TissueEnrich: A tool to calculate tissue-specific gene enrichment

Ashish Jain, Geetu Tuteja
Bioinformatics and Computational Biology
Genetics, Development, and Cell Biology
Iowa State University, Ames, Iowa

08/04/2018

TissueEnrich

The `TissueEnrich` package is used to calculate enrichment of tissue-specific genes in a set of input genes. For example, the user can input the most highly expressed genes from RNA-Seq data, or gene co-expression modules to determine which tissue-specific genes are enriched in those datasets. Tissue-specific genes were defined by processing RNA-Seq data from the Human Protein Atlas (HPA) (Uhlén et al. 2015), GTEx (Ardlie et al. 2015), and mouse ENCODE (Shen et al. 2012) using the algorithm from the HPA (Uhlén et al. 2015). The hypergeometric test is being used to determine if the tissue-specific genes are enriched among the input genes. Along with tissue-specific gene enrichment, the `TissueEnrich` package can also be used to define tissue-specific genes from expression datasets provided by the user, which can then be used to calculate tissue-specific gene enrichments. `TissueEnrich` has the following three functions.

- `teEnrichment` : Given a gene list as input, this function calculates the tissue-specific gene enrichment using tissue-specific genes from either human or mouse RNA-Seq datasets.
- `teGeneRetrieval` : Given gene expression data across tissues, this function defines tissue-specific genes by using the algorithm from the HPA.
- `teEnrichmentCustom` : Given a gene list and tissue-specific genes from teGeneRetrieval as input, this function calculates the tissue-specific gene enrichment.

How to get help for TissueEnrich

Please post all the questions or queries related to TissueEnrich package on the **Bioconductor support website**. This will help us to build an information repository which can be used by other users.

https://support.bioconductor.org (https://support.bioconductor.org)

Please **do not** email your questions directly to the package authors.

`teEnrichment`: Tissue-specific gene enrichment using human or mouse genes

The `teEnrichment` function is used to calculate the enrichment of tissue-specific genes in an input gene set. It uses tissue-specific genes defined by processing RNA-Seq datasets from human and mouse. The user must specify the organism using the `organism` ("Homo Sapiens" (default) or "Mus Musculus") parameter in the input GeneSet object. More details about the RNA-Seq datasets and tissue-specific genes are discussed in the next sections.

RNA-Seq datasets

`TissueEnrich` defines tissue-specific genes using RNA-Seq data from the HPA, GTEx, and mouse ENCODE. In order to make the tissue-specific gene calculations more robust, we only used tissues that had ≥2 biological replicates. The datasets used in the tool are:

- **HPA Dataset:** RNA-Seq data across 35 human tissues (Uhlén et al. 2015).
- **GTEx Dataset:** RNA-Seq data across 29 human tissues (Ardlie et al. 2015).
- **Mouse ENCODE Dataset:** RNA-Seq data across 17 mouse tissues (Shen et al. 2012).

When using `teEnrichment`, the user can specify the RNA-Seq dataset (`rnaSeqDataset`) to be used for the tissue-specific gene enrichment analysis.

- 1 for "Human Protein Atlas" (default)
- 2 for "GTEx"
- 3 for "Mouse ENCODE"

Defining Tissue-specific Genes

Tissue-specific genes are defined using the algorithm from the HPA (Uhlén et al. 2015), and can be grouped as follows:

- **Tissue Enriched:** Genes with an expression level greater than 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to all other tissues.
- **Group Enriched:** Genes with an expression level greater than 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a group of 2-7 tissues compared to all other tissues, and that are not considered Tissue Enriched.
- **Tissue Enhanced:** Genes with an expression level greater than 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to the average levels in all other tissues, and that are not considered Tissue Enriched or Group Enriched.

In `teEnrichment` , the user can specify the type of tissue-specific genes (`tissueSpecificGeneType`) to be used for the tissue-specific gene enrichment analysis.

- 1 for “All” (default)
- 2 for “Tissue-Enriched”
- 3 for “Tissue-Enhanced”
- 4 for “Group-Enriched”

Hypergeometric test

The hypergeometric test is used to calculate tissue-specific gene enrichment. The p-value is calculated as:

$$P(X > k) = \sum_{i=k+1}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$$

Where, N is the total number of genes, K is the total number of tissue-specific genes for a tissue, n is the number of genes in the input gene set, k is the number of tissue-specific genes in the input gene set. The p-values can be corrected for multiple hypothesis testing using the Bonferroni Hochberg correction by setting `multiHypoCorrection = TRUE` (It is `TRUE` by default).

Example: Tissue-specific gene enrichment

This example uses trophectoderm (TE) specific genes identified from single cell RNA-Seq analyses, performed on human blastocysts on days 5, 6, and 7 of preimplantation development (Petropoulos et al. 2016). The single cells are assigned to either the inner cell mass (epiblast plus emerging extraembryonic endoderm) or the TE using PCA. After that, a list of 100 TE-specific genes was generated using differential gene expression analysis (Petropoulos et al. 2016). We used those 100 genes as the input gene set and carried out tissue-specific gene enrichment using the tissue-specific genes defined by the HPA dataset.

Note: The input gene set can either contain Ensembl Ids (`ENSEMBLIdentifier()`) or Gene Symbols (`SymbolIdentifier()`) (specify this using the `geneIdType` parameter in the input GeneSet object).

```
library(TissueEnrich)
genes<-system.file("extdata", "inputGenes.txt", package = "TissueEnrich")
inputGenes<-scan(genes,character())
gs<-GeneSet(geneIds=inputGenes,organism="Homo Sapiens",geneIdType=SymbolIdentifier())
output<-teEnrichment(inputGenes = gs)
```

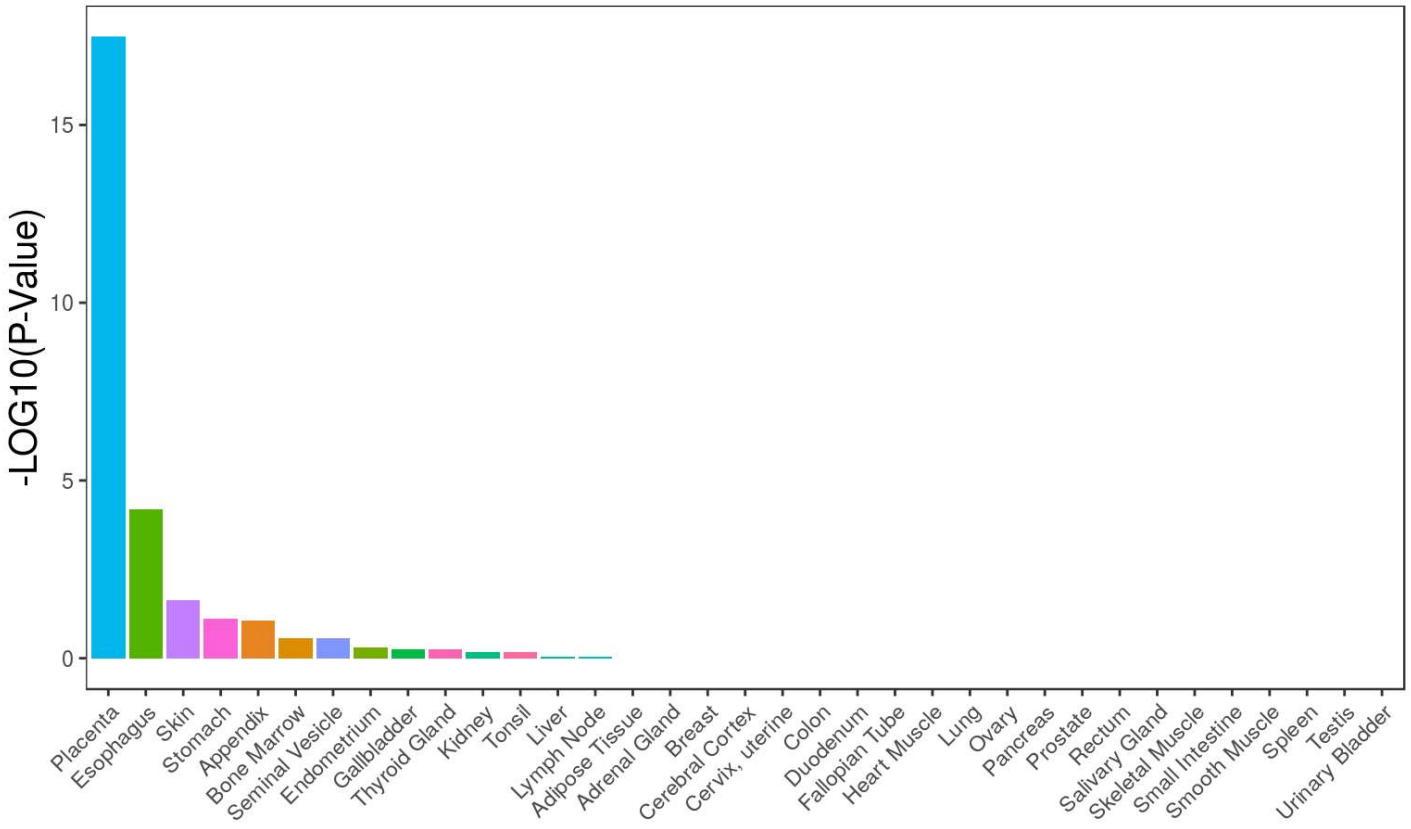
The `output` is a list object containing the enrichment results. These results are explained in the next section.

Exploring tissue-specific gene enrichment results

Tissue-specific gene enrichment bar chart using `ggplot2`

The first object in the `output` list is a `SummarizedExperiment` object containing the $-Log_{10}(P-Value)$, corresponding to the tissue-specific gene enrichment, along with the number of tissue-specific genes in the input gene set. This object can be used to visualize tissue-specific gene enrichment in the form of a bar chart.

```
seEnrichmentOutput<-output[[1]]
enrichmentOutput<-setNames(data.frame(assay(seEnrichmentOutput),row.names = rowData(seEnrichment
  Output)[,1]), colData(seEnrichmentOutput)[,1])
enrichmentOutput$Tissue<-row.names(enrichmentOutput)
ggplot(enrichmentOutput,aes(x=reorder(Tissue,-Log10PValue),y=Log10PValue,label = Tissue.Specifi
  c.Genes,fill = Tissue))+
      geom_bar(stat = 'identity')+
      labs(x='', y = '-LOG10(P-Value)')+
      theme_bw()+
      theme(legend.position="none")+
      theme(plot.title = element_text(hjust = 0.5,size = 20),axis.title = element_text(size=15))
  +
      theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),panel.grid.major= eleme
  nt_blank(),panel.grid.minor = element_blank())
```



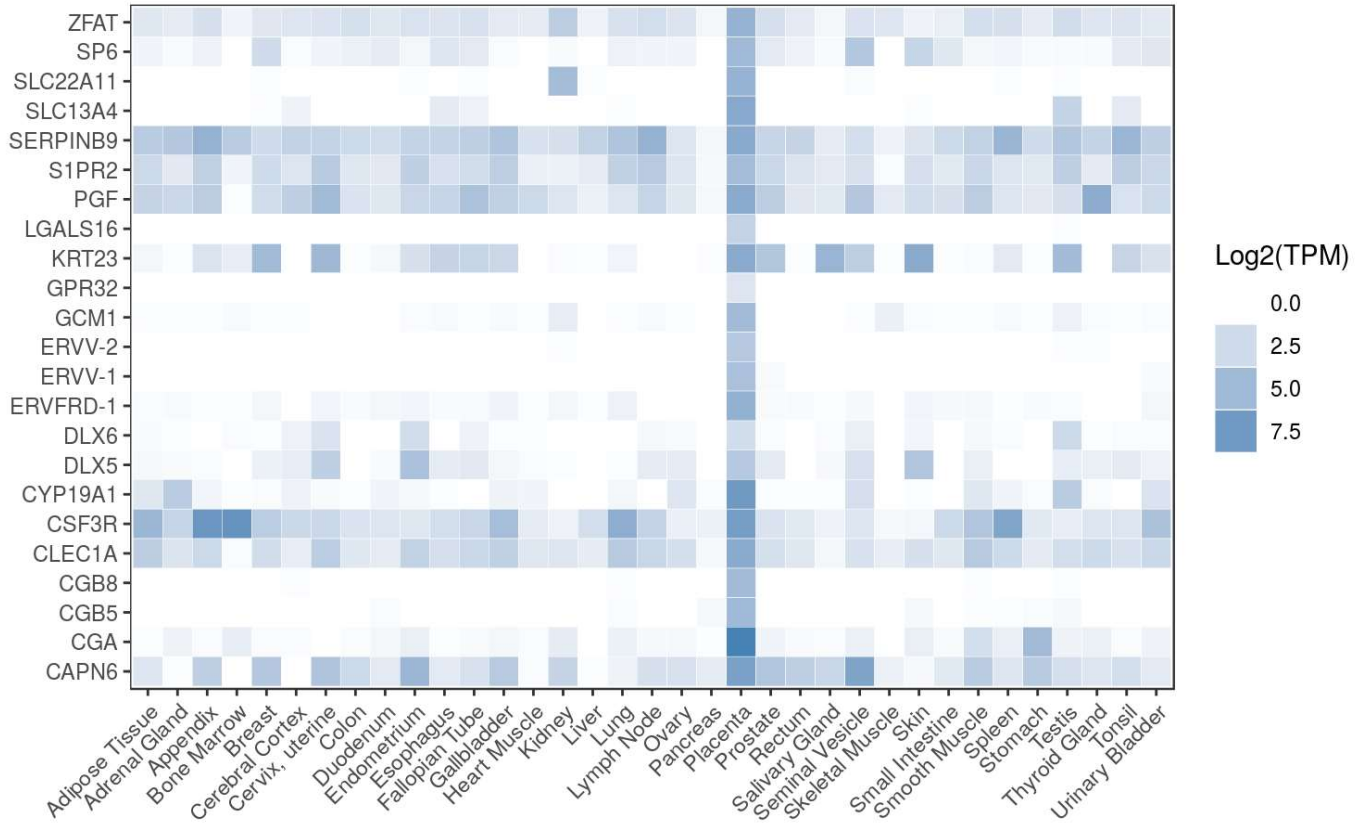
In the plot above, the x-axis shows each of the tissues, and the y-axis represents the tissue-specific gene enrichment ($-Log_{10}(P-Value)$) values. As expected, the 100 TE-specific genes show enrichment for placenta specific genes.

Heatmap to show expression profiles of tissue-specific genes using `ggplot2`

The second object in the `output` is a list containing the expression values of the tissue-specific genes identified from the input gene set. The expression values can be visualized in the form of a heatmap. For example, the code below generates a heatmap showing the expression of the placenta specific genes across all the tissues.

```
library(tidyr)
seExp<-output[[2]][["Placenta"]]
exp<-setNames(data.frame(assay(seExp), row.names = rowData(seExp)[,1]), colData(seExp)[,1])
exp$Gene<-row.names(exp)
exp<-exp %>% gather(Tissue=1:(ncol(exp)-1))

ggplot(exp, aes(key, Gene)) + geom_tile(aes(fill = value),
     colour = "white") + scale_fill_gradient(low = "white",
     high = "steelblue")+
     labs(x='', y = '')+
      theme_bw()+
      guides(fill = guide_legend(title = "Log2(TPM)"))+
      #theme(legend.position="none")+
      theme(plot.title = element_text(hjust = 0.5,size = 20),axis.title = element_text(size=15))
  +
      theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),panel.grid.major= eleme
  nt_blank(),panel.grid.minor = element_blank())
```



Retrieval of input tissue-specific genes

The third object in the `output` is a list containing the tissue-specificity information for the input genes. The code below retrieves the tissue-specific genes along with the type of tissue-specificity in placenta tissue.

```
seGroupInf<-output[[3]][["Placenta"]]
groupInf<-data.frame(assay(seGroupInf))
print(head(groupInf))
#>      Gene           Group
#> 1     CGA Tissue-Enriched
#> 2    GCM1 Tissue-Enriched
#> 3 CYP19A1 Tissue-Enriched
#> 4   GPR32 Tissue-Enriched
#> 5  CLEC1A Tissue-Enriched
#> 6 SLC13A4 Tissue-Enriched
```

Retrieval of tissue-specific genes that could not be mapped

The fourth object in the `output` list is a character vector that has a list of input genes that were not identified in the tissue-specific gene data.

```
print(geneIds(output[[4]]))
#> [1] "C10ORF54" "CGB"      "GRAMD3"   "PVRL4"
```

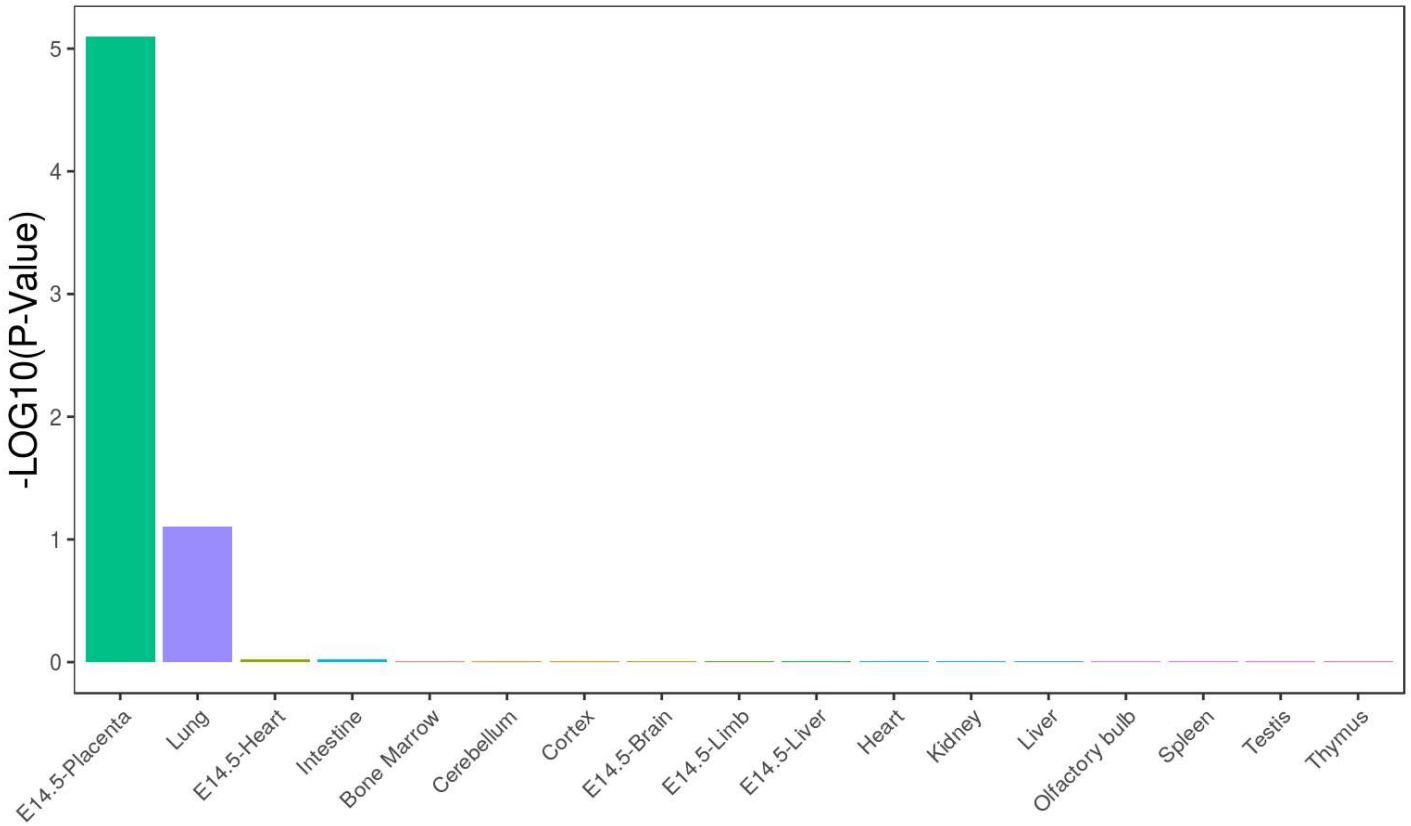
Orthologous gene enrichment

The `teEnrichment` function can calculate mouse tissue-specific gene enrichment from a human gene list or vice versa. The user simply specifies if the input data is from mouse or human, and selects the tissue-specific gene data of interest, whether it is from mouse or human. The function will automatically carry out orthologous tissue-specific gene enrichment using one-to-one protein coding orthologous genes between human and mouse, downloaded from Ensembl V91 database (Aken et al. 2016).

Example: Tissue-specific gene enrichment of mouse tissues using input human genes

In this example, the list of 100 TE-specific genes from the **Tissue-specific gene enrichment** example is used to carry out tissue-specific gene enrichment using the mouse ENCODE data.

```
library(TissueEnrich)
library(ggplot2)
genes<-system.file("extdata", "inputGenes.txt", package = "TissueEnrich")
inputGenes<-scan(genes,character())
gs<-GeneSet(geneIds=inputGenes,organism="Homo Sapiens",geneIdType=SymbolIdentifier())
output<-teEnrichment(inputGenes = gs,rnaSeqDataset = 3)
seEnrichmentOutput<-output[[1]]
enrichmentOutput<-setNames(data.frame(assay(seEnrichmentOutput), row.names = rowData(seEnrichmen
  tOutput)[,1]), colData(seEnrichmentOutput)[,1])
enrichmentOutput$Tissue<-row.names(enrichmentOutput)
ggplot(enrichmentOutput,aes(x=reorder(Tissue,-Log10PValue),y=Log10PValue,label = Tissue.Specifi
  c.Genes,fill = Tissue))+
      geom_bar(stat = 'identity')+
      labs(x='', y = '-LOG10(P-Value)')+
      theme_bw()+
      theme(legend.position="none")+
      theme(plot.title = element_text(hjust = 0.5,size = 20),axis.title = element_text(size=15))
  +
      theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),panel.grid.major= eleme
  nt_blank(),panel.grid.minor = element_blank())
```



This result shows that human TE-specific genes also show enrichment for mouse placenta-specific genes.

`teGeneRetrieval`: Identification of tissue-specific genes

The `teGeneRetrieval` function is used to define tissue-specific genes, using the algorithm from the HPA (Uhlén et al. 2015). It takes an `SummarizedExperiment` object containing expression information as input (rows as genes and columns as tissue) and classifies the genes into different gene groups and returns the information in another `SummarizedExperiment` object. The users also have the options of changing the default thresholds to vary the degree of tissue specificity of genes. More details about the gene groups and HPA thresholds are provided below.

Gene groups

The genes are divided into six groups based on their gene expression across the tissues. These groups are:

- **Not Expressed:** Genes with an expression level less than 1 (TPM or FPKM) across all the tissues.
- **Tissue Enriched:** Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to all other tissues.
- **Group Enriched:** Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a group of 2-7 tissues compared to all other tissues, and that are not considered Tissue Enriched.
- **Tissue Enhanced:** Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to the average levels in all other tissues, and that are not considered Tissue Enriched or Group Enriched.
- **Expressed in all:** Genes with an expression level greater than or equal to 1 (TPM or FPKM) across all of the tissues that are not in any of the above 4 groups.
- **Mixed:** Genes that are not assigned to any of the above 5 groups.

Genes from the **Tissue Enriched**, **Group Enriched**, and **Tissue Enhanced** groups are classified as tissue-specific genes.

Example: Tissue-specific gene retrieval

In the example below, we supplied a subset of mouse ENCODE data, consisting of expression data of 36 genes across 17 tissues.

```
library(TissueEnrich)
library(SummarizedExperiment)
data<-system.file("extdata", "test.expressiondata.txt", package = "TissueEnrich")
expressionData<-read.table(data,header=TRUE,row.names=1,sep='\t')
se<-SummarizedExperiment(assays = SimpleList(as.matrix(expressionData)),rowData = row.names(expr
  essionData),colData = colnames(expressionData))
output<-teGeneRetrieval(se)
head(assay(output))
#>      Gene                  Tissue           Group
#> [1,] "ENSMUSG00000003200" "All"            "Expressed-In-All"
#> [2,] "ENSMUSG00000003206" "Bone.Marrow"    "Tissue-Enhanced"
#> [3,] "ENSMUSG00000003208" "All"            "Mixed"
#> [4,] "ENSMUSG00000004530" "All"            "Expressed-In-All"
#> [5,] "ENSMUSG00000004535" "All"            "Expressed-In-All"
#> [6,] "ENSMUSG00000004540" "E14.5.Placenta" "Tissue-Enriched"
```

As seen above, the `output` consists of the tissue-specific genes information in a `SummarizedExperiment` object with columns for Gene name, Tissue name, and Tissue-Specific group.

`teEnrichmentCustom`: Tissue-specific gene enrichment in custom expression datasets

The `teEnrichmentCustom` function is used to calculate tissue-specific gene enrichment using tissue-specific genes defined using the `teGeneRetrieval` function.

Example: Tissue-specific gene enrichment in custom dataset

The example uses 10 genes, randomly selected from the 36 genes used in the **Tissue-specific gene retrieval** example. The tissue-specific genes identified from the custom gene expression are used to calculate tissue-specific gene enrichment in the input gene set.

```
library(TissueEnrich)
library(ggplot2)
genes<-system.file("extdata", "inputGenesEnsembl.txt", package = "TissueEnrich")
inputGenes<-scan(genes,character())
gs<-GeneSet(geneIds=inputGenes)
output2<-teEnrichmentCustom(gs,output)
enrichmentOutput<-setNames(data.frame(assay(output2[[1]]), row.names = rowData(output2[[1]])[,1
  ]), colData(output2[[1]])[,1])
ggplot(enrichmentOutput,aes(x=reorder(Tissue,-Log10PValue),y=Log10PValue,label = Tissue.Specifi
  c.Genes,fill = Tissue))+
      geom_bar(stat = 'identity')+
      labs(x='', y = '-LOG10(P-Value)')+
      theme_bw()+
      theme(legend.position="none")+
      theme(plot.title = element_text(hjust = 0.5,size = 20),axis.title = element_text(size=15))
  +
      theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),panel.grid.major= eleme
  nt_blank(),panel.grid.minor = element_blank())
```

As seen above, the metadata of the output consists of the tissue-specific gene enrichment information in a list object containing the enrichment results.

References

Aken, Bronwen L, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. “The Ensembl gene annotation system.” *Database : The Journal of Biological Databases and Curation* 2016. Oxford University Press. https://doi.org/10.1093/database/baw093 (https://doi.org/10.1093/database/baw093).

Ardlie, Kristin G., David S. Deluca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, et al. 2015. “The Genotype-Tissue Expression (Gtex) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235). American Association for the Advancement of Science:648–60. https://doi.org/10.1126/science.1262110 (https://doi.org/10.1126/science.1262110).

Petropoulos, Sophie, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. “Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos.” *Cell* 165 (4). Elsevier:1012–26. https://doi.org/10.1016/j.cell.2016.03.023 (https://doi.org/10.1016/j.cell.2016.03.023).

Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. “A map of the cis-regulatory sequences in the mouse genome.” *Nature* 448 (7409). http://www.nature.com/articles/nature11243 (http://www.nature.com/articles/nature11243).

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Asa Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” *Science* 347 (6220). American Association for the Advancement of Science. https://doi.org/10.1126/science.1260419 (https://doi.org/10.1126/science.1260419).