Package 'GESE'

November 16, 2016

Type Package

Title Gene-Based Segregation Test

Version 2.0.0

Date 2016-11-15

Author Dandi Qiao, Michael H. Cho

Maintainer Dandi Qiao <redaq@channing.harvard.edu>

Description Implements the gene-based segregation test(GESE) and the weighted GESE test for identifying genes with causal variants of large effects for family-based sequencing data. The methods are described in Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies. More details can be found at <http://scholar.harvard.edu/dqiao/gese>.

Depends kinship2

License GPL-2

NeedsCompilation yes

Suggests knitr, rmarkdown

VignetteBuilder knitr

Repository CRAN

Date/Publication 2016-11-16 23:56:52

R topics documented:

GESE-package *Gene-Based Segregation Test*

Description

Implements the gene-based segregation test(GESE) and the weighted GESE test for identifying genes with causal variants of large effects for family-based sequencing data. The methods are described in Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies. More details can be found at <http://scholar.harvard.edu/dqiao/gese>.

Details

The DESCRIPTION file:

Package:	GESE
Type:	Package
Title:	Gene-Based Segregation Test
Version:	2.0.0
Date:	2016-11-15
Author:	Dandi Qiao, Michael H. Cho
Maintainer:	Dandi Qiao <redaq@channing.harvard.edu>
Description:	Implements the gene-based segregation test(GESE) and the weighted GESE test for identifying genes w
Depends:	kinship2
License:	GPL-2
NeedsCompilation:	yes
Suggests:	knitr, rmarkdown
VignetteBuilder:	knitr

Index of help topics:

```
GESE                    Gene-Based Segregation Test
GESE-internal functions
                        GESE package internal functions
GESE-package            Gene-Based Segregation Test
condSegProbF            Computes conditional segregation probability
                        for the family
dataRaw                 dataRaw - a data frame containing the pedigree,
                        phenotype and genotype information
database                database file in example
getSegInfo              Computes segregation information for different
                        mode of inheritance.
mapInfo                 mafInfo - example data
pednew                  pednew - an example pedigree structure
trim_oneLineage         Trims the pedigree structure to include one
```

	lineage only.
trim_unrelated	Trims the pedigree structure to exclude multiple founder cases

computes gene-based segregation tests(GESE and weighted GESE) for family-based sequencing data. The main functions are: GESE: computes gene-based segregation information and GESE test p-values (unweighted and weighted version). trim_oneLineage: trims the pedigree so that for any subject, either the paternal family or the maternal family is included. Minimal set of sequenced subjects may be removed to ensure one lineage per pedigree only. trim_unrelated: trims the pedigree so that only one founder case is kept for each pedigree, and pedigrees with no cases are removed. condSegProbF: computes the conditional probability that a variant in the gene is segregating in the family specified, conditional on that the variant is present in the family.

Author(s)

Dandi Qiao, Michael H. Cho

Maintainer: Dandi Qiao <redaq@channing.harvard.edu>

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

http://scholar.harvard.edu/dqiao/gese

See Also

GESE

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
results <- GESE(pednew, database, 1000000, dataRaw, mapInfo, threshold=1e-2)
results
```

---

condSegProbF *Computes conditional segregation probability for the family*

Description

Computes the conditional probability that a variant is segregating in the family conditional on that the variant is present in one of the founders in the family.

Usage

```
condSegProbF(pedTemp, subjInfo)
```

Arguments

pedTemp	The data frame that includes the complete pedigree structure for the family
subjInfo	A data frame that contains the subject phenotype information for the sequenced subjects. it should include the columns FID, IID, and PHENOTYPE.

Value

returns the conditional segregating probability of a variant in the family

Author(s)

Dandi Qiao

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

See Also

GESE

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
library(kinship2)
pedigrees = kinship2::pedigree(pednew$IID, pednew$faID, pednew$moID,pednew$sex,famid=pednew$FID)
subjects= dataRaw[,c(1,2,6)]
condSegProbF(pedigrees['93'], subjects)
condSegProbF(pedigrees['412'], subjects)
results2 <- GESE(pednew, database, 1000000, dataRaw, mapInfo, threshold=1e-2)
results2$condSegProb
```

---

database *database file in example*

Description

a data frame containing the GENE and MAF information for the variants under consideration in the public reference database.

Usage

```
data("database")
```

Format

A data frame of 20 observations on the following 3 variables.

`SNP` an unique identifier for variant

`GENE` a character vector: Gene name

`MAF` a numeric vector: minor allele frequency of the variants in the referecne database

Details

A data frame containing the information for all the variants satisfying the same filtering criteria in the chosen reference genome. It should include at least three columns with these names: SNP (unique SNP ID), GENE (gene name), MAF (minor allele frequency for the variant in reference database for the corresponding population).

Source

Randomly simulated data.

Examples

```
data(database)
```

---

dataRaw *dataRaw - a data frame containing the pedigree, phenotype and genotype information*

Description

A data frame that can be created from the `.raw` formatted filed generated by PLINK.

Usage

```
data("dataRaw")
```

Format

A data frame with 198 observations on the following 26 variables.

`FID` Family iD

`IID` Individual ID

`PAT` Father ID

`MAT` Mother ID

SEX sex

PHENOTYPE Affection status

X1 Genotype for variant 1

X2 Genotype for variant 2

X3 Genotype for variant 3

X4 Genotype for variant 4

X5 Genotype for variant 5

X6 Genotype for variant 6

X7 Genotype for variant 7

X8 Genotype for variant 8

X9 Genotype for variant 9

X10 Genotype for variant 10

Details

The number of rows equal the number of subjects in the data and the number of columns equas the number of markers M + 6. The first six columns with specific column names include: the Family ID (FID), Individual ID (IID), father ID(PAT), mother ID (MAT), sex (SEX) and affection status (PHENOTYPE). The rest of the columns containing the genotypes for the variants listed in the coreesponding mapInfo file. It is also important to make sure that the recoding is with respect to the minor allele in the population. The affection status of this file will be used as the phenotype.

Examples

```
data(dataRaw)
```

GESE *Gene-Based Segregation Test*

Description

Computes the gene-based segregation information and tests for family-based sequencing data.

Usage

```
GESE(pednew, variantInformation, dbSize, dataPed, mapInfo,
threshold = 1e-7, onlySeg = FALSE, familyWeight = NA )
```

Arguments

pednew	A data frame of the complete pedigree information for all families in the dataset. The required column names of this data frame include: FID (family ID), IID (individual ID, must be of class character), faID (father ID, NA if unavailable), moID (mother ID, NA if unavailable), and sex.
variantInformation	A data frame containing the information for all the variants satisfying the same filtering criteria in the chosen reference genome. It should include at least three columns with these names: SNP (unique SNP ID), GENE (gene name), MAF (minor allele frequency for the variant in reference database for the corresponding population).
dbSize	An integer indicating the sample size of the reference database used.
dataPed	A data frame in the raw file format generated by PLINK. The number of rows equal the number of subjects in the data and the number of columns equas the number of markers M + 6. The first six columns with specific column names include: the Family ID (FID), Individual ID (IID), father ID(PAT), mother ID (MAT), sex (SEX) and affection status (PHENOTYPE). The rest of the columns containing the genotypes for the variants listed in the coreesponding mapInfo file. It is also important to make sure that the recoding is with respect to the minor allele in the population. The affection status of this file will be used as the phenotype.
mapInfo	A data frame that contains at least two columns (required column names): variant ID (SNP) and Gene name (GENE). The number of rows equal to the number of SNPs/markers to be considered (M).
threshold	Specifies the precision needed to be reached for significant p-values. Default value is 1e-7.
onlySeg	True if only the segregation information (number of pedigrees segregating in each gene) is needed, else FALSE (DEFUALT), which computes the GESE p-values too.
familyWeight	An optional data frame. It gives the weight for the families. If it is NA, no weighting scheme is used. Otherwise, its dimenstion could be (number of families)x(number of genes+1) or (number of families)x2. The first column should be family name (column name FID). If the weights for the families are the same for all the genes, the second column should just be weight (columns name "weight"), otherwise the second column and above should be the gene names (columns names are corresponding GENE names).

Details

This is the main function in the GESE package. The gene-based segregation tests (GESE) described in Qiao et al (2016) is a segregation-based test extending the work of Bureau et al (2014) by computing the marginal probability of segregation events within a gene. The first step in this function is to trim the families such that only one lineage (with the most possible number of cases) is included (i.e. for any subject, only the information of either the parental pedigree or the maternal pedigree would be included). In addition, if multiple founder cases are present, remove the (smallest set of) founder(s) that are unrelated most other sequenced subjects. Then this function computes the

gene-based segregating information and p-values for multiple families. If only the segregation information (number of families segregating in each gene) is needed, set onlySeg = TRUE. If different family weights will be used to boost the power, assign the weights to familyWeight parameter.

Value

`segregation` a data frame containing the information about whether each gene is segregating in each family. The number of columns equals the number of families +3. The last column is the number of families the gene is segregating in. The number of rows equals the number of genes. Only this data frame and `varSeg` will be returned if onlySeg is set to TRUE.

`varSeg` a data frame containing the information about whether each variant is segregating in each family. The number of columns equals the number of families +3. The last column is the number of families the variant is segregating in. The number of rows equals the number of variants. Only this data frame and `segregation` will be returned if onlySeg is set to TRUE.

`results` This is available when onlySeg = FALSE. The datat frame contains the columns: GENE (gene name), obs_prob (the observed segregating probability for the gene), pvalue (gene-based p-value for GESE), numSim (The number of simulations used to compute the p-value if resampling-based method is used), N_seg (the number of families that are segregating in the gene). If familyWeight is not NA, obs_weight_stat (the observed weighted test statistic) and pvalue_weighted (the p-value for the weighted test statistic) will also be returned.

`condSegProb` A vector of length equals the number of families. The conditional probability of at least one variant in the gene is segregating in the family condition on at least one variant (among the set of variants to be considered) is present in the familiy.

`segProbGene` A matrix of the segregating probability for the gene and for each family. This is a working matrix that could be used in other functions.

Author(s)

Dandi Qiao

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. *Gene-based segregation method for identifying rare variants for family-based sequencing studies.*

http://scholar.harvard.edu/dqiao/gese

Bureau, A., Younkin, S.G., Parker, M.M., Bailey-Wilson, J.E., Marazita, M.L., et al. 2014. *Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives*. Bioinformatics 30, 2189-2196

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
```

```
results <- GESE(pednew, database, 1000000, dataRaw, mapInfo, threshold=1e-3)
results
```

GESE-internal *GESE package internal functions*

Description

GESE package internal functions.

Details

computeP_resampling findIntermediateFounder findMostRecentCommonFounder findMostRecentCommonFounderCo
getFounder getProb getPvalue_resampling getTranProb_dv isRelated oneSetSim segProb
getProb_weight

Author(s)

Dandi Qiao

Maintainer: Dandi Qiao <redaq@channing.harvard.edu>

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

See Also

GESE

---

getSegInfo *Computes segregation information for different mode of inheritance.*

Description

Computes variant-based and gene-based segregation information for different mode of inheritance.

Usage

```
getSegInfo(pednew, dataPed, mapInfo, mode="recessive")
```

Arguments

pednew	A data frame of the complete pedigree information for all families in the dataset. The required column names of this data frame include: FID (family ID), IID (individual ID, must be of class character), faID (father ID, NA if unavailable), moID (mother ID, NA if unavailable), and sex.
dataPed	A data frame in the raw file format generated by PLINK. The number of rows equal the number of subjects in the data and the number of columns equas the number of markers M + 6. The first six columns with specific column names include: the Family ID (FID), Individual ID (IID), father ID(PAT), mother ID (MAT), sex (SEX) and affection status (PHENOTYPE). The rest of the columns containing the genotypes for the variants listed in the coreesponding mapInfo file. It is also important to make sure that the recoding is with respect to the minor allele in the population. The affection status of this file will be used as the phenotype.
mapInfo	A data frame of at least two columns (required column names): variant ID (SNP) and Gene name (GENE). The number of rows equal to the number of SNPs/markers to be considered (M).
mode	The mode of inheriance assumed to compute the segregation information. The options are "dominant", "recessive", and "CH" (compound heterozygous). The default value is "recessive".

Details

This function is used to compute the segregation information for different mode of inheritance without computing the GESE test. The mode of inheritance supported here are: dominant, recessive and compound heterozygous (CH). For dominant mode of inheritance, a variant is segregating if all the cases in the family carry at least one alternative allele (genotype X>0), and all the controls in the family do not carry any alternative allele (X=0). For recessive mode of inheritance, a variant is segregating if all the cases in the family carry two alternative alleles (X=2), and all the controls in the family carry less than 2 alternative alleles (X=0 or X=1). For compound heterozygous mode of inheritance, a variant is segregating at two variant position if all the cases in the family carry at least one alternative allele at the two positions (X1>0 and X2>0), and all the controls in the family do not carry any alternative allele at either of the two positions (X1 = 0 or X2 = 0).

Value

varSeg	For dominant and recessive mode of inheriancce, this is a data frame containing the information about whether each variant is segregating in each family. The number of columns equals the number of families +3. The last column is the number of families the variant is segregating in. The number of rows equals the number of variants. For compound heterozygous mode of inheritance, this is a data frame containing the information of whether each pair of variants is segregating in each of the families. We consider all pairs in the dataset, if the pair of variants are not included in this data frame, they are not segregating in any families.
geneSeg	For dominant and recessive mode of inheriancce, this is a data frame containing the information about whether each gene is segregating in each family. The

number of columns equals the number of families +3. The last column is the number of families the gene is segregating in. The number of rows equals the number of genes. For compound heterozygous mode of inheritance, this is a data frame containing the information of whether any pair of variants in this gene are segregating in each of the families. The last columns is the number of families with the presence of any pair of variants segregating in the gene.

`genePairSeg` This data frame is returned only for compound heterozygous mode of inheritance. This considers any pair of genes in the data. It returns a data frame containing the information of whether any pair of variants, each in a different gene, is segregating in each of the families considered. Each row represents the information for each gene pair, summed over all possible pairs of variants in the two genes, one in each gene.

Author(s)

Dandi Qiao

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

See Also

GESE

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
result <- getSegInfo(pednew, dataRaw, mapInfo)
result$varSeg
result$geneSeg

result <- getSegInfo(pednew, dataRaw, mapInfo, mode="recessive")
result$varSeg
result$geneSeg

result <- getSegInfo(pednew, dataRaw, mapInfo, mode="CH")
result$varSeg
result$geneSeg
result$genePairSeg
```

mapInfo *mafInfo - example data*

Description

a data frame containing the gene information for the variants in the study.

Usage

```
data("mapInfo")
```

Format

A data frame of 20 observations on the following 2 variables.

`GENE` The gene name

`SNP` An unique SNP identifier

Examples

```
data(mapInfo)
```

pednew *pednew - an example pedigree structure*

Description

A data frame of the complete pedigree strucutre for the families included

Usage

```
data("pednew")
```

Format

A data frame of 1700 observations on the following 26 variables.

`FID` Family ID of class character

`IID` Individual ID of class character

`faID` Father ID, NA if missing

`moID` Mother ID, NA if missing

`sex` Sex, 1 for male, 2 for female and NA if missing.

Examples

```
data(pednew)
```

`trim_oneLineage` *Trims the pedigree structure to include one lineage only.*

Description

Trims the families to include only one lineage.

Usage

```
trim_oneLineage(seqSub, pednew)
```

Arguments

`seqSub`	A data frame that should include three columns FID (family ID), IID (individual ID), and PHENOTYPE (affection status) for the sequenced subjects in the data. One example is the 1st, 2nd and 6th columns from the plink raw format.
`pednew`	A data frame includes the complete pedigree structure information for all sequenced families in the dataset. The required column names of this data frame include: FID (family ID), IID (individual ID, must be of class character), faID (father ID, NA if unavailable), moID (mother ID, NA if unavailable), and sex.

Details

For each subject, only the maternal or the paternal family is included, since the rare variant should be present in only the related subjects. The lineage with the maximal set of sequenced cases will be used as the final pedigree.

Value

`pedInfoUpdate`	the complete pedigrees with only the paternal or maternal lineage
`seqSubjUpdate`	The sequenced subjects that are in the selected lineage are returned for the rest of the analysis.

Note

This function can be used for other analysis of family-based data processing. For example, the pre-processing step for PVAAST analysis.

Author(s)

Dandi Qiao

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

See Also

GESE, trim_unrelated

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
subjects <- dataRaw[,c(1:2, 6)]
cat("Trimming the families...\n")
cat("Trimming step 1: keep only one lineage \n")
trim <- trim_oneLineage(seqSub=subjects, pednew)
```

---

trim_unrelated *Trims the pedigree structure to exclude multiple founder cases*

Description

Trims the families to include only one founder case in each pedigree. It also exclude families with only one control subject.

Usage

```
trim_unrelated(seqSub, pednew2)
```

Arguments

seqSub	A data frame that includes at least three columns: the Family ID (FID), Individual ID (IID), and affection status (PHENOTYPE). This input values should be output from the trim_oneLineage function.
pednew2	A data frame includes the complete pedigree structure information for all sequenced families in the dataset. The required column names of this data frame include: FID (family ID), IID (individual ID, must be of class character), faID (father ID, NA if unavailable), moID (mother ID, NA if unavailable), and sex.

Details

For each pedigree,each there are multiple case founders in the pedigree, to satisfy the assumption that the causal rare variant must be introduced by one founder only, we need to keep only one case fonder that is related to most of the sequened subjects in the pedigree. We also want to remove families with no case.

Value

The modified dataPed2 file after the trimming.

Author(s)

Dandi Qiao

References

Qiao, D. Lange, C., Laird, N.M., Won, S., Hobbs, B., et al. 2016. Gene-based segregation method for identifying rare variants for family-based sequencing studies.

See Also

GESE, trim_oneLineage

Examples

```
data(pednew)
data(mapInfo)
data(dataRaw)
data(database)
subjects <- dataRaw[,c(1:2, 6)]
cat("Trimming the families...\n")
cat("Trimming step 1: keep only one lineage \n")
trim <- trim_oneLineage(seqSub=subjects, pednew)
subjects2 <- trim_unrelated(trim$seqSubjUpdate, trim$pedInfoUpdate)
```

Index