Package: DoubletFinder (via r-universe)

August 14, 2024

Type Package

Title DoubletFinder is a suite of tools for identifying doublets in single-cell RNA sequencing data.

Version 2.0.4

Imports fields, KernSmooth, ROCR, parallel

Description DoubletFinder identifies doublets by generating artificial doublets from existing scRNA-seq data and defining which real cells preferentially co-localize with artificial doublets in gene expression space. Other DoubletFinder package functions are used for fitting DoubletFinder to different scRNA-seq datasets. For example, ideal DoubletFinder performance in real-world contexts requires (I) Optimal pK selection and (2) Homotypic doublet proportion estimation. pK selection is achieved using pN-pK parameter sweeps and maxima identification in mean-variance-normalized bimodality coefficient distributions. Homotypic doublet proportion estimation is achieved by finding the sum of squared cell annotation frequencies. For more information, see our Cell Sysmtes paper https://www.cell.com/cell-systems/fulltext/S2405-4712(19)30073-0 and our github https://github.com/chris-mcginnis-ucsf/DoubletFinder

License CC-0

Encoding UTF-8

LazyData true

Repository https://blaserlab.r-universe.dev

RemoteUrl https://github.com/blaserlab/DoubletFinder

RemoteRef HEAD

RemoteSha cff84e94d50b1acbda5878800e30a20bd236dbd0

Contents

`bimodality_coefficient` *bimodality_coefficient*

Description

Internal function to compute bimodality coefficient during BCmvn computation and pK estimation.

Usage

```
bimodality_coefficient(x)
```

Arguments

x	Guassian kernel density estimation representing pANN distribution

Value

Bimodality coefficient value

References

Taken from the 'modes' R package (v0.7)

Examples

```
## Internal to summarizeSweep
gkde <- approxfun(bkde(res.temp$pANN, kernel="normal"))
x <- seq(from=min(res.temp$pANN), to=max(res.temp$pANN), length.out=nrow(res.temp))
sweep.stats$BCreal[i] <- bimodality_coefficient(gkde(x))
```

doubletFinder *doubletFinder*

Description

Core doublet prediction function of the DoubletFinder package. Generates artifical doublets from an existing, pre-processed Seurat object. Real and artificial data are then merged and pre-processed using parameters utilized for the existing Seurat object. PC distance matrix is then computed and used the measure the proportion of artificial nearest neighbors (pANN) for every real cell. pANN is then thresholded according to the number of expected doublets to generate final doublet predictions.

Usage

```
doubletFinder(seu, PCs, pN = 0.25, pK, nExp, reuse.pANN = FALSE, sct = FALSE)
```

Arguments

seu	A fully-processed Seurat object (i.e., After NormalizeData, FindVariableGenes, ScaleData, and RunPCA have all been performed).
PCs	Number of statistically-significant principal components (e.g., as estimated from PC elbow plot)
pN	The number of generated artificial doublets, expressed as a proportion of the merged real-artificial data. Default is set to 0.25, based on observation that DoubletFinder performance is largely pN-invariant (see McGinnis, Murrow and Gartner 2019, Cell Systems).
pK	The PC neighborhood size used to compute pANN, expressed as a proportion of the merged real-artificial data. No default is set, as pK should be adjusted for each scRNA-seq dataset. Optimal pK values can be determined using mean-variance-normalized bimodality coefficient.
nExp	The total number of doublet predictions produced. This value can best be estimated from cell loading densities into the 10X/Drop-Seq device, and adjusted according to the estimated proportion of homotypic doublets.
reuse.pANN	Seurat metadata column name for previously-generated pANN results. Argument should be set to FALSE (default) for initial DoubletFinder runs. Enables fast adjusting of doublet predictions for different nExp.
sct	Logical representing whether SCTransform was used during original Seurat object pre-processing (default = FALSE).

Value

Seurat object with updated metadata including pANN and doublet classifications.

Examples

```
## Initial run, nExp set to 0.15 Poisson loading estimate (e.g., 1000 total doublet predictions)
nExp_poi <- round(0.15*nrow(seu@meta.data))
seu <- doubletFinder(seu, PCs = 1:10, pN = 0.25, pK = 0.01, nExp = nExp_poi, reuse.pANN = FALSE, sct=FALSE)

## With homotypic adjustment
homotypic.prop <- modelHomotypic(annotations)
nExp_poi.adj <- round(nExp_poi*(1-homotypic.prop))
seu <- doubletFinder(seu, PCs = 1:10, pN = 0.25, pK = 0.01, nExp = nExp_poi.adj, reuse.pANN = "pANN_0.25_0.01_1000",
```

`find.pK` *find.pK*

Description

Computes and visualizes the mean-variance normalized bimodality coefficient (BCmvn) score for each pK value tested during doubletFinder_ParamSweep. Optimal pK for any scRNA-seq data can be manually discerned as maxima in BCmvn distributions. If ground-truth doublet classifications are available, BCmvn is plotted along with mean ROC AUC for each pK.

Usage

```
bcmvn <- find.pK(sweep.stats)
```

Arguments

`sweep.stats`	pN-pK bimodality coefficient dataframe as produced by summarizeSweep.

Value

Dataframe of mean BC, BC variance, and BCmvn scores for each pK value. Includes mean AUC for each pK value if ground-truth doublet classifications are utilized during summarizeSweep.

Examples

```
sweep.list <- paramSweep(seu)
sweep.stats <- summarizeSweep(sweep.list, GT = FALSE)
bcmvn <- find.pK(sweep.stats)
```

kurtosis *kurtosis*

Description

Internal function to compute bimodality coefficient during BCmvn computation and pK estimation.

Usage

```
kurtosis(x)
```

Arguments

x	Guassian kernel density estimation representing pANN distribution

Value

Kurtosis value

References

Taken from the 'modes' R package (v0.7).

Examples

```
## Internal to bimodality_coefficient
sample.excess.kurtosis <- kurtosis(x)
```

modelHomotypic *modelHomotypic*

Description

Leverages user-provided cell annotations to model the proportion of homotypic doublets. Building on the assumption that literature-supported annotations reflect real transcriptional divergence, homotypic doublet proportions are modeled as the sum of squared annotation frequencies.

Usage

```
modelHomotypic(annotations)
```

Arguments

annotations	An nCell-length character vector of annotations.

Value

Numeric proportion of homotypic doublets.

Examples

```
## Initial run, nExp set to Poisson loading estimate (e.g., 913 total doublet predictions)
nExp_poi <- round(0.15*length(seu@cell.names))
seu <- doubletFinder(seu, pN = 0.25, pK = 0.01, nExp = nExp_poi, reuse.pANN = FALSE)

## With homotypic adjustment
homotypic.prop <- modelHomotypic(annotations)
nExp_poi.adj <- round(nExp_poi*(1-homotypic.prop))
seu <- doubletFinder(seu, pN = 0.25, pK = 0.01, nExp = nExp_poi.adj, reuse.pANN = "pANN_0.25_0.01_913")
```

---

parallel_paramSweep *parallel_paramSweep*

Description

Internal parallelization function for paramSweep.

Usage

```
**NOT RUN**
parallel_paramSweep(n, n.real.cells, real.cells, pK, pN, data, orig.commands, PCs, sct)
```

Arguments

n	pN iteration counter.
n.real.cells	Number of real cells. Set automatically during paramSweep_v3.
real.cells	Vector of real cell IDs. Set automatically during paramSweep_v3.
pK	The PC neighborhood size used to compute pANN, expressed as a proportion of the merged real-artificial data. No default is set, as pK should be adjusted for each scRNA-seq dataset. Optimal pK values can be determined using mean-variance-normalized bimodality coefficient.
pN	The number of generated artificial doublets, expressed as a proportion of the merged real-artificial data. Default is set to 0.25, based on observation that DoubletFinder performance is largely pN-invariant (see McGinnis, Murrow and Gartner 2019, Cell Systems).
data	Count matrix. Set automatically during paramSweep_v3.
orig.commands	Count matrix. Set automatically during paramSweep_v3.
PCs	Number of statistically-sigificant PCs. Set according to paramSweep_v3 arguments.
sct	Logical representing whether Seurat object was pre-processed using 'sctransform'. Set according to paramSweep_v3 arguments (default = F).

Value

Parallelization function compatible with mclapply.

Author(s)

Implemented by Nathan Skeene, June 2019.

`paramSweep` *paramSweep*

Description

Performs pN-pK parameter sweeps on a 10,000-cell subset of a pre-processed Seurat object. Will use all cells if Seurat object contains less than 10,000 cells. Results are fed into 'summarizeSweep' and 'find.pK' functions during optimal pK parameter selection workflow. Parameters tested: pN = 0.05-0.3, pK = 0.0005-0.3.

Usage

```
sweep.list <- paramSweep(seu, PCs, sct=FALSE)
```

Arguments

`seu`	A fully-processed Seurat object (i.e., After NormalizeData, FindVariableGenes, ScaleData, RunPCA, and RunTSNE have all been performed).
`PCs`	Number of statistically-significant principal components (e.g., as estimated from PC elbow plot)
`sct`	Logical representing whether SCTransform was used during original Seurat object pre-processing (default = FALSE).
`num.cores`	Number of cores to use for parallelization, default=1.

Value

List of pANN vectors for every pN and pK combination. Output also contains pANN information for artificial doublets.

Examples

```
sweep.list <- paramSweep(seu, PCs = 1:10, sct=FALSE)
sweep.stats <- summarizeSweep(sweep.list, GT = FALSE)
bcmvn <- find.pK(sweep.stats)
```

skewness *skewness*

Description

Internal function to compute skewness during BCmvn computation and pK estimation.

Usage

```
skewness(x)
```

Arguments

x	Guassian kernel density estimation representing pANN distribution

Value

Skewness value

References

Taken from the 'modes' R package (v0.7).

Examples

```
## Internal to bimodality_coefficient
G <- skewness(x)
```

summarizeSweep *summarizeSweep*

Description

Summarizes results from doubletFinder_ParamSweep, computing the bimodality coefficient across pN and pK parameter space. If ground-truth doublet classifications are available, then ROC analysis is performed, enabling optimal DoubletFinder parameter selection.

Usage

```
sweep.stats <- summarizeSweep(sweep.list, GT = FALSE)
sweep.stats <- summarizeSweep(sweep.list, GT = TRUE, GT.calls = classifications)
```

Arguments

sweep.list	List of pANN vectors across pN-pK space, as produced by doubletFinder_ParamSweep.
GT	Logical set to TRUE when ground-truth doublet classifications are available for ROC analysis. Default set to FALSE.
GT.calls	An nCell-length character vector of ground-truth doublet classifications (e.g., "Singlet" or "Doublet") used to gauge performance of logistic regression models trained using pANN vectors during ROC analysis.

Value

Dataframe with bimodality coefficient values at each pN-pK parameter set. If GT = TRUE, dataframe also includes AUC for each pN-pK parameter set computed during ROC analysis.

Examples

```
sweep.list <- paramSweep(seu)
sweep.stats <- summarizeSweep(sweep.list, GT = FALSE)
bcmvn <- find.pK(sweep.stats)
```

Index