Package 'GloScope'

May 10, 2024

Type Package

Title Population-level Representation on scRNA-Seq data

Version 1.2.0

Description This package aims at representing and summarizing the entire single-cell profile of a sample. It allows researchers to perform important bioinformatic analyses at the sample-level such as visualization and quality control. The main functions Estimate sample distribution and calculate
statistical divergence among samples, and visualize the distance matrix through MDS plots.

BugReports https://github.com/epurdom/GloScope/issues

License Artistic-2.0

Encoding UTF-8

Imports utils, stats, MASS, mclust, ggplot2, RANN, FNN, BiocParallel, mvnfast, SingleCellExperiment, rlang

Depends R (>= 4.3.0)

Suggests BiocStyle, testthat (>= 3.0.0), knitr, rmarkdown, zellkonverter

VignetteBuilder knitr

LazyData false

biocViews DataRepresentation, QualityControl, RNASeq, Sequencing, Software, SingleCell

RoxygenNote 7.2.3

Config/testthat/edition 3

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

git_branch RELEASE_3_19

git_last_commit a6d8a5a

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-05-10

Author William Torous [aut, cre] (<https://orcid.org/0000-0001-5668-5510>),
Hao Wang [aut] (<https://orcid.org/0000-0002-0749-474X>),
Elizabeth Purdom [aut],
Boying Gong [aut]

Maintainer William Torous <wtorous@berkeley.edu>

Contents

example_SCE 2
gloscope 3
plotMDS 4

Index **6**

`example_SCE` *SingleCellExperiment containing example inputs to GloScope*

Description

'example_SCE' is a SingleCellExperiment object which contains PCA embeddings and metadata for PBMCs from 20 COVID-infected and healthy control patients. Each sample is reduced to a random subset of 500 cells, for a total of 10,000 cells. The 'colData' slot of the object contains the metadata for each cell, its sample ID and phenotype. The dimensionality reductions slot contains the first 50 PCs, and these embeddings are provided by the authors of "Single-cell multi-omics analysis of the immune response in COVID-19" (Stephenson et al., 2021; doi: 10.1038/s41591-021-01329-2).

'example_SCE_small' is a SingleCellExperiment with the same structure as 'example_SCE', but only containing data from the first five samples. This is a smaller set for examples.

Format

A SingleCellExperiment object with metadata and PCA embeddings

Value

A SingleCellExperiment object

Examples

```
# Code to create the small SCE from the full sample
# Reduction to 5 samples demonstrates data extraction from SCE objects
data(example_SCE)
sample_ids <- SingleCellExperiment::colData(example_SCE)$sample_id
whKeep <- which(sample_ids %in% unique(sample_ids)[seq_len(5)])
example_SCE_small <- SingleCellExperiment::SingleCellExperiment(
 assays=list(counts=matrix(rep(0,2500),ncol=2500)),
 colData=SingleCellExperiment::colData(example_SCE)[whKeep,],
 reducedDims=list("PCA"=SingleCellExperiment::reducedDim(example_SCE,"PCA")[whKeep,]))
```

gloscope *Calculate statistical divergence between all sample pairs*

Description

This function calculates a matrix of pairwise divergences between input samples of single cell data.

Usage

```
gloscope(
  embedding_matrix,
  cell_sample_ids,
  dens = c("GMM", "KNN"),
  dist_mat = c("KL", "JS"),
  r = 10000,
  num_components = seq_len(9),
  k = 50,
  BPPARAM = BiocParallel::SerialParam(),
  prefit_density = NULL,
  return_density = FALSE
)
```

Arguments

embedding_matrix	a matrix of latent embeddings with rows corresponding to cells and columns to dimensions
cell_sample_ids	a vector of the samples IDs each cell comes from. Length must match the number of rows in ‘embedding_matrix‘
dens	the density estimation. One of c("GMM","KNN")
dist_mat	distance metric to calculate the distance. One of c("KL","JS")
r	number of Monte Carlo simulations to generate
num_components	a vector of integers for the number of components to fit GMMS to, default is seq_len(9)
k	number of nearest neighbours for KNN density estimation, default k = 50.
BPPARAM	BiocParallel parameters, default is running in serial. Set random seed with ‘RNGseed‘ argument
prefit_density	a named list of pre-fit ‘densityMclust‘ objects for each sample, default is NULL
return_density	return the GMM parameter list or not (if applicable), default is FALSE

Value

A matrix containing the pairwise divergence or distance between all pairs of samples

Examples

```
# Bring in small example data of single cell embeddings
data(example_SCE_small)
sample_ids <- SingleCellExperiment::colData(example_SCE_small)$sample_id
pca_embeddings <- SingleCellExperiment::reducedDim(example_SCE_small,"PCA")
# Run gloscope on first 10 PCA embeddings
# We use 'KNN' option for speed ('GMM' is slightly slower)
pca_embeddings_subset <- pca_embeddings[,seq_len(10)] # select the first 10 PCs
dist_result <- gloscope(pca_embeddings_subset, sample_ids,
         dens="KNN", BPPARAM = BiocParallel::SerialParam(RNGseed=2))
dist_result
```

`plotMDS` *Plot the multidimensional scaling of the GloScope represenation*

Description

This function creates a multidimensional scaling plot for a set of samples using their GloScope divergence. Each sample's scatter will be color-coded based on their phenotype. The function calls the 'isoMDS' function from the 'MASS' package.

'paletteBig' is a small helper function to create a large color palette for plotting

Usage

```
plotMDS(dist_mat, metadata_df, sample_id, group_id, k = 10)

paletteBig()
```

Arguments

`dist_mat`	The divergence matrix output of 'gloscope()'
`metadata_df`	A data frame contains each sample's metadata. Note this is NOT at the cell-level.
`sample_id`	The column name or index in metadata_df that contains the sample ID
`group_id`	The column name or index in metadata_df that contains the patient condition
`k`	Number of MDS dimension to generate, default = 10

Value

A list containing the MDS embedding and plot of the distance matrix

- mds - A data.frame containing the MDS embedding, with the number of rows equal to the number of samples.
- plot - A ggplot object containing the plot object. 'print' of the object will create a plot.

Examples

```
data(example_SCE_small)
sample_ids <- SingleCellExperiment::colData(example_SCE_small)$sample_id
# Run gloscope on first 10 PCA embeddings
# We use 'KNN' option for speed ('GMM' is slightly slower)
pca_embeddings <- SingleCellExperiment::reducedDim(example_SCE_small,"PCA")
pca_embeddings_subset <- pca_embeddings[,seq_len(10)] # select the first 10 PCs
dist_result <- gloscope(pca_embeddings_subset, sample_ids,
   dens="KNN",
   BPPARAM = BiocParallel::SerialParam(RNGseed=2))
# make a per-sample metadata
sample_metadata <- as.data.frame(unique(SingleCellExperiment::colData(example_SCE_small)[,c(1,2)]))
mds_result <- plotMDS(dist_mat = dist_result, metadata_df = sample_metadata ,
"sample_id", "phenotype",k=2)
mds_result$plot
head(mds_result$mds)
```

Index

example_SCE, 2
example_SCE_small *(*example_SCE*)*, 2

gloscope, 3

paletteBig *(*plotMDS*)*, 4
plotMDS, 4