Package 'pRolocdata'

October 23, 2025

Type Package

Title Data accompanying the pRoloc package

Version 1.47.0

Description Mass-spectrometry based spatial proteomics data sets and protein complex separation data. Also contains the time course expression experiment from Mulvey et al. 2015.

Depends R (>= 3.6.3), MSnbase

Imports Biobase, utils

Suggests pRoloc (>= 1.13.8), testthat, SummarizedExperiment

License GPL-2

BugReports https://github.com/lgatto/pRolocdata/issues

URL https://github.com/lgatto/pRolocdata

biocViews ExperimentData, Homo_sapiens_Data, MassSpectrometryData, Arabidopsis_thaliana_Data, Drosophila_melanogaster_Data, Mus_musculus_Data, StemCell, Proteome

RoxygenNote 6.1.1

Encoding UTF-8

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

git_branch devel

git_last_commit 40b364e

git_last_commit_date 2025-04-15

Repository Bioconductor 3.22

Date/Publication 2025-10-23

Author Laurent Gatto [aut] (ORCID: <https://orcid.org/0000-0002-1520-2268>), Oliver Crook [aut] (ORCID: <https://orcid.org/0000-0001-5669-8506>), Lisa Breckels [cre, aut] (ORCID: <https://orcid.org/0000-0001-8918-7171>)

Maintainer Lisa Breckels <lms79@cam.ac.uk>

Contents

andreyev2010 *Six sub-cellular fraction data from mouse macrophage-like RAW264.7 cells from Andreyev et al. (2009)*

Description

Data from Andreyev AY, Shen Z, Guan Z, Ryan A, Fahy E, Subramaniam S, Raetz CR, Briggs S, Dennis EA. Application of proteomic marker ensembles to subcellular organelle identification. Mol Cell Proteomics. 2010 Feb;9(2):388-402. DOI:http://dx.doi.org/10.1074/mcp.M900432-MCP200. PubMed PMID:19884172; PubMed Central PMCID:PMC2830848.

The 6 subcellular fractions are nuclei, mitochondria, cytoplasm, endoplasmic reticulum, plasma membrane and dense microsomal fraction.

Usage

```
data("andreyev2010")
data("andreyev2010rest")
data("andreyev2010activ")
```

Details

`andreyev2010` is the full data where missing values were replaced by 0. `andreyev2010rest` and `andreyev2010activ` contain the resting (control) and Kdo2-lipid A-treated (activated) data respectively, which have been normalised (each reporter intensity was normalised by the sum over all replicates).

Source

These data were generated from supplemental tables S1 (quantitative data) and 2 (organelle markers) (http://www.mcponline.org/content/9/2/388/suppl/DC1). See `inst/scripts/andreyev2010.R` for details.

Examples

```
data(andreyev2010rest, verbose = TRUE)
data(andreyev2010activ, verbose = TRUE)

library("pRoloc")
par(mfrow = c(1, 2))
plot2D(andreyev2010rest, main = "Resting (control)")
plot2D(andreyev2010activ, main = "Kdo2-lipid A-treated")
addLegend(andreyev2010activ)
```

andy2011 *LOPIT experiment on Human Embryonic Kidney fibroblast HEK293T cells from Breckels et al. (2013)*

Description

This is a LOPIT dataset from a standard LOPIT experimental design on Human Embryonic Kidney (HEK293T) fibroblast cells. See below for more details.

Note: this data was originally called `andy2011`. It is still available under that name but might be deprecated in the future and hence it is advised to use `HEK293T2011`.

Usage

```
data(HEK293T2011)
data(HEK293T2011hpa)
data(HEK293T2011goCC)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

This is a LOPIT experiment. Normalised intensities for 1371 proteins for eight iTRAQ 8-plex labelled fractions. This dataset was used in testing the phenotype discovery algorithm from Breckels et al., *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation*, J Proteomics, 2013, 88:129-40, see `phenoDisco`. New phenotype clusters identified from algorithm application are available as `pd.2013` feature meta-data and the markers used as input for the analysis are available as `markers` feature meta-data.

The `HEK293T2011goCC` instance contains binary assay data. Its columns represent GO CC terms that have been observed for the object's features. A 1 indicates that a GO term has been associated to a given feature (protein); a 0 means not such association was found in the GO ontology.

The `HEK293T2011hpa` instance contains binary assay data. Its columns represent subcellular locations that have been observed from microscopy images from the Human Protein Atlas, for each protein. A 1 indicates that a subcellular term has been associated to a given feature (protein); a 0 means not such association was found. This matrix of terms was generated from version 13, released on 11/06/2014 of the Human Protein Atlas.

Source

The data was generated by A. Christoforou at the Cambridge Centre for Proteomics.

http://www.bio.cam.ac.uk/proteomics/.

References

Breckels LM, Gatto L, Christoforou A, Groen AJ, Lilley KS and Trotter MWB. *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation.* J Proteomics. 2013 Aug 2;88:129-40. doi: 10.1016/j.jprot.2013.02.019. Epub 2013 Mar 21. PubMed PMID: 23523639

Examples

```
data(HEK293T2011)
HEK293T2011
pData(HEK293T2011)
head(exprs(HEK293T2011))
## Organelle marker proteins
table(fData(HEK293T2011)$markers)
## PhenoDisco assignment results
table(fData(HEK293T2011)$pd.2013)

data(HEK293T2011goCC)
dim(HEK293T2011goCC)
head(featureNames(HEK293T2011goCC))
exprs(HEK293T2011goCC)[1:10, 1:5]
```

---

at_chloro *The AT_CHLORO data base*

Description

AT_CHLORO is a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins.

The `assayData` contains the raw spectral counts for 3 chloroplastic fractions (the envelope, the stroma and the thylakoids) and for a complete chloroplast sample. The percentage of occurrence in each of the sub-chloroplast fraction as calculated in Ferro et al. (2010) are available as feature meta data (`Percent_ENV`, `Percent_STR` and `Percent_THY`).

Usage

```
data(at_chloro)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Source

Myriam Ferro Exploring the Dynamics of Proteomes (EDyP) Laboratoire Biologie a Grande Echelle (BGE) U1038 INSERM/CEA/UJF Institut de Recherches en Technologies et Sciences pour le Vivant (iRTSV) CEA/Grenoble

References

Ferro M, Brugiere S, Salvi D, Seigneurin-Berny D, Court M, Moyet L, Ramus C, Miras S, Mellal M, Le Gall S, Kieffer-Jaquinod S, Bruley C, Garin J, Joyard J, Masselon C, Rolland N. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol Cell Proteomics. 2010 Jun;9(6):1063-84. Epub 2010 Jan 10. PubMed PMID: 20061580; PubMed Central PMCID: PMC2877971

Examples

```
data(at_chloro)
dim(at_chloro)
pData(at_chloro)
head(exprs(at_chloro))
fvarLabels(at_chloro)
table(fData(at_chloro)$markers)
## check exprs data and 'TotalSpectralCount' feature meta data
all(fData(at_chloro)$TotalSpectralCount == rowSums(exprs(at_chloro)))
## create a set with the percentage of occurrence, as in Ferro et al. 2010
## rows that have no 'TOT' in the feature vars of interest
sel <- apply(fData(at_chloro)[, c("Percent_ENV", "Percent_STR", "Percent_THY")],
             1, function(.x) length(grep("TOT", .x)) == 0)
## new MSnSet
at_chloro2 <- at_chloro[sel, 1:3]
## columns of interest
perc <- c("Percent_ENV", "Percent_STR", "Percent_THY")
## create a new intensity matrix
exprs2 <-  matrix(as.numeric(as.matrix(fData(at_chloro2)[, perc])), ncol
= 3)
colnames(exprs2) <- sampleNames(at_chloro2)
rownames(exprs2) <- featureNames(at_chloro2)
summary(rowSums(exprs2))
exprs(at_chloro2) <- exprs2
validObject(at_chloro2)
```

`baers2018` *Synechocystis spatial proteomics*

Description

Data from 'Spatial mapping of a cyanobacterial proteome reveals distinct subcellular compartment organisation and dynamic metabolic pathways' (submitted).

Cyanobacteria are complex prokaryotes, incorporating a Gram-negative cell wall and internal thylakoid membranes. However, localisation of proteins within cyanobacterial cells is poorly understood. Using subcellular fractionation and quantitative proteomics we report the most extensive subcellular map of the proteome of a cyanobacterial cell, identifying ~67% of Synechocystis sp. PCC 6803 proteins, ~1000 more than previous studies. 1711 proteins were assigned to six specific subcellular regions.

This dataset is composed of two combined replicated 10-plex LOPIT experiments.

Protein markers for the plasma membrane, thylakoid membrane, cytosol, and small and la rge ribosomal subunits were curated from a literature review. A Support Vector Machine (SVM) classifier was employed on the combined dataset, with a radial basis function kernel, using class specific weights for classification of unassigned proteins to one of the 5 defined sub-cellular niches, TM, PM, soluble, small ribosomal subunit, large ribosomal subunit. The weights used in classification were set to be inversely proportional to the subcellular class frequencies to account for class imbalance. Algorithmic performance of the SVM on the dataset was estimated (as described in Trotter et al 2010). Scoring thresholds were calculated per subcellular niche and were set based on concordance with existing subcellular knowledge annotation to attain a 7.5% false discovery rate (FDR). Unassigned proteins were then classified to 1 of the 5 compartments according to the SVM prediction if greater than the calcul ated class threshold.

Usage

```
data("baers2018")
```

Examples

```
data(baers2018)

library("pRoloc")
par(mfrow = c(1, 2))
plot2D(baers2018, main = "Markers")
addLegend(baers2018, where = "bottomright")
plot2D(baers2018, fcol = "Localisation", main = "Localisation")
```

---

beltran2016 *Data from Beltran et al. 2016*

Description

The data contain the spatial proteomics data for 5 time points (24, 48, 72, 96 and 120) and two conditions (HMCV infection and MOCK), totalling 10 `MSnSet` object. Each contains expression data for 1748 to 2220 proteins along 6 fractions quantified by TMT 6-plex.

Usage

```
data(beltran2016HCMV24)
data(beltran2016HCMV48)
data(beltran2016HCMV72)
data(beltran2016HCMV96)
data(beltran2016HCMV120)
data(beltran2016MOCK24)
data(beltran2016MOCK48)
data(beltran2016MOCK72)
data(beltran2016MOCK96)
data(beltran2016MOCK120)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Jean Beltran PM, Mathias RA, Cristea IM. *A Portrait of the Human Organelle Proteome In Space and Time during Cytomegalovirus Infection*. Cell Syst. 2016 Oct 26;3(4):361-373.e6. doi: 10.1016/j.cels.2016.08.012. Epub 2016 Sep 15. PubMed PMID: 27641956; PubMed Central PMCID: PMC5083158.

Examples

```
## load the two 24 hours datasets
data(beltran2016MOCK24)
data(beltran2016HCMV24)
beltran2016MOCK24
beltran2016HCMV24
```

```
## the expression data
head(exprs(beltran2016MOCK24))
head(exprs(beltran2016HCMV24))

## abstract
abstract(beltran2016HCMV24)

## plotting
library("pRoloc")
par(mfrow = c(1, 2))
plot2D(beltran2016HCMV24, main = "HCMV 24hpi")
plot2D(beltran2016MOCK24, main = "MOCK 24hpi")

## Combine the date as a list and keep only common features
ml <- MSnSetList(list(beltran2016HCMV24, beltran2016MOCK24))
ml <- commonFeatureNames(ml)

par(mfrow = c(1, 2))
plot2D(ml[[1]], main = "HCMV 24hpi")
plot2D(ml[[2]], main = "MOCK 24hpi")
```

`courtland_control` *Genetic Disruption of WASHC4 Drives Endo-lysosomal Dysfunction and Cognitive-Movement Impairments in Mice and Humans*

Description

Data from ' Genetic Disruption of WASHC4 Drives Endo-lysosomal Dysfunction and Cognitive-Movement Impairments in Mice and Humans'

Mutation of the WASH complex subunit, SWIP, is implicated in human intellectual disability, but the cellular etiology of this association is unknown. We identify the neuronal WASH complex proteome, revealing a network of endosomal proteins. To uncover how dysfunction of endosomal SWIP leads to disease, we generate a mouse model of the human WASHC4 c.3056C>G mutation. Quantitative spatial proteomics analysis of SWIP P1019R mouse brain reveals that this mutation destabilizes the WASH complex and uncovers significant perturbations in both endosomal and lysosomal pathways. Cellular and histological analyses confirm that SWIP P1019R results in endo-lysosomal disruption and uncover indicators of neurodegeneration. We find that SWIP P1019R not only impacts cognition, but also causes significant progressive motor deficits in mice. Remarkably, a retrospective analysis of SWIP P1019R patients confirms motor deficits in humans. Combined, these findings support the model that WASH complex destabilization, resulting from SWIP P1019R, drives cognitive and motor impairments via endo-lysosomal dysfunction in the brain.

Usage

```
data("courtland_control")
data("courtland_mutant")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Examples

```
data(courtland_control)
courtland_control
pData(courtland_control)
exprs(courtland_control)[1:3,1:3]

library("pRoloc")
plot2D(courtland_control, main = "mouse brain control")
```

`davies2018` *AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A*

Description

Data from 'AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A' Nature Communications.

Adaptor protein 4 (AP-4) is an ancient membrane trafficking complex, whose function has largely remained elusive. In humans, AP-4 deficiency causes a severe neurological disorder of unknown aetiology. We apply unbiased proteomic methods, including 'Dynamic Organellar Maps', to find proteins whose subcellular localisation depends on AP-4. We identify three transmembrane cargo proteins, ATG9A, SERINC1 and SERINC3, and two AP-4 accessory proteins, RUSC1 and RUSC2. We demonstrate that AP-4 deficiency causes missorting of ATG9A in diverse cell types, including patient-derived cells, as well as dysregulation of autophagy. RUSC2 facilitates the transport of AP-4-derived, ATG9A-positive vesicles from the trans-Golgi network to the cell periphery. These vesicles cluster in close association with autophagosomes, suggesting they are the 'ATG9A reservoir' required for autophagosome biogenesis. Our study uncovers ATG9A trafficking as a ubiquitous function of the AP-4 pathway. Furthermore, it provides a potential molecular pathomechanism of AP-4 deficiency, through dysregulated spatial control of autophagy.

Usage

```
data("davies2018ap4b1")
data("davies2018ap4e1")
data("davies2018wt")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A Alexandra K. Davies, Daniel N. Itzhak, James R. Edgar, Tara L. Archuleta, Jennifer Hirst, Lauren P. Jackson, Margaret S. Robinson & Georg H. H. Borner https://doi.org/10.1038/s41467-018-06172-7

Examples

```
data(davies2018wt)
davies2018wt
pData(davies2018wt)
exprs(davies2018wt)[1:3,1:3]

library("pRoloc")
plot2D(davies2018wt,, main = "Davies 2018 HeLa - wt")
```

`dunkley2006` *LOPIT data from Dunkley et al. (2006)*

Description

This is the data from Dunkley et al., *Mapping the Arabidopsis organelle proteome*, PNAS 2006 (PMID 16618929). See below for more details.

Usage

```
data(dunkley2006)
data(dunkley2006goCC)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

This is a LOPIT experiment. Normalised intensities for 689 proteins for four iTRAQ 4-plex labelled fraction and 2 membrane preparation in duplicate (16 samples, see `phenoData(dunkley2006)` for more details) are provided.

Partial least square discriminant analysis (PLSDA) has originally been applied to the test data `fData(dunkley)$markers)`; assignment results are available with `fData(dunkley)$assigned)` for 5 organelles.

This dataset was also used in testing the phenotype discovery algorithm from Breckels et al., *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation*, J Proteomics, *In Press.*, see `phenoDisco`. New phenotype clusters identified from algorithm application are available as `pd.2013` feature meta-data.

The `dunkley2006goCC` instance contains binary assay data. Its columns represent GO CC terms that have been observed for the object's features. A 1 indicates that a GO term has been associated to a given feature (protein); a 0 means not such association was found in the GO ontology.

Source

Supporting Information on http://www.pnas.org/content/103/17/6518.abstract.

References

Dunkley TP, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, Griffin JL, Bessant C, Brandizzi F, Hawes C, Watson RB, Dupree P, Lilley KS. *Mapping the Arabidopsis organelle proteome.* Proc Natl Acad Sci U S A. 2006 Apr 25;103(17):6518-23. Epub 2006 Apr 17. PubMed PMID: 16618929; PubMed Central PMCID: PMC1458916.

Breckels LM, Gatto L, Christoforou A, Groen AJ, Lilley KS and Trotter MWB. *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation* J Proteomics. *In Press.*

Examples

```
data(dunkley2006)
dunkley2006
phenoData(dunkley2006)
## Input training data (organelle markers)
table(fData(dunkley2006)$markers)
## PLSDA assignment results
table(fData(dunkley2006)$assigned)
## PhenoDisco results
table(fData(dunkley2006)$pd.2013)
```

`E14TG2a` *LOPIT experiment on Mouse E14TG2a Embryonic Stem Cells from Breckels et al. (2016)*

Description

This is data from a standard LOPIT experimental design on Mouse E14TG2a embryonic stem cells. See below for more details.

Usage

```
data(E14TG2aS1)
data(E14TG2aS2)
data(E14TG2aR)
data(E14TG2aS1yLoc)
data(E14TG2aS1goCC)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

This is a LOPIT experiment. Normalised intensities of proteins from eight iTRAQ 8-plex labelled fractions are available for 2 replicates (indexed 1 and 2) using stringent and relaxed setting (S and R, respectively).

The `E14TG2aS1goCC` instance contains binary assay data. Its columns represent GO CC terms that have been observed for the object's features. A 1 indicates that a GO term has been associated to a given feature (protein); a 0 means not such association was found in the GO ontology.

The `E14TG2aS1yLoc` instance contains 34 sequence and annotation features obtained from a feature selection of the sequence and annotation features from the computational classifier YLoc. These

features include: variants of psuedo amino acid counts, autocorrelation, sum of charge, prosite patterns, Gene Ontoloy terms and the number of signal peptides. These features are described in detail in Breckels et al (2015).

Source

The data was generated by A. Christoforou at the Cambridge Centre for Proteomics, Cambridge. http://www.bio.cam.ac.uk/proteomics/.

Examples

```
data(E14TG2aS1)
E14TG2aS1
pData(E14TG2aS1)
head(exprs(E14TG2aS1))
```

---

fabre2015r1 *Data from Fabre et al. 2015*

Description

Duplicated experimental data from Fabre et al. 2015, *Deciphering preferential interactions within supramolecular protein complexes: the proteasome case*. Protein complexes were separated by glycerol density gradient centrifugation. Proteins have been quantified by label-free (iBAQ) mass spectrometry.

Usage

```
data("fabre2015r1")
data("fabre2015r2")
```

References

Fabre B, Lambour T, Garrigues L, Amalric F, Vigneron N, Menneteau T, Stella A, Monsarrat B, Van den Eynde B, Burlet-Schiltz O, Bousquet-Dubouch MP. Deciphering preferential interactions within supramolecular protein complexes: the proteasome case. Mol Syst Biol. 2015 Jan 5;11(1):771. doi: 10.15252/msb.20145497. PubMed PMID: 25561571.

Examples

```
data(fabre2015r1)
experimentData(fabre2015r1)
library("pRoloc")
plot2D(fabre2015r1)
addLegend(fabre2015r1, where = "topright")
```

`foster2006` *PCP data from Foster et al. (2006)*

Description

This is the data from Foster et al., *A Mammalian Organelle Map by Protein Correlation Profiling*, Cell 2006 (PMID 16615899). See below for more details.

Usage

```
data(foster2006)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

This is a PCP experiment. Label-free quantification has been done on a totla of 26 high and low density-separated fractions (see `pData(foster2006)`). A total of 1555 proteins have been quantified in a subset of the fractions. The proteins are described in the `featureData` slot. Chi^2 calculations, as defined in the PCP experiment, has been performed using marker proteins for a total of 8 organelles, as well as the authors' original assignment and notes are available in the `featureData` slot.

Source

Supplemental data on http://www.cell.com/abstract/S0092-8674(06)00369-2.

References

Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, Mann M. *A mammalian organelle map by protein correlation profiling.* Cell. 2006 Apr 7;125(1):187-99. PubMed PMID: 16615899.

Examples

```
data(foster2006)
foster2006
phenoData(foster2006)
featureData(foster2006)
## organelle marker proteins
table(fData(foster2006)$train)
```

groen2014 *LOPIT experiments on Arabidopsis thaliana roots, from Groen et al. (2014)*

Description

This is the data from Groen et al. *Identification of Trans Golgi Network proteins in Arabidopsis thaliana root tissue* J. Proteome Res, 2014, Feb 7; 13(2):763-776. See below for more details.

Usage

```
data(groen2014r1)
data(groen2014r2)
data(groen2014r3)
data(groen2014cmb)
data(groen2014r1goCC)
```

Format

An instance of class `MSnSet` from package `MSnbase`.

Details

This is a LOPIT experiment. Normalised intensities for proteins for four iTRAQ 4-plex labelled fractions are available for 3 replicates (`r1`, `r2` and `r3` respectively). The 3 replicates have also been combined as described in Groen et al. and Trotter et al. (2010) to generate a fourth dataset (`cmb`), also shown in the example code below.

The `groen2014r1goCC` instance contains binary assay data. Its columns represent GO CC terms that have been observed for the object's features. A 1 indicates that a GO term has been associated to a given feature (protein); a 0 means not such association was found in the GO ontology.

Source

http://pubs.acs.org/doi/abs/10.1021/pr4008464

References

Groen AJ, Sancho-Andres G, Breckels LM, Gatto L, Aniento F, and Lilley KS. *Identification of Trans Golgi Network proteins in Arabidopsis thaliana root tissue*. J. Proteome Res, 2014, Feb 7; 13(2):763-776. DOI:10.1021/pr4008464, PMID: 24344820.

Trotter MWB, Sadowski PG, Dunkley TPJ, Groen AJ and Lilley KS. *Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions.* Proteomics 2010 10(23):4213-4219. PMID 21058340.

Sadowski PG, Groen AJ, Dupree P and Lilley KS. *Sub-cellular localization of membrane proteins.* Proteomics 2008 8(19):3991-4011. PMID 18780351.

Dunkley TP, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, Griffin JL, Bessant C, Brandizzi F, Hawes C, Watson RB, Dupree P, Lilley KS. *Mapping the Arabidopsis organelle proteome.* Proc Natl Acad Sci U S A. 2006 Apr 25;103(17):6518-23. Epub 2006 Apr 17. PubMed PMID: 16618929; PubMed Central PMCID: PMC1458916.

Examples

```
data(groen2014r1)
data(groen2014r2)
data(groen2014r3)
data(groen2014cmb)

## The combine dataset can generated manually using
cmb <- combine(groen2014r1, updateFvarLabels(groen2014r2))
cmb <- filterNA(cmb)
cmb <- combine(cmb, updateFvarLabels(groen2014r3))
cmb <- filterNA(cmb)
fData(cmb) <- fData(cmb)[, c(1,2,5)]
cmb

## or can simply be loaded directly
data(groen2014cmb)

## check datsets are the same
all.equal(cmb, groen2014cmb, check.attributes=FALSE)
```

---

hall2009 *LOPIT data from Hall et al. (2009)*

Description

This is the data from Hall et al. *The Organelle Proteome of the DT40 Lymphocyte Cell Line* Mol Cell Proteomics. 2009 Jun;8(6):1295-305. (PMID: PMC2690488).

Usage

```
data(hall2009)
```

Format

An instance of class `MSnSet` from package `MSnbase`.

Details

See reference.

Source

http://www.mcponline.org/content/8/6/1295.abstract

References

Hall SL, Hester S, Griffin JL, Lilley KS, Jackson AP. *The organelle proteome of the DT40 lymphocyte cell line* Mol Cell Proteomics. 2009 Jun;8(6):1295-305. doi: 10.1074/mcp.M800394-MCP200. Epub 2009 Jan 30. PubMed PMID: 19181659; PubMed Central PMCID: PMC2690488.

Examples

```
data(hall2009)
pData(hall2009)
library("pRoloc")
plot2D(hall2009)
```

`havugimana2012` *Data from Havugimana et al. 2012*

Description

Data from Havugimana et al. 2012, *A census of human soluble protein complexes*. The protein complexes were fractionated by ion exchange chromatography, Isoelectric focusing and sucrose density gradient centrifugation. Proteins were quantified by spectral counting.

Usage

```
data("havugimana2012")
```

References

Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ER, Paccanaro A, Marcotte EM, Emili A. A census of human soluble protein complexes. Cell. 2012 Aug 31;150(5):1068-81. doi: 10.1016/j.cell.2012.08.011. PubMed PMID: 22939629; PubMed Central PMCID: PMC3477804.

Examples

```
data(havugimana2012)
experimentData(havugimana2012)
```

`hirst2018` *Data from Hirst et al. 2018*

Description

From the supplementary file notes:

These are the SILAC ratio data from 2046 proteins with complete profiles across all nine organellar maps.

Each profile consists of five ratios, corresponding to five subcellular fractions obtained by differential centrifugation (3000 x g pellet, 6000 x g pellet, 12000 x g pellet, 24000 x pellet, 80000 x g pellet). The centrifugation speeds are available `centriguation` in the `MSnSet` object.

Each ratio shows the abundance of the total membrane SILAC heavy spike-in relative to the abundance in a given subfraction.

Maps were made from three cell lines (control HeLa, and two independent AP5Z1 KO HeLa cell lines, called AP5KNC2 and AP5KOC6), each in triplicate (replicates R1, R2, and R3). The sample are code as `"CTRL"` (HeLa control), `"C2"` (AP5KNC2 AP5Z1 KO cells) and `"C6"` (AP5KNC6 AP5Z1 KO cells).

Marker proteins used to define organellar clusters in Supplemental Figure 1 in the manuscript are annotated as feature variable `markers`.

Finally, the ratios in the `hirst2018` data where normalised by their sum (using `normalise(, method = "sum")`).

The feature data also contains information about the comparison of organellar maps made from control or AP5 ablated cells, revealing putative proteins that undergo subcellular localisation shifts. Each protein receives an M score (magnitude of movement), and an R score (reproducibility of movement, i.e. correlation between replicates). In addition, the reproducibility of movement between the two AP5 KO cell lines is scored (Correlation C2 vs C6). Note however that the authors themselves claim that:

'The cutoffs chosen in Fig 1C (M > 1.5, R > 0.5) correspond to an estimated FDR of 23%. Please note that the actual FDR is probably lower than this estimated FDR, because the mock data lack the additional cell line and the clonal correlation filter.'

The re-localisation candidates are those that have an M score > 1.5 and a R score > 0.5, and are marked with a hit feature variable set to `TRUE`.

Usage

```
data(hirst2018)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Hirst J, Itzhak DN, Antrobus R, Borner GHH, Robinson MS. Role of the AP-5 adaptor protein complex in late endosome-to-Golgi retrieval. PLoS Biol. 2018 Jan 30;16(1):e2004411. doi: 10.1371/journal.pbio.2004411. eCollection 2018 Jan. PubMed PMID: 29381698; PubMed Central PMCID: PMC5806898.

Examples

```
## load the two 24 hours datasets
data(hirst2018)
hirst2018

## experimental design
table(pData(hirst2018)[, -2])

## the expression data
exprs(hirst2018)[1:5, 1:3]

## abstract
abstract(hirst2018)

## split data by samples
x <- split(hirst2018, "sample")

## These are the relocalisation hits
hits <- which(fData(hirst2018)$Hits)
reloc <- FeaturesOfInterest(description = "Relocation hits",
       featureNames(hirst2018)[hits])
reloc
```

```
## plotting
library("pRoloc")
par(mfrow = c(1, 3))
plot2D(x[[1]], main = "AP5KNC2")
highlightOnPlot(x[[1]], reloc)
plot2D(x[[2]], main = "AP5KNC6")
highlightOnPlot(x[[1]], reloc)
plot2D(x[[3]], main = "HeLa control")
highlightOnPlot(x[[1]], reloc)
addLegend(x[[3]], where = "topleft")
```

`hyperLOPIT2015` *Protein and PMS-level hyperLOPIT datasets on Mouse E14TG2a embryonic stem cells from Christoforou et al. (2016).*

Description

This is a spatial proteomics dataset from a hyperLOPIT experimental design on Mouse E14TG2a embryonic stem cells.

Usage

```
data(hyperLOPIT2015)
data(hyperLOPIT2015_se)
data(hyperLOPIT2015ms3r1)
data(hyperLOPIT2015ms3r2)
data(hyperLOPIT2015ms3r3)
data(hyperLOPIT2015ms2)
data(hyperLOPIT2015ms3r1psm)
data(hyperLOPIT2015ms3r2psm)
data(hyperLOPIT2015ms2psm)
```

Format

The data are an instance of class `MSnSet` from package `MSnbase`. Those ending in `_se` are of class `SummarizedExperiment`.

Details

This is a hyperLOPIT experiment. Normalised intensities for proteins for TMT 10-plex labelled fractions are available for 3 replicates acquired in MS3 mode (`hyperLOPIT2015ms3r1`, `hyperLOPIT2015ms3r2` and `hyperLOPIT2015ms3r3`) using an Orbitrap Fusion mass-spectrometer. The first two replicates have also been combined as described in Trotter et al (2010) to generate dataset `hyperLOPIT2015` to increase organellar resolution. A dataset acquired in MS2 mode has also been acquired (`hyperLOPIT2015ms2`) which was also generated using the Orbitrap Fusion and using a TMT 10-plex experimental design.

The PSM-level cvs file are available in the `extdata` directory and have been processed as follows: imported `MSnSet` instances using `readMSnSet2`, PSMs with missing values were filtered out with `filterNA`, only PSMs with feature variable `Quan.Usage "Used"` and a TMT6plex modification were retained and the phenoData was matched and assigned from the respective protein-level data. Finally, marker proteins are annotated based on the combined protein-level data `hyperLOPIT2015` and reporter tags are normalised using the `"sum"` method. The processing script is `scripts/hyperlopit2015psm.R`.

The TAGM feature data contains the allocation results from the Baysian T-augmented Gaussian Mixture modelling approach as described in Crook et al. (2018).

Source

The data was generated by A. Christoforou and C. Mulvey in the Cambridge Centre for Proteomics. http://www.bio.cam.ac.uk/proteomics/.

References

A draft map of the mouse pluripotent stem cell spatial proteome. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, Naake T, Gatto L, Viner R, Martinez Arias A, Lilley KS. Nat Commun. 2016 Jan 12;7:8992. doi: 10.1038/ncomms9992. PubMed PMID: 26754106; PubMed Central PMCID: PMC4729960.

A Bayesian Mixture Modelling Approach For Spatial Proteomics Oliver M Crook, Claire M Mulvey, Paul D. W. Kirk, Kathryn S Lilley, Laurent Gatto bioRxiv 282269; doi: https://doi.org/10.1101/282269

Examples

```
data(hyperLOPIT2015)
hyperLOPIT2015
pData(hyperLOPIT2015)
head(exprs(hyperLOPIT2015))

data(hyperLOPIT2015ms3r1psm)
x <- combineFeatures(hyperLOPIT2015ms3r1psm,
                     groupBy = fData(hyperLOPIT2015ms3r1psm)$Protein.Group.Accession,
                     fun = median)
library("pRoloc")
par(mfrow = c(1, 2))
plot2D(hyperLOPIT2015ms3r1psm, main = "PSM-level")
plot2D(x, main = "Protein-level (using mean)")

## SummarizedExperiment data
library(SummarizedExperiment)
data(hyperLOPIT2015_se)
```

---

hyperLOPITU2OS2017 *2017 and 2018 hyperLOPIT on U2OS cells*

Description

This data contains 4 different datasets generated from U2OS cells. The lopitdcU2OS2018 was generated using the LOPIT-DC method and all other datasets have been generated using the hyperLOPIT protocol (see Christoforou et al. 2016 and Mulvey et al. 2017). The lopitdcU2OS2018 dataset contains 3 replicates, 10 fractions per replicate. The hyperLOPITU2OS2017 dataset contains 2 replicates, in which the quantitation was obtained using two sets of TMT 10-plex per replicate, producing a total of 40 quanitation channels, while in hyperLOPITU2OS2017b, 3 fractions with low protein yields have been remove (see example below). The hyperLOPITU2OS2018 dataset contains a third replicate, thus giving 57 quantitation channels in total.

Usage

```
data("hyperLOPITU2OS2017")
```

Format

An object of class `MSnSet`, defined in the `MSnbase` package.

Details

The data (expression and feature variable) contain:

- UniProt Accession for Protein Group (no isoform information): Unique UniProt accession for quantified protein group reported by Proteome Discoverer (1% FDR) - isoform information not retained.
- Normalized TMT 10-plex Reporter Ion Distribution: ReplicateX TMT SetX-126 Normalized TMT 10-plex reporter ion values, representing the distribution of each protein across the fractionation scheme for each experiment. Protein-level reporter ion values were calculated by taking the median of all quantifiable PSMs for the protein group, then normalized so that the sum of all 10 channels was equal to 1. The numeric value in the tag name corresponds to the nominal mass of each TMT reporter ion. The N and C suffixes differentiates between the 15N or 13C isotopologue variants of TMT 10-plex reporter ions of the same nominal mass.
- UniProt Accession for Protein Group (with isoform information): Unique UniProt accession for quantified protein group reported by Proteome Discoverer (1% FDR) - isoform information retained.
- UniProt Protein Description: UniProt description for protein accession.
- Coverage: Percentage of protein sequence covered by identified peptides.
- Quantified Proteins: Number of quantified protein groups.
- Quantified Unique Peptides: Number of unique quantified peptides. Only these peptides were used for quantification.
- Quantified Peptides: Number of quantified peptides. Only peptides that were unique to a single protein group were used for quantification.
- Quantified PSMs: Number of quantified peptide-spectrum matches.
- Score - ReplicateX TMT SetX: Total score of identified protein group for each experiment. This score is equal to the sum of the individual peptide scores.
- Coverage - ReplicateX TMT SetX: Percentage of protein sequence covered by identified peptides for each experiment.
- Quantified Peptides - ReplicateX TMT SetX: Number of quantified peptides for each experiment. Only peptides that were unique to a single protein group were used for quantification.
- Quantified PSMs - ReplicateX TMT SetX: Number of quantified peptide-spectrum matches for each experiment.
- SVM Marker Set: Final marker set used for SVM classification of protein subcellular localization to 14 subcellular compartments.
- SVM Classification: Subcellular class to which the protein group was assigned by SVM classification. All proteins are assigned to a single class by SVM.
- SVM Score: Confidence score for localization assignment, ranging from 0 to 1. A score close to 0 represents a very low confidence assignment, whereas a score of 1 indicates a very high confidence assignment.

- Final SVM Classification (5% FDR) (assignment): Predicted localization, with SVM score thresholds determined empirically by comparison to GO annotation and protein database annotation. The SVM score thresholds were set individually for each class so that the false discovery rate of the SVM classification was equal or lower than (5%).

References

Thul PJ et al. *A subcellular map of the human proteome*. Science. 2017 May 26;356(6340). pii: eaal3321. doi: 10.1126/science.aal3321. Epub 2017 May 11. PubMed PMID: 28495876.

Examples

```
data(hyperLOPITU2OS2017)
hyperLOPITU2OS2017

library("pRoloc")
plot2D(hyperLOPITU2OS2017, addLegend = "bottomleft")

## removing low intensity fractions
sort(colSums(exprs(hyperLOPITU2OS2017)))
i <- order(colSums(exprs(hyperLOPITU2OS2017)))[1:3]
x <- hyperLOPITU2OS2017[, -i]
plot2D(x, mirrorY = TRUE)

data(hyperLOPITU2OS2017b)
## only difference if subsetting date
all.equal(hyperLOPITU2OS2017b, x)
processingData(hyperLOPITU2OS2017b)
processingData(x)
```

---

itzhak2016 *Data from Itzhak et al. (2016)*

Description

Data from Daniel N Itzhak, Stefka Tyanova, Jurgen Cox and Georg HH Borner. Global, quantitative and dynamic mapping of protein subcellular localization. DOI:http://dx.doi.org/10.7554/eLife.16950 Published June 9, 2016 Cite as eLife 2016;10.7554/eLife.16950

It currently contains

- The second sheet contains the 6 replicates of the SILAC static data (*Static* data were used to genrate six deep organellar maps) and is made available as `itzhak2016stcSILAC`.

Usage

```
data("itzhak2016stcSILAC")
```

Source

This data was generated from Supplementary file 9 (https://elifesciences.org/content/5/e16950/supp-material9). See `inst/scripts/itzhak2016.R` for details.

Examples

```
data(itzhak2016stcSILAC)
itzhak2016stcSILAC
dim(itzhak2016stcSILAC)
pData(itzhak2016stcSILAC)

## only 1st replicate
dim(itzhak2016stcSILAC[, itzhak2016stcSILAC$rep == 1])


## filter out features with missing values
itzhak2016stcSILAC <- filterNA(itzhak2016stcSILAC)

library("pRoloc")
## Cell map
plot2D(itzhak2016stcSILAC)
## as in the paper
plot2D(itzhak2016stcSILAC, dims = c(1, 3))
```

`itzhak2016dynamic` *Global, quantitative and dynamic mapping of protein subcellular localization*

Description

Data from 'Global, quantitative and dynamic mapping of protein subcellular localization.

Subcellular localization critically influences protein function, and cells control protein localization to regulate biological processes. We have developed and applied Dynamic Organellar Maps, a proteomic method that allows global mapping of protein translocation events. We initially used maps statically to generate a database with localization and absolute copy number information for over 8700 proteins from HeLa cells, approaching comprehensive coverage. All major organelles were resolved, with exceptional prediction accuracy (estimated at >92%). Combining spatial and abundance information yielded an unprecedented quantitative view of HeLa cell anatomy and organellar composition, at the protein level. We subsequently demonstrated the dynamic capabilities of the approach by capturing translocation events following EGF stimulation, which we integrated into a quantitative model. Dynamic Organellar Maps enable the proteome-wide analysis of physiological protein movements, without requiring any reagents specific to the investigated process, and will thus be widely applicable in cell biology.

Usage

```
data("itzhak2016helaCtrl")
data("itzhak2016helaEgf")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. Elife. 2016 Jun 9;5:e16950.

Examples

```
data("itzhak2016helaCtrl")
helaCtrl <- itzhak2016helaCtrl
pData(helaCtrl)
exprs(helaCtrl)[1:3,1:3]

library("pRoloc")
plot2D(helaCtrl, main = "HeLa Ctrl", dims = c(1, 3))
```

`itzhak2017` *Data from Itzhak et al. 2017*

Description

The data from Itzhal et al. 2017 defines a spatial map for mouse primary neurons. The data are composed of 5 spatial maps, each containing 6 differential centrifugation fraction (as described in Itzhak et al. 2016, see `itzhak2016stcSILAC`).

The annotatied marker proteins are available in the `itzhak2017markers` dataset.

This data corresponds to the *Mouse neuron Intensity and LFQ data* data from supplemental table S3 (Proteomic Mouse Neuron Data Generated in This Study, Related to Figure 5).

Usage

```
data(itzhak2017)
data(itzhak2017markers)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Itzhak DN, Davies C, Tyanova S, Mishra A, William son J, Antrobus R, Cox J, Weekes MP, Borner GHH. A Mass Spectrometry-Based Approach for Mapping Protein Subcellular Localization Reveals the Spatial Proteome of Mouse Primary Neurons. Cell Rep. 2017 Sep 12;20(11):2706-2718. doi: 10.1016/j.celrep.2017.08.063. PubMed PMID: 28903049; PubMed Central PMCID: PMC5775508.

Examples

```
data(itzhak2017)
itzhak2017

## experimental design
table(pData(itzhak2017))

## the expression data
exprs(itzhak2017)[1:5, 1:5]

## abstract
abstract(itzhak2017)
```

```
## split data by samples
x <- split(itzhak2017, "map")

## plotting
library("pRoloc")
par(mfrow = c(2, 3))
for (i in 1:5)
    plot2D(x[[i]], main = paste("Map", i))
plot2D(itzhak2017, main = "All maps")
addLegend(itzhak2017, where = "bottomleft")
```

`kirkwood2013` *Data from Kirkwood et al. 2013.*

Description

Data from Kirkwood et al. 2013, *Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics*. Protein complexes were separated by size exclusion chromatography and proteins were quantified by spectral counting.

Usage

```
data("kirkwood2013")
```

References

Kirkwood KJ, Ahmad Y, Larance M, Lamond AI. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. Mol Cell Proteomics. 2013 Dec;12(12):3851-73. doi: 10.1074/mcp.M113.032367. Epub 2013 Sep 16. PubMed PMID: 24043423; PubMed Central PMCID: PMC3861729.

Examples

```
data(kirkwood2013)
experimentData(kirkwood2013)
```

`Kozik_con` *Small molecule enhancers of endosome-to-cytosol import augment anti-tumour immunity*

Description

Data from 'Small molecule enhancers of endosome-to-cytosol import augment anti-tumour immunity'

Efficient cross-presentation of antigens by dendritic cells (DCs) is critical for initiation of anti-tumour immune responses. Yet, several steps of antigen intracellular traffic during cross-presentation are incompletely understood: in particular, the molecular mechanisms and the relative importance of antigen import from endocytic compartments into the cytosol. Here, we asked whether antigen import into the cytosol is rate-limiting for cross-presentation and anti-tumour immunity. By screening 700 FDA-approved drugs, we identified 37 import enhancers. We focused on prazosin and

tamoxifen, and generated proteomic organellar maps of drug-treated DCs, covering the subcellular localisations of over 2000 proteins. By combining organellar mapping, quantitative proteomics, microscopy, and bioinformatics, we conclude that import enhancers undergo lysosomal trapping leading to membrane permeation and antigen release into the cytosol. Enhancing antigen import facilitates cross-presentation of both soluble and cell-associated antigens. Systemic administration of prazosin also led to reduced growth of MC38 tumours and to a synergistic effect with checkpoint immunotherapy in a melanoma model. Thus, inefficient antigen import into the cytosol limits antigen cross-presentation, restraining the potency of anti-tumour immune responses and efficacy of checkpoint blockers.

Usage

```
data("Kozik_con")
data("Kozik_pra")
data("Kozik_tam")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Examples

```
data(Kozik_con)
Kozik_con
pData(Kozik_con)
exprs(Kozik_con)[1:3,1:3]

library("pRoloc")
plot2D(Kozik_con, main = "denderitic cells control")
```

`krahmer2018pcp` *Subcellular Reorganization in Diet-Induced Hepatic Steatosis*

Description

Data from 'Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis' Developmental cell.

Lipid metabolism is highly compartmentalized between cellular organelles that dynamically adapt their compositions and interactions in response to metabolic challenges. Here, we investigate how diet-induced hepatic lipid accumulation, observed in non-alcoholic fatty liver disease (NAFLD), affects protein localization, organelle organization, and protein phosphorylation in vivo. We develop a mass spectrometric workflow for protein and phosphopeptide correlation profiling to monitor levels and cellular distributions of 6,000 liver proteins and 16,000 phosphopeptides during development of steatosis. Several organelle contact site proteins are targeted to lipid droplets (LDs) in steatotic liver, tethering organelles orchestrating lipid metabolism. Proteins of the secretory pathway dramatically redistribute, including the mis-localization of the COPI complex and sequestration of the Golgi apparatus at LDs. This correlates with reduced hepatic protein secretion. Our systematic in vivo analysis of subcellular rearrangements and organelle-specific phosphorylation reveals how nutrient overload leads to organellar reorganization and cellular dysfunction.

Usage

```
data("krahmer2018pcp")
data("krahmer2018phosphopcp")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Organellar proteomics and phospho-proteomics reveal subcellular reorganization in diet-induced hepatic steatosis. Krahmer N, Najafi B, Schueder F, Quagliarini F, Steger M, Seitz S, Kasper R, Salinas F, Cox J, Uhlenhaut NH, Walther TC. Developmental Cell. 2018 Oct 22;47(2):205-21. https://doi.org/10.1016/j.devcel.2018.09.017

Examples

```
data(krahmer2018pcp)
krahmer2018pcp
pData(krahmer2018pcp)
exprs(krahmer2018pcp)[1:3,1:3]

library("pRoloc")
plot2D(krahmer2018pcp, fcol = "Organelle" , main = "Krahmer 2018")

data(krahmer2018phosphopcp)
krahmer2018phosphopcp
exprs(krahmer2018phosphopcp)[1:3,1:3]

plot2D(krahmer2018phosphopcp, main = "Krahmer 2018")
```

---

kristensen2012r1 *Data from Kristensen et al. 2012*

Description

Triplicated experimental data from Kristensen et al. 2012, *A high-throughput approach for measuring temporal changes in the interactome*. Protein complexes were separated by size exclusion chromatography and protein were quantified using SILAC.

Usage

```
data("kristensen2012r1")
data("kristensen2012r2")
data("kristensen2012r3")
```

References

Kristensen AR, Gsponer J, Foster LJ. A high-throughput approach for measuring temporal changes in the interactome. Nat Methods. 2012 Sep;9(9):907-9. doi: 10.1038/nmeth.2131. Epub 2012 Aug 5. PubMed PMID: 22863883; PubMed Central PMCID: PMC3954081.

Examples

```
data(kristensen2012r1)
experimentData(kristensen2012r1)
```

---

lopimsSyn2 *LOPIMS data for the Synapter 2.0 paper*

Description

TODO

Usage

```
data("lopimsSyn1")
data("lopimsSyn2")
data("lopimsSyn2_0frags")
```

Format

These data are `MSnSet` instances, defined in the `MSnbase` package.

Examples

```
data(lopimsSyn1)
data(lopimsSyn2)
data(lopimsSyn2_0frags)

## Visualisation
library("pRoloc")
par(mfrow = c(1, 3))
plot2D(lopimsSyn1, main = "Synapter 1", addLegend = "topleft")
plot2D(lopimsSyn2, main = "Synapter 2")
plot2D(lopimsSyn2_0frags, main = "Synapter 2 (0 fragments)")
```

---

lpsTimecourse_mulvey2021 *Protein and PMS-level datasets from temporal abundance profiling experiments of THP-1 human leukaema cells stimulated with LPS*

Description

These are timecourse proteomics datasets output from temporal abundance experiments on THP-1 human leukaema cells stimulated with LPS at 0, 2, 4, 6, 12 and 24 hours.

Usage

```
data(lpsTimecourse_mulvey2021)
data(lpsTimecourse_rep1_mulvey2021)
data(lpsTimecourse_rep2_mulvey2021)
data(lpsTimecourse_rep3_mulvey2021)
data(psms_lpsTimecourse_rep1_mulvey2021)
data(psms_lpsTimecourse_rep2_mulvey2021)
data(psms_lpsTimecourse_rep3_mulvey2021)
```

Format

These data are instances of class MSnSet from package MSnbase.

Details

These are triplicate timecourse proteomics datasets output from temporal abundance experiments on THP-1 human leukaema cells stimulated with LPS. THP-1 total cell lysates were harvested at specific time-points of LPS stimulation; 0, 2, 4, 6, 12 and 24 hours. These were then digested, labelled with TMT 6plex, and analysed by LC-MS/MS on the Q-Exactive MS (Thermo). The resulting datasets were processed by Proteome Discoverer software (v2.1, Thermo) and imported and analysed in R using the packages MSnbase and pRoloc.

PSM-level and protein-level csv files are available in the codeextdata directory and have been imported as MSnSet instances using readMSnSet2. There are 3 replicates and the PSM-level and protein-level datasets are denoted accordingly.

The dataset lpsTimecourse_mulvey2021 contains the triplicates concatenated thus including only proteins common across all experiments, and is that which was used in the core analayis in the manuscipt by Mulvey et al 2021. The featureData columns contain the output (adjusted) p-values output from the temporal abundance analysis that was performed using limma and as described in Mulvey et al 2021.

Source

The data was generated by C. Mulvey at the Cambridge Centre for Proteomics. http://www.bio.cam.ac.uk/proteomics/.

Examples

```
data(lpsTimecourse_mulvey2021)
lpsTimecourse_mulvey2021
pData(lpsTimecourse_mulvey2021)
head(exprs(lpsTimecourse_mulvey2021))
```

---

moloneyTbBSF *Spatial proteomics datasets from two African trypanosome species*

Description

Spatial proteomics datasets from a hyperLOPIT experimental design on two on two African trypanosome species, *Trypanosoma brucei* and *Trypanosoma congolense*, which have been mapped across two life-stages.

Usage

```
data(moloneyTbBSF)
data(moloneyTbPCF)
data(moloneyTcBSF)
data(moloneyTcPCF)
```

Format

These data are instances of class MSnSet from package MSnbase.

Details

Protein function is intimately linked with localisation and as a consequence the subcellular distribution of a protein provides information on its role in the cell. We have optimised a method for resolving subcellular compartments in *Trypanosoma brucei* and *Trypanosoma congolense* and implemented it in the spatial proteomics strategy of hyperLOPIT (hyperplexed localisation of organelle proteins by isotope tagging) (Christoforou et al. 2016; Mulvey et al. 2017). Between the vertebrate and insect stages of these parasites, represented by bloodstream and procyclic forms respectively, we have detected over 7000 proteins in each species across three biological iterations. Of these, 6182 *T. brucei* proteins and 6324 *T. congolense* proteins are included in a spatial proteome characterisation (Trotter et al., 2010). Classification to 19-23 subcellular compartments was performed using a machine learning approach based on a T-augmented Gaussian mixture model (Crook et al. 2019; Crook et al. 2018). With 713-852 compartment marker proteins, this has yielded localisation information for 2504-2795 proteins in each organism.

The data (expression and feature variable) contain:

- MW: TriTrypDB based molecular mass (Aslett et al. 2010)
- Signal_peptide: TriTrypDB based signal peptide prediction.
- TM: TriTrypDB based predicted number of transmembrane domains.
- Curated_GO_Processes: TriTrypDB based curated gene ontology biological process term.
- Computed_GO_Processes: TriTrypDB based computed gene ontology biological process term.
- Curated_GO_Components: TriTrypDB based curated gene ontology cellular component term.
- Computed_GO_Components: TriTrypDB based computed gene ontology cellular component term.
- PFam_Description: TriTrypDB based Pfam description.
- HDBSCAN_cluster: Cluster number according to HDBSCAN unsupervised clustering (Campello et al. 2013)
- HDBSCAN_cluster_probability: Probability of membership associated with HDBSCAN clustering.
- NetGPI: Binary prediction of GPI anchor according to NetGPI (Gíslason et al. 2021)
- DeepLoc_location: Subcellular localisation predicted by DeepLoc (Almagro Armenteros et al. 2017)
- computed_pI: pI computed by pIR (Perez-Riverol et al. 2012, Audain et al. 2016)
- markers: The marker set used for TAGM-MAP classification of protein subcellular localisation.
- tagm.map.allocation: The TAGM-MAP prediction for the most probable subcellular allocation.
- tagm.map.probability: The posterior probability for the master protein subcellular allocations computed by TAGM-MAP.
- tagm.map.outlier: The posterior probability for the master protein to belong to the outlier component rather than any of the annotated components.
- tagm.map.localisation.probability: The localisation probability for the master protein to belong a subcellular class; defined as the product of the
- tagm.map.probability: and 1 - tagm.map.outlier
- tagm.map.localisation.prediction: The final prediction of the master protein subcellular localisation based on its localisation probability; only proteins with a localisation probability of greater than 99.9 percent and outlier probability of less than 5E-5 were retained.

- tagm.mcmc.allocation: The TAGM-MCMC prediction for the most probable subcellular allocation.
- tagm.mcmc.probability: The mean posterior probability for the master protein subcellular allocations computed by TAGM-MCMC.
- tagm.mcmc.probability.lowerquantile: The lower boundary to the equitailed 95-credible interval of tagm.mcmc.probability.
- tagm.mcmc.probability.upperquantile: The upper boundary to the equitailed 95-credible interval of tagm.mcmc.probability.
- tagm.mcmc.mean.shannon: A Monte-Carlo averaged Shannon entropy, which is a measure of uncertainty in the allocations.
- tagm.mcmc.outlier: The posterior probability for the master protein to belong to the outlier component rather than any of the annotated components.
- tagm.mcmc.joint: The posterior probability for the master protein allocation to each of the subcellular classes determined by TAGM-MCMC.

Source

The data was generated by N. Moloney at the Cambridge Centre for Proteomics. http://www.bio.cam.ac.uk/proteomics/.

References

Christoforou, A., Mulvey, C.M., Breckels, L.M., Geladaki, A., Hurrell, T., Hayward, P.C., Naake, T., Gatto, L., Viner, R., Martinez Arias, A., and Lilley, K.S. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. Nat Commun 7, 8992. 10.1038/ncomms9992.

Crook, O.M., Breckels, L.M., Lilley, K.S., Kirk, P.D.W., and Gatto, L. (2019). A Bioconductor workflow for the Bayesian analysis of spatial proteomics. F1000Research 8, 446. 10.12688/f1000research.18636.1.

Crook, O.M., Mulvey, C.M., Kirk, P.D.W., Lilley, K.S., and Gatto, L. (2018). A Bayesian mixture modelling approach for spatial proteomics. PLOS Computational Biology 14, e1006516. 10.1371/journal.pcbi.1006516.

Mulvey, C.M., Breckels, L.M., Geladaki, A., Britovsek, N.K., Nightingale, D.J.H., Christoforou, A., Elzek, M., Deery, M.J., Gatto, L., and Lilley, K.S. (2017). Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. Nat Protoc 12, 1110-1135. 10.1038/nprot.2017.026.

Trotter, M.W., Sadowski, P.G., Dunkley, T.P., Groen, A.J., and Lilley, K.S. (2010). Improved subcellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. Proteomics 10, 4213-4219.

Examples

```
## load the data
data(moloneyTbBSF)

## View a summary of the data MSnSet data container
moloneyTbBSF

## PCA plot of the data
library("pRoloc")
plot2D(moloneyTbBSF, fcol = "tagm.map.localisation.prediction",
       main = "PCA plot map of T. Brucei bloodstream form")

## Pass pre-computed t-SNE coords and plot the data
```

```
tsne_coords <- as.matrix(fData(moloneyTbBSF)[, c("Dim.1", "Dim.2")])
plot2D(tsne_coords, fcol = "tagm.map.localisation.prediction",
       method = "none", methargs = list(moloneyTbBSF),
       main = "t-SNE map of T. Brucei bloodstream form")
```

`mulvey2015` *Data from Mulvey et al. 2015*

Description

This is the data from Mulvey et al., *Dynamic proteomic profiling of extra-embryonic endoderm differentiation in mouse embryonic stem cells.* , Stem Cell. (PMID 26059426). See below for more details.

Usage

```
data(mulvey2015)
data(mulvey2015norm)
```

Format

The data are instances of class `MSnSet` from package `MSnbase`. This ending with `_se` are of class `SummarizedExperiment`.

Details

While not a spatial proteomics data, it was analysed with the `pRoloc` package.

During mammalian preimplantation development, the cells of the blastocyst's inner cell mass differentiate into the epiblast and primitive endoderm lineages, which give rise to the fetus and extra-embryonic tissues, respectively. Extra-embryonic endoderm (XEN) differentiation can be modeled in vitro by induced expression of GATA transcription factors in mouse embryonic stem cells. Here, we use this GATA-inducible system to quantitatively monitor the dynamics of global proteomic changes during the early stages of this differentiation event and also investigate the fully differentiated phenotype, as represented by embryo-derived XEN cells. Using mass spectrometry-based quantitative proteomic profiling with multivariate data analysis tools, we reproducibly quantified 2,336 proteins across three biological replicates and have identified clusters of proteins characterized by distinct, dynamic temporal abundance profiles. We first used this approach to highlight novel marker candidates of the pluripotent state and XEN differentiation. Through functional annotation enrichment analysis, we have shown that the downregulation of chromatin-modifying enzymes, the reorganization of membrane trafficking machinery, and the breakdown of cell-cell adhesion are successive steps of the extra-embryonic differentiation process. Thus, applying a range of sophisticated clustering approaches to a time-resolved proteomic dataset has allowed the elucidation of complex biological processes which characterize stem cell differentiation and could establish a general paradigm for the investigation of these processes.

References

Mulvey CM, Schr\"oter C, Gatto L, Dikicioglu D, Fidaner IB, Christoforou A, Deery MJ, Cho LT, Niakan KK, Martinez-Arias A, Lilley KS. Dynamic Proteomic Profiling of Extra-Embryonic Endoderm Differentiation in Mouse Embryonic Stem Cells. Stem Cells. 2015 Sep;33(9):2712-25. doi: 10.1002/stem.2067. Epub 2015 Jun 23. PubMed PMID: 26059426.

Examples

```
data(mulvey2015)
library("pRoloc")
plot2D(mulvey2015)

data(mulvey2015norm)
heatmap(exprs(mulvey2015))

library(SummarizedExperiment)
data(mulvey2015_se)
mulvey2015_se
```

---

nikolovski2012 *Meta-analysis from Nikolovski et al. (2012)*

Description

This is the data used in Nikolovksi et al. (2012). See below for details and references.

Usage

```
data(nikolovski2012)
data(nikolovski2012imp)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

These data are a concatenation of 4 LOPIT experiments. Experiments 1 and 2 are from Dunkley et al. 2006 (see also `dunkley2006`). Exepriments 3 and 4 are new.

In the LOPIT experiments by Dunkley et al. (2006), peripheral membrane proteins were removed by carbonate washing of the isolated membranes, while for experiments 3 and 4, no carbonate wash was performed and are, as such, enriched in peripheral and luminal proteins. See figure 1 in Nikolovski 2012 for a description of the design.

In `nikolovksi2012imp` missing values have been imputed using partial least-squares regression.

The training set used for the Naive Bayesian classifier is available as the `markers` feature meta-data. Note that Nikolovksi included a group of markers labelled 'others', which has been retained in these data sets. The results produced in this work are available in the `preds` feature variable (note that some organelles are marked with a '*', which is undefined here).

Source

Supporting Information on http://www.plantphysiol.org/content/160/2/1037.long, also available in the package's `extdata` directory.

References

Nikolovski N, Rubtsov D, Segura MP, Miles GP, Stevens TJ, Dunkley TP, Munro S, Lilley KS, Dupree P. *Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics.* Plant Physiol. 2012 Oct;160(2):1037-51. doi: 10.1104/pp.112.204263. Epub 2012 Aug 24. PMID: 22923678; PMCID: PMC3461528.

Dunkley TP, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, Griffin JL, Bessant C, Brandizzi F, Hawes C, Watson RB, Dupree P, Lilley KS. *Mapping the Arabidopsis organelle proteome.* Proc Natl Acad Sci U S A. 2006 Apr 25;103(17):6518-23. Epub 2006 Apr 17. PubMed PMID: 16618929; PubMed Central PMCID: PMC1458916.

Examples

```
data(nikolovski2012)
data(nikolovski2012imp)
table(is.na(nikolovski2012))
table(is.na(nikolovski2012imp))
phenoData(nikolovski2012)
table(fData(nikolovski2012)$markers)
all.equal(sort(featureNames(nikolovski2012)),
          sort(featureNames(nikolovski2012imp)))
library("pRoloc")
plot2D(nikolovski2012imp)
addLegend(nikolovski2012imp, where = "topright", bty = "n", cex =.7)
```

---

nikolovski2014 *LOPIMS data from Nikolovski et al. (2014)*

Description

This is the data used in Nikolovksi et al. (2014). See below for details and references.

Usage

```
data(nikolovski2014)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

Abstract: The proteomic composition of the Arabidopsis Golgi apparatus is currently reasonably well documented; however little is known about the relative abundances between different proteins within this compartment. Accurate quantitative information of Golgi resident proteins is of great importance: it facilitates a better understanding of the biochemical processes which take place within this organelle, especially those of different polysaccharide synthesis pathways. Golgi resident proteins are challenging to quantify since the abundance of this organelle is relatively low within the cell. In this study an organelle fractionation approach, targeting the Golgi apparatus, was combined with a label free quantitative mass spectrometry (MS), data-independent acquisition (DIA) method employing ion mobility separation known as LC-IMS-MSE (or HDMSE), to simultaneously localize proteins to the Golgi apparatus and assess their relative quantity. In total 102 Golgi localised proteins were quantified. These data provide new insight into Golgi apparatus organization and

demonstrate that organelle fractionation in conjunction with label free quantitative MS is a powerful and relatively simple tool to access protein organelle localisation and their relative abundances. The findings presented open a unique view on the organization of the plant Golgi apparatus, leading towards novel hypotheses centered on the biochemical processes of this organelle.

These data are a concatenation of 2 LOPIMS gradients, labelled gradient A and B, each with 10 fractions.

Source

Supplemental Data downloaded from http://www.plantphysiol.org/content/early/2014/08/13/pp.114.245589/suppl/DC1, also available in the package's extdata directory.

References

Nikolovski N, Shliaha PV, Gatto L, Dupree P, Lilley KS. Label free protein quantification for plant Golgi protein localisation and abundance. Plant Physiol. 2014 Aug 13. pii: pp.114.245589. [Epub ahead of print] PubMed PMID: 25122472.

Examples

```
data(nikolovski2014)
pData(nikolovski2014)
library("pRoloc")
plot2D(nikolovski2014)
addLegend(nikolovski2014, where = "topright", bty = "n", cex =.7)

A <- pData(nikolovski2014)$gradient == "A"
par(mfrow = c(1, 2))
plot2D(nikolovski2014[, A], main = "Gradient A")
plot2D(nikolovski2014[, !A], main = "Gradient B")
```

---

orre2019 *SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization*

Description

Data from 'SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization' Molecular cell.

Subcellular localization is a main determinant of protein function; however, a global view of cellular proteome organization remains relatively unexplored. We have developed a robust mass spectrometry-based analysis pipeline to generate a proteome-wide view of subcellular localization for proteins mapping to 12,418 individual genes across five cell lines. Based on more than 83,000 unique classifications and correlation profiling, we investigate the effect of alternative splicing and protein domains on localization, complex member co-localization, cell-type-specific localization, as well as protein relocalization after growth factor inhibition. Our analysis provides information about the cellular architecture and complexity of the spatial organization of the proteome; we show that the majority of proteins have a single main subcellular location, that alternative splicing rarely affects subcellular location, and that cell types are best distinguished by expression of proteins exposed to the surrounding environment. The resource is freely accessible via www.subcellbarcode.org.

Usage

```
data("orre2019a431")
data("orre2019h322")
data("orre2019hcc827")
data("orre2019hcc827gef")
data("orre2019hcc827rep1")
data("orre2019hcc827rep2")
data("orre2019hcc827rep3")
data("orre2019mcf7")
data("orre2019u251")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization Lukas Minus Orre, Mattias Vesterlund, Yanbo Pan, Taner Arslan, Yafeng Zhu, Alejandro Fernandez Woodbridge, Oliver Frings, Erik Fredlund, and Janne Lehtio https://doi.org/10.1016/j.molcel.2018.11.035

Examples

```
data(orre2019a431)
orre2019a431
pData(orre2019a431)
exprs(orre2019a431)[1:3,1:3]

library("pRoloc")
plot2D(orre2019a431,, main = "Orre 2019 A431")
```

---

pRolocdata *List of pRolocdata data sets*

Description

This function lists the data sets available in `pRolocdata` package by calling `data(package = "pRolocdata")`.

Usage

```
pRolocdata()
```

Author(s)

Laurent Gatto <lg390@cam.ac.uk>

References

See in the respective data sets' manual pages for references to publications.

Examples

```
pRolocdata()
```

pRolocmetadata *Extract pRoloc metadata*

Description

Extracts relevant metadata from an MSnSet instance. See README.md for a description and explanation of the metadata fields.

Usage

```
pRolocmetadata(x)
```

Arguments

x	A pRolocdata data.

Value

An instance of class pRolocmetadata.

Author(s)

Laurent Gatto

Examples

```
library("pRolocdata")
data(dunkley2006)
data(dunkley2006)
pRolocmetadata(dunkley2006)
```

rodriguez2012r1 *Spatial proteomics of human inducible goblet-like LS174T cells from Rodriguez-Pineiro et al. (2012)*

Description

Data from Rodriguez-Pineiro AM, van der Post S, Johansson ME, Thomsson KA, Nesvizhskii AI, Hansson GC. Proteomic study of the mucin granulae in an intestinal goblet cell model. J Proteome Res. 2012 Mar 2;11(3):1879-90. doi: 10.1021/pr2010988. Epub 2012 Feb 2. PubMed PMID:22248381; PubMed Central PMCID:PMC3292267.

Usage

```
data("rodriguez2012r1")
data("rodriguez2012r2")
data("rodriguez2012r3")
```

Details

As no marker were provided with the data, we transferred markers from the hyperLOPIT2015 (mouse) data using gene names to match between experiments. To validate our marker annotations, we compared the relative distributions of our markers (see figure below) to the PCA plot provided by the authors (Figure 3). Both show a similar separation of mitochondion/ER vs the rest along PC1 and ribosomes/lysosome vs rest along PC2. The data do not match exactly as the different marker protein are used.

Source

The supplementary file is pr2010988_si_003.xls. See scripts/rodriguez-pineiro2012.R for data preparation.

Examples

```
data(rodriguez2012r1)
data(rodriguez2012r2)
data(rodriguez2012r3)

library("pRoloc")
par(mfrow = c(2, 2))
plot2D(rodriguez2012r1)
addLegend(rodriguez2012r1, cex = .7, where = "topleft")
plot2D(rodriguez2012r2)
plot2D(rodriguez2012r3)


## compare to figure 3
dev.new()
plot2D(markerMSnSet(rodriguez2012r1),
       mirrorX = TRUE, mirrorY = TRUE,
       main = "Our markers")
addLegend(markerMSnSet(rodriguez2012r1), where = "bottomright")
```

Shin2020 — *Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers*

Description

Data from 'Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers'

Intracellular traffic between compartments of the secretory and endocytic pathways is mediated by vesicle-based carriers. The precise and complete proteomes of carriers destined for many organelles are ill-defined because the vesicular intermediates are transient, low-abundance and difficult to purify. Here, we combine vesicle relocalisation with organelle proteomics and Bayesian analysis to define the content of different endosome-derived vesicles destined for the trans-Golgi network (TGN). The golgin coiled-coil proteins golgin-97, golgin-245 and GCC88, shown previously to capture endosome-derived vesicles at the TGN, were individually relocalised to mitochondria and the content of subsequently re-routed vesicles was determined by organelle proteomics. Our findings revealed 45 integral and 51 peripheral membrane proteins re-routed by golgin-97, evidence for a distinct class of vesicles shared by golgin-97 and GCC88, and various cargoes specific to individual golgins. These results illustrate a general strategy for analysing intracellular sub-proteomes by combining acute cellular re-wiring with high-resolution spatial proteomics.

Usage

```
data("Shin2019MitoControlrep1")
data("Shin2019MitoGcc88rep1")
data("Shin2019MitoGol97rep1")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

References

Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers. John J.H. Shin, Oliver M. Crook, Alicia C. Borgeaud, Jerome Cattin-Ortola, Sew-Yeu Peak- Chew, Jessica Chadwick, Kathryn S. Lilley, Sean Munro.

Examples

```
data(Shin2019MitoControlrep1)
Shin2019MitoControlrep1
pData(Shin2019MitoControlrep1)
exprs(Shin2019MitoControlrep1)[1:3,1:3]

library("pRoloc")
plot2D(Shin2019MitoControlrep1, main = "Shin 2019 HeK 293T WT")
```

`stekhoven2014` *Data from Stekhoven et al. 2014*

Description

Proteomics data provide unique insights into biological systems, including the predominant subcellular localization (SCL) of proteins, which can reveal important clues about their functions. Here we analyzed data of a complete prokaryotic proteome expressed under two conditions mimicking interaction of the emerging pathogen Bartonella henselae with its mammalian host. Normalized spectral count data from cytoplasmic, total membrane, inner and outer membrane fractions allowed us to identify the predominant SCL for 82 proteins. The spectral count proportion of total membrane versus cytoplasmic fractions indicated the propensity of cytoplasmic proteins to co-fractionate with the inner membrane, and enabled us to distinguish cytoplasmic, peripheral inner membrane and bona fide inner membrane proteins. Principal component analysis and k-nearest neighbor classification training on selected marker proteins or predominantly localized proteins, allowed us to determine an extensive catalog of at least 74 expressed outer membrane proteins, and to extend the SCL assignment to 94% of the identified proteins, including 18 silico methods gave no prediction. Suitable experimental proteomics data combined with straightforward computational approaches can thus identify the predominant SCL on a proteome-wide scale. Finally, we present a conceptual approach to identify proteins potentially changing their SCL in a condition-dependent fashion.

Usage

```
data("stekhoven2014")
```

References

Stekhoven DJ, Omasits U, Quebatte M, Dehio C, Ahrens CH. *Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism.* J Proteomics. 2014 Mar 17;99:123-37. doi:10.1016/j.jprot.2014.01.015. Epub 2014 Jan 28. PubMed PMID: 24486812.

Examples

```
data(stekhoven2014)
library("pRoloc")
plot2D(stekhoven2014)
```

`tan2009` *LOPIT data from Tan et al. (2009)*

Description

This is the data from Tan et al., *Mapping organelle proteins and protein complexes in Drosophila melanogaster*, J Proteome Res. 2009 Jun;8(6):2667-78. See below for more details.

Usage

```
data(tan2009r1)
data(tan2009r2)
data(tan2009r3)
data(tan2009r1goCC)
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Details

This is a LOPIT experiment. Normalised intensities for proteins for four iTRAQ 4-plex labelled fractions are available for 3 replicates (r1, r2 and r3 respectively). The partial least square discriminant analysis results from the paper are available as `PLSDA` feature meta-data and the markers used in analysis are available as `markers` feature meta-data (Note: the ER and Golgi organelle markers were combined in original PLSDA analysis).

Replicate 1 was also used in testing the phenotype discovery algorithm from Breckels et al., *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation*, J Proteomics, *In Press.*, see `phenoDisco`. New phenotype clusters identified from algorithm application are available as `pd.2013` feature meta-data.

The `tan2009r1goCC` instance contains binary assay data. Its columns represent GO CC terms that have been observed for the object's features. A 1 indicates that a GO term has been associated to a given feature (protein); a 0 means not such association was found in the GO ontology.

Source

Supporting Information on http://pubs.acs.org/doi/full/10.1021/pr800866n

References

Mapping organelle proteins and protein complexes in Drosophila melanogaster. Tan DJ, Dvinge H, Christoforou A, Bertone P, Martinez Arias A, Lilley KS. J Proteome Res. 2009 Jun;8(6):2667-78. PMID: 19317464

Breckels LM, Gatto L, Christoforou A, Groen AJ, Lilley KS and Trotter MWB. *The Effect of Organelle Discovery upon Sub-Cellular Protein Localisation* J Proteomics. *In Press.*

Examples

```
data(tan2009r1)
tan2009r1
pData(tan2009r1)
head(exprs(tan2009r1))
# Organelle markers
table(fData(tan2009r1)$markers)
# PLSDA assignment results
table(fData(tan2009r1)$PLSDA)
```

---

thpLOPIT_lps_mulvey2021

Protein and PMS-level hyperLOPIT datasets from THP-1 human leukaema cells

---

Description

These are spatial proteomics datasets from a hyperLOPIT experimental design on the THP-1 human leukaema cell line.

Usage

```
data(thpLOPIT_lps_mulvey2021)
data(thpLOPIT_unstimulated_mulvey2021)
data(thpLOPIT_unstimulated_rep1_mulvey2021)
data(thpLOPIT_unstimulated_rep2_mulvey2021)
data(thpLOPIT_unstimulated_rep3_mulvey2021)
data(thpLOPIT_lps_rep1_mulvey2021)
data(thpLOPIT_lps_rep2_mulvey2021)
data(thpLOPIT_lps_rep3_mulvey2021)
data(psms_thpLOPIT_lps_rep1_set1)
data(psms_thpLOPIT_lps_rep1_set2)
data(psms_thpLOPIT_lps_rep2_set1)
data(psms_thpLOPIT_lps_rep2_set2)
data(psms_thpLOPIT_lps_rep3_set1)
data(psms_thpLOPIT_lps_rep3_set2)
data(psms_thpLOPIT_unstim_rep1_set1)
data(psms_thpLOPIT_unstim_rep1_set2)
data(psms_thpLOPIT_unstim_rep2_set1)
data(psms_thpLOPIT_unstim_rep2_set2)
data(psms_thpLOPIT_unstim_rep3_set1)
data(psms_thpLOPIT_unstim_rep3_set2)
```

Format

These data are instances of class `MSnSet` from package `MSnbase`.

Details

These datasets are from a spatiotemporal proteomic profiling experiment (Mulvey et al 2021) in which the dynamic pro-inflammatory response to lipopolysaccharide (LPS) in the THP-1 human leukaemia cell line is mapped.

Triplicate hyperLOPIT experiments using subcellular fractionation were conducted and fractions were digested at either 0h-LPS or 12h-LPS post-stimulation. For each replicate 20 fractions were selected and labelled with 2 x TMT 10plex then acquired with SPS-MS3 acquisition on the Orbitrap Fusion Lumos Tribrid instrument (Thermo). The resulting datasets were processed by Proteome Discoverer software (v2.1, Thermo) and analysed using the Bioconductor packages `MSnbase` and `pRoloc`.

PSM-level csv files are available in the codeextdata directory and have been imported as `MSnSet` instances using `readMSnSet2`. In total there are 12 PSM-level datasets, 6 for each conidition and 2 sets per replicate (2 x TMT-10plex). A maximum to 2 missing values per PSM was allowed and are designated as NA.

There are 8 protein-level datasets; 3 replicates for each condition, e.g. `thpLOPIT_lps_rep1_mulvey2021`, `thpLOPIT_lps_rep2_mulvey2021`, etc. then a final 2 datasets `thpLOPIT_lps_mulvey2021` and `thpLOPIT_unstim_mulvey2021` in which the 3 replicates per condition have been concatenated. This last 2 datasets form part of the main analysis in the manuscript from Mulvey et al 2021.

The protein-level data was generated from the PSM-data in the following steps: any PSMs with missing values were assessed by examining if there was any trend in missing values. Barplots of the data suggest the few missing values that appear accumulate in the first few fractions and we deduce they are not missing at random and on the whole reflect the gradient distributions. A left-cenored method was used for imputation using `MinDet` in `MSnbase`. PSMs were normalised by sum across the fractions and then combined to protein according to the median PSM per protein group.

Marker proteins were annotated based on the combined protein-level data in `thpLOPIT_lps_mulvey2021` and `thpLOPIT_unstimulated_mulvey2021` and can be found in the `fData` slot. A list of well annotated, unambiguous resident organelle marker proteins from 11 subcellular niches: mitochondria, ER, Golgi apparatus, lysosome, peroxisome, PM, nucleus, nucleolus, chromatin, ribosome and cytosol, were curated from the Uniprot database (The UniProt, 2017), Gene Ontology (Ashburner et al., 2000) and from mining the literature. Only proteins known to localise to a single location were included as markers. The processing script is `scripts/thpLOPIT2021.R`.

The `fData` slot of the two combined datasets also contains the results from the data analysis as described in Mulvey et al (2021).

The data (expression and feature variable) contain:

- UniProt Accession: found in the `featureNames` e.g. `featureNames(thpLOPIT_lps_mulvey2021)`. The Protein Group (no isoform information):Unique UniProt accession for quantified protein group reported by Proteome Discoverer (1% FDR).
- Expression data slot: Normalized TMT 10-plex reporter distributions representing the normalised abundance of each protein across the fractionation scheme for each experiment. Protein-level reporter ion values were calculated by taking the median of all quantifiable PSMs for the protein group, then normalized so that the sum of all 10 channels was equal to 1 before being concatenated across replicates.
- GN: Gene name for protein accession.
- Description: UniProt description for protein accession.

- Confidence_x: The confidence level of protein identification FDR determined in hyperLOPIT experiment. Only proteins with Medium (Q ² 5 %) and High (Q ² 1 %) FDR confidence levels were retained; Percolator v. 2.05 was used to determine FDR; for details, see L. KSll et al., Nat. Methods 2007, 4, 923-925, L. KSll et al., J. Proteome Res. 2008, 7, 29-34, and L. Kall et al., Bioinformatics 2008, 24, i42-i48.
- Coverage: Percentage of protein sequence covered by identified peptides for each experiment.
- Quantified Peptides: Number of quantified peptides for each experiment. Only peptides that were unique to a single protein group were used for quantification.
- Quantified PSMs: Number of quantified peptide-spectrum matches for each experiment.
- tagm.xxx.xx: `TAGM` allocation results from the Baysian T-augmented Gaussian Mixture modelling approach as described in Crook et al. (2018). See `??tagmMcmcTrain`.
- L2_distance: the natural L2 distance between the TAGM joint posterior probabilities
- non_movers: proteins predicted to not change location
- type1_translocation: proteins predicted from one organelle class in the unstimulated condition to a different organelle class in the LPS-stimulated dataset i.e. organelle to organelle
- type2_translocation: proteins precicted to move from an unknown localisation in the unstimulated dataset to a predicted organelle class in the LPS-stimulated dataset i.e. unknown to organelle
- type3_translocation: proteins predicted to move from a organelle localisation in the unstimulated dataset to an unknown location in the LPS-stimulated dataset i.e. organelle to unknown
- type4_translocation: a translocation event within the unknown class i.e. a protein that exhibits a large change between posterior probabilities in both conditions and is classified to an unknown location. For more information please see Mulvey et al 2021.
- markers: annotated protein location based on the combined protein-level data used for training the `TAGM` MCMC classifier.
- localisation.prob: assignment score, product of the `tagm.mcmc.probability` * 1 - `tagm.mcmc.outlier`.
- localisation.pred: predicted localisation as filtered by the `localisation.prob`. Proteins were assigned the localisation predicted from `tagm.mcmc.allocation` if their `localisation.prob` was lower than .99 (1% FDR)

Source

The data was generated by C. Mulvey and L. Breckels in the Cambridge Centre for Proteomics. http://www.bio.cam.ac.uk/proteomics/.

References

Using hyperLOPIT to perform high-resolution mapping of the spatial proteome Mulvey, Claire M and Breckels, Lisa M and Geladaki, Aikaterini and Britovšek, Nina Kočevar and Nightingale, Daniel J H and Christoforou, Andy and Elzek, Mohamed and Deery, Michael J and Gatto, Laurent and Lilley, Kathryn S Nature Protocols 12, 1110–1135 (2017). https://doi.org/10.1038/nprot.2017.026

A Bayesian Mixture Modelling Approach For Spatial Proteomics Oliver M Crook, Claire M Mulvey, Paul D. W. Kirk, Kathryn S Lilley, Laurent Gatto bioRxiv 282269; doi: https://doi.org/10.1101/282269

Examples

```
## load a THP data (unstimulated combined tripcate)
data(thpLOPIT_unstimulated_mulvey2021)
thpLOPIT_unstimulated_mulvey2021
```

```
pData(thpLOPIT_unstimulated_mulvey2021)
head(exprs(thpLOPIT_unstimulated_mulvey2021))

## simple protocol for combining psm to protein
## load LPS stimulated data for replicate 1

library("MSnbase")
data(psms_thpLOPIT_lps_rep1_set1)
data(psms_thpLOPIT_lps_rep1_set2)

## impute missing values with "MinDet"
set1 <- impute(psms_thpLOPIT_lps_rep1_set1, method = "MinDet")
set2 <- impute(psms_thpLOPIT_lps_rep1_set2, method = "MinDet")

## normalise to sum
set1 <- normalise(set1, "sum")
set2 <- normalise(set2, "sum")

## combine to protein
set1 <- combineFeatures(set1,
       groupBy = fData(set1)$Master.Protein.Accessions,
       method = median)
set2 <- combineFeatures(set2,
       groupBy = fData(set2)$Master.Protein.Accessions,
       method = median)

## update fvarLabels for set 2 to differentiate them from set 1
set2 <- updateFvarLabels(set2)

## combine sets to form one replicate
xx <- combine(set1, set2)

## keep on proteins common in both sets
xx <- filterNA(xx)

library("pRoloc")
plot2D(thpLOPIT_unstimulated_mulvey2021, main = "Protein-level unstimulated data")
```

`ToxoLopit` *Whole-cell spatial proteome of Toxoplasma: molecular anatomy of an apicomplexan cell*

Description

Data from 'Whole-cell spatial proteome of Toxoplasma: molecular anatomy of an apicomplexan cell'

Apicomplexan parasites are the causative agents of major human diseases and food insecurity and owe their considerable success to novel, highly specialized cell compartments and structures. These adaptations facilitate the recognition and non-destructive penetration of their host cells, and elaborate reengineering of these cells to promote growth, dissemination and active countering of host defense responses. The evolution of apicomplexan compartments is concomitant with great proteomic novelty that defines these new cell organizations and functions and, hence, approximately

half of their proteins are unique and uncharacterized. Consequently, apicomplexan cells are relatively poorly understood. Here we employ the hyperLOPIT cell spatial proteomic method to the apicomplexan Toxoplasma gondii and define the steady-state subcellular location of thousands of proteins simultaneously giving comprehensive definition to these cells and their compartments. These data, moreover, provide new insight into the spatial organizations of protein expression, adaptation to hosts, and the underlying evolutionary trajectories of these parasites.

Usage

```
data("Barylyuk2020ToxoLopit")
```

Format

The data is an instance of class `MSnSet` from package `MSnbase`.

Examples

```
data(Barylyuk2020ToxoLopit)
Barylyuk2020ToxoLopit
pData(Barylyuk2020ToxoLopit)
exprs(Barylyuk2020ToxoLopit)[1:3,1:3]

library("pRoloc")
plot2D(Barylyuk2020ToxoLopit, main = "Davies 2018 HeLa - wt")
```

---

trotter20010 *LOPIT data sets used in Trotter et al. (2010)*

Description

The two Arabidobsis LOPIT data sets `trotter2010shallow` and `trotter2010steep` have been used in Trotter et al. (2010) to illustrate improvement of sub-cellular resolution upon data fusion. The data have originally been published in Dunkley et al. (2006) and Sadowski et al. (2008), respectively.

The feature metadata contains the cellular compartment from TAIR8 and the pRoloc Arabidopsis markers (see `pRolocmarkers`).

Usage

```
data(trotter2010)
data(trotter2010shallow)
data(trotter2010steep)
```

Format

The data are instances of class `MSnSet` from package `MSnbase`. `trotter2010` corresponds to the combined steep and shallow data.

Source

Supporting information available on http://onlinelibrary.wiley.com/doi/10.1002/pmic.201000359/abstract

References

Trotter MWB, Sadowski PG, Dunkley TPJ, Groen AJ and Lilley KS. *Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions.* Proteomics 2010 10(23):4213-4219. PMID 21058340.

Sadowski PG, Groen AJ, Dupree P and Lilley KS. *Sub-cellular localization of membrane proteins.* Proteomics 2008 8(19):3991-4011. PMID 18780351.

Examples

```
library(pRoloc)
## Replication of figure 4 from Trotter et al.
## individual data sets
data(trotter2010)
data(trotter2010steep)
data(trotter2010shallow)

par(mfrow = c(2,3))
plot2D(trotter2010shallow, fcol = "TAIR8", main = "Shallow (TAIR8)")
plot2D(trotter2010steep, fcol = "TAIR8", main = "Steep (TAIR8)")
plot2D(trotter2010, fcol = "TAIR8", main = "Combined (TAIR8)")
addLegend(trotter2010, where = "bottomleft", fcol = "TAIR8", ncol = 2)
plot2D(trotter2010shallow, main = "Shallow (markers)")
plot2D(trotter2010steep, main = "Steep (markers)")
plot2D(trotter2010, main = "Combined (markers)")
addLegend(trotter2010, where = "bottomleft", ncol = 2)
```

---

yeast2018 *Saccharomyces cerevisiae spatial proteomics (2018)*

Description

Data from 'The subcellular organisation of Saccharomyces cerevisiae' (submitted).

This dataset represents four biological replicate hyperLOPIT experiments performed in Saccharomyces cerevisiae cultured to early-mid exponential phase, in synthetic media with glucose as sole carbon source (SD-His media (Breker et al 2013)). These were carried out to produce a map of the spatial proteome of this organism under no-perturbed conditions. The associated quantitation data from these experiments were combined using the method described in reference. This dataset contains quantitative information for 2,847 proteins that were common across our four biological replicate experiments and information regarding localisation for all of the proteins in the combined experiment. Overall this dataset describes 936 proteins that localise to one of 12 subcellular locations in S. cerevisiae under our experimental conditions.

Usage

```
data("yeast2018")
```

Examples

```
data(yeast2018)

library("pRoloc")
par(mfrow = c(1, 2))
plot2D(yeast2018, main = "Markers")
addLegend(yeast2018, where = "bottomleft", cex = .7)
plot2D(yeast2018, fcol = "predicted.location", main = "Localisation")
```

Index