PCP: An R-based application for visualization data from protein correlation profiling

Prepared for Noah Dephoure, Weill Cornell Medical College
by the Applied Bioinformatics Core (ABC), Weill Cornell Medical College

revised June 13, 2017

Contents

1 Installation and updating

The routine for installing and updating the package is the same (with the exception of the `devtools` installation, which must be done only once).

1. Open RStudio.
2. In the Console section of RStudio, type:

```
# install helper package (must be done only once!)
> install.packages("devtools")

# download the source of the package and install it
# note that the version number will change depending on the update
# you can check which version is appropriate at http://chagall.med.cornell.edu/deplab-pcp/
> devtools::install_url(url = "http://chagall.med.cornell.edu/deplab-pcp/DepLab_0.1.1.9000.tar.gz")

# attach the package to your current R environment
> library(DepLab)

# start the interactive app
> runPCP()
```

2 Start

If you have installed the most recent version of the package and you just want to continue using the app, all you need to do is:

1. Open RStudio.
2. In the Console section of RStudio, type:

```
> library(DepLab)
> runPCP()
```

   This will start the application in a new window.

3 Using the app - overview

There are three main parts to the app:

Data Import	specify the file that contains the results of a protein correlation profiling experiment and add it to the data base
Visualization	generate plots for all values that `MaxQuant` reported per eluted fraction
Tables	explore the values underlying a specific plot in tabular format

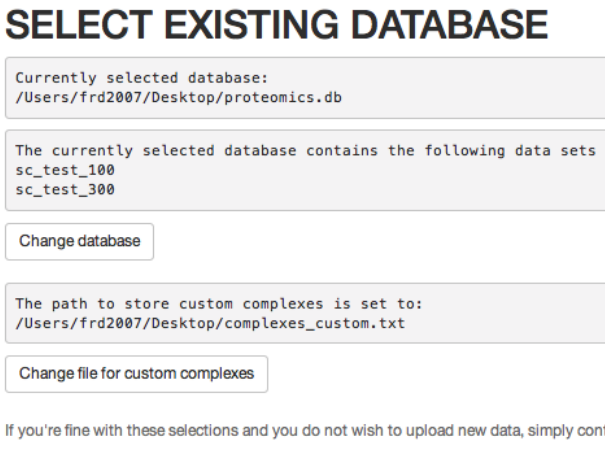
While the Data import panel will always be shown at the start of the app, you can skip right to the visualization or tables part without uploading a new file. You will, however, need to make sure you have selected the correct data base.

4 Data import

The app allows for the import of a table of proteins that were identified during a protein correlation profiling experiment. Every file that is uploaded will be stored in a relational `sqlite` database. This allows for storing (and retrieval) of multiple data files in a consistent manner including customized metadata entries, such as information about the specific experiment, the date of data acquistion etc. (see Section **??**). This also means that the database will grow over time as more and more files are added.

> ! The data base should be regularly backed up.

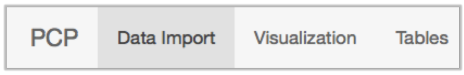
Once you start the app, you will be shown the path to the database that you used during your last session. You will also see the experiment IDs of the data sets that were previously stored. You can change the path to the data base via the Select database button, but we strongly recommend to keep all the data in *one* database rather than scattering information from different experiments into different databases.



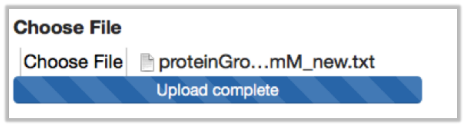
In addition, you can check and change the path to the text file where the custom-defined complexes are stored.

To upload a file, the current workflow is as follows:

1. Select the "Data Import" tab.



2. Click on "Choose File" and navigate to the `MaxQuant` output file. Currently, the app accepts the `proteinGroups*.txt` output of `MaxQuant` [4, 6]. For each protein group, only the founder protein will be extracted. There should be a bar indicating that the upload was successful.



3. Specify a *unique experimental ID* which will help you identify the experiment later on.

4. Select the *organism* from which the samples of that particular experiment originated from.

5. If you know the UniProt ID of the protein you have used as a *standard* for that particular experiment, indicate that in the field below the experimental ID. If you do not specify a UniProt ID, the default is to retrieve the values related to trypsin (from pig) in that data set.

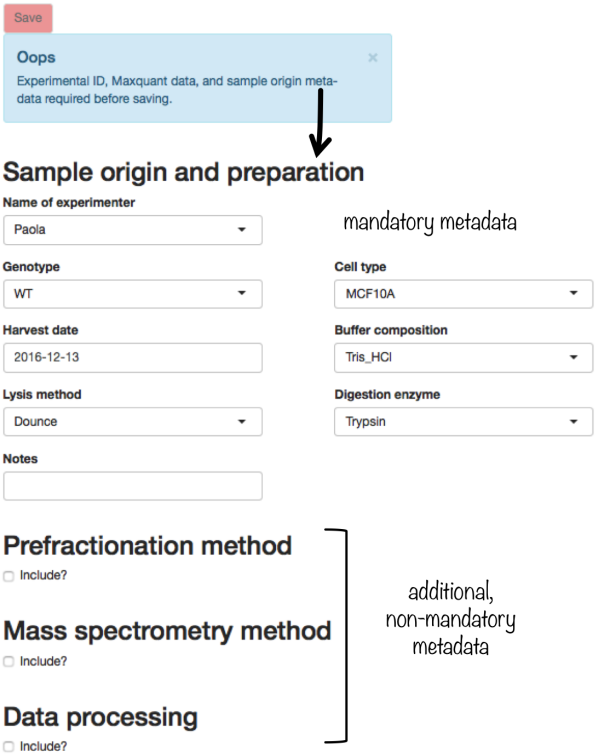
6. Fill out all the relevant metadata connected to that particular experiment (for details, see Section 4.1).

7. Click "Save". Since this will load all the entries from the `MaxQuant` output into the database, this may take a while. Once the updating of the database has succeeded, you will see a corresponding message.

It is not possible to upload a file with a previously used experimental ID. Metadata, however, can be updated and corrected after upload (see Section 4.1.1).

4.1 Metadata

As of `DepLab v0.1.1`, you will be required to provide some basic metadata, i.e., additional information about the experiment for which you are uploading the data. The mandatory entries include the name of the experimenter, the genotype, harvest data, lysis method etc.

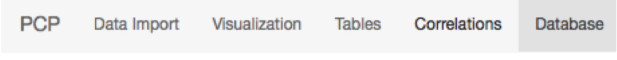


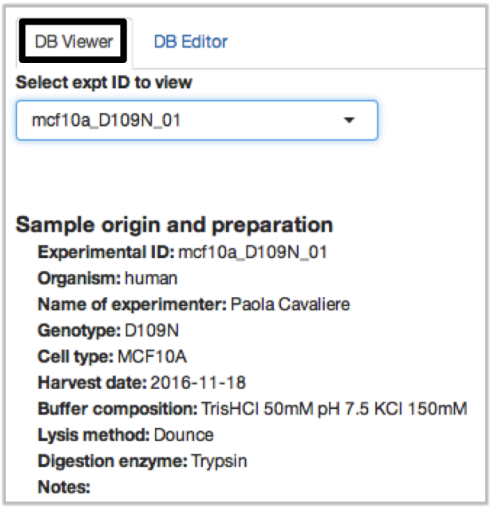
Additional types of metadata can be added:

- Prefraction method captures information such as column ID, amount of loaded protein, sample volume, time per and number of fractions.
- Mass spectrometry method expects information about the Instrument ID, the run date (default will be the current day) and the length of the method.
- Data processing lets you keep track of basic software settings that were used for the processing of the spectra, such as the search and filtering algorithms.

If you need more pre-defined entries – e.g., for the name of the experimenter – get in touch with us (`frd2007@med.cornell.edu`).

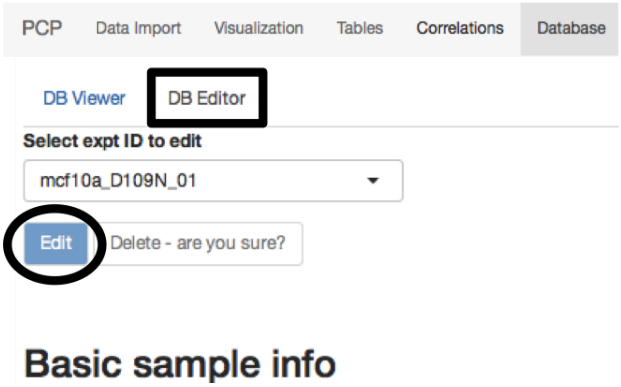
You can check the metadata entered for the experiments within the selected database in the Database section (DB Viewer).





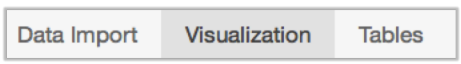
4.1.1 Editing the metadata already stored in the data base

To edit the metadata related to experimental data within the selected database, go to the DB Editor in the Database section. Just select the experiment ID of the experiment for which you would like to change the metadata. Then add the respective changes and hit "Save changes".



You can also delete all information related to an entire experiment by clicking theDelete button.

5 Visualizing protein profiles

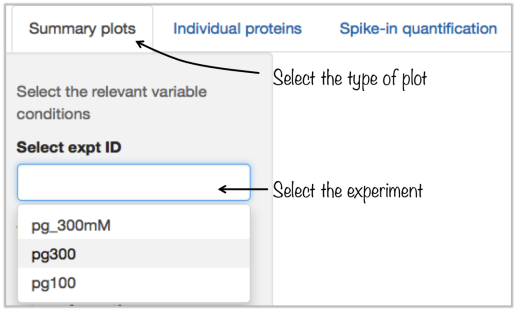


There are three kinds of plots that can currently be generated to visualize the fraction-wise protein profiles:

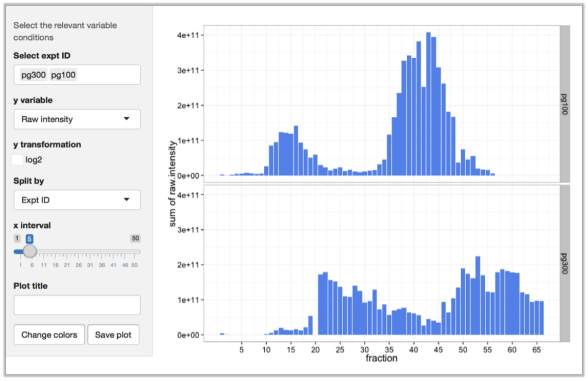
1. **Summary plot:** Depicting the sums of values for all protein groups per fraction.
2. **Individual proteins:** Tracking the measured values across all fraction for individual proteins and complexes.
3. **Spike-in:** Tracking the values for the protein spike-in controls across the fractions.

The general steps to obtain a plot are very similar for all three types:

1. Select the *type of plot* that you would like to generate.
2. Select *one or more experiments*. The field "Select expt ID" will contain all experimental IDs stored in the database. You can choose by scrolling through the list or typing the specific ID.



Once you select the experiment ID, a plot will immediately be generated using default settings that currently include to use the raw intensity for the y variable and to make a separate plot for each selected experiment ID.



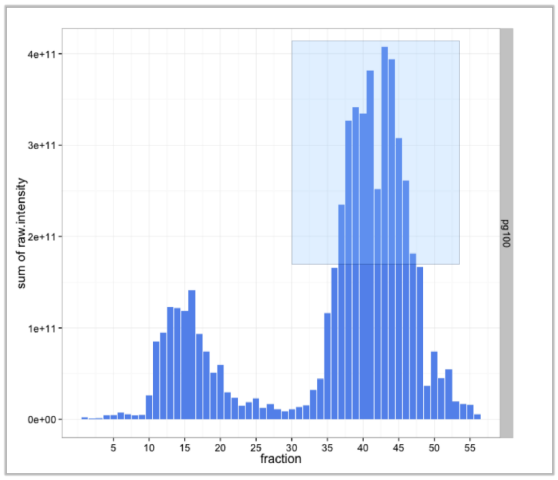
All images can be saved as pdf or png files ("Save plot").

5.1 Interacting with the protein profiles

The plots can be customized in various ways. Every change will immediately refresh the image.

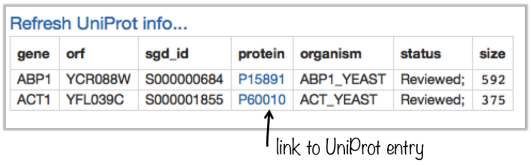
The *size of the plot* is connected to the size of the app's display. If you reduce the size of the app's window, the plot will be adjusted accordingly.

To *zoom* into certain fractions, simply use the mouse to draw a rectangle around the region of interest. Then double-click into the light blue blox.



To zoom out to the original scale, double click anywhere within the image.

To *retrieve basic information* from UniProt for each depicted molecule, click on the respective link underneath each plot (scroll down if you don't see it immediately). Since the download from UniProt takes some time, this list is not automatically updated if you change the selection of proteins. You will have to refresh that list manually. If you click on the individual UniProt IDs within the result table, the corresponding web site will be opened in your Internet browser.



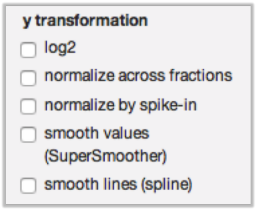
5.2 Values for display

The following measurements are available for all three kinds of plots ("y variable"):

- Raw intensity
- LFQ intensity
- MS count
- Peptides count
- Unique peptides only
- Razor and unique peptides
- Sequence coverage

These are all the numerical values that `MaxQuant` supplies per fraction.

For the *individual protein plots*, these numerical values can be modified to improve the effectiveness of the visualization.



- Normalize across fractions: Each measured value, m_f, is normalized so that proteins with equal measurements in each fraction, f, have a normalized value of 1 per fraction: $m_{norm} = \frac{m_f \times F}{\sum_{f=1}^{F} m}$.

- Spike-in control: Each measured value, m_f, is normalized to the corresponding value, s_f, of the spiked-in protein: $m_{norm} = \frac{m_f}{s_f}$.

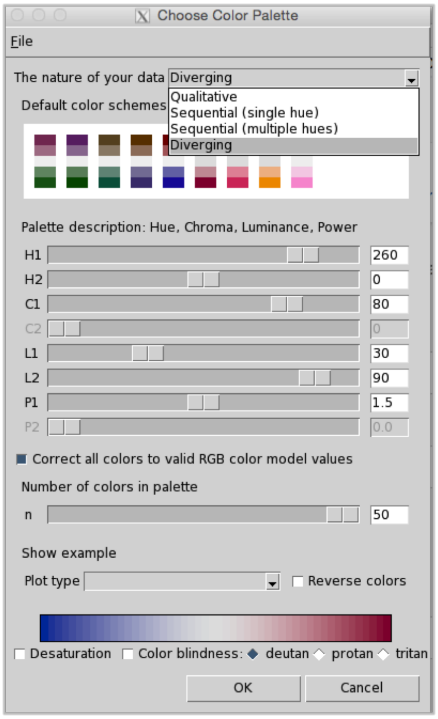
- Smooth values (SuperSmoother): This will apply Friedman's SuperSmoother, which is a sophisticated moving average method [5].

- Smooth lines (spline): This applies a function that interpolates the numerical values using polynomial functions [3].

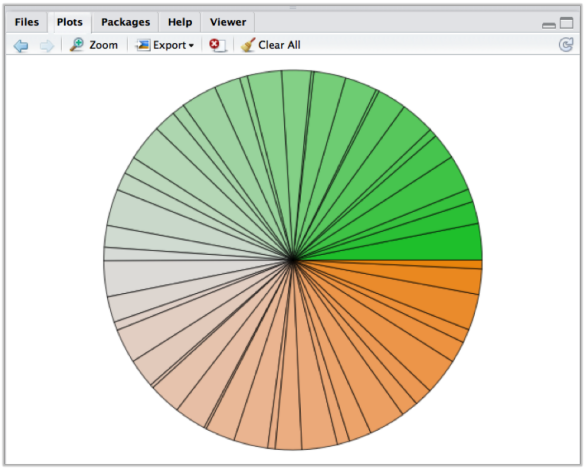
5.3 Customizing the colors

You can change the color scheme via the button "Change colors". Once you click on it, a separate window will appear which will allow you to change the type of color scheme (qualitative, sequential, diverging), choose the colors, manipulate hue[1], chroma[2], luminance[3] and power, and specify the numbers of colors that will be used.

1. Hue: The degrees on the RGB color wheel where Red = 0, Green = 120, Blue = 240.
2. Chroma: The saturation of the hue indicated with values 0 (black) to 255.
3. Luminance: Perceived brightness of a color; depends on the hue as well as the saturation.

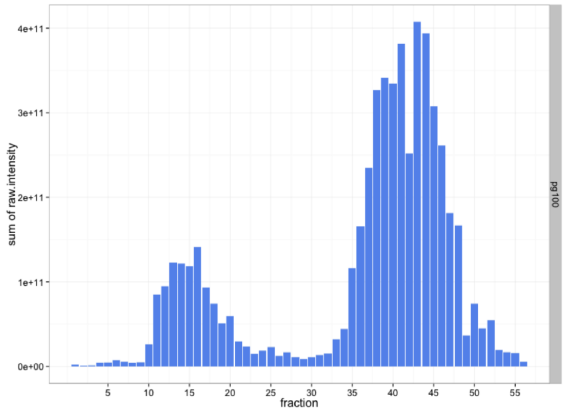


The bar at the bottom will reflect your choices and should give you a good idea of how your plot will be affected. In addition, you can choose to see an example plot, e.g. a pie chart. Note that this example will not be shown within the app, but in RStudio's Plot environment!



5.4 Summary plots

Summary plots display the sum of all measurements per fraction. The default setting will make one plot per selected experiment ID. If you would like to visually distinguish the different fractions, choose a different color scheme.

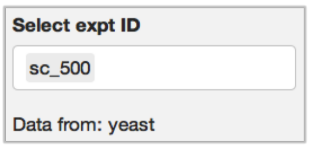


Since the summary plots will have to retrieve and compute *all* values per experiment, the generation of the image may take a couple of seconds.

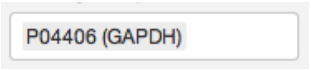
5.5 Individual proteins

The individual protein plots are useful to track the measurements for single protein across the different fractions.

As for the summary plots, start by *selecting the experiment ID(s)*. Once you have selected the first experiment ID, the list of possible additional experiment IDs will be reduced based on the organism of your first selection. This way, you will always only be able to compare samples from the same organism. This also means that you should wait for the successful loading of the first data set which will be indicated by a message stating the organism of your choice.



To display protein profiles of interest, select them by their gene name or their UniProt ID, e.g. "GAPDH".

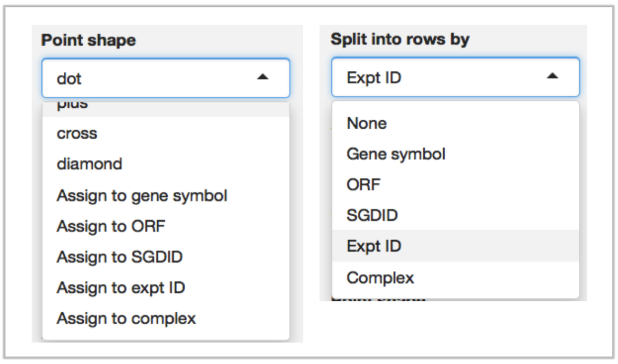


Customizing the plots There are numerous options to customize the display to allow for useful comparisons, e.g., across experiments and between the members of distinct protein complexes.

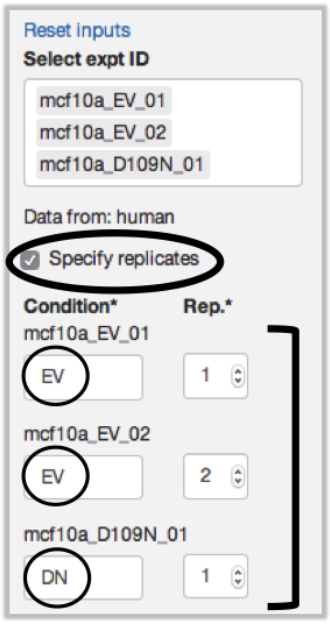
Typically, it will make sense to *assign the color based on the gene symbol*. The legend will always be sorted alphabetically, and so will the table of basic information that you can retrieve from UniProt.

To compare multiple complexes across different experiments, we recommend to *split by rows* using the experiment IDs and to *split by columns* using the complex' IDs.

You can also choose to differentiate different gene symbols by the point shape or line type, but note that this makes only sense for very limited numbers of variables (up to 5 different genes, for example).



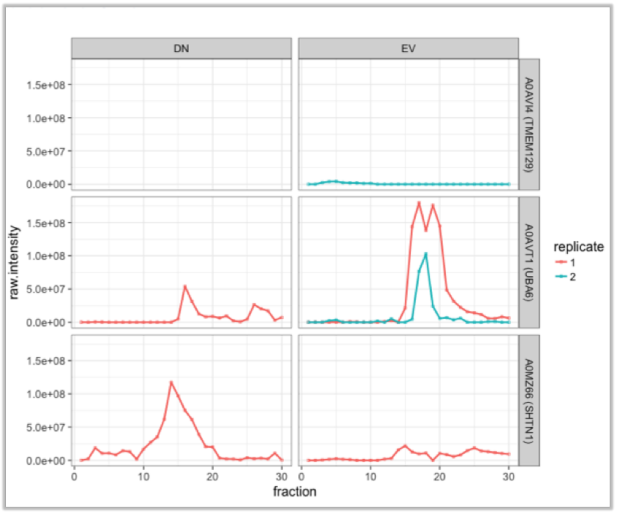
If you have replicates for the same experimental condition, you can specify both, the name of the condition and numeric replicate labels. For this, you will have to check the box Specify replicates underneath the box for the selection of the experiment IDs. You can then manually define the name of the experimental condition, which can then be used to, for example, split the plots.



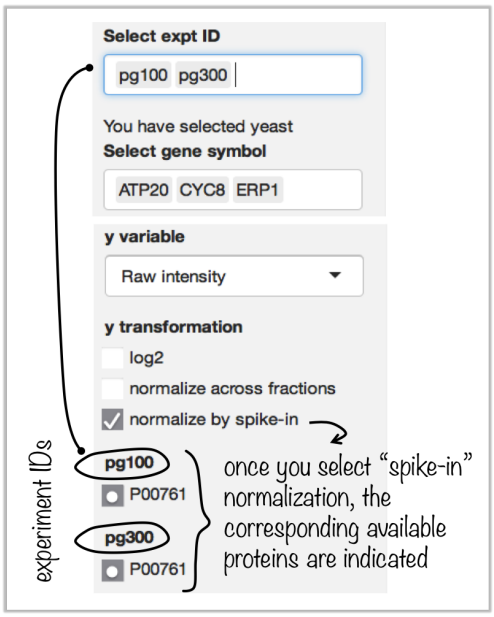
A useful combination of settings may be:

- Split into rows by: Gene symbol
- Split into columns by: Condition
- Color by: Replicate

This would result in the following figure:



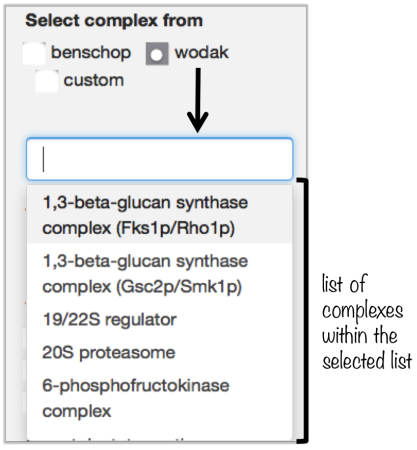
Normalization The values displayed on the y-axis can be *log*-transformed or normalized using two approaches: either across fractions or by using a spike-in control. When normalizing based on the values for spike-in controls, you will be able to choose the specific protein if more than one standard was indicated in the metadata section for any experiment.



For details on the normalization of the y variable, see Section 5.2.

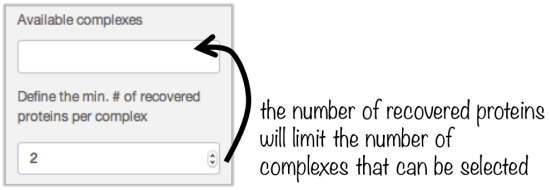
Complexes Instead of specifying individual gene symbols that should be displayed, you can also select a set of proteins that have been reported to be part of common complexes. We currently supply three lists of complexes: two for yeast based on publications by Benschop et al. [2], and Wodak et al. [7], and one for human complexes, CORUM Core [1].

To see the profiles for proteins of a complex defined by either one, e.g., select "Wodak", then use the empty field below to search for the complex of interest, e.g. "20S Proteasome". You can select multiple protein complexes. To add additional proteins, use the field "Select gene symbol".



Many complexes may encompass proteins for which your experiment does not provide any values. This will cause NA entries.

If you want to limit the complexes that are available in the auto-select box to those for which many proteins were recovered in your experiment(s), you can specify the minimum number of proteins per complex that should be present in *every* experiment.

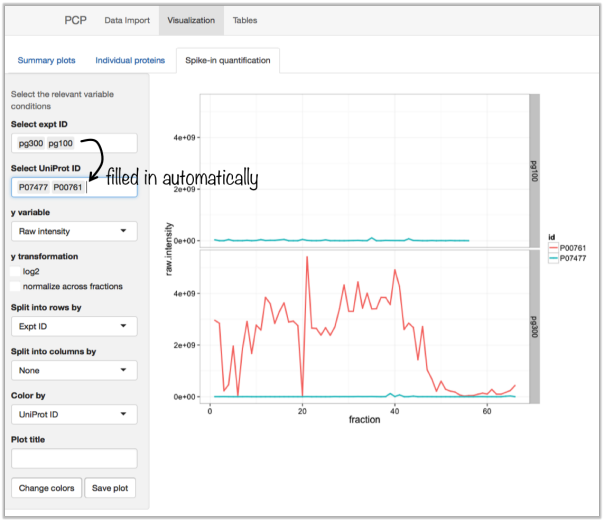


In addition to the pre-defined complexes, you can *generate your own groups of proteins*. Simply choose the individual proteins that you would like to group together via the "Select gene symbol" line, then hit "save as complex...". You will be asked to specify a name for the protein group. Once you have saved a group, it will be available in all future sessions: simply select the "custom" field and use the empty line below to select your custom-made group of interest.

5.6 Spike-in quantification

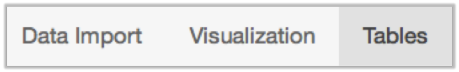
Ideally, each experimental ID should have the corresponding UniProt ID for the spike-in protein that was used. Therefore, all you will have to do here is to indicate the experiment(s) of interest and the profiles for the spike-ins related to the experiment(s) will be shown.

If no UniProt ID was specified during the upload of the file (as described in Section 4), Trypsin from pig (UniProt ID P00761) will be used.

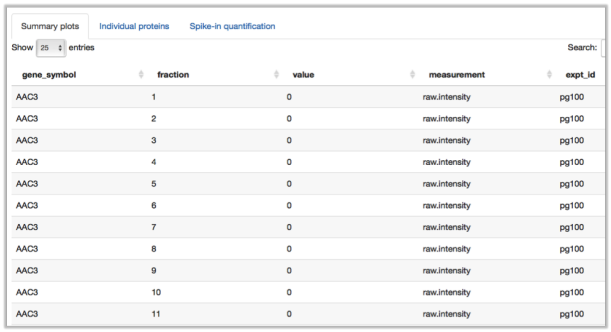


The field "Select UniProt ID" will be automatically populated depending on the experiment ID. You can delete individual entries, but you cannot add spike-ins. Use the "Individual proteins" plots for assessing individual protein profiles.

6 Tables



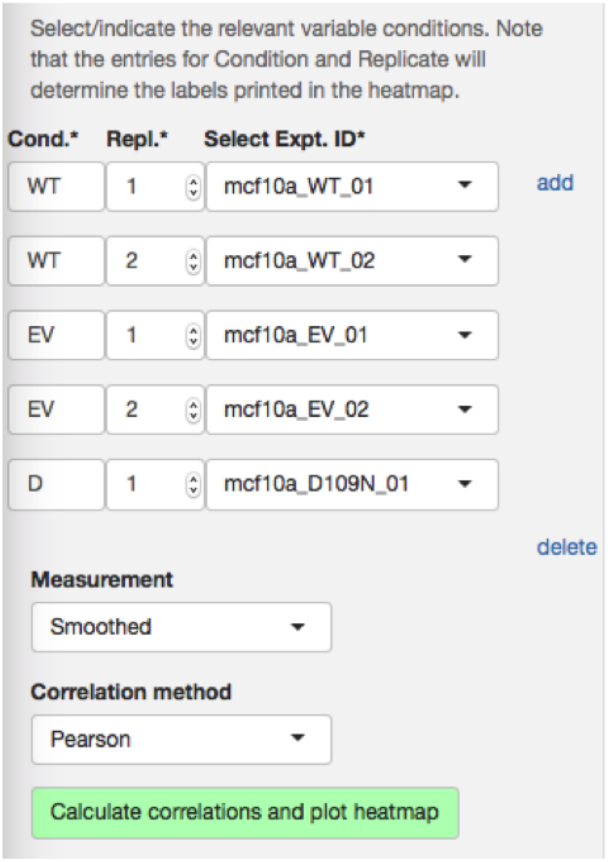
The tables section allows you to see the values that underlie the summary plots and individual protein profiles. Conversely, if the visualization section is empty (e.g., because you haven't selected an experimental ID), the corresponding table section will be empty, too. The following example shows the table for the summary plot.



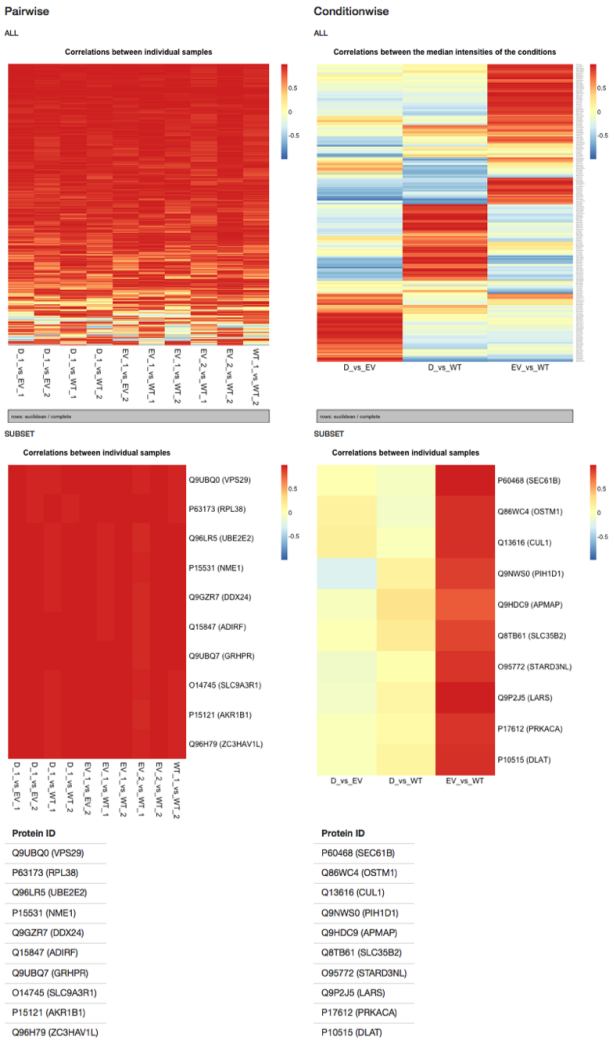
Each row corresponds to *one* fraction per protein per experiment ID, which is different from `MaxQuant`'s table where each row corresponds to *all* fractions per protein and experiment ID.

7 Correlations

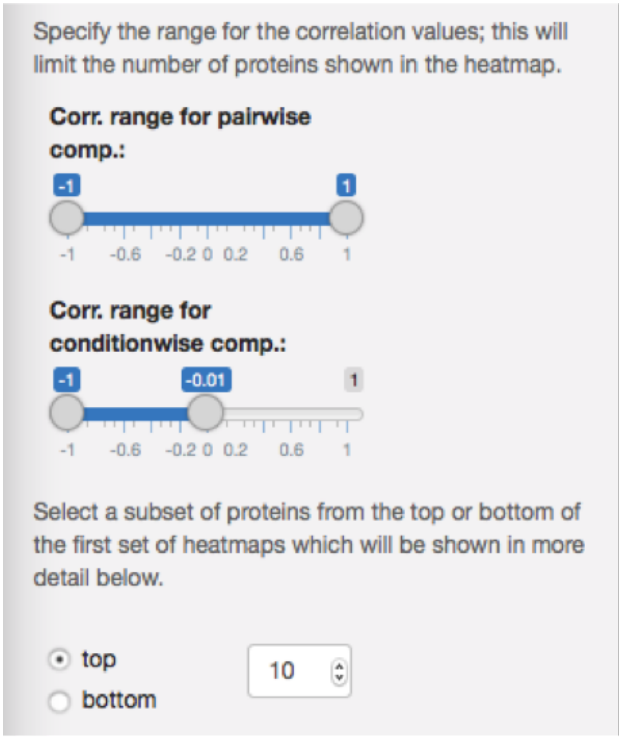
Define Condition (Cond) and replicate numbers (Repl.) that correspond to the experimental ID.



Currently saving has to be done via right-mouse click



7.1 Reducing the numbers of displayed proteins



Limit the range of correlation values

specify the number of proteins shown in the subsets

7.2 Defining proteins of interest for protein profile plots

References

[1] CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, **38**(SUPPL.1), 2009. doi:10.1093/nar/gkp914.

[2] Benschop JJ, Brabers N, van Leenen D, Bakker LV, van Deutekom HWM, van Berkum NL, Apweiler E, Lijnzaad P, Holstege FCP, and Kemmeren P. A consensus of core protein complex compositions for Saccharomyces cerevisiae. *Molecular Cell*, **38**(6):916–928, 2010. doi:10.1016/j.molcel.2010.06.002.

[3] Blanc C and Schlick C. X-splines: A spline model designed for the end-user. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 377–386, 1995. doi:http://doi.acm.org/10.1145/218380.218488.

[4] Cox J and Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, **26**(12):1367–1372, 2008. doi:10.1038/nbt.1511.

[5] Friedman JH. A Variable Span Scatterplot Smoother. Tech. Rep. October, Stanford University, 1984.

[6] Kirchner M and Selbach M. In vivo quantitative proteome profiling: Planning and evaluation of SILAC experiments. *Methods in Molecular Biology*, **893**:175–199, 2012. doi:10.1007/978-1-61779-885-6{_}13.

[7] Pu S, Wong J, Turner B, Cho E, and Wodak SJ. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, **37**(3):825–31, 2009. doi:10.1093/nar/gkn1005.