Package 'HDCytoData'

October 14, 2018

Version 1.0.0

Title Data package of high-dimensional cytometry data sets in SummarizedExperiment and flowSet formats

Description Data package containing a set of high-dimensional cytometry data sets saved in SummarizedExperiment and flowSet Bioconductor object formats, including row and column metadata describing samples, cell populations (clusters), and protein markers.

URL https://github.com/lmweber/HDCytoData

BugReports https://github.com/lmweber/HDCytoData/issues

License MIT + file LICENSE

biocViews ExperimentData, ExpressionData, Proteome, FlowCytometryData, Homo_sapiens_Data

Depends ExperimentHub, SummarizedExperiment, flowCore

Imports utils, methods

VignetteBuilder knitr

Suggests BiocStyle, knitr, rmarkdown, FlowSOM, Rtsne, ggplot2

RoxygenNote 6.0.1

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

git_branch RELEASE_3_7

git_last_commit 01a83be

git_last_commit_date 2018-04-30

Date/Publication 2018-10-14

Author Lukas M. Weber [aut, cre], Charlotte Soneson [aut]

Maintainer Lukas M. Weber <lukmweber@gmail.com>

R topics documented:

Bodenmiller_BCR_XL *'Bodenmiller_BCR_XL' data set*

Description

Mass cytometry (CyTOF) data set from Bodenmiller et al. (2012), consisting of paired samples of healthy immune cells in unstimulated and BCR-XL stimulated condition.

Usage

```
Bodenmiller_BCR_XL_SE(metadata = FALSE)
Bodenmiller_BCR_XL_flowSet(metadata = FALSE)
```

Arguments

metadata	logical value indicating whether ExperimentHub metadata (describing the overall data set) should be returned only, or if the whole data set should be loaded. Default = FALSE, which loads the whole data set.

Details

This is a mass cytometry (CyTOF) data set originally from Bodenmiller et al. (2012). The data set consists of paired samples of healthy peripheral blood mononuclear cells (PBMCs), where one sample from each pair was stimulated with B cell receptor / Fc receptor cross-linker (BCR-XL). This creates a strong differential expression signal for several signaling markers in several cell populations, especially B cells. The strongest observed differential signal is for the signaling marker phosphorylated S6 (pS6) in B cells (see Nowicka et al., 2017, Figure 29).

The data set contains 16 samples (8 paired samples); a total of 172,791 cells; and a total of 24 protein markers. The markers consist of 10 'cell type' markers (which can be used to define cell populations or clusters), and 14 'cell state' or signaling markers.

Reference cell population or cluster labels are available from Nowicka et al. (2017), where these were generated using a strategy of expert-guided manual merging of automatically generated clusters from the FlowSOM clustering algorithm (Van Gassen et al., 2015).

The data set is provided in two Bioconductor object formats: SummarizedExperiment and flowSet. In each case, cells are stored in rows, and protein markers in columns (this is the usual format used for flow and mass cytometry data).

For the link{SummarizedExperiment}, row and column meta-data can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains group IDs, patient IDs, sample IDs, and reference population or cluster IDs. The column data contains channel names, protein marker names, and a factor marker_class to identify the class of each protein marker ('cell type' or 'cell state'). In this data set, the cell type markers are used to define cell populations, and the cell state markers define signaling states. The expression values for each cell can be accessed with assay. The expression values are formatted as a single table (i.e. the 16 samples are stacked into a single matrix).

For the flowSet, the row meta-data is stored as additional columns of data within the flowFrame object for each sample. The row meta-data includes group IDs, patient IDs, sample IDs, and reference population or cluster IDs. Note that the factor values are converted to numeric values, since the tables of expression values must be numeric matrices. The column meta-data consists of protein marker names only, which are stored in the column names of the flowFrame object for each sample (since column names cannot contain multiple entries, the marker_class information to identify

cell type and cell state markers cannot be included). The expression values for each cell are stored in the remaining columns of the `flowFrame` object for each sample. The tables of expression values can be accessed with the `exprs` function from the `flowCore` package. The expression values are stored in one table per sample (i.e. one `flowFrame` object per sample, within the overall `flowSet` object).

Prior to performing any analysis, the expression values should be transformed. A standard transformation used for mass cytometry data is the `arcsinh` with `cofactor = 5`. See the vignette for an example.

Raw data files for the original data set are available from Cytobank, experiment 15713 (https://community.cytobank.org/cy additional information is also available from the `Citrus` wiki page at: https://github.com/nolanlab/citrus/wiki/PBMC-Example-1).

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Bodenmiller, B., Zunder, E. R., Finck, R., Chen, T. J., Savig, E. S., Bruggner, R. V., Simonds, E. F., Bendall, S. C., Sachs, K., Krutzik, P. O., and Nolan, G. P. (2012). *Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators.* Nature Biotechnology, 30(9):858-867.

Nowicka, M., Krieg, C., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., Levesque, M. P., and Robinson, M. D. (2017). *CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets.* F1000Research.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). *FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data.* Cytometry Part A, 87A:636-645.

Examples

```
Bodenmiller_BCR_XL_SE()
Bodenmiller_BCR_XL_flowSet()
```

`HDCytoData` *Data package of high-dimensional cytometry data sets*

Description

Data package containing high-dimensional cytometry data sets saved in `SummarizedExperiment` and `flowSet` Bioconductor object formats, hosted on Bioconductor ExperimentHub.

Details

This package contains a set of publicly available high-dimensional flow cytometry and mass cytometry (CyTOF) data sets, which have been formatted into the `SummarizedExperiment` and `flowSet` Bioconductor object formats.

The objects contain the cell-level expression values, as well as row and column meta-data. The row meta-data includes sample IDs, group IDs, and true cell population or cluster labels (where available). The column meta-data includes channel names, protein marker names, and protein marker classes.

These data sets have been used for benchmarking purposes in our previous work and publications, e.g. to evaluate the performance of clustering algorithms. They are provided here in the `SummarizedExperiment` and `flowSet` formats to make them easier to access.

Currently, the package contains the following data sets:

- `Bodenmiller_BCR_XL`

For additional details on each data set, including references and raw data sources, see the help files for the data sets.

For a short example workflow demonstrating how to load the data objects and use them in an analysis workflow, see the package vignette.

The steps to prepare each data object from the raw data files are included in the `make-data` scripts in the directory `inst/scripts`.

Index