Package ‘Dino’

March 20, 2024

Type Package

Title Normalization of Single-Cell mRNA Sequencing Data

Version 1.8.0

biocViews Software, Normalization, RNASeq, SingleCell, Sequencing,
       GeneExpression, Transcriptomics, Regression, CellBasedAssays

Description Dino normalizes single-cell, mRNA sequencing data to correct for
technical variation, particularly sequencing depth, prior to downstream
analysis. The approach produces a matrix of corrected expression for which
the dependency between sequencing depth and the full distribution of
normalized expression; many existing methods aim to remove only the
dependency between sequencing depth and the mean of the normalized
expression. This is particularly useful in the context of highly sparse
datasets such as those produced by 10X genomics and other unique
molecular identifier (UMI) based microfluidics protocols for which the
depth-dependent proportion of zeros in the raw expression data can
otherwise present a challenge.

Depends R (>= 4.0.0)

License GPL-3

Encoding UTF-8

LazyData false

RoxygenNote 7.1.1

Suggests testthat (>= 2.1.0), knitr, rmarkdown, BiocStyle, devtools,
ggplot2, gridExtra, ggpubr, grid, magick, hexbin

VignetteBuilder knitr

Imports BiocParallel, BiocSingular, SummarizedExperiment,
       SingleCellExperiment, S4Vectors, Matrix, Seurat, matrixStats,
       parallel, scran, grDevices, stats, methods


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

Normalize scRNAseq data

Description

Dino removes cell-to-cell variation in observed counts due to the effects of sequencing depth from single-cell mRNA sequencing experiments. Dino was particularly designed with UMI based protocols in mind, but is applicable to non-UMI based chemistries in the library preparation stage of sequencing.

Usage

Dino(counts, nCores = 2, prec = 3, minNZ = 10,
    nSubGene = 1e4, nSubCell = 1e4, depth = NULL, slope = NULL,
    minSlope = 1/2, maxSlope = 2, clusterSlope = TRUE,
    returnMeta = FALSE, doRQS = FALSE,
    emPar = list(maxIter = 100, tol = 0.1, conPar = 15, maxK = 100), ...)

Arguments

counts A numeric matrix object of expression counts - usually in dgCMatrix format for memory efficiency. Column names denote cells (samples or droplets) and row names denote genes.

nCores A non-negative integer scalar denoting the number of cores which should be used. Setting nCores to 0 uses all cores as determined by running `parallel::detectCores()`
Dino

prec
A positive integer denoting the number of decimals to which to round depth (if estimated internally via depth = NULL) and normalized counts for computational efficiency.

minNZ
A positive integer denoting the minimum number of non-zero counts for a gene to be normalized by the Dino algorithm. It is recommended to pre-filter the counts matrix such that all genes meet this threshold. Otherwise, genes with fewer than minNZ non-zeros will be scaled by depth for normalization.

nSubGene
A positive integer denoting the number of genes to subset for calculation of slope.

nSubCell
A positive integer denoting the number of samples to subset for calculation of slope and the EM algorithm.

depth
A numeric vector of length equal to the columns of counts. depth denotes a median-centered, log-scale measure of cell-wise sequencing depth. Dino defaults to defining depth as the (within-cell) sum of counts across genes, followed by a log and median-centering transformation.

slope
A numeric scalar denoting the count-depth relationship on the log-log scale. Typical values are close to 1 (implying a unit increase in depth corresponds to a unit increase in expected counts on the log-log scale), but may be higher, particularly in the case of non-UMI protocols. Dino defaults to estimating slope internally.

minSlope
A numeric scalar denoting the minimum slope. Fitted slopes below this value will return a warning and be set to 1

maxSlope
A numeric scalar denoting the maximum slope. Fitted slopes above this value will return a warning and be set to 1

clusterSlope
A logical indicating whether cells should be pre-clustered prior to calculation of slope. Under the default where cells are pre-clustered, cluster is used as a factor in the regression.

returnMeta
A logical indicating whether metadata (sequencing depth and slope) should be returned.

doRQS
A logical indicating how normalization resampling is to be done. By default (F), normalization is done by resampling from the full posterior distribution. Alternately, restricted quantile sampling (RQS) can be performed to enforce stronger preservation of expression ranks in normalized data. Currently RQS is considered experimental.

emPar
A list of parameters to send to the EM algorithm. maxIter denotes the maximum number of model updates. tol denotes the cutoff threshold for reductions in the log likelihood function. conPar denotes the concentration parameter for the resampling. conPar = 1 implies full resampling from the fitted distribution. As conPar increases, the normalized expression converges to the scale-factor normalized values. maxK denotes the maximum number of mixture components in the mixture model.

... Additional parameters to pass to Scran::quickCluster.

Value
Dino by default returns a matrix of normalized expression with identical dimensions as counts. If returnMeta = TRUE, then Dino returns a list of normalized expression, sequencing depth, and slope.
Dino_SCE

Run Dino normalization on a SingleCellExperiment dataset

Description

Dino_SCE is a wrapper simplifying the application of the Dino method to data formatted as a SingleCellExperiment.

Usage

Dino_SCE(SCE, ...)

Arguments

SCE A SingleCellExperiment object with unnormalized count data (eg. raw UMIs) in the assays slot under the name counts.

... Further arguments to pass to Dino

Value

Dino_SCE returns a SingleCellExperiment object using Dino normalized expression in the assays slot under the normcounts name for downstream analysis.

If returnMeta = T is passed to Dino, then depth and slope results are stored in the metadata slot under the names depth and slope respectively.

Author(s)

Jared Brown
**multimodalDat**

**References**


**Examples**

```r
# raw data
data("pbmcSmall")
str(pbmcSmall)

# format as SingleCellExperiment
library(SingleCellExperiment)
pbmc_SCE <- SingleCellExperiment(assays = list("counts" = pbmcSmall))

# Run Dino
pbmc_SCE <- Dino_SCE(pbmc_SCE)
str(pbmc_SCE)
str(normcounts(pbmc_SCE))
```

---

**Description**

This data is used in the vignette to demonstrate the flexibility of the Dino model to smoothly estimate arbitrary latent multimodal expression distributions. These data are intended for internal use only.

**Usage**

```r
data("multimodalDat")
```

**Format**

Object of class "gtable".

**Examples**

```r
data("multimodalDat")
```
Subset of 500 peripheral blood mononuclear cells (PBMCs) from a healthy donor

This dataset derives from the "3k PBMCs from a Healthy Donor" public dataset from 10X Genomics.

data("pbmcSmall")

An object of class "dgCMatrix".

3k PBMCs from a Healthy Donor

data("pbmcSmall")
str(pbmcSmall)

SeuratFromDino is a wrapper simplifying the export of Dino normalized counts to a Seurat object for secondary analysis.

SeuratFromDino(counts, doNorm = TRUE, doLog = TRUE, ...)

A numeric matrix of count data, either raw (eg. UMIs) or normalized expression.

A logical indicating whether to normalize the input counts data before exporting results to a Seurat object. By default, it is assumed that the contents of counts raw expression which should be normalized.

A logical indicating whether normalized counts should be log transformed with a pseudocount of 1 prior to export.

Further arguments to pass to Dino
**unimodalDat**

**Value**

SeuratFromDino returns a Seurat object using Dino normalized and log transformed expression (default) for downstream analysis in the Seurat pipeline.

If `returnMeta = T` is passed to Dino, then `depth` and `slope` results are stored in the Misc slot under the names `depth` and `slope` respectively.

**Author(s)**

Jared Brown

**References**


**Examples**

```r
# raw data
data("pbmcSmall")
str(pbmcSmall)

# run Dino on raw expression matrix, output Seurat object
pbmcSmall_Seurat <- SeuratFromDino(pbmcSmall)
str(pbmcSmall_Seurat)
```

---

**unimodalDat**

*Plot data from simulated expression*

**Description**

This data is used in the vignette to demonstrate the flexibility of the Dino model to smoothly estimate arbitrary latent unimodal expression distributions. These data are intended for internal use only.

**Usage**

data("unimodalDat")

**Format**

Object of class "gtable".

**Examples**

data("unimodalDat")
Index

* datasets
  * multimodalDat, 5
  * pbmcSmall, 6
  * unimodalDat, 7

Dino, 2
Dino_SCE, 4

multimodalDat, 5
pbmcSmall, 6
SeuratFromDino, 6
unimodalDat, 7