Package 'pRolocdata'

April 11, 2019

Type Package

Title Data accompanying the pRoloc package

Version 1.20.0

Author Laurent Gatto, Olive M. Crook and Lisa M. Breckels

Maintainer Laurent Gatto <lg390@cam.ac.uk>

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 (>= 2.15), MSnbase

Imports Biobase, utils

Suggests pRoloc (>= 1.13.8), testthat

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

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

git_branch RELEASE_3_8

git_last_commit 229fec9

git_last_commit_date 2018-10-30

Date/Publication 2019-04-11

R topics documented:

`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 condidions (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")
```

---

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

Format

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

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)")
```

`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)
```

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))
```

---

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 `itzhak2016`).

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

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)
```

---

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)")
```

---

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 is an instance of class MSnSet from package MSnbase.

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.

Source

Supporting Information on

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))
```

---

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")
```

---

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")
```

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)
```

---

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