Package 'RVFamSq'

March 6th, 2019

Type Package

Title RVFamSq: Rare Variant-Family-Based Score Test for Quantitative Traits

Version 1.0

Data 2019-3-6

Author Zhihui Zhang and Suzanne Leal

Contact Zhihui Zhang zhihuiz@bcm.edu

Maintainer Zhihui Zhang zhihuiz@bcm.edu

Depends R (>=3.0.0), bbmle (>=1.0.20), mvtnorm (>=1.0.10), rlist (>=0.4.6.1)

Imports bbmle, mvtnorm, rlist, kinship2

Description A package for family-based association analysis of rare variants and quantitative traits

License GPL-3

URL *https://github.com/zhangzhhcb/RVFamSq*

LazyLoad yes

R topics documented:

RVFamSq-package *RVFamSq: Rare Variant-Family-Based Score Test for Quantitative Traits*

Description

The RVFamSq package provides an efficient approach to examine the association between the region-based rare variants and the quantitative traits in family-based data. The RVFamsSq can be broadly applied to diverse pedigrees with members missing sequence data. In addition, qualitative and quantitative covariates, e.g., age, sex, and body mass index, can be flexibly included. The RVFamSq extends the method proposed by Chen et al. , and the speed of the package is significantly optimized to analyze large-scale data sets.

Details

Package: RVFamSq
Type: Package
Version: 1.0
Data: 2019-3-6
License: GPLv3

Author(s)

Zhihui Zhang zhihuiz@bcm.edu
Suzanne Leal

References

Chen, W.-M., and Abecasis, G.R. (2007). Family-Based Association Tests for Genomewide Association Scans. Am. J. Hum. Genet. 81, 913–926.

Example.data *A small family data set for association analysis*

Description

Test data is generated for 100 nuclear families with sib-pairs and 100 sites of variants for each individual. The quantitative traits and MAF of variants are sampled randomly. The genotype of these samples are generated according to the MAF of each variant.

test.pheno A phenotype file of 100 nuclear families and 100 sites of variants for each individual. The first five columns of the file should be family IDs, individual IDs, father IDs, mother IDs, sex, respectively. The interested covariates and trait for each sample could be listed in the rest columns of the file. RVFamSq is capable to analyze multiple covariates. Each family should have information of two founders, and all individuals in test.geno should be included in this file.

test.geno A genotype file of samples included in the test.pheno file. The first two columns of the file should be family IDs and individual IDs that have the same format as in test.pheno. The rest columns of the file are genotypes of each individual and codes as 0, 1, and 2.

test.sfs A file with descriptive information for 100 genetic variants in test.geno. The file contains four columns representing the name of the interested gene, chromosome, position, and MAF of the variant, respectively.

Usage

Download the files to you working directory. The path to these files is required to use the package.

RVFamSq *Rare Variant-Family-Based Score Test for Quantitative Traits*

Description

A regional association analysis of rare variants and quantitative train in family data

Usage

RV_FamSq (genofile, phenofile, maffile, parafile, covar_col, trait_col, out,kin, start_par=NULL)

Arguments

genofile referring to the path of the genotype file. The columns in the file represent family IDs, individual IDs and genotypes of each individual. The file should not have header.

phenofile referring to the path of the phenotype file. The first five columns in the file should be family IDs, individual IDs, father IDs, mother IDs, and sex, respectively. The rest columns of the file are covariates and trait. The covariates may be multi-columns and trait should be one column. The file should not have header.

maffile the path to the .sfs file describing the information of variants in the genotype file.

parafile the path to the file saving parameters estimated in null model. If the file is not existed, the package will estimate the parameters using the available data and save the values of the parameters in parafile.

covar_col a vector defining the index of columns of covariates in phenofile.

trait_col an integer number defining the column of interested trait in phenofile.

out the directory that save the results. RVFamSq produces output file with the name of the gene, score, p-value of the gene, number of sample size and number of families.

kin a kinship matrix calculated from the pedigree using the "kinship2" package.

start-par an optional argument that defines the starting values of the parameters fitting. Users can define specific values to start the parameters fitting, otherwise, the starting values will generate randomly.

Details

RVFamSq performed association analysis by estimating the parameters under the null model and calculating the statistical score using these parameters. The parameters are estimated by maximizing the multivariate normal likelihood:

$$L = \prod_i (2\pi)^{-n_i/2} |\Omega_i|^{-1/2} e^{[y_i - E(y_i)]' \Omega_i^{-1} [y_i - E(y_i)]}$$

where n_i is the number of individuals in family i and $|\Omega_i|$ is the determinant of matrix Ω_i. $E(y_i)$ in the above function is a vector of expected phenotype of all individuals within each family and is calculated by:

$$E(y_i) = \mu + \beta_x x_i$$

where x_i is a $n_i \times q$ matrix of q covariates included in the association analysis. The parameters μ and β_x are population mean and a vector of covariate effects. The two parameters are estimated by maximizing the above likelihood function. The matrix Ω_i in the likelihood function is calculated within each family i and defined as:

$$\Omega_{ijk} \begin{cases} \sigma_g^2 + \sigma_e^2 & if\ j = k \\ 2\varphi_{ijk}\sigma_g^2 & if\ j \neq k \end{cases}$$

Here, the parameter $\sigma_g^2, and\ \sigma_e^2$ are variance components that are account for background polygenic effects and environmental effects, respectively. Additionally, φ_{ijk} denote the kinship coefficient between individuals j and k. The values of parameters $\sigma_g^2, and\ \sigma_e^2$ are also estimated by maximizing the above likelihood. With

the values of these parameters (β_x, μ, $\sigma_g^2, and\ \sigma_e^2$), we can obtain the variance-covariance matrix $\Omega_i^{(base)}$ and vector $E(y_i)^{(base)}$ based on the above equations for each family. Using these two quantities, the statistic score is defined by:

$$T^{SCORE} = \frac{\{\sum_i [G_i - E(G_i)]' [\Omega_i^{(base)}]^{-1} [y_i - E(y_i)^{(base)}]\}^2}{\sum_i [G_i - E(G_i)]' [\Omega_i^{(base)}]^{-1} [G_i - E(G_i)^{(base)}]}$$

where $E(G_i)$ is defined as:

$$E(G_i) = \sum_{m=1}^{M} 2p_m$$

where p_m is the minor allele frequency (MAF) which is obtain from the .sfs file.

Value

results a data frame results containing the name of the gene, the value of the score, the number of sample size, the number of family size and p-value write into the file named by the name of the gene under the directory defined by the argument "out".

paras a list of parameters estimated by maximizing the likelihood. The values of the "paras" save to the file defined by the argument "parafile". If the "parafile" is existed, the values of the "paras" load from the file "parafile".

Author(s)

Zhihui Zhang zhihuiz@bcm.edu
Suzanne Leal

References

Chen, W.-M., and Abecasis, G.R. (2007). Family-Based Association Tests for Genomewide Association Scans. Am. J. Hum. Genet. 81, 913–926.

Examples

```
library (kinship2)
library (RVFamSq)

##  Define variables pointing to the path to the example data
maffile<-"/directory-to-example-data/test.sfs"
genofile<-"/ directory-to-example-data/test.geno"
phenofile<-"/ directory-to-example-data/test.pheno"

##  Define output directory that save the results
out<-"/directory-to-save-the-results"

## Define the files that save the parameters estimated under the null model.
## If the file is not existed, the package will estimate the parameters based on the available data.
parafile<-"/ directory-to-example-data /paras.rds"

## Load phenotype data and calculate the kinship matrix of the pedigree.
ped_pheno<-read.table(phenofile)
```

```
## Detached “kinship” package if it is pre-loaded.
if ("kinship" %in% (.packages())){
   print("Detaching kinship library")
  detach("package:kinship", unload=TRUE)
}
kin_pre<-data.frame(id=ped_pheno[,2], mom=ped_pheno[,4], dad=ped_pheno[,3], sex=ped_pheno[,5])
tped <- with(kin_pre, pedigree(id, dad, mom, sex, famid=ped_pheno[,1]))
kin <- kinship(tped)

## Run RVFamsSq package and calculate the statistical score of the interested gene.
RV_FamSq(pedfile, phenofile, maffile, parafile, c(5,7), 6, out,kin)
```