\begin{AbsFilterGSEA}

title: AbsFilterGSEA: Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering
author: Sora Yoon | Ulsan National Institute of Science and Technology | yoonsora@unist.ac.kr
date: March 30, 2016

Introduction

AbsFilterGSEA provides an efficient gene-permuting GSEA methods for small replicate RNA-seq data

Gene-set enrichment analysis (GSEA) has been popularly used for assessing the enrichment of differential signal in a pre-defined gene-set without using a cutoff threshold for differential expression. The significance of enrichment is evaluated through sample- or gene-permutation method. Although the sample-permutation approach is highly recommended due to its good false positive control, ***gene-permuting method is the only choice for small replicate data***. However, such gene-permuting GSEA (or preranked GSEA) generates a lot of false positive gene-sets due to the inter-gene correlation in each gene set. These false positives can be effectively reduced by filtering with the ***one-tailed absolute GSEA*** results. This package provides a function that performs gene-permuting GSEA with or without the absolute filtering. One-tailed absolute GSEA is also provided.

A Quick Example

1. Load an expression matrix.

The GenePermGSEA function accepts gene expression matrix composed of raw read counts. Already normalized counts or microarray data are also acceptable using the option *normalization='AlreadyNormalized'*. Here, the example dataset, *example* contains raw read counts generated from negative biomial distribution using 'rnbinom' function. It contains 10000 genes, and the case and control groups are composed of three replicates, respectively.

```
library(AbsFilterGSEA, quietly = TRUE)
data(example)
head(example)
```

```
##       groupA1 groupA2 groupA3 groupB1 groupB2 groupB3
## Gene1    5642     900     236    1239   20673    2699
## Gene2    2401    1264    1464    2413    2456    1969
## Gene3     611     352     364     548     651     332
## Gene4     140      50     111      97     111      80
## Gene5    2224     915    2129    1930    2720    1757
## Gene6    5987    2959    6094    6700    8630    4974
```

2. Prepare the gene-set file.

Next, a gene-set file should be prepared. It must be **tab-delimited**. For example, the ***.gmt file*** stored on the Molecular Signature Database (mSigDB - http://software.broadinstitute.org/gsea/msigdb) is directly applicable. Here, an example gene-set file **(geneset.txt)** is generated and stored on your local directory. Each gene-set contains 50 non-overlapping genes, and the inter-gene correlation of each gene-set is varied from 5% to 60%. The 41st to 50th gene-sets are set as differentially expressed gene-sets.

```
# Example gene-set generation: It contains 50 gene-sets each having 100 genes.
# Geneset_41 ~ Geneset_50 are differentially expressed and others are not.
# It will stored in your working directory with the file name 'geneset.txt'.
for(Geneset in 1:50)
{
  GenesetName = paste("Geneset", Geneset, sep = "_")
  Genes = paste("Gene", (Geneset*100-99):(Geneset*100), sep="", collapse = '\t')
  Geneset = paste(GenesetName, Genes, sep = '\t')
  write(Geneset, file = "geneset.txt", append = TRUE, ncolumns = 1)
}
```

The example gene-set looks like this... Each line contains gene-set name followed by the member genes which are tab-delimited.

3. Run the gene-permuting GSEA

Now you can run gene-permuting GSEA using **GenePermGSEA** function as shown in below code.

```
# If you want to perform absolute filtering GSEA...
res = GenePermGSEA(countMatrix = example, GeneScoreType = 'moderated_t', idxCase=1:3,
idxControl = 4:6, GenesetFile = './geneset.txt', normalization = 'DESeq',
GSEAtype = 'absFilter', minCount = 3, FDR = 0.05)
res
```

```
##    GenesetName Size       NES Nominal.P.value FDR.Q.value Direction
## 12  Geneset_15   98 -2.543179         0.00000     0.00000      DOWN
## 33  Geneset_43   97  2.215153         0.00000     0.00000        UP
## 35  Geneset_45   98  2.453150         0.00000     0.00000        UP
## 36  Geneset_46   99 -2.708949         0.00000     0.00000      DOWN
## 40  Geneset_50   97 -2.267864         0.00000     0.00000      DOWN
## 32  Geneset_42   98  1.373412         0.03512     0.04234        UP
```

Three GESA modes

When gene scores are chosen, GSEA implements a (weighted) Kolmogorov-Smirnov (K-S) statistic to calculate the enrichment score (ES) of each pre-defined gene set. AbsFilterGSEA provides three modes of gene-permuting GSEA methods: (1) original two-tailed GSEA, (2) absolute one-tailed GSEA and (3) the ordinary GSEA filtered with absolute GSEA results.

1. Original two-tailed GSEA (GSEAtype = 'original')

This is the standard GSEA method introduced by the original GSEA paper[1]. When S is a gene-set of size N_H, and r_i is a gene score of a gene g_i, the enrichment score $ES(S)$ is defined as the maximum deviation of $p_{hit} - p_{miss}$ from zero, that is

$$ES(S) = \begin{cases} \max_i(p_{hit,i} - p_{miss,i}), & \text{if } |max_i(p_{hit,i} - p_{miss,i})| \geq |min_i(p_{hit,i} - p_{miss,i})| \\ min_i(p_{hit,i} - p_{miss,i}), & \text{otherwise} \end{cases}$$

where

$$p_{hit,i} = \sum_{g_j \in S, j \leq i} \frac{|r_j|^q}{N_R}$$

$$p_{miss,i} = \sum_{g_j \in S^c, j \leq i} \frac{1}{(N - N_H)}$$

and

$$N_R = \sum_{g_j \in S} |r_j|^q$$

Features

- It provides the direction of regulation for each gene-set.
- It shows good statistical power.
- However, it suffers from high false positive rate caused by the inter-gene correlation. [2]

2. Absolute one-tailed GSEA (GSEAtype = 'absolute')

It is shown in [3] and the main manuscript that the absolute gene statistic effectively reduces false positives and maintains a good statistical power for gene-permuting GSEA methods.In this approach, the absolute gene scores are used to calculate one-tailed K-S statistics that only consider the positive deviation. Thus, the enrichment score for the absolute one-tailed GSEA is simply defined as

$$ES(S) = max_i(p_{hit,i} - p_{miss,i})$$

Features

- It reduces false positives effectively.
- Its statistical power is rather lower than the original two-tailed GSEA.
- Its overall discriminatory ability (AUC) is improved.

3. Absolute filtering GSEA (GSEAtype = 'absFilter')

To provide a robust GSEA result with the directionality of regulation, we suggest users to use the absolute filtering GSEA method. It filters the result obtained from two-tailed GSEA with that obtained from absolute one-tailed GSEA. In other words, It returns ***gene-set significant in both two-tailed and absolute one-tailed GSEA***.

Option descriptions

There are 14 options in the GenePermGSEA function.

```
head(GenePermGSEA, 4)
```

```
##
## 1 function (countMatrix, GeneScoreType, idxCase, idxControl, GenesetFile,
## 2     normalization, minGenesetSize = 10, maxGenesetSize = 300,
## 3     q = 1, nPerm = 1000, GSEAtype = "absFilter", FDR = 0.05,
## 4     FDRfilter = 0.05, minCount = 3)
```

1. countMatrix: Input gene expression matrix. Both microarray and (normalized) RNA-seq count data can be used.

2. GeneScoreType: There are four gene scores provided as listed below.

- moderated_t : This is moderated t-statistics that uses shrinkage variation. Moderated t-statistics is defined as

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{\nu_g}}$$

where $\hat{\beta}_g$ is the difference in means between two groups, and $\tilde{s}_g^2$ is the shrinkage variation defined as

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

where $\frac{d_0}{d_0+d_g}$ is weight coefficient associated with all genes and $\frac{d_g}{d_0+d_g}$ is associated with gene g. Here, s_0^2 is overall estimate vatiation evaluated from Baysian approach and s_g^2 is per-gene deviation variation. This is good statistic for small replicate data.

- SNR : This is signal-to-noise ratio defined as

$$SNR_g = \frac{\mu_{g,1} - \mu_{g,2}}{s_{g,1} + s_{g,2}}$$

where $\mu_{g,i}$ is the mean of expression level of gene g in group i, and $s_{g,i}$ is the standard deviation of expression level of gene g in group i.

- RANKSUM : This is two-sample Wilcoxon ranksum statistic introduced by Li and Tibshirani. For a gene g, the ranksum test statistic (T_g) is calculated as,

$$T_g = \sum_{j \in C_1} R_{gj} - \frac{n_1 \cdot (n+1)}{2}$$

where R_{gj} is the rank of the j^{th} samples' expression level among all counts of gene a gene g, C_1 is a set of sample indices in the first phehotypic group, n_1 is the sample size of C_1 and n is teh total sample size.

- FC : Log fold change. For a gene g, the log fold change (LFC_g) is calculated as

$$LFC_g = log_2 \frac{\mu_g^1}{\mu_g^2}$$

where μ_g^i is the mean expression value of g in group i.

3. idxCase: Indexes of case samples. In the case of example data, idxCase = c(1,2,3).

4. idxControl: Indexes of control samples. In the case of example data, idxControl = c(4,5,6).

5. GenesetFile: The path for the gene-set file. ex) "C:/Users/A/Documents/geneset.txt".

6. normalization: If your data is RNA-seq raw count data, type ***normalization='DESeq'***, then your data will be normalized by DESeq method [4], and pseudocount (5% quantile of positive normalized counts) will be added to the data to stabilize the log fold change. Otherwise, if your data is already normalized, just type ***normalization='AlreadyNormalized'***. Also, pseudocount is also added in the same way in this case.

7. minGenesetSize: The minimum gene-set size used in the analysis. Defalut = 10.

8. maxGenesetSize: The maximum gene-set size used in the analysis. Defalut = 300.

9. q: Weighting exponent of gene score used for the enrichment score calculation.

$$p_{hit,i} = \sum_{g_j \in S, j \leq i} \frac{|r_j|^q}{N_R}$$

'q' value in this equation. q must be equal or greater than zero. Defaule = 1. If you set q=0 for the pre-ranked test.

10. nPerm: The number of gene-label permutation. Default = 1000.

11. GSEAtype: The type of GSEA. "original" for two-tailed GSEA, "absolute" for absolute one-tailed GSEA, and "absFilter" for absolute filtering GSEA.

12. FDR: FDR cutoff for two-tailed or absolute one-tailed GSEA result (used when GSEAtype="original" or "absolute"). Default=0.05.

13. FDRfilter: FDR cutoff for absolute one-tailed GSEA result for absolute filtering (used when GSEAtype="absFilter". In this case, FDR cutoff for two-tailed GSEA is applied to the 'FDR' option).

14. minCount: Minimum median count of a gene to be included in the analysis. It is used for gene-filtering to avoid genes having small read counts. Default = 0

Reference

[1] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(43):15545-15550.

[2] Wu D, Smyth GK: Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research 2012, 40(17).

[3] Nam, D. Effect of the absolute statistic on gene-sampling gene-set analysis methods. Statistical methods in medical research 2015

[4] Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biology 2010, 11(10).

\end{AbsFilterGSEA}