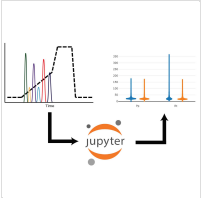


Aug 02, 2022

# Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files

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

[1]Lawrence Berkeley National Laboratory



dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

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.

DOI

dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

PROTOCOL CITATION

Yan Chen, Christopher J Petzold 2022. Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files. **protocols.io** https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

KEYWORDS

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

CREATED

Mar 18, 2022

LAST MODIFIED

Aug 02, 2022

PROTOCOL INTEGER ID

59616

PARENT PROTOCOLS

In steps of

Discovery proteomic (DIA) LC-MS/MS data acquisition and analysis

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:
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

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 TEXT

Required:

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

Optional:

- A list of selected proteins for bar chart visualization
- A list of two-sample comparisons of different samples (Sample A vs. Sample B; Sample B vs. Sample C, etc.)

DISCLAIMER:



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

This protocol is for research purposes only.

BEFORE STARTING

**INPUTS:**

Required:

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

Optional:

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

**OUTPUTS:**

- Abundance values correspond to summed peptide peak area in arbitrary units
- 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.

- **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
- **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)

If applicable:

- **Summary data table of a selected list of proteins (CSV)** - a list of selected protein quantitative values averaged over the sample replicates
- **Bar charts of selected proteins in .png, .svg, and .plotly formats**
- **Individual bar charts of protease, heat shock, or insoluble expression marker protein(s) abundance if they are present (.png, .svg, and .plotly formats)**
- **Data tables with the Welch's t-test results**
- **Data tables with a list of the significantly UP and DOWN regulated proteins** if there are any
- **Volcano plots visualizing the Welch's t-test p-value significance and log(2)**



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

**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)**

## Data processing

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
- Sample
- Replicate
- Abundance value (Peptide peak area in arbitrary units)

2 The peptide abundance values are summed to the protein abundances. The resulting data table is exported as:
**Full_list_proteins_XXXXXXXXX-xxxxxxx.csv**



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

| | 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

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 abundances of the sample replicates are averaged (mean), the standard deviation (SD), and percent coefficient of variation (CV%) are calculated. The resulting data table is exported as:
**Full_list_proteins_summary_XXXXXXXXX-xxxxxxx.csv**



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

| | 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 |

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Protein.Group | Protein.Names | Proteins | Sample | Counts_mean | Counts_std | CV% |
| 2 | P0A6C5 | ARGA_ECOLI | Arga | StrainA | 5596840 | 5401065.386 | 96.50205091 |
| 3 | P0A6C8 | ARGB_ECOLI | Argb | StrainA | 4656066.567 | 4306032.232 | 92.48218792 |
| 4 | P11446 | ARGC_ECOLI | Argc | StrainA | 4685392 | 4357092.59 | 92.99312822 |

Output file: Full_list_proteins_summary_XXXXXXXXX-xxxxxxx.csv

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**

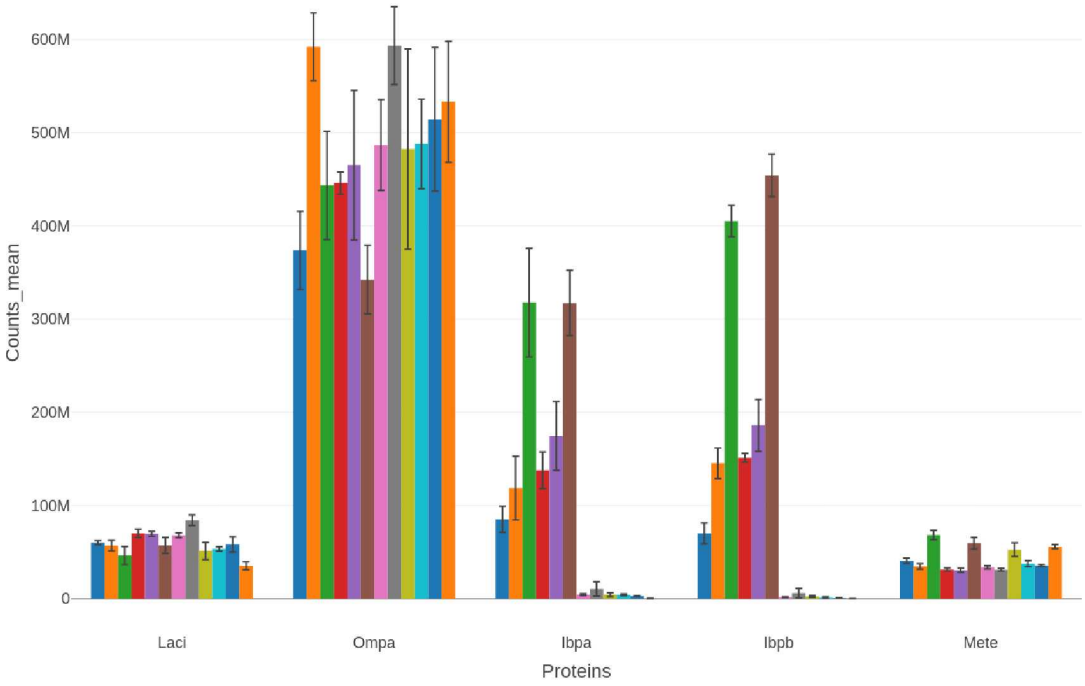
## Selected Proteins: Bar Charts

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

- Proteins
  selectproteins-bar_XXXXXXXX-xxxxxx.png
  selectproteins-bar_XXXXXXXX-xxxxxx.svg
  selectproteins-bar_XXXXXXXX-xxxxxx.plotly



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1




Proteins labels

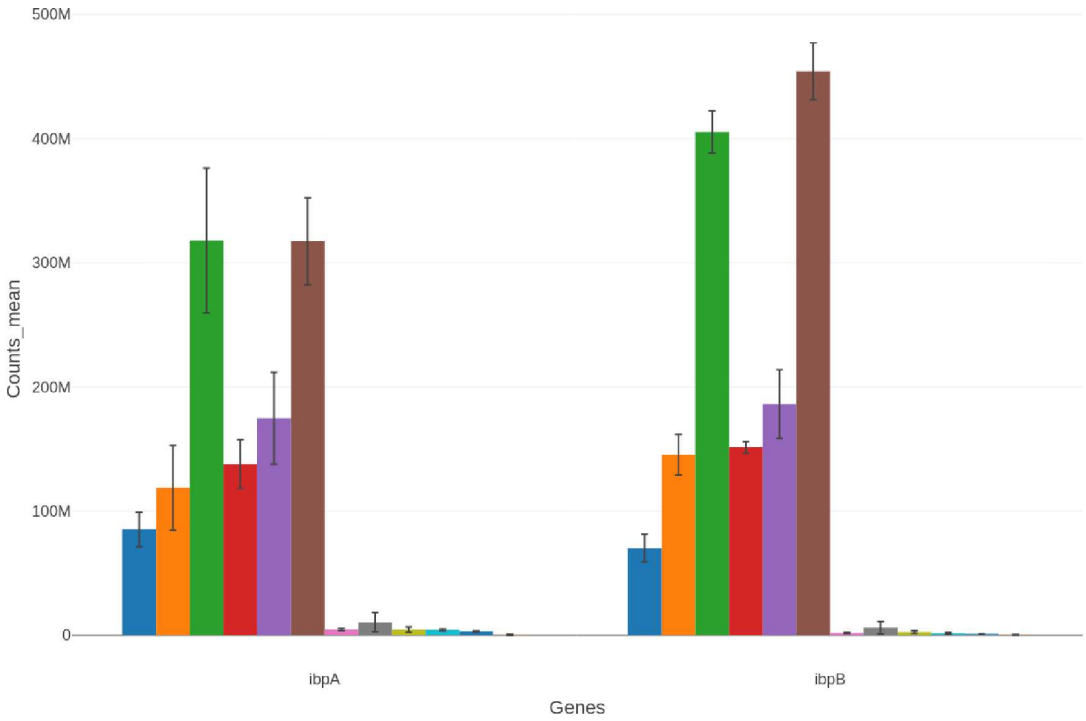
**NOTE**: You can visualize the .plotly files by using Plotly or a Colab jupyter notebook. This provides an interactive view that you can see data labels, zoom, and save parts of the plot as separate .png files.

5 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



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1




Insoluble protein bar chart

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

6 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 filled with '1000', a value that is just below our level of detection.
- 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:
Full t-test export:
t-Test_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv

Significant changing proteins (p-value <0.05):
t-Test_signifDOWN_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv
t-Test_signifUP_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1

Significant changing proteins (adjusted p-value <0.05):
t-Test_signifDOWN_adj-p_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv
t-Test_signifUP_adj-p_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.csv

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.

7 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:

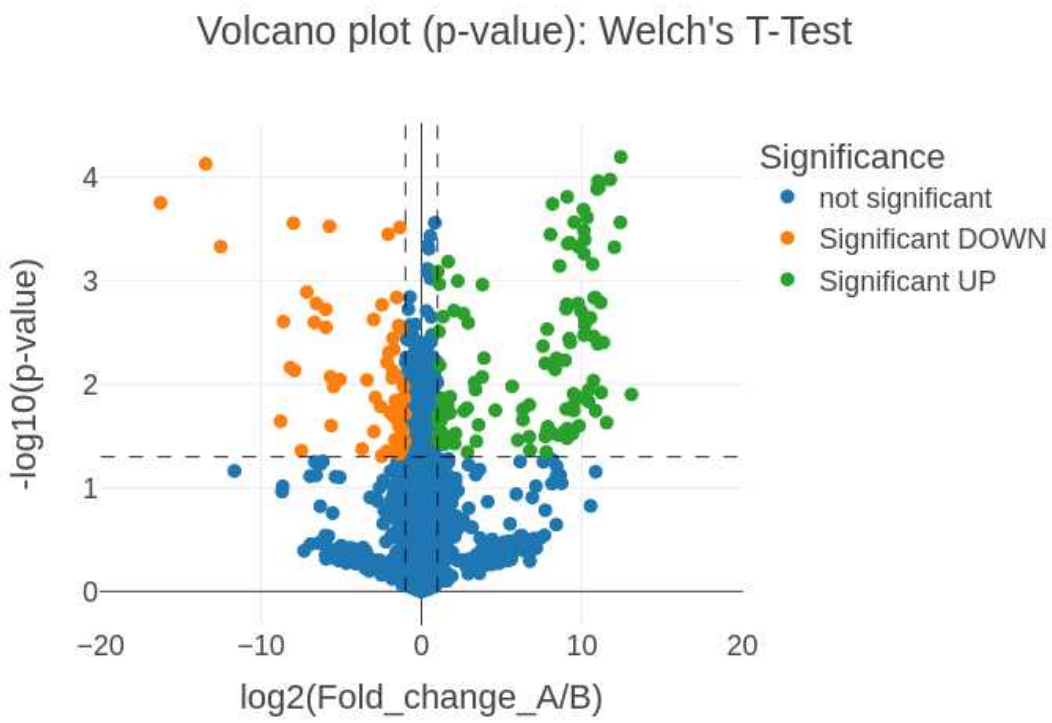
- log2 (Fold change) vs. -log10(p-value) plots
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.png
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.svg
  Volcano_plot_p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.plotly
- log2 (Fold change) vs. -log10(adjusted-p-value) plots
  Volcano_plot_adj-p-value_SampleA_OVER_SampleB_XXXXXXXX-xxxxxx.pn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



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1




Volcano Plot

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: 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.



**Citation:** Yan Chen, Christopher J Petzold Label-free quantification (LFQ) proteomic data analysis from DIA-NN output files https://dx.doi.org/10.17504/protocols.io.5qpvobk7xl4o/v1