Babraham Bioinformatics

Proteomics Exercises

Version 2024-09

Licence

Exercise 1: Obtaining Raw Data

In this exercise we're going to look for the data for a particular study which will be the data we're going to use for the subsequent exercises.

We're going to find both the data and the reference sequences we'd need to be able to analyse this study.

Finding Data in Proteome Central

Go to https://proteomecentral.proteomexchange.org/ in your browser.

We want to find a yeast (Saccharomyces cerevisiae) study performing label free quantitation.

From the full list of studies you can use the links on the left to limit your view to only yeast studies

How many yeast datasets are present in the repository?

Use the search box to search for eif2a – this should now restrict your results to a single study (PXD043985) which is the dataset we are going to use.

Click on the accession code to go through to the details page for the study. Check that you can find the following information:

- When was this data submitted?
- What sort of mass spec generated the data?
- Does this data include measurements of post-translational modifications?
- Which underlying database does this data come from?

You will see that the data in proteome central doesn't give any details of the samples deposited, nor the files which are available. For those you'll need to move through to the actual database in which the data resides.

Collecting data from PRIDE

Use the link on proteome central to move through to the PRIDE entry for this data (https://www.ebi.ac.uk/pride/archive/projects/PXD043985). Read through the details of the submission to see what samples you expect to see.

Have a look and see how the data was originally processed and analysed. Which program was used for the searches?

Look at the list of files available to download. Can you identify the original raw data, and the quantified results? How many samples are there, and can you match these to those described in the text at the top? In this study there is an "experimentalDesignTemplate.txt" file, which is not a required file. Have a look at the contents of this file to see if that helps in interpreting which file is which.

If you wanted more details about the study then can you find the paper which describes this study.

Finding Uniprot Reference Proteomes

If we wanted to re-process this data then we will need a reference proteome against which to search. In the project description can you see an exact description of which sequences were originally used to process the data?

We are going to find a suitable reference sequence set in Uniprot so that we can reprocess the data.

Go to https://www.uniprot.org/proteomes/ in your browser.

See if you can find the Saccharomyces cerevisiae proteome.

- How many proteins are in this proteome?
- How does this compare to the number of genes?

You will see that you can download the full proteome from

https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000002311/UP000002311_559292.fasta.gz

Exercise 2: Examining QC Reports

Examining QC Reports

If you visit:

https://www.bioinformatics.babraham.ac.uk/training/proteomics/Proteomics_Course_Data/QC%20Reports/

You will see links to a number of PTXQC reports which you can examine.

- **ptxqc_demo_report.html** is the demo file which comes with the program
- **ptxqc_yeast_report.html** is from PXD043985 and compares a full proteome to an affinity pulldown with eIF2a
- **ptxqc_worm_report.html** is from PXD025694 and compares WT worms to cey3 KO worms
- **ptxqc_mouse_report.html** is from PXD003155 and compares WT and ATXN2 knockdown animals

For the mouse sample we also have ptxqc_mouse_relaxed_report.html which is a reprocessed version of the same file following identification of a problem in the original processing.

You can have a look at all of these datasets to see if there are any issues you might want to talk to whoever created the data about, or take into account during your analysis.

Demo Report

This is an example report which the authors of the program distribute. You can look through the different plots which are provided and read the information associated with each of them. Some of the plots show properties which are very specific to the way the mass spec was run to create the data, but others show the properties of the samples themselves, or the output from the search results.

Yeast Report

There are a few things to note on this report. You will see that the biggest flags on the QC are due to the low total number of peptides and proteins detected. The expected value here was probably set on a more complex proteome such as human or mouse. The lower number of data points in this data is going to largely be driven by the fact that this is a yeast, with a much less complex proteome.

You will also see in some of the plots that there are two distinct sets of samples. One has much lower complexity, signal intensity and higher contamination that the other. Given the design of the experiment can you see why this could occur?

Worm Report

The worm data again shows a somewhat lower amount of data than the QC program would like, but this will also be because of the lower proteome complexity in worm.

The samples here are much more homogeneous than the yeast, but you should see that one sample is behaving somewhat differently to the others. This would probably be something you'd want to note in your later analysis in case any changes you saw were being driven by the oddly behaving sample.

Mouse Report

In the mouse report we see problems with the alignment of the LC timings in the different runs. Look at the alignment plots and see if the green line appears to run through the centre of the flat portion of the run. If the difference in retention times is too large then maxquant may not have found the correct offset for the run.

In the relaxed report we reran the analysis but allowed 30mins difference in retention time, rather than the default 10. Have a look at the equivalent plots in the relaxed report and see if allowing the additional flexibility helped. Do you think there might still be problems with these samples?

Exercise 3: Running MSstats Shiny

Introduction

In this exercise we're going to use the simple automated interface to running MSstats. The Shiny interface provides a simple web front end to analysing the data which makes it very easy to run but gives you little control over the specifics of the different steps, nor does it allow you to explore the data in other ways.

Launching the MSstats Shiny interface

To launch the interface you need to log into your RStudio server, then in console you need to run this single command.

```
MSstatsShiny::launch_MSstatsShiny(host="0.0.0.0", port=8080)
```

You will see a new window open, which looks greyed out. You **need to leave this window open** but this is ***NOT*** the window in which you will be running the analysis.

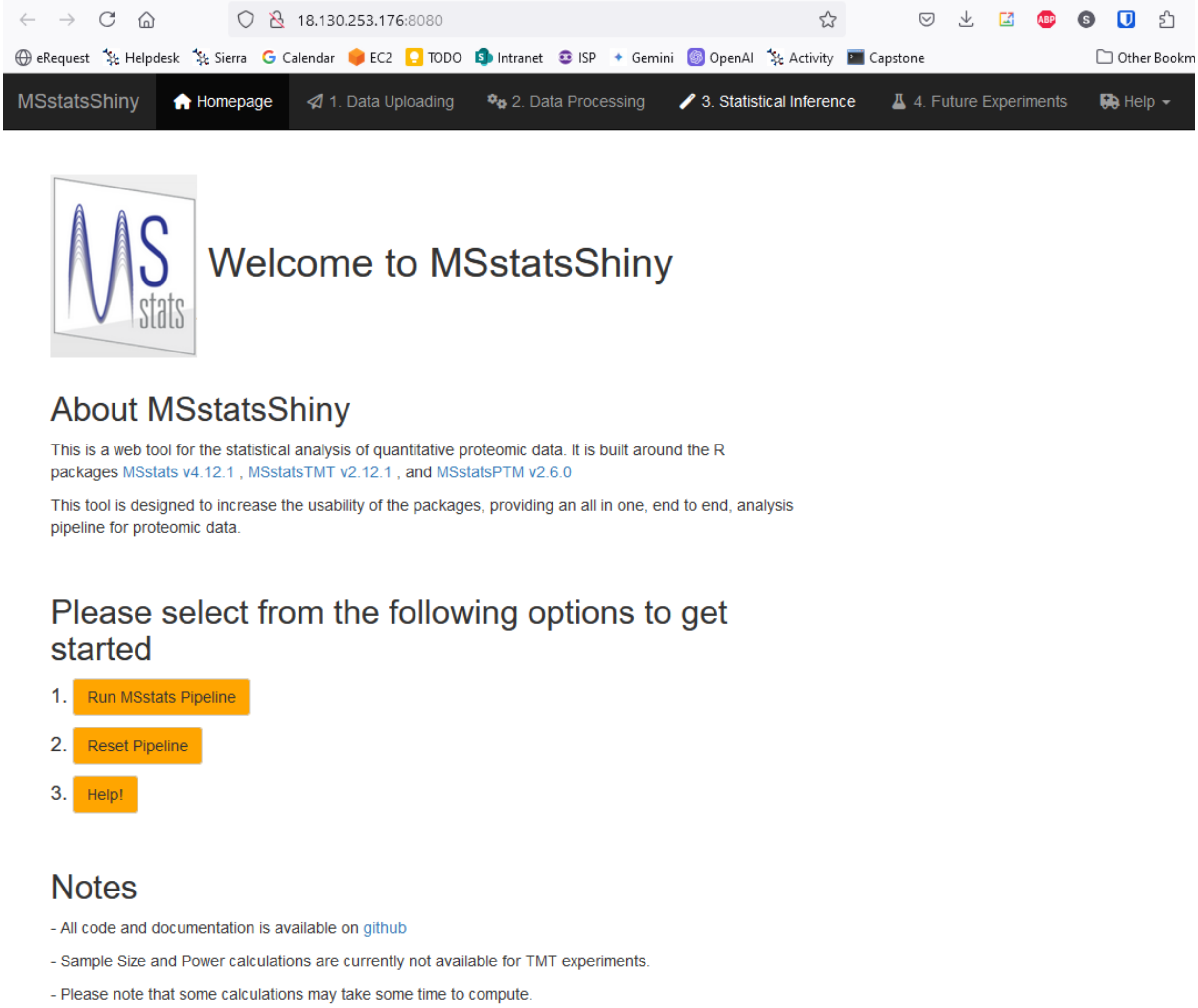
Instead you need to open a new browser window and open a new URL which is related to the address of the server you are using.

The address you need is

http://a.b.c.d:8080

Where a.b.c.d is the address of your server, so if your server is 18.130.253.176 then you need to go to http:// 18.130.253.176:8080

You should see something like this:



Welcome to MSstatsShiny

About MSstatsShiny

This is a web tool for the statistical analysis of quantitative proteomic data. It is built around the R packages MSstats v4.12.1 , MSstatsTMT v2.12.1 , and MSstatsPTM v2.6.0

This tool is designed to increase the usability of the packages, providing an all in one, end to end, analysis pipeline for proteomic data.

Please select from the following options to get started

1. Run MSstats Pipeline
2. Reset Pipeline
3. Help!

Notes

- All code and documentation is available on github

- Sample Size and Power calculations are currently not available for TMT experiments.

- Please note that some calculations may take some time to compute.

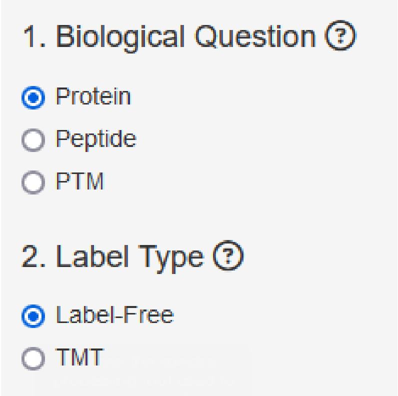
Running and analysis in the program

Start the analysis

To start the analysis you need to press the button. You will then need to set the options for the type of experiment you're running and the type of data you're using.

Select the analysis options

We are doing a label free MaxQuant dataset, which we want to analyse at the protein level.

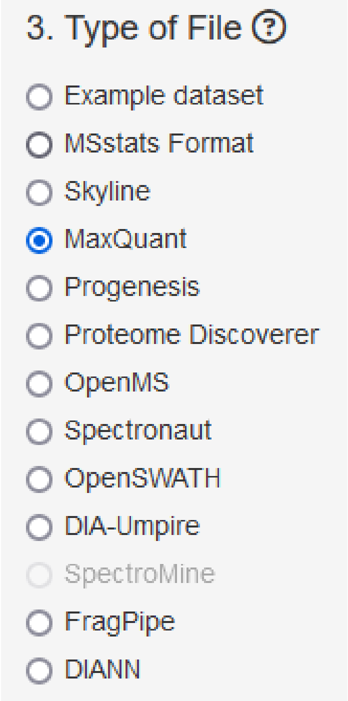


1. Biological Question ?

● Protein
○ Peptide
○ PTM

2. Label Type ?

● Label-Free
○ TMT



3. Type of File ?

○ Example dataset
○ MSstats Format
○ Skyline
● MaxQuant
○ Progenesis
○ Proteome Discoverer
○ OpenMS
○ Spectronaut
○ OpenSWATH
○ DIA-Umpire
○ SpectroMine
○ FragPipe
○ DIANN

Upload the data files

You will need to upload 3 files, which can be found in the MSstats_Shiny folder of the data file you were given.

1. `evidence.txt` comes from MaxQuant and holds the quantitated peptide values
2. `proteinGroups.txt` comes from MaxQuant and says how peptides are joined into protein groups
3. `annotation.csv` is not provided and is an annotation file you would need to create

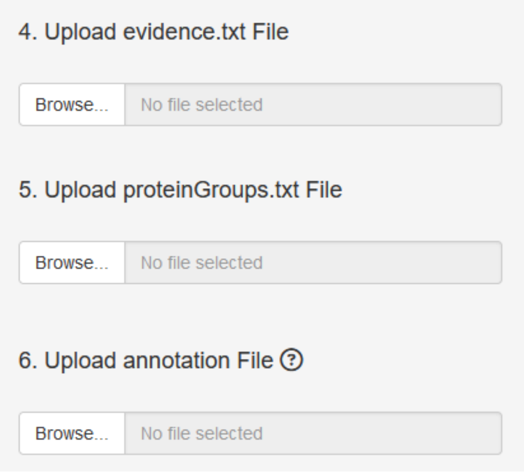
This is what the annotation file looks like:

Raw.file	Condition	BioReplicate	IsotypeLabelType
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	ProtTot	1	L
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	ProtTot	2	L
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	ProtTot	3	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep1	TAP	1	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep2	TAP	2	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep3	TAP	3	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	ProtTot	4	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep2	ProtTot	5	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	ProtTot	6	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep1	TAP	4	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep2	TAP	5	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep3	TAP	6	L

The columns are fairly self-explanatory. The Raw.file column names must match the names used in the evidence file. The groups can be any names you like and will tell the program how to combine the raw files into biological conditions. The replicates should just be an increasing integer within each group.

For label free quantitation then the IsotypeLabelType should always be L (light) because none of the samples were labelled with heavy isotopes. Other mass spec data types (eg SILAC) would require that you say which samples were labelled with which isotopes.

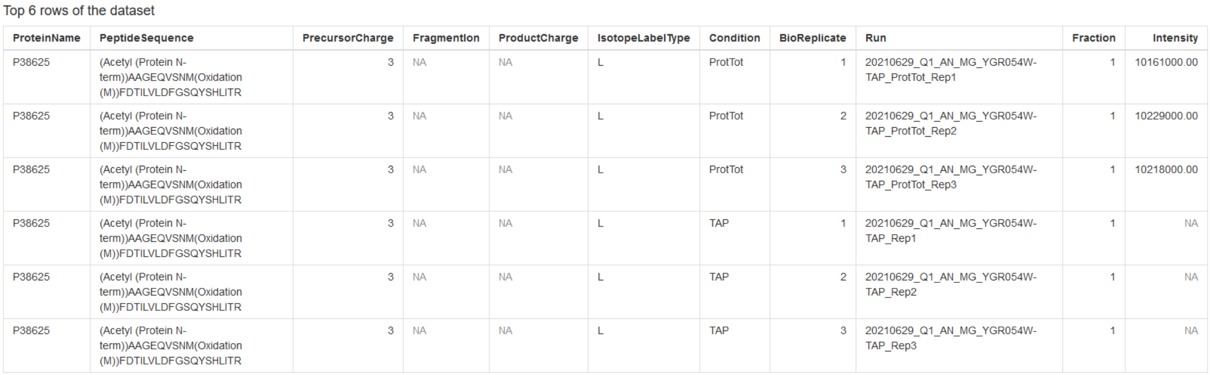
You can use the upload fields at the bottom to upload these three files to the system


4. Upload evidence.txt File
Browse...
No file selected
5. Upload proteinGroups.txt File
Browse...
No file selected
6. Upload annotation File ?
Browse...
No file selected

Other pre-filtering options

You have the option to turn on some additional filtering which will make your data smaller, but will remove some of the less well observed proteins from the data. The protein identifications most likely to be incorrect are those which are supported by only a single peptide, so these are often removed. We'll leave these values set as default, but be aware that these more tenuous identifications will still be present.

Press the Upload Data button to complete the upload. You will see a summary of the first few rows of data to check it did things correctly.



Top 6 rows of the dataset

ProteinName	PeptideSequence	PrecursorCharge	FragmentIon	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Fraction	Intensity
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	ProtTot	1	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	1	10161000.00
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	ProtTot	2	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	1	10229000.00
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	ProtTot	3	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	1	10218000.00
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	TAP	1	20210629_Q1_AN_MG_YGR054W-TAP_Rep1	1	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	TAP	2	20210629_Q1_AN_MG_YGR054W-TAP_Rep2	1	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILVLDFGSQYSHLITR	3	NA	NA	L	TAP	3	20210629_Q1_AN_MG_YGR054W-TAP_Rep3	1	NA

Scroll to the top of the page and press the Next step button.

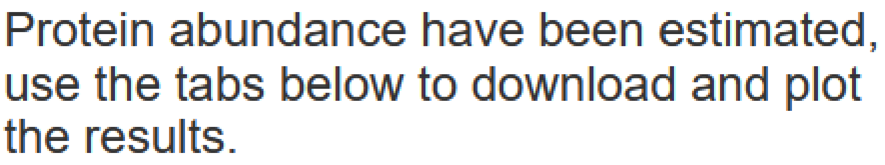
Running protein summarisation

Now the program is going to use the information in protein groups along with the quantitation in the evidence file to make protein level quantitations.

From the bar on the left you can see the options you have.

- The log base won’t really affect your results, but log2 values will be more sensible numbers (log10 numbers get very small very quickly)
- We’ll leave normalisation at it’s default value of “equalize medians” but we’ll look at more options in the next exercise
- We will use all features since we won’t have a huge number to start with
- We will impute missing values
- We will use the default TMP summarisation (we don’t have a choice about this anyway)
- We will not remove runs with >50% missing values. We should have detected these in the QC stage anyway

Once you’ve set the options you can press Run protein summarization to perform the calculation. It will take a minute or two to complete. When it’s finished the text at the top of the page should change to show



Protein abundance have been estimated, use the tabs below to download and plot the results.

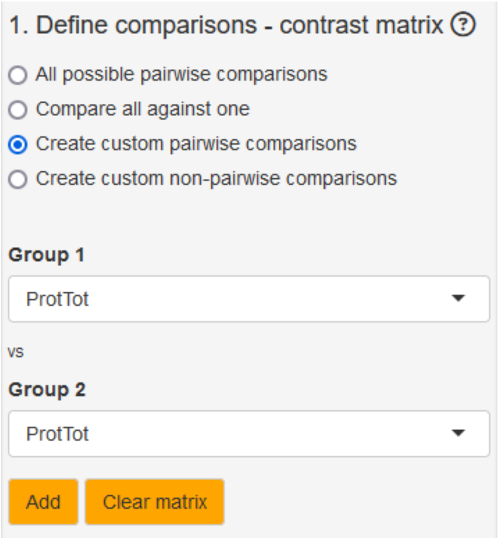
There are options to download tables containing the summarised quantitations which might be useful if you wanted to take this data into another analysis program.

There is also a tab which allows you to make summarised plots but these are drawn per protein, so they're only really useful if you know which protein you want to see. You can pick some at random to see what you get but these will be more useful later.

Scroll to the top of the page and press the Next step button.

Perform a statistical comparison

Now you need to define the comparison you want to make. In our case this is simple because we only have two conditions so there is only one comparison to make (ProtTot vs TAP). You can either say you want to make all comparisons, or you can make a custom comparison and select the two groups. The only difference is that you can select the order for the comparison which will reflect the direction of change (whether you want up in TAP to be positive or negative). The p-values will be the same in either case.



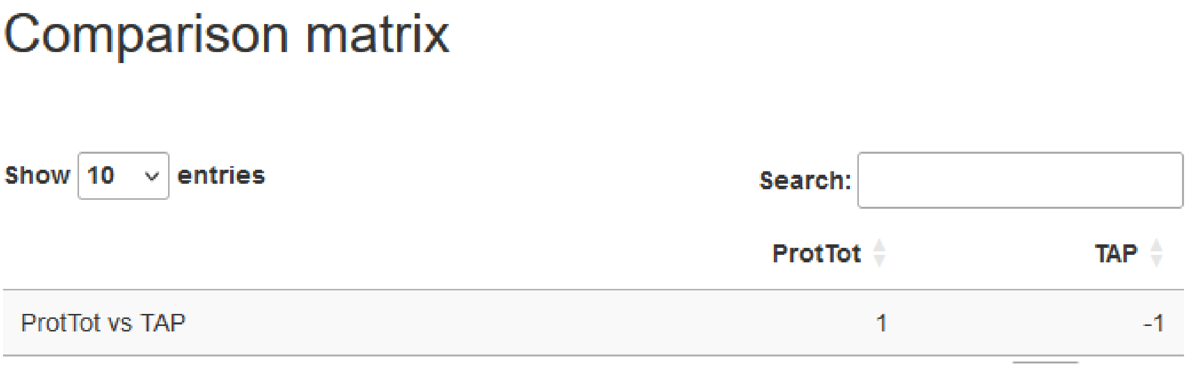
1. Define comparisons - contrast matrix ?

○ All possible pairwise comparisons
○ Compare all against one
● Create custom pairwise comparisons
○ Create custom non-pairwise comparisons

Group 1: ProtTot

vs

Group 2: ProtTot



Comparison matrix

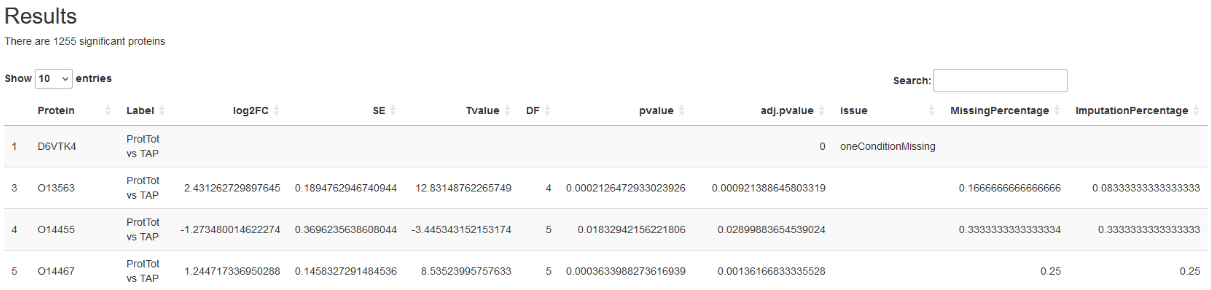
Show 10 entries

Search:

	ProtTot	TAP
ProtTot vs TAP	1	-1

Once you've created the contrast matrix you can press Start to run the analysis. Again, this will take a minute or so.

You will then see a table of results appear. You can sort this table by any of the columns. The useful ones to sort will be adj.pvalue or log2FC



Results

There are 1255 significant proteins

Show 10 entries

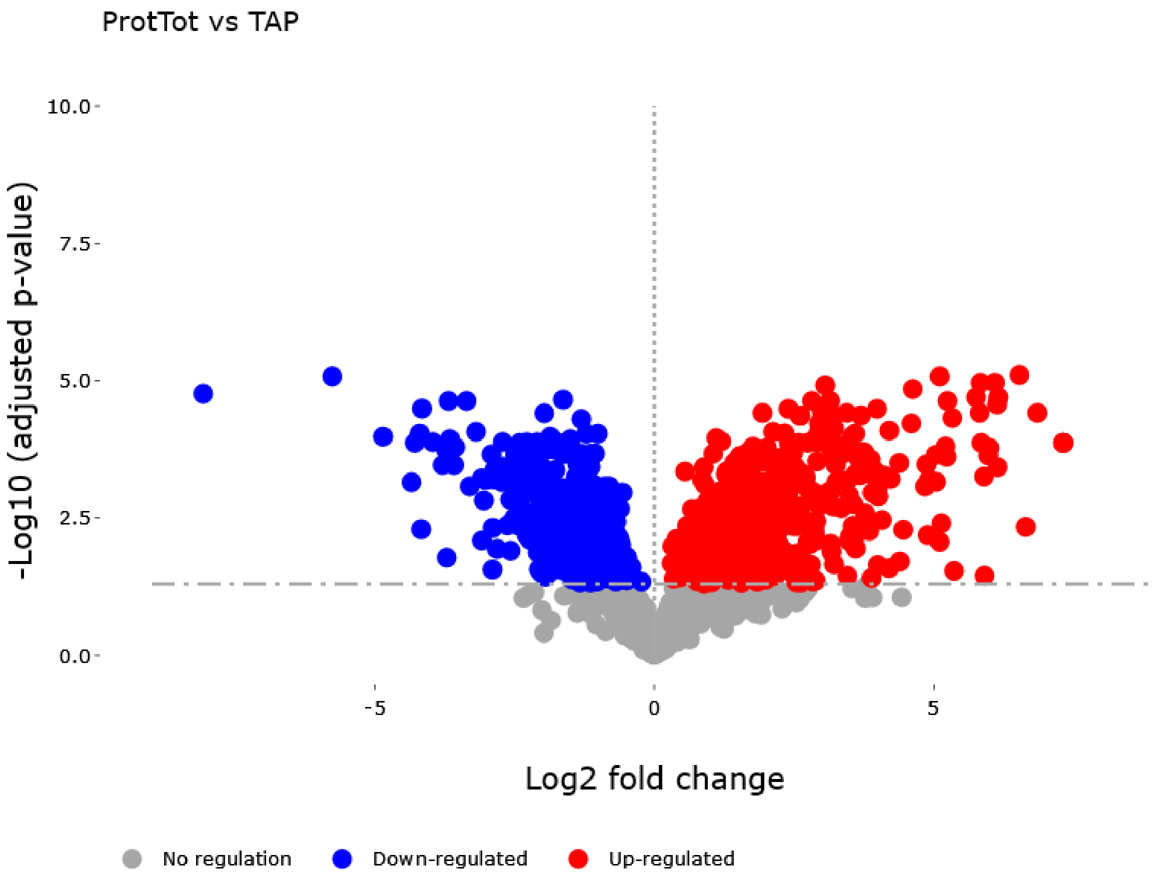
Search:

	Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	issue	MissingPercentage	ImputationPercentage
1	D6VTK4	ProtTot vs TAP						0	oneConditionMissing		
3	O13563	ProtTot vs TAP	2.431262729897645	0.1894762946740944	12.83148762265749	4	0.0002126472933023926	0.000921388645803319		0.1666666666666666	0.08333333333333333
4	O14455	ProtTot vs TAP	-1.273480014622274	0.3696235638608044	-3.445343152153174	5	0.01832942156221806	0.02899883654539024		0.3333333333333334	0.3333333333333333
5	O14467	ProtTot vs TAP	1.244717336950288	0.1458327291484536	8.53523995757633	5	0.0003633988273616939	0.00136166833335528		0.25	0.25

Try sorting the table and have a look at some of the top hits. Do they change in the same direction (is log2FC always either positive or negative?). Why do lots of hits have no reported log2FC and a zero value for adj.pvalue?

Draw a volcano plot

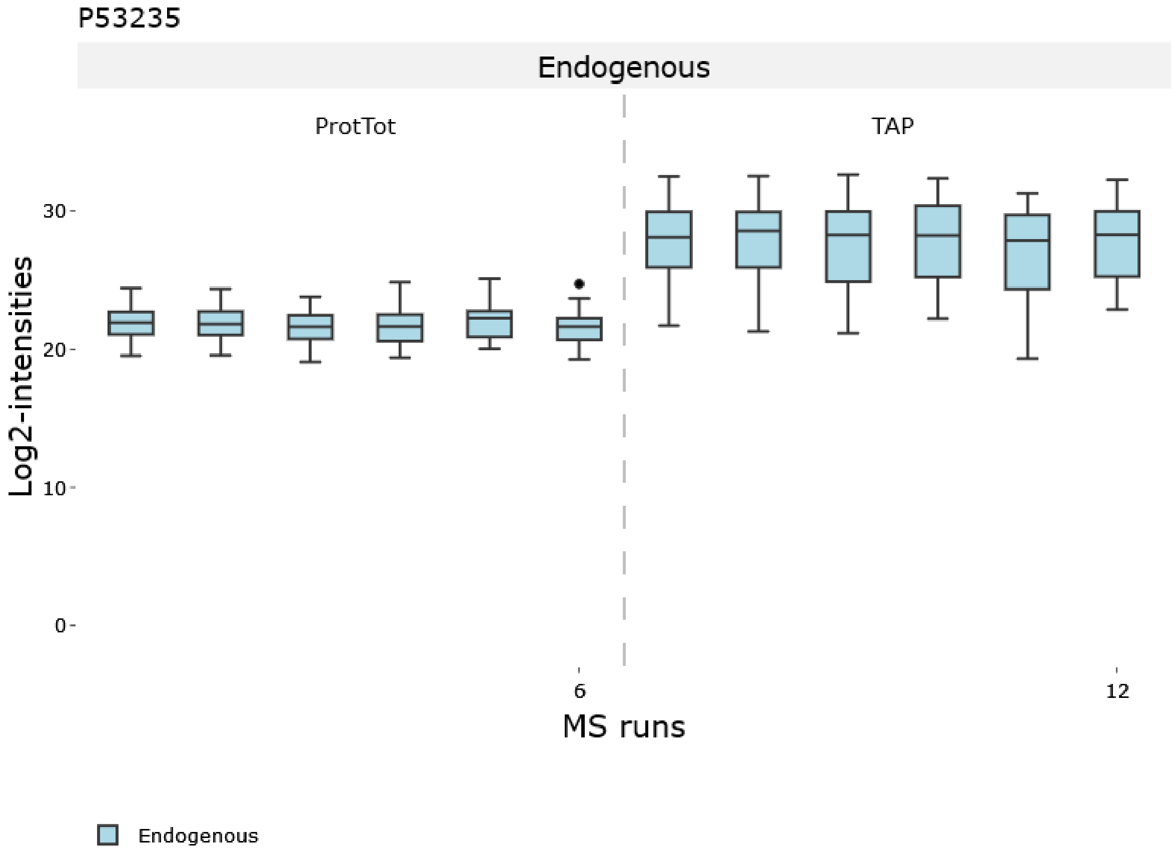
At the bottom of the results panel you have the option to draw some additional plots. Use this to draw a volcano plot.

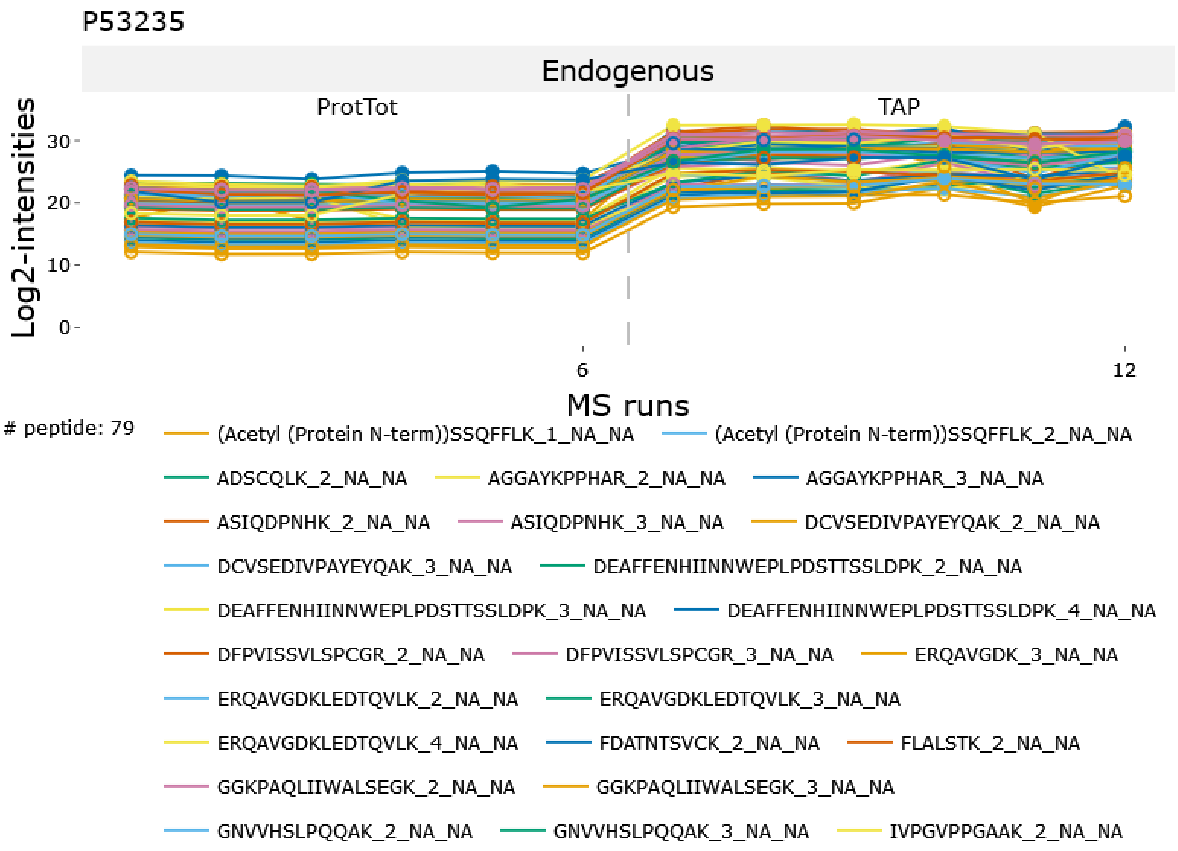


You should be able to go back and draw other plots, but sadly this appears to be broken in the current version of the program (you get an error saying “Cannnot generate multiple plots in a screen”

Look at the details for top hits

From the results table you should be able to find some of the top hits reported in the analysis. You can take these accession and then go back to the previous “Data Processing” tab and then look at either the Quality Control Plots or the profile plots for any of the hits to see how much evidence was present for them and how consistent the changes were between replicates.




P53235
Endogenous
ProtTot
TAP
Log2-intensities
30
20
10
0
6
12
MS runs
# peptide: 79
(Acetyl (Protein N-term))SSQFFLK_1_NA_NA
(Acetyl (Protein N-term))SSQFFLK_2_NA_NA
ADSCQLK_2_NA_NA
AGGAYKPPHAR_2_NA_NA
AGGAYKPPHAR_3_NA_NA
ASIQDPNHK_2_NA_NA
ASIQDPNHK_3_NA_NA
DCVSEDIVPAYEYQAK_2_NA_NA
DCVSEDIVPAYEYQAK_3_NA_NA
DEAFFENHIINNWEPLPDSTTSSLDPK_2_NA_NA
DEAFFENHIINNWEPLPDSTTSSLDPK_3_NA_NA
DEAFFENHIINNWEPLPDSTTSSLDPK_4_NA_NA
DFPVISSVLSPCGR_2_NA_NA
DFPVISSVLSPCGR_3_NA_NA
ERQAVGDK_3_NA_NA
ERQAVGDKLEDTQVLK_2_NA_NA
ERQAVGDKLEDTQVLK_3_NA_NA
ERQAVGDKLEDTQVLK_4_NA_NA
FDATNTSVCK_2_NA_NA
FLALSTK_2_NA_NA
GGKPAQLIIWALSEGK_2_NA_NA
GGKPAQLIIWALSEGK_3_NA_NA
GNVVHSLPQQAK_2_NA_NA
GNVVHSLPQQAK_3_NA_NA
IVPGVPPGAAK_2_NA_NA

Exercise 4: Analysing Data in R

Introduction

For this final exercise we are going to use the same data as we used for the MSStats Shiny exercise but we will look at the additional levels of control and visualisation we can do if we step outside the Shiny interface and work directly in R.

The basic steps of the analysis will be the same as before but we can do further examinations at each intermediate step, and we can look more carefully at the properties of the data after each step.

You can interact with this exercise in one of two ways.

1. You can open up your RStudio server environment, and in there you should see an R notebook document called “msstats.Rmd” in which the code for the analysis is laid out. You can work through this document, running each of the blocks and making changes to try out variations of the main code if you wish. This would be the best route if you have some familiarity with R

2. If you are not familiar with R but want to see what you can do with a programmatic approach to this type of analysis then you can open a compiled version of the analysis which can be found here and read through the analysis to see what is possible.