Package ‘pSI’

February 20, 2015

Type Package

Title Specificity Index Statistic

Version 1.1

Date 2014-01-30

Author Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Maintainer Alan B. Wells <awells@dsgmail.wustl.edu>

Description This package contains functions to calculate the Specificity Index statistic, which can be used for comparative quantitative analysis to identify genes enriched in specific cell populations across a large number of profiles, as well as perform numerous post-processing operations. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in pSI.data package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

License GPL (>= 2)

LazyData TRUE

Depends R (>= 2.10), gdata

NeedsCompilation no

Repository CRAN

Date/Publication 2014-01-30 21:12:44

R topics documented:

pSI-package 2
candidate.genes 2
candidate.overlap 4
dataset.s1 5
fisher.iteration 6
pSI.count 7
pSI.list 8
sample.data 9
specificity.index 9

Index **11**

pSI-package *Specificity Index Statistic*

Description

This package contains functions to calculate the Specificity Index statistic, which can be used for comparative quantitative analysis to identify genes enriched in specific cell populations across a large number of profiles, as well as perform numerous post-processing operations

NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Details

Package:	pSI
Type:	Package
Version:	1.1
Date:	2014-1-30
License:	GPL (>= 2)

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Maintainer:Alan B. Wells <awells@dsgmail.wustl.edu>

References

Joseph D. Dougherty, Eric F. Schmidt, Miho Nakajima, and Nathaniel Heintz Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells http://nar.oxfordjournals.org/content/38/13/4218

candidate.genes *Candidate Gene Lists*

Description

This list contains 5 candidate gene lists. They are provided as sample gene lists to be used with `fisher.iteration` & `candidate.overlap` functions found in this package.

- `candidate.genes$AutDB`
  Autism spectrum disorder candidate gene list from AutDB (N=328)

- `candidate.genes$protein.disrupting.rdnv`
  Autism spectrum disorder candidate gene list collected from 4 studies published in 2012 (N = 122)
- `candidate.genes$silent.rdnv`
  Autism spectrum disorder negative control gene list collected from 4 studies published in 2012 (N = 122)
- `candidate.genes$hcrt.genes`
  Narcolepsy Candidate Gene List (N=63)
- `candidate.genes$retinopathy.genes`
  Human Congenital Retinopathies Disease Gene List (N=120)

NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Format

5 candidate gene lists, each in the form of a character vector, which are contained within one R list.

Details

- `candidate.genes$AutDB`
  Hand-curated list of Autism Spectrum Disorder (ASD) candidate genes derived from human genetics studies downloaded from AutDB (N=328)
- `candidate.genes$protein.disrupting.rdnv`
  List of Protein-Disrupting rare de novo variant affected genes in ASD Probands (N = 122)
- `candidate.genes$silent.rdnv`
  List of Silent rare de novo variant affected genes in ASD unaffected siblings (N = 122)
- `candidate.genes$hcrt.genes`
  List of differentially dysregulated genes from narcoleptic mice with Hcrt neuron ablation versus control (N=63)
- `candidate.genes$retinopathy.genes`
  List of genes identified in human congenital retinopathies downloaded from the curated Ret-Net database (N=120)

Source

`AutDB`
Basu SN, Kollu R, Banerjee-Basu S (2009): AutDB: a gene reference resource for autism research. Nucleic Acids Research. 37:D832-D836. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13379

`protein.disrupting.rdnv` & `silent.rdnv`
Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. (2012): De novo gene disruptions in children on the autistic spectrum. Neuron. 74:285-299.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. (2012): Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 485:242-245.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. (2012): De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 485:237-241.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. (2012): Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 485:246-250.

`hcrt.genes`
Honda M, Eriksson KS, Zhang S, Tanaka S, Lin L, Salehi A, et al. (2009): IGFBP3 colocalizes with and regulates hypocretin (orexin). PLoS One. 4:e4254. `http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004254`

`retinopathy.genes`
Daiger, SP. RetNet, the Retinal Information Network. `https://sph.uth.edu/RetNet/`

Examples

```
data(candidate.genes)

names(candidate.genes)

candidate.genes[[5]]
```

`candidate.overlap` *Candidate Gene Overlap*

Description

`candidate.overlap` Extracts genes specific to samples which overlap with a candidate gene list at various pSI thresholds

Usage

```
candidate.overlap(pSIs, candidate.genes, write.csv = FALSE)
```

Arguments

`pSIs` data frame output from `specificity.index` function with the number of columns equal to the number of samples and genes as rows.

`candidate.genes` candidate gene list tested for overrepresentation in cell types/samples. Comprised of official gene symbols.

`write.csv` logical variable indicating if csv files will be written to the current working directory (default value is FALSE)

Details

Returns list consisting of 6 data frames, one for each pSI threshold. Each data frame contains genes specific to each sample which overlap with a candidate gene list and whose pSI values fall below each respective threshold for each cell type/sample included in the analysis. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Examples

```
##load sample pSI output
data(sample.data)
##load sample candidate gene lists
data(candidate.genes)
##Generates lists of overlapping genes
candidate.gene.overlap.AutDB <- candidate.overlap(pSIs=sample.data$pSI.output,
                                                   candidate.genes=candidate.genes$AutDB)
```

`dataset.s1` *Supplementary Data Set 1*

Description

Differentially expressed genes (up-regulated & down-regulated) in cortices of autsitic vs control subjects from human transcriptomic data. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Format

List containing two character vectors

Details

`dataset.s1` is a list which contains two character vectors of differentially expressed genes. Differential expression was assessed using the SAM package (Significance Analysis of Microarrays), for FDR <.05 and fold change >1.3.

- `dataset.s1$up.regulated`
  Up-regulated genes (N=234)
- `dataset.s1$down.regulated`
  - Down-regulated genes (N=229)

Source

Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. (2011): Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474:380-384.

Examples

```
data(dataset.s1)
```

`fisher.iteration` *Fisher's Exact Test Across All Cell Types & pSI Thresholds*

Description

`fisher.iteration` will test a candidate gene list for overrepresenation in the various cell type/pSI threshold combinations produced by the specificty.index function. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: `http://genetics.wustl.edu/jdlab/psi_package/`

Usage

```
fisher.iteration(pSIs, candidate.genes, background = "data.set",
  p.adjust = TRUE)
```

Arguments

`pSIs` data frame output from `specificity.index` function with the number of columns equal to the number of samples and genes as rows.

`candidate.genes` candidate gene list tested for overrepresentation in cell types/samples. Comprised of official gene symbols.

`background` character string used to indicate what background gene list should be used in Fisher's exact test for overrepresentation. The default value is `"data.set"` which indicates that the gene list of the input pSI data set will be used to represent the background gene list. This would be used in the case when the input pSI data set is comprised of genes derived from the same species as the genes found in the candidate gene list. `background` can take on two other values, the first of which is `"human.mouse"`. `"human.mouse"` indicates that the background gene list will be comprised of intersection of two lists: 1) all genes in the input pSI dataset (all are human genes), 2) all genes with clear human-mouse homologs. This option would be used in the case when the input data set is comprised of human genes (i.e. genes from a human microarray) and the candidate gene list being tested is comprised of mouse genes. The last value `background` can take on is `"mouse.human"`. `"mouse.human"` indicates that the background gene list will be comprised of intersection of two lists: 1) all genes in the input pSI dataset (all are mouse genes), 2) all genes with clear mouse-human homologs. This option would be used in the case when the input data set is comprised of mouse genes (i.e. genes from a mouse microarray) and the candidate gene list being tested is comprised of human genes.

`p.adjust` logical. default output is bonferroni corrected p-value but if `p.adjust` is `FALSE`, nominal p-values will be output.

Details

This function is used to answer the question of what is the probability that a certain number of genes specific to a certain cell type/sample occured by chance (as usual with low probabilities corresponding to high statistical significance). This is accomplished with a binary variable for each gene in the population with two mutual exclusive values: 1) The gene is specific to the cell type/sample in question or 2) The gene is not specific to the cell type/sample in question

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Examples

```
##load sample pSI output
data(sample.data)
##load sample candidate gene lists
data(candidate.genes)
##run Fisher's exact test for overrperesentation on pSI.out for the AutDB
##candidate gene list across all cell types/sample types & pSI thresholds
fisher.out.AutDB <- fisher.iteration(pSIs=sample.data$pSI.output,
                                     candidate.genes=candidate.genes$AutDB)
```

`pSI.count` *Convert pSI output to gene count list*

Description

`pSI.count` This functions counts number of genes specific to each sample type

Usage

```
pSI.count(pSIs, write.csv = FALSE)
```

Arguments

`pSIs` data frame output from `specificity.index` function with the number of columns equal to the number of samples and genes as rows.

`write.csv` logical variable indicating if csv files will be written to the current working directory (default value is FALSE)

Details

Returns data frame consisting of 6 rows, one for each pSI threshold, and as many columns as cell types/samples were included in the analysis. Each cell type/sample will have a count of many genes whose pSI values fall below each respective threshold for each cell type/sample. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Examples

```
##load sample pSI output
data(sample.data)
##Count the number of genes specific to each cell type/sample type across all pSI thresholds
pSI.out.count <- pSI.count(pSIs=sample.data$pSI.output, write.csv=TRUE)
```

`pSI.list` *Convert pSI output to gene list*

Description

`pSI.list` returns list consisting of 6 data frames, one for each pSI threshold.

Usage

```
pSI.list(pSIs, write.csv = TRUE)
```

Arguments

`pSIs` data frame output from `specificity.index` function with the number of columns equal to the number of samples and genes as rows.

`write.csv` logical variable indicating if csv files will be written to the current working directory (default value is FALSE)

Details

Each data frame contains genes whose pSI values fall below each respective threshold for each cell type/sample included in the analysis. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: http://genetics.wustl.edu/jdlab/psi_package/

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

Examples

```
##load sample pSI output
data(sample.data)
##List the genes specific to each cell type/sample type across all pSI thresholds
pSI.out.list <- pSI.list(pSIs=sample.data$pSI.output, write.csv=FALSE)
```

`sample.data` *Sample Input & Output pSI Data Sets*

Description

Sample expression matrix and sample specificity index statistic (pSI) output data sets used to illustrate the `pSI` package functions. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: `http://genetics.wustl.edu/jdlab/psi_package/`

Format

List comprised of two data frames.

Details

Sample expression matrix is the Brainspan RNAseq data condensed into 6 major regional divisions for just the young adult regions. Sample pSI output data set is calculated from the young adulthood Brainspand RNAseq data just previously described.

Source

`http://www.brainspan.org`

Examples

```
data(sample.data)
```

`specificity.index` *Specificity Index Statistic*

Description

`specificity.index` Calculates specificity index statistic (pSI) values of input expression matrix which can be used for comparative quantitative analysis to identify genes enriched in specific cell populations across a large number of profiles. This measure correctly predicts in situ hybridization patterns for many cell types. `specificity.index` returns a data frame of equal size as input data frame, with pSI values replacing the expression values. NOTE:Supplementary data (human & mouse expression sets, calculated pSI datasets, etc.) can be found in `pSI.data` package located at the following URL: `http://genetics.wustl.edu/jdlab/psi_package/`

Usage

```
specificity.index(pSI.in, pSI.in.filter, bts = 50, p_max = 0.1,
  e_min = 0.3, hist = FALSE, SI = FALSE)
```

Arguments

`pSI.in` data frame with expresion values for genes in rows, and samples or cell types in columns (at this point replicate arrays have been averaged, so one column per cell type)

`pSI.in.filter` matched array (same genes and samples) but with NA's for any genes that should be excluded for a particular cell type.

`bts` numeric. number of distributions to average for permutation testing

`p_max` numeric. maximum pvalue to be calculated

`e_min` numeric. minimum expression value for a gene to be included. For microarray studies, a value of 50 has been the default value and for RNAseq studies, a value of 0.3 has been used as the default.

`hist` logical. option for producing histograms of actual & permuted distributions of gene rank

`SI` logical. option to output SI value instead of default pSI value

Details

$$SI_{n,1} = \frac{\sum_{k=2}^{m} rank(\frac{IP_{1,n}}{IP_{k,n}})}{m-1}$$

Author(s)

Xiaoxiao Xu, Alan B. Wells, David OBrien, Arye Nehorai, Joseph D. Dougherty

References

Joseph D. Dougherty, Eric F. Schmidt, Miho Nakajima, and Nathaniel Heintz Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells Nucl. Acids Res. (2010)

Examples

```
##load sample expression matrix
data(sample.data)
##calculate specificity index on expression matrix
##(Normally for RNAseq data, and e_min of 0.3, microarrays: e_min= 50)
pSI.output <- specificity.index(pSI.in=sample.data$pSI.input, e_min=20)
```

Index

∗Topic **datasets**
    candidate.genes, 2
    dataset.s1, 5
    sample.data, 9

candidate.genes, 2
candidate.overlap, 4

dataset.s1, 5

fisher.iteration, 6

pSI *(*pSI-package*)*, 2
pSI-package, 2
pSI.count, 7
pSI.list, 8

sample.data, 9
specificity.index, 9