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

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

October 13, 2014

1 Introduction

Epistasis is the primary factor in molecular evolution (Breen et al. 2012) and plays an important role in quantitative genetic analysis (Steen 2011). Epistasis is a phenomenon in which the effect of one genetic variant is masked or modified by one or other genetic variants and is often defined as the departure from additive effects in a linear model (Fisher 1918).
The critical barrier in interaction analysis for rare variants is that most traditional statistical methods for testing interaction were originally designed for testing the interaction between common variants and are difficult to apply to rare variants because of their prohibitive computational time and low power. The great challenges for successful detection of interactions with next-generation sequencing data are:

- (1) lack of methods for interaction analysis with rare variants.
- (2) severe multiple testing.
- (3) time consuming computations.

To meet these challenges, we shift the paradigm of interaction analysis between two loci to interaction analysis between two sets of loci or genomic regions and collectively test interaction between all possible pairs of SNPs within two genomic regions. In other words, we take a genome region as a basic unit of interaction analysis and use high dimensional data reduction and functional data analysis techniques to develop a novel functional regression model to collectively test interaction between all possible pairs of SNPs within two genome regions.
To support our method, we developed a R package named FRGEpistasis. FRGEpistasis is designed to detect the epistasis between genes or genomic regions for both common variants and rare variants. Currently FRGEpistasis was developed by Futao Zhang with R language and maintained in Xiong's lab 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

This package is memory efficient with high performance.

- `Memory efficiency`: Only store reduced expansion data of genotypes instead of raw data of genotypes. In real dataset the genotypes on different chromosome are always organized into different files. And each genotype file is very large. Reading all the files into memory is unacceptable. This package reads the files one by one and reduces the genotype dimension with Fourier expansion. After dimension reduction, the whole genome expansion genotype data can be easily stored in the memory.

- `high performance`: Each data file only needs to read once and reduce dimension once. So I/O times are reduced and repeated computing of data reduction was avoided; This method is a kind of group test. We take a gene(or genomic region) as the test unit. The number of Test is sharply reduced comparing with point-wise interaction (SNP-SNP) test; The dimension of genotype is reduced by functional expansion, So the time of each test is reduced.

In this version we implemented FDR (False Discovery Rate) for a multiple testing threshold to our package. When the FDR parameter == 1, FDR control is turned off. So users can use this parameter to switch FDR control on and off.

At present our package can not do multi-loci interaction test, that is it can not test 3-gene interaction ($G \times G \times G$) or more.

2 Installing FRGEpistasis

To install the R package FRGEpistasis, you can install it through:
library(BiocInstaller)
biocLite("FRGEpistasis")
Or download the source code from the bioconductor website.

3 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 "extdata" directory. This sample data is extracted from exome sequence data (the NHLBI's Exome Sequencing Project).

3.1 Genotype file format

The format of Genotype file is recoded from PLINK PED file with command:
plink –file DATA –recodeA
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.table(system.file("extdata", "simGeno-chr2.raw", package="FRGEpistasis"),header=TRU
> 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
```

3.2 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

>
```

3.3 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
```

3.4 Gene Annotation file format

This package takes a genome region as a basic unit of interaction analysis instead of a SNP. And this Gene Annotation file is used to set the scope of each region. This file can be selfdefined or derived from the Consensus CDS (CCDS) project if gene as the test unit.
Gene Annotation file contains 4 columns indicate the gene name, chromosome, gene start position and gene end position. Each line represent the region of a gene. The start position is 0-based and end position is 1-based. Thus the length of a gene is equal to pos(end) - pos(start).

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

  Gene_Symbol Chromosome     Start       End
1       gene1          1 159824106 159832447
2       gene2          2  74017030  74063042
3       gene3          2  78365582  78385273
4       gene4          3  88182642  88222051
5       gene5          3 100054649 100063454
6       gene6          3 153071932 153080898
7       gene7          4  40380093  40401075
8       gene8          5  34493060  34503988
```

3.5 genotype files index and map files index

Because the NGS data are large, They are always organized in many files. For example, In real dataset the genotypes on different chromosome are always organized into different files. In order to bring convenience to users and alleviate the burden of the memory, FRGEpistasis can handle a bunch of genotype files. These indices indicate how many and where to read the genotype files and the genetic map files.

genotype files index:

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

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

map files index:

```
> 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
```

4 Implementation

4.1 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.

4.2 Run

```
> library("FRGEpistasis")
> 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
> mapFiles<-read.table(system.file("extdata", "list_map.txt", package="FRGEpistasis"))
> ##read the phenotype file
> phenoInfo <- read.csv(system.file("extdata", "phenotype.csv", package="FRGEpistasis"),header=TRUE)
> ##read the gene annotation file
> gLst<-read.csv(system.file("extdata", "gene.list.csv", package="FRGEpistasis"))
> ##define the extension scope of gene region
> rng=0
> fdr=0.05
> ## output data structure
> out_epi <- data.frame( )
> ##log transformation
> phenoInfo [,2]=log(phenoInfo [,2])
> ##rank transformation
```

```
> #c=0.5
> #phenoInfo[,2]=rankTransPheno(phenoInfo[,2],c)
> # test epistasis with Functional Regression Model
> out_epi = fRGEpistasis(work_dir,phenoInfo,geno_files,mapFiles,gLst,fdr,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 epistasis test inner chromsome 1"
[1] "1 of 1 with other genes both on 1 chromsome!"
[1] "Performing epistasis test outer  1 : 2 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 2 chromsome(2genes)"
[1] "Performing epistasis test outer  1 : 3 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 3 chromsome(3genes)"
[1] "Performing epistasis test outer  1 : 4 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 4 chromsome(1genes)"
[1] "Performing epistasis test outer  1 : 5 chromsomes!"
[1] "1 of 1 on  1 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing epistasis test inner chromsome 2"
[1] "1 of 2 with other genes both on 2 chromsome!"
[1] "2 of 2 with other genes both on 2 chromsome!"
[1] "Performing epistasis 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 epistasis 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 epistasis 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 epistasis test inner chromsome 3"
[1] "1 of 3 with other genes both on 3 chromsome!"
[1] "2 of 3 with other genes both on 3 chromsome!"
[1] "3 of 3 with other genes both on 3 chromsome!"
[1] "Performing epistasis 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 epistasis 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 epistasis test inner chromsome 4"
[1] "1 of 1 with other genes both on 4 chromsome!"
[1] "Performing epistasis test outer  4 : 5 chromsomes!"
[1] "1 of 1 on  4 chromsome with other genes on 5 chromsome(1genes)"
[1] "Performing epistasis test inner chromsome 5"
```

```
[1] "1 of 1 with other genes both on 5 chromsome!"

> ## output the result to physical file
> write.csv(out_epi,"Output_Pvalues_Epistasis_Test.csv ")
> ##if you want to test epistasis with PCA method and pointwise method then
> ##implement the following command. This method is more slow than FRG method.
> #out_pp <- data.frame( )
> #out_pp <- pCAPiontwiseEpistasis(wDir,out_epi,phenoInfo,gnoFiles,mapFiles,gLst,rng)
```

5 Questions and Bug Reports

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