gskat package

Xuefeng Wang

November 3, 2013

1 Overview

gskat package contains functions to test for assocation between SNP/SNV sets (with both binary and continuous phenotypes) based on collected family samples.

2 Family based assocation test with SNP data (binary traits)

2.1 Example Dataset

The package contains an example dataset (gskatda), which has a genotype matrix (Z) of 500 sib-pairs (a total of 1000 individuals) and 12 SNPs, a vector of binary phenotype (y) and a covariate matrix (X).

```
> library(gskat)
> data(gdata)
> names(gdata)

[1] "ID" "y"  "X"  "Z"

> attach(gdata)
> #head(ID);head(y);head(X);head(Z)
```

To test for an association, simply run the gskat_score function to get a p-value.For continuous phenotype, use function score_FSKAT_IC_cont instead (see instructions below).

```
> gskat_score(gdata) #using the kinship working correlation matrix

$pval
[1] 0.01121034

$ifault
[1] 0
```

```
> gskat_score(gdata,F1=FALSE) #using identity working correlation

$pval
[1] 0.01121034

$ifault
[1] 0
```

When the sample size is small, gskat tends to give conservative results. We developed a resampling method to adjust for the small sample and other issues that may affect the type I error (size) of the family based association test. Run function gskat_score_pert to compute the pertubation based p-value.

```
> gskat_score_pert(gdata)$pval

           [,1]
[1,] 0.01026639
```

By default, we use 10,000 resampling replicates in the test. We recommend the default number for the genome wide scan. For the top ranked genes, we recommend to apply more replicates to acheive more accurate results (by changing the arugments "np")

```
> # gskat_score_pert(gdata,np=100000) #100000 resampling trials
>
```

If the genome-wide or simulation-based Q-Q plot (of log10 p-values) looks funky, it will help in detecting potential issues by trying other random distribution in the perturbation process.

```
> # gskat_score_pert(gdata,pw="Norm") #using normal distribtion instead of Rademacher
>
```

2.2 Data preperation

plink file: The data can be easily prepared based on plink file format. The following are example codes for getting the required format of gskat from a plink PED file. VCF file: VCF file should be converted to plink file using vcftools.

```
> # fileName=foo #plink PED file name
> # system(paste("plink --noweb --file ", fileName, " --recodeA --out TEMP/",
> #        fileName,sep="") #plink additive coding
> # RAW=read.table(paste("TEMP/",fileName,".raw",sep=""),as.is=T,
> #        header=T) #read into R the plink RAW file
> # RAW=RAW[order(RAW$FID),] #sort according to Family ID
> # ID=RAW[,("FID","IID","PAT","MAT")]
> # y=RAW["PHENOTYPE"]
```

```
> # Z=as.matrix(RAW[,-1:-6])
> # #X prepared by users
> # mydata<-list(ID,y,X,Z)
> # gskat_score(mydata) #done
```

3 Family based assocation test with sequencing data

3.1 Binary traits

gskat package also provides a function (gskat_seq) for assocation test with rare variants (in its beta version and will be updated frequently). One may use the software vcftools (v2plink) to convert a VCF file to the plink file format. The following is a toy example.

```
> # attach(gdata) #same format as before
> # gskat_seq(y,XC=X,Z,ID)
> # gskat_seq(y,XC=X,Z,ID,resampling=FALSE) # get asymptoptic p-value only
> # gskat_seq(y,XC=X,Z,ID,pw="Norm") #using normal r.v. in pertubation
> # gskat_seq(y,XC=X,Z,ID,SNP.weigts)#using customized SNV weighting scores
>
```

3.2 Continuous traits

Currenlty, gskat package implements the test for continous traits in a separate fucntion gskat_seq_cont, which uses the same data format and argument setting as in the function for binary traits. The function can also be used for assoation tests with common varaints by setting SNP weight on equal SNP to 1.

```
> # attach(gdata)
> # gskat_seq_cont(y,XC=X,Z,ID)
> # gskat_seq_cont(y,XC=X,Z,ID, w_a=1,w_b=1) #common variants with no/equal weights
```