FRGEpistasis: A Tool for Epistasis Analysis Based on Functional Regression Model

Futao Zhang,Momiao Xiong
School of Public Health
University of Texas Health Science Center at Houston

March 17, 2014

1 Overview

FRGEpistasis is designed to detect the epistasis between genes or genomic regions for both common variants and rare variants. Currently FRGepi was developed by Futao Zhang with R language and maintained in Xiong'slab at UTSPH. This tool is friendly, convenient and memory efficient. It currently has the following functional modules:

- Epistasis test using Functional Regression Model
- Epistasis test using Principal Components Analysis
- Epistasis test of Pointwise

2 Data Formats

In order to process large-scale NGS data we have done a lot optimazation work.This package can take the genotype on one chromesome as the input genotype unit, that means it can deal with a genotyoe file list.And the package expands the whole genotype at one time. After this step only expansion data is stored, so a lot memory space is saved. The sample data are located in the "data" directory.

2.1 Genotype file format

The format of Genotype file is very similar to PLINK PED file: the first six columns are Family ID, Individual ID, Paternal ID, Maternal ID, Sex and Phenotype. The data column 7 onwards are genotypes coding in 0,1,2 where the title of the column is RS and missing value is coded as 3.

```
> geno_info <- read.csv(system.file("extdata", "simGeno-chr2.csv", package="FRGEpistasis"),header=TRUE)
> geno_info[1:5,1:9]
```

```
      FID     IID PAT MAT SEX PHENOTYPE rs74017040_1 rs74017445_1 rs74017463_1
1 sample0 sample0   0   0   0    44.298            0            0            0
2 sample1 sample1   0   0   0    31.779            0            0            0
3 sample2 sample2   0   0   0    40.446            0            0            0
4 sample3 sample3   0   0   0    61.632            0            0            0
5 sample4 sample4   0   0   0    32.742            0            0            0
```

2.2 Map file format

Map file Contains 4 columns: Chromosome, snp identifier, Genetic distance, base-pair genomic position. The map file has no header line.

```
> map_info <- read.table(system.file("extdata", "chr2.map", package="FRGEpistasis"))
> map_info[1:5,]

  V1       V2 V3       V4
1  2 74017040  0 74017040
2  2 74017445  0 74017445
3  2 74017463  0 74017463
4  2 74017499  0 74017499
5  2 74017589  0 74017589

>
```

2.3 Phenotype file format

Phenotype file contains 2 columns: Individual ID and phenotype.

```
> pheno_info <- read.csv(system.file("extdata", "phenotype.csv", package="FRGEpistasis"),header=TRUE)
> pheno_info[1:5,]

      IID PHENOTYPE
1 sample0    44.298
2 sample1    31.779
3 sample2    40.446
4 sample3    61.632
5 sample4    32.742
```

2.4 Gene Annotation file format

Gene Annotation file Contains contains 4 columns indicate the gene name, chromosome, gene start position and gene end position.

```
> gene.list<-read.csv(system.file("extdata", "gene.list.csv", package="FRGEpistasis"))
> gene.list

  Gene_Symbol Chromosome     Start       End
1 SimuRegion1          1 159824106 159832447
2 SimuRegion2          2  74017030  74063042
3 SimuRegion3          2  78365582  78385273
4 SimuRegion4          3  88182642  88222051
5 SimuRegion5          3 100054649 100063454
6 SimuRegion6          3 153071932 153080898
7 SimuRegion7          4  40380093  40401075
8 SimuRegion8          5  34493060  34503988
```

2.5 genotype file list and map file list

Because the NGS data are large, They are always organized in many files. For example, some data are organized in chromosome form. In order to bring convenience to users and alleviate the burden of the memory, FRGEpistasis can handle a list of genotype files.

genotype file list:

```
> geno_files<-read.table(system.file("extdata", "list_geno.txt", package="FRGEpistasis"))
> geno_files

               V1
1 simGeno-chr1.csv
2 simGeno-chr2.csv
3 simGeno-chr3.csv
4 simGeno-chr4.csv
5 simGeno-chr5.csv
```

map file list:

```
> map_files<-read.table(system.file("extdata", "list_map.txt", package="FRGEpistasis"))
> map_files

       V1
1 chr1.map
2 chr2.map
3 chr3.map
4 chr4.map
5 chr5.map
```

3 Implementation

3.1 Environment Requirement

- a: R version 3.0.1 or later needed.
- b: fda package is needed.
- c: In Windows system Environment Variable "PATH" should be set to let Operating System know where to find the R executable files.

3.2 Run

```
> library(MASS)
> library(fda)
> library(FRGEpistasis)
> #work_dir<-"./"
> work_dir <-paste(system.file("extdata", package="FRGEpistasis"),"/",sep="")
> #read the list of genotype files
> geno_files<-read.table(system.file("extdata", "list_geno.txt", package="FRGEpistasis"))
> #read the list of map files
> map_files<-read.table(system.file("extdata", "list_map.txt", package="FRGEpistasis"))
> #read the phenotype file
> pheno_info <- read.csv(system.file("extdata", "phenotype.csv", package="FRGEpistasis"),header=TRUE)
> #read the gene annotation file
> gene.list<-read.csv(system.file("extdata", "gene.list.csv", package="FRGEpistasis"))
> #define the extension scope of gene region
> rng=0
> # output data structure
> out_epi <- data.frame( )
> ##log transformation
```

```
> pheno_info [,2]=log(pheno_info [,2])
> ##rank transformation
> #c=0.5
> #pheno_info [,2]=rank_trans_pheno(pheno_info [,2],c)
>
> # test epistasis with Functional Regression Model
> out_epi = FRG_interaction(work_dir,pheno_info,geno_files,map_files,gene.list,rng)

[1] "Expansion gene1 of 1 on chromsome1!"
[1] "Expansion gene1 of 2 on chromsome2!"
[1] "Expansion gene2 of 2 on chromsome2!"
[1] "Expansion gene1 of 3 on chromsome3!"
[1] "Expansion gene2 of 3 on chromsome3!"
[1] "Expansion gene3 of 3 on chromsome3!"
[1] "Expansion gene1 of 1 on chromsome4!"
[1] "Expansion gene1 of 1 on chromsome5!"
[1] "Performing the gene interaction test inner chromsome 1"
[1] "1 of 1 with other gene both on 1 chromsome!"
[1] "Performing the gene interaction test outer  1 : 2 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 2 chromsome(2genes)"
[1] "Performing the gene interaction test outer  1 : 3 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 3 chromsome(3genes)"
[1] "Performing the gene interaction test outer  1 : 4 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 4 chromsome(1genes)"
[1] "Performing the gene interaction test outer  1 : 5 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing the gene interaction test inner chromsome 2"
[1] "1 of 2 with other gene both on 2 chromsome!"
[1] "2 of 2 with other gene both on 2 chromsome!"
[1] "Performing the gene interaction test outer  2 : 3 chromsomes!"
[1] "1 of 2 on  2 chromsome with other genes on 3 chromsome(3genes)"
[1] "2 of 2 on  2 chromsome with other genes on 3 chromsome(3genes)"
[1] "Performing the gene interaction test outer  2 : 4 chromsomes!"
[1] "1 of 2 on  2 chromsome with other genes on 4 chromsome(1genes)"
[1] "2 of 2 on  2 chromsome with other genes on 4 chromsome(1genes)"
[1] "Performing the gene interaction test outer  2 : 5 chromsomes!"
[1] "1 of 2 on  2 chromsome with other genes on 5 chromsome(1genes)"
[1] "2 of 2 on  2 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing the gene interaction test inner chromsome 3"
[1] "1 of 3 with other gene both on 3 chromsome!"
[1] "2 of 3 with other gene both on 3 chromsome!"
[1] "3 of 3 with other gene both on 3 chromsome!"
[1] "Performing the gene interaction test outer  3 : 4 chromsomes!"
[1] "1 of 3 on  3 chromsome with other genes on 4 chromsome(1genes)"
[1] "2 of 3 on  3 chromsome with other genes on 4 chromsome(1genes)"
[1] "3 of 3 on  3 chromsome with other genes on 4 chromsome(1genes)"
[1] "Performing the gene interaction test outer  3 : 5 chromsomes!"
[1] "1 of 3 on  3 chromsome with other genes on 5 chromsome(1genes)"
[1] "2 of 3 on  3 chromsome with other genes on 5 chromsome(1genes)"
[1] "3 of 3 on  3 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing the gene interaction test inner chromsome 4"
[1] "1 of 1 with other gene both on 4 chromsome!"
```

```
[1] "Performing the gene interaction test outer  4 : 5 chromsomes!"
[1] "1 of 1 on  4 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing the gene interaction test inner chromsome 5"
[1] "1 of 1 with other gene both on 5 chromsome!"

> # if you want to test epistasis with PCA method and pointwise method then implement the following comm
> out_epi = PCA_Piontwise_interaction(work_dir,out_epi,pheno_info,geno_files,map_files,gene.list,rng)

[1] "Performing the gene interaction test inner chromsome 1"
[1] "Performing the gene interaction test outer  1 : 2 chromsomes!"
[1] "Performing the gene interaction test outer  1 : 3 chromsomes!"
[1] "Performing the gene interaction test outer  1 : 4 chromsomes!"
[1] "Performing the gene interaction test outer  1 : 5 chromsomes!"
[1] "Performing the gene interaction test inner chromsome 2"
[1] "Performing the gene interaction test outer  2 : 3 chromsomes!"
[1] "Performing the gene interaction test outer  2 : 4 chromsomes!"
[1] "Performing the gene interaction test outer  2 : 5 chromsomes!"
[1] "Performing the gene interaction test inner chromsome 3"
[1] "Performing the gene interaction test outer  3 : 4 chromsomes!"
[1] "Performing the gene interaction test outer  3 : 5 chromsomes!"
[1] "Performing the gene interaction test inner chromsome 4"
[1] "Performing the gene interaction test outer  4 : 5 chromsomes!"
[1] "Performing the gene interaction test inner chromsome 5"

> # output the result to physical file
> write.csv(out_epi,"Output_Pvalues_Epistasis_Test.csv ")
>
```

4 Questions and Bug Reports

For any questions and bug reports, please contact the package maintainer Futao Zhang (futoaz@gmail.com)