R documentation

of ‘`generateBulkSamples.Rd`’

September 3, 2020

`generateBulkSamples`

Generate training and test simulated bulk RNA-seq samples.

Description

Generate training and test bulk profiles using the cell composition matrix built by `generateTrainAndTestBulkPro` function. These samples are generated using the assumption that the expression of gene i in sample j is given by the sum of the cell type specific expression X_{ijk} weighted by the proportions of cell type k in the sample determined by the probability matrix. In practice, as described in Torroja et al., 2019, these profiles are generated by the summation of 100 cells from different cell types determined by cell composition matrix. The number of bulk samples is determined by dimensions of cell composition matrix. See `generateTrainAndTestBulkProbMatrix` for details.

Usage

```
generateBulkSamples(
  object,
  type.data = "both",
  file.backend = NULL,
  threads = 1,
  compression.level = NULL,
  verbose = TRUE
)
```

Arguments

`object` `DigitalDLSorter` object with `single.cell.final` and `prob.cell.types` slots.

`type.data` Type of data to generate among 'train', 'test' or 'both' (the last by default).

`file.backend` Valid file path where to save the HDF5 file used as backend. If it is equal to `NULL` (by default), the data are produced and loaded in memory.

`threads` Number of threads used during the generation of bulk samples (2 by default).

`compression.level`
The compression level used if file.backend provided (6 by default). It is an integer value between 0 (no compression) and 9 (highest and slowest compression).

`verbose` Show informative messages during the execution.

Details

digitalDLSorteR allows the use of HDF5 files as back-end for the resulting data using DelayedArray and HDF5Array packages in cases of generating too large bulk expression matrix. This functionality allows you to work without keeping the data loaded in memory, which will be of vital importance during some computationally heavy steps such as neural network training. You must provide a valid file path in file.backend argument to store the resulting file with '.h5' extension. The data will be accessible from R without being loaded into memory. This option slightly slows down execution times, since subsequent transformations of data will be carried out by chunks instead of using all data. We recommend this option due to the large size of the simulated matrices.

Value

A DigitalDLSorter object with bulk.sim slot containing a list with one or two entries (depending on selected type.data argument): 'train' and 'test'. Each entry contains a SummarizedExperiment object with simulated bulk samples in assay slot, sample names in colData slot and feature names in rowData slot.

References

Pagès H, Hickey wcfP, Lun A (2020). DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets. R package version 0.14.1.

Pagès H (2020). HDF5Array: HDF5 backend for DelayedArray objects. R package version 1.16.1.

See Also

generateTrainAndTestBulkProbMatrix ProbMatrixCellTypes

Examples

```
## loading all data in memory
DDLSChung <- generateBulkSamples(
  DDLSChung,
  threads = 2,
  type.data = "both"
)
## Not run:
## using HDF5 as backend
DDLSChung <- generateBulkSamples(
  DDLSChung,
  threads = 2,
  type.data = "both",
  file.backend = "DDLSChung.bulk.sim.h5"
)

## End(Not run)
```

Index