = 0.5)

Arguments

geno**type**  An n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of variants. Note that the required format of the genotype matrix is general column-oriented numeric sparse matrix.

phenotype  A phenotype vector of length n.

X  A covariate matrix of dimension n*q.

family  either 'gaussian', for continuous data or 'binomial' for 0/1 outcomes.

weightstype  the type of weight function. weightstype=1 stands for beta(maf;1,1) and weightstype=2 stands for beta(maf;1,25).

Lmax  The maximum number of variants in searching windows.

Lmin  The minimum number of variants in searching windows.

maf  A vector of minor allele frequency for all the variants in the target sequence. The length is p.

steplength  difference of number of variants in searching windows, that is, the number of variants in searching windows are Lmin, Lmin+steplength, Lmin+steplength,...,Lmax. The default value is 5.

rnt  logical. Should do rank normal transformation of phenotype?

times  Monte-Carlo simulation times for generating the empirical threshold. Default value is 2000.

alpha  genome-wide significane level.

f  An overlap fraction, which controls for the overlapping of detected regions, for example, when f=0, the detected regions are non-overlapped with each other, and when f=1, we keep every susceptive region as detected regions. The default value is 0.5.

Value

a list with the following members:

SCANG_B_res: A matrix which summarized the significant region detected by SCANG-B. The first column is the -log(p-value) of the detected region. The next two columns are the location of
SCANG_O

the detected region (in sense of variants order). The last column is the family-wise/genome-wide error rate of the detected region. The result (0,0,0,1) means there is no significant region.

SCANG_B_top1: A vector of length 4 which summarized the top 1 region detected by SCANG-B. The first element is the -log(p-value) of the region. The next two elements are the location of the detected region (in sense of variants order). The last element is the family-wise/genome-wide p-value.

SCANG_B_thres Empirical: threshold of SCANG-B for controlling the family-wise type I error at alpha level.

SCANG_B_thres_boot: A vector of Monte Carlo simulation sample for generating the empirical threshold. The 1-alpha quantile of this vector is the empirical threshold.

Examples

SCANG_B(genotype, phenotype, X, family, weightstype, Lmax, Lmin, maf)

---

**Description**

Take in genotype, phenotype and covariates, and analyzes the target sequence by using SCANG-O procedure. For each region, the scan statistic is the set-based p-value of an omnibus test that aggregated SKAT(1,1), SKAT(1,25), Burden(1,1) and Burden(1,25) using Cauchy method.

**Usage**

```r
SCANG_O(genotype, phenotype, X, family, Lmax, Lmin, maf, steplength = 5, 
rnt = FALSE, times = 2000, alpha = 0.05, filter = 1e-04, 
f = 0.5)
```

**Arguments**

- `genotype`: An n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of variants. Note that the required format of the genotype matrix is general column-oriented numeric sparse matrix.
- `phenotype`: A phenotype vector of length n.
- `X`: A covariate matrix of dimension n*q.
- `family`: either 'gaussian', for continuous data or 'binomial' for 0/1 outcomes.
- `Lmax`: The maximum number of variants in searching windows.
- `Lmin`: The minimum number of variants in searching windows.
- `maf`: A vector of minor allele frequency for all the variants in the target sequence. The length is p.
- `steplength`: difference of number of variants in searching windows, that is, the number of variants in searching windows are Lmin, Lmin+steplength, Lmin+steplength,...,Lmax. The default value is 5.
- `rnt`: logical. Should do rank normal transformation for continuous trait?
- `times`: Monte-Carlo simulation times for generating the empirical threshold. Default value is 2000.
alpha: family-wise/genome-wide significance level.

filter: a filtering threshold of screening method for SKAT. SKAT p-values are calculated for regions whose p-value is possibly smaller than the filtering threshold. Default value is $1e^{-4}$. For the whole genome, we recommend to set it as $1e^{-6}$.

f: An overlap fraction, which controls for the overlapping of detected regions, for example, when $f=0$, the detected regions are non-overlapped with each other, and when $f=1$, we keep every susceptible region as detected regions. The default value is 0.5.

Value

a list with the following members:

SCANG_O_res: A matrix which summarized the significant region detected by SCANG-O. The first column is the $-\log(p$-value) of the detected region. The next two columns are the location of the detected region (in sense of variants order). The last column is the family-wise/genome-wide error rate of the detected region. The result (0,0,0,1) means there is no significant region.

SCANG_O_top1: A vector of length 4 which summarized the top 1 region detected by SCANG-O. The first element is the $-\log(p$-value) of the region. The next two elements are the location of the detected region (in sense of variants order). The last element is the family-wise/genome-wide p-value.

SCANG_O_thres: Empirical threshold of SCANG-O for controlling the family-wise type I error at alpha level.

SCANG_O_thres_boot: A vector of Monte Carlo simulation sample for generating the empirical threshold. The 1-alpha quantile of this vector is the empirical threshold.

SCANG_S_1_25_res, SCANG_S_1_25_thres, SCANG_S_1_25_top1, SCANG_S_1_25_thres_boot: Analysis results using SCANG-S(1,25). Details see SCANG-O.

SCANG_B_1_25_res, SCANG_B_1_25_thres, SCANG_B_1_25_top1, SCANG_B_1_25_thres_boot: Analysis results using SCANG-B(1,25). Details structure see SCANG-O.

Examples

SCANG_O(genotype, phenotype, X, family, Lmax, Lmin, maf)

---

SCANG_S: SCANG procedure using SKAT

Description

Take in genotype, phenotype and covariates, and analyzes the target sequence by using SCANG-S procedure. For each region, the scan statistic is the set-based p-value of SKAT(1,1) or SKAT(1,25).

Usage

SCANG_S(genotype, phenotype, X, family, weightstype, Lmax, Lmin, maf, steplength = 5, rnt = FALSE, times = 2000, alpha = 0.05, filter = 1e-04, f = 0.5)
Arguments

- **genotype**: An n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of variants. Note that the required format of the genotype matrix is general column-oriented numeric sparse matrix.

- **phenotype**: A phenotype vector of length n.

- **X**: A covariate matrix of dimension n*q.

- **family**: either 'gaussian', for continuous data or 'binomial' for 0/1 outcomes.

- **weightstype**: the type of weight function. weightstype=1 stands for beta(maf;1,1) and weightstype=2 stands for beta(maf;1,25).

- **Lmax**: The maximum number of variants in searching windows.

- **Lmin**: The minimum number of variants in searching windows.

- **maf**: A vector of minor allele frequency for all the variants in the target sequence. The length is p.

- **steplength**: difference of number of variants in searching windows, that is, the number of variants in searching windows are Lmin, Lmin+steplength, Lmin+steplength,...,Lmax. The default value is 5.

- **rnt**: logical. Should do rank normal transformation of phenotype?

- **times**: Monte-Carlo simulation times for generating the empirical threshold. Default value is 2000.

- **alpha**: genome-wide significance level.

- **filter**: a filtering threshold for screening method. SKAT p-values are calculated for regions whose p-value is smaller than the filtering threshold. Default value is 1e-4. For the whole genome, we recommend to set 1e-6.

- **f**: An overlap fraction, which controls for the overlapping of detected regions, for example, when f=0, the detected regions are non-overlapped with each other, and when f=1, we keep every susceptible region as detected regions. The default value is 0.5.

Value

A list with the following members:

- **SCANG_S_res**: A matrix which summarized the significant region detected by SCANG-S. The first column is the -log(p-value) of the detected region. The next two columns are the location of the detected region (in sense of variants order). The last column is the family-wise/genome-wide error rate of the detected region. The result (0,0,0,1) means there is no significant region.

- **SCANG_S_top1**: A vector of length 4 which summarized the top 1 region detected by SCANG-S. The first element is the -log(p-value) of the region. The next two elements are the location of the detected region (in sense of variants order). The last element is the family-wise/genome-wide p-value.

- **SCANG_S_thres**: Empirical threshold of SCANG-S for controlling the family-wise type I error at alpha level.

- **SCANG_S_thres_boot**: A vector of Monte Carlo simulation sample for generating the empirical threshold. The 1-alpha quantile of this vector is the empirical threshold.

Examples

```
SCANG_S(genotype,phenotype,X,family,weightstype,Lmax,Lmin,maf)
```
Index

SCANG_B, 2
SCANG_O, 3
SCANG_S, 4