Package 'MsDataHub'

May 1, 2025

Title Mass Spectrometry Data on ExperimentHub

Version 1.8.0

Description The MsDataHub package uses the ExperimentHub infrastructure to distribute raw mass spectrometry data files, peptide spectrum matches or quantitative data from proteomics and metabolomics experiments.

License Artistic-2.0

BugReports https://github.com/RforMassSpectrometry/MsDataHub/issues

URL https://rformassspectrometry.github.io/MsDataHub

Imports ExperimentHub, utils

Suggests ExperimentHubData, DT, BiocStyle, knitr, rmarkdown, testthat (>= 3.0.0), Spectra, mzR, PSMatch, QFeatures (>= 1.13.3)

biocViews ExperimentHubSoftware, MassSpectrometry, Proteomics, Metabolomics

Encoding UTF-8

VignetteBuilder knitr

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Config/testthat/edition 3

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

git_branch RELEASE_3_21

git_last_commit a7636af

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-04-30

Author Laurent Gatto [aut, cre] (ORCID: <https://orcid.org/0000-0002-1520-2268>), Kristina Gomoryova [ctb] (ORCID: <https://orcid.org/0000-0003-4407-3917>), Johannes Rainer [aut] (ORCID: <https://orcid.org/0000-0002-6977-7147>)

Maintainer Laurent Gatto <laurent.gatto@uclouvain.be>

Contents

Ai2025 . . . 2
benchmarkingDIA . . . 3
cdf . . . 3
cptac . . . 4
cRAP . . . 5
MsDataHub . . . 5
PXD000001 . . . 6
Report.Derks2022.plexDIA . . . 7
sciex . . . 7
TripleTOF . . . 8

Index **9**

`Ai2025` *Ai et al (2025) single-cell data*

Description

Single-cell proteomics captures the Proteome Heterogeneity in Human iPSC-Derived Cardiomyocytes and Adult Cardiomyocytes.

Project description (from MassIVE): Human induced pluripotent stem cell (IPSC)-derived cardiomyocytes (iCMs) have become important tools to model cardiovascular diseases and drug toxicology. Despite suggested transcriptomic heterogeneity in both iPSC and iCMs, the cellular proteome heterogeneity is poorly understood. Using cutting-edge single cell proteomics, we quantify the maturation from IPSC to iCMs and observed two distinct populations of iCMs with different metabolism, which recapitulates the single adult cardiomyocyte proteome populations albeit less mature.

The two DIA-NN report files are downloaded from the MassIVE dataset MSV000094438 (doi:10.25345/C5T727S7Q) are redistruted here are:

- Adult cardiomyocyte (aCMs): 299 cells
- iPSC-derived cardiomyocytes (iCMs): 2184 cells

Dataset license: CC0 1.0 Universal (CC0 1.0)

Author(s)

EuBIC 2025 developper meeting SCP hackathon members

References

Ai, Lizhuo, Aleksandra Binek, Vladimir Zhemkov, Jae Hyung Cho, Ali Haghani, Simion Kreimer, Edo Israely, et al. 2025. "Single Cell Proteomics Reveals Specific Cellular Subtypes in Cardiomyocytes Derived from Human iPSCs and Adult Hearts." Mol. Cell. Proteomics, no. 100910 (January): 100910. https://doi.org/10.1016/j.mcpro.2025.100910.

benchmarkingDIA *DIA benchmarking data*

Description

These data were generated based on publicly available DIA benchmarking dataset from Gotti et al. (2021). A subset of raw data, containing "overlapped" in the File.Name were searched using the DIA-NN software, and the resulting report.tsv (here labelled as 'benchmarkingDIA.tsv') is provided.

The dataset contains 8 conditions containing a mix of E.coli and Universal Standard Protein-1 (UPS1) peptides. Per 1 ug of E.coli protein (equal in all samples), UPS1 proteins are diluted to final concentration of 50, 25, 10, 5, 2.5, 1, 0.25 and 0.1 fmol.

Each sample was prepared in 3 replicates, so altogether there are 24 samples in the dataset.

Author(s)

Kristina Gomoryova and Laurent Gatto

References

- Gotti C, Roux-Dalvai F, Joly-Beauparlant C, Mangnier L, Leclercq M, Droit A. Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. J Proteome Res. 2021 Oct 1;20(10):4801-4814. doi: 10.1021/acs.jproteome.1c00490. Epub 2021 Sep 2. PMID: 34472865.

cdf *MS data in CDF format*

Description

This data set represents a single CDF file in (AIA/ANDI) NetCDF format from a larger experiment in which the metabolic consequences of knocking the fatty acid amide hydrolase (FAAH) gene in mice was investigated. The file contains datain centroid mode acquired in positive ion mode from 200-600 m/z and 2500-4500 seconds.

Data file:

- *ko15.CDF* file in NetCDF format.

References

- Saghatelian, A et al. *Assignment of endogenous substrates to enzymes by global metabolite profiling*, Biochemistry, 2004. http://dx.doi.org/10.1021/bi0480335

cptac *CPTAC label-free data*

Description

This case-study is a subset of the data of the 6th study of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) (Paulovich et al. 2010). In this experiment, the authors spiked the Sigma Universal Protein Standard mixture 1 (UPS1) containing 48 different human proteins in a protein background of 60 micro g/micro L Saccharomyces cerevisiae strain BY4741.

Five different spike-in concentrations were used:

- 6A: 0.25 fmol UPS1 proteins/micro L
- 6B: 0.74 fmol UPS1 proteins/micro L
- 6C: 2.22 fmol UPS1 proteins/micro L
- 6D: 6.67 fmol UPS1 proteins/micro L
- 6E: 20 fmol UPS1 proteins/micro L

Three replicates are available for each concentration.

The data were searched with MaxQuant version 1.5.2.8 (Cox et al. 2008) including matching between runs. Detailed search settings were described in Goeminne et al. (2016).

Three files are readily available as tab-delimited spreadsheets:

- cptac_a_b_peptides.txt: triplicates from lab 3 for groupes 6A and 6B.
- cptac_a_b_c_peptides.txt: triplicates from labs 1, 2 and 3 for groupes 6A, 6B and 6C.
- cptac_peptides.txt: triplicates from labs 1, 2, and 3 for all groups.

Author(s)

Laurent Gatto and Lieven Clement

References

- Paulovich, Amanda G, Dean Billheimer, Amy-Joan L Ham, Lorenzo Vega-Montoto, Paul A Rudnick, David L Tabb, Pei Wang, et al. 2010. *Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance.* Mol. Cell. Proteomics 9 (2): 242–54.
- Cox, J, and M Mann. 2008. *MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification.* Nat Biotechnol 26 (12): 1367–72. https://doi.org/10.1038/nbt.1511.
- Goeminne, LJ, Gevaert K and Clement, L. 2016. *Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics*, Mol Cell Proteomics, 15:2 657-668.

cRAP *Contaminant and cRAP databases*

Description

These 3 fasta files are widely used proteomics contaminants. The files are:

1. crap_gpm.fasta: the common Repository of Adventitious Proteins (cRAP) from the Global Proteome Machine (GPM) organisation.
2. crap_ccp.fasta: Cambridge Centre for Proteomics' own cRAP fasta database.
3. crap_maxquant.fasta.gz: MaxQuant's contaminant database.

These files are extracted from the `camprotR` package and described in the cRAP databases vignette (see References).

These files are added to the `MsDataHub` package via the corresponding Zenodo repository to raciliate re-use with minimal dependecies and avoid repeated downloaded using caching.

All credit for compiling the fasta files goes to Charlotte Dawson, maintainer of the `camprotR` package.

Author(s)

Laurent Gatto

References

- cRAP databases vignette: https://cambridgecentreforproteomics.github.io/camprotR/articles/crap.html
- cRPA protein sequences (GPM): https://www.thegpm.org/crap/
- camprotR package: https://cambridgecentreforproteomics.github.io/camprotR/index.html
- Gatto, L. (2025). Proteomics contaminant databases (1.0). Zenodo. https://doi.org/10.5281/zenodo.15115102

MsDataHub *All MsDataHub datasets*

Description

The MsDataHub package provides example mass spectrometry data, peptide spectrum matches or quantitative data from proteomics and metabolomics experiments.

The `MsDataHub()` function returns a `data.frame` with all the annotated datasets provided in the package. For details on these individual datasets, refer to their respective manual pages.

See the vignette and the respective manuals pages for more details about the package and the data themselves.

Usage

```
MsDataHub()
```

Value

A `data.frame` describing the data available in `MsDataHub`.

Author(s)

Laurent Gatto

Examples

```
MsDataHub()
```

PXD000001 *PXD000001 Proteomics Data*

Description

The PXD000001 files are part of the first ProteomeXchange submission (Vizcaíno J.A. et al, 2014), and contain the following files.

- TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01-20141210.mzML.gz: an TMT6 6-plex LC-MSMS data containing 6 human spiked-in proteins in a constant *Erwinia carotovora* protein backgroud. The data is described in more details in Gatto and Christoforou (2013).
- TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01-20141210.mzid: generated searching the raw data against the *Erwinia carotovora* fasta database

References

- Vizcaíno J.A. et al. *ProteomeXchange: globally co-ordinated proteomics data submission and dissemination*, Nature Biotechnology 2014, 32, 223–226. http://www.ncbi.nlm.nih.gov/pubmed/24727771
- Gatto L. and Christoforou A. *Using R and Bioconductor for proteomics data analysis*, Biochim Biophys Acta - Proteins and Proteomics, 2013. http://www.ncbi.nlm.nih.gov/pubmed/23692960

See Also

The rpx package can be used to access and download any PRIDE/ProteomeXchange files.

`Report.Derks2022.plexDIA` *Derks 2022 plexDIA data*

Description

Single cell proteomics data acquired by the Slavov Lab using the plexDIA protocol. It contains quantitative information from pancreatic ductal acinar cells (PDAC; HPAF-II), melanoma cells (WM989-A6-G3) and monocytes (U-937) at precursor and protein level. The each run acquired 3 samples thanks to mTRAQ multiplexing.

The data were downloaded from the Slavov lab google drive:

- https://drive.google.com/drive/folders/1pUC2zgXKtKYn22mlor0lmUDK0frgwL_-
- DIANN_outputs
- wJD1146_1193_1200_tsvLib
- Report.tsv

For more details about the data: https://plexdia.slavovlab.net/

The file is reshare here allow its dissemination via the MsDataHub package.

Author(s)

Laurent Gatto

References

Derks, J., Leduc, A., Wallmann, G. et al. Increasing the throughput of sensitive proteomics by plexDIA. Nat Biotechnol (2022). 10.1038/s41587-022-01389-w.

`sciex` *AB Sciex LC-MS data files*

Description

The *sciex* mzML files represent profile-mode LC-MS data of pooled human serum samples (the same pool being measured). The samples were analyzed by ultra high-performance liquid chromatography (UHPLC; Agilent 1290) coupled to a Q-TOF mass spectrometer (TripleTOF 5600+ AB Sciex). The chromatographic separation was based in hydrophilic interaction liquid chromatography (HILIC) and performed using an Waters Acquity BEH Amide, 100 x 2.1 mm column.

The mass spectrometer was operated in full scan mode in the mass range from 50 to 1000 m/z and with an accumulation time of 250 ms. The files represent a subset of spectra/scans from m/z 105 to 134 and from retention time 0 to 260 seconds. The files were generated in the same LC-MS run, but from different injections. Details on the individual files are provided below.

Files:

- *20171016_POOL_POS_1_105-134.mzML*: profile-mode LC-MS data of pooled human serum samples. Injection index: 1.
- *20171016_POOL_POS_3_105-134.mzML*: profile-mode LC-MS data of pooled human serum samples. Injection index: 19.

Author(s)

Sigurdur Smarason, Giuseppe Paglia and Johannes Rainer

`TripleTOF` *Triple TOF SWATH Data*

Description

These files represent data from reverse-phased LC-MS/MS runs on the Agilent Pesticide mix obtained from a Sciex 6600 Triple ToF operated either in Sequential Window Acquisition of all THeoretical mass spectra (SWATH) or Data Dependent Acquisition (DDA) acquisition mode.

The data files are:

- *PestMix1_DDA.mzML*: mzML file with MS1 and MS2 spectra from the Agilent Pesticide Mix acquired in DDA mode.
- *PestMix1_SWATH.mzML*: mzML file with MS1 and MS2 spectra from the Agilent Pesticide Mix acquired in SWATH mode.

Author(s)

Micheal Witting, Johannes Rainer

Index

20171016_POOL_POS_1_105-134.mzML (sciex), 7
20171016_POOL_POS_3_105-134.mzML (sciex), 7

Ai2025, 2
Ai2025_aCMs_report.tsv (Ai2025), 2
Ai2025_iCMs_report.tsv (Ai2025), 2

benchmarkingDIA, 3

cdf, 3
contaminatants (cRAP), 5
cptac, 4
cptac_a_b_c_peptides.txt (cptac), 4
cptac_a_b_peptides.txt (cptac), 4
cptac_peptides.txt (cptac), 4
cRAP, 5
crap (cRAP), 5
crap_ccp.fasta (cRAP), 5
crap_gpm.fasta (cRAP), 5
crap_maxquant.fasta.gz (cRAP), 5

ko15.CDF (cdf), 3

MsDataHub, 5
MsDataHub(), 5

PestMix1_DDA.mzML (TripleTOF), 8
PestMix1_SWATH.mzML (TripleTOF), 8
PXD000001, 6

Report.Derks2022.plexDIA, 7

sciex, 7

TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01-20141210.mzid (PXD000001), 6
TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01-20141210.mzML.gz (PXD000001), 6
TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01.20141210.mzid (PXD000001), 6
TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01.2014 (PXD000001), 6
TripleTOF, 8

X20171016_POOL_POS_1_105.134.mzML (sciex), 7
X20171016_POOL_POS_3_105.134.mzML (sciex), 7