Package 'scpdata'

October 18, 2022

Type Package

Title Single-Cell Proteomics Data Package

Version 1.4.0

Description The package disseminates mass spectrometry (MS)-based single-cell proteomics (SCP) datasets. The data were collected from published work and formatted using the `scp` data structure. The data sets contain quantitative information at spectrum, peptide and/or protein level for single cells or minute sample amounts.

Depends R (>= 4.1), QFeatures, ExperimentHub

Imports utils, AnnotationHub, SingleCellExperiment, S4Vectors

Suggests scp, magrittr, dplyr, knitr, BiocStyle, BiocCheck, rmarkdown, testthat

biocViews ExperimentData, ExpressionData, ExperimentHub, ReproducibleResearch, MassSpectrometryData, Proteome, SingleCellData

License GPL-2

Encoding UTF-8

LazyData false

VignetteBuilder knitr

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

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

git_branch RELEASE_3_15

git_last_commit c3dcf64

git_last_commit_date 2022-04-26

Date/Publication 2022-10-18

Author Christophe Vanderaa [aut, cre] (<https://orcid.org/0000-0001-7443-5427>), Laurent Gatto [aut] (<https://orcid.org/0000-0002-1520-2268>)

Maintainer Christophe Vanderaa <christophe.vanderaa@uclouvain.be>

R topics documented:

cong2020AC . . . 2
dou2019_boosting . . . 4
dou2019_lysates . . . 6
dou2019_mouse . . . 8
liang2020_hela . . . 10
schoof2021 . . . 12
scpdata . . . 14
specht2019v2 . . . 15
specht2019v3 . . . 17
williams2020_lfq . . . 19
williams2020_tmt . . . 21
zhu2018MCP . . . 22
zhu2018NC_hela . . . 24
zhu2018NC_islets . . . 25
zhu2018NC_lysates . . . 27
zhu2019EL . . . 28

Index **31**

`cong2020AC` *Cong et al. 2020 (Ana. Chem.): HeLa single cells*

Description

Single-cell proteomics using the nanoPOTS sample processing device in combination with ultranarrow-bore (20um i.d.) packed-column LC separations and the Orbitrap Eclipse Tribrid MS. The dataset contains label-free quantitative information at PSM, peptide and protein level. The samples are single Hela cells. Bulk samples (100 and 20 cells) were also included in the experiment to increase the idendtification rate thanks to between-run matching (cf MaxQuant).

Usage

```
cong2020AC
```

Format

A QFeatures object with 9 assays, each assay being a SingleCellExperiment object:

- `100/20 HeLa cells`: 2 assays containing PSM data for a bulk of 100 or 20 HeLa cells, respectively.
- `Blank`: assay containing the PSM data for a blank sample
- `Single cell X`: 4 assays containing PSM data for a single cell. The `X` indicates the replicate number.
- `peptides`: quantitative data for 12590 peptides in 7 samples (all runs combined).
- `proteins`: quantitative data for 1801 proteins in 7 samples (all runs combined).

Sample annotation is stored in `colData(cong2020AC())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: The HeLa cells were diluted and aspired using a microcapillary with a pulled tip.
- **Sample preparation** performed using the nanoPOTs device. Protein extraction using RapiGest (+ DTT) + alkylation (IAA) + Lys-C digestion + cleave RapiGest (formic acid)
- **Separation**: UltiMate 3000 RSLCnano pump with a home-packed nanoLC column (60cm x 20um i.d.; approx. 20 nL/min)
- **Ionization**: ESI (2,000V; Nanospray Flex)
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Eclipse. MS1 settings: accumulation time = 246ms; resolution = 120,000; AGC = 1E6. MS/MS settings depend on quantity. All: AGC = 1E5. 20-100 cels: accumulation time = 246ms; resolution = 120,000. Single cells: accumulation time = 500ms; resolution = 240,000.
- **Data analysis**: MaxQuant (v1.6.3.3) + Excel

Data collection

The PSM, peptide and protein data were collected from the PRIDE repository (accession ID: PXD016921). We downloaded the `evidence.txt` file containing the PSM identification and quantification results. The sample annotation was inferred from the samples names. The data were then converted to a QFeatures object using the scp::readSCP function.

The peptide data were processed similarly from the `peptides.txt` file. The quantitative column names were adpated to match the PSM data. The peptide data were added to QFeatures object and link between the features were stored.

The protein data were similarly processed from the `proteinGroups.txt` file. The quantitative column names were adapted to match the PSM data. The peptide data were added to QFeatures object and link between the features were stored.

Source

All files can be downloaded from the PRIDE repository PXD016921. The source link is: ftp://ftp.pride.ebi.ac.uk/pride/data/ar

References

Cong, Yongzheng, Yiran Liang, Khatereh Motamedchaboki, Romain Huguet, Thy Truong, Rui Zhao, Yufeng Shen, Daniel Lopez-Ferrer, Ying Zhu, and Ryan T. Kelly. 2020. "Improved Single-Cell Proteome Coverage Using Narrow-Bore Packed NanoLC Columns and Ultrasensitive Mass Spectrometry." Analytical Chemistry, January. (link to article).

Examples

```
cong2020AC()
```

`dou2019_boosting` *Dou et al. 2019 (Anal. Chem.): testing boosting ratios*

Description

Single-cell proteomics using nanoPOTS combined with TMT isobaric labeling. It contains quantitative information at PSM and protein level. The cell types are either "Raw" (macrophage cells), "C10" (epihelial cells), or "SVEC" (endothelial cells). Each cell is replicated 2 or 3 times. Each cell type was run using 3 levels of boosting: 0 ng (no boosting), 5 ng or 50 ng. When boosting was applied, 1 reference well and 1 boosting well were added, otherwise 1 empty well was added. Each boosting setting (0ng, 5ng, 50ng) was run in duplicate.

Usage

```
dou2019_boosting
```

Format

A QFeatures object with 7 assays, each assay being a SingleCellExperiment object:

- `Boosting_X_run_Y`: PSM data with 10 columns corresponding to the TMT-10plex channels. The `X` indicates the boosting amount (0ng, 5ng or 50ng) and `Y` indicates the run number (1 or 2).
- `peptides`: peptide data containing quantitative data for 13,462 peptides in 60 samples (run 1 and run 2 combined).
- `proteins`: protein data containing quantitative data for 1436 proteins and 60 samples (all runs combined).

Sample annotation is stored in `colData(dou2019_boosting())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: single-cells from the three murine cell lines were isolated using FACS (BD Influx II cell sorter). Boosting sample were prepared (presumably in bulk) from 1:1:1 mix of the three cell lines.
- **Sample preparation** performed using the nanoPOTs device. Protein extraction (DMM + TCEAP) + alkylation (IAA) + Lys-C digestion + trypsin digestion + TMT-10plex labeling and pooling.
- **Separation**: nanoLC (Dionex UltiMate with an in-house packed 50cm x 30um LC columns; 50nL/min)
- **Ionization**: ESI (2,000V)
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid (MS1 accumulation time = 50ms; MS1 resolution = 120,000; MS1 AGC = 1E6; MS2 accumulation time = 246ms; MS2 resolution = 60,000; MS2 AGC = 1E5)
- **Data analysis**: MS-GF+ + MASIC (v3.0.7111) + RomicsProcessor (custom R package)

Data collection

The PSM data were collected from the MassIVE repository MSV000084110 (see `Source` section). The downloaded files are:

- `Boosting_*ng_run_*_msgfplus.mzid`: the MS-GF+ identification result files.
- `Boosting_*ng_run_*_ReporterIons.txt`: the MASIC quantification result files.

For each batch, the quantification and identification data were combined based on the scan number (common to both data sets). The combined datasets for the different runs were then concatenated feature-wise. To avoid data duplication due to ambiguous matching of spectra to peptides or ambiguous mapping of peptides to proteins, we combined ambiguous peptides to peptides groups and proteins to protein groups. Feature annotations that are not common within a peptide or protein group are are separated by a `;`. The sample annotation table was manually created based on the available information provided in the article. The data were then converted to a QFeatures object using the scp::readSCP function.

We generated the peptide data. First, we removed PSM matched to contaminants or decoy peptides and ensured a 1% FDR. We aggregated the PSM to peptides based on the peptide (or peptide group) sequence(s) using the median PSM instenity. The peptide data for the different runs were then joined in a single assay (see QFeatures::joinAssays), again based on the peptide sequence(s). We then removed the peptide groups. Links between the peptide and the PSM data were created using QFeatures::addAssayLink. Note that links between PSM and peptide groups are not stored.

The protein data were downloaded from `Supporting information` section from the publisher's website (see `Sources`). The data is supplied as an Excel file `ac9b03349_si_004.xlsx`. The file contains 7 sheets from which we took the 2nd, 4th and 6th sheets (named `01 - No Boost raw data`, `03 - 5ng boost raw data`, `05 - 50ng boost raw data`, respectively). The sheets contain the combined protein data for the duplicate runs given the boosting amount. We joined the data for all boosting ration based on the protein name and converted the data to a SingleCellExperiment object. We then added the object as a new assay in the QFeatures dataset (containing the PSM data). Links between the proteins and the corresponding PSM were created. Note that links to protein groups are not stored.

Source

The PSM data can be downloaded from the massIVE repository MSV000084110. FTP link: ftp://massive.ucsd.edu/MSV0000

The protein data can be downloaded from the ACS Publications website (Supporting information section).

References

Dou, Maowei, Geremy Clair, Chia-Feng Tsai, Kerui Xu, William B. Chrisler, Ryan L. Sontag, Rui Zhao, et al. 2019. "High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform." Analytical Chemistry, September (link to article).

See Also

dou2019_lysates, dou2019_mouse

Examples

```
dou2019_boosting()
```

`dou2019_lysates` *Dou et al. 2019 (Anal. Chem.): HeLa lysates*

Description

Single-cell proteomics using nanoPOTS combined with TMT multiplexing. It contains quantitative information at PSM and protein level. The samples are commercial Hela lysates diluted to single-cell amounts (0.2 ng). The boosting wells contain the same digest but at higher amount (10 ng).

Usage

```
dou2019_lysates
```

Format

A QFeatures object with 3 assays, each assay being a SingleCellExperiment object:

- `Hela_run_1`: PSM data with 10 columns corresponding to the TMT-10plex channels. Columns hold quantitative information for HeLa lysate samples (either 0, 0.2 or 10ng). This is the data for run 1.
- `Hela_run_1`: PSM data with 10 columns corresponding to the TMT-10plex channels. Columns hold quantitative information for HeLa lysate samples (either 0, 0.2 or 10ng). This is the data for run 2.
- `peptides`: peptide data containing quantitative data for 13,934 peptides in 20 samples (run 1 and run 2 combined).
- `proteins`: protein data containing quantitative data for 1641 proteins in 20 samples (run 1 and run 2 combined).

Sample annotation is stored in `colData(dou2019_lysates())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: commercially available HeLa protein digest (Thermo Scientific).
- **Sample preparation** performed using the nanoPOTs device. Protein extraction (DMM + TCEAP) + alkylation (IAA) + Lys-C digestion + trypsin digestion + TMT-10plex labeling and pooling.
- **Separation**: nanoLC (Dionex UltiMate with an in-house packed 50cm x 30um LC columns; 50nL/min)
- **Ionization**: ESI (2,000V)

- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid (MS1 accumulation time = 50ms; MS1 resolution = 120,000; MS1 AGC = 1E6; MS2 accumulation time = 246ms; MS2 resolution = 60,000; MS2 AGC = 1E5)
- **Data analysis**: MS-GF+ + MASIC (v3.0.7111) + RomicsProcessor (custom R package)

Data collection

The PSM data were collected from the MassIVE repository MSV000084110 (see `Source` section). The downloaded files are:

- `Hela_run_*_msgfplus.mzid`: the MS-GF+ identification result files
- `Hela_run_*_ReporterIons.txt`: the MASIC quantification result files

For each batch, the quantification and identification data were combined based on the scan number (common to both data sets). The combined datasets for the different runs were then concatenated feature-wise. To avoid data duplication due to ambiguous matching of spectra to peptides or ambiguous mapping of peptides to proteins, we combined ambiguous peptides to peptides groups and proteins to protein groups. Feature annotations that are not common within a peptide or protein group are are separated by a `;`. The sample annotation table was manually created based on the available information provided in the article. The data were then converted to a QFeatures object using the scp::readSCP function.

We generated the peptide data. First, we removed PSM matched to contaminants or decoy peptides and ensured a 1% FDR. We aggregated the PSM to peptides based on the peptide (or peptide group) sequence(s) using the median PSM instenity. The peptide data for the different runs were then joined in a single assay (see QFeatures::joinAssays), again based on the peptide sequence(s). We then removed the peptide groups. Links between the peptide and the PSM data were created using QFeatures::addAssayLink. Note that links between PSM and peptide groups are not stored.

The protein data were downloaded from `Supporting information` section from the publisher's website (see `Sources`). The data is supplied as an Excel file `ac9b03349_si_003.xlsx`. The file contains 7 sheets from which we only took the sheet 6 (named `5 - Run 1 and 2 raw data`) with the combined protein data for the two runs. We converted the data to a SingleCellExperiment object and added the object as a new assay in the QFeatures dataset (containing the PSM data). Links between the proteins and the peptides were created. Note that links to protein groups are not stored.

Source

The PSM data can be downloaded from the massIVE repository MSV000084110. FTP link: ftp://massive.ucsd.edu/MSV0000

The protein data can be downloaded from the ACS Publications website (Supporting information section).

References

Dou, Maowei, Geremy Clair, Chia-Feng Tsai, Kerui Xu, William B. Chrisler, Ryan L. Sontag, Rui Zhao, et al. 2019. "High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform." Analytical Chemistry, September (link to article).

See Also

dou2019_mouse, dou2019_boosting

Examples

```
dou2019_lysates()
```

`dou2019_mouse` *Dou et al. 2019 (Anal. Chem.): murine cell lines*

Description

Single-cell proteomics using nanoPOTS combined with TMT isobaric labeling. It contains quantitative information at PSM and protein level. The cell types are either "Raw" (macrophage cells), "C10" (epihelial cells), or "SVEC" (endothelial cells). Out of the 132 wells, 72 contain single cells, corresponding to 24 C10 cells, 24 RAW cells, and 24 SVEC. The other wells are either boosting channels (12), empty channels (36) or reference channels (12). Boosting and reference channels are balanced (1:1:1) mixes of C10, SVEC, and RAW samples at 5 ng and 0.2 ng, respectively. The different cell types where evenly distributed across 4 nanoPOTS chips. Samples were 11-plexed with TMT labeling.

Usage

```
dou2019_mouse
```

Format

A QFeatures object with 13 assays, each assay being a SingleCellExperiment object:

- `Single_Cell_Chip_X_Y`: PSM data with 11 columns corresponding to the TMT channels (see `Notes`). The `X` indicates the chip number (from 1 to 4) and `Y` indicates the row name on the chip (from A to C).
- `peptides`: peptide data containing quantitative data for 15,492 peptides in 132 samples (run 1 and run 2 combined).
- `proteins`: protein data containing quantitative data for 2331 proteins in 132 samples (all runs combined).

Sample annotation is stored in `colData(dou2019_mouse())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: single-cells from the three murine cell lines were isolated using FACS (BD Influx II cell sorter).

- **Sample preparation** performed using the nanoPOTs device. Protein extraction (DMM + TCEAP) + alkylation (IAA) + Lys-C digestion + trypsin digestion + TMT-10plex labeling and pooling.
- **Separation**: nanoLC (Dionex UltiMate with an in-house packed 50cm x 30um LC columns; 50nL/min)
- **Ionization**: ESI (2,000V)
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid (MS1 accumulation time = 50ms; MS1 resolution = 120,000; MS1 AGC = 1E6; MS2 accumulation time = 246ms; MS2 resolution = 60,000; MS2 AGC = 1E5)
- **Data analysis**: MS-GF+ + MASIC (v3.0.7111) + RomicsProcessor (custom R package)

Data collection

The PSM data were collected from the MassIVE repository MSV000084110 (see `Source` section). The downloaded files are:

- `Single_Cell_Chip_*_*_msgfplus.mzid`: the MS-GF+ identification result files.
- `Single_Cell_Chip_*_*_ReporterIons.txt`: the MASIC quantification result files.

For each batch, the quantification and identification data were combined based on the scan number (common to both data sets). The combined datasets for the different runs were then concatenated feature-wise. To avoid data duplication due to ambiguous matching of spectra to peptides or ambiguous mapping of peptides to proteins, we combined ambiguous peptides to peptides groups and proteins to protein groups. Feature annotations that are not common within a peptide or protein group are are separated by a `;`. The sample annotation table was manually created based on the available information provided in the article. The data were then converted to a QFeatures object using the scp::readSCP function.

We generated the peptide data. First, we removed PSM matched to contaminants or decoy peptides and ensured a 1% FDR. We aggregated the PSM to peptides based on the peptide (or peptide group) sequence(s) using the median PSM instenity. The peptide data for the different runs were then joined in a single assay (see QFeatures::joinAssays), again based on the peptide sequence(s). We then removed the peptide groups. Links between the peptide and the PSM data were created using QFeatures::addAssayLink. Note that links between PSM and peptide groups are not stored.

The protein data were downloaded from `Supporting information` section from the publisher's website (see `Sources`). The data is supplied as an Excel file `ac9b03349_si_005.xlsx`. The file contains 7 sheets from which we only took the 2nd (named `01 - Raw sc protein data`) with the combined protein data for the 12 runs. We converted the data to a SingleCellExperiment object and added the object as a new assay in the QFeatures dataset (containing the PSM data). Links between the proteins and the corresponding PSM were created. Note that links to protein groups are not stored.

Note

Although a TMT-10plex labeling is reported in the article, the PSM data contained 11 channels for each run. Those 11th channel contain mostly missing data and are hence assumed to be empty channels.

Source

The PSM data can be downloaded from the massIVE repository MSV000084110. FTP link: ftp://massive.ucsd.edu/MSV0000

The protein data can be downloaded from the ACS Publications website (Supporting information section).

References

Dou, Maowei, Geremy Clair, Chia-Feng Tsai, Kerui Xu, William B. Chrisler, Ryan L. Sontag, Rui Zhao, et al. 2019. "High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform." Analytical Chemistry, September (link to article).

See Also

dou2019_lysates, dou2019_boosting

Examples

```
dou2019_mouse()
```

`liang2020_hela` *Liang et al. 2020 ((Anal. Chem.)): HeLa cells (MaxQuant preprocessing)*

Description

Single-cell proteomics data from HeLa cells using the autoPOTS acquisition workflow. The samples contain either no cells (blanks), 1 cell, 10 cells, 150 cells or 500 cells. Samples containing between 0 and 10 cells are isolated using micro-pipetting while samples containing between 150 and 500 cells were prepared using dilution of a bulk sample.

Usage

```
liang2020_hela
```

Format

A QFeatures object with 17 assays, each assay being a SingleCellExperiment object:

- `HeLa_*`: 15 assays containing PSM data.
- `peptides`: quantitative data for 48705 peptides in 15 samples (all runs are combined).
- `proteins`: quantitative data for 3970 protein groups in 15 samples (all runs combined).

Sample annotation is stored in `colData(liang2020_hela_MQ())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: The HeLa cells come from a commercially available cell line. Samples containing between 0 and 10 cells were isolated using micro-manipulation and the counts were validated using a microscope. Samples containing between 150 and 500 cells were prepared by diluting a bulk sample and the exact counts were evaluated by obtaining phtotmicrographs.
- **Sample preparation** performed using the autoPOTS worflow that relied on the OT-2 pipeting robot. Cell are lysed using sonication. Samples are then processed by successive incubation with DTT (reduction), then IAA (alkylation), then Lys-C and trypsin (protein digestion).
- **Separation**: Samples were injected on the column using a modified Ultimate WPS-3000 TPL autosampler coupled to an UltiMate 3000 RSLCnano pump. The LC column is a home-packed nanoLC column (45cm x 30um; 40nL/min)
- **Ionization**: Nanospray Flex ion source (2,000V)
- **Mass spectrometry**: Orbitrap Exploris 480. MS1 settings: accumulation time = 250 ms (0-10 cells) or 100 ms (150-500 cells); resolution = 120,000; AGC = 100\ duration = 90 s (0-10 cells) or 60 s (150-500 cells) ; accumulation time = 500 ms (0-1 cell), 250 ms (10 cells), 100 ms (150 cells) or 50 ms (500 cells); resolution = 60,000 (0-10 cells) or 30,000 (150-500 cells); AGC = 5E3 (0-1 cells) or 1E4 (10-500 cells).
- **Data analysis**: MaxQuant (v1.6.7.0) and the search database is Swiss-Prot (July 2020).

Data collection

All data were collected from the PRIDE repository (accession ID: PXD021882).

The sample annotations were collected from the methods section and from table S3 in the paper.

The PSM data were found in the `evidence.txt` file. The data were converted to a QFeatures object using the scp::readSCP function.

The peptide data were found in the `peptides.txt` file. The column names holding the quantitative data were adapted to match the sample names in the QFeatures object. The data were then converted to a SingleCellExperiment object and then inserted in the QFeatures object. Links between the PSMs and the peptides were added

A similar procedure was applied to the protein data. The data were found in the `proteinGroups.txt` file. The column names were adapted, the data were converted to a SingleCellExperiment object and then inserted in the QFeatures object. Links between the peptides and the proteins were added

Source

The PSM data can be downloaded from the PRIDE repository PXD021882 The source link is: http://ftp.pride.ebi.ac.uk/pride/data/archive/2020/12/PXD021882/

References

Liang, Yiran, Hayden Acor, Michaela A. McCown, Andikan J. Nwosu, Hannah Boekweg, Nathaniel B. Axtell, Thy Truong, Yongzheng Cong, Samuel H. Payne, and Ryan T. Kelly. 2020. "Fully Automated Sample Processing and Analysis Workflow for Low-Input Proteome Profiling." Analytical Chemistry, December. (link to article).

Examples

```
liang2020_hela()
```

`schoof2021` *Schoof et al. 2021 (Nat. Comm.): acute myeloid leukemia differentiation*

Description

Single-cell proteomics data from OCI-AML8227 cell culture to reconstruct the cellular hierarchy. The data were acquired using TMTpro multiplexing. The samples contain either no cells, single cells, 10 cells (reference channel) 200 cells (booster channel) or are simply empty wells. Single cells are expected to be one of progenitor cells (`PROG`), leukaemia stem cells (`LSC`), CD38- blast cells (`BLAST CD38-`) or CD38+ blast cells (`BLAST CD38+`). Booster are either a known 1:1:1 mix of cells (PROG, LSC and BLAST) or are isolated directly from the bulk sample. Samples were isolated and annotated using flow cytometry.

Usage

```
schoof2021
```

Format

A QFeatures object with 194 assays, each assay being a SingleCellExperiment object:

- `F*`: 192 assays containing PSM quantification data for 16 TMT channels. The quantification data contain signal to noise ratios as computed by Proteome Discoverer.
- `proteins`: quantitative data for 2898 protein groups in 3072 samples (all runs combined). The quantification data contain signal to noise ratios as computed by Proteome Discoverer.
- `logNormProteins`: quantitative data for 2723 protein groups in 2025 single-cell samples. This assay is the protein datasets that was processed by the authors. Dimension reduction and clustering data are also available in the `reducedDims` and `colData` slots, respectively

Sample annotation is stored in `colData(schoof2021())`. The cell type annotation is stored in the `Population` column. The flow cytometry data is also available: FSC-A, FSC-H, FSC-W, SSC-A, SSC-H, SSC-W, APC-Cy7-A (= CD34) and PE-A (= CD38).

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Sample isolation**: cultured AML 8227 cells were stained with anti-CD34 and anti-CD38. The sorting was performed by FACSAria instrument and deposited in 384 well plates.

- **Sample preparation**: cells are lysed using freeze-boil and sonication in a lysis buffer (TFE) that also includes reduction and alkylation reagents (TCEP and CAA), followed by trypsin (protein) and benzonase (DNA) digestion, TMT-16 labeling and quenching, desalting using SOLAμ C18 plate, peptide concentration, pooling and peptide concentration again. The booster channel contains 200 cell equivalents.
- **Liquid chromatography**: peptides are separated using a C18 reverse-phase column (50cm x 75 µm i.d., Thermo EasySpray) combined to a Thermo EasyLC 1200 for 160 minute gradient with a flowrate of 100nl/min.
- **Mass spectrometry**: FAIMSPro interface is used. MS1 setup: resolution 60.000, AGC target of 300%, accumulation of 50ms. MS2 setup: resolution 45.000, AGC target of 150, 300 or 500%, accumulation of 150, 300, 500, or 1000ms.
- **Raw data processing**: Proteome Discoverer 2.4 + Sequest spectral search engine and validation with Percolator

Data collection

All data were collected from the PRIDE repository (accession ID: PXD020586). The data and metadata were extracted from the `SCeptre_FINAL.zip` file.

We performed extensive data wrangling to combine al the metadata available from different files into a single table available using `colData(schoof2021)`.

The PSM data were found in the `bulk_PSMs.txt` file. Contaminants were defined based on the protein accessions listed in `contaminant.txt`. The data were converted to a QFeatures object using the scp::readSCP function.

The protein data were found in the `bulk_Proteins.txt` file. Contaminants were defined based on the protein accessions listed in `contaminant.txt`.The column names holding the quantitative data were adapted to match the sample names in the QFeatures object. Unnecessary feature annotations (such as in which assay a protein is found) were removed. Feature names were created following the procedure in SCeptre: features names are the protein symbol (or accession if missing) and if duplicated symbols are present (protein isoforms), they are made unique by appending the protein accession. Contaminants were defined based on the protein accessions listed in `contaminant.txt`. The data were then converted to a SingleCellExperiment object and inserted in the QFeatures object.

The log-normalized protein data were found in the `bulk.h5ad` file. This dataset was generated by the authors by running the notebook called `bulk.ipynb`. The `bulk.h5ad` was loaded as an `AnnData` object using the `scanpy` Python module. The object was then converted to a `SingleCellExperiment` object using the `zellkonverter` package. The column names holding the quantitative data were adapted to match the sample names in the QFeatures object. The data were then inserted in the QFeatures object.

The script to reproduce the `QFeatures` object is available at `system.file("scripts", "make-data_schoof2021.R", package = "scpdata")`

Source

The PSM and protein data can be downloaded from the PRIDE repository PXD020586 The source link is: https://www.ebi.ac.uk/pride/archive/projects/PXD020586

References

Schoof, Erwin M., Benjamin Furtwängler, Nil Üresin, Nicolas Rapin, Simonas Savickas, Coline Gentil, Eric Lechman, Ulrich auf Dem Keller, John E. Dick, and Bo T. Porse. 2021. "Quantitative Single-Cell Proteomics as a Tool to Characterize Cellular Hierarchies." Nature Communications 12 (1): 745679. (link to article).

Examples

```
schoof2021()
```

`scpdata` *Single-Cell Proteomics Data Package*

Description

The `scpdata` package distributes mass spectrometry-based single-cell proteomics datasets. The datasets were collected from published work and formatted to a standardized data framework. The `scp` frameworks stores the expression data for different MS levels (identified spectrum, peptide, or protein) in separate assays. Each assay is an object of class SingleCellExperiment that allows easy integration with state-of-the-art single-cell analysis tools. All assays are contained in a single object of class QFeatures. An overview of the data structure is shown provided in the `scp` package.

The `scpdata()` function returns a summary table with all currently available datasets in the package. More information about the data content and the data collection can be found in the corresponding manual pages.

Usage

```
scpdata()
```

Value

A `DataFrame` table containing a summary of the available datasets.

Author(s)

Christophe Vanderaa

See Also

More information about the data manipulation can be found in the `scp` package.

Examples

```
## List available datasets and their metadata
scpdata()

## Load data using the ExperimentHub interface
hub <- ExperimentHub()

## Download the data set of interest using ExperimentHub indexing
hub[["EH3899"]]
## Download the same data set using the build-in function
specht2019v2()
```

`specht2019v2` *Specht et al. 2019 - SCoPE2 (biorRxiv): macrophages vs monocytes (version 2)*

Description

Single cell proteomics data acquired by the Slavov Lab. This is the version 2 of the data released in December 2019. It contains quantitative information of macrophages and monocytes at PSM, peptide and protein level.

Usage

```
specht2019v2
```

Format

A QFeatures object with 179 assays, each assay being a SingleCellExperiment object:

- Assay 1-63: PSM data for SCoPE2 sets acquired with a TMT-11plex protocol, hence those assays contain 11 columns. Columns hold quantitative information from single-cell channels, carrier channels, reference channels, empty (blank) channels and unused channels.
- Assay 64-177: PSM data for SCoPE2 sets acquired with a TMT-16plex protocol, hence those assays contain 16 columns. Columns hold quantitative information from single-cell channels, carrier channels, reference channels, empty (blank) channels and unused channels.
- `peptides`: peptide data containing quantitative data for 9208 peptides and 1018 single-cells.
- `proteins`: protein data containing quantitative data for 2772 proteins and 1018 single-cells.

The `colData(specht2019v2())` contains cell type annotation and batch annotation that are common to all assays. The description of the `rowData` fields for the PSM data can be found in the `MaxQuant` documentation.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: flow cytometry (BD FACSAria I).
- **Sample preparation** performed using the SCoPE2 protocol. mPOP cell lysis + trypsin digestion + TMT-11plex or 16plex labelling and pooling.
- **Separation**: online nLC (DionexUltiMate 3000 UHPLC with a 25cm x 75um IonOpticksAurora Series UHPLC column; 200nL/min).
- **Ionization**: ESI (2,200V).
- **Mass spectrometry**: Thermo Scientific Q-Exactive (MS1 resolution = 70,000; MS1 accumulation time = 300ms; MS2 resolution = 70,000).
- **Data analysis**: DART-ID + MaxQuant (1.6.2.3).

Data collection

The PSM data were collected from a shared Google Drive folder that is accessible from the SlavovLab website (see `Source` section). The folder contains the following files of interest:

- `ev_updated.txt`: the MaxQuant/DART-ID output file
- `annotation_fp60-97.csv`: sample annotation
- `batch_fp60-97.csv`: batch annotation

We combined the the sample annotation and the batch annotation in a single table. We also formated the quantification table so that columns match with those of the annotation and filter only for single-cell runs. Both table are then combined in a single QFeatures object using the scp::readSCP function.

The peptide data were taken from the Slavov lab directly (`Peptides-raw.csv`). It is provided as a spreadsheet. The data were formated to a SingleCellExperiment object and the sample metadata were matched to the column names (mapping is retrieved after running the SCoPE2 R script) and stored in the `colData`. The object is then added to the QFeatures object (containing the PSM assays) and the rows of the peptide data are linked to the rows of the PSM data based on the peptide sequence information through an `AssayLink` object.

The protein data (`Proteins-processed.csv`) is formated similarly to the peptide data, and the rows of the proteins were mapped onto the rows of the peptide data based on the protein sequence information.

Source

The data were downloaded from the Slavov Lab website via a shared Google Drive folder. The raw data and the quantification data can also be found in the massIVE repository `MSV000083945`: ftp://massive.ucsd.edu/MSV000083945.

References

Specht, Harrison, Edward Emmott, Aleksandra A. Petelski, R. Gray Huffman, David H. Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. 2019. "Single-Cell Mass-Spectrometry Quantifies the Emergence of Macrophage Heterogeneity." bioRxiv. (link to article).

Examples

```
specht2019v2()
```

`specht2019v3` *Specht et al. 2019 - SCoPE2 (biorRxiv): macrophages vs monocytes (version 3)*

Description

Single cell proteomics data acquired by the Slavov Lab. This is the version 3 of the data released in October 2020. It contains quantitative information of macrophages and monocytes at PSM, peptide and protein level.

Usage

```
specht2019v3
```

Format

A QFeatures object with 179 assays, each assay being a SingleCellExperiment object:

- Assay 1-63: PSM data for SCoPE2 sets acquired with a TMT-11plex protocol, hence those assays contain 11 columns. Columns hold quantitative information from single-cell channels, carrier channels, reference channels, empty (blank) channels and unused channels.
- Assay 64-177: PSM data for SCoPE2 sets acquired with a TMT-16plex protocol, hence those assays contain 16 columns. Columns hold quantitative information from single-cell channels, carrier channels, reference channels, empty (blank) channels and unused channels.
- `peptides`: peptide data containing quantitative data for 9208 peptides and 1018 single-cells.
- `proteins`: protein data containing quantitative data for 2772 proteins and 1018 single-cells.

The `colData(specht2019v2())` contains cell type annotation and batch annotation that are common to all assays. The description of the `rowData` fields for the PSM data can be found in the `MaxQuant` documentation.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: flow cytometry (BD FACSAria I).
- **Sample preparation** performed using the SCoPE2 protocol. mPOP cell lysis + trypsin digestion + TMT-11plex or 16plex labeling and pooling.
- **Separation**: online nLC (DionexUltiMate 3000 UHPLC with a 25cm x 75um IonOpticksAurora Series UHPLC column; 200nL/min).
- **Ionization**: ESI (2,200V).
- **Mass spectrometry**: Thermo Scientific Q-Exactive (MS1 resolution = 70,000; MS2 accumulation time = 300ms; MS2 resolution = 70,000).
- **Data analysis**: DART-ID + MaxQuant (1.6.2.3).

Data collection

The PSM data were collected from a shared Google Drive folder that is accessible from the SlavovLab website (see `Source` section). The folder contains the following files of interest:

- `ev_updated_v2.txt`: the MaxQuant/DART-ID output file
- `annotation_fp60-97.csv`: sample annotation
- `batch_fp60-97.csv`: batch annotation

We combined the the sample annotation and the batch annotation in a single table. We also formated the quantification table so that columns match with those of the annotation and filter only for single-cell runs. Both table are then combined in a single QFeatures object using the scp::readSCP function.

The peptide data were taken from the Slavov lab directly (`Peptides-raw.csv`). It is provided as a spreadsheet. The data were formated to a SingleCellExperiment object and the sample metadata were matched to the column names (mapping is retrieved after running the SCoPE2 R script) and stored in the `colData`. The object is then added to the QFeatures object (containing the PSM assays) and the rows of the peptide data are linked to the rows of the PSM data based on the peptide sequence information through an `AssayLink` object.

The protein data (`Proteins-processed.csv`) is formated similarly to the peptide data, and the rows of the proteins were mapped onto the rows of the peptide data based on the protein sequence information.

Note

Since version 2, a serious bug in the data were corrected for TMT channels 12 to 16. Many more cells are therefore contained in the data. Version 2 is maintained for backward compatibility. Although the final version of the article was published in 2021, we have kept `specht2019v3` as the data set name for consistency with the previous data version `specht2019v2`.

Source

The data were downloaded from the Slavov Lab website via a shared Google Drive folder. The raw data and the quantification data can also be found in the massIVE repository `MSV000083945`: ftp://massive.ucsd.edu/MSV000083945.

References

Specht, Harrison, Edward Emmott, Aleksandra A. Petelski, R. Gray Huffman, David H. Perlman, Marco Serra, Peter Kharchenko, Antonius Koller, and Nikolai Slavov. 2021. "Single-Cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity Using SCoPE2." Genome Biology 22 (1): 50. (link to article).

Examples

```
specht2019v3()
```

`williams2020_lfq` *Williams et al. 2020 (Anal. Chem.): MCF10A cell line*

Description

Single-cell label free proteomics data from a MCF10A cell line culture. The data were acquired using a label-free quantification protocole based on the nanoPOTS technology. The objective was to test 2 elution gradients for single-cell applications and to demonstrate successful use of the new nanoPOTS autosampler presented in the article. The samples contain either no cells, single cells, 3 cells, 10 cells 50 cells.

Usage

```
williams2020_lfq
```

Format

A QFeatures object with 9 assays, each assay being a SingleCellExperiment object:

- `peptides_[30 or 60]min_[intensity or LFQ]`: 3 assays containing peptide intensities or LFQ normalized quantifications (see `References`) for either a 30min or a 60 min gradient.
- `proteins_[30 or 60]min_[intensity or iBAQ or LFQ]`: 6 assays containing protein intensities, iBAQ normalized or LFQ normalized quantifications (see `References`) for either a 30min or a 60 min gradient.

Sample annotation is stored in `colData(williams2020_lfq())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Sample isolation**: cultured MCF10A cells were isolated using flow-cytometry based cell sorting and deposit on nanoPOTS microwells
- **Sample preparation**: cells are lysed using using a DDM+DTT lysis buffer. Alkylation was then performed using an IAA solution. Proteins are digested with Lys-C and trypsin followed by acidification with FA. Sample droplets are then dried until LC-MS/MS analysis.
- **Liquid chromatography**: peptides are loaded using the new autosampler described in the paper. Samples are loaded using a a homemade miniature syringe pump. The samples are then desalted and concentrated through a SPE column (4cm x 100µm i.d. packed with 5µm C18) with microflow LC pump. The peptides are then eluted from a long LC column (60cm x 50 µm i.d. packed with 3µm C18) coupled to a nanoflox LC pump at 150nL/mL with either a 30 min or a 60 min gradient.
- **Mass spectrometry**: MS/MS was performed on an Orbitrap Fusion Lumos Tribrid MS coupled to a 2kV ESI. MS1 setup: Orbitrap analyzer at resolution 120.000, AGC target of 1E6, accumulation of 246ms. MS2 setup: ion trap with CID at resolution 60.000, AGC target of 2E4, accumulation of 120ms (50 cells) or 250ms (0-10 cells).
- **Raw data processing**: preprocessing using Maxquant v1.6.2.10 that use Andromeda search engine (with UniProtKB 2016-21-29), MBR and LFQ normalization were enabled.

Data collection

All data were collected from the MASSIVE repository (accession ID: MSV000085230).

The peptide and protein data were extracted from the `Peptides_[...].txt` or `ProteinGroups[...].txt` files, respectively, in the `MCF10A_LC_[30 or 60]minutes` folders.

The tables were duplicated so that peptide intensisities, peptide LFQ, protein intensities, protein LFQ and protein intensities are contained in separate tables. Tables are then converted to SingleCellExperiment objects. Sample annotations were infered from the sample names and from the paper. All data is combined in a QFeatures object. AssayLinks were stored between peptide assays and their corresponding proteins assays based on the leading razor protein (hence only unique peptides are linked to proteins).

The script to reproduce the `QFeatures` object is available at `system.file("scripts", "make-data_williams2020_lfq.R' package = "scpdata")`

Suggestion

See `QFeatures::joinAssays` if you want to join the 30min and 60min assays in a single assay for an integrated analysis.

Source

The PSM and protein data can be downloaded from the MASSIVE repository MSV000085230.

References

Source article: Williams, Sarah M., Andrey V. Liyu, Chia-Feng Tsai, Ronald J. Moore, Daniel J. Orton, William B. Chrisler, Matthew J. Gaffrey, et al. 2020. "Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics." Analytical Chemistry 92 (15): 10588–96. (link to article).

LFQ normalization: Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. "Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ." Molecular & Cellular Proteomics: MCP 13 (9): 2513–26. (link to article).

iBAQ normalization: Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. "Global Quantification of Mammalian Gene Expression Control." Nature 473 (7347): 337–42. (link to article).

Examples

```
williams2020_lfq()
```

`williams2020_tmt` *Williams et al. 2020 (Anal. Chem.): 3 AML cell line*

Description

Single-cell label data from three acute myeloid leukemia cell line culture (MOLM-14, K562, CMK). The data were acquired using a TMT-based quantification protocole and the nanoPOTS technology. The objective was to demonstrate successful use of the new nanoPOTS autosampler presented in the source article. The samples contain either carrier (10 ng), reference (0.2ng), empty or single-cell samples..

Usage

```
williams2020_tmt
```

Format

A QFeatures object with 4 assays, each assay being a SingleCellExperiment object:

- `peptides_[intensity or corrected]`: 2 assays containing peptide reporter ion intensities or corrected reporter ion intensities as computed by MaxQuant.
- `proteins_[intensity or corrected]`: 2 assays containing protein reporter ion intensities or corrected reporter ion intensities as computed by MaxQuant.

Sample annotation is stored in `colData(williams2020_tmt())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Sample isolation**: cultured MOLM-14, K562 or CMK cells were isolated using flow-cytometry based cell sorting and deposit on nanoPOTS microwells
- **Sample preparation**: cells are lysed using using a DDM lysis buffer. Proteins are digested with trypsin followed by TMT labelling and quanching with HA. The samples are then acidified with FA, pooled in a single samples (adding carrier and reference peptide mixtures), and dried until LC-MS/MS analysis.
- **Liquid chromatography**: peptides are loaded using the new autosampler described in the paper. Samples are loaded using a a homemade miniature syringe pump. The samples are then desalted and concentrated through a SPE column (4cm x 100µm i.d. packed with 5µm C18) with microflow LC pump. The peptides are then eluted from a long LC column (60cm x 50 µm i.d. packed with 3µm C18) coupled to a nanoflox LC pump at 150nL/mL (elution time is not expliceted).
- **Mass spectrometry**: MS/MS was performed on an Orbitrap Fusion Lumos Tribrid MS coupled to a 2kV ESI. MS1 setup: Orbitrap analyzer at resolution 120.000, AGC target of 1E6, accumulation of 246ms. MS2 setup: Orbitrap with HCD at resolution 120.000, AGC target of 1E6, accumulation of 246ms.
- **Raw data processing**: preprocessing using Maxquant v1.6.2.10 that use Andromeda search engine (with UniProtKB 2016-21-29).

Data collection

All data were collected from the MASSIVE repository (accession ID: MSV000085230).

The peptide and protein data were extracted from the `Peptides_AML_SingleCell.txt` or `ProteinGroups_AML_SingleCell` files, respectively, in the `AML_SingleCell` folders.

The tables were duplicated so that intensisities and corrected intensities are contained in separate tables. Tables are then converted to SingleCellExperiment objects. Sample annotations were inferred from the sample names, from table S2 and from the Experimental Section of the paper. All data is combined in a QFeatures object. AssayLinks were stored between peptide assays and their corresponding proteins assays based on the leading razor protein (hence only unique peptides are linked to proteins).

The script to reproduce the `QFeatures` object is available at `system.file("scripts", "make-data_williams2020_tmt.R' package = "scpdata")`

Source

The PSM and protein data can be downloaded from the MASSIVE repository MSV000085230.

References

Source article: Williams, Sarah M., Andrey V. Liyu, Chia-Feng Tsai, Ronald J. Moore, Daniel J. Orton, William B. Chrisler, Matthew J. Gaffrey, et al. 2020. "Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics." Analytical Chemistry 92 (15): 10588–96. (link to article).

Examples

```
williams2020_tmt()
```

---

zhu2018MCP *Zhu et al. 2018 (Mol. Cel. Prot.): rat brain laser dissections*

Description

Near single-cell proteomics data of laser captured micro-dissection samples. The samples are 24 brain sections from rat pups (day 17). The slices are 12 um thick squares of either 50, 100, or 200 um width. 5 samples were dissected from the corpus callum (`CC`), 4 samples were dissected from the corpus collosum (`CP`), 13 samples were extracted from the cerebral cortex (`CTX`), and 2 samples are labeled as (`Mix`).

Usage

```
zhu2018MCP
```

Format

A QFeatures object with 4 assays, each assay being a SingleCellExperiment object:

- `peptides`: quantitative information for 13,055 peptides from 24 samples
- `proteins_intensity`: protein intensities for 2,257 proteins from 24 samples
- `proteins_LFQ`: LFQ intensities for 2,257 proteins from 24 samples
- `proteins_iBAQ`: iBAQ values for 2,257 proteins from 24 samples

Sample annotation is stored in `colData(zhu2018MCP())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the original article (see `References`).

- **Cell isolation**: brain patches were collected using laser-capture microdissection (PALM MicroBeam) on flash frozen rat (*Rattus norvergicus*) brain tissues. Note that the samples were stained with H&E before dissection for histological analysis. DMSO is used as sample collection solution
- **Sample preparation** performed using the nanoPOTs device: DMSO evaporation + protein extraction (DMM + DTT) + alkylation (IAA)
  - Lys-C digestion + trypsin digestion.
- **Separation**: nanoLC (Dionex UltiMate with an in-house packed 60cm x 30um LC columns; 50nL/min)
- **Ionization**: ESI (2,000V)
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid (MS1 accumulation time = 246ms; MS1 resolution = 120,000; MS1 AGC = 3E6). The MS/MS settings depend on the sample size, excepted for the AGC = 1E5. 50um (time = 502ms; resolution = 240,000), 100um (time = 246ms; resolution = 120,000), 200um (time = 118ms; resolution = 60,000).
- **Data analysis**: MaxQuant (v1.5.3.30) + Perseus (v1.5.6.0) + Origin Pro 2017

Data collection

The data were collected from the PRIDE repository (accession ID: PXD008844). We downloaded the `MaxQuant_Peptides.txt` and the `MaxQuant_ProteinGroups.txt` files containing the combined identification and quantification results. The sample annotations were inferred from the names of columns holding the quantification data and the information in the article. The peptides data were converted to a SingleCellExperiment object. We split the protein table to separate the three types of quantification: protein intensity, label-free quantitification (LFQ) and intensity based absolute quantification (iBAQ). Each table is converted to a SingleCellExperiment object along with the remaining protein annotations. The 4 objects are combined in a single QFeatures object and feature links are created based on the peptide leading razor protein ID and the protein ID.

Source

The PSM data can be downloaded from the PRIDE repository PXD008844. FTP link ftp://ftp.pride.ebi.ac.uk/pride/data/archiv

References

Zhu, Ying, Maowei Dou, Paul D. Piehowski, Yiran Liang, Fangjun Wang, Rosalie K. Chu, William B. Chrisler, et al. 2018. "Spatially Resolved Proteome Mapping of Laser Capture Microdissected Tissue with Automated Sample Transfer to Nanodroplets." Molecular & Cellular Proteomics: MCP 17 (9): 1864–74 (link to article).

Examples

```
zhu2018MCP()
```

`zhu2018NC_hela` *Zhu et al. 2018 (Nat. Comm.): HeLa titration*

Description

Near single-cell proteomics data of HeLa samples containing different number of cells. There are three groups of cell concentrations: low (10-14 cells), medium (35-45 cells) and high (137-141 cells). The data also contain measures for blanks, HeLa lysates (50 cell equivalent) and 2 cancer cell line lysates (MCF7 and THP1, 50 cell equivalent).

Usage

```
zhu2018NC_hela
```

Format

A QFeatures object with 4 assays, each assay being a SingleCellExperiment object:

- `peptides`: quantitative information for 37,795 peptides from 21 samples
- `proteins_intensity`: protein intensities for 3,984 proteins from 21 samples
- `proteins_LFQ`: LFQ intensities for 3,984 proteins from 21 samples
- `proteins_iBAQ`: iBAQ values for 3,984 proteins from 21 samples

Sample annotation is stored in `colData(zhu2018NC_hela())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the original article (see `References`).

- **Cell isolation**: HeLa cell concentration was adjusted by serial dilution and cell counting was performed manually using an inverted microscope.
- **Sample preparation** performed using the nanoPOTs device. Protein extraction using RapiGest (+ DTT) + alkylation (IAA) + Lys-C digestion + cleave RapiGest (formic acid).
- **Separation**: nanoACQUITY UPLC pump (60nL/min) with an Self-Pack PicoFrit 70cm x 30um LC columns.

- **Ionization**: ESI (1,900V).
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid. MS1 settings: accumulation time = 246ms; resolution = 120,000; AGC = 1E6. MS/MS settings, depend on the sample size, excepted for the AGC = 1E5. Blank and approx. 10 cells (time = 502ms; resolution = 240,000), approx. 40 cells (time = 246ms; resolution = 120,000), approx. 140 cells (time = 118ms; resolution = 60,000).
- **Data analysis**: MaxQuant (v1.5.3.30) + Perseus + OriginLab 2017

Data collection

The data were collected from the PRIDE repository (accession ID: PXD006847). We downloaded the `CulturedCells_peptides.txt` and the `CulturedCells_proteinGroups.txt` files containing the combined identification and quantification results. The sample annotations were inferred from the names of columns holding the quantification data and the information in the article. The peptides data were converted to a SingleCellExperiment object. We split the protein table to separate the three types of quantification: protein intensity, label-free quantitification (LFQ) and intensity based absolute quantification (iBAQ). Each table is converted to a SingleCellExperiment object along with the remaining protein annotations. The 4 objects are combined in a single QFeatures object and feature links are created based on the peptide leading razor protein ID and the protein ID.

Source

The PSM data can be downloaded from the PRIDE repository PXD006847. FTP link: ftp://ftp.pride.ebi.ac.uk/pride/data/arch

References

Zhu, Ying, Paul D. Piehowski, Rui Zhao, Jing Chen, Yufeng Shen, Ronald J. Moore, Anil K. Shukla, et al. 2018. "Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10-100 Mammalian Cells." Nature Communications 9 (1): 882 (link to article).

See Also

The same experiment was conducted on HeLa lysates: zhu2018NC_lysates.

Examples

```
zhu2018NC_hela()
```

`zhu2018NC_islets` *Zhu et al. 2018 (Nat. Comm.): human pancreatic islets*

Description

Near single-cell proteomics data human pancreas samples. The samples were collected from pancreatic tissue slices using laser dissection. The pancreata were obtained from organ donors through the JDRFNetwork for Pancreatic Organ Donors with Diabetes (nPOD) program. The sample come either from control patients (n=9) or from type 1 diabetes (T1D) patients (n=9).

Usage

```
zhu2018NC_islets
```

Format

A QFeatures object with 4 assays, each assay being a SingleCellExperiment object:

- `peptides`: quantitative information for 24,321 peptides from 18 islet samples
- `proteins_intensity`: quantitative information for 3,278 proteins from 18 islet samples
- `proteins_LFQ`: LFQ intensities for 3,278 proteins from 18 islet samples
- `proteins_iBAQ`: iBAQ values for 3,278 proteins from 18 islet samples

Sample annotation is stored in `colData(zhu2018NC_islets())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: The islets were extracted from the pacreatic tissues using laser-capture microdissection.
- **Sample preparation** performed using the nanoPOTs device. Protein extraction using RapiGest (+ DTT) + alkylation (IAA) + Lys-C digestion + cleave RapiGest (formic acid)
- **Separation**: nanoACQUITY UPLC pump with an Self-Pack PicoFrit 70cm x 30um LC columns; 60nL/min)
- **Ionization**: ESI (1,900V)
- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid. MS1 settings: accumulation time = 246ms; resolution = 120,000; AGC = 1E6. MS/MS settings: accumulation time = 118ms; resolution = 60,000; AGC = 1E5.
- **Data analysis**: MaxQuant (v1.5.3.30) + Perseus + OriginLab 2017

Data collection

The data were collected from the PRIDE repository (accession ID: PXD006847). We downloaded the `Islet_t1d_ct_peptides.txt` and the `Islet_t1d_ct_proteinGroups.txt` files containing the combined identification and quantification results. The sample types were inferred from the names of columns holding the quantification data. The peptides data were converted to a SingleCellExperiment object. We split the protein table to separate the three types of quantification: protein intensity, label-free quantitification (LFQ) and intensity based absolute quantification (iBAQ). Each table is converted to a SingleCellExperiment object along with the remaining protein annotations. The 4 objects are combined in a single QFeatures object and feature links are created based on the peptide leading razor protein ID and the protein ID.

Source

The PSM data can be downloaded from the PRIDE repository PXD006847. The source link is: ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2018/01/PXD006847

References

Zhu, Ying, Paul D. Piehowski, Rui Zhao, Jing Chen, Yufeng Shen, Ronald J. Moore, Anil K. Shukla, et al. 2018. "Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10-100 Mammalian Cells." Nature Communications 9 (1): 882 (link to article).

Examples

```
zhu2018NC_islets()
```

`zhu2018NC_lysates` *Zhu et al. 2018 (Nat. Comm.): HeLa lysates*

Description

Near single-cell proteomics data of HeLa lysates at different concentrations (10, 40 and 140 cell equivalent). Each concentration is acquired in triplicate.

Usage

```
zhu2018NC_lysates
```

Format

A QFeatures object with 4 assays, each assay being a SingleCellExperiment object:

- `peptides`: quantitative information for 14,921 peptides from 9 lysate samples
- `proteins_intensity`: quantitative information for 2,199 proteins from 9 lysate samples
- `proteins_LFQ`: LFQ intensities for 2,199 proteins from 9 lysate samples
- `proteins_iBAQ`: iBAQ values for 2,199 proteins from 9 lysate samples

Sample annotation is stored in `colData(zhu2018NC_lysates())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the original article (see `References`).

- **Cell isolation**: HeLas were collected from cell cultures.
- **Sample preparation** performed in bulk (5E5 cells/mL). Protein extraction using RapiGest (+ DTT) + dilution to target concentration + alkylation (IAA) + Lys-C digestion + trypsin digestion + cleave RapiGest (formic acid).
- **Separation**: nanoACQUITY UPLC pump (60nL/min) with an Self-Pack PicoFrit 70cm x 30um LC columns.
- **Ionization**: ESI (1,900V).

- **Mass spectrometry**: Thermo Fisher Orbitrap Fusion Lumos Tribrid. MS1 settings: accumulation time = 246ms; resolution = 120,000; AGC = 1E6. MS/MS settings, depend on the sample size, excepted for the AGC = 1E5. Blank and approx. 10 cells (time = 502ms; resolution = 240,000), approx. 40 cells (time = 246ms; resolution = 120,000), approx. 140 cells (time = 118ms; resolution = 60,000).
- **Data analysis**: MaxQuant (v1.5.3.30) + Perseus + OriginLab 2017.

Data collection

The data were collected from the PRIDE repository (accession ID: PXD006847). We downloaded the `Vail_Prep_Vail_peptides.txt` and the `Vail_Prep_Vail_proteinGroups.txt` files containing the combined identification and quantification results. The sample annotations were inferred from the names of columns holding the quantification data and the information in the article. The peptides data were converted to a SingleCellExperiment object. We split the protein table to separate the three types of quantification: protein intensity, label-free quantitification (LFQ) and intensity based absolute quantification (iBAQ). Each table is converted to a SingleCellExperiment object along with the remaining protein annotations. The 4 objects are combined in a single QFeatures object and feature links are created based on the peptide leading razor protein ID and the protein ID.

Source

The PSM data can be downloaded from the PRIDE repository PXD006847. The source link is: ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2018/01/PXD006847

References

Zhu, Ying, Paul D. Piehowski, Rui Zhao, Jing Chen, Yufeng Shen, Ronald J. Moore, Anil K. Shukla, et al. 2018. "Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10-100 Mammalian Cells." Nature Communications 9 (1): 882 (link to article).

See Also

The same experiment was conducted directly on HeLa cells samples rather than lysates. The data is available in zhu2018NC_hela.

Examples

```
zhu2018NC_lysates()
```

Description

Single-cell proteomics data from chicken utricle acquired to study the hair-cell development. The cells are isolated from peeled utrical epithelium and separated into hair cells (FM1-43 high) and supporting cells (FM1-43 low). The sample contain either 1 cell (n = 28), 3 cells (n = 7), 5 cells (n = 8) or 20 cells (n = 14).

Usage

```
zhu2019EL
```

Format

A QFeatures object with 62 assays, each assay being a SingleCellExperiment object:

- `XYZw`: 60 assays containing PSM data. The sample are annotated as follows. `X` indicates the experiment, either 1 or 2. `Y` indicated the FM1-43 signal, either high (H) or low (L). `Z` indicates the number of cells (0, 1, 3, 5 or 20). `w` indicates the replicate, starting from `a`, it can go up to `j`.
- `peptides`: quantitative data for 3444 peptides in 60 samples (all runs are combined).
- `proteins_intensity`: protein intensities for 840 proteins from 24 samples
- `proteins_iBAQ`: iBAQ values for 840 proteins from 24 samples

Sample annotation is stored in `colData(zhu2019EL())`.

Acquisition protocol

The data were acquired using the following setup. More information can be found in the source article (see `References`).

- **Cell isolation**: The cells were taken from the utricles of E15 chick embryos. Samples were stained with FM1-43FX and the cells were dissociated using enzymatic digestion. Cells were FACS sorted (BD Influx) and split based on their FM1-43 signal, while ensuring no debris, doublets or dead cells are retained.
- **Sample preparation** performed using the nanoPOTs device. Cell lysis and protein extraction and reduction are performed using dodecyl beta-D-maltoside + DTT + ammonium bicarbonate. Protein were then alkylated using IAA. Protein digestion is performed using Lys-C and trypsin. Finally samples acidification is performed using formic acid.
- **Separation**: Dionex UltiMate pump with an C18-Packed column (50cm x 30um; 60nL/min)
- **Ionization**: ESI (2,000V)
- **Mass spectrometry**: Orbitrap Fusion Lumos Tribrid. MS1 settings: accumulation time = 246ms; resolution = 120,000; AGC = 3E6. MS/MS settings: accumulation time = 502ms; resolution = 120,000; AGC = 2E5.
- **Data analysis**: Andromeda & MaxQuant (v1.5.3.30) and the search database is NCBI GRCg6a.

Data collection

All data were collected from the PRIDE repository (accession ID: PXD014256).

The sample annotation information is provided in the `Zhu_2019_chick_single_cell_samples_CORRECTED.xlsx` file. This file was given during a personal discussion and is a corrected version of the annotation table available on the PRIDE repository.

The PSM data were found in the `evidence.txt` (in the `Experiment 1+ 2`) folder. The PSM data were filtered so that it contains only samples that are annotated. The data were then converted to a QFeatures object using the scp::readSCP function.

The peptide data were found in the `peptides.txt` file. The column names holding the quantitative data were adapted to match the sample names in the QFeatures object. The data were then converted to a SingleCellExperiment object and then inserted in the QFeatures object. Links between the PSMs and the peptides were added

A similar procedure was applied to the protein data. The data were found in the `proteinGroups.txt` file. We split the protein table to separate the two types of quantification: summed intensity and intensity based absolute quantification (iBAQ). Both tables are converted to SingleCellExperiment objects and are added to the QFeatures object as well as the `AssayLink` between peptides and proteins.

Source

The PSM data can be downloaded from the PRIDE repository PXD014256. The source link is: ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2019/11/PXD014256

References

Zhu, Ying, Mirko Scheibinger, Daniel Christian Ellwanger, Jocelyn F. Krey, Dongseok Choi, Ryan T. Kelly, Stefan Heller, and Peter G. Barr-Gillespie. 2019. "Single-Cell Proteomics Reveals Changes in Expression during Hair-Cell Development." eLife 8 (November). (link to article).

Examples

```
zhu2019EL()
```

Index

∗ **datasets**
  cong2020AC, 2
  dou2019_boosting, 4
  dou2019_lysates, 6
  dou2019_mouse, 8
  liang2020_hela, 10
  schoof2021, 12
  specht2019v2, 15
  specht2019v3, 17
  williams2020_lfq, 19
  williams2020_tmt, 21
  zhu2018MCP, 22
  zhu2018NC_hela, 24
  zhu2018NC_islets, 25
  zhu2018NC_lysates, 27
  zhu2019EL, 28

AssayLinks, *20*, *22*

cong2020AC, 2

dou2019_boosting, 4, *8*, *10*
dou2019_lysates, *5*, 6, *10*
dou2019_mouse, *5*, *8*, 8

liang2020_hela, 10

QFeatures, *2–30*
QFeatures::addAssayLink, *5*, *7*, *9*
QFeatures::joinAssays, *5*, *7*, *9*

schoof2021, 12
scp::readSCP, *3*, *5*, *7*, *9*, *11*, *13*, *16*, *18*, *30*
scpdata, 14
scpdata-package *(*scpdata*)*, 14
SingleCellExperiment, *2*, *4–30*
specht2019v2, 15
specht2019v3, 17

williams2020_lfq, 19
williams2020_tmt, 21

zhu2018MCP, 22
zhu2018NC_hela, 24, *28*
zhu2018NC_islets, 25
zhu2018NC_lysates, *25*, 27
zhu2019EL, 28