GIGSEAdata: Gene set collection used for the Genotype Imputed Gene Set Enrichment Analysis (GIGSEA)

Shijia Zhu
Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology

June 30, 2018

Contents

Abstract 1
1. Description of gene sets 1
2 Load data of gene sets: 2

Abstract

GIGSEAdata is the gene set collection used for GIGSEA (Genotype Imputed Gene Set Enrichment Analysis), which is a novel SNP enrichment method that uses GWAS-and-eQTL-imputed differential gene expression to interrogate gene set enrichment for the trait-associated SNPs. The gene sets are saved as matrices. Such matrices are largely sparse, so, in order to save space, we used the functions provided by the R package "Matrix" to build the sparse matrices and saved into the GIGSEAdata package.

1. Description of gene sets

GIGSEA is built on the weighted linear regression model, so it permits both discrete-valued and continuous-valued gene sets. In the GIGSEA package, we already included four categories of gene sets: "MSigDB.KEGG.Pathway", "MSigDB.miRNA", "MSigDB.TF", and "TargetScan.miRNA". Here, we added two more categories in the GIGSEAdata package:

1) **discrete-valued gene sets**:

- `org.Hs.eg.GO`: Gene sets that contain genes annotated by the same Gene Ontology (GO) term. For each GO term, we not only incorporate its own gene sets, but also incorporate the gene sets belonging to its offsprings. See the database "org.Hs.eg.GO.db" and "GO.db" in R.

2) **continuous-valued gene sets**:

- `Fantom5.TF`: The human transcript promoter locations were obtained from Fantom5. Based on the promoter locations, the tool MotEvo was used to predict the human transcriptional factor (TF) target sites. The dataset contains 500 Positional Weight Matrices (PWM) and 21964 genes. For each PWM, there is a list of associated human TFs, ordered by percent identity of TFs known to bind sites of the PWM. The list of associations was checked manually. The entire set of PWMs and mapping to associated TFs is available from the SwissRegulon website http://www.swissregulon.unibas.ch.

2 Load data of gene sets:

We first take as an example of the gene set “org.Hs.eg.GO”“, where the row represents the gene, and the column represents the GO term. Each entry takes discrete values of 0 or 1, where 1 represents the gene (row) belongs to the GO term (column), and otherwise, not.

```r
library(GIGSEAdata)
data(org.Hs.eg.GO)
class(org.Hs.eg.GO)
```

```
## [1] "list"
```

```r
names(org.Hs.eg.GO)
```

```
## [1] "net"   "annot"
```

```r
dim(org.Hs.eg.GO$net)
```

```
## Loading required package: Matrix
```

```
## NULL
```

```r
head(colnames(org.Hs.eg.GO$net))
```

```
## [1] "GO:0008150" "GO:0001869" "GO:0002576" "GO:0007264" "GO:0007596"
## [6] "GO:0007597"
```

```r
head(rownames(org.Hs.eg.GO$net))
```

```
## [1] "A1BG"   "A1CF"   "A2M"    "A2ML1"  "A4GALT" "A4GNT"
```

```r
head(org.Hs.eg.GO$annot)
```

```
##         goid ontology
## 1 GO:0008150       BP
## 2 GO:0001869       BP
## 3 GO:0002576       BP
## 4 GO:0007264       BP
## 5 GO:0007596       BP
## 6 GO:0007597       BP
##                                                        term totalGenes
## 1                                                biological      16655
## 2 negative regulation of complement activation, lectin pathway          2
## 3                                    platelet degranulation         87
## 4                 small GTPase mediated signal transduction        922
## 5                                         blood coagulation        566
## 6                      blood coagulation, intrinsic pathway         19
```

```r
head(org.Hs.eg.GO$net[,1:30])
```

```
## 6 x 30 sparse Matrix of class "dgCMatrix"
```

```
##    [[ suppressing 30 column names 'GO:0008150', 'GO:0001869', 'GO:0002576' ... ]]
```

```
##
## A1BG   1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
## A1CF   1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
## A2M    1 1 1 1 1 1 1 1 1 1 1 1 . . . 1 . . . . . . 1 . . . . . . .
## A2ML1  1 . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . .
## A4GALT 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
## A4GNT  1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```