Package 'Dino'

November 15, 2024

Type Package

Title Normalization of Single-Cell mRNA Sequencing Data

Version 1.12.0

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 particuarly useful in the context of highly sparse datasets such as those produced by 10X genomics and other uninque 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

URL https://github.com/JBrownBiostat/Dino

BugReports https://github.com/JBrownBiostat/Dino/issues

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

git_branch RELEASE_3_20

git_last_commit ba125c5

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-14

Author Jared Brown [aut, cre] (<https://orcid.org/0000-0002-9151-4386>),
Christina Kendziorski [ctb]

Maintainer Jared Brown <brownj@ds.dfci.harvard.edu>

Contents

`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()`
`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.

Author(s)

Jared Brown

References

Brown, J., Ni, Z., Mohanty, C., Bacher, R. and Kendziorski, C. (2020) "Normalization by distributional resampling of high throughput single-cell RNA-sequencing data." bioRxiv. https://doi.org/10.1101/2020.10.28.359

Examples

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

# run Dino on raw expression matrix
pbmcSmall_Norm <- Dino(pbmcSmall)
str(pbmcSmall_Norm)
```

`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

References

Brown, J., Ni, Z., Mohanty, C., Bacher, R. and Kendziorski, C. (2020). "Normalization by distributional resampling of high throughput single-cell RNA-sequencing data." bioRxiv. https://doi.org/10.1101/2020.10.28.359

Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). "Orchestrating single-cell analysis with Bioconductor." Nat. Methods, 17, 137–145. https://doi.org/10.1038/s41592-019-0654-x

Examples

```
# 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))
```

---

multimodalDat *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 multimodal expression distributions. These data are intended for internal use only.

Usage

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

Format

Object of class "gtable".

Examples

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

---

pbmcSmall *Subset of 500 peripheral blood mononuclear cells (PBMCs) from a healthy donor*

Description

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

Usage

```
data("pbmcSmall")
```

Format

An object of class "dgCMatrix".

Source

3k PBMCs from a Healthy Donor

Examples

```
data("pbmcSmall")
str(pbmcSmall)
```

SeuratFromDino *Create Seurat object from Dino normalized data*

Description

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

Usage

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

Arguments

counts	A numeric matrix of count data, either raw (eg. UMIs) or normalized expression.
doNorm	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.
doLog	A logical indicating whether normalized counts should be log transformed with a psuedocount of 1 prior to export.
...	Further arguments to pass to *Dino*

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

Brown, J., Ni, Z., Mohanty, C., Bacher, R. and Kendziorski, C. (2020). "Normalization by distributional resampling of high throughput single-cell RNA-sequencing data." bioRxiv. https://doi.org/10.1101/2020.10.28.359

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015). "Spatial reconstruction of single-cell gene expression data." Nat. Biotechnol., 33, 495–502. https://doi.org/10.1038/nbt.3192

Examples

```
# 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 uninmodal expression distributions. These data are intended for internal use only.

Usage

```
data("unimodalDat")
```

Format

Object of class "gtable".

Examples

```
data("unimodalDat")
```

Index