VIQoR

Manual version 1.0

Vasileios Tsiamis [*, 1], Veit Schwämmle [1]

[1] Department of Biochemistry & Molecular Biology and VILLUM Center for Bioanalytical Sciences, University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark.

* Contact: vasileios@bmb.sdu.dk

What is VIQoR?

VIQoR, is a user-friendly web service for **V**isually supervised protein **I**nference and protein **Q**uantification implemented in **R**. The Shiny web interface integrates all the post-identification processes involved in protein inference and relative protein abundance summarization, along with smart and novel interactive visualization modules to support the common researchers with a straight-forward tool for protein quantification, data browsing and data inspection. The VIQoR's visualization modules additionally support modified peptides. Two parsimonious algorithms are implemented to solve the protein inference problem, while protein summarization is facilitated by a factor analysis method called fast-FARMS followed by a weighted average summarization function that minimizes the effect of missing values [1,2].

The user interface explicitly separates the tasks carried out by the analysis and arranges them in 5 tabs. These are the '*Input*', '*Conditions and Samples*', '*Modifications and Normalization*', '*Protein Inference*' and '*Protein Quantification'* tabs. Each tab consists of a sidebar to let the user operate with the tool and a main body to render the data inspection and visualization modules.

***Hint**: Hover the mouse over the following symbol distributed in the user interface for further information and tips.*

Input Tab

In '*Input*' tab the user can import the data required for the analysis, inspect them in interactive tables and manually validate them.

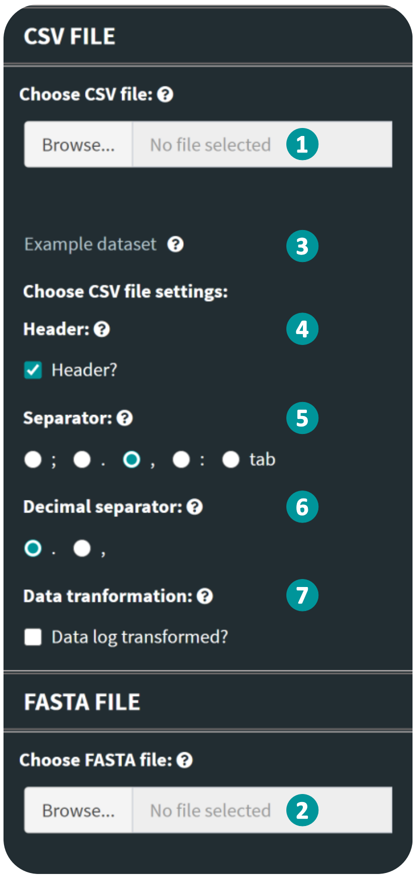


Figure 1: Sidebar of the 'Input' tab.

1 Click '*Browse*' and import the quantitative peptide or PSM report. The file should be in *.csv* or *.txt* format. The data can be of any type of proteomics experiment, either labeled or label-free. The first column of the data table should contain the identified peptide/PSM amino acid sequences, that should not necessarily be unique. The sequences additionally can contain modifications. PTMs can be annotated on the sequences by the name of the modification within parenthesis or brackets. For instance, the tryptic peptide CESGGFLSK with a phosphoserine at the third position of the sequence can be annotated as CE(ph)SGGFLSK or CE[ph]SGGFLSK.

The rearmost columns of the table, that individually correspond to a single biological sample, should contain the peptide/PSM intensities in either linear or logarithmic scale. The minimum requirement for protein abundance summarization is 4 samples and consequently 4 at least quantitative columns must be contained in the data table. In case of replicated experiments, each condition should have the same number of replicates. Additionally, any number of missing measurements (NA values) is allowed in the file.

Between the sequence identifier column and the quantitative columns any number of columns may intermediate, but they are not considered in any computational process.

Note: *the abundances of the modified peptides are used only for visualization purposes.*

2 Click '*Browse*' to import the protein sequence database. Preferably the one that have used for the peptide identification. The file should be in *.fasta* format and can contain entries of any NCBI identifier type. The file may consist of sequences of a proteome or a mixture of proteomes or a set of selected proteins that will serve as a reference in protein inference.

3 Click the '*Example dataset*' action link to load the *example_data.csv* and *example_fasta.fasta* files provided as an example. The peptide report is taken from a hybrid proteome label-free study of 2 conditions (HYE110_A and HYE110_B) in triplicates with known protein concentration ratios of 1:1, 10:1 and 1:10 for human, yeast and E. coli respectively [3]. Additionally, 4 modified peptides have been added artificially in the dataset as an example for PTMs visualization modules. The modified peptides correspond to E. coli protein P0A6Y8 and are the following:

- TIAVYDLGGG(Phosphorylation)TFDISIIEIDEVDGEK (concentration ratio 1:1)
- LINYLVEEF(Acetylation)K (concentration ratio 10:1)
- QVEEAGD(Acetylation)K (concentration ratio 10:1)
- NTTIPT(Acetylation)K ((concentration ratio 10:1)

The protein sequence database contains the proteomes of the three species. Preselected settings for the analysis are provided in all following tabs.

Note: *the example dataset, and more specifically the protein P0A6Y8 is used in all the examples presented in this manual.*

4 Check the box if the imported *.csv* file provides the column identifiers in the first row.
5 Select the character that separates the values in the imported *.csv* file.
6 Select the character used as decimal separator for the values in the imported *.csv* file.

***Hint**: when in doubt about the .csv file settings open and inspect the file in notepad or any text editor.*

***Note**: in case of wrong settings for the imported .csv file the application will not report any error but will not display the imported data in the body of Input tab.*

❼ Check the box if the peptide intensities are already log2-tranformed.

Once both files are imported and loaded correctly, the interactive '*Peptide abundance table*' and '*Fasta file table*' will unfold in the body of the '*Input tab*' (*Figure 2*).

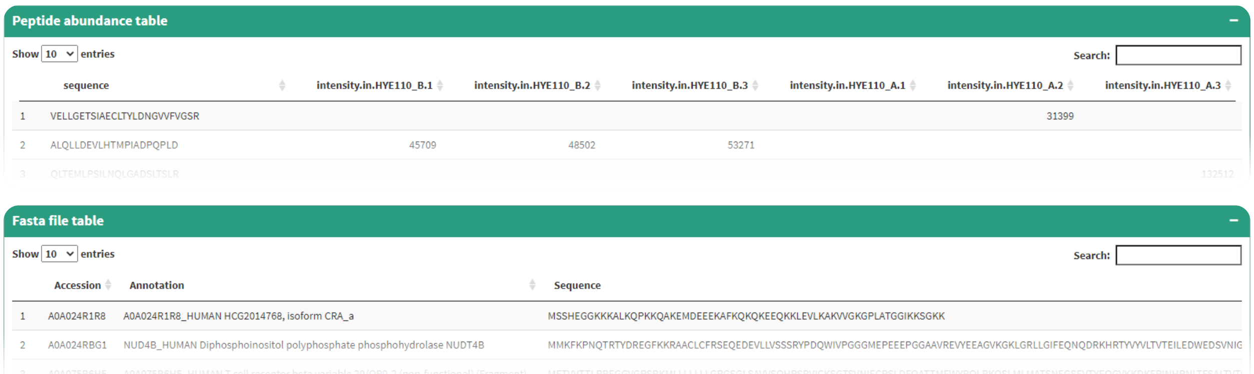
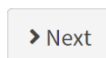


Figure 2: Body of the 'Input tab'. Interactive tables of the imported .csv and .fasta files.

To continue to '*Conditions and Samples*' tab click 

***Hint**: the button to proceed to 'Conditions and Samples' tab will not appear if the input data are not loaded correctly.*

***Note**: To go back to Input tab, click the header of the tab located at the sidebar menu.*

Conditions and Samples tab

In '*Conditions and Samples*' tab the user can specify the experimental design of the study, select the reference condition to enable relative protein summarization and perform missing value filtering. Multiple measurements of the same peptide/PSM are summarized as the sum of all intensities. Additionally, due to identical molecular mass, the Isoleucine (I) is substituted by Leucine (L) in all the sequences to ease the protein inference process.

1 Use the slider to specify the number of quantitative columns in the imported report. The value should correspond to the total number of samples in the experiment. Make sure that the background of the '*Conditions and Sample table*' have changed to pink color for the columns corresponding to the quantitative data, as it is shown in *Figure 4A*.

2 Once the number of samples is set, the range of values in the following slider is readjusted. Use the second slider to specify the total number of conditions in the experiment. Make sure that the samples in the '*Conditions and Samples table*' have grouped according to the condition they belong (different background color), as it is illustrated in *Figure 4B*.

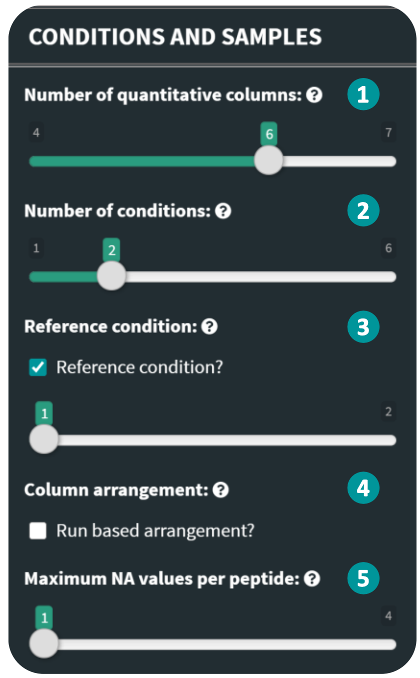


Figure 3: Sidebar of the 'Conditions and Samples' tab.

3 Check the box if one of the conditions is used as a reference. Once the box is checked, select the reference condition in the appearing slider. The relative peptide abundances are calculated relatively to the average of the peptide abundances of the samples of the selected condition. The selected condition is highlighted with yellow colored background in the '*Conditions and Samples table*', as it is shown in *Figure 4C*. If the box is not checked, the relative peptide abundance is calculated relatively to the average of the peptide abundances of all samples.

4 Check the box if the samples in the peptide/PSM report are arranged based on the replication indices (Example: C1_R1, C2_R1, C1_R2, C2_R2 where C stands for conditions and R for replicate). If the samples are arranged based on the conditions (Example: C1_R1, C1_R2, C1_R1, C2_R2) keep the box unchecked.

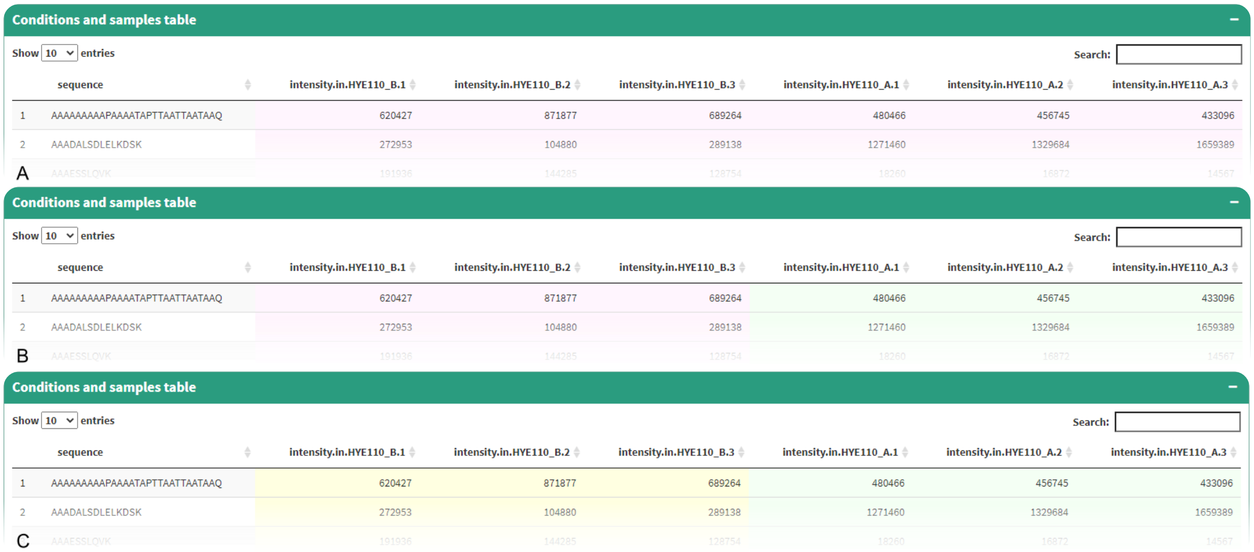


Figure 4: Body of the 'Conditions and Samples' tab. A – 6 quantitative columns are selected (pink). B – 2 conditions are selected with 3 replicates each (pink and green). C – The first condition is selected as a reference (yellow).

5 Use the slider to select the maximum number of missing abundance values allowed per peptide entry. Peptides with more missing values than the selected number are removed. The range of values in the slider is readjusted based on the number of quantitative columns, with minimum of 1 and maximum of (number of samples) - 2. Consequently, peptides with less than 3 valid measurements are removed by default.

***Note**: abundance values of 0, -Inf and Inf are replaced by missing values (NA). This substitution should be considered during missing value filtering.*

To continue to '*Modifications and Normalization*' tab click › Next

***Hint**: the button to proceed to 'Modifications and Normalization' tab will not appear if the selected numbers of quantitative columns and conditions are not compatible. For instance, 7 total samples and 2 conditions do not fulfill the requirement of equal number of replicates per condition.*

***Note**: To go back to 'Conditions and Samples' tab, click the header of the tab located at the sidebar menu.*

Modifications and Normalization tab

In '*Modifications and Normalization'* tab, the user can select the types of modifications to be separated for PTMs visualization analysis and perform log2-tranformation and per sample zero-center normalization of the peptide abundances.

1 Check the box(es) of the modification type(s) that should proceed to PTMs analysis and click the '*Choose!*' button. The sidebar interface for the modification type selection is dynamic and depends on the modification types found in the imported dataset. If the dataset contains only unmodified peptides, the message '*No modifications were detected*' appears in the sidebar instead. The modified peptides of the selected types are separated and used only for visualization purposes. The modified peptides of unselected types are treated as unmodified and proceed for protein inference and summarization along with the other unmodified peptides.

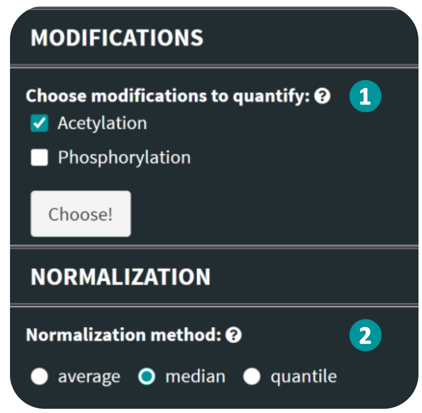


Figure 5: Sidebar of the 'Modifications and Normalization' tab.

Note*: Modified peptides that are not separated, are aggregated with their counterparts to maintain unique entries for each peptide sequence.*

Note*: the PTM annotations on modified peptide sequences that are not separated in 'Modification and Normalization' tab, are not displayed in 'Protein Inference' and 'Protein Quantification' tabs.*

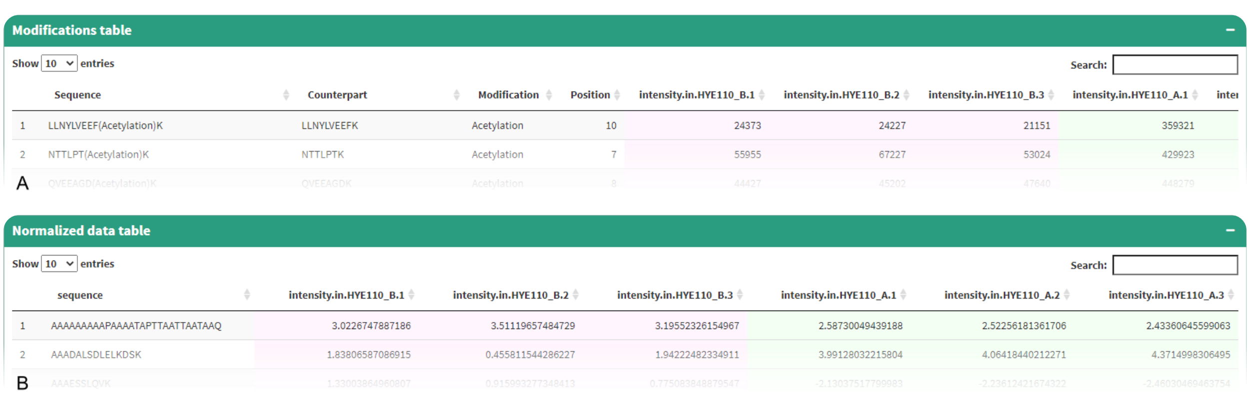


Figure 6: The interactive 'Modifications table' and 'Normalized data table' in the body of the tab.

2 Select the normalization method. The log2-transformed peptide abundances can be zero-center normalized by the average, median or quantile. Alternation of the normalization method will update the result automatically.

The peptides that contain the selected modification types are displayed in the '*Modifications table*' (*Figure 6A*). Each entry in this table corresponds to a single modified peptide and provides information about the modified sequence, the counterpart sequence (without the PTMs), the modification type(s) and the site(s) on the peptide sequence. The normalized abundances are displayed in the '*Normalized data table*' (*Figure 6B*). To download the data displayed in the tables, click the download button Download located under each table.

To continue to '*Protein Inference*' tab click › Next

Hint*: the button to proceed to '*Protein Inference*' tab will not appear if the modification types are not separated ('Choose!' button).*

Note*: To go back to 'Modifications and Normalization' tab, click the header of the tab located at the sidebar menu.*

Protein Inference tab

In '*Protein Inference*' tab the user can assemble proteins from the peptides found in the imported dataset, or in other words, to perform a protein identification process. The peptide sequences are mapped by the PeptideMapper [4] on the protein sequences of the imported *.fasta* file. Based on that mapping, a bipartite graph is constructed, where peptide nodes are connected to protein nodes to denote their relation. For each connected component of the graph, the user can apply different inference approaches to handle degenerate peptides and report a list of protein groups. It is referred to as a protein group because the implemented methods promote the merging of overlapping proteins and therefore a protein group can contain more than one protein accession.

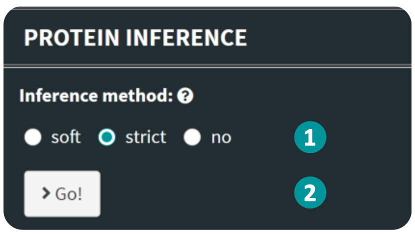


Figure 7: Sidebar of the 'Protein Inference' tab.

1 Select the protein inference method. The *Soft* and *Strict* approaches rely on the principle of parsimony. For a given connected component, parsimonious algorithms iteratively report the proteins related to the most peptides, until the presence of all peptides is explained. The two implemented methods extend the classic parsimony with the addition of criteria to address degenerate peptides.

The *Strict* parsimony groups together the proteins that share the same peptides. The result is a minimal protein group list that explains the presence of all peptides in the connected component uniquely.

The *Soft* parsimony, groups together two proteins if the peptides related to the latter is a subset of the peptides related to the former. In addition, the degenerate peptides assigned already to a protein group can be shared with all the following protein groups that are related to these peptides. Consequently, the result is a minimal protein group list that explains the presence of all peptides in the connected component at least once.

The third approach (*No* parsimony) does not rely on the principle of parsimony but infers proteins directly from the connected components. The result is a maximal protein group list that explains the presence of all peptides in the connected component at least once.

Note*: the ability of the three methods to distribute the degenerate peptides to proteins is differentiated by a spectrum of 'strictness'. Strict parsimony does not allow the sharing of such peptides, while Soft* *parsimony handles them with more tolerance yet in a more conservative way compared to the No parsimony approach. Additionally, No parsimony will report more protein groups.*

2 Click the button › Go! to start the protein inference analysis. A progress bar (*Figure 8*) appears at the bottom right corner of the tab to indicate the proportion of connected components that have been analyzed.

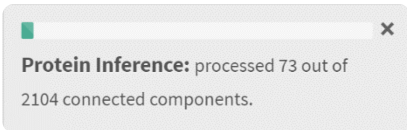


Figure 8: Progress bar during protein inference analysis.

***Hint**: if the progress bar remains unchanged for some time, be patient, it might take some time. Connected components that are very large need more time to be processed.*

Once the protein inference is completed the inferred protein group list is displayed in the '*Protein inference table*' located in the tab's body (*Figure 9A*). The protein accessions of a protein group but also the peptide sequences assigned to a group are separated by a vertical slash (|).

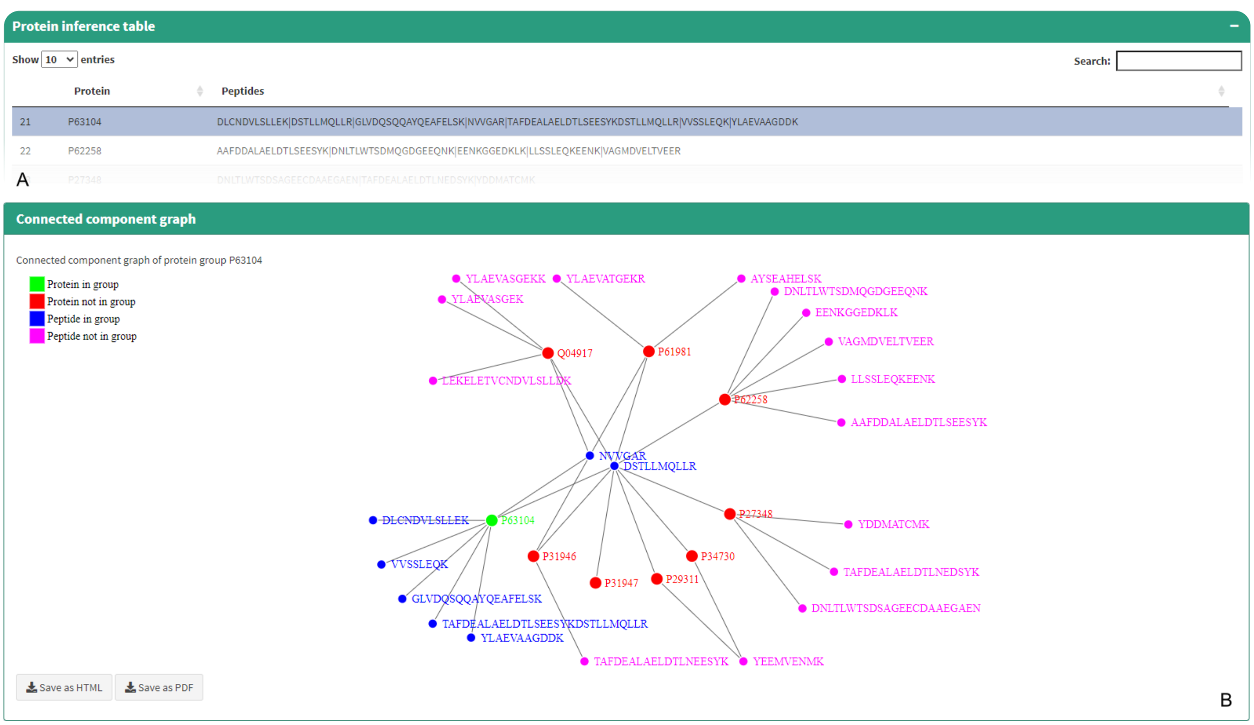


Figure 9: The 'Protein inference table' and 'Connected component graph' in the body of the 'Protein Inference' tab. Example for the protein group P63104.

The protein accessions in a protein group are ordered based on the number of peptides they are related to. When two proteins in a protein group are related to the same number of peptides, they are ordered alphabetically. Similarly, the peptide sequences assigned to a protein group follow an alphabetically order too. The inferred protein groups can be downloaded in *.csv* format by the button [Download] located under the table.

Connected component graph

Click on any protein group entry in the interactive '*Protein inference table*' to visualize the corresponding '*Connected component graph*' (*Figure 9B*) in a panel that automatically unfolds under the table. The protein nodes of the selected protein group are colored green and the peptide nodes blue. The nodes that do not belong to the selected group are colored red and pink, respectively for the proteins and the peptides. In this graph the user can inspect the similarities between protein groups on the same connected component and observe how the degenerate peptides are distributed (peptide nodes connected to multiple protein nodes). To download the graph as interactive graphics in *.html* format or as vector graphics in *.pdf* format click [Save as HTML] or [Save as PDF]

To continue to '*Protein Quantification*' tab click [> Next]

***Hint**: the button to proceed to 'Protein Quantification' tab will not appear unless the protein inference is completed.*

***Note**: To go back 'Protein Inference tab', click the header of the tab located at the sidebar menu.*

Protein Quantification tab

In the '*Protein Quantification*' tab the user can perform relative protein abundance summarization. For each inferred protein group, a factor analysis is applied on the normalized and relatively expressed abundances of the corresponding peptides to access their quality [1,2]. The analysis assigns individual weights to the peptides, which denote their coherence to the protein they belong. The weight values vary between 0 and 1. Low weights correspond to unrepresentative peptides and high weights to coherent peptides. The relative protein expression is then calculated as the weighted average of the relative peptide measurements. To remove unrepresentative peptides, a peptide weight threshold can be applied to eliminate their contribution during the summarization.

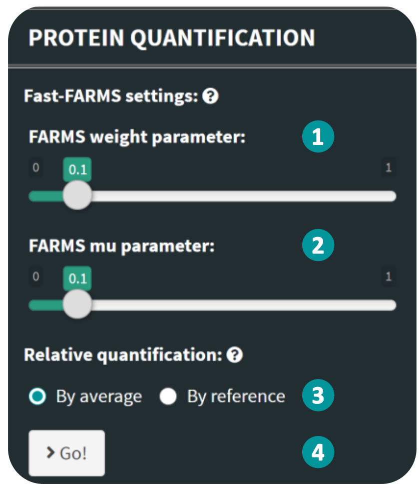


Figure 10: Sidebar of the 'Protein Quantification' tab.

1 Select the value of the *weight* parameter for the factor analysis. This parameter is a weight applied on the variance of the prior distribution to determine its influence during the factor analysis. Default value is set to 0.1.

Note*: to avoid any confusion, the weight parameter is used during factor analysis, while the peptide weight threshold is used after the factor analysis as a parameter in the protein abundance summarization, therefore they are two different parameters.*

2 Select the value of the *mu* parameter for the factor analysis. This parameter determines the influence of prior knowledge to the analysis. For example, a low *mu* value assumes that the peptide measurements do not contain any signal. It is advisable to keep *mu* value low for proteomics data analysis. The default value is set to 0.1.

3 Select how the relative peptide abundance should be calculated. For each peptide this could be by either the average of the measurements in all samples or by the average of the samples of a selected reference condition. This option is available only if a reference condition is selected in the '*Conditions and Samples*' tab, otherwise the relative abundances are calculated by the average.

4 Click the '*Go!*' button to start the analysis. Progress bars like the one in *Figure 8* appear in the bottom right corner of the tab to track the progress of the peptide weight estimation, protein quantification and protein annotation retrieval.

After the quantitative analysis is completed, the results are displayed in a panel that appears in the body of '*Protein Quantification*' tab (*Figure 11*). The panel has two sub-tabs, the '*Per sample*' and the '*Per condition*' tabs, where the results are presented per sample or per the condition average of samples, respectively (*Figure 11A*).

***Note**: if the number of samples is equal to the number of conditions only the 'Per condition' tab appears. In case the study has a single condition, only the 'Per Sample' tab is available.*

The two tabs of the panel provide the same data inspection and visualization modules. The '*Per sample*' tab additionally provides a slider that enables the alternation of the peptide weight threshold and the Global Correlation Indicator (GCI) module to indicate the optimal threshold value (*Figure 11B*).

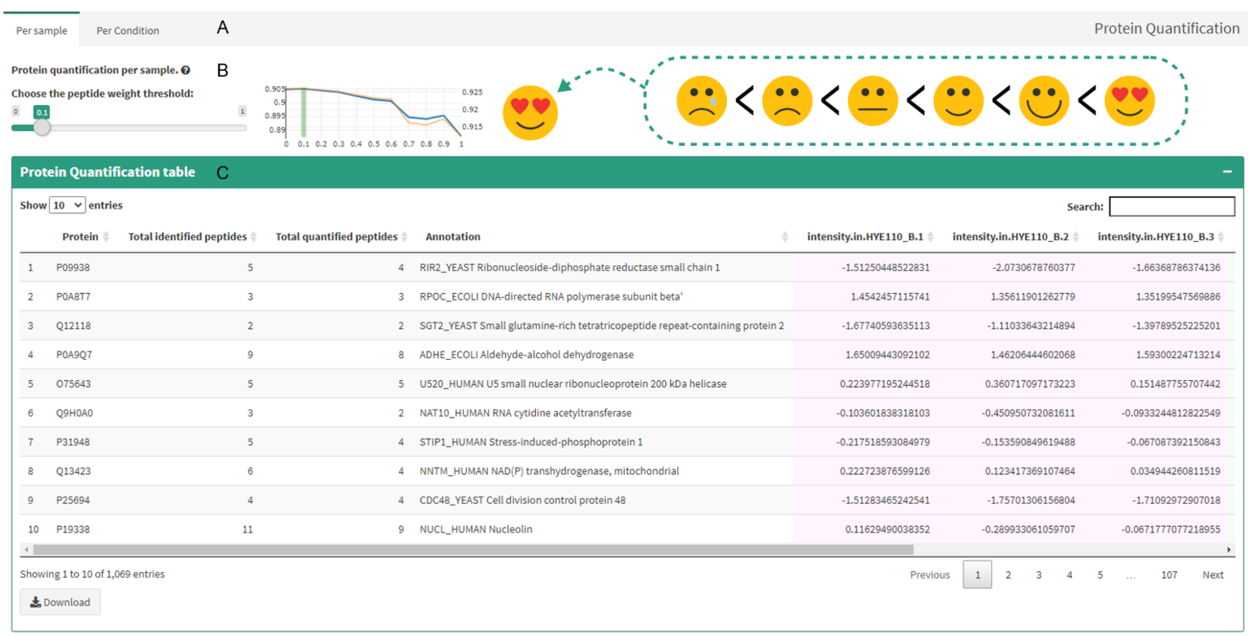


Figure 11: Body of the 'Protein Quantification' tab. GCI (B) and 'Protein Quantification table' (C).

Global Correlation Indicator (GCI)

The GCI is a line plot that summarizes the pairwise Pearson correlation coefficients of the protein expressions for samples within conditions (Blue trace) and between conditions (Orange trace), according to the experimental design set in '*Condition and Samples*' tab, and for all possible peptide weight threshold values [0, 0.1, … 1]. The weight threshold that maximizes the correlation within conditions trace (Blue) in the CGI module, results in the optimal protein summarization. Additionally, a series of smiley icons are provided to ease the parameter tunning visually. The 'heart eyes' icon indicates the optimal peptide weight threshold while the 'crying face' icon corresponds to the weight value that minimizes the CGI trace. For instance, in *Figure 11B*, the selected threshold value is 0.1 and is the optimal one according to the smiley icon.

Note*: alternation of the peptide weight threshold updates all the inspection and visualization modules in both 'Per sample' and 'Per condition' tabs according to the selected value.*

The summarized relative protein abundances of the inferred protein groups (except the one-hit wonders i.e., the single peptide protein groups) and for the selected peptide weight threshold are displayed in the '*Protein Quantification table'*. For each protein group entry, the table provides information about the protein group accessions, the number of the peptides assigned to the protein group (*Total identified peptides*), the number of the peptides that are not eliminated and are considered in protein summarization (*Total quantified peptides*), the description of the first protein accession in the group (*Annotation*) and the protein expressions for all samples/conditions (*Figure 11C*). To download the table in *.csv* format click the button Download

Note*: a missing value in the 'Protein Quantification table' means that all the corresponding peptides are not found in that sample/condition (NA).*

Hint*: to find a specific protein group, use the search bar located at top right corner of the table.*

***Hint**: the data can be sorted by a specific column. Click the identifier of a column of interest once or twice to sort in increasing or decreasing order.*

Protein expression correlation heatmap

The '*Protein expression correlation heatmap*' is located under the '*Protein Quantification table*' (*Figure 12*). It provides the pairwise Pearson correlation scores of the protein group expression profiles between all samples/conditions. The heatmap is combined with a hierarchical clustering dendrogram. The user can select the clustering method (Ward's, Single linkage, Complete linkage, UPGMA or WPGMA), the distance measurement to use for the clustering (Euclidean, Maximum, Manhattan, Canberra, Binary or Minkowski) and the number of the desired clusters (minimum of 1 to maximum equal to the total number of samples/conditions or Any i.e., the optimal number of clusters). Additionally, there are 4 different color gradients available to select. The heatmap is a Plotly object and offers all the interactive functionalities of Plotly.

To download the heatmap as vector graphics in *.pdf* format select the desired width and height and click the button Download

***Note**: for all graphics available to download, the maximum width and height is set to 1000 pixels. Values higher than 1000 will not render the complete graphics in the .pdf file.*

Figure 12: The 'Protein expression correlation heatmap' in the 'Protein Quantification' tab.

The user can select any entry in the interactive '*Protein Quantification table*'. For the selected protein group, three visualization modules unfold under the '*Protein expression correlation heatmap*'. These are the '*Protein group line plot*', the '*Protein group VIQoR plot*' and the '*Peptide abundance correlation heatmap*'.

Protein group line plot

In the '*Protein group line plot*', the summarized abundance of a selected protein group over all samples/conditions is visualized along with the zero-center normalized and relatively expressed abundances of the corresponding peptides (*Figure 13*). The user can inspect the peptide abundance trends over the different samples/conditions and how those influence the summarized protein expression. The peptides are displayed with different colors and their opacity depends on the assigned peptide weights. The eliminated peptides are colored grey and illustrated with dashed lines. For instance, in *Figure 13* the line plot of the selected protein group P0A6Y8 is presented for a peptide weight threshold of 0.1. The peptide with sequence AVLTVPAYFNDAQR that is following an opposite concentration change compared to the rest of the peptides is eliminated and its grey dashed line trace fades. The line plot is implemented in Plotly and maintains the interactive functionalities of Plotly. Hover the mouse over the traces for additional information.

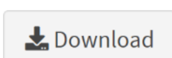
To download the line plot as vector graphics in *.pdf* format select the desired width and height and click the button 

Figure 13: The 'Protein group line plot' in the 'Protein Quantification' tab.

Protein group VIQoR plot

The '*Protein group VIQoR plot*' for the selected protein group unfolds under the '*Protein group line plot*'. VIQoR plot combines quantitative data with amino acid sequence information, to reveal the expression patterns of modified and unmodified peptides projected on top of the protein expression changes. The vertical axis shows the log2-fold changes of the protein and the peptide relative abundances between two comparing samples/conditions. In the horizontal axis the modified and unmodified peptide sequences (green segments for unmodified and pink segments for modified) are aligned on the sequence of the first protein in the selected group (blue segment). Additionally, the individual PTMs are annotated on the peptide segments according to their site. The eliminated peptides due to the peptide weight threshold are colored grey. The opacity of all the segments depends on the assigned weights. VIQoR plot is implemented with Plotly and therefore any interactions of the mouse with the components of the graph can reveal hover labels with additional information regarding the amino acid sequences, the assigned weights, the position of peptides on the protein sequence and the modification sites.

Hint*: the end parts of the unmodified peptide segments are colored with the same colors as their traces in the* '*Protein group line plot*'.

Hint*: the PTM annotations on the modified peptides are colored according to the modification type. For example, all phosphorylation sites are colored with the same color.*

Note*: the VIQoR plot will not be generated if the protein expression of one of the two comparing samples/conditions is missing (NA). Similarly, the segments of peptides that are not found in both comparing states are not visualized.*

For example, *Figure 14* illustrates the VIQoR plot for the protein group P0A6Y8. The two comparing samples are the first replicate of the condition HYE110_B (as denominator/reference) and the first replicate of the condition HYE110_A (as nominator/testing). The unmodified peptides with weights higher than 0.1 (green segments) fall within the expected log2-fold change of around -3, that corresponds to the real E. coli concentration ratio of 1:10. The peptide with sequence AVLTVPAYFNDAQR, is eliminated similarly as in the '*Protein group line plot*' and is colored grey. The three artificially added acetylated peptides (pink segments with orange annotations) with known concentration ratio of 10:1 have log2-fold change of around 3. The artificially added phosphorylated peptide (pink segment with purple annotation) remains unchanged between the two samples due to the known 1:1 concentration ratio.

To download the VIQoR plot as vector graphics in *.pdf* format select the desired width and height and click the button Download

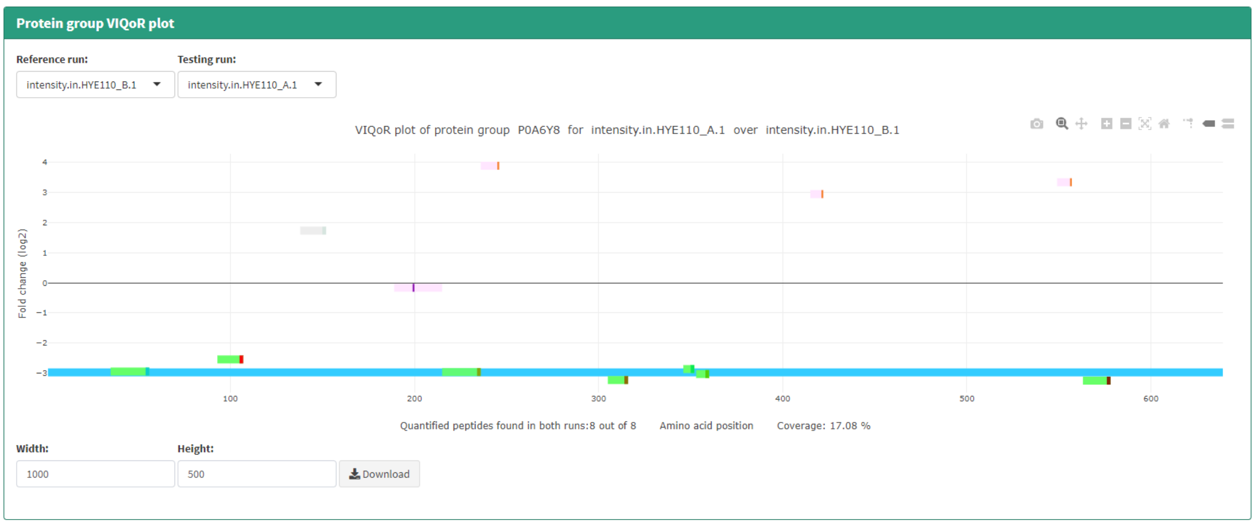


Figure 14: The 'Protein group VIQoR plot' in the 'Protein Quantification' tab.

Peptide abundance correlation heatmap

The '*Peptide abundance correlation heatmap*' for the selected protein group unfolds under the '*Protein group VIQoR plot*' (*Figure 15*). This heatmap illustrates the pairwise Pearson correlations of the peptides of the selected protein group. The user can observe the similarities between the peptides and inspect the incoherent peptides with low correlation scores. The heatmap provides the same functionalities as the '*Protein expression correlation heatmap*'. *Figure 15* illustrates the heatmap for the peptide profile of the protein group P0A6Y8 for weight threshold of 0. The peptide with sequence AVLTVPAYFNDAQR appears to follow an opposite expression compared to the other peptides.

To download the heatmap as vector graphics in *.pdf* format select the desired width and height and click the button Download

***Note**: the selected P0A6Y8 protein group is regulated therefore in the 'Peptide abundance correlation heatmap' the coherent peptides are highly correlated. For a non-regulated protein, the correlation scores are expected to be less homogeneous, since the variance introduced to the peptide abundances is noise.*

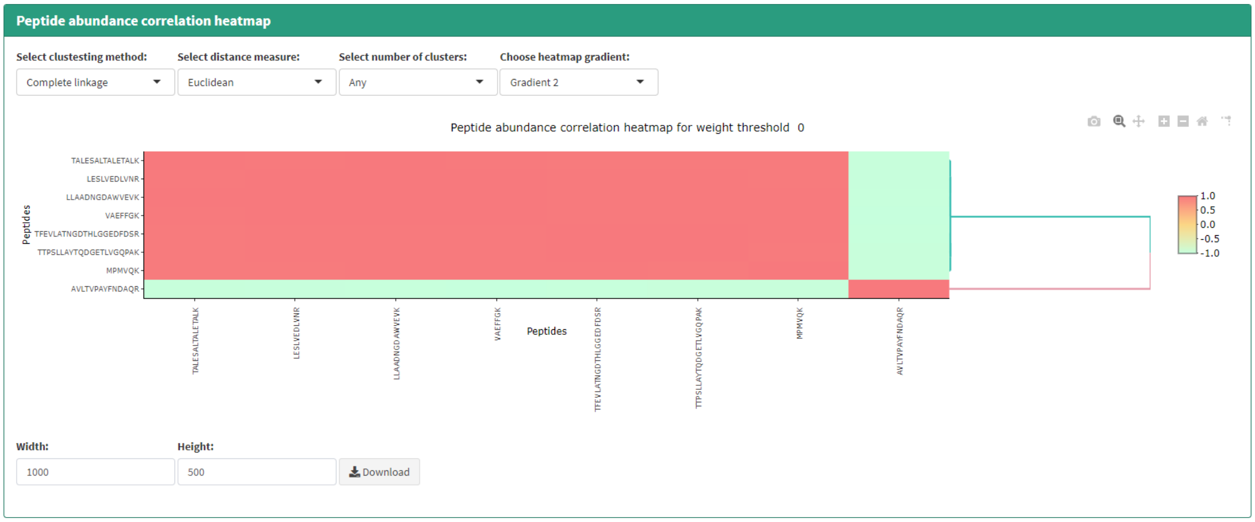


Figure 15: The 'Peptide abundance correlation heatmap' in the 'Protein Quantification' tab.

References

1. Hochreiter S, Clevert DA, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics. 2006 Apr 15;22(8):943-9.

2. Zhang B, Pirmoradian M, Zubarev R, Käll L. Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences. Mol Cell Proteomics. 2017 May;16(5):936-948.

3. Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, Tate SA, Tsou CC, Reiter L, Distler U, Rosenberger G, Perez-Riverol Y, Nesvizhskii AI, Aebersold R, Tenzer S. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 2016 Nov;34(11):1130-1136.

4. Kopczynski D, Barsnes H, Njølstad PR, Sickmann A, Vaudel M, Ahrends R. PeptideMapper: efficient and versatile amino acid sequence and tag mapping. Bioinformatics. 2017 Jul 1;33(13):2042-2044.