SCRABBLE

June 7, 2018

Type Package

Title Impute scRNAseq data method

Version 0.1.0

Author Tao Peng

Maintainer The package maintainer Tao Peng <tpengmath@gmail.com>

Description SCRABBLE imputes drop-out data by optimizing an objective function that consists of three terms. The first term ensures that imputed values for genes with nonzero expression remain as close to their original values as possible, thus minimizing unwanted bias towards expressed genes. The second term ensures the rank of the imputed data matrix to be as small as possible. The rationale is that we only expect a limited number of distinct cell types in the samples. The third term operates on the bulk RNA-Seq data. It ensures consistency between the average gene expression of the aggregated imputed data and the average gene expression of the bulk RNA-Seq data. We developed a convex optimization algorithm to minimize the objective function.

License GPL-3

Encoding UTF-8

Depends R(>= 3.3)

LazyData true

Imports Rcpp (>= 0.12.13), rARPACK, pracma

LinkingTo Rcpp, RcppEigen

RoxygenNote 6.0.1

R topics documented:

data *Test data for scrabble*

Description

"data" is a data list with the length of 3. The first element in the list is generated drop-out scRNAseq data with 732 genes and 1000 cells. The second element in the list is the generated bulk RNAseq data with 732 genes. The third element is the true scRNAseq data without dropouts. The steps of generating the data is shown in Details section.

Usage

```
data_sc <- data[[1]]
data_bulk <- data[[2]]
data_true <- data[[3]]
```

Format

An object of class `list` of length 3.

Details

The data set was generated from down sampling from bulk RNAseq data. We used the bulk RNA-Seq data set of mouse hair follicles (GSE85039). In total, the dataset contains 20 different combinations of anatomic sites and developmental time points, thus constituting a high dimensional measurement space. We used the following procedures to generate the drop-out datasets. 1) We selected 732 genes that are differentially expressed in the 20 conditions based on ANOVA analysis. 2) We randomly selected 10 out of the 20 conditions. 3) For each condition, we generated 100 resampled datasets. The means and standard deviations of genes were calculated for each condition based on the 100 resampled datasets. 4) 100 new datasets were generated based on the mean and the standard deviation of each gene. 5) The final data set was obtained by combining 1000 samples representing the 10 conditions. This 1000x732 matrix now represents 1000 cells and 732 genes. 6) we make the drop-out rate of each gene in each cell following a double exponential function . Zero values are introduced into the simulated data for each gene in each cell based on the Bernoulli distribution defined by the corresponding drop-out rate.

Author(s)

Tao Peng, Kai Tan

scrabble *Runs SCRABBLE*

Description

SCRABBLE imputes drop-out data by optimizing an objective function that consists of three terms. The first term ensures that imputed values for genes with nonzero expression remain as close to their original values as possible, thus minimizing unwanted bias towards expressed genes. The second term ensures the rank of the imputed data matrix to be as small as possible. The rationale is that we only expect a limited number of distinct cell types in the samples. The third term operates on the bulk RNA-Seq data. It ensures consistency between the average gene expression of the aggregated imputed data and the average gene expression of the bulk RNA-Seq data. We developed 58 a convex optimization algorithm to minimize the objective function.

Usage

```
scrabble(data, parameter, nIter, error_out_threshold = 1e-05,
  nIter_inner = 5, error_inner_threshold = 1e-05)
```

Arguments

`data`	the input data list. There are two cases SCRABBLE could handle. The first one is that the input data is a list of two datasets, scRNAseq and bulk RNAseq. The second one is scRNAseq only.
`parameter`	the vector of parameters. The first parameter is the value of alpha in the mathematical model and the second one is the value of beta in the mathematical model.
`nIter`	the maximum iterations.
`error_out_threshold`	the threshold of the error between the current imputed matrix and the previous one. Default is 1e-5.
`nIter_inner`	the maximum interations of calculating the sub-optimization problem. Default is 5.
`error_inner_threshold`	the threshold of the error between the current updated matrix and the previous one. Default is 1e-5.

Value

A data matrix with the same size of the input scRNAseq data

Examples

```
# Set up the parameter used in SCRABBLE
parameter <- c(100,2e-7)
nIter <- 100

# Run SCRABLE
result <- scrabble(data,parameter = parameter, nIter = nIter)
```

Index