Package 'HDCytoData'

October 27, 2026

Version 1.32.1

Title Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats

Description Data package containing a set of publicly available high-dimensional cytometry benchmark datasets, formatted into SummarizedExperiment and flowSet Bioconductor object formats, including all required metadata. Row metadata includes sample IDs, group IDs, patient IDs, reference cell population or cluster labels (where available), and labels identifying 'spiked in' cells (where available). Column metadata includes channel names, protein marker names, and protein marker classes (cell type or cell state).

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

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

License MIT + file LICENSE

biocViews ExperimentHub, ExperimentData, ExpressionData, FlowCytometryData, SingleCellData, Homo_sapiens_Data, ImmunoOncologyData

Depends ExperimentHub, SummarizedExperiment, flowCore

Imports utils, methods

VignetteBuilder knitr

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

RoxygenNote 7.0.0

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

git_branch RELEASE_3_23

git_last_commit ea8560b

git_last_commit_date 2026-07-14

Repository Bioconductor 3.23

Date/Publication 2026-10-27

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

Maintainer Lukas M. Weber <weberlm3@gmail.com>

Contents

Bodenmiller_BCR_XL 2
HDCytoData 4
Krieg_Anti_PD_1 5
Levine_13dim 7
Levine_32dim 9
Mosmann_rare 11
Nilsson_rare 12
Samusik_01 14
Samusik_all 16
Weber_AML_sim 17
Weber_BCR_XL_sim 21

Index **24**

`Bodenmiller_BCR_XL` *'Bodenmiller_BCR_XL' dataset*

Description

Mass cytometry (CyTOF) dataset from Bodenmiller et al. (2012), consisting of 8 paired samples (16 samples) of stimulated (BCR-XL) and unstimulated peripheral blood cells from healthy individuals. This dataset can be used to benchmark differential analysis algorithms used to test for differential states within cell populations.

Usage

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

Arguments

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

Details

This is a mass cytometry (CyTOF) dataset from Bodenmiller et al. (2012), consisting of paired samples of peripheral blood cells from healthy individuals, where one sample from each pair was stimulated with B cell receptor / Fc receptor cross-linker (BCR-XL), and the other sample is the reference. The dataset contains strong differential expression of several signaling markers in several cell populations; one of the strongest effects is differential expression of phosphorylated S6 (pS6) in the population of B cells.

This dataset can be used to benchmark differential analysis algorithms used to test for differential states within cell populations (e.g. differential expression of pS6 in B cells).

There are 8 paired samples (i.e. 16 samples in total), and a total of 172,791 cells. The dataset contains expression levels of 24 protein markers (10 surface lineage markers used to define cell populations or clusters, and 14 intracellular signaling functional markers). The surface markers are classified as 'cell type' markers, and the signaling markers as 'cell state' markers.

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 dataset 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 metadata 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 cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses). The expression values for each cell can be accessed with `assay`. The expression values are formatted as a single table.

For the `flowSet`, the expression values are stored in a separate table for each sample. Each sample is represented by one `flowFrame` object within the overall `flowSet`. The expression values can be accessed with the `exprs` function from the `flowCore` package. Row metadata is stored as additional columns of data within the `flowFrame` for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the `description` slots, which can be accessed with the `description` accessor function for the individual `flowFrames`; this includes filenames, additional sample information, additional marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the `asinh` with `cofactor = 5`.

File sizes: 24.6 MB (`SummarizedExperiment`), 24.8 MB (`flowSet`).

Original source: Bodenmiller et al. (2012): https://www.ncbi.nlm.nih.gov/pubmed/22902532

Original link to raw data (Cytobank, experiment 15713): https://community.cytobank.org/cytobank/experiments/15713/down

Additional information (`Citrus` wiki page): https://github.com/nolanlab/citrus/wiki/PBMC-Example-1

Cell population labels from: Nowicka et al. (2017), v2: https://f1000research.com/articles/6-748/v2

This dataset has previously been used to benchmark algorithms for differential analysis by ourselves and other authors, including Bruggner et al. (2014) (https://www.ncbi.nlm.nih.gov/pubmed/24979804/), Nowicka et al. (2017) (https://f1000research.com/articles/6-748/v2), and Weber et al. (2019) (https://www.ncbi.nlm.nih.gov/pubmed/31098416).

Data files are also available from FlowRepository (FR-FCM-ZYL8): http://flowrepository.org/id/FR-FCM-ZYL8

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Bodenmiller et al. (2012). "Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators." Nature Biotechnology, 30(9), 858-867: https://www.ncbi.nlm.nih.gov/pubmed/22902532

Bruggner et al. (2014), "Automated identification of stratifying signatures in cellular subpopulations." PNAS, 111(26), E2770-E2777: https://www.ncbi.nlm.nih.gov/pubmed/24979804/

Nowicka et al. (2017). "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets." F1000Research, v2: https://f1000research.com/articles/6-748/v2

Van Gassen et al. (2015). "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data." Cytometry Part A, 87A, 636-645: https://www.ncbi.nlm.nih.gov/pubmed/25573116

Weber et al. (2019). "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering." Communications Biology, 2:183: https://www.ncbi.nlm.nih.gov/pubmed/31098416

Examples

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

`HDCytoData` *Data package of high-dimensional cytometry datasets*

Description

Data package containing a collection of high-dimensional cytometry datasets saved in `SummarizedExperiment` and `flowSet` Bioconductor object formats, hosted on Bioconductor ExperimentHub.

Details

Overview

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

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

These datasets have been used in our previous work and publications for benchmarking purposes, e.g. to benchmark clustering algorithms or methods for differential analysis. They are provided here in the `SummarizedExperiment` and `flowSet` formats to make them easier to access.

The package contains the following datasets, which can be grouped into datasets useful for benchmarking either (i) clustering algorithms or (ii) methods for differential analysis.

Clustering:

- `Levine_32dim`
- `Levine_13dim`

- Samusik_01
- Samusik_all
- Nilsson_rare
- Mosmann_rare

Differential analysis:

- Krieg_Anti_PD_1
- Bodenmiller_BCR_XL

Programmatic access to list of datasets

An updated list of all available datasets can also be obtained programmatically using the ExperimentHub accessor functions, as follows. This retrieves a table of metadata from the ExperimentHub database, which includes information such as the ExperimentHub ID, title, and description for each dataset.

```
ehub <- ExperimentHub() # create ExperimentHub instance
ehub <- query(ehub, "HDCytoData") # find HDCytoData datasets
md <- as.data.frame(mcols(ehub)) # retrieve metadata table
```

Additional details

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

For a short tutorial showing how to load the data objects, see the "HDCytoData package" vignette.

Note that flow and mass cytometry datasets should be transformed prior to performing any downstream analyses, such as clustering. Standard transforms include the asinh with cofactor parameter equal to 5 (for mass cytometry data) or 150 (for flow cytometry data).

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

File sizes are listed in the help files for the datasets. The removeCache function from the ExperimentHub package can be used to clear the local download cache.

`Krieg_Anti_PD_1` *'Krieg_Anti_PD_1' dataset*

Description

Mass cytometry (CyTOF) dataset from Krieg et al. (2018), consisting of 20 baseline samples (prior to treatment) of peripheral blood from melanoma skin cancer patients subsequently treated with anti-PD-1 immunotherapy. The samples are split across 2 conditions (non-responders and responders) and 2 batches. This dataset can be used to benchmark differential analysis algorithms used to test for differentially abundant rare cell populations.

Usage

```
Krieg_Anti_PD_1_SE(metadata = FALSE)
Krieg_Anti_PD_1_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a mass cytometry (CyTOF) dataset from Krieg et al. (2018), who used mass cytometry to characterize immune cell subsets in peripheral blood from melanoma skin cancer patients treated with anti-PD-1 immunotherapy. This study found that the frequency of CD14+CD16-HLA-DRhi monocytes in baseline samples (taken from patients prior to treatment) was a strong predictor of survival in response to immunotherapy treatment. In particular, the frequency of a small subpopulation of CD14+CD33+HLA-DRhiICAM-1+CD64+CD141+CD86+CD11c+CD38+PD-L1+CD11b+ monocytes in baseline samples was strongly associated with responder status following immunotherapy treatment. Note that this dataset contains a strong batch effect, due to sample acquisition on two different days (Krieg et al., 2018).

This dataset can be used to benchmark differential analysis algorithms used to test for differentially abundant rare cell populations (i.e. the small subpopulation of CD14+CD33+HLA-DRhiICAM-1+CD64+CD141+CD86+CD11c+CD38+PD-L1+CD11b+ monocytes).

The dataset contains 20 baseline samples (i.e. samples taken prior to treatment), from patients subsequently classified into 2 groups (9 non-responders and 11 responders). Samples are also split across 2 batches ('batch23' and 'batch29'), due to sample acquisition on two different days. The total number of cells is 85,715.

There are 24 'cell type' markers used to characterize cell subpopulations. (One additional cell type marker – CD45 – is also available, but should be excluded from most analyses since almost all cells show very high expression of CD45; so it does not help distinguish subpopulations, and may dominate other signals. Therefore, CD45 has been classified as 'none' in the `marker_info` table.)

The dataset 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 metadata can be accessed with the `rowData` and `colData` accessor functions from the `SummarizedExperiment` package. The row data contains group IDs, batch IDs, and sample 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with `assay`. The expression values are formatted as a single table.

For the `flowSet`, the expression values are stored in a separate table for each sample. Each sample is represented by one `flowFrame` object within the overall `flowSet`. The expression values can be accessed with the `exprs` function from the `flowCore` package. Row metadata is stored as additional columns of data within the `flowFrame` for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the `description` slots, which can be accessed with the `description` accessor function for the individual `flowFrames`; this includes filenames, additional sample information, and additional marker information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the asinh with cofactor = 5.

File sizes: 12.3 MB (SummarizedExperiment), 12.6 MB (flowSet).

Original source: Krieg et al. (2018): https://www.ncbi.nlm.nih.gov/pubmed/29309059

Original link to raw data (FlowRepository, FR-FCM-ZY34): http://flowrepository.org/id/FR-FCM-ZY34

This dataset was previously used to benchmark algorithms for differential analysis in our article, Weber et al. (2019): https://www.ncbi.nlm.nih.gov/pubmed/31098416. (For additional details on the dataset, see Supplementary Note 1: Benchmark datasets.)

Data files are also available from FlowRepository (FR-FCM-ZYL8): http://flowrepository.org/id/FR-FCM-ZYL8

Value

Returns a SummarizedExperiment or flowSet object.

References

Krieg et al. (2018), "High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy." Nature Medicine, 24, 144-153: https://www.ncbi.nlm.nih.gov/pubmed/29309059

Weber et al. (2019). "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering." Communications Biology, 2:183: https://www.ncbi.nlm.nih.gov/pubmed/31098416

Examples

```
Krieg_Anti_PD_1_SE()
Krieg_Anti_PD_1_flowSet()
```

Levine_13dim *'Levine_13dim' dataset*

Description

Mass cytometry (CyTOF) dataset from Levine et al. (2015), containing 13 dimensions (surface protein markers). Manually gated cell population labels are available for 24 populations. Cells are human bone marrow cells from a single healthy donor. This dataset can be used to benchmark clustering algorithms.

Usage

```
Levine_13dim_SE(metadata = FALSE)
Levine_13dim_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 13-dimensional mass cytometry (CyTOF) data set, consisting of expression levels of 13 surface marker proteins. Cell population labels are available for 24 manually gated populations. Cells are human bone marrow cells from a single healthy donor. Manually gated cell population labels were provided by the original authors.

This dataset can be used to benchmark clustering algorithms.

The dataset contains cells from a single patient; a total of 167,044 cells (81,747 manually gated and 85,297 unclassified); 24 manually gated cell population IDs (as well as 'unassigned'); and a total of 13 surface marker proteins.

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains the manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet (note that for this dataset, there is only one sample). The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as additional columns of data within the flowFrame for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the description slots, which can be accessed with the description accessor function for the individual flowFrames; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the asinh with cofactor = 5.

File sizes: 10.0 MB (SummarizedExperiment and flowSet).

Original source: "benchmark data set 1" in Levine et al. (2015): https://www.ncbi.nlm.nih.gov/pubmed/26095251

Original link to raw data: https://www.cytobank.org/cytobank/experiments/46259 (download the FCS files with Actions -> Export -> Download Files -> All FCS Files)

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a SummarizedExperiment or flowSet object.

References

Levine et al. (2015), "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis", Cell, 162, 184-197: https://www.ncbi.nlm.nih.gov/pubmed/26095251

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Levine_13dim_SE()
Levine_13dim_flowSet()
```

Levine_32dim *'Levine_32dim' dataset*

Description

Mass cytometry (CyTOF) dataset from Levine et al. (2015), containing 32 dimensions (surface protein markers). Manually gated cell population labels are available for 14 populations. Cells are human bone marrow cells from 2 healthy donors. This dataset can be used to benchmark clustering algorithms.

Usage

```
Levine_32dim_SE(metadata = FALSE)
Levine_32dim_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 32-dimensional mass cytometry (CyTOF) data set, consisting of expression levels of 32 surface marker proteins. Cell population labels are available for 14 manually gated populations. Cells are human bone marrow cells from 2 healthy donors. Manually gated cell population labels were provided by the original authors.

This dataset can be used to benchmark clustering algorithms.

The dataset contains cells from 2 patients ('H1' and 'H2'); a total of 265,627 cells (104,184 manually gated and 161,443 unclassified); 14 manually gated cell population IDs (as well as 'unassigned'); and a total of 32 surface marker proteins.

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains patient IDs and manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet. The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as additional columns of data within the flowFrame for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the description slots, which can be accessed with the description accessor function for the individual flowFrames; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the asinh with cofactor = 5.

File sizes: 44.2 MB (SummarizedExperiment and flowSet).

Original source: "benchmark data set 2" in Levine et al. (2015): https://www.ncbi.nlm.nih.gov/pubmed/26095251

Original link to raw data: https://www.cytobank.org/cytobank/experiments/46102 (download the .zip file shown under "Exported Files")

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a SummarizedExperiment or flowSet object.

References

Levine et al. (2015), "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis", Cell, 162, 184-197: https://www.ncbi.nlm.nih.gov/pubmed/26095251

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Levine_32dim_SE()
Levine_32dim_flowSet()
```

Mosmann_rare *'Mosmann_rare' dataset*

Description

Flow cytometry dataset from Mosmann et al. (2014), containing 14 dimensions (7 surface protein markers and 7 signaling markers). Manually gated cell population labels are available for one rare population of activated (cytokine-producing) memory CD4 T cells. Cells are human peripheral blood cells exposed to influenza antigens, from a single healthy donor. This dataset can be used to benchmark clustering algorithms for rare cell populations.

Usage

```
Mosmann_rare_SE(metadata = FALSE)
Mosmann_rare_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 14-dimensional flow cytometry dataset, consisting of expression levels of 7 surface protein markers and 7 signaling markers. Cell population labels are available for one rare population of activated (cytokine-producing) memory CD4 T cells. Cells are human peripheral blood cells exposed to influenza antigens, from a single healthy donor.

This dataset can be used to benchmark clustering algorithms for rare cell populations.

The dataset contains cells from a single patient; a total of 396,460 cells (including 109 manually gated cells from the rare population of interest); and a total of 14 protein markers (7 surface protein markers and 7 signaling markers).

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains the manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet (note that for this dataset, there is only one sample). The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as additional columns of data within the flowFrame

for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the `description` slots, which can be accessed with the `description` accessor function for the individual `flowFrames`; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for flow cytometry data is the `asinh` with `cofactor = 150`.

File sizes: 23.1 MB (`SummarizedExperiment`), 23.0 MB (`flowSet`).

Original source: Figure 4 in Mosmann et al. (2014): https://www.ncbi.nlm.nih.gov/pubmed/24532172

Original link to raw data: http://flowrepository.org/id/FR-FCM-ZZ8J (filename: "JMW034-J16OFVQX_G2 0o1 3_D07.fcs"; see Supplementary Information file 3 for full list of filenames)

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Mosmann et al. (2014), "SWIFT - Scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation", Cytometry Part A, 85A, 422-433: https://www.ncbi.nlm.nih.gov/pubmed/24532172

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Mosmann_rare_SE()
Mosmann_rare_flowSet()
```

Nilsson_rare *'Nilsson_rare' dataset*

Description

Flow cytometry dataset from Nilsson et al. (2013), containing 13 dimensions (surface protein markers). Manually gated cell population labels are available for one rare population of hematopoietic stem cells (HSCs). Cells are human bone marrow cells from a single healthy donor. This dataset can be used to benchmark clustering algorithms for rare cell populations.

Usage

```
Nilsson_rare_SE(metadata = FALSE)
Nilsson_rare_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 13-dimensional flow cytometry dataset, consisting of expression levels of 13 surface protein markers. Cell population labels are available for one rare population of hematopoietic stem cells (HSCs). Cells are human bone marrow cells from a single healthy donor.

This dataset can be used to benchmark clustering algorithms for rare cell populations.

The dataset contains cells from a single patient; a total of 44,140 cells (including 358 manually gated cells from the rare population of interest); and a total of 13 surface marker proteins.

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains the manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet (note that for this dataset, there is only one sample). The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as additional columns of data within the flowFrame for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the description slots, which can be accessed with the description accessor function for the individual flowFrames; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for flow cytometry data is the asinh with cofactor = 150.

File sizes: 2.4 MB (SummarizedExperiment and flowSet).

Original source: Figure 2 in Nilsson et al. (2013): https://www.ncbi.nlm.nih.gov/pubmed/23839904

Original link to raw data: http://flowrepository.org/id/FR-FCM-ZZ6L

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a SummarizedExperiment or flowSet object.

References

Nilsson et al. (2013), "Frequency determination of rare populations by flow cytometry: A hematopoietic stem cell perspective", Cytometry Part A, 83A, 721-727: http://www.ncbi.nlm.nih.gov/pubmed/23839904

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Nilsson_rare_SE()
Nilsson_rare_flowSet()
```

`Samusik_01` *'Samusik_01' dataset*

Description

Mass cytometry (CyTOF) dataset from Samusik et al. (2016), containing 39 dimensions (surface protein markers). Manually gated cell population labels are available for 24 populations. The full dataset ('Samusik_all') contains cells from 10 replicate bone marrow samples from C57BL/6J mice (i.e. samples from 10 different mice); this dataset ('Samusik_01') contains the data from sample 01 only. This dataset can be used to benchmark clustering algorithms.

Usage

```
Samusik_01_SE(metadata = FALSE)
Samusik_01_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 39-dimensional mass cytometry (CyTOF) data set, consisting of expression levels of 39 surface marker proteins. Cell population labels are available for 24 manually gated populations. Manually gated cell population labels were provided by the original authors. The full dataset ('Samusik_all') contains cells from 10 replicate bone marrow samples from C57BL/6J mice (i.e. samples from 10 different mice); this dataset ('Samusik_01') contains the data from sample 01 only.

This dataset can be used to benchmark clustering algorithms.

The 'Samusik_01' dataset contains cells from 1 mouse (sample '01'); a total of 86,864 cells (53,173 manually gated and 33,691 unclassified); 24 manually gated cell population IDs (as well as 'unassigned'); and a total of 39 surface marker proteins.

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains sample IDs and manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet (note that for this dataset, there is only one sample). The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as additional columns of data within the flowFrame for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the description slots, which can be accessed with the description accessor function for the individual flowFrames; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the asinh with cofactor = 5.

File sizes: 19.9 MB (SummarizedExperiment), 20.0 MB (flowSet).

Original source: Samusik et al. (2016): https://www.ncbi.nlm.nih.gov/pubmed/27183440

Original link to raw data (.zip file): "https://web.stanford.edu/~samusik/Panorama BM 1-10.zip"

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a SummarizedExperiment or flowSet object.

References

Samusik et al. (2016), "Automated mapping of phenotype space with single-cell data", Nature Methods, 13(6), 493-496: https://www.ncbi.nlm.nih.gov/pubmed/27183440

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Samusik_01_SE()
Samusik_01_flowSet()
```

Samusik_all *'Samusik_all' dataset*

Description

Mass cytometry (CyTOF) dataset from Samusik et al. (2016), containing 39 dimensions (surface protein markers). Manually gated cell population labels are available for 24 populations. This dataset contains cells from 10 replicate bone marrow samples from C57BL/6J mice (i.e. samples from 10 different mice). This dataset can be used to benchmark clustering algorithms.

Usage

```
Samusik_all_SE(metadata = FALSE)
Samusik_all_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a 39-dimensional mass cytometry (CyTOF) data set, consisting of expression levels of 39 surface marker proteins. Cell population labels are available for 24 manually gated populations. Manually gated cell population labels were provided by the original authors. This dataset contains cells from 10 replicate bone marrow samples from C57BL/6J mice (i.e. samples from 10 different mice).

This dataset can be used to benchmark clustering algorithms.

The 'Samusik_all' dataset contains cells from 10 mice (samples '01' to '10); a total of 841,644 cells (514,386 manually gated and 327,258 unclassified); 24 manually gated cell population IDs (as well as 'unassigned'); and a total of 39 surface marker proteins.

The dataset 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 metadata can be accessed with the rowData and colData accessor functions from the SummarizedExperiment package. The row data contains sample IDs and manually gated cell population 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', 'cell state', as well as 'none' for any non protein marker columns that are not needed for downstream analyses; for this dataset, all proteins are cell type markers). The expression values for each cell can be accessed with assay. The expression values are formatted as a single table.

For the flowSet, the expression values are stored in a separate table for each sample. Each sample is represented by one flowFrame object within the overall flowSet. The expression values can be accessed with the exprs function from the flowCore package. Row metadata is stored as

additional columns of data within the `flowFrame` for each sample; note that factor values are converted to numeric values, since the expression tables must be numeric matrices. Channel names are stored in the column names of the expression tables. Additional row and column metadata is stored in the `description` slots, which can be accessed with the `description` accessor function for the individual `flowFrames`; this includes additional sample information (where available), marker information, and cell population information.

Prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the `asinh` with `cofactor = 5`.

File sizes: 192.2 MB (`SummarizedExperiment`), 194.5 MB (`flowSet`).

Original source: Samusik et al. (2016): https://www.ncbi.nlm.nih.gov/pubmed/27183440

Original link to raw data (.zip file): "https://web.stanford.edu/~samusik/Panorama BM 1-10.zip"

This dataset was previously used to benchmark clustering algorithms for high-dimensional cytometry in our article, Weber and Robinson (2016): https://www.ncbi.nlm.nih.gov/pubmed/27992111

Data files are also available from FlowRepository (FR-FCM-ZZPH): http://flowrepository.org/id/FR-FCM-ZZPH

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Samusik et al. (2016), "Automated mapping of phenotype space with single-cell data", Nature Methods, 13(6), 493-496: https://www.ncbi.nlm.nih.gov/pubmed/27183440

Weber and Robinson (2016), "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data", Cytometry Part A, 89A, 1084-1096: https://www.ncbi.nlm.nih.gov/pubmed/27992111

Examples

```
Samusik_all_SE()
Samusik_all_flowSet()
```

`Weber_AML_sim` *'Weber_AML_sim' semi-simulated datasets*

Description

Semi-simulated mass cytometry (CyTOF) datasets from Weber et al. (2019), constructed by computationally 'spiking in' small percentages of AML (acute myeloid leukemia) blast cells into samples of healthy BMMCs (bone marrow mononuclear cells), simulating the phenotype of minimal residual disease (MRD) in AML patients. These datasets can be used to benchmark differential analysis algorithms used to test for differentially abundant rare cell populations. Raw data sourced from Levine et al. (2015), and data generation strategy modified from Arvaniti et al. (2017). See Weber et al. (2019), Supplementary Note 1, for more details.

Usage

```
Weber_AML_sim_main_5pc_SE(metadata = FALSE)
Weber_AML_sim_main_5pc_flowSet(metadata = FALSE)
Weber_AML_sim_main_1pc_SE(metadata = FALSE)
Weber_AML_sim_main_1pc_flowSet(metadata = FALSE)
Weber_AML_sim_main_0.1pc_SE(metadata = FALSE)
Weber_AML_sim_main_0.1pc_flowSet(metadata = FALSE)
Weber_AML_sim_main_blasts_all_SE(metadata = FALSE)
Weber_AML_sim_main_blasts_all_flowSet(metadata = FALSE)
Weber_AML_sim_null_SE(metadata = FALSE)
Weber_AML_sim_null_flowSet(metadata = FALSE)
Weber_AML_sim_random_seeds_5pc_SE(metadata = FALSE)
Weber_AML_sim_random_seeds_5pc_flowSet(metadata = FALSE)
Weber_AML_sim_random_seeds_1pc_SE(metadata = FALSE)
Weber_AML_sim_random_seeds_1pc_flowSet(metadata = FALSE)
Weber_AML_sim_random_seeds_0.1pc_SE(metadata = FALSE)
Weber_AML_sim_random_seeds_0.1pc_flowSet(metadata = FALSE)
Weber_AML_sim_less_distinct_5pc_SE(metadata = FALSE)
Weber_AML_sim_less_distinct_5pc_flowSet(metadata = FALSE)
Weber_AML_sim_less_distinct_1pc_SE(metadata = FALSE)
Weber_AML_sim_less_distinct_1pc_flowSet(metadata = FALSE)
Weber_AML_sim_less_distinct_0.1pc_SE(metadata = FALSE)
Weber_AML_sim_less_distinct_0.1pc_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a set of semi-simulated mass cytometry (CyTOF) datasets, generated for benchmarking purposes in our paper introducing the 'diffcyt' framework (Weber et al., 2019).

The datasets are constructed by computationally 'spiking in' small percentages of AML (acute myeloid leukemia) blast cells into samples of healthy BMMCs (bone marrow mononuclear cells), simulating the phenotype of minimal residual disease (MRD) in AML patients. Blast cells are spiked in at 3 different thresholds of abundance (5%, 1%, and 0.1%), to create multiple simulations with varying levels of difficulty.

These datasets can be used to benchmark differential analysis algorithms used to test for differentially abundant rare cell populations.

The raw data consists of 5 healthy samples (H1-H5), 1 AML (diseased) sample from condition CN (cytogenetically normal), and 1 AML sample from condition CBF (core-binding factor translocation). The dataset is constructed by splitting the healthy samples into 3 parts; one part is kept as the healthy reference condition, and small proportions of either CN or CBF blast cells are spiked into the other two parts to create the semi-simulated MRD conditions CN and CBF. We are then interested in detecting the differentially abundant rare population of either CN or CBF blasts in differential comparisons between CN and healthy, or CBF and healthy.

The dataset contains 16 surface protein markers used to define cell populations. For additional details, including numbers of cells per sample, see Weber et al. (2019), Supplementary Note 1 (in particular Supplementary Tables 1 and 2).

Multiple simulations are available, as described in our paper (Weber et al., 2019). These are stored in the objects listed below.

In each case, the objects are available in both `SummarizedExperiment` and `flowSet` formats, with cells stored in rows, and protein markers in columns (i.e. the usual format for cytometry data). After loading the datasets, they can be inspected using the standard accessor functions for either `SummarizedExperiments` or `flowSets` (e.g. for `SummarizedExperiments`: `rowData`, `colData`, `assays`, and `metadata`).

For the `SummarizedExperiments`: `assays` contain tables of expression values (multiple `assays` for datasets with multiple replicates); `rowData` contains group IDs, patient IDs, sample IDs, and a column identifying spike-in cells; `colData` contains channel names, marker names, and marker classes; and `metadata` contains experiment information and number of cells.

For the `flowSets`: individual `flowFrames` within the `flowSet` contain tables of expression values (multiple `flowFrames` for datasets with multiple replicates); row data is stored as additional columns of numeric values within the expression tables; column data is stored in the `pData(parameters())` slot of the individual `flowFrames`; and additional information (e.g. experiment information, marker information, replicate information, and lookup tables to identify row data values) is stored in the `description()` slot of the `flowFrames`.

Main simulations

Separate files for each threshold of abundance (5%, 1%, and 0.1%), as well additional objects containing all blast cells.

- Weber_AML_sim_main_5pc_SE (31.2 MB)
- Weber_AML_sim_main_5pc_flowSet (31.2 MB)
- Weber_AML_sim_main_1pc_SE (30.4 MB)
- Weber_AML_sim_main_1pc_flowSet (30.4 MB)
- Weber_AML_sim_main_0.1pc_SE (30.2 MB)
- Weber_AML_sim_main_0.1pc_flowSet (30.2 MB)
- Weber_AML_sim_main_blasts_all_SE (11.0 MB)
- Weber_AML_sim_main_blasts_all_flowSet (11.0 MB)

Additional simulations: null simulations

- Weber_AML_sim_null_SE (90.3 MB)
- Weber_AML_sim_null_flowSet (90.3 MB)

Additional simulations: modified random seeds

Separate files for each threshold of abundance (5%, 1%, and 0.1%).

- Weber_AML_sim_random_seeds_5pc_SE (93.6 MB)
- Weber_AML_sim_random_seeds_5pc_flowSet (93.6 MB)
- Weber_AML_sim_random_seeds_1pc_SE (91.1 MB)

- Weber_AML_sim_random_seeds_1pc_flowSet (91.1 MB)
- Weber_AML_sim_random_seeds_0.1pc_SE (90.6 MB)
- Weber_AML_sim_random_seeds_0.1pc_flowSet (90.6 MB)

Additional simulations: 'less distinct' spike-in cells

Separate files for each threshold of abundance (5%, 1%, and 0.1%).

- Weber_AML_sim_less_distinct_5pc_SE (64.1 MB)
- Weber_AML_sim_less_distinct_5pc_flowSet (64.1 MB)
- Weber_AML_sim_less_distinct_1pc_SE (61.1 MB)
- Weber_AML_sim_less_distinct_1pc_flowSet (61.1 MB)
- Weber_AML_sim_less_distinct_0.1pc_SE (60.4 MB)
- Weber_AML_sim_less_distinct_0.1pc_flowSet (60.4 MB)

Note that prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the `asinh` with `cofactor = 5`.

The raw data is sourced from Levine et al. (2015), and the data generation strategy is modified from Arvaniti et al. (2017). See Weber et al. (2019), Supplementary Note 1, for more details.

Original links to raw data from Cytobank:

- all cells (also contains gating scheme for CD34+CD45mid cells, i.e. blasts): https://community.cytobank.org/cytobank/expe

- blasts (repository cloned from the one for 'all cells' above, using the gating scheme for CD34+CD45mid cells; this allows .fcs files for the subset to be exported): https://community.cytobank.org/cytobank/experiments/63534/illustra

Data files are also available from FlowRepository (FR-FCM-ZYL8): http://flowrepository.org/id/FR-FCM-ZYL8

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Arvaniti and Claassen (2017), "Sensitive detection of rare disease-associated cell subsets via representation learning", Nature Communications, 8:14825: https://www.ncbi.nlm.nih.gov/pubmed/28382969

Levine et al. (2015), "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis", Cell, 162, 184-197: https://www.ncbi.nlm.nih.gov/pubmed/26095251

Weber et al. (2019). "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering." Communications Biology, 2:183: https://www.ncbi.nlm.nih.gov/pubmed/31098416

Examples

```
Weber_AML_sim_main_5pc_SE()
Weber_AML_sim_main_5pc_flowSet()
```

Weber_BCR_XL_sim *'Weber_BCR_XL_sim' semi-simulated datasets*

Description

Semi-simulated mass cytometry (CyTOF) datasets from Weber et al. (2019), constructed by randomly splitting unstimulated (reference) samples of PBMCs (peripheral blood mononuclear cells) into two halves, and replacing B cells in one half with stimulated (BCR-XL) B cells from corresponding paired samples. These datasets can be used to benchmark differential analysis algorithms used to test for differential states within cell populations. Raw data sourced from Bodenmiller et al. (2012); cell population labels reproduced from Nowicka et al. (2017). See Weber et al. (2019) Supplementary Note 1, for more details.

Usage

```
Weber_BCR_XL_sim_main_SE(metadata = FALSE)
Weber_BCR_XL_sim_main_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_null_rep1_SE(metadata = FALSE)
Weber_BCR_XL_sim_null_rep1_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_null_rep2_SE(metadata = FALSE)
Weber_BCR_XL_sim_null_rep2_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_null_rep3_SE(metadata = FALSE)
Weber_BCR_XL_sim_null_rep3_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep1_SE(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep1_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep2_SE(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep2_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep3_SE(metadata = FALSE)
Weber_BCR_XL_sim_random_seeds_rep3_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_less_distinct_less_50pc_SE(metadata = FALSE)
Weber_BCR_XL_sim_less_distinct_less_50pc_flowSet(metadata = FALSE)
Weber_BCR_XL_sim_less_distinct_less_75pc_SE(metadata = FALSE)
Weber_BCR_XL_sim_less_distinct_less_75pc_flowSet(metadata = FALSE)
```

Arguments

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

Details

This is a set of semi-simulated mass cytometry (CyTOF) datasets, generated for benchmarking purposes in our paper introducing the 'diffcyt' framework (Weber et al., 2019).

The datasets are constructed by randomly splitting unstimulated (reference) samples of PBMCs (peripheral blood mononuclear cells) into two halves, and replacing B cells in one half with stimulated (BCR-XL) B cells from corresponding paired samples. Strong differential expression signals exist

for several signaling state markers in B cells between the stimulated (BCR-XL) and unstimulated (reference) conditions; in particular phosphorylated S6 (pS6).

These datasets can be used to benchmark differential analysis algorithms used to test for differential states within cell populations.

The raw data consists of 8 paired samples (i.e. 16 samples in total), and a total of 172,791 cells. The dataset contains expression levels of 24 protein markers (10 surface markers used to define cell populations, and 14 intracellular signaling markers). Cell population labels are reproduced from Nowicka et al. (2017). For more details, see Weber et al. (2019), Supplementary Note 1 (in particular Supplementary Tables 3 and 4).

Multiple simulations are available, as described in our paper (Weber et al., 2019). These are stored in the objects listed below.

In each case, the objects are available in both `SummarizedExperiment` and `flowSet` formats, with cells stored in rows, and protein markers in columns (i.e. the usual format for cytometry data). After loading the datasets, they can be inspected using the standard accessor functions for either `SummarizedExperiments` or `flowSets` (e.g. for `SummarizedExperiments`: `rowData`, `colData`, `assays`, and `metadata`).

For the `SummarizedExperiments`: `assays` contain tables of expression values (with multiple objects for datasets with multiple replicates; note that the replicates cannot be combined as multiple `assays` within a single object because each replicate has different row data); `rowData` contains group IDs, patient IDs, sample IDs, cell population IDs, and columns identifying B cells and spike-in cells; `colData` contains channel names, marker names, and marker classes; and `metadata` contains experiment information and number of cells.

For the `flowSets`: individual `flowFrames` within the `flowSet` contain tables of expression values (with multiple `flowSet` objects for datasets with multiple replicates); row data is stored as additional columns of numeric values within the expression tables; column data is stored in the `pData(parameters())` slot of the individual `flowFrames`; and additional information (e.g. experiment information, marker information, replicate information, and lookup tables to identify row data values) is stored in the `description()` slot of the `flowFrames`.

Main simulations

- Weber_BCR_XL_sim_main_SE (12.7 MB)
- Weber_BCR_XL_sim_main_flowSet (12.7 MB)

Additional simulations: null simulations

Separate files for each replicate.

- Weber_BCR_XL_sim_null_rep1_SE (12.7 MB)
- Weber_BCR_XL_sim_null_rep1_flowSet (12.7 MB)
- Weber_BCR_XL_sim_null_rep2_SE (12.7 MB)
- Weber_BCR_XL_sim_null_rep2_flowSet (12.7 MB)
- Weber_BCR_XL_sim_null_rep3_SE (12.7 MB)
- Weber_BCR_XL_sim_null_rep3_flowSet (12.7 MB)

Additional simulations: modified random seeds

Separate files for each replicate.

- Weber_BCR_XL_sim_random_seeds_rep1_SE (12.7 MB)
- Weber_BCR_XL_sim_random_seeds_rep1_flowSet (12.7 MB)
- Weber_BCR_XL_sim_random_seeds_rep2_SE (12.7 MB)
- Weber_BCR_XL_sim_random_seeds_rep2_flowSet (12.7 MB)
- Weber_BCR_XL_sim_random_seeds_rep3_SE (12.7 MB)
- Weber_BCR_XL_sim_random_seeds_rep3_flowSet (12.7 MB)

Additional simulations: 'less distinct' spike-in cells

Separate files for each replicate.

- Weber_BCR_XL_sim_less_distinct_less_50pc_SE (13.1 MB)
- Weber_BCR_XL_sim_less_distinct_less_50pc_flowSet (13.1 MB)
- Weber_BCR_XL_sim_less_distinct_less_75pc_SE (13.1 MB)
- Weber_BCR_XL_sim_less_distinct_less_75pc_flowSet (13.1 MB)

Note that prior to performing any downstream analyses, the expression values should be transformed. A standard transformation used for mass cytometry data is the asinh with `cofactor = 5`.

The raw data is sourced from Bodenmiller et al. (2012), and cell population labels are reproduced from Nowicka et al. (2017). See Weber et al. (2019), Supplementary Note 1, for more details.

Original link to raw data (Cytobank, experiment 15713): https://community.cytobank.org/cytobank/experiments/15713/down

Additional information (`Citrus` wiki page): https://github.com/nolanlab/citrus/wiki/PBMC-Example-1

Data files are also available from FlowRepository (FR-FCM-ZYL8): http://flowrepository.org/id/FR-FCM-ZYL8

Value

Returns a `SummarizedExperiment` or `flowSet` object.

References

Bodenmiller et al. (2012). "Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators." Nature Biotechnology, 30(9), 858-867: https://www.ncbi.nlm.nih.gov/pubmed/22902532

Nowicka et al. (2017). "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets." F1000Research, v2: https://f1000research.com/articles/6-748/v2

Weber et al. (2019). "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering." Communications Biology, 2:183: https://www.ncbi.nlm.nih.gov/pubmed/31098416

Examples

```
Weber_BCR_XL_sim_main_SE()
Weber_BCR_XL_sim_main_flowSet()
```

Index

∗ **datasets**
  Bodenmiller_BCR_XL, 2
  HDCytoData, 4
  Krieg_Anti_PD_1, 5
  Levine_13dim, 7
  Levine_32dim, 9
  Mosmann_rare, 11
  Nilsson_rare, 12
  Samusik_01, 14
  Samusik_all, 16
  Weber_AML_sim, 17
  Weber_BCR_XL_sim, 21

asinh, *3*, *5*, *7, 8*, *10*, *12, 13*, *15*, *17*, *20*, *23*
assay, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*
assays, *19*, *22*

Bodenmiller_BCR_XL, 2, *5*
Bodenmiller_BCR_XL_flowSet
  *(*Bodenmiller_BCR_XL*)*, 2
Bodenmiller_BCR_XL_SE
  *(*Bodenmiller_BCR_XL*)*, 2

colData, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*, *19*, *22*

description, *3*, *6*, *8*, *10*, *12, 13*, *15*, *17*

exprs, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*

flowCore, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*
flowFrame, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*
flowSet, *3, 4*, *6–17*, *19, 20*, *22, 23*

HDCytoData, 4
HDCytoData-package *(*HDCytoData*)*, 4

Krieg_Anti_PD_1, *5*, 5
Krieg_Anti_PD_1_flowSet
  *(*Krieg_Anti_PD_1*)*, 5
Krieg_Anti_PD_1_SE *(*Krieg_Anti_PD_1*)*, 5

Levine_13dim, *4*, 7
Levine_13dim_flowSet *(*Levine_13dim*)*, 7
Levine_13dim_SE *(*Levine_13dim*)*, 7
Levine_32dim, *4*, 9
Levine_32dim_flowSet *(*Levine_32dim*)*, 9
Levine_32dim_SE *(*Levine_32dim*)*, 9

metadata, *19*, *22*
Mosmann_rare, *5*, 11
Mosmann_rare_flowSet *(*Mosmann_rare*)*, 11
Mosmann_rare_SE *(*Mosmann_rare*)*, 11

Nilsson_rare, *5*, 12
Nilsson_rare_flowSet *(*Nilsson_rare*)*, 12
Nilsson_rare_SE *(*Nilsson_rare*)*, 12

rowData, *3*, *6*, *8*, *10, 11*, *13*, *15, 16*, *19*, *22*

Samusik_01, *5*, 14
Samusik_01_flowSet *(*Samusik_01*)*, 14
Samusik_01_SE *(*Samusik_01*)*, 14
Samusik_all, *5*, 16
Samusik_all_flowSet *(*Samusik_all*)*, 16
Samusik_all_SE *(*Samusik_all*)*, 16
SummarizedExperiment, *3, 4*, *6–17*, *19, 20*, *22, 23*

Weber_AML_sim, 17
Weber_AML_sim_less_distinct
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_0.1pc
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_0.1pc_flowSet
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_0.1pc_SE
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_1pc
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_1pc_flowSet
  *(*Weber_AML_sim*)*, 17
Weber_AML_sim_less_distinct_1pc_SE
  *(*Weber_AML_sim*)*, 17

Weber_AML_sim_less_distinct_5pc (Weber_AML_sim), 17
Weber_AML_sim_less_distinct_5pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_less_distinct_5pc_SE (Weber_AML_sim), 17
Weber_AML_sim_main (Weber_AML_sim), 17
Weber_AML_sim_main_0.1pc (Weber_AML_sim), 17
Weber_AML_sim_main_0.1pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_main_0.1pc_SE (Weber_AML_sim), 17
Weber_AML_sim_main_1pc (Weber_AML_sim), 17
Weber_AML_sim_main_1pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_main_1pc_SE (Weber_AML_sim), 17
Weber_AML_sim_main_5pc (Weber_AML_sim), 17
Weber_AML_sim_main_5pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_main_5pc_SE (Weber_AML_sim), 17
Weber_AML_sim_main_blasts_all (Weber_AML_sim), 17
Weber_AML_sim_main_blasts_all_flowSet (Weber_AML_sim), 17
Weber_AML_sim_main_blasts_all_SE (Weber_AML_sim), 17
Weber_AML_sim_null (Weber_AML_sim), 17
Weber_AML_sim_null_flowSet (Weber_AML_sim), 17
Weber_AML_sim_null_SE (Weber_AML_sim), 17
Weber_AML_sim_random_seeds (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_0.1pc (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_0.1pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_0.1pc_SE (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_1pc (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_1pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_1pc_SE (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_5pc (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_5pc_flowSet (Weber_AML_sim), 17
Weber_AML_sim_random_seeds_5pc_SE (Weber_AML_sim), 17
Weber_BCR_XL_sim, 21
Weber_BCR_XL_sim_less_distinct (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_50pc (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_50pc_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_50pc_SE (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_75pc (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_75pc_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_less_distinct_less_75pc_SE (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_main (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_main_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_main_SE (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep1 (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep1_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep1_SE (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep2 (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep2_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep2_SE (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep3 (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep3_flowSet (Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_null_rep3_SE

(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep1
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep1_flowSet
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep1_SE
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep2
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep2_flowSet
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep2_SE
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep3
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep3_flowSet
(Weber_BCR_XL_sim), 21
Weber_BCR_XL_sim_random_seeds_rep3_SE
(Weber_BCR_XL_sim), 21