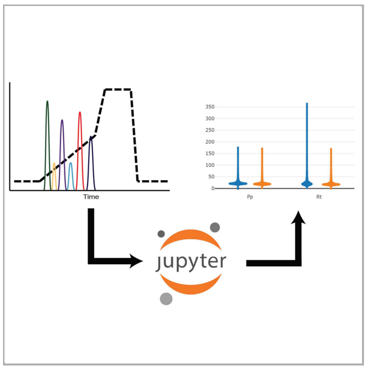


Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files V.2

Yan Chen[1], Christopher J Petzold[1]

[1]Lawrence Berkeley National Laboratory

LBNL omics

Agile BioFoundry

1 more workspace ↓

Christopher J Petzold
Lawrence Berkeley National Laboratory

DISCLAIMER

This protocol is for research purposes only.

ABSTRACT

This protocol details the analysis of label-free quantification (LFQ) data from data independent acquisition (DIA) discovery (shotgun) proteomic experiments and generates a series of outputs.

VERSION 2

MAR 01, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v2

Protocol Citation: Yan Chen, Christopher J Petzold 2023. Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files. **protocols.io** https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v2Version created by Christopher J Petzold

Protocol status: Working
We use this protocol and it's working

Created: Oct 11, 2022

Last Modified: Mar 01, 2023

PROTOCOL integer ID: 71194

Keywords: Data analysis, proteomics, DIA, volcano plots, Welch's t-Test

GUIDELINES

- Abundance values correspond to summed peptide peak area in arbitrary units
- SVG files are provided for easy editing with Adobe Illustrator or similar programs
- .plotly files can be can be visualized by using Plotly or a Colab jupyter notebook.

Helpful references and links:

DIA-NN reference:

- Demichev, V., Messner, C.B., Vernardis, S.I. et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods 17, 41–44 (2020). https://doi.org/10.1038/s41592-019-0638-x

Top3 absolute protein quantification references:

- Ludwig C, Claassen M, Schmidt A, Aebersold R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. Mol Cell Proteomics. 2012 Mar;11(3):M111.013987. doi: 10.1074/mcp.M111.013987
- Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics. 2006 Jan;5(1):144-56. doi: 10.1074/mcp.M500230-MCP200
- Ahrné E, Molzahn L, Glatter T, Schmidt A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics. 2013 Sep;13(17):2567-78. doi: 10.1002/pmic.201300135
- Grossmann J, Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S, Rutishauser D, Schlapbach R. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J Proteomics. 2010 Aug 5;73(9):1740-6. doi: 10.1016/j.jprot.2010.05.011.

T-test references:

- Benjamini, Y. and Hochberg, Y. (1995), Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), 57: 289-300.https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind_from_stats.html

- https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html

MATERIALS

Required:

- DIANN peptide report file (Experiment report.pr_matrix.csv)

Optional:

- A list of selected proteins for bar and strip chart visualization
- Sample timepoints in the sample name (e.g., CJP1234_24hr-R1) for line chart visualization
- A list of two-sample comparisons of different samples (Sample A vs. Sample B; Sample B vs. Sample C, etc.)

SAFETY WARNINGS

- Missing proteins - Some proteins may not meet the Top3 criteria (at least three peptides detected across all samples), so they won't be quantified and shown on the bar, line, and strip charts. If you do not see the protein you are looking for, search for them in the "Full_list_proteins_XXXXX-xxxxx.csv" file. If they are not listed in the Full_list_proteins file they were not detected in any of the data acquisition runs.

BEFORE START INSTRUCTIONS

INPUTS:

Required:

- DIANN peptide report file (Experiment_report.pr_matrix.csv)

Optional:

- **selected_proteins.csv** - A list of selected proteins for bar chart visualization with Protein.Group identifiers (e.g., P0C054, P0C058)
- **selected-ttest-vol-samples.csv** - A list of two-sample comparisons of different samples (Sample A vs. Sample B; Sample B vs. Sample C, etc.)

OUTPUTS:

Note

- Abundance values correspond to summed peptide peak area in arbitrary units
- Top3 absolute protein quantification is based on the “best flyer” hypothesis, which assumes that the specific MS signal intensity of the most intense tryptic peptides per protein is approximately constant throughout a whole proteome (references in Guidelines Section)
- Top3 protein amount consists of the averaged peptide intensity (counts) for the top 3 peptides of each protein presented as a percentage of the total amount of all detected proteins
- SVG files are provided for easy editing with Adobe Illustrator or similar programs
- You can visualize the .plotly files by using Plotly or a Colab jupyter notebook. This provides an interactive view that you can see data labels, zoom, and save parts of the plot as separate .png files.

Top level folder:

- **DIA-NN peptide report output file (CSV)** - a full list of precursor ion quantitative values
- **Protein data table (CSV)** - a full list of protein quantitative values from the summed peptide abundances
- **Summary Protein data table (CSV)** - a full list of protein quantitative values averaged over the sample replicates
- User provided selected_proteins and selected-ttest-vol-sample (CSV) files

If applicable:

- **Summary of Bar charts** (PDF)
- **Summary of Strip charts** (PDF)
- **Summary of protein abundance histograms** (PDF)
- **Summary of Line charts** (PDF; if timepoints are included in the sample names)

EDD_files folder:

- **Protein data table in EDD upload format (CSV)** - a full list of protein quantitative values from the summed peptide abundances in EDD data upload format with Time (e.g., 24h) and Units (e.g., counts)
- **Top3 quantitative protein data table in EDD upload format (CSV)** - a full list of protein quantitative values from the Top3 quantitative method in EDD data upload format with Time (e.g., 24h) and Units (e.g., % protein abundance)

QC_files folder:

- **QC Protein counts bar chart (png)** - a bar chart showing the number of

proteins identified and quantified in each individual sample replicate, the cumulative number of proteins found in all the samples, and the number of proteins that meet the criteria for the Top3 quantitative method
- **QC Peptide counts bar chart (png)** - a bar chart showing the number of peptides identified and quantified in each individual sample replicate and the cumulative number of peptides found in all the samples
- **QC Box plot (png)** - of the relative peptide abundance (log2 counts) data for each sample replicate
- **QC peptide CVs violin plot (png)** - a violin plot showing the distribution of peptide CVs for each sample

Top3_quant_files folder:

- **Top3 Summary Protein data table (CSV)** - a full list of protein absolute abundance values averaged over the sample replicates
- **Top3_Full_list_peptides_used_for_quant (CSV)** - a full list of peptides and corresponding intensity values used for the Top3 absolute protein calculations.
- **Top3 jitter plot (PNG, .plotly)** - a plot detailing the distribution of proteins across the percentiles of abundance. Groups of proteins from the selected_proteins.csv file are highlighted.

If applicable:

Bar_Charts folder:

- **Summary data table of a selected list of proteins (XLSX)** - a list of selected protein quantitative values averaged over the sample replicates
- **Individual bar charts of selected protein groups in .png, .svg, and .plotly formats**

Strip_Charts folder:

- **Individual strip charts of selected protein groups in .png, .svg, and .plotly formats**

t-test_files folder:

- **Excel file with the Welch's t-test results for each comparison**
- **Volcano plots visualizing the Welch's t-test p-value significance and log(2) normalized Fold Change (FC) between the two samples (.png, .svg, and .plotly formats)**
- **Volcano plots visualizing the t-test adjusted p-value (Benjamini-Hochberg) significance and log(2) normalized Fold Change (FC) between the two samples (.png, .svg, and .plotly formats)**

Line_Charts folder (if timepoints are included in the sample names):

- **Individual line charts of selected protein groups in .png, .svg, and .plotly formats**

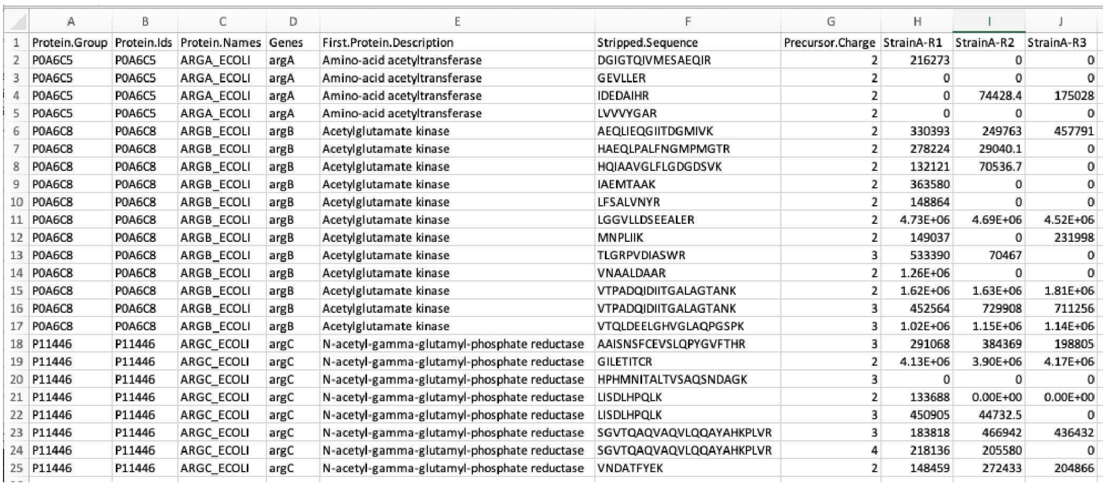
Data processing: Relative Counts

1 We start with a DIA data acquisition peptide search output file the DIANN search (DIA; link to DIA-NN paper) and we trim out unused columns in the reports to simplify the analysis.

DIA-NN report restricted to:

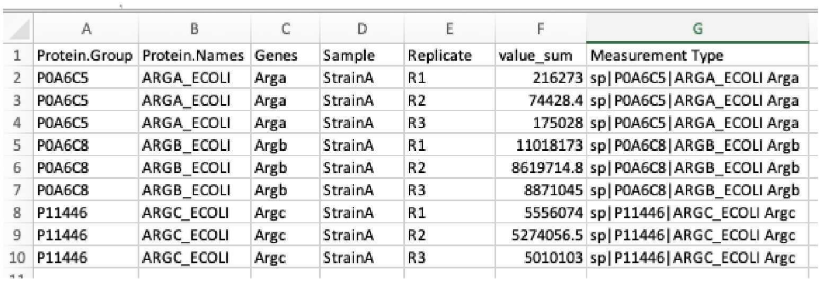
- Protein Group
- Protein Name
- Genes
- Protein Description
- Peptide Sequence
- Sample
- Replicate
- Intensity value (counts, peptide peak area in arbitrary units)

2 All of the peptide intensity values (counts) are summed to the protein intensity (counts). The resulting data table is exported as:
Full_list_proteins_XXXXXXXXX-xxxxxxx.csv



	A	B	C	D	E	F	G	H	I	J
1	Protein.Group	Protein.Ids	Protein.Names	Genes	First.Protein.Description	Stripped.Sequence	Precursor.Charge	StrainA-R1	StrainA-R2	StrainA-R3
2	P0A6C5	P0A6C5	ARGA_ECOLI	argA	Amino-acid acetyltransferase	DGIGTQIVMESAEQIR	2	216273	0	0
3	P0A6C5	P0A6C5	ARGA_ECOLI	argA	Amino-acid acetyltransferase	GEVLLER	2	0	0	0
4	P0A6C5	P0A6C5	ARGA_ECOLI	argA	Amino-acid acetyltransferase	IDEDAIHR	2	0	74428.4	175028
5	P0A6C5	P0A6C5	ARGA_ECOLI	argA	Amino-acid acetyltransferase	LVVVYGAR	2	0	0	0
6	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	AEQLIEQGIITDGMIVK	2	330393	249763	457791
7	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	HAEQLPALFNGMPMGTR	2	278224	29040.1	0
8	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	HQIAAVGLFLGDGDSVK	2	132121	70536.7	0
9	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	IAEMTAAK	2	363580	0	0
10	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	LFSALVNYR	2	148864	0	0
11	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	LGGVLLDSEEALER	2	4.73E+06	4.69E+06	4.52E+06
12	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	MNPLIIK	2	149037	0	231998
13	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	TLGRPVDIASWR	3	533390	70467	0
14	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	VNAALDAAR	2	1.26E+06	0	0
15	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	VTPADQIDIITGALAGTANK	2	1.62E+06	1.63E+06	1.81E+06
16	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	VTPADQIDIITGALAGTANK	3	452564	729908	711256
17	P0A6C8	P0A6C8	ARGB_ECOLI	argB	Acetylglutamate kinase	VTQLDEELGHVGLAQPGSPK	3	1.02E+06	1.15E+06	1.14E+06
18	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	AAISNSFCEVSLQPYGVFTHR	3	291068	384369	198805
19	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	GILETITCR	2	4.13E+06	3.90E+06	4.17E+06
20	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	HPHMNITALTVSAQSNDAGK	3	0	0	0
21	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	LISDLHPQLK	2	133688	0.00E+00	0.00E+00
22	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	LISDLHPQLK	3	450905	44732.5	0
23	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	SGVTQAQVAQVLQQAYAHKPLVR	3	183818	466942	436432
24	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	SGVTQAQVAQVLQQAYAHKPLVR	4	218136	205580	0
25	P11446	P11446	ARGC_ECOLI	argC	N-acetyl-gamma-glutamyl-phosphate reductase	VNDATFYEK	2	148459	272433	204866

↓



	A	B	C	D	E	F	G
1	Protein.Group	Protein.Names	Genes	Sample	Replicate	value_sum	Measurement Type
2	P0A6C5	ARGA_ECOLI	Arga	StrainA	R1	216273	sp|P0A6C5|ARGA_ECOLI Arga
3	P0A6C5	ARGA_ECOLI	Arga	StrainA	R2	74428.4	sp|P0A6C5|ARGA_ECOLI Arga
4	P0A6C5	ARGA_ECOLI	Arga	StrainA	R3	175028	sp|P0A6C5|ARGA_ECOLI Arga
5	P0A6C8	ARGB_ECOLI	Argb	StrainA	R1	11018173	sp|P0A6C8|ARGB_ECOLI Argb
6	P0A6C8	ARGB_ECOLI	Argb	StrainA	R2	8619714.8	sp|P0A6C8|ARGB_ECOLI Argb
7	P0A6C8	ARGB_ECOLI	Argb	StrainA	R3	8871045	sp|P0A6C8|ARGB_ECOLI Argb
8	P11446	ARGC_ECOLI	Argc	StrainA	R1	5556074	sp|P11446|ARGC_ECOLI Argc
9	P11446	ARGC_ECOLI	Argc	StrainA	R2	5274056.5	sp|P11446|ARGC_ECOLI Argc
10	P11446	ARGC_ECOLI	Argc	StrainA	R3	5010103	sp|P11446|ARGC_ECOLI Argc

Output file: Full_list_proteins_XXXXXXXXX-xxxxxxx.csv

Note

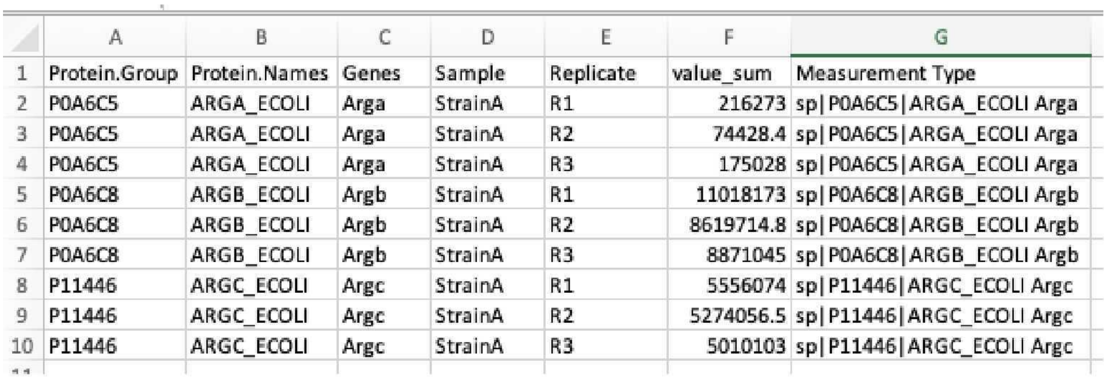
Protein of interests in "selected_proteins" file that are not shown in reports at "Bar_Charts", "Strip_Charts", "Line_charts", and "Top3_quant_files" may be identified and quantified in these "Full_list_proteins_XXXXX-xxxxx.csv" files.

Note

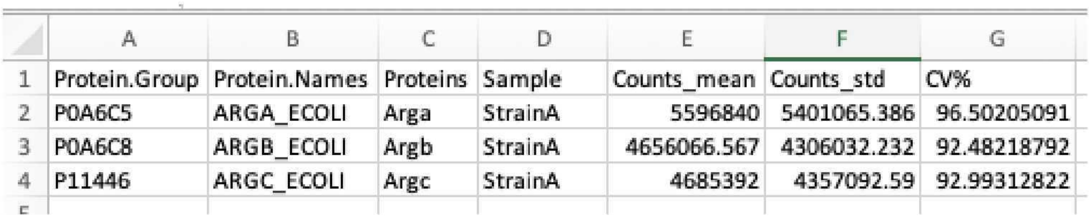
A file for Experiment Data Depot (EDD) data import is also generated with the name: **Full_list_proteins_EDDformat_XXXXXXXXX-xxxxxxx.csv**

Directions for the EDD import process can be found **here**.

3 Then the protein intensities (counts) of the sample replicates are averaged (mean), the standard deviation (SD), percent coefficient of variation (CV%), and Z-scores (across all samples) are calculated. The resulting data table is exported as:
Full_list_proteins_summary_XXXXXXXXX-xxxxxxx.csv



Protein.Group	Protein.Names	Genes	Sample	Replicate	value_sum	Measurement Type
P0A6C5	ARGA_ECOLI	Arga	StrainA	R1	216273	sp|P0A6C5|ARGA_ECOLI Arga
P0A6C5	ARGA_ECOLI	Arga	StrainA	R2	74428.4	sp|P0A6C5|ARGA_ECOLI Arga
P0A6C5	ARGA_ECOLI	Arga	StrainA	R3	175028	sp|P0A6C5|ARGA_ECOLI Arga
P0A6C8	ARGB_ECOLI	Argb	StrainA	R1	11018173	sp|P0A6C8|ARGB_ECOLI Argb
P0A6C8	ARGB_ECOLI	Argb	StrainA	R2	8619714.8	sp|P0A6C8|ARGB_ECOLI Argb
P0A6C8	ARGB_ECOLI	Argb	StrainA	R3	8871045	sp|P0A6C8|ARGB_ECOLI Argb
P11446	ARGC_ECOLI	Argc	StrainA	R1	5556074	sp|P11446|ARGC_ECOLI Argc
P11446	ARGC_ECOLI	Argc	StrainA	R2	5274056.5	sp|P11446|ARGC_ECOLI Argc
P11446	ARGC_ECOLI	Argc	StrainA	R3	5010103	sp|P11446|ARGC_ECOLI Argc



Protein.Group	Protein.Names	Proteins	Sample	Counts_mean	Counts_std	CV%
P0A6C5	ARGA_ECOLI	Arga	StrainA	5596840	5401065.386	96.50205091
P0A6C8	ARGB_ECOLI	Argb	StrainA	4656066.567	4306032.232	92.48218792
P11446	ARGC_ECOLI	Argc	StrainA	4685392	4357092.59	92.99312822

Output file: Full_list_proteins_summary_XXXXXXXXX-xxxxxxx.csv

Note

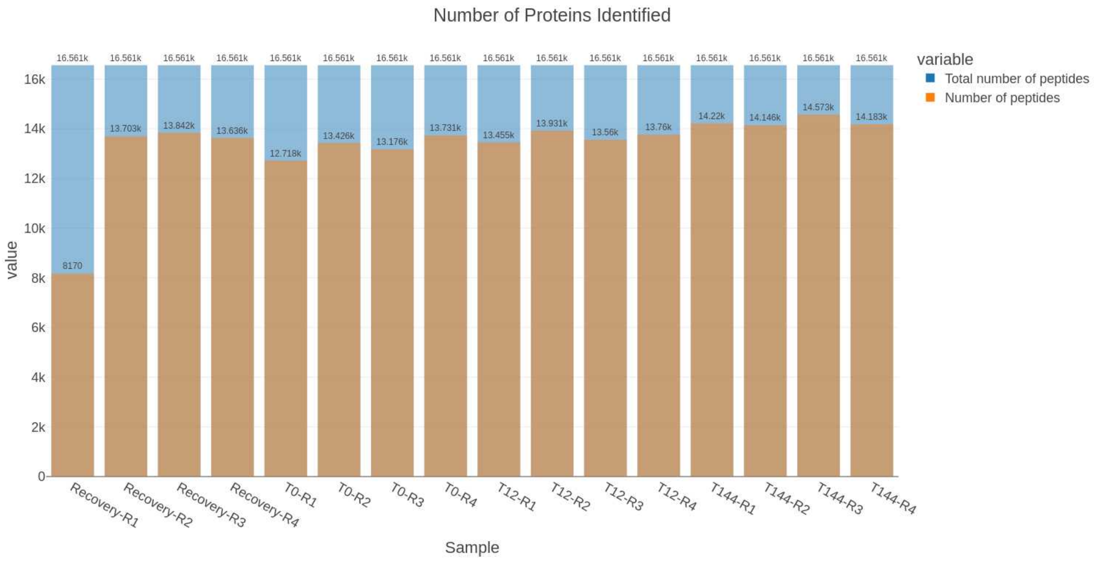
A similar output file is generated for a select list of proteins if one is provided. The resulting data table is exported as:
Selected_proteins_summary_XXXXXXXXX-xxxxxxx.csv

QC plots

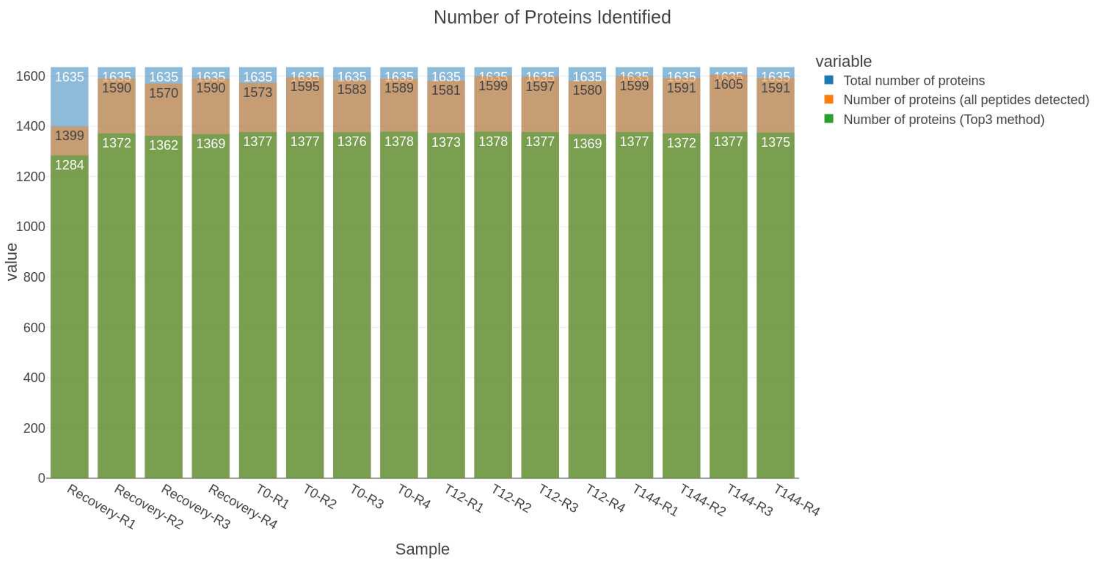
Found in the **QC_files** folder:

4

Bar plots of total proteins and peptides: The bar charts show the number of peptides or proteins identified and quantified by DIA-NN from each sample and the cumulative number for all the samples in the dataset. The protein plot also includes the number of proteins that meet the criteria for the Top3 protein quantification method.



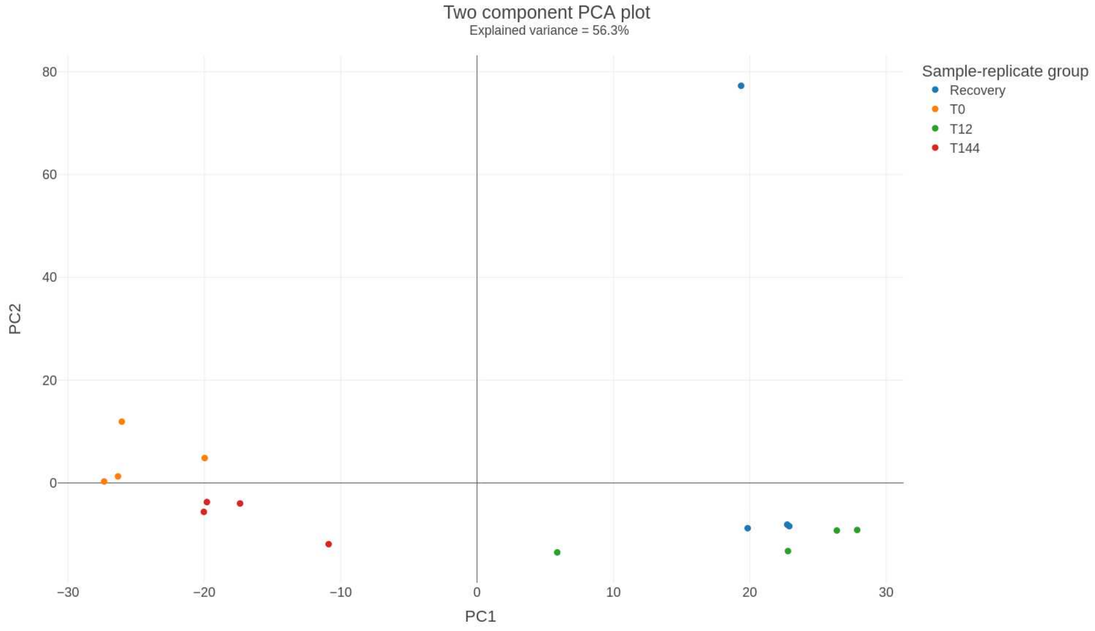
Example QC peptide bar plot



Example QC protein bar plot

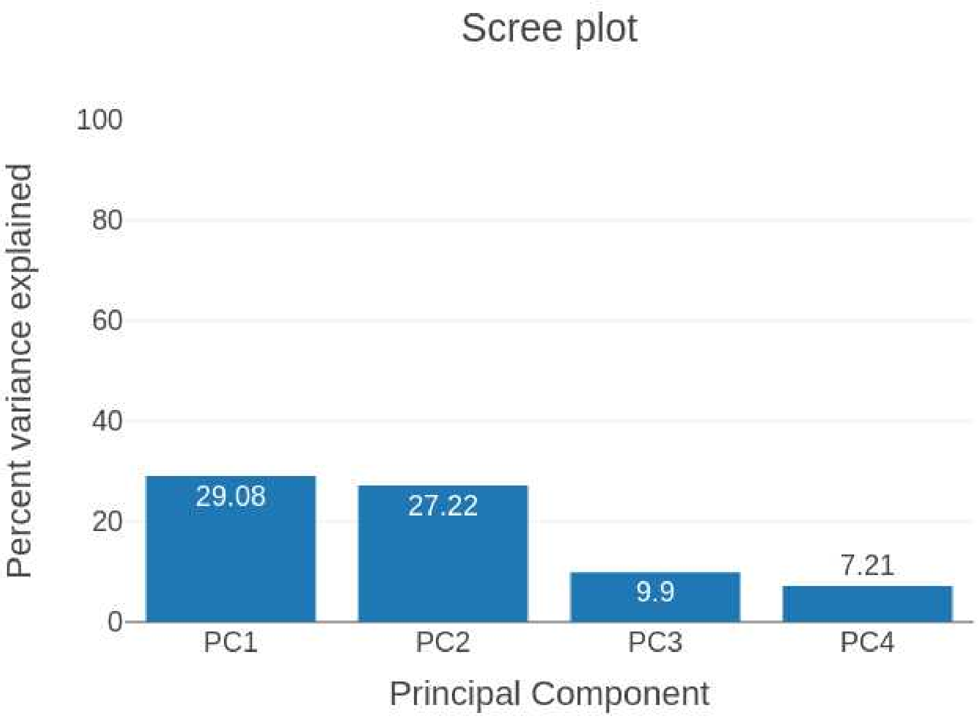
5 Found in the **PCA_plot** folder:

PCA plot: The PCA plot shows clusters of individual sample replicates based on their similarity. The amount of explained variance contributed by the first two principal components (PC1 + PC2) is shown as the subtitle. This plot can help identify outliers and the overall precision of the data.



Example PCA plot

Scree plot: The scree plot displays the variation contributed by the top four principal components from the data.

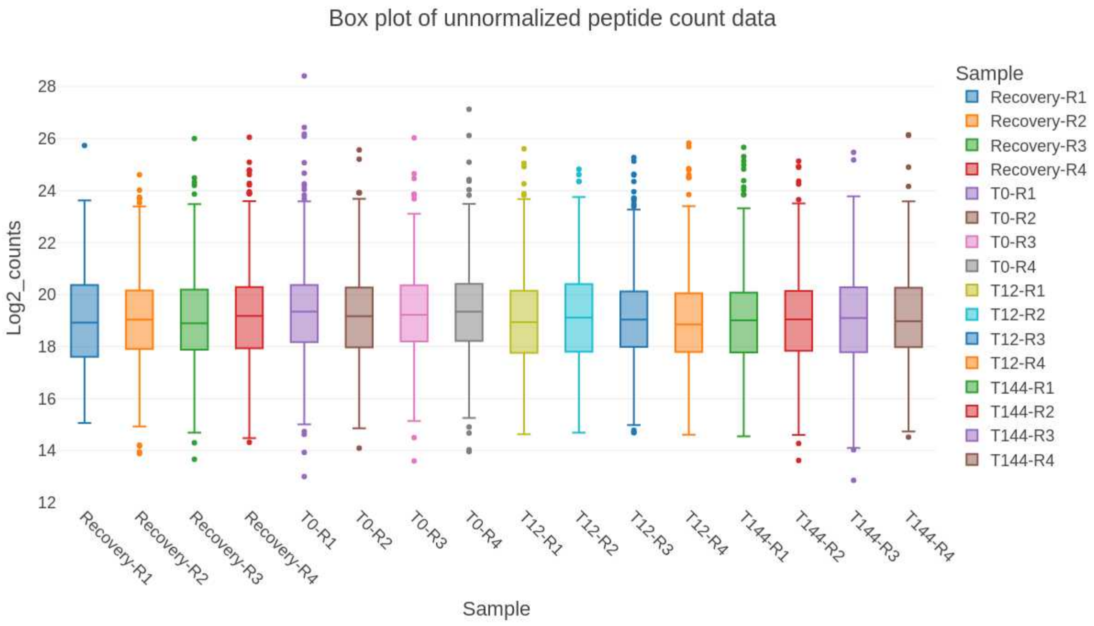


Example Scree plot

PCA plot calculations:

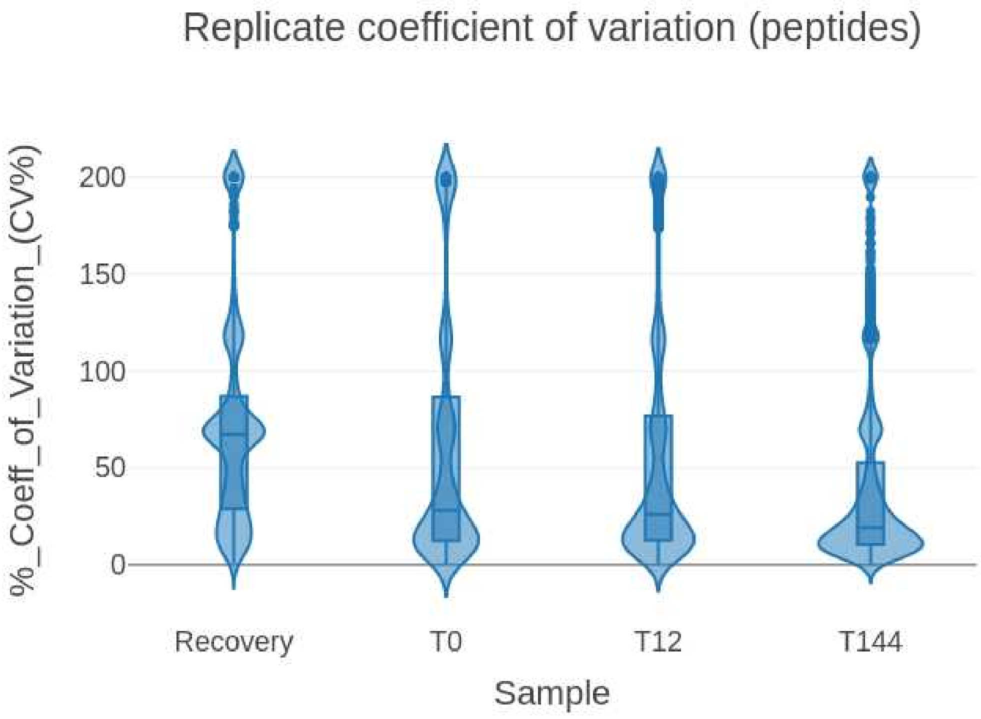
1. The data is scaled with the sklearn StandardScaler fit_transform method.
2. The PCA is implemented with the sklearn PCA method. The number of principal components are limited to 4.
3. Calculate the explained variance and the cumulative variance for the top two components
4. Plot the 2D PCA graph
5. Plot the Scree graph

6



Example box plots of the log2 peptide data across all sample replicates

7 Coefficient of variation (CV) violin plot:



Example violin plot of peptide CVs across samples

Top3 absolute protein abundance quantification

8 We use the Top3 quantification method (references below) to calculate the absolute protein abundance as fractions of total protein mass in each sample. Briefly, the Top3 quantification method is based on the “best flyer” hypothesis, which assumes that the specific MS signal intensity of the most intense tryptic peptides per protein is approximately constant throughout a whole proteome (ref: Ludwig et al. Mol. Cell. Proteomics 2012).

Our Top3 quantification analysis consists of:

1. Filter the DIA-NN peptide report data (from step 1) to only proteins that have three or more peptides identified across all samples
2. For each protein, rank the top 3 peptides by intensity (counts) in each of the samples
3. Calculate the mean rankings of the peptides in each protein across all samples
4. Filter the data to the three highest ranked peptides in each protein
5. Calculate protein intensity (counts) by averaging the intensity (counts) of the Top3 peptides
6. Calculate the percent of the total protein abundance
   ((intensity of individual protein / sum of all protein intensities in a given sample) * 100)

The resulting data tables are exported as:
Top3 full peptide list:
Top3_Full_list_peptides_used_for_quant_XXXXXXXX-xxxxxx.csv

Top3 full protein list for each replicate:
Top3_Full_list_proteins_XXXXXXXX-xxxxxx.csv

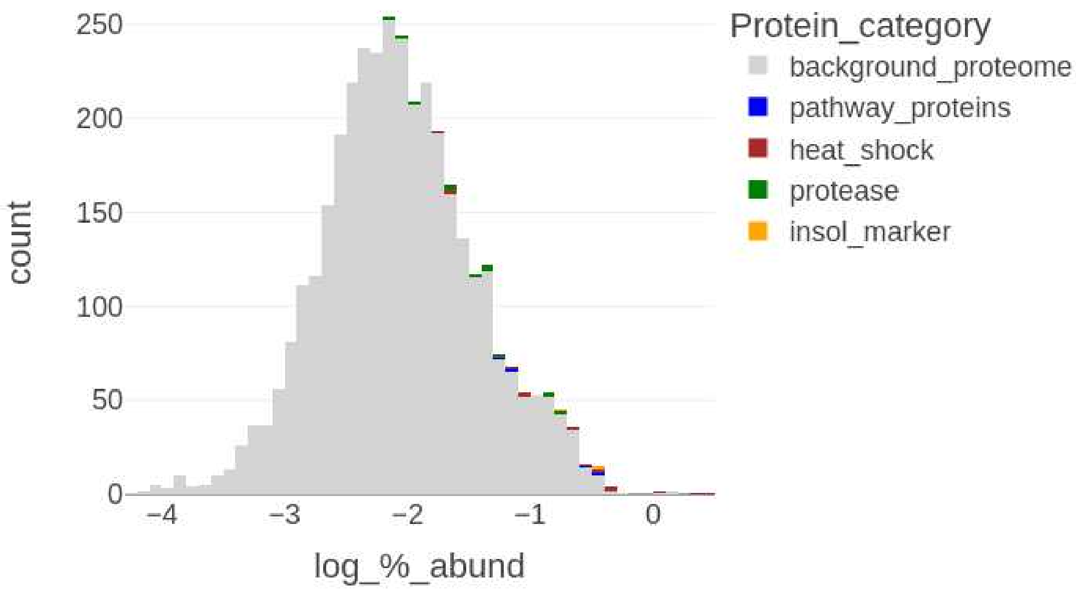
Top3 full list of proteins averaged across replicates:
Top3_Full_list_proteins_summary_XXXXXXXX-xxxxxx.csv

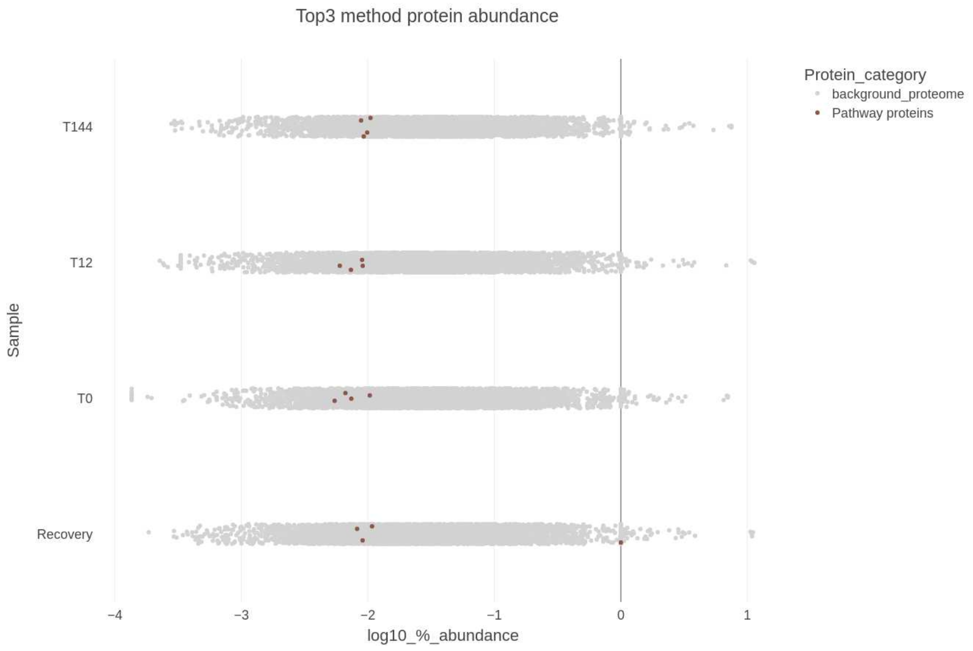
References:
Ludwig et al. DOI 10.1074/mcp.M111.013987
Silva et al. DOI 10.1074/mcp.M500230-MCP200
Ahrne et al. DOI 10.1002/pmic.201300135
Grossman et al. DOI 10.1016/j.jprot.2010.05.011

9 If a list of selected proteins is provided then a histogram for each sample is generated in .png, format with the categories of selected proteins highlighted with the background proteome:

- X-axis: log10 (% protein abundance) bins
- Y-axis: count of proteins in the bins

- Files generated:
  SampleID-histogram_XXXXXXXX-xxxxxx.png





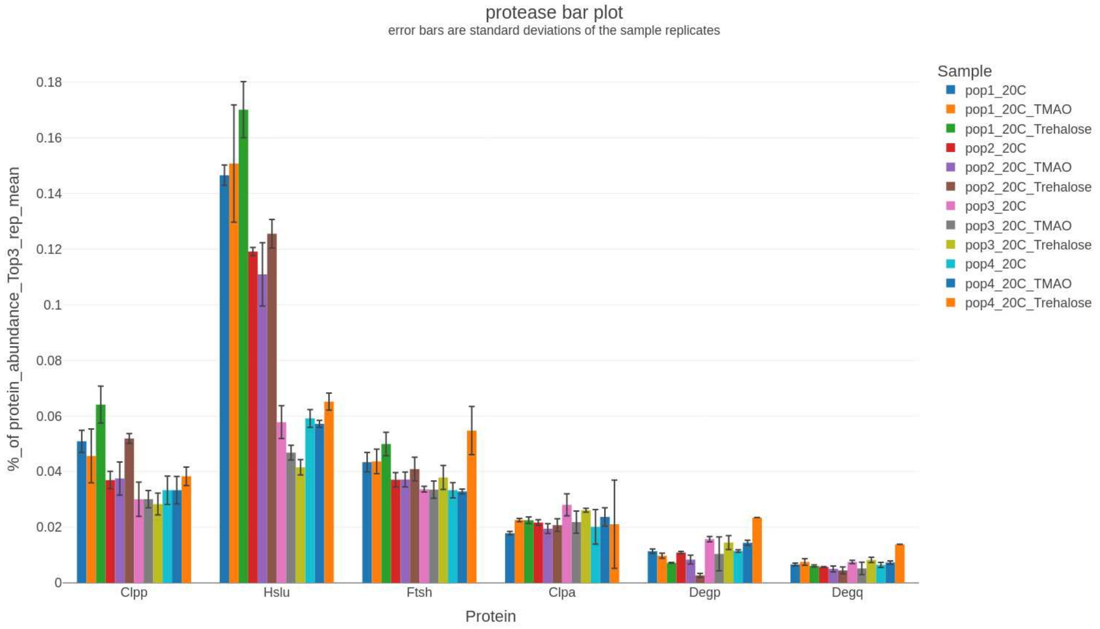
Example jitter plot showing the distribution of proteins across the percentiles of abundance. Groups of proteins from the selected_proteins.csv file are highlighted.

- Files generated:
  Top3_allsamples_jitterplot_XXXXXXXX-xxxxxx.png
  Top3_allsamples_jitterplot_XXXXXXXX-xxxxxx.plotly

Selected Proteins: Bar Charts

10 If a list of selected proteins is provided bar charts are generated in .png, .svg, and .plotly formats:

- X-axis: Proteins
- Y-axis: % protein abundance (from Top3 quantification method) averaged over replicates
- Error bars: standard deviation of % protein abundance (from Top3 quantification method) from the replicates
- Files generated:
  Full_and_select_proteins_summary_XXXXXXXX-xxxxxx.xlsx
  selectproteincategory-bar_XXXXXXXX-xxxxxx.png
  selectproteincategory-bar_XXXXXXXX-xxxxxx.svg
  selectproteincategory-bar_XXXXXXXX-xxxxxx.plotly



Safety information

Note: Missing proteins - Some proteins may not meet the Top3 criteria (at least three peptides detected across all samples), so they won't be quantified and shown on the bar and strip charts.

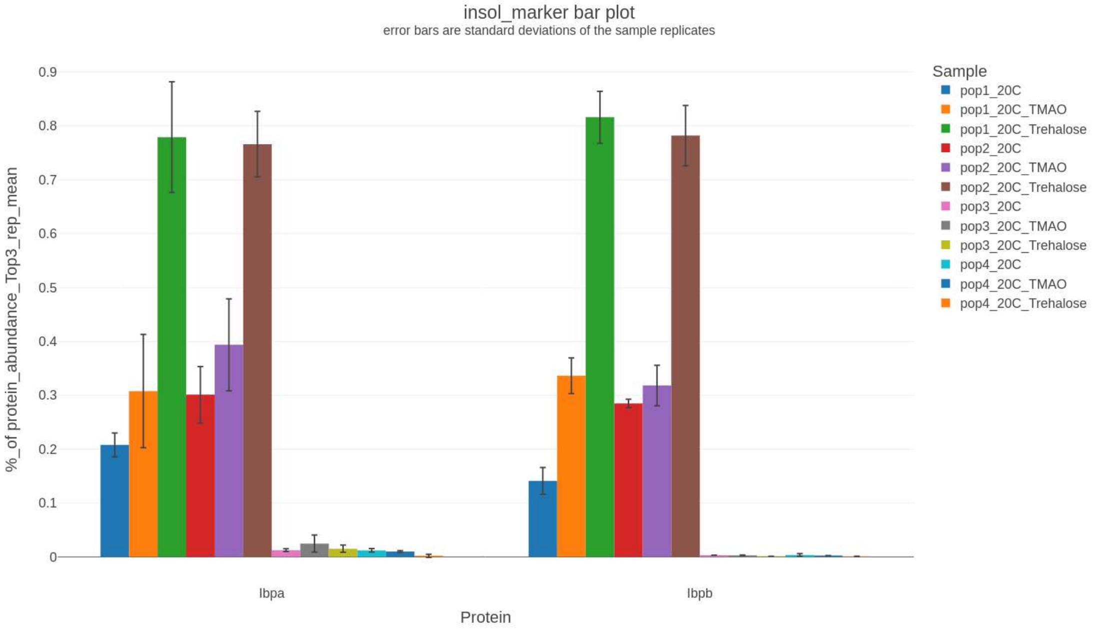
Note

NOTE: You can visualize the .plotly files by using Plotly or a Colab jupyter notebook. This provides an interactive view that you can see data labels, zoom, and save parts of the plot as separate .png files.

11 If other commonly analyzed proteins (e.g., insoluble protein diagnsotic marker proteins, proteases, heat shock proteins) are detected and quantified then a bar chart is generated in .png, .svg, and .plotly formats with only the corresponding data:

Example filenames:
insol-marker-bar_XXXXXXXX-xxxxxx.png

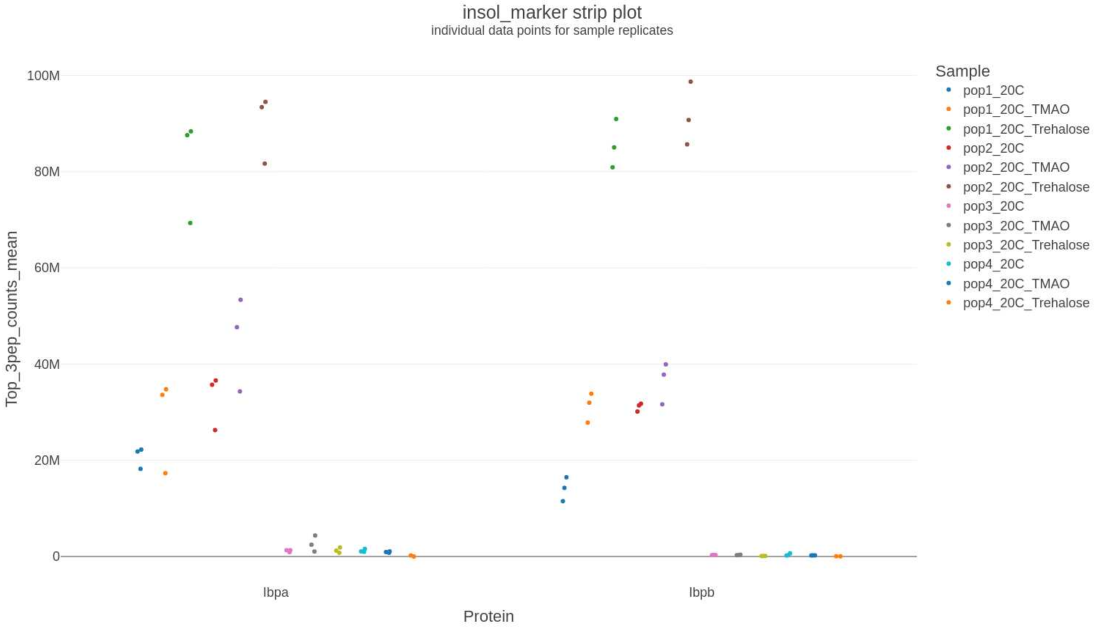
insol-marker-bar_XXXXXXXX-xxxxxx.svg
insol-marker-bar_XXXXXXXX-xxxxxx.plotly



Selected Proteins: Strip Charts

12 If a list of selected proteins is provided strip charts are generated in .png, .svg, and .plotly formats to show the individual data points for each sample:

- X-axis: Proteins
- Y-axis: Protein intensity (counts) calculated from the mean of the top 3 peptides for each sample replicate
- Error bars: none
- Files generated:
  selectproteincategory-strip_XXXXXXXX-xxxxxx.png
  selectproteincategory-strip_XXXXXXXX-xxxxxx.svg
  selectproteincategory-strip_XXXXXXXX-xxxxxx.plotly

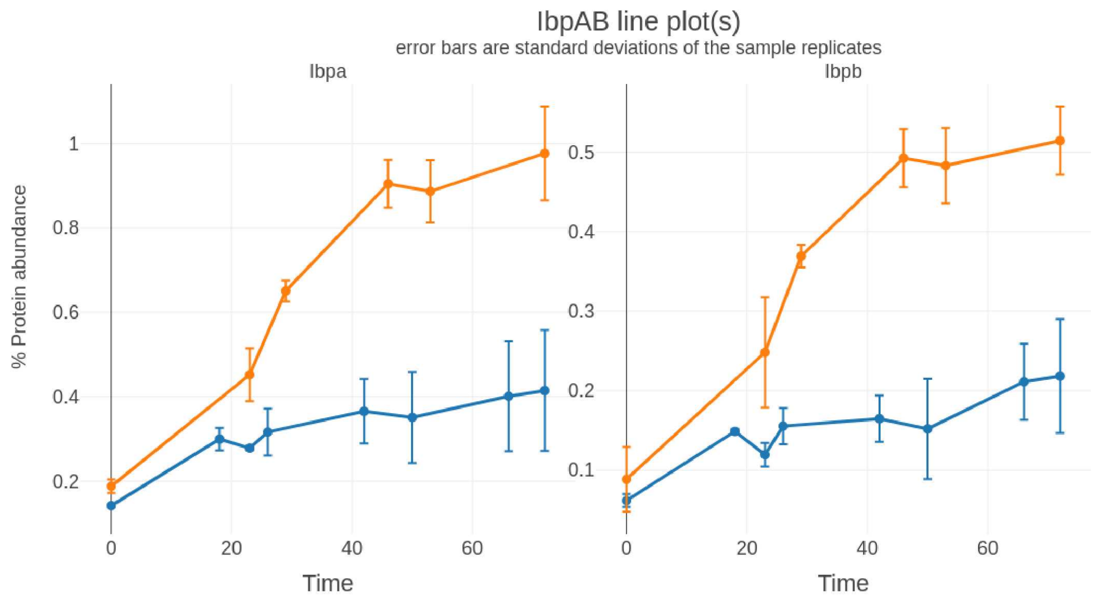


Note

Note: Missing proteins - Some proteins may not meet the Top3 criteria (at least three peptides detected across all samples), so they won't be quantified and shown on the bar and strip charts.

Selected Proteins: Line Charts

13 If a list of selected proteins is provided **AND** the sample names contain timepoint information (e.g., CJP1234_24hr-R1) line charts are generated in .png, .svg, and .plotly formats to show the individual data points for each sample:



- Sub-plot: Protein
- X-axis: Time
- Y-axis: % protein abundance (Top3 method)
- Error bars: standard deviation of % protein abundance (Top3 method) from the replicates
- Files generated:
  selectproteincategory-line_XXXXXXXX-xxxxxx.png
  selectproteincategory-line_XXXXXXXX-xxxxxx.svg
  selectproteincategory-line_XXXXXXXX-xxxxxx.plotly

Sample A-B comparisons: t-Test and volcano plots

14

If applicable, two samples (A and B) are selected for comparison then a Welch's t-Test is performed by using the ttest_ind_from_stats function from scipy (details here). This is comparable to the Excel function t-Test: Two-Sample Assuming Unequal Variances.

For this analysis:

- Missing values and zero abundance values are imputed with the lowest of detected (LOD) value in each sample.
- Abundance values are log2 transformed prior to the t-Test
- The False Discovery Rate (FDR; adjusted p-value; q-value) is calculated by the Benjamini-Hochberg method by using the statsmodels.stats.multitest.multipletests function.

Significantly changing proteins are defined as:

- a p-value (or adjusted p-value) < 0.05
- a fold change of > 2 (UP) or < 0.5 (DOWN)

The resulting data tables are exported as an Excel file (xlsx):
t-Test_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv

with five sheets corresponding to:

1. Full t-test output
2. p-value Significant UP changing proteins (p-value <0.05)
3. p-value Significant DOWN changing proteins (p-value <0.05)
4. adjusted p-value Significant UP changing proteins (adjusted p-value <0.05)
5. adjusted p-value Significant DOWN changing proteins (adjusted p-value <0.05)

Note

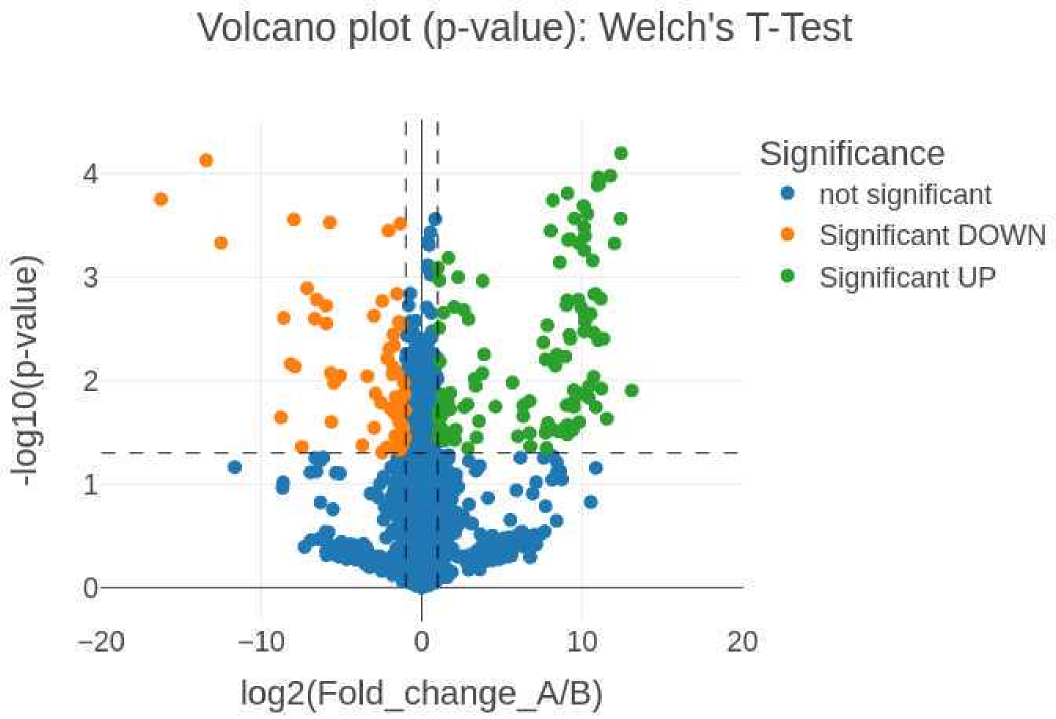
Note: The definition of 'significance' for your experiment may be different from these values. You can use the full t-test output to select data based on your criteria or process the full dataset as needed.

15 If a list of selected proteins is provided two volcano plots are generated in .png, .svg, and .plotly formats (six total volcano plot visualization outputs) for the two sample comparisons:

- log2 (Fold change) vs. -log10(p-value) plots
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.png
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.svg
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.plotly
- log2 (Fold change) vs. -log10(adjusted-p-value) plots
  Volcano_plot_adj-p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.png
  Volcano_plot_adj-p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.svg
  Volcano_plot_adj-p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.plotly

The significance cutoffs are defined as:

- Fold Change = 0.5x and 2x (-1 and 1 on the log2 axis)
- p-value and adj-p-value = 0.05 (1.3 on the -log10 axis)



Volcano Plot

Note

NOTE: You can visualize the .plotly files by using Plotly or a Colab jupyter notebook. This provides an interactive view that you can see data labels, zoom, and save parts of the plot as separate .png files.

Note

NOTE: Typically there are more significantly changing (UP & DOWN) proteins observed in the p-value plot than the adjusted-p-value plot. Which plot is most applicable for your experiment will depend on the questions of interest.