TSEA: Tissue-Specific Enrichment Analysis to decode tissue heterogeneity

Guangsheng Pei[1], Yulin Dai[1], Zhongming Zhao[1, 2, 3,*] and Peilin Jia[1,*]

[1]Center for Precision Health, School of Biomedical Informatics, [2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

August 21, 2018

1. Introduction

Genome-wide association studies (GWAS) and next-generation sequencing technologies have identified hundreds of thousands of disease-associated variants and genes. Interpretation of these variants could be greatly enhanced in tissue-specific systems. However, there are many diseases or traits where the causal tissues or cell types remain unknown. In many studies, tissue transcriptome data are generated for research, which include both genes that are ubiquitously expressed (e.g., housekeeping genes) and genes that are specifically expressed in a range of tissues. This documentation introduces Tissue-Specific Enrichment Analysis (TSEA), an R package to identify the most relevant tissues for candidate genes or for gene expression profiles. TSEA builds on two pre-processed reference panels. We implemented different statistic tests for different forms of query data. We demonstrate TSEA using multi-trait GWAS data and cancer RNA-sequencing data.

2. Usage

2.1 Installation TSEA

Requirements

TSEA relies on *R* (>= 3.4), *pheatmap* (>= 1.0.10), *RColorBrewer* (>= 1.1)

The *pheatmap* relies on CRAN. Please follow their installation instruction.

```
> install.packages("pheatmap")
```

To download the codes, please do:

```
git clone https://github.com/bsml320/TSEA.git
cd TSEA
### Then open the R:
> install.packages("TSEA_1.0.tar.gz")
```

TSEA loading

```
### Load the TSEA package and dependent library
> library(TSEA)
> library(pheatmap)
```

2.2 Built-in data

TSEA requires two reference panels to conduct the enrichment test: one from GTEx and the other from ENCODE. For GTEx, a matrix including the summary statistics for each tissue is also needed. All datasets have been included in the package. After installation of the package, one can load the data using the following commands:

```
### Load the t-statistic matrix for the GTEx panel
> load("data/GTEx_t_score.rda")
```

```
### Load the z-score matrix for the ENCODE panel
> load("data/ENCODE_z_score.rda")
```

Then "`GTEx_t_score`" and "`ENCODE_z_score`" will be loaded to R enviroment.

2.3 Input data

TSEA deals with two types of enrichment analysis for different forms of query data. For convenience, we provide two TSEA functions for query gene lists (single sample and multiple samples), and another function for RNA-Seq expression profiles tissue-specific enrichment analysis.

2.3.1 TSEA for candidate genes

When the query data are lists of genes, the Fisher's Exact Test is implemented. The function is

`tsea.analysis()`. The input is a vector of gene symbols. Here we used disease-associated genes identified from GWAS summary statistics as an example. The gene symbols can be found here:

```
### Load gene symbol from TSEA package.
> load("data/GWAS_gene.rda")
> query.genes = GWAS_gene

### Or you can read your own gene symbol list from a text file
> dat = read.table("Gene_list.txt", head = F)
> query.genes = as.character(dat[,1])

### Tissue-specific analysis for query gene list.
> tsea_t = tsea.analysis(query.genes, GTEx_t_score, ratio = 0.05,
p.adjust.method = "bonferroni")
```

Here, the parameter *`ratio`* is to define tissue-specific genes and provides the first way of categorizing genes. The second way of categorizing genes is based on the query genes. The two ways of category form a two by two table, which is used in the Fisher's Exact Test (FET). P-values from FET will be stored in *`tsea_t`*. To explore the results, we provide a plot function and a summary function.

```
### Check tissue-specific enrichment analysis result.
> head(tsea_t)
                                  query
Adipose - Subcutaneous       1.00000000
Adipose - Visceral (Omentum) 0.01095850
Adrenal Gland                1.00000000
Artery - Aorta               0.21208614
Artery - Coronary            0.01095850
Artery - Tibial              0.00257813
### TSEA result plot and summary
> tsea.plot(tsea_t, threshold = 0.05)
> tsea.summary(tsea_t)
```

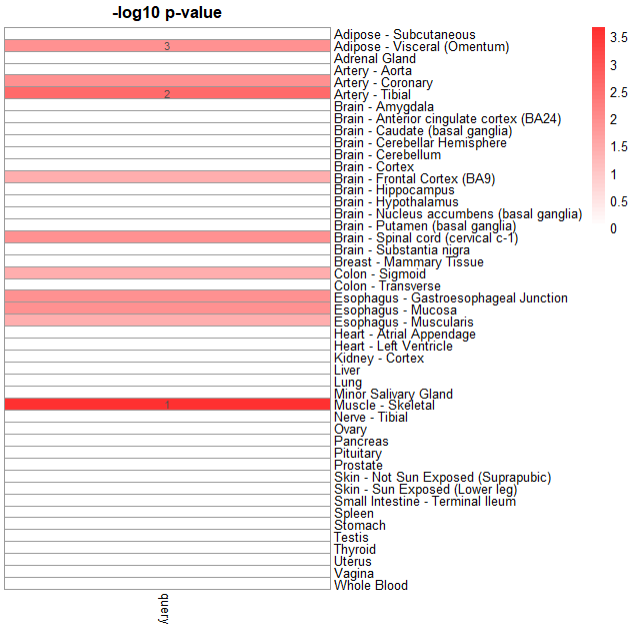

Fig. 1 Tissue-specific enrichment result for query gene list. Adjusted p-values from Fisher's Exact Text in the $\log_{10}$ form for each tissue are used for the heatmap plot. The top 3 most significantly associated tissues were labeled with "1", "2" and "3" in their corresponding cells.

2.3.2 TSEA for multiple gene lists

In most condition, you might want to analysis multiple samples together, then you can upload a 0~1 table. In the table, gene labeled with 1 indicated significant associate within a sample, while 0 indicated not in a given sample. You can check the format of example data.

```
### Load multiple gene symbol from TSEA package.
> load("data/GWAS_gene_multiple.rda")
> query.gene.list = GWAS_gene_multiple

### Or you can read your own multiple gene symbol from a text file.
> dat = read.table("Gene_list_multiple.txt", head = T, row.names = 1)
> query.gene.list = dat

### To keep result reliable, please keep at least 20 genes for each
samples. You can check the total genes number for each sample:
> colSums(query.gene.list)
```

Then, we can make tissue specific enrichment analysis for multiple samples by `tsea.analysis.multiple()` and plot the result by `tsea.plot()` as showed in Fig. 2. You can

summary the top 3 most associated tissues by `tsea.summary()` function and save your result in to a text-format spreadsheet, simply type:

```
### Tissue-specific enrichment analysis in GTEx panel
> tsea_t_multi = tsea.analysis.multiple(query.gene.list,
          GTEx_t_score, ratio = 0.05, p.adjust.method = "BH")
### Save tissue-specific enrichment analysis result
> write.csv(tsea_t_multi, "GWAS_multi_TSEA_in_GTEx_panel.csv")
### Save the tissue-specific enrichment analysis plot
> pdf ("GWAS_multi_TSEA_in_GTEx_panel.pdf",6,6,onefile = FALSE)
> tsea.plot(tsea_t_multi, threshold = 0.05)
> dev.off()
### Save summary result in to a spreadsheet
> tsea_t_multi_summary = tsea.summary(tsea_t_multi)
> write.csv(tsea_t_multi_summary,"GWAS_multi_summary_GTEx_panel.csv")
```

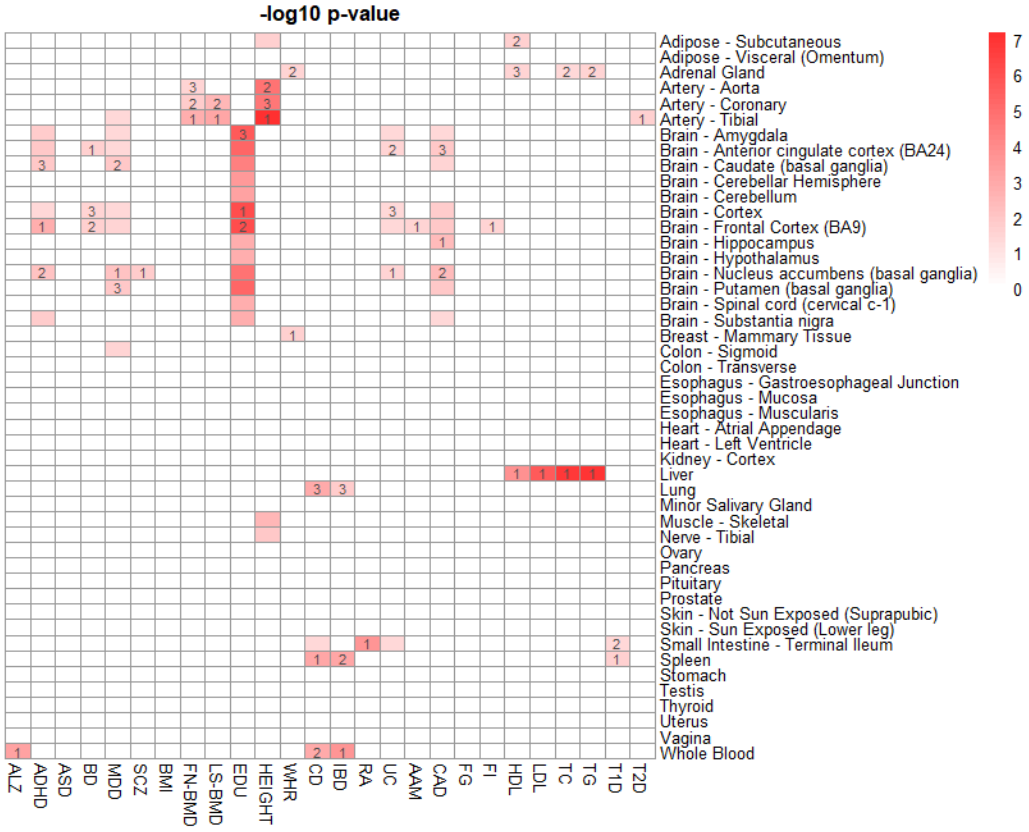

Fig. 2. Tissue-specific enrichment analysis for multiple samples.

2.3.3 TSEA for RNA-Seq profiles

For query data as RNA-sequencing profiles, RPKM values are required in the format of matrix, with genes on rows and samples on columns. As an example, we use the ENCODE panel data as query and GTEx panel data as reference to demonstrate:

```
### Load ENCODE query data
> load("data/query_ENCODE.rda")
> query.matrix = query_ENCODE

### Load correction variable
> load("data/correction_factor.rda")
```

As RNA-Seq samples are often heterogeneous, before in-depth analysis, it is necessary to decode tissue heterogeneity to avoid samples with confounding effects. However, the raw discrete RPKM value should be normalized to continuous variable meet the normal distribution before t-test. We provided two normalization approaches: `"z-score"` and `"abundance"` in function `tsea.expression.normalization()`:

(1)`z-score` normalization will calculate a z-score for the query sample for each tissue in the reference panel as below: $e_i = (e_0 - \mu_t))/sd_t$, where μ_t and sd_t were the mean and SD of tissue t.

(2)`abundance` normalization will provide an abundance correction approach for the query sample for each tissue in the reference panel as below: $e_i = \log_2(e_0 + 1)/(\log_2(u_t + 1) + 1)$.

We have the preloaded the test RPKM variable in `query.matrix` and correction variable in `correction_factor`, we take `"abundance"` normalization approach as an example, simply type:

```
### RNA-Seq profiles scale by abundance normalization
> query_mat_abundance_nor =
tsea.expression.normalization(query.matrix, correction_factor,
normalization = "abundance")
```

After normalization, we submit it for `tsea.expression.decode()`.

```
> tseaed_in_GTEx = tsea.expression.decode(query_mat_abundance_nor,
            GTEx_t_score, ratio = 0.05, p.adjust.method = "BH")
```

Then, the tissue specific enrichment analysis for query RNA-Seq is finish. After tissue specific enrichment decode analysis, one-side *t*-test results between query RNA-Seq sample tissue specific genes (top 5%) versus remains genes (95%) is stored in variable `tseaed_in_GTEx`. Further analysis for top 3 most associated tissues is similar to previous analysis, and results were plotted in Fig. 3.

```
> tsea.plot(tseaed_in_GTEx, threshold = 0.05)
> tseaed_in_GTEx_summary = tsea.summary(tseaed_in_GTEx)
> write.csv(tseaed_in_GTEx_summary, "RNAseq_in_GTEx_panel.csv")
```

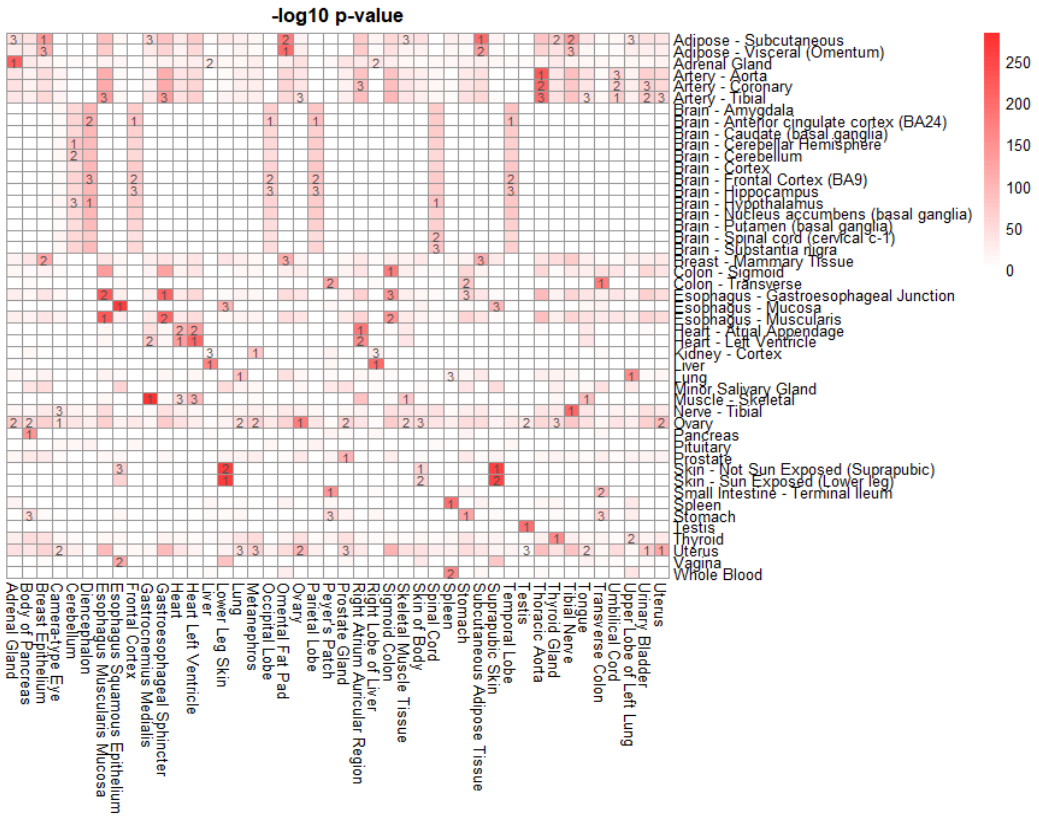

Fig 3. Tissue-specific enrichment analysis for RNA-Seq expression profiles in ENCODE panel

To prove the robustness of our proposed pipeline, user can validate the two reference panels through self-validation. Simply, load GTEx example RNA-Seq profiles and perform tissue-specific enrichment analysis in ENCODE panel.

```
### Load GTEx query data
> load("data/query_GTEx.rda")
> query_matrix = query_GTEx
```

Usually, in GTEx panel, we suggest take abundance normalization approach; while in ENCODE panel, we suggest take z-score normalization approach.

```
### RNA expression profiles z-score normalization
> query_mat_zscore_nor = tsea.expression.normalization(query_matrix,
          GTEx_ave_sd, normalization = "z-score")
### RNA expression profiles TSEA in ENCODE panel
> tseaed_in_ENCODE = tsea.expression.decode(query_mat_zscore_nor,
          ENCODE_z_score, ratio = 0.05, p.adjust.method = "BH")
```

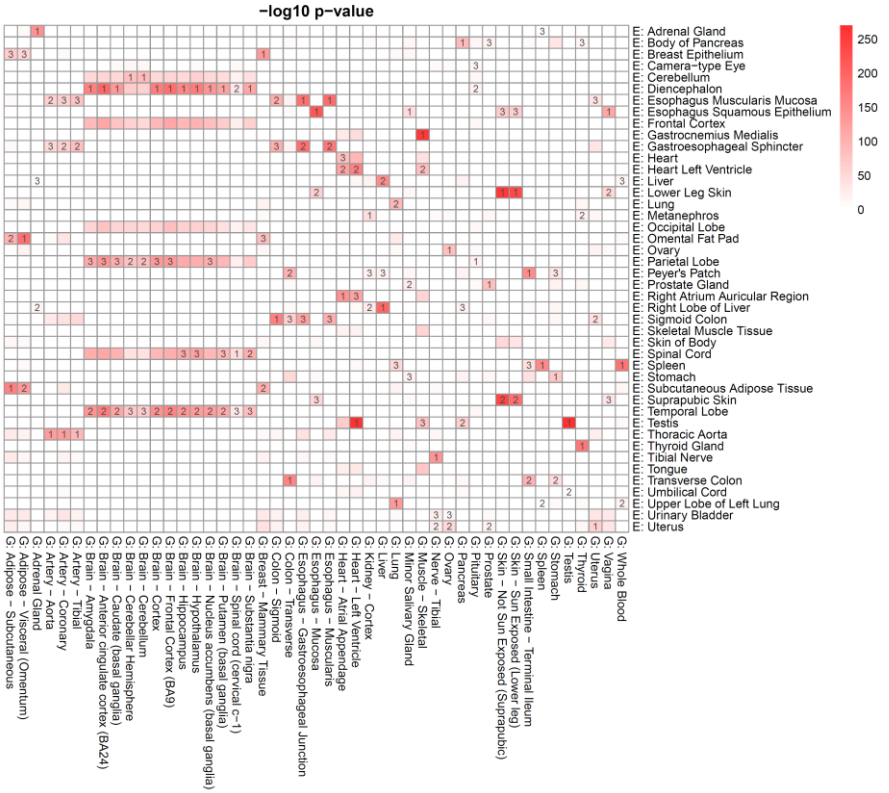

Fig. 4. . Tissue-specific enrichment analysis for RNA-Seq expression profiles in GTEx panel

Further analysis for top 3 most associated tissues is similar to previous analysis, and results were plotted in Fig. 4. The reader is encouraged to open and view the file in a spreadsheet software, or inspect it directly within R using the command `fix(tseaed_in_ENCODE)`. In addition, sometime, you might

want to edit some parameters for your own data, e.g., you can change the `GTEx_t_score` to `ENCODE_z_score` for ENCODE tissue specific enrichment analysis, you can also change the tissue specific genes `ratio` from `0.05` to `0.2`, or change the `p.adjust.method` to "`bonferroni`".

In addition, we provide `tsea.plot()` to facilitate interpretation and visualization of the results, as showed in Fig. 3 and Fig. 4. Further analysis for top 3 most associated tissues is similar to previous analysis:

```
> tsea.plot(tseaed_in_ENCODE, threshold = 0.05)
> tseaed_in_ENCODE_summary = tsea.summary(tseaed_in_ENCODE)
> write.csv(tseaed_in_ENCODE_summary, "RNAseq_in_ENCODE_panel.csv")
```

Citation

Pei G., Dai Y., Zhao Z, Jia P. (2018) Tissue-Specific Enrichment Analysis (TSEA) to decode tissue heterogeneity. Bioinformatics, in submission.

References

Cavalli, F.M., *et al.* (2011) SpeCond: a method to detect condition-specific gene expression, *Genome Biol*, **12**, 2011-2012.

Eisenberg, E. and Levanon, E.Y. (2013) Human housekeeping genes, revisited, *Trends in genetics : TIG*, **29**, 569-574.

Kim, P., *et al.* (2018) TissGDB: tissue-specific gene database in cancer, *Nucleic acids research*, **46**, D1031-D1038.

Lamparter, D., *et al.* (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics, *PLoS computational biology*, **12**, e1004714.