Package ‘ZetaSuite’

October 12, 2022

Type Package

Title Analyze High-Dimensional High-Throughput Dataset and Quality Control Single-Cell RNA-Seq

Version 1.0.1

Date 2022-05-22

Maintainer Junhui Li <ljh.biostat@gmail.com>

Description The advent of genomic technologies has enabled the generation of two-dimensional or even multi-dimensional high-throughput data, e.g., monitoring multiple changes in gene expression in genome-wide siRNA screens across many different cell types (E Robert McDonald 3rd (2017) <doi:10.1016/j.cell.2017.07.005> and Tsherniak A (2017) <doi:10.1016/j.cell.2017.06.010>) or single cell transcriptomics under different experimental conditions. We found that simple computational methods based on a single statistical criterion is no longer adequate for analyzing such multi-dimensional data. We herein introduce 'ZetaSuite', a statistical package initially designed to score hits from two-dimensional RNAi screens.We also illustrate a unique utility of 'ZetaSuite' in analyzing single cell transcriptomics to differentiate rare cells from damaged ones (Vento-Tormo R (2018) <doi:10.1038/s41586-018-0698-6>). In 'ZetaSuite', we have the following steps: QC of input datasets, normalization using Z-transformation, Zeta score calculation and hits selection based on defined Screen Strength.

Imports RColorBrewer, Rtsne, dplyr, e1071, ggplot2, reshape2, gridExtra, mixtools

License GPL-2 | GPL-3

Depends R (>= 2.10)

RoxygenNote 7.1.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

Author Yajing Hao [aut] (<https://orcid.org/0000-0003-1384-4176>),
Shuyang Zhang [ctb] (<https://orcid.org/0000-0002-8428-1828>),
Junhui Li [cre],
Guofeng Zhao [ctb],
Xiang-Dong Fu [cph, fnd] (<https://orcid.org/0000-0001-5499-8732>)

NeedsCompilation no

Repository CRAN

Date/Publication 2022-05-24 19:40:02 UTC

R topics documented:

`countMat` *Subsampled data from in-house HTS2 screening for global splicing regulators.*

Description

A data frame with 1609 individual screened genes and 100 functional readouts. The data was generated from a siRNA screen for global splicing regulators. In this screen, we interrogated ~400 endogenous alternative splicing (AS) events by using an oligo ligation-based strategy to quantify 18,480 pools of siRNAs against annotated protein-coding genes in the human genome.

Usage

```
data("countMat")
```

Format

A data frame with 1609 observations on the following 100 variables

A data frame with 1609 observations on the following 100 maker variables.Each row represents gene with specific knocking-down siRNA pool, each column is an AS event. The values in the matrix are the processed foldchange values between included exons and skipping exons read counts.

Details

This data frame is the raw output data from large-scale screening.

Examples

```
data(countMat)
```

`countMatSC` *The cell x gene matrix from single-cell RNA-seq.*

Description

A scRNA-seq dataset generated from placenta that has been analyzed with CellRanger and used to develop EmptyDrops. We have subsampled the genes from the real datasets to generated the matrix.

Usage

```
data("countMatSC")
```

Format

A data frame with 1090 cells and 10000 genes. This is the subset of data obtained from single-cell RNAseq for package testing. Each row represents one cell detected in single-cell RNA-seq, each column is one gene in detected cells. The values in the matrix are the raw read counts from single-cell RNAseq.

> A data frame with 1090 cells and 10000 genes.This is the subset of data obtained from single-cell RNAseq for package testing.

Details

This data frame is the generated by single-cell RNA-seq.

Examples

```
data(countMatSC)
```

`EventCoverage` *Generation of Zeta Plot.*

Description

A zeta plot is generated from the input Z-score matrix. Zeta plot labels: x-axis: Z-score cutoffs, y-axis: the percentage of readouts that survived at a given Z-score cutoff over the total scored readouts. In order to generate this plot, the range of Z-scores is determined by ranking the absolute value of Zij (Z-score value in row i and column j) from the smallest to the largest. Z-cutoffs next are selected in the range of (-|Znxmx0.9999|, -2) to (2, |Znxmx0.9999|) to excluded the insignificant changes that may result from experimental noise(|Z| < 2, which equals to p-value >0.05). Then, for all Zij within the selected range (both positive range and negative range), the range is divided equally into x bins (the recommended input of x is 100). Thus, the percentage of readouts scored above the Z-cutoff in each bin is determined.

Usage

```
EventCoverage(ZscoreVal, negGene, posGene, binNum, combine = TRUE)
```

Arguments

`ZscoreVal`	zscore value
`negGene`	negative control dataset, the siRNAs/genes used as negative controls in screening.
`posGene`	positive control dataset, the siRNAs/genes used as positive controls in screening.
`binNum`	bin number
`combine`	combine two direction zeta together(TRUE or FALSE),default FALSE

Value

A list of data.frames and plots, the data.frame includes 'ZseqList', 'EC_N_I', 'EC_N_D', 'EC_P_I' and 'EC_P_D'. The plot 'EC_jitter_D' and 'EC_jitter_I' are the zeta plot for positive and negative samples.'ZseqList', 'EC_N_I', 'EC_N_D', 'EC_P_I' and 'EC_P_D' are the inputfiles for zeta plot and SVM.R. ZseqList describs the bin size in the zeta plot.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(countMat)
data(negGene)
data(posGene)
ZscoreVal <- Zscore(countMat,negGene)
ECList <- EventCoverage(ZscoreVal,negGene,posGene,binNum=100,combine=TRUE)
```

`FDRcutoff` *Find a cut-off according to screen strength.*

Description

Find a cutoff according to the Screen Strength (SS) and graph the Screen Strength plot. Zeta score is used to rank genes, and then, SS is calculated to define a suitable cutoff so that the cutoff can define hits at different confidence intervals. Formula of SS: SS = 1 - aFDR/bFDR, where aFDR (apparent FDR) = number of non-expressors identified at hits divided by the total number of hits, bFDR (baseline FDR) = total number of non-expressors divided by all screened genes. SS plot labels: x-axis: zeta score, y-axis: Screen Strength, SS value is determined at each bin (m bin in total), then connect individual SS value to generate a simulated SS curve based on balance points. Users may choose one or multiple balance point as the different SS intervals.

Usage

```
FDRcutoff(zetaData, negGene, posGene, nonExpGene, combine = FALSE)
```

Arguments

`zetaData` ZetaScore file calculated by ZetaSuite.

`negGene` negative control dataset, the siRNAs/genes used as negative controls in screening.

`posGene` positive control dataset, the siRNAs/genes used as positive controls in screening.

`nonExpGene` non-expressed gene

`combine` combine two direction zeta together(TRUE or FALSE),default FALSE

Value

A list of data.frame and plots, the data.frame is cut off matrix with 6 columns including "Cut_Off","aFDR", "SS","TotalHits","Num_nonExp" and "Type". Plots includes 'Zeta_type' and 'SS_cutOff'.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(nonExpGene)
data(negGene)
data(posGene)
data(ZseqList)
data(countMat)
ZscoreVal <- Zscore(countMat,negGene)
zetaData <- Zeta(ZscoreVal,ZseqList,SVM=FALSE)
cutoffval <- FDRcutoff(zetaData,negGene,posGene,nonExpGene,combine=TRUE)
```

`negGene` *Input negative file.*

Description

A data frame with 510 different well IDs in which the cells treated with non-specific siRNAs.If users did not have the build-in negative controls, the non-expressed genes should be provided here.

Usage

```
data("negGene")
```

Format

A data frame with 510 different well IDs in which the cells treated with non-specific siRNAs.

> A data frame with 510 different well IDs in which the cells treated with non-specific siRNAs. These wells were served as negative control.

Details

These wells were designed by the authors in the large-scale screen.

Examples

```
data(negGene)
```

nonExpGene *Input internal negative control file.*

Description

A data frame with 722 different well IDs in which the cells treated with siRNAs targeting to non-expressed genes in HeLa cells.It the subset of total non-expressed genes in HeLa cells.

Usage

```
data("nonExpGene")
```

Format

A data frame with 722 different well IDs in which the cells treated with siRNAs targeting to non-expressed genes in HeLa cells.

> A data frame with 722 different well IDs in which the cells treated with siRNAs targeting to non-expressed genes in HeLa cells. These wells were served as internal negative controls.

Details

These non-expressed genes can be obtained from a prior expression profile.

Examples

```
data(nonExpGene)
```

posGene *Input positive file.*

Description

A data frame with 299 different well IDs in which the cells treated with siRNAs targeting to PTB.If users didn't have the build-in positive controls, choose the parameters -withoutsvm and the filename can use any name such as 'NA'.

Usage

```
data("negGene")
```

Format

A data frame with 299 different well IDs in which the cells treated with siRNAs targeting to PTB.

A data frame with 299 different well IDs in which the cells treated with siRNAs targeting to PTB. These wells were served as positive control.

Details

These wells were designed by the authors in the large-scale screen.

Examples

```
data(posGene)
```

QC *Quality control of input datasets.*

Description

Quality Control (QC) is a step in evaluating the experiment design. For all two-dimension high throughput data, the t-SNE plot is firstly used to evaluate whether features are sufficient to separate positive and negative controls. The SSMD score (See reference Zhang) is further generated for each readout to evaluate the percentage of high-quality readouts.

Usage

```
QC(countMat, negGene, posGene)
```

Arguments

`countMat` input data set. The siRNA/gene x readouts matrix from HTS2 or large-scale RNAi screens

`negGene` negative control data set, the siRNAs/genes used as negative controls in screening.

`posGene` positive control data set, the siRNAs/genes used as positive controls in screening.

Value

A list of plots, and their names are 'score_q', 'tSNE_QC', 'QC_box' and 'QC_SSMD'. 'tSNE_QC' is the global evaluation based on all the readouts. This figure can evaluate whether the positive and negative samples are well separated based on current all readouts. And the other 3 plots are the quality evaluation of the individual readouts.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

References

Laurens van der Maaten GH: Visualizing Data using t-SNE. JournalofMachineLearningResearch 2008,9(2008):2579-2605.

Zhang XD: A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. Genomics 2007, 89:552-561.

Examples

```
data(countMat)
data(negGene)
data(posGene)
QC(countMat,negGene,posGene)
```

SVM *Find a SVM curve to separate positive and negative controls.*

Description

Radical kernel SVM is constructed to maximally separate positive controls from negative controls in the prior defined Z range using e1071 packages of R, and therefore, the SVM curve is generated.

Usage

```
SVM(ECdataList)
```

Arguments

ECdataList	data list of output EventCoverage, names of list shoule be ’EC_N_D’, ’EC_P_D’, ’EC_N_I’, ’EC_P_I’ and ’ZseqList’

Value

A list of data.frame, including ’cutOffD’ and ’cutOffI’.cutOffD and cutOffI are the deduced SVM.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(countMat)
data(negGene)
data(posGene)
ZscoreVal <- Zscore(countMat,negGene)
ECdataList <- EventCoverage(ZscoreVal,negGene,posGene,binNum=10,combine=TRUE)
SVM(ECdataList)
```

SVMcurve *The SVM curve lines in Zeta-plot.*

Description

The SVM curves were calculated from raw input matrix files. They were designed to maximally seperate the positive and negative genes.

Usage

```
data("SVMcurve")
```

Format

A data frame with 24 rows and 4 features.

A data frame with 24 rows and 4 features.The first column is the bins cut-offs for decresed direction. The second column is the values of percentage with different cut-offs in column 1. The third column is the bins cut-offs for increased direction. The fourth column is the values of percentage with different cut-offs in column 3.

Details

This data frame is the generated by SVM.R.

Examples

```
data(SVMcurve)
```

`Zeta` *Calculation of zeta and weighted zeta score.*

Description

This step calculates the Zeta Score based on the two curvecs. Firstly, this step provides another curve above the SVM curve to set a value to represent the regulatory function of gene i. Then, the area between the two curves (the one mentioned above and the SVM curve) is calculated as the Zeta score for this gene. Since the graph of the curves is divided into m bins, then the Zeta score can be calculated as the sum of all the bins' areas that exist between the two curves.

Usage

```
Zeta(ZscoreVal, ZseqList, SVMcurve = NULL, SVM = FALSE)
```

Arguments

`ZscoreVal`	input file name.
`ZseqList`	the list of bins.
`SVMcurve`	SVM curves for decrease and increase direction.###not always use
`SVM`	do SVM or not, default is FALSE

Value

A data.frame where zeta values for all tested knockding-down genes including positive and negative controls. The first column is the direction which knockding-down gene will lead to exon inclusion, whereas the second column is the knock-down genes will lead to exon skipping.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(ZseqList)
data(SVMcurve)
data(countMat)
data(negGene)
ZscoreVal <- Zscore(countMat,negGene)
zetaData <- Zeta(ZscoreVal,ZseqList,SVM=FALSE)
```

ZetaSuitSC *Calculation of zeta score for single cell RNA-seq.*

Description

This tool is used to evalucate the quality of cells detected in the single-cell RNA-seq. A zeta score will be assigned to each cell. And a cut-off for low quality and broken cells will be provided. The users can based on the selected cut-off to select the high quality cells for further analysis.

Usage

```
ZetaSuitSC(countMatSC, binNum = 10, filter = TRUE)
```

Arguments

`countMatSC`	Shalek input matrix
`binNum`	bin number for ZetaScore calculation.
`filter`	Whether to filter the extreme low read counts cells with nCount <100. default is TRUE

Value

A list of data.frame and plots. The data.frame is the Cell matrix with column name 'Cell' and 'Zeta'. The plot is the distribution of Zeta score for the detected cells and including a cut-off for removing the broken and empty cells.

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(countMatSC)
zetaDataSC <- ZetaSuitSC(countMatSC,binNum=50,filter=TRUE)
```

Zscore *Z-transformation for input matrix.*

Description

In this step, the input matrix is transformed to Z-score matrix.

Usage

```
Zscore(countMat, negGene)
```

Arguments

`countMat`	input data set. The siRNA/gene x readouts matrix from HTS2 or large-scale RNAi screens.
`negGene`	negative control dataset, the siRNAs/genes used as negative controls in screening. Z-transfromation according to thses negative control siRNAs/genes for each readout.

Details

The initial input matrix is arranged in N x M dimension, where each row contains individual functional readouts against a siRNA pool and each column corresponds to individually siRNA pools tested on a given functional readout. Readouts in each column may be thus considered as the data from one-dimensional screen (many-to-one), and thus, the typical Z statistic can be used to evaluate the relative function of individual genes in such column. The conversion is repeated on all columns, thereby converting the raw activity matrix into a matrix. Suppose Nij are the values in the original matrix i (1 <= i <= N siRNA pool) row and j (1 <= j <= M readout) column, then Zij = (Nij - uj) / sigma(j), where uj and sigma(j) are the mean and standard deviation of negative control samples in column j.

Value

A Z-transformated matrix, where each row represents each knocking-down condition and each column is a specific readout (AS event). The values in the matrix are the normalized values(Z-scores).

Author(s)

Yajing Hao, Shuyang Zhang, Junhui Li, Guofeng Zhao, Xiang-Dong Fu

Examples

```
data(countMat)
data(negGene)
ZscoreVal <- Zscore(countMat,negGene)
ZscoreVal[1:5,1:5]
```

ZseqList *The bin size for Zeta calculation.*

Description

A data frame with 11 different cut-offs and 2 directions. We divided the ranges of input values into bins. The number of bins is determined by the users.

Usage

```
data("ZseqList")
```

Format

A data frame with 11 different cut-offs and 2 directions.

A data frame with 11 different cut-offs and 2 directions.We divided the ranges of input values into bins. The number of bins is determined by the users.

Details

This data frame is the generated by EventCoverage.R.

Examples

```
data(ZseqList)
```

Index