

Package ‘CSCDRNA’

November 1, 2024

Title Covariance Based Single-Cell Decomposition of Bulk Expression Data

Version 2.0.1

Author Ali Karimnezhad [aut, cre, ctb]

Maintainer Ali Karimnezhad <ali.karimnezhad@gmail.com>

Description Provides accurate cell type proportion estimation by incorporating covariance structure in both single-cell and bulk RNA-seq datasets into the analysis. For more detail, see Karimnezhad, A. (2022) <<https://doi.org/10.1093/bioadv/vbac049>>.

License GPL-3

Encoding UTF-8

biocViews MAST

Imports nlshrink, limSolve, Biobase, BisqueRNA, methods, plyr, Seurat, MAST, SeuratObject, Matrix

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Depends R (>= 4.4)

LazyData true

BugReports <https://github.com/empiricalbayes/CSCDRNA/issues>

NeedsCompilation no

Contents

CSCD	2
example_data	3
Index	5

CSCD

*Performs gene expression decomposition.***Description**

Provides accurate cell-type proportion estimation by incorporating covariance structure in given single-cell RNA-seq (scRNA-seq) and bulk RNA-seq datasets, see Karimnezhad (2022). The approach uses an extension of the transformation used in Jew et al. (2020) implemented in the `BisqueRNA::ReferenceBasedDecomposition()` function.

Usage

```
CSCD(
  bulk.eset,
  sc.eset,
  min.p = NULL,
  markers = NULL,
  cell.types = "cellType",
  subj.names = "SubjectName",
  verbose = TRUE
)
```

Arguments

<code>bulk.eset</code>	ExpressionSet with bulk data. Bulk RNA-seq data can be converted from a matrix with samples in columns and genes in rows to an ExpressionSet. See example_data for an example on how to create a bulk.eset object.
<code>sc.eset</code>	ExpressionSet with single-cell data. Single-cell data requires additional information in the ExpressionSet, specifically cell-type labels and individual labels. See example_data for an example on how to create a sc.eset object.
<code>min.p</code>	A percentage. This parameter is passed to the <code>Seurat::FindAllMarkers()</code> function (Butler et al., 2019) to pick the most relevant genes to each cell-type cluster. Users may pick a number between 0.3 and 0.5 for best results. The higher the value, the more genes to be excluded from the analysis.
<code>markers</code>	A character vector containing marker genes to be used in decomposition. If NULL provided, the method will use all available genes for decomposition.
<code>cell.types</code>	Character string. A vector of cell-type labels.
<code>subj.names</code>	Character string. A vector of individual labels that correspond to cells.
<code>verbose</code>	Boolean. Whether to print log info during decomposition. Errors will be printed regardless.

Value

A list. Slot **bulk.props** contains a matrix of cell-type proportion estimates with cell types as rows and individuals as columns. Slot **sc.props** contains a matrix of cell-type proportions estimated directly from counting single-cell data. Slot **transformed.bulk** contains the covariance-based transformed bulk expression used for decomposition. These values are generated by applying a linear transformation to the CPM expression. Slot **genes.used** contains a vector of genes used in decomposition. Slot **rnorm** contains Euclidean norm of the residuals for each individual's proportion estimates.

References

- Butler, A. et al. (2019). Seurat: Tools for Single Cell Genomics. R package version, 4.1.1.
- Jew, B. et al. (2020) Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun 11, 1971. <https://doi.org/10.1038/s41467-020-15816-6>
- Jew, B. and Alvarez, M. (2020). BisqueRNA: Decomposition of Bulk Expression with Single-Cell Sequencing. R package version, 1.0.5.
- Karimnezhad, A. (2022) More accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Bioinformatics Advances, Volume 2, Issue 1, 2022, vbac049. <https://doi.org/10.1093/bioadv/vbac049>

example_data

Example data

Description

An example data containing synthetic bulk and single-cell datasets. This example illustrates how to build ExpressionSets and run [CSCD](#).

Usage

```
example_data
```

Format

A list. Slot **bulk.matrix** contains a sample bulk data matrix with 100 rows (genes) and 5 columns (individuals). Slot **sc.counts.matrix** contains a sample single-cell data matrix with 100 rows (genes) and 20 columns (a combination of cells assigned to 4 different cell types and 5 individuals). Slot **individual.labels** contains individual labels in the single-cell data. Slot **cell.type.labels** contains cell-type labels in the single-cell data. Slot **sample.ids** contains sample ids in the single-cell data. Note that individual.labels and cell.types should be in the same order as in sample.ids.

Examples

```

# Load example data.
data(example_data)

# Build ExpressionSet with bulk data.
bulk.eset <- Biobase::ExpressionSet(assayData = example_data$bulk.matrix)

# Build ExpressionSet with single-cell data.
sc.counts.matrix=example_data$sc.counts.matrix
individual.labels=example_data$individual.labels
cell.type.labels=example_data$cell.type.labels
sample.ids <- colnames(sc.counts.matrix)
# individual.labels and cell.types should be in the same order as in sample.ids.
sc.pheno <- data.frame(check.names=FALSE, check.rows=FALSE,
                      stringsAsFactors=FALSE, row.names=sample.ids,
                      SubjectName=individual.labels, cellType=cell.type.labels)
sc.meta <- data.frame(labelDescription=c("SubjectName", "cellType"),
                     row.names=c("SubjectName", "cellType"))
sc.pdata <- new("AnnotatedDataFrame", data=sc.pheno, varMetadata=sc.meta)
sc.eset <- Biobase::ExpressionSet(assayData=sc.counts.matrix, phenoData=sc.pdata)

# Run CSCD on the example data.
analysis <- CSCD(bulk.eset=bulk.eset, sc.eset= sc.eset,
               min.p=0.3, markers=NULL, cell.types="cellType",
               subj.names="SubjectName", verbose=TRUE)

# Estimated cell-type proportions.
analysis$bulk.props

# Cell-type proportions estimated directly by counting single-cell data.
analysis$sc.props

# The covariance based transformed bulk expression used for decomposition.
analysis$transformed.bulk.

# Genes used in the decomposition.
analysis$genes.used

# Euclidean norm of the residuals for each individual's proportion estimates.
analysis$rnorm

```

Index

* **datasets**

example_data, [3](#)

BisqueRNA::ReferenceBasedDecomposition(),
[2](#)

CSCD, [2](#), [3](#)

example_data, [2](#), [3](#)

Seurat::FindAllMarkers(), [2](#)