Package 'dsb'

September 3, 2021

Type Package

Title Normalize & Denoise Droplet Single Cell Protein Data (CITE-Seq)

Version 0.2.0

Description This lightweight R package provides a method for normalizing and denoising protein expression data from droplet based single cell experiments. Raw protein Unique Molecular Index (UMI) counts from sequencing DNA-conjugated antibodies in droplets (e.g. 'CITE-seq') have substantial measurement noise. Our experiments and computational modeling revealed two major components of this noise: 1) protein-specific noise originating from ambient, unbound antibody encapsulated in droplets that can be accurately inferred via the expected protein counts detected in empty droplets, and 2) droplet/cell-specific noise revealed via the shared variance component associated with isotype antibody controls and background protein counts in each cell. This package normalizes and removes both of these sources of noise from raw protein data derived from methods such as 'CITE-seq', 'REAP-seq', 'ASAP-seq', 'TEA-seq', 'proteogenomic' data from the Mission Bio platform, etc. See the vignette for tutorials on how to integrate dsb with 'Seurat', 'Bioconductor' and the AnnData class in 'Python'. Please also see our preprint Mulè M.P., Martins A.J., and Tsang J.S. (2020) <https://www.biorxiv.org/content/10.1101/2020.02.24.963603v3> for more details on the dsb method.

License BSD_3_clause + file LICENSE | file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Depends R (>= 2.10)

biocViews

Imports magrittr, limma, mclust, stats

Suggests testthat, knitr, rmarkdown, ggplot2, cowplot, spelling

URL https://github.com/niaid/dsb

BugReports https://github.com/niaid/dsb/issues

VignetteBuilder knitr

Language en-US

NeedsCompilation no

Author Matthew Mulè [aut, cre] (<https://orcid.org/0000-0001-8457-2716>),
Andrew Martins [aut] (<https://orcid.org/0000-0002-1832-1924>),
John Tsang [pdr] (<https://orcid.org/0000-0003-3186-3047>)

Maintainer Matthew Mulè <mattmule@gmail.com>

Repository CRAN

Date/Publication 2021-09-03 00:20:05 UTC

R topics documented:

`cells_citeseq_mtx` *small example CITE-seq protein dataset for 87 surface protein in 2872 cells*

Description

A matrix of raw UMI counts for surface proteins for surface proteins measured with CITE-seq antibodies. This data is used for example scripts in the dsb package. Raw data was processed with CITE-seq-Count.

Usage

```
cells_citeseq_mtx
```

Format

An R matrix, rows: 87 proteins, columns: 2872 cells

cells_citeseq_mtx R matrix of cells by proteins; a random distribution of a maximum of 100 cells per cluster from the 30 clusters reported in Kotliarov et. al. 2020

References

Kotliarov et. al. 2020 Nat. Medicine

DSBNormalizeProtein *Normalize single cell antibody derived tag (ADT) protein data with the DSBNormalizeProtein function. This single function runs step I (ambient protein background correction) and step II (defining and removing cell to cell technical variation) of the dsb normalization method. See <https://www.biorxiv.org/content/10.1101/2020.02.24.963603v3> for details of the algorithm.*

Description

Normalize single cell antibody derived tag (ADT) protein data with the DSBNormalizeProtein function. This single function runs step I (ambient protein background correction) and step II (defining and removing cell to cell technical variation) of the dsb normalization method. See <https://www.biorxiv.org/content/10.1101/2020.02.24.963603v3> for details of the algorithm.

Usage

```
DSBNormalizeProtein(
  cell_protein_matrix,
  empty_drop_matrix,
  denoise.counts = TRUE,
  use.isotype.control = TRUE,
  isotype.control.name.vec = NULL,
  define.pseudocount = FALSE,
  pseudocount.use,
  quantile.clipping = FALSE,
  quantile.clip = c(0.001, 0.9995),
  return.stats = FALSE
)
```

Arguments

`cell_protein_matrix`
: Raw protein ADT count data to be normalized with cells as columns and proteins as rows. See vignette, this is defined after quality control outlier cell removal based on the filtered output from Cell Ranger. Any CITE-seq count alignment tool can be used to define this as well.

`empty_drop_matrix`
: Raw empty droplet protein count data used for background correction with cells as columns and proteins as rows. This can easily be defined from the raw output from Cell Ranger (see vignette). Any count alignment tool for CITE-seq can be used to align and define these background drops.

`denoise.counts`
: TRUE (default) recommended to keep this TRUE and use with use.isotype.control = TRUE. This runs step II of the dsb algorithm to define and remove cell to cell technical noise.

`use.isotype.control`
: TRUE (default) recommended to use this with denoise.counts = TRUE. This includes isotype controls in defining the dsb technical component.

`isotype.control.name.vec`
: A vector of the names of the isotype control proteins in the rows of the cells and background matrix e.g. isotype.control.name.vec = c('isotype1', 'isotype2') or rownames(cells_citeseq_mtx)[grepl('sotype', rownames(cells_citeseq_mtx))]

`define.pseudocount`
: FALSE (default) uses the value 10 optimized for protein ADT data.

`pseudocount.use`
: the pseudocount to use if overriding the default pseudocount by setting define.pseudocount = TRUE

`quantile.clipping`
: FALSE (default), if outliers or a large range of values for some proteins is seen (e.g. -50 to 50) re-run with quantile.clipping = TRUE. This applies 0.001 and 0.998th quantile value clipping to handle low and high magnitude outliers.

`quantile.clip`
: if quantile.clipping = TRUE, one can provide a vector of the lowest and highest quantile to clip, these can be tuned to the dataset size. The default c(0.001, 0.9995) optimized to clip only a few of the most extreme outliers.

`return.stats`
: if TRUE, returns a list, element 1 $dsb_normalized_matrix is the normalized adt matrix element 2 $dsb_stats is the internal stats used by dsb during denoising (the background mean, isotype control values, and the final dsb technical component that is regressed out of the counts)

Value

The normalized ADT data are returned as a standard R "matrix" of cells (columns) by proteins (rows) that can be added to any Seurat, SingleCellExperiment or python anndata object - see vignette.

Author(s)

Matthew P. Mulè, <mattmule@gmail.com>

Examples

```
library(dsb) # lazy load example data cells_citeseq_mtx and empty_drop_matrix included in package

# use a subset of cells and background droplets from example data
cells_citeseq_mtx = cells_citeseq_mtx[ ,1:400]
empty_drop_matrix = empty_drop_citeseq_mtx[ ,1:400]

# example I
adt_norm = dsb::DSBNormalizeProtein(
  # step I: remove ambient protein noise reflected in counts from empty droplets
  cell_protein_matrix = cells_citeseq_mtx,
  empty_drop_matrix = empty_drop_matrix,

 # recommended step II: model and remove the technical component of each cell's protein data
```

```
  denoise.counts = TRUE,
  use.isotype.control = TRUE,
  isotype.control.name.vec = rownames(cells_citeseq_mtx)[67:70]
)

# example II - experiments without isotype controls
adt_norm = dsb::DSBNormalizeProtein(
  cell_protein_matrix = cells_citeseq_mtx,
  empty_drop_matrix = empty_drop_matrix,
  denoise.counts = FALSE
)

# example III - return dsb internal stats used during denoising for each cell
# returns a 2 element list - the normalized matrix and the internal stats
dsb_object = dsb::DSBNormalizeProtein(
   cell_protein_matrix = cells_citeseq_mtx,
   empty_drop_matrix = empty_drop_matrix,
   isotype.control.name.vec = rownames(cells_citeseq_mtx)[67:70],
   return.stats = TRUE
)

# the dsb normalized matrix to be used in downstream analysis
dsb_object$dsb_normalized_matrix

# the internal dsb stats; can be examined for outliers see vignette FAQ
dsb_object$dsb_stats
```

`empty_drop_citeseq_mtx` *small example CITE-seq protein dataset for 87 surface protein in 8005 empty droplets*

Description

A matrix of empty background droplet counts for surface proteins measured with CITE-seq antibodies. This data is used for example scripts in the dsb package. Raw data was processed with CITE-seq-Count.

Usage

```
empty_drop_citeseq_mtx
```

Format

An R matrix, rows: 87 proteins, columns: 8005 empty droplets.

empty_drop_citeseq_mtx R matrix of empty / background droplets from a CITE-seq experiment. Negative drops were called on cell hashing data with Seurat's HTODemux function and cross referencing mRNA in droplets against patient genotypes with Demuxlet. Ambiguous drops, and with less than 80 unique mRNA were removed. This is used for robust estimation of the background distribution of each protein

References

Kotliarov et. al. 2020 Nat. Medicine

Index