Package 'GWsignif'

September 12, 2016

Type Package

Title Estimating Genome-Wide Significance for Whole Genome Sequencing Studies, Either Single SNP Tests or Region-Based Tests

Version 1.2

Date 2016-09-10

Author ChangJiang Xu and Celia M.T. Greenwood

Maintainer ChangJiang Xu <changjiang.h.xu@gmail.com>

Description

The correlations and linkage disequilibrium between tests can vary as a function of minor allele frequency thresholds used to filter variants, and also varies with different choices of test statistic for region-based tests. Appropriate genome-wide significance thresholds can be estimated empirically through permutation on only a small proportion of the whole genome.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2016-09-12 20:55:00

R topics documented:

GWsignif-package 1
GWsignif 2

Index **6**

`GWsignif-package` *Estimating Genome-Wide Significance for Whole Genome Sequencing Studies, Either Single SNP Tests or Region-Based Tests*

Description

The correlations and linkage disequilibrium between tests can vary as a function of minor allele frequency thresholds used to filter variants, and also varies with different choices of test statistic for region-based tests. Appropriate genome-wide significance thresholds can be estimated empirically through permutation on only a small proportion of the whole genome.

Details

Package:	GWsignif
Type:	Package
Version:	1.2
Date:	2016-09-10
License:	GLP 2.0 or greater

Author(s)

ChangJiang Xu and Celia M.T. Greenwood Maintainer: ChangJiang Xu <changjiang.h.xu@gmail.com>

References

ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, Celia M.T. Greenwood (2014) Estimating genome-wide significance for whole genome sequencing studies. Genetic Epidemiology

---

`GWsignif` *Estimating Genome-Wide Significance for Whole Genome Sequencing Studies, Either Single SNP Tests or Region-Based Tests*

Description

The correlations and linkage disequilibrium between tests can vary as a function of minor allele frequency thresholds used to filter variants, and also varies with different choices of test statistic for region-based tests. Appropriate genome-wide significance thresholds can be estimated empirically through permutation on only a small proportion of the whole genome.

Usage

```
GWsignif(pvalues, files = NULL, readFun = read.table, header = FALSE,
ntest.genome, K = 5, alpha = 0.05, plot.it = TRUE)
```

Arguments

pvalues	A matrix or data frame of permutation testing pvalues. These p-values could arise from single SNP tests or from region-based tests. The rows of the matrix or data frame should correspond to permutations, and the columns to the results for each SNP or tested region. The columns should be ordered by genomic position. Usually, permutation analysis would be undertaken for only a small proportion of the whole genome, such as one moderately-sized chromosome.
files	A vector of the names of the files which the permutation testing pvalues may be read from. Each file contains a table of the permutation pvalues in a sub-genomic region. The rows correspond to permutations and columns to the SNP and/or region based testings in the sub-genome region. The file names should be ordered by genomic position. If the variable files is not NULL, the permutation testing pvalues are to be importd by reading the files using the read function readFun.
readFun	Data input function. readFun = read.table, read.csv, or read.delim for reading tabular data txt files (.txt), comma separated value files (.csv), or delimited files (defaulting to the TAB character for the delimiter).
header	A logical value of TRUE or FALSE indicating whether the files contain the names of the variables as the first line.
ntest.genome	The total number of tests for which there is a goal of estimating the genome-wide significance threshold. For example it could be the number of genome-wide tests.
K	See details below. This parameter determines, for the part of the genome that was used for permutation analysis, how it is divided in order to do the extrapolation. Values of K around 5-9 are typical. Default set to 5.
alpha	Family-wise error rate (FWER), default 0.05.
plot.it	Plot a figure if TRUE.

Details

The function estimates the significance threshold needed to control the family-wise error rate (FWER) for a large number of tests, based on an extrapolation approach described in Xu et al. 2014, Genetic Epidemiology 38(4): 281-290. The required input is a set of p-values obtained via permutation, for a subset of the total number of tests to be performed. The tests used for permutation should lie on the same chromosome and be ordered by position in the input matrix; permutations should be performed to maintain the correlation structure between the genetic data, i.e. by permuting the phenotype and then repeating all analyses along the chromosome.

To give an example, suppose there are m=500,000 tests to be performed on one chromosome, and 10 million tests to be performed genome-wide. A significance threshold controlling FWER at 0.05 is desired. GWsignif will divide the m=500,000 tests into 2^(k-1) subset(s) of approximately equal size, each consisting of m/(2^(k-1)) tests, for k=1,2,...,K. Ideally, K should be chosen such that there are still several thousand tests in the smallest subset. For example, if K=8, then m/(2^(K-1)) = 3906.25. The 500,000 tests will therefore be divided into 128 sets of approximately 3906 tests, then 64 sets of 7812 tests (2*3906), etc., and finally 2 sets of 250,000 tests and one set of 500,000 tests. Within each of these sets, and for each permutation, the minimum p-value is calculated, and the alpha percentile of these minimum p-values is then estimated across the permutations. (There

should be sufficient permutations to estimate the alpha-th percentile with some accuracy). It is then possible to fit a regression line between the naive Bonferroni threshold that would be required for m/(2^(k-1)) tests and the alpha percentile of the minimum p-values (on the negative log10 scale); this line is then used for extrapolation.

If the variable files is not NULL, the permutation testing pvalues are to be importd by reading the files using the read function readFun. This option is useful for large permutations that need large memory size, but the running time will be much longer.

Value

`qminp`	a list of alpha-quantiles of minimum pvalues within a subset of tests for k = 1,2,...,K. The k-th element in the list consists of a vector of 2^(k-1) alpha-quantiles.
`mlogq`	a matrix consisting of ntest (number of tests in a subset), bonf (-log10 of Bonferroni correction, i.e., -log10(alpha/ntest)), mean and standard error of -log10(qminp). Each row represents a different subset.
`alpha`	a desired family-wise error rate (FWER)
`ntest.region`	the number of tests for which permutation results are provided in the pvalues input matrix.
`ntest.genome`	the number of tests in the larger region of interest, i.e. genome-wide.
`GWsignif.threshold`	estimated significance threshold in the larger region

Author(s)

ChangJiang Xu, Celia M.T. Greenwood

References

ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, Celia M.T. Greenwood (2014) Estimating genome-wide significance for whole genome sequencing studies. Genetic Epidemiology

Examples

```
nr <- 100
nSubgenome <- 3
pvalues <- NULL
for (i in 1:nSubgenome){
nc <- round(runif(1, 1000, 1500))
dat <- matrix(runif(nr*nc), nr, nc)
write.table(dat, file = paste0("testdata", i, ".txt"),
row.names = FALSE, col.names = FALSE, quote = FALSE, sep = "\t")
pvalues <- cbind(pvalues, dat)
}

pvaluefiles <- paste0("testdata", 1:nSubgenome, ".txt")

ptm <- proc.time()
```

```
s <- GWsignif(pvalues)
proc.time() - ptm
s$mlogq

ptm <- proc.time()
sf <- GWsignif(files = pvaluefiles)
proc.time() - ptm
sf$mlogq
```

Index

∗Topic **package**
GWsignif-package, 1

GWsignif, 2
GWsignif-package, 1