deTS: tissue-specific enrichment analysis to decode tissue specificity

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

February 05, 2019

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 deTS: Tissue-Specific Enrichment Analysis, an R package to identify the most relevant tissues for candidate genes or for gene expression profiles. deTS builds on two pre-processed reference panels. We implemented different statistic tests for different forms of query data. We demonstrate deTS using multi-trait GWAS data and cancer RNA-sequencing data.

2. Usage

2.1 Installation

deTS is available at: https://github.com/bsml320/deTS. deTS requires the package *pheatmap*. Before installing deTS, please install R package pheatmap.

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

2.2 Built-in data

deTS 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:

```
### the t-statistic matrix for the GTEx panel
> data(GTEx_t_score)
```

```
### the z-score matrix for the ENCODE panel
> data(ENCODE_z_score)
```

```
### The summary statistics for GTEx tissues
> data(correction_factor)
```

2.3 Input data

deTS deals with two types of enrichment analysis for different forms of query data.

2.3.1 deTS 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 deTS package.
> data(GWAS_gene)
> query.genes = GWAS_gene

# Or you can read gene symbol 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, 0.05)
> tsea.summary(tsea_t)
Top1  tissue1_p-value   Top2  tissue2_p-value   Top3  tissue3_p-value
query "Muscle - Skeletal"     "0.000213565633284591"  "Artery - Tibial"
"0.00257812976565265"   "Adipose - Visceral (Omentum)"
      "0.0109584998307628"
```

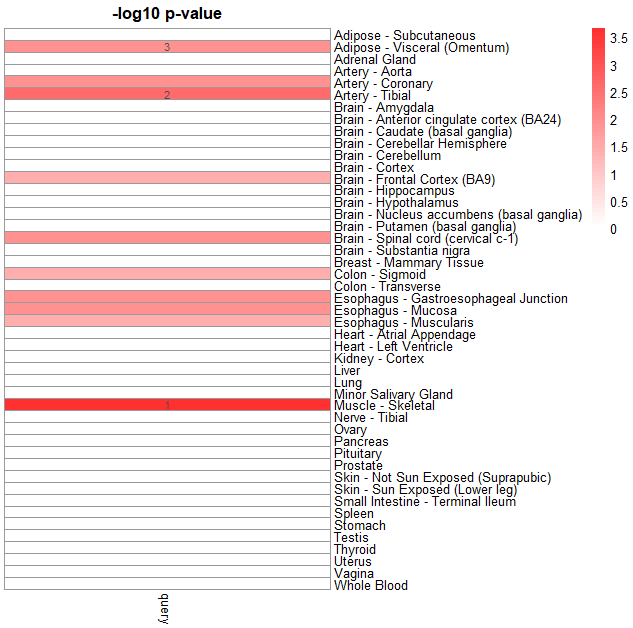

Fig. 1 Tissue-specific enrichment result for query gene list. Adjusted p-values from Fisher's Exact Text in the $\log_{10}$ form for each tissue are used for the heatmap plot. The 3 most significantly associated tissues were labeled with "1", "2" and "3" in their corresponding cells.

2.3.2 deTS 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 deTS package.
> data(GWAS_gene_multiple)
> query.gene.list = GWAS_gene_multiple
# Or you can read multiple gene symbol from a text file.
> dat = read.table("Gene_list_multiple.txt", head = T, row.names = 1)
> query.gene.list = dat
> query.gene.list[1:5,1:5]
         ALZ ADHD ASD BD MDD
A1BG       0    0   0  0   0
A1BG-AS1   0    0   0  0   0
A1CF       0    0   0  0   0
A2M        0    0   0  0   0
A2M-AS1    0    0   0  0   0
> colSums(query.gene.list)
   ALZ   ADHD    ASD     BD    MDD    SCZ    BMI FN-BMD LS-BMD
   121     20     20     25     20     88    233    109    107
   EDU HEIGHT    WHR     CD    IBD     RA     UC    AAM    CAD
   487   3311     98    336    480    213    275    270     71
    FG     FI    HDL    LDL     TC     TG   T1D    T2D
    69     26    368    305    329    223    329     43
```

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, 0.05, p.adjust.method = "BH")
> #write.csv(tsea_t_multi,"GWAS_multi_TSEA_in_GTEx_panel.csv")
```

```
# Heatmap Plot for TSEA result
> #pdf ("GWAS_multi_TSEA_in_GTEx_panel.pdf",6,6,onefile = FALSE)
> tsea.plot(tsea_t_multi, 0.05)
> #dev.off()
# Save your 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")
```

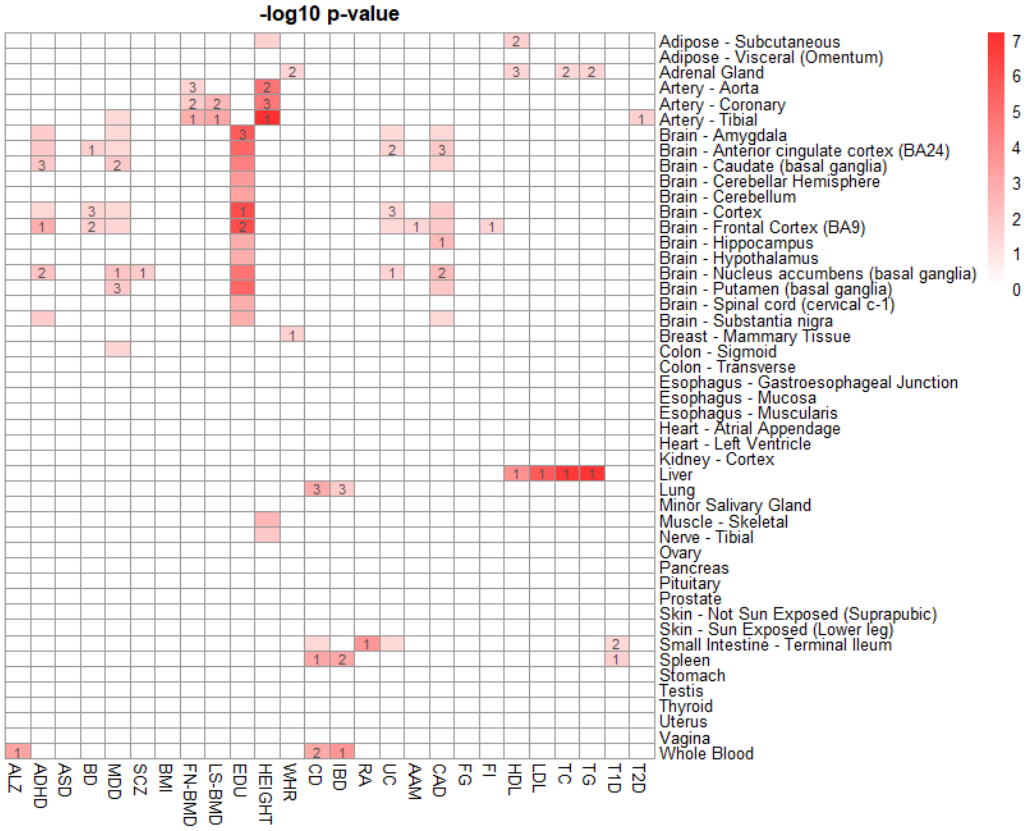

Fig. 2. Tissue-specific enrichment analysis for multiple samples.

2.3.3 deTS 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
```

```
> data(query_ENCODE)
> query.matrix = query_ENCODE
> query.matrix[1:5,1:4]
         Adrenal Gland Body of Pancreas Breast Epithelium Camera-type Eye
TSPAN6       11.639167           5.3900         11.038333         24.6475
TNMD          0.010000           0.1475          2.243333         12.4325
DPM1         18.819167           9.8125         14.215000         24.0250
SCYL3         3.812500           2.5925          5.626667         10.4975
C1orf112      1.643333           2.3075          2.858333         14.6125
```

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:

```
# RNA-Seq profiles scale by abundance normalization
> query_mat_abundance_nor =
tsea.expression.normalization(query.matrix, correction_factor,
normalization = "abundance")
> query_mat_abundance_nor[1:5,1:5]
         Adrenal Gland Body of Pancreas Breast Epithelium Camera-type Eye Cerebellum
C1orf112     0.7427561        0.9140352         1.0317423       2.0998577  1.6212056
FGR          0.4884476        0.2615171         0.5977016       0.2500974  0.1617040
CFH          0.8805367        0.5989594         0.8540355       0.8547878  0.8363093
FUCA2        1.0366443        0.9182102         0.7530246       0.7801565  0.6394422
NFYA         0.6386405        0.5508964         0.8167324       1.1198416  0.9170538
```

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

```
> tseaed_in_GTEx = tsea.expression.decode(query_mat_abundance_nor,
            GTEx_t_score, 0.05, p.adjust.method = "BH")
> tseaed_in_GTEx[1:5,1:3]
                              Adrenal Gland Body of Pancreas Breast Epithelium
Adipose - Subcutaneous         7.093272e-49     4.636686e-44     5.779239e-142
Adipose - Visceral (Omentum)   2.199051e-33     9.641733e-32     1.313532e-112
Adrenal Gland                 9.925492e-220     3.404965e-33      1.476023e-25
Artery - Aorta                 2.749910e-39     1.075081e-21      5.919623e-47
Artery - Coronary              6.702724e-42     4.048237e-22      6.188863e-44
```

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, 0.05)
> tseaed_in_GTEx_summary = tsea.summary(tseaed_in_GTEx)
> write.csv(tseaed_in_GTEx_summary,"RNAseq_summary_in_GTEx_panel.csv")
```

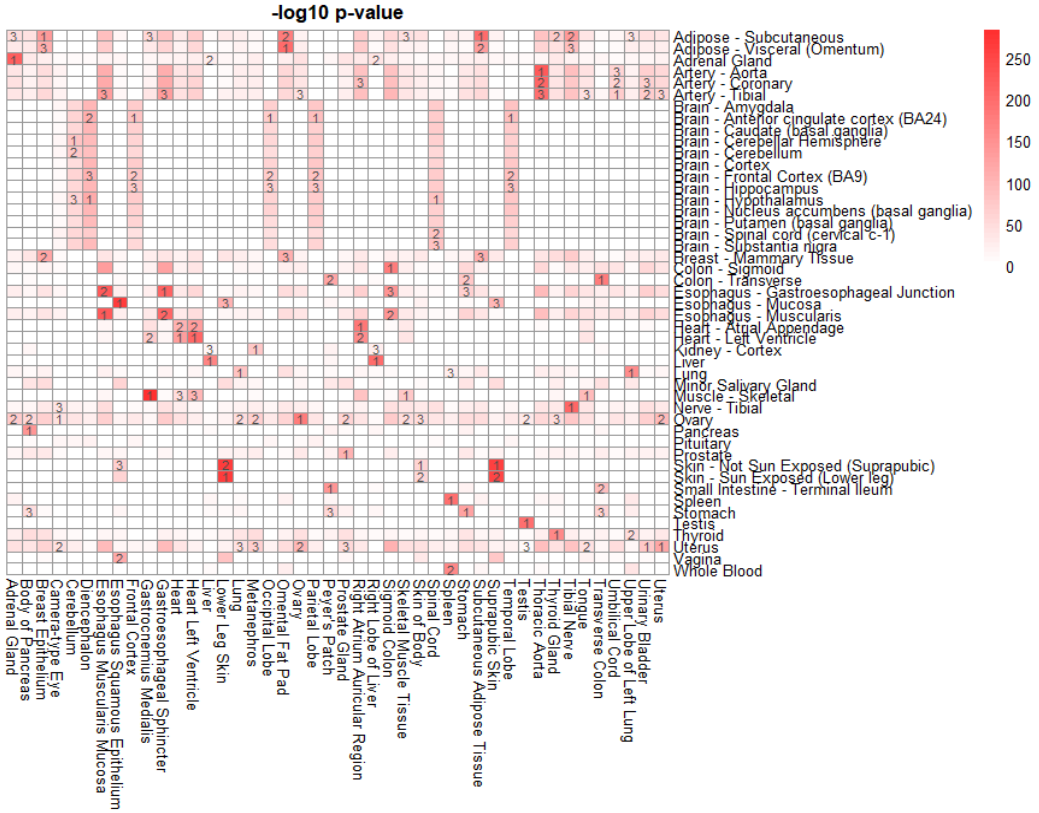

Fig 3. Tissue-specific enrichment analysis for RNA-seq expression profiles

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
> data(query_GTEx)
> 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, 0.05, p.adjust.method = "BH")
```

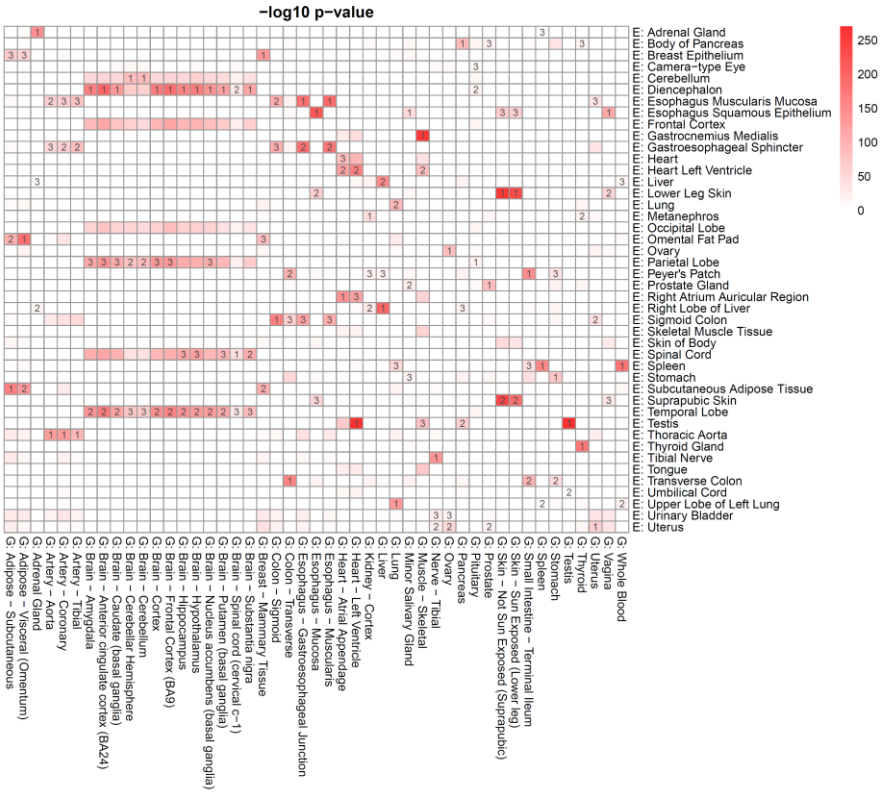

Fig. 4. Enrichment test of ENCODE tissues using GTEx as the reference panel

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 `threshold` 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.

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

Citation

Pei G., Dai Y., Zhao Z, Jia P. (2019) deTS: tissue-specific enrichment analysis to decode tissue specificity. 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.