

IBIP 2021

PROLINE TUTORIAL: QUALITY ASSESMENT OF QUANTITATIVE PROTEOMICS DATA

I/ PRESENTATION OF THE DATASET

In this tutorial, we will explore the quantitative analysis of a “ground truth” dataset generated from a yeast lysate in which different levels of the equimolar UPS1 mix of 48 human proteins were spiked.

1.1 SAMPLE PROCESSING PROTOCOL

Samples (1µg of yeast cell lysate + different spiked level of UPS1) were analyzed in quadruplicate by nanoLC-MS/MS using a nanoRS UHPLC system coupled to a QExactive Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) using a top20 data-dependent acquisition method.

See the references below for additional information.

1.2 DATA PROCESSING PROTOCOL

MS/MS data were searched in a yeast database from UniprotKB (S_cerevisiae_20160915.fasta, 6729 sequences), a compiled database containing the UPS1 human sequences (48 sequences) and a common contaminants database. Peaklists were generated using Proline and were submitted to database search with Mascot (version 2.5.1). ESI-FTMS-HCD was chosen as the instrument, trypsin/P as the enzyme and 1 missed cleavage was allowed. Precursor and fragment mass error tolerances were set at 5 ppm and 70mmu, respectively. Peptide variable modifications allowed during the search were: acetyl (Protein N-ter), oxidation (M), whereas carbamidomethyl (C) was set as fixed modification.

1.3 DATASET DESCRIPTION

The used dataset contains the MS analysis of two samples of the yeast cell lysate spiked respectively with 50fmol and 5fmol of UPS1. Here is a table summarizing the names of injected samples, and corresponding output files:

Sample Name	Raw File	Mascot .dat file
50fmol DDA inj1	QEKAC160601_20_0910	F085233.dat
50fmol DDA inj2	QEKAC160601_50_0910	F085231.dat
50fmol DDA inj3	QEKAC160601_81_0910	F085235.dat
50fmol DDA inj4	QEKAC160601_131_0910	F085240.dat
5fmol DDA inj1	QEKAC160601_14_0910	F085230.dat
5fmol DDA inj2	QEKAC160601_44_0910	F085229.dat
5fmol DDA inj3	QEKAC160601_75_0910	F085234.dat
5fmol DDA inj4	QEKAC160601_125_0910	F085238.dat

References:

- <https://www.ebi.ac.uk/pride/archive/projects/PXD009815>
- Bouyssié D, Hesse AM, Mouton-Barbosa E, Rompais M, Macron C, Carapito C, de Peredo AG, Couté Y, Dupierris V, Burel A, Menetrey JP, Kalaitzakis A, Poisat J, Romdhani A, Burlet-Schiltz O, Cianférani S, Garin J, Bruley C. Proline: an efficient and user-friendly software suite for large-scale proteomics. Bioinformatics. 2020; <https://pubmed.ncbi.nlm.nih.gov/32096818>


II/ GETTING STARTED WITH THE PROLINE WEB INTERFACE


Proline was executed to import, validate then quantify the MS and MS/MS output files. During this tutorial, we will use its web interface to browse and inspect the obtained results.

2.1 LOGIN

On the link <http://proteomique.ipbs.fr:3000> to open the interface. Please use the Google Chrome (or Chromium), since some visualization bugs were recently reported with other web browsers.

The following login form should appear:





Name

Password

OK

User: IBIPx (e.g. 'IBIP1')

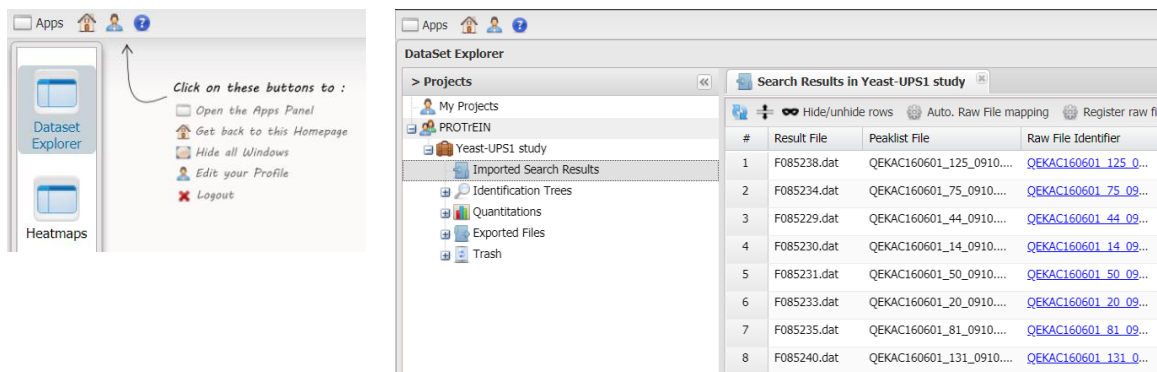
Password: IBIP

2.2 BROWSING THE DATASET

Action

To visualize the dataset, click on the “Dataset Explorer” icon (under the Apps menu), then in the new window expand the “IBIP” node and the “Yeast-UPS1 study” one. Double click on the “Imported Search Results” item to display the list of files that were processed.

You should see something similar to the following screenshots:



III/ BRIEF INSPECTION OF THE IDENTIFICATION RESULTS

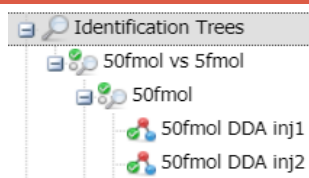
3.1 GLOBAL RESULTS

The menu on the left can be used to browse both the identification and quantification results.

Action

Expand all nodes of the “Identification Trees” to reveal the different levels of results aggregation.

Then double click on the two intermediate levels, named “50fmol” and “5fmol” (they correspond to the compared samples).



Files belonging to the 50fmol sample:

Aggregated Identifications								
#	Validated	Identification Na...	Result File Nam	Raw File Name	Sample Name	#Val. Protein sets	#Val. Peptides	#Queries
1	✓	50fmol DDA inj1	F085233.dat	QEKAC160601_20_0910	50fmol DDA inj1	2119	12343	35737
2	✓	50fmol DDA inj2	F085231.dat	QEKAC160601_50_0910	50fmol DDA inj2	2076	12044	34968
3	✓	50fmol DDA inj3	F085235.dat	QEKAC160601_81_0910	50fmol DDA inj3	2166	12130	35122
4	✓	50fmol DDA inj4	F085240.dat	QEKAC160601_131_0910	50fmol DDA inj4	2106	11848	32953

Files belonging to the 5fmol sample:

Aggregated Identifications								
#	Validated	Identification Na...	Result File Name	Raw File Name	Sample Name	#Val. Protein sets	#Val. Peptides	#Queries
1	✓	5fmol DDA inj1	F085230.dat	QEKAC160601_14_0910	5fmol DDA inj1	2190	12441	34479
2	✓	5fmol DDA inj2	F085229.dat	QEKAC160601_44_0910	5fmol DDA inj2	2194	12397	34200
3	✓	5fmol DDA inj3	F085234.dat	QEKAC160601_75_0910	5fmol DDA inj3	2242	12778	34747
4	✓	5fmol DDA inj4	F085238.dat	QEKAC160601_125_0910	5fmol DDA inj4	2279	12609	34031

These tables give us a quick overview of the identification performance, and of the proteome coverage of our analyzed samples. It can also be used to quickly verify if got similar results in the different replicates.

Questions

Do you see differences in terms of the number of validated peptides and proteins (columns named #Val. Protein Sets, #Val. Peptides)? Is it logical, given the experimental design?

3.2 IDENTIFIED HUMAN PROTEINS

Now that we checked the global performance, we will inspect the results specifically for our spiked UPS human proteins. Since we know precisely that we added 48 proteins, we have a reference that we can use to perform a more specific verification.

Let's first compare the number of identified proteins in our two conditions:

Action

In each aggregated result ("50fmol" and "5fmol") click on the "Proteins" tab and expand the hidden menu on the left using the double arrow button.

This will display a list of available filters. Click on "Text data", select "Accession" in the list then type "HUMAN" in the "Value" column. Click on "Apply" to filter the proteins table.

Questions

How many human proteins were identified in each condition? Is it what you would expect?

Now let's doing the same at the peptide level:

Action

Sort the filtered proteins tables by descending number of validated peptides (click on the header of the column "#Val. Peptides").

Questions

On an average, can you give a rough ratio of the number of identified peptides for the top 5 proteins? Is this correlating with the proportions of spiked UPS1 proteins?

Out of curiosity, let's inspect the detailed identification results of a single protein:

Action

In the 50fmol and 5fmol proteins tables, click on the magnifier icon (🔍) in the row corresponding to ALBU_HUMAN_UPS.

A new web browser tab should display detailed information about this protein as follow (for 50fmol):

584
aa

[sp|P02768|ALBU_HUMAN_UPS Serum albumin \(Chain 26-609\) - Homo sapiens \(Human\)](#)

Search in sequence.. (Rege)

AA per line Select

```

1  AHKSEVAHRF KDLGEENFKA LVLIATAQYL QQCPFEDHVK LVNEVTEFAK TCVADESAEN CDKSLHTLFG DKLCTVATLR ETYGMADCC AKQEPERNEC FLQHKDDNPN LPRLVVRPEVD
121 VMCTAFHDNE ETFLKKYLYE IARRHPYFYA PELLFFAKRY KAAFTCECQA ADKAAACLLPK LDELDEGKA SSAKQRLKCA SLQKFGERA F KAWAVARLSQ RFPKAEFAEV SKLVTDLTKV
241 HTECCGDDL ECADDRADLA KYICENQDSI SSKLKECEK PLEKSHCIA EVENDEMPAD LPSLAADFVE SKDVCKNYAE AKDVLGMFL YEYARRHPDY SVVLLRLAK TYETTLKCC
361 AAADPHCEYA KVFDEFKPLV EEPQNLIKQN CELFEQLGEY KFQNALLVRY TKKVPQVSTP TLVEVSRNLG KVGSKCKHP EAKRMPCAED YLSVVLNQLC VLHEKTPVSD RVTCKCTESL
481 VNRPRPCFSAL EVDETYVPKE FNAETTFEHA DICTLSEKER QIKKQTALVE LVKHKPKATK EQLKAVMDDF AAFVEKCKKA DDKETCFAE GKGLVAASQA ALGL
                    
```

Questions

Some regions of the protein sequence are colored and some others are not. What these regions correspond to? Can you find in the previous discussed “proteins” table a column summarizing this information?
Why do you identify different peptides between the two conditions (5fmol and 50fmol)?

IV/ QUALITY CONTROL OF THE QUANTITATIVE RESULTS

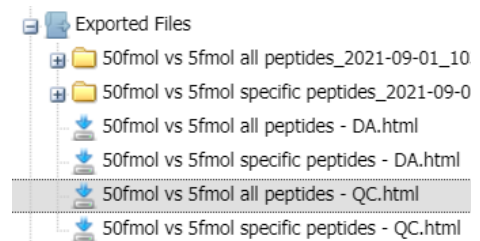
We will now inspect the label-free results, and start by doing some quality controls. LC-MS/MS measurements are not always 100% reproducible, and it is thus of major importance to check the consistency of performed acquisitions. Several parameters can be monitored, such as for instance the stability of intensity measurements and the retention time shifts (captured by the chromatographic alignment procedure).

4.1 QC OF INTENSITY MEASUREMENTS

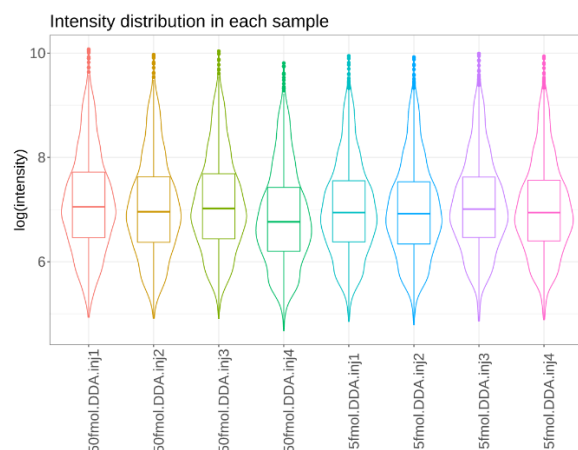
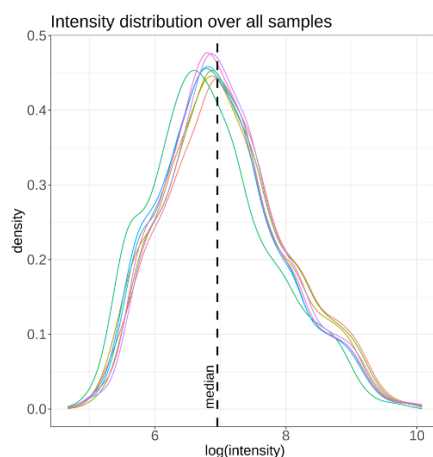
We will start to verify that we got reproducible intensity measurements across our eight acquisition files. This will tell us if our mass spectrometer was working with the same sensitivity during the whole study.

Action

In the left tree panel, expand the “Exported files” node and double click on the item named “50fmol vs 5fmol all peptides – QC.html”.



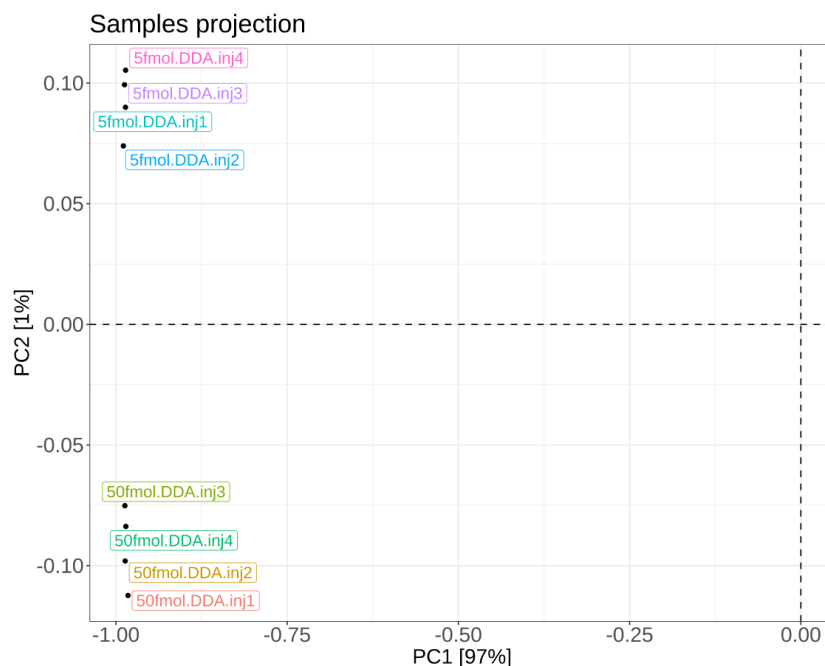
The two first figures summarize the intensity distributions of the detected proteins in the different raw files. These data were obtained before any kind of data normalization.



Questions

According to these results, can you spot differences in terms of total intensity measurements? Is this correlating the number of peptides that were identified in individual files (see results discussed in the previous section 3.1).

Additional information: the dimension reduction plots (PCA, Principal Component Analysis) are useful figures to highlight outliers, and inspect similarities between raw files and samples. Note: the displayed PCA plots were obtained without normalization and/or transformation of the data.



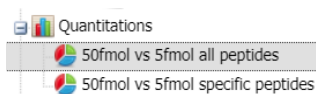
The PCA applied to the projection of the different samples of our dataset highlights two groups of data: one group on the top left composed of 5fmol injection replicates, and a second group on the bottom left composed of 50fmol injection replicates. If one of the replicates was an outlier in terms of intensity measurements, it would be separated from a given group. Fortunately, it is not the case in our dataset.

4.2 QC OF CHROMATOGRAPHIC ALIGNMENTS

While the previous QC gave us information about the mass spectrometer performance, we have no clue about the reproducibility of our chromatographic system coupled to the instrument. It is a very important information in the context of label-free studies, since the software has to map the intensities of peptides in different runs using the retention time information. We will now make this verification.

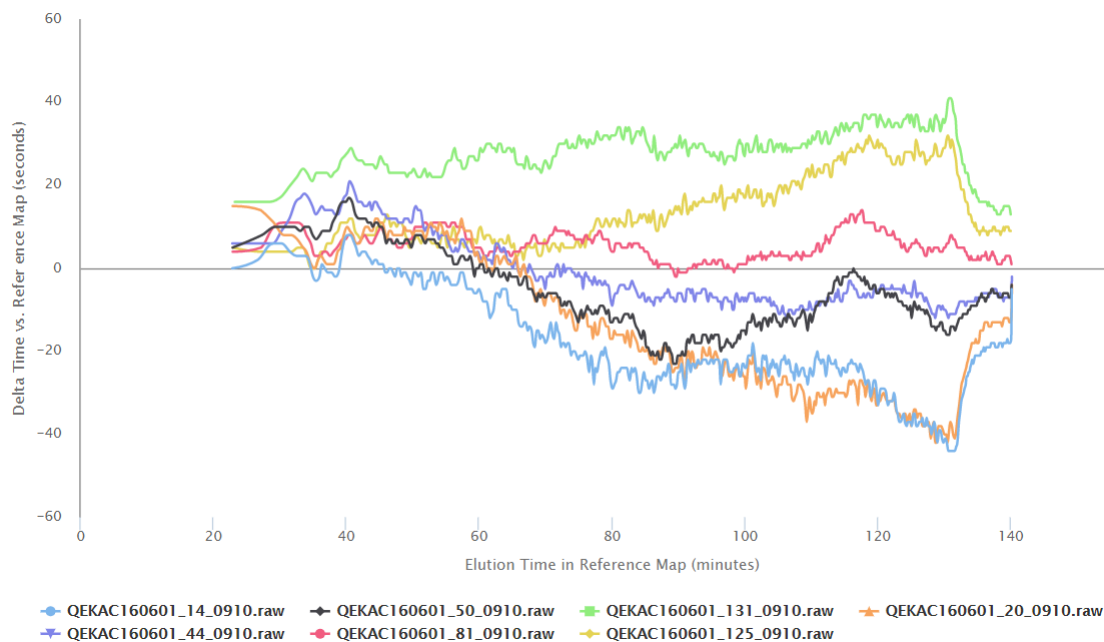
Action

In the left tree panel, expand the “Quantitations” node and double click on the item named “50fmol vs 5fmol all peptides”. Then click on the tab “LC-MS Maps”.



The displayed panel shows the pairwise alignment curves computed between the reference (QEKAC160601_75_0910.raw) and all other runs as follow:

LC-MS Map Alignments to QEKAC160601_75_0910.raw



These curves can be used to assess the chromatographic reproducibility of the compared runs, and to verify that the software was able to perform a consistent correction.

The x-axis corresponds to the elution time scale of the reference run (QEKAC160601_75_0910.raw). The y-axis shows the moving average difference in seconds between the elution times of observed peptides in a given run and the elution times of the same peptides in the reference.

Questions

Try to sort (by hand) the raw file names according to the order of the curves from the bottom to the top (don't forget to include the reference). Do you see a trend? What is your assumption?
Note: each raw file has a specific number (14, 20, 44, etc...) corresponding to the acquisition order.
Do you see any outlier?

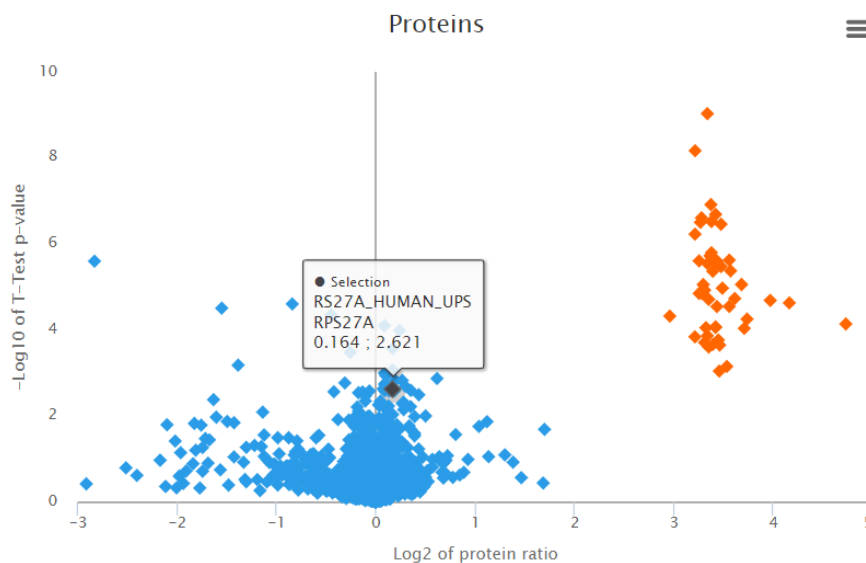
Additional information: the alignment curves are just an indicator of the reproducibility of the LC system but it doesn't tell us if software was able to correct the retention time (RT) shifts. To check the residual errors you can have a look at the RT standard deviations histograms before and after RT correction (i.e. using the alignment curves above) by visiting the "LC correction" tab.

V/ IN-DEPTH EXPLORATION OF THE QUANTITATIVE RESULTS

5.1 VOLCANO PLOTS

Action	In the quantitation named "50fmol vs 5fmol all peptides", click on the tab "Quantitation Stats". This will display the "Proteins Volcano Plot". In the color palette section at the bottom, click a given color, then type "HUMAN" in the Accession column.	Log2 Fold Change: <input type="text" value="Ratio"/> <input type="button" value="v"/> P-Value: <input type="text" value="P-Value"/> <input type="button" value="v"/>
		Accession HUMAN

You should see a Volcano Plot similar to this one:



This scatter plot shows two complementary information for each single quantified protein:

- the log₂ ratio on the x-axis, i.e. the value obtained by dividing the intensity averages in the two compared conditions (50fmol / 5fmol)
- the t-test p-value on the y-axis, transformed using the formula “-10 x log(p-value)” to ease the visualization

This figure is very useful to assess the significant variations of the proteins. Here we clearly see two groups of proteins: the blue dots correspond to the Yeast proteins and the orange dots to the Human ones (i.e. UPS1).

Questions	Are these results consistent with our experimental design, and why?
	What is the log ₂ ratio of the UPS protein having the best p-value? Try to compute the inverse log transformation to estimate the real fold change. Is it also consistent with our experimental design?

Some Yeast proteins have a very good p-value, i.e. with $-\log_{10}(p\text{-value})$ greater than 2 (equivalent to a p-value of 0.01). These hits can be considered as false positives.

In the middle of the Yeast proteins, we can see a dot corresponding to a Human hit, which is thus a false negative. During the next steps, we will try to understand why this UPS1 protein was not well quantified.

Action

Click on the dot corresponding to the Human protein named "RS27A_HUMAN_UPS".

5.2 QUANTIFIED PEPTIDES: SPECIFIC AND SHARED PEPTIDES

You should see a new web page containing the following table:

Quantified peptides											
<input checked="" type="checkbox"/> Select all in Dataset <input type="checkbox"/> Deselect all in Dataset <input checked="" type="checkbox"/> Select all in Protein set <input type="checkbox"/> Deselect all in Protein set											
#	Dataset selection	Protein selection	Sequence	PTMs	Missed clv.	Calculated mass	#PSMs	#Ions	#Protein sets	Best score	Max. abundance
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	TLSDYNIQK	0		1080.54512	1	1	2	61.5	3.37e+8
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	ESTLHLVLR	0		1066.61348	1	2	2	71.4	1.85e+8
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	TITLEVEPSDTIENVKAK	1		1986.05208	1	1	1	69.1	4.29e+6
4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	TITLEVEPSDTIENVK	0		1786.92001	1	1	1	74.4	2.80e+6
5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	IQDKEGIPPDQQR	1		1522.77396	1	2	2	95.4	2.27e+6
6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	TLSDYNIQKESTLHLVLR	1		2129.14804	1	2	2	43.2	2.25e+6

This table shows the list of quantified peptides belonging to the current protein. The column "#Protein sets" gives the number of protein sets (or protein groups) the peptide belongs to. As you can see, some peptides are found in a single protein set while some other ones belong to two different protein sets.

Action

Go back to the "Proteins" tab of the current quantitation and open the filters panel on the left side using the double arrow button.

Filters

Numeric data

Text data

Boolean data

Button data

Accession

Param

Accession

Must/not

must

Condition

contain

Value

RS27A

Protein Sets

#	Accession
1	RS27A_HUMAN_UPS
2	RS27A_YEAST

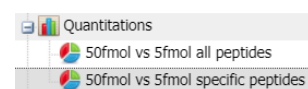
Click on "Text data", select "Accession" in the list then type "RS27A" in the "Value" column. Click on "Apply" to filter the proteins table. Double click on the entry named RS27A_YEAST.

Questions

How many peptides are not specific to RS27A_YEAST? Is it the same number for RS27A_HUMAN_UPS?
Do you think these peptides can bias the quantification of these two proteins?

Action

Go back to left tree panel and under the "Quantifications" node double click on the item named "50fmol vs 5fmol specific peptides".



Open the "Proteins" tab and filter again the "Accession" value with the "RS27A" text. Double click on the RS27A_HUMAN_UPS and RS27A_YEAST entries.

Questions

Have a look at the peptide selection checkboxes. What is the difference?
In the "protein set profiles" table at the top of the page, note the value of the column "1/2" (which represents the quantitative ratio between the two compared biological groups) for each protein set (RS27A_HUMAN_UPS and RS27A_YEAST), and in each quantitation (all peptides vs specific peptides).
Which results are better considering the experimental design?
Verify this by looking again the Volcano plot of the "specific peptides" quantitation (filter the color palette with the HUMAN keyword).

5.3 QUANTIFICATION ERRORS AND CHROMATOGRAMS VISUALIZATION

Action

Go back to the “Proteins Volcano Plot” of the “specific peptides” quantitation. On the left part of the plot, there are two extreme values: “YL363_YEAST” and “GCY1_YEAST”. Click on each on them to open the quantification details. Then double click on wrongly quantified peptides (VGHLELLR for YL363_YEAST and AVGVSNFSINNLIK for GCY1_YEAST).


AVGVSNFSINNLIK ions quantification results

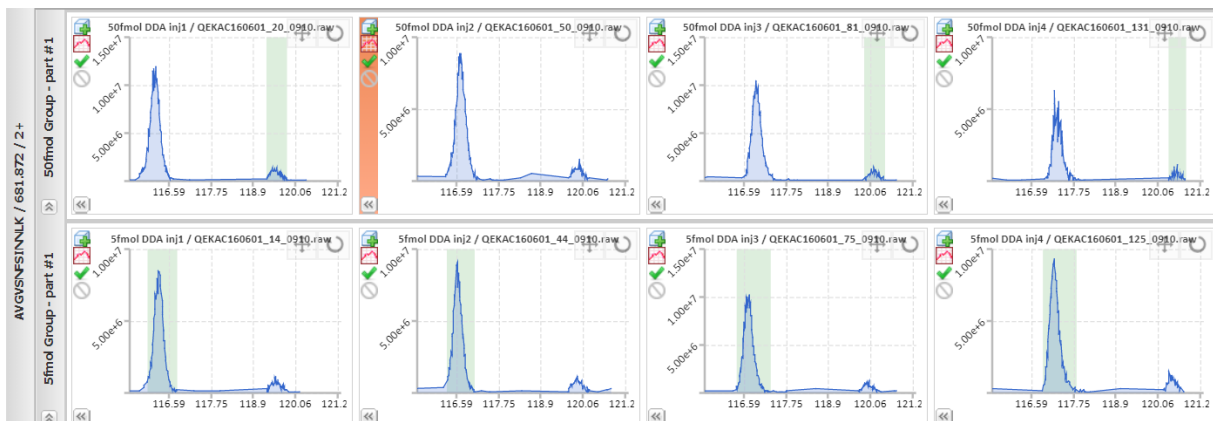
Z	Selection	#	Condition	Raw File	Sample Name	m/z	Score	#PSMs	Ret. time	Adjusted RT	Duration	Raw Ab.	Abundance
2+	<input checked="" type="checkbox"/> In dataset <input checked="" type="checkbox"/> In protein	1	50fmol Gro...	QEKAC160601_20_0910	50fmol DDA inj1	681.87353	37.7	1	119.42	119.96	0.57	1.37e+6	1.37e+6
		2	50fmol Gro...	QEKAC160601_50_0910	50fmol DDA inj2								
		3	50fmol Gro...	QEKAC160601_81_0910	50fmol DDA inj3	681.87338	37.7	1	120.14	119.99	0.57	1.47e+6	1.47e+6
		4	50fmol Gro...	QEKAC160601_131_0910	50fmol DDA inj4	681.87308	18.5	1	120.58	119.99	0.49	1.13e+6	1.13e+6
		5	5fmol Group	QEKAC160601_14_0910	5fmol DDA inj1	681.87248	36.1	1	116.31	116.7	0.85	8.50e+6	8.50e+6
		6	5fmol Group	QEKAC160601_44_0910	5fmol DDA inj2	681.87254	57.1	1	116.62	116.7	0.86	9.07e+6	9.07e+6
		7	5fmol Group	QEKAC160601_75_0910	5fmol DDA inj3	681.87252	58.9	1	116.74	116.74	0.93	1.02e+7	1.02e+7
		8	5fmol Group	QEKAC160601_125_0910	5fmol DDA inj4	681.8727	53.2	1	117.22	116.72	0.92	9.31e+6	9.31e+6

Questions

For each peptide, verify in the results table if they were identified (scored) in each raw file.

Action

Finally click on the “blue eye” button  to display the interactive viewer.



Questions

What are these plots? How are they constructed from the raw data? If you click on the numbers ranging from 1 to 4 on the bottom left, below the slider, you should see additional series on the plots. What are they representing? For each peptide, can you infer what kind of error the software has made? First, try to update the XICs by expanding the RT range, changing the RT scale and eventually the kind of Zoom. Then, compare the RT of displayed signals to the RT obtained in the results table.

5.4 MISSING VALUES

You may have noticed that the “match between runs” procedure is not able to recover signals in given raw files, where the MS/MS identification was missing.

Depending on the method used for the protein intensity summarization this may have an impact on the quality of our data.

Action

Download then open the Excel file named “50fmol vs 5fmol specific peptides - ions exported.xlsx” in a spreadsheet editor. Open the tab named “Quantified peptide ions” and copy paste the lines containing “UB2E1_HUMAN_UPS” in a new document. For each column prefixed with “abundance_”, sum up the abundances to obtain the abundance of the protein in the 8 files. Calculate for each condition the coefficient of variation (CV).

Questions

Compare the two computed CVs to those displayed in the Proline Web interface. How is it changing?
The displayed results in the web interface were summarized with a method different from the sum of ion intensities. Which method would you prefer in this context and why?

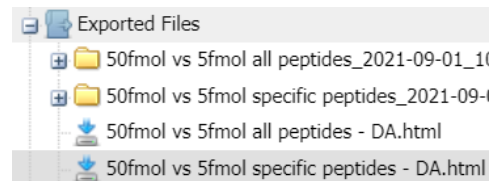
5.5 FINAL EVALUATION

The Volcano plot is a nice tool to highlight at a glance the most important hits. However, it only shows the t-test p-value and the fold change of each quantified protein. Thus, it does not take into account the cumulated errors resulting of individual statistical tests. It is nevertheless possible to adjust the obtained p-values and thus try to control the final level of false positives. This procedure is called the Multiple Testing Correction. We won't cover the theoretical aspects here but will quickly see the impact of such a treatment.

Action

Go back to the tree panel and in below the “Exported Files” node, double click on the item named “50fmol vs 5fmol specific peptides – DA.html”.

Go to the last table and use the “Search” text field to look for HUMAN and YEAST proteins.



Accession	Label	50fmol_1_50fmol Group_mean	50fmol_2_50fmol Group_mean	50fmol_3_50fmol Group_mean	50fmol_4_50fmol Group_mean	50fmol_5_50fmol Group_mean	50fmol_6_50fmol Group_mean
ALBU_HUMAN_UPS	ALBU_HUMAN_UPS	1586037513.44912	1557475186.78903	1561714057.61202	1671396685.38797	161373990.283867	
ANT3_HUMAN_UPS	ANT3_HUMAN_UPS	695829123.512586	683276748.9159	698741728.282129	694712051.711358	73259020.3603221	
ANXA5_HUMAN_UPS	ANXA5_HUMAN_UPS	518304904.010017	513986009.600833	497503567.737778	486078318.879336	49658008.6722653	
ATP9_YEAST	OLI1	1596368.56741494	1521857.26449387	1595945.15085393	1292301.09279253	2655692.21554218	

Questions

The p-values were adjusted using the “Benjamini & Hochberg” procedure (R package multtest). Proteins were considered significant if their fold change was greater than 1.5 and their q-value (adjusted p-value) lower than 0.05 (5%).
How many HUMAN and YEAST proteins are considered significant in this dataset?
Try to estimate the False Discovery Rate (number of false hits divided by the total number of hits). Is it close to our theoretical value (5% q-value cutoff)?