Package 'SEPIRA'

October 17, 2020

Type Package

Title Systems EPigenomics Inference of Regulatory Activity

Description SEPIRA (Systems EPigenomics Inference of Regulatory Activity) is an algorithm that infers sample-specific transcription factor activity from the genome-wide expression or DNA methylation profile of the sample.

Version 1.8.0

Date 2019-07-03

Author Yuting Chen [aut, cre],
Andrew Teschendorff [aut]

Maintainer Yuting Chen <cytwarmmay@hotmail.com>

License GPL-3

biocViews GeneExpression, Transcription, GeneRegulation, GeneTarget, NetworkInference, Network, Software

Depends R (>= 3.5.0)

Encoding UTF-8

Imports limma (>= 3.32.5), corpcor (>= 1.6.9), parallel (>= 3.3.1), stats

RoxygenNote 6.1.1

Suggests knitr, rmarkdown, testthat, igraph

VignetteBuilder knitr

NeedsCompilation no

git_url https://git.bioconductor.org/packages/SEPIRA

git_branch RELEASE_3_11

git_last_commit adb2d35

git_last_commit_date 2020-04-27

Date/Publication 2020-10-16

R topics documented:

ComputePCOR 2
GeneExp 3
InferTFact 3
InferTFactPRL 4

LimmaFn 5
LUSCexp 6
LUSCmeth 6
sepiraInfNet 7
sepiraRegAct 9
TFeid 10
VALexp 11

Index **12**

`ComputePCOR` *Compute partial correlation coefficient*

Description

`ComputePCOR` computes partial correlation between transcription regulators and their targets.

Usage

```
ComputePCOR(idx, mapTG.idx, mapTF.idx, selbinNET.m, exp)
```

Arguments

`idx`	Numeric, the index of target genes in the row of a binarized network `selbinNET.m`.
`mapTG.idx`	A vector of indexes output from `match` function when mapping the target genes in the network to rows of the original expression data matrix.
`mapTF.idx`	A vector of indexes output from `match` function when mapping the transcription factors (TFs) in the network to rows of the original expression data matrix.
`selbinNET.m`	A matrix, the binarized network with rows referring to TF target genes, and columns to TFs (the regulators). 0s means no regulation between TF-gene, while 1s means significant regulation (either positive or negative).
`exp`	A matrix, the original gene expression data matrix across different tissue types with rows referring to genes, columns to samples.

Details

This function is used to calculate partial correlations between a TF target gene and all its regulators.

User needs to first provide the index (`idx`) of the gene in the binarized matrix, then all its regulators are detected from the network. Then the expression profile of this gene and its regulators are retrieved from the expression data matrix (`exp`) with `mapTG.idx` and `mapTF.idx` indexes which indicates the number of rows they are in `exp` (from `match` function).

The output could be: 1) a matrix storing the partial correlation coefficients; 2) NULL if the selected gene has only one regulator according to the binarized network.

Value

A matrix with partial correlation coefficients between TF targets and their regulators.

GeneExp — *RNA-seq data matrix from GTEx*

Description

GeneExp is a gene expression data matrix with rows (467 genes) and 520 columns. This data matrix is a subset of GTEx RNA-seq data which is a very big dataset. In this subset we only randomly chose 20 samples from each tissue type with more than 50 samples in the original GTEx data set.

Usage

```
data("GeneExp")
```

Format

The format is: num [1:467, 1:520] 0.0361 0.758 0.5484 2.3569 6.1973 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:467] "9496" "22797" "7080" "6943" ... ..$: chr [1:520] "Adipose Tissue" "Adipose Tissue" "Adipose Tissue" "Adipose Tissue" ...

Source

https://www.gtexportal.org/home/

Examples

```
data(GeneExp)
## view the first 5 rows and columns
GeneExp[1:5,1:5]
```

InferTFact — *Estimate TF activity score in a sample*

Description

InferTFact is an auxiliary function for function sepiraRegAct.

Usage

```
InferTFact(exp, regnet)
```

Arguments

exp	A numeric vector of gene expression levels for all TF target genes in a sample.
regnet	A matrix, the network estimated by function sepiraInfNet, with +1 referring to positive regulation, -1 negative regulation, and 0 no regulation.

Details

InferTFact regresses the expression profiles of TF target genes against the binding profile of this TF on these genes. The output t-statistics are taken as the TF activity scores.

Value

A vector storing the activity scores of all TFs in a specific sample.

Examples

```
# TF regulatory network with 100 genes and 5 TFs.
regnet.m <- matrix(sample(c(-1, 0, 1), 500, replace = TRUE), nrow = 100)
# gene expression vector (for one sample)
exp.v <- rnorm(100)
# TF activity score
TFact.v <- SEPIRA:::InferTFact(exp.v, regnet.m)
```

InferTFactPRL *Estimate TF activity score in different samples*

Description

InferTFactPRL is an auxiliary function for function sepiraRegAct.

Usage

```
InferTFactPRL(idx, tmp, regnet)
```

Arguments

idx	A numeric vector indicating the samples from which you want to estimate the TF activity score.
tmp	A numeric matrix of gene expression levels for all TF target genes, with rows referring to genes, columns to samples.
regnet	The network estimated from function sepiraInfNet, with +1 referring to positive regulation, -1 negative regulation, and 0 no regulation.

Details

The user needs to input the index of the sample in the data matrix from which InferTFactPRL estimates TF activity. InferTFactPRL accomplishes the work by regressing the expression profile of TF target genes against the binding profile of this TF on these genes. The output t-statistics are taken as the TF activity score.

Value

A vector storing the activity scores of all TFs from user specified sample in the data matrix.

Examples

```
# TF regulatory network with 100 genes and 5 TFs.
regnet.m <- matrix(sample(c(-1, 0, 1), 500, replace = TRUE), nrow = 100)
# gene expression data matrix with 100 genes and 10 samples.
tmp.m <- matrix(rnorm(1000), nrow = 100)
# TF activity score
TFact.v <- SEPIRA:::InferTFactPRL(1, tmp.m, regnet.m)
```

`LimmaFn` *Do differential gene expression/ differential DNA methylation analysis*

Description

`LimmaFn` uses functions in `limma` package to easily compute the moderated t-statistics and p-values from differential gene expression/methylation tests comparing between different phenotypes even when sample size is small.

Usage

```
LimmaFn(pheno, data)
```

Arguments

`pheno`	A vector of sample phenotypes. Sample phenotype in a scientific research could be treatment/control, normal/cancer or smoker/non-smoker. Different phenotypes could each be encoded as 0/1 when inputting to `LimmaFn`, for example, Normal-0; Cancer-1.
`data`	A matrix, the normalized gene expression or DNA methylation dataset, should be a numeric matrix, with rows referring to genes and columns to samples. In this matrix you could use any type of gene IDs, like Entrez ID, Ensembl ID, HUGO gene symbol... But make sure to use the same gene annotation through out your analysis.

Details

This function computes the moderated t-statistic for users using empirical Bayes method, it is especially useful when the sample size is too small to perform parametric tests.

Given a normalized gene expression or DNA methylation data matrix and a vector indicating sample phenotype, `LimmaFn` first fits a linear model using `lmFit`, then it refits the model and do comparisons between any two different phenotypes with `contrasts.fit`, finally it estimates moderated t-statistics for each comparison from the fitted model using empirical Bayes method (`eBayes`) and output the result from the `topTable` function.

Note that doing the `contrasts.fit` step will not make a difference if you do comparison between two different sample status (treatment/control). However, When there are more than two sample status in your data set, this step will do comparison between every two status. And resulted summary tables will be stored in a list.

Value

A table with rows for all genes (ranked by significance) and columns of log2 fold-change, average expression, moderated t-statistic, p-value, adjusted p-value (Benjamini–Hochberg procedure). The table is the output of `topTable` function.

See Also

`lmFit` for fitting a linear model, `contrasts.fit` for refitting, `eBayes` for Bayes method, `topTable` for the output table.

Examples

```
# prepare the phenotype info ("C"-control; "T"-treatment)
pheno.v <- c("C","C","C","T","T","T")
# prepare your normalized data matrix.
data.m <- matrix(rnorm(120),nrow=20,ncol=6)
# run function
lim.o <- LimmaFn(pheno.v, data.m)
```

LUSCexp *LUSC RNA-seq dataset*

Description

LUSCmeth is a subset of TCGA LUSC RNA dataset. It contains 391 genes and 518 samples with 45 normal sample and 473 cancer samples.

Usage

```
data("LUSCexp")
```

Format

The format is: num [1:391, 1:518] -0.0842 0.7222 1.1949 -1.1637 0.2717 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:391] "100009676" "100129842" "100130417" "100131726" ... ..$: chr [1:518] "TCGA-22-5481-11" "TCGA-56-7222-11" "TCGA-77-7338-11" "TCGA-77-7138-11" ...

Source

https://cancergenome.nih.gov/

Examples

```
data(LUSCexp)
```

LUSCmeth *LUSC DNA methylation dataset*

Description

LUSCmeth is a subset of TCGA LUSC DNA methylation dataset. It contains 316 genes and 370 samples with 41 normal sample and 275 cancer samples.

Usage

```
data("LUSCmeth")
```

Format

The format is: num [1:333, 1:316] 0.154246 0.197384 0.000585 -0.033967 -0.000737 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:333] "100130417" "100287284" "10105" "1016" ... ..$: chr [1:316] "TCGA-43-3394-11A" "TCGA-18-3417-11A" "TCGA-18-4721-11A" "TCGA-22-4599-11A" ...

Source

https://cancergenome.nih.gov/

Examples

```
data(LUSCmeth)
```

---

sepiraInfNet *Infer tissue-specific regulatory network from gene expression data*

Description

`sepiraInfNet()` is one of the two main functions in package `SEPIRA`. Using it you can estimate tissue-specific regulatory networks in any tissue type of interest.

Usage

```
sepiraInfNet(data, tissue, toi, cft = NULL, TFs, sdth = 0.25,
  sigth = NULL, pcorth = 0.2, degth = c(0.05, 0.05), lfcth = c(1,
  log2(1.5)), minNtgts = 10, ncores = 4)
```

Arguments

`data`	A matrix, the normalized gene expression data matrix, with rows referring to unique genes and columns to samples from different tissue types.
`tissue`	A phenotype vector, indicating the tissue types of samples. It should have the same order as the columns of the matrix.
`toi`	A character, the tissue type of interest, a character telling the function the tissue type for which a user wants to estimate the network.
`cft`	A vector of tissue types to be used to adjust for confounding by immune or stromal cells infiltration in `toi`. It can be blood and/or spleen, which we found using `ESTIMATE` package that they contain extremely high proportion of immune and stromal cells.
`TFs`	A vector of TFs. Note that one should use the same annotation in different data sets throughout the analysis.
`sdth`	A numeric, the standard deviation threshold used to remove genes with little or zero standard deviation of its expression levels.
`sigth`	A numeric, the unadjusted p-value threshold used to call significant interactions after calculating the correlation coefficients between TFs and target genes. This threshold is used to binarize the correlation coefficient matrix. If this value is not specified by user, the function will do Bonferroni correction and then use 0.05 as the threshold.

pcorth	A numeric, the partial correlation threshold, in the range between 0 and 1, used to remove indirect interactions between TFs and their target genes.
degth	A vector of length three, thresholds of adjusted p-value to call significant TFs in 1) comparison between toi and all other tissue types; 2) & 3) comparison between toi and blood/spleen in cft.
lfcth	A vector of length three, thresholds of log2(fold-change) to call significant TFs in 1) comparison between toi and all other tissue types; 2) & 3) comparison between toi and blood/spleen in cft.
minNtgts	An integer used to filter out TFs with few targets. Only TFs with more than 'minNtgts' target genes can be kept in the network.
ncores	An integer, the number of cores to use when computing partial correlation. See mclapply.

Details

sepiraInfNet generates tissue specific TF regulatory networks from gene expression data across multi-tissue types.

The gene expression data set data should be normalized by user before inputting to sepiraInfNet, with rows referring to genes and columns to samples from different tissue types. Duplicated gene names/IDs should be averaged before normalization.

The user needs to input the tissue type of samples (tissue) in the data set as well as the tissue type of interest (toi). Please make sure the toi is in the tissue and spelled correctly.

Using differential gene expression analysis, we detect TFs that are highly active in toi and less active in other tissue types. When doing such analyses, the results could be confounded due to cell-type heterogeneity. sepiraInfNet provides a way to adjust for immune/ stromal cell contamination by doing additional comparisons between toi and 1) blood; 2)spleen as long as expression data for any one/ both of the tissue types are available in data.

TFs is a vector containing the identifiers of all TFs (regulators). In our paper we used the 1313 TFs annotated as "transcription factors" in MSigDB. You could input your own list of TFs to sepiraInfNet.

sdth is a standard deviation threshold that is used to remove genes in user provided data set which are with small or close to zero standard deviation. By default the threshold is 0.25.

From the gene expression data matrix sepiraInfNet estimates Pearson correlation coefficient between every TF-gene pair as well as corresponding p-value. The p-value threshold sigth binarizes the network into "regulation" (1) /"no regulation" (0). This binarized network is used to determine the covariants when estimating the partial correlation between target genes and their regulators (TFs).

pcorth is the partial correlation coefficient threshold for calling significant direct TF-gene interactions. By default pcorth equals 0.2.

degth and lfcth are vectors each contains the 3 thresholds for adjusted p-value/log2 fold-change to call significant TFs in comparisons between toi and 1) all other tissue types; 2) the 1st tissue type (blood) in cft; 3) the 2nd tissue type (spleen) in cft. These differential expression analyses are done to find tissue-specific TFs that are only highly activated in tissue type of interest.

When having detected tissue-specific TFs, we could get a network with only these TFs and their target genes. However sepiraInfNet further remove TFs with less than minNtgts target genes. By default the minimal number of TF targets in the final network is 10.

The step of calculating partial correlation coefficients is done by in parallel, by default sepiraInfNet splits the work into 4 sub-processes. User could use more cores by specifying parameter ncores.

Value

A list with three entries:

`$netTOI` the tissue specific network, rows refer to TF target genes, while columns refer to TFs.

`$sumnet` a matrix summarizing the number of TF target genes and the number of positively/negatively regulated target genes for each TF in the inferred network.

`$top` a list, entries are the tables summarizing the results of differential expression analyses. The first is the table from comparison between `toi` and 1) all other tissue; The rest tables are resulted from comparison to 2) the blood or/ and 3) spleen.

Examples

```
# gene expression data set (a subset of GTEx data set)
data("GeneExp")
# TFs
data("TFeid")
# run the function
cf <- "Blood"
coln <- colnames(GeneExp)
degth <- c(0.3,0.3) # 'degth = c(0.05, 0.05)' is recommended
# The resulted network is small due to the limited size of the 'GeneExp' data set
net.o <- sepiraInfNet(GeneExp,coln,"Lung",cf,TFeid,sigth=0.05,degth=degth,minNtgts=5,ncores=1)
```

`sepiraRegAct` *Infer TF activity from gene expression/ DNA methylation profile*

Description

`sepiraRegAct` calculates TF activity scores in user input data set. It could be a gene expression dataset or a DNA methylation dataset

Usage

```
sepiraRegAct(data, type = c("mRNA", "DNAm"), regnet, norm = c("c",
  "z"), ncores = 4)
```

Arguments

`data`	A gene expression or DNA methylation data matrix, with rows referring to genes and columns to samples.
`type`	A character, "mRNA" for gene expression data; "DNAm" for DNA methylation data.
`regnet`	A matrix, the regulatory network inferred from `sepiraInfNet` function.
`norm`	A character indicating the method used to normalize your input data set, "c" for "centering"; "z" for "z-score normalization".
`ncores`	A numeric, the number of cores to use. See `mclapply`.

Details

sepiraRegAct is one of the two main functions in SEPIRA package. It takes the output regulatory network from sepiraInfNet as input, and computes the activity of all TFs in this network from user provided data.

The data matrix could be gene expression data or DAN methylation data, with rows are genes and columns are samples. Duplicated row names are not allowed, so you should average the these rows before running sepiraRegAct.

Note that it's very important that you use the same gene identifier through out the whole analysis.

Value

A matrix of TF activity score with rows referring to TFs, columns to samples.

Examples

```
# gene expression dataset
data("GeneExp")
# TFs
data("TFeid")
# run the function
cf <- "Blood"
coln <- colnames(GeneExp)
degth <- c(0.3,0.3) # 'degth = c(0.05, 0.05)' is recommended
net.o <- sepiraInfNet(GeneExp,coln,"Lung",cf,TFeid,sigth=0.05,degth=degth,minNtgts=5,ncores=1)
# normalized LSCC DNAm data set from TCGA
data("LUSCmeth")
# estimate TF activity
TFact.lusc <- sepiraRegAct(LUSCmeth,type="DNAm",regnet=net.o$netTOI,norm="z",ncores=1)
TFact.gtex <- sepiraRegAct(GeneExp,type="exp",regnet=net.o$netTOI,norm="z",ncores=1)
```

TFeid *Transcription factor genes annotated using Entrez gene ID*

Description

TFeid is a vector of transcription factor Entrez IDs. This list of TFs is from MSigDB with 1385 TFs in it. It is of different length as the one we used in our paper because of different "MSigDB" versions.

Usage

```
data("TFeid")
```

Format

A vecto with 1385 observations. chr [1:1385] "9684" "59269" "2965" "6935" "9247" "3661" "6239" "11151" ...

Examples

```
data(TFeid)
## maybe str(TFeid) ; plot(TFeid) ...
```

VALexp *RNA-seq data matrix from ProteinAtlas project*

Description

VALexp is a gene expression data matrix with rows (369 genes) and 35 columns (samples from 7 tissue types). This data matrix was obtained from EBI arrayexpress and is part of ProteinAtlas project. It is a very big dataset, so we only used a small subset as an example dataset here.

Usage

```
data("VALexp")
```

Format

The format is: num [1:467, 1:200] 0 2.379 0 0.379 5.392 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:467] "9496" "22797" "7080" "6943" ... ..$: chr [1:200] "saliva-secreting gland" "lung" "liver" "heart" ...

Source

https://www.ebi.ac.uk/arrayexpress/

Examples

```
data(VALexp)
## view the first 5 rows and columns
VALexp[1:5,1:5]
```

Index

∗ **datasets**
GeneExp, 3
LUSCexp, 6
LUSCmeth, 6
TFeid, 10
VALexp, 11

ComputePCOR, 2
contrasts.fit, *5*

eBayes, *5*

GeneExp, 3

InferTFact, 3
InferTFactPRL, 4

LimmaFn, 5
lmFit, *5*
LUSCexp, 6
LUSCmeth, 6

sepiraInfNet, 7
sepiraRegAct, 9

TFeid, 10
topTable, *5*

VALexp, 11